INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Open Source Business Intelligence

João Nuno da Silva Bernardino Martins Severino

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores (MEIC)

Júri

Presidente:	Prof. Ana Almeida e Paiva
Orientador:	Prof. Andreas Miroslaus Wichert
Vogais:	Prof. João Pereira

Setembro 2008

Resumo

Com a crescente competitividade do mercado, as empresas sentem a necessidade de recolher mais informações sobre o seu modo de funcionamento e analisá-lo para produzir informações úteis. Por esta razão a procura por soluções que suportem o processo de tomada de decisão, como é o caso de Business Intelligence, está a aumentar.

A arquitectura genérica de Business Intelligence (BI) é composta por uma componente de base de dados, ferramentas de integração de dados, motores analíticos e por ferramentas de geração de relatórios, por isso não é de estranhar que só o custo das licenças de software atinja valores proibitivos para muitas empresas. Como alternativa estas recorrem a ferramentas já existentes na empresa, como folhas de cálculo ou módulos de relatórios de sistemas transaccionais. Actualmente estas empresas já têm à sua disposição um variado número de aplicações open source para todos os componentes da arquitectura tradicional de um sistema de Business Intelligence. No entanto, o uso destas ferramentas ainda não é pratica comum.

O objectivo desta tese é de efectuar o levantamento e analisar os projectos de open source business intelligence (OSBI) existentes, e distinguir aqueles que mais garantias oferecem.

Para alem da analise das principais ofertas BI no mercado das ferramentas open source, foi desenvolvido um sistema de business intelligence composto apenas por ferramentas open source e que prova que o open source bi é uma alternativa viável às ferramentas comerciais.

Palavras-Chave:

Business Intelligence (BI), Open Source, Sistemas de Apoio à Decisão (SAD), Data Warehouse (DW), Extract Transform and Load (ETL), Data Mining, Dashboards, On-line Analytical Processing (OLAP), Relatórios.

Abstract

Recently, due to the increasing focus on information as a valuable resource and to the increasing market pressures, organizations are building Business Intelligence (BI) infrastructures to help them achieve “better" business decisions by making precise and relevant information available to the point of action.

The traditional Business Intelligence (BI) architecture requires a database, ‘Extract, Transform and Load’ (ETL), on-line Analytical Processing (OLAP) and Reporting tools, therefore the initial licensing cost can easily extend to inhibitive values for many businesses. As an alternative, organizations delegate the information analysis to office productivity software such as spreadsheets or to reporting modules of core transactional applications. However, Open Source Business Intelligence (OSBI) tools such as ETL, OLAP or Reporting tools are available to organizations at very low or no cost,. Since Open Source software is a recent trend, the use of such OSBI tools is not yet common, therefore it is of interest to analyze and evaluate the existing OSBI solutions.

The goal of this thesis is to survey the existent open source business intelligence (OSBI) projects, and from a list of selected projects choose the “best in class”.

Additionally to the evaluation of the current open source software offering, a BI system, based on OSBI software, was build to proof that it is possible to implement an end-to-end BI solution using only open source tools and that it could be a viable alternative to the classical commercial software.

Keywords

Business Intelligence (BI), Open Source, Decision Support Systems (DSS), Data Warehouse (DW), Extract Transform and Load (ETL), Data Mining, Dashboards, On-line Analytical Processing (OLAP), Reports.

Acknowledgements

I would like to thank my two supervisors: Andreas Wichert at Instituto Superior Técnico; and João Damásio at Link Consulting SA.

I would specifically like to thank, Ana Montes and Marcelino Moreno for helping me out with practical details during my work, for taking the time to proofread my report and for giving me valuable feedback.

Finally I would like to thank the support provided by my family and closest friends that has been tireless throughout the duration of the thesis.

Table of Contents

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface
BAM	Business Activity Monitoring
BI	Business Intelligence
BIRT	Business Intelligence and Reporting Tools
BPM	Business Process Management
CAS	Central Authentication Service
CDF	Community Dashboard Framework
CRM	Customer Relationship Management
CSV	Comma Separated Values
CWM	Common Warehouse Metamodel
DB	Database
DOM	Document Object Model
DSS	Decision Support System
DW	Data Warehouse
EAI	Enterprise Application Integration
EDA	Event-Driven Architecture
EIS	Executive Information System
EJB	Enterprise Java Beans
EMEA	Europe, Middle East and Africa
ERP	Enterprise Resource Planning
ESB	Enterprise Service Bys
ETL	Extract, Transform and Load
FTP	File Transfer Protocol
GIS	Geographic Information System
GUI	Graphical User Interface
HOLAP	Hybrid On-Line Analytical Processing
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IDC	International Data Corporation
IT	Information Technology
J2EE	Java 2 Enterprise Edition
JDBC	Java Database Connectivity
JDK	Java Development Kit
JSP	Java Server Pages

KETTLE	Kettle Extraction, Transport, Transformation and Loading Environment
KPI	Key Performance Indicator
MDDB	Multidimensional Database
MDX	Multidimensional Expressions
MOLAP	Multidimensional On-Line Analytical Processing
MOF	Meta Object Facility
OLAP	On-Line Analytical Processing
OLTP	Online Transaction Processing
OSBI	Open Source Business Intelligence
OS	Open Source
OSS	Open Source Solution
OTLIS	Operadores de Transportes da Região de Lisboa
PDF	Portable Document Format
PDI	Pentaho Data Integration
PME	Pentaho Metadata Editor
POJO	Plain Old Java Object
QRA	Query, Reporting and Analysis
QoS	Quality of Service
RDBMS	Relational Database Management System
RFID	Radio-Frequency Identification
RIA	Rich Internet Application
ROI	Return on Investment
ROLAP	Relational On-Line Analytical Processing
RSS	Really Simple Syndication
RTF	Rich Text Format
SAD	Sistemas de Apoio à Decisão
SIMIP	Serviço Integrado de Mensagens de Informação ao Passageiro
SGO	Sistema de Gestão de Ocorrências
SMC	Small to Medium sized Company
SMS	Short Message Service
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SVG	Scalable Vectorial Graphics
UML	Unified Modelling Language
XMI	XML Metadata Interchange
XML	Extensible Markup Language

XMLA	XML for Analysis
XLS	Excel Workbook

Chapter 1 – Introduction

In order to keep competitive, organizations gather information that helps them assessing their business environment. However just collecting information is not enough to gain any competitive advantage, it is necessary to analyze that information and provide knowledge (i.e. added value). Business Intelligence is a concept for analyzing collected data that aims to help people achieve “better” business decisions by making available more accurate and relevant information when needed. It is a very popular concept in the software industry and many consulting companies have realized the need for these services.

One of the consulting companies that are offering BI services is Link Consulting SA, and this master thesis was conducted at their local office in Lisboa. Link Consulting SA is a part of the AITEC group and offers Information Technology (IT) Solutions in several sectors.

1.1. Study Background

The BI concept can be roughly comprised of three stages: data collection, data storage and data visualization.

Normally the data collection is done by building a data warehouse that stores data from multiple data sources. These external data sources can be relational databases, text files, spreadsheets or any other data container. The process of transfer data from data sources to the data warehouse is called Extract, Transform and Load (ETL) process. Since that multiple data sources may have multiple data structures to define the same content it is necessary to consolidate this data and load it into the Data Warehouse. This process is known as data cleansing or data profiling.

Once the data is available at the data warehouse is necessary to produce information from it. It is common to structure the warehouse data into a cube, which is a multidimensional data structure. The cube is essentially used to group data by several dimensions and selecting a subset of interest.

Finally, either the multidimensional or relational data can be used to feed the third stage of a BI Solution – Data Visualization. Data Visualization is comprised of tools such as reporting, dashboards and scorecard. This data can also be used with analysis tools like data mining. Data mining is a concept for finding trends and patterns in the data. The concept of data mining is outside the scope of this thesis and will just be discussed in the literature review.

1.2. Motivation

For a long time, Open Source Software (OSS) is being used by IT organizations, however an area that has been poorly adopted, in contrast to other areas, such as Customer Relationship Managers (CRM) or Application Servers.

Given the increasing market competitiveness, companies feel the need to gather more information about their way of operation and analyze it to produce useful information. Therefore the number of organizations in need of OSS for BI and DW is increasing.

Until recently, the variety of available OSS for this area was limited and not comprehensive. The most mature projects were in Relational Databases Management Systems (RDMS), but neither of them was optimized for large-scale databases. Fortunately, a company can choose today from more than twenty open source projects that support each component of BI.

With the increasing demand for open source projects, and with the rise of such solutions, it is necessary to understand the level of maturity of these tools as well as to validate if they can be a viable alternative to commercial software and provide extra value to organizations.

1.3. Goals

The goals of this thesis, is to survey the existent Open Source Business Intelligence (OSBI) projects, and from a list of selected projects choose the "best in class".

Another goal, besides evaluating the current open source offer is to proof that it is possible to implement an end-to-end BI solution using only open source tools and that it could be a viable alternative to any commercial BI software.

1.4. Methodology

To achieve a successful BI implementation it is necessary to involve the project team and its end users in an iterative development process. This is a requirement because the core of a BI system is to condense data into useful information, and consequently it is needed to get the feedback from those that are going to use and extract knowledge from the information presented.

The methodology followed can be breakdown in six phases:

- **Startup:** During this phase, a sketch of the initial architecture and pilot goals are drawn. It is also when business requirements and pilot end-users are identified. The objective of this step is to do some initial architecture "thinking".
- **Analysis:** The result of this step should be an analysis of the data access tools, data models and technical architecture. During this stage, it is usual to **Control and Validate** the work done with Link Consultants.
- **Design:** During this phase, the data warehouse architecture and the technological environment are specified based on the requirements defined at the previous step. Like the previous step, it is usual to **Control and Validate** the work done with Link Consultants.
- **Implementation:** It consists in the development of the system planned in previous stages.

- **Deploy & Test:** In this stage each component build is deployed and tested. Minor bugs were solved “just-in-time”.
- **Control:** This stage occurs in parallel with the previous stages and can be characterized by being the “simulation” of end-users interaction.

1.5. Development Environment

The tools selected for the project were Kettle (Pentaho Data Integration), MySQL, Mondrian (Pentaho Analysis), Pentaho BI Suite and JBoss Application Server. A large part of the practical part of the project was to familiarize with these tools, which was mainly done by reading forums or analyzing existing examples on the open source project webpage.

- Kettle is an easy-to-use graphical tool for ETL process design. With this tool, a data integration model can be developed and executed using a simple drag-and-drop interface.
- MySQL is the most known open source relational database server.
- Mondrian is an On-Line Analytical Processing (OLAP) engine. It contains tools for processing and deploying the cube as well as designing the cube structure, dimensions, aggregates and other cube related entities.
- The Pentaho BI Suite, integrates the above mentioned applications, allows automation of processes by scheduling tasks that are to be performed regularly and is comprised of Reporting and Dashboards engines.
- The JBoss application server is used to deploy the several applications being used.

1.6. Major Contributions

The main goal of this thesis is to prove that it is possible to build a business intelligence solution only comprised of open source tools and that it could be a viable alternative to commercial tools. Therefore it was developed a business intelligence solution for the public transportation area which allows the transport operator or a regional consortium to analyze their business with special focus on quality of service and incident management. This area was chosen because only recently this industry has made efforts to modernize their information and operational systems, for example nowadays operators are migrating from paper tickets to electronic smartcards[1].

The developed prototype consists in a web application, accessible from any device with an internet connection, and a web browser. It was built as a modular system so that any component can be changed for another one or serve as a basis for future work.

[1] *http://www.fertagus.pt/*

Another goal of this thesis is to identify any relevant open source project in the Business Intelligence area and select those that are the best in class. A survey of each OSBI solution was made and since that there are a large variety of choices it was necessary to focus the evaluation only in projects that cover more than one component of a BI architecture – BI Suites. Pentaho BI Suite is currently the one that offers more comprehensive and mature solutions, however during the last months the JasperSoft project has merged with other open source projects and provides now a more flexible application, and even in some components such as data integration posing a serious threat to Pentaho. SpagoBI is now offering a solution as good as the Pentaho one, however it fails to deliver a critical "feature" of any open source project, the active community, comparing with Pentaho the SpagoBi community is almost inexistent.

1.7. Dissertation Structure

This thesis is divided into seven chapters:

- **Chapter 2 – Literature Review:** This chapter concerns to the topic of Business Intelligence with special focus on open source solutions for this area. An introduction to the evolution, concept and goals of Business Intelligence is given, as well as the impact of Open Source on Business Intelligence. An overview about data warehousing and the open source tools available to each stage are presented, as well as some statistics and studies about the overall situation of the Business Intelligent market and trends. Special attention is given to open source vendors, their solutions and contributions for the worldwide market.
- **Chapter 3 – Case Study:** In this chapter it is presented the case study chosen to prove that it is possible to build an open source BI solution. It also described the methodology followed to develop the prototype.
- **Chapter 4 – Architecture:** In this chapter it is presented the prototype conceptual and technical architecture that were designed during the startup and analysis phase. It also contains the description of the high-level architecture idealized.
- **Chapter 5 – Design:** This chapter details the result of each design stage. It contains the data model design for the data staging and data warehouse, as well as the specification of ETL processes and dashboards.
- **Chapter 6 – Implementation:** In this chapter is presented an overview of the prototype developed, plus an analysis of each tool that was used to develop and support the designed architecture.
- **Chapter 7 – Conclusion:** This chapter contains the main conclusions of this Dissertation, emphasizing on the goals achieved and on future developments.

Chapter 2 – Literature Review

2.1. Introduction

Business Intelligence (BI) concept dates since 1958 (1), however it started “officially” in the eighties. In the beginnings was known as Decision Support Systems (DSS) and was limited to large enterprises and used only by a limited group of highly trained users.[2] Alternative solutions emerged, but they were restricted to expensive statistical analysis or standard reporting tools (usually associated with the 132 column green-bar paper). With the rise of applications as Lotus 1-2-3 and later Excel, the Spreadsheets became more popular, however they weren’t connected to corporate data. In order to analyze the company data, users had to, in a time-consuming and error-prone process, re-enter the existing data from the hard-copy reports into new spreadsheets.[3] [4]

In the early 1990s with the emergence of Windows platforms, the DSS became more specialized, with emphasis on graphical displays and easy-to-use user interfaces. This new “version” was known as Executive Information System (EIS), it offered strong reporting tools and drill-down capabilities which helped top-level executives to analyze, compare and highlight trends, simplifying the process of identification the opportunities and problems within the enterprise. (2)

Since those days, the influence of technology and growth of information volume have challenged the industry leaders to improve their performance in order to meet the constantly changing customer demands [5] (3) .In this way, Business intelligence became a powerful instrument that can support them to sustain and improve their competitive position. This is easily observed through the evolution of the Key Performance Indicators (KPI) in the past years.

Today, more companies are implementing BI solutions and assessing their business through the use of Key Performance Indicators. As result, data is becoming available to business faster and more efficiently. Not too long ago, organizations had to wait months for the results, and as a consequence it limited the review of the company’s operational capabilities. Nowadays, many companies can obtain data in real-time, daily, or weekly allowing them to adjust their strategies faster, leading to increased customer satisfaction.

[2] **Eckerson, Wayne.** *Business Intelligence 2006 - Only the Beginning! TDWI. [Online] 21 May 2006. [Cited: 29 November 2007.] http://www.tdwi.org/Publications/WhatWorks/display.aspx?id=7977*

[3] **Power, D.J.** *A Brief History of Decision Support Systems. DSSResources.COM. [Online] 31 May 2003. http://DSSResources.COM/history/dsshistory.html.*

[4] **Howard Dresner.** *The Insights: A Short History of Business Intelligence and Where It's Headed. Hyperion. [Online] [Cited: 02 December 2007.] http://www.hyperion.be/leaders/insights/organization_culture/history_bi.cfm.*

[5] **Wittschen, Lee.** *Why Business Intelligence?*
http://www.businessintelligence.com/print_content.asp?code=29&pagenum=1. [Online] [Cited: 11 November 2007.]

2.1.1. Business Intelligence Concepts

Nowadays and because of the large range of capabilities that a BI system offers, many definitions have emerged. However, I think that the one suggested by Business Intelligence Institute is the more adequate.

"*Business Intelligence (BI) is the management and analysis of vast amounts of information in order to gain valuable insights to drive strategic business decisions, and to support operational processes with new functions.* BI is about managing information that is conclusive, fact based, and actionable. It includes technology practices like data warehouses, data marts, data mining, text mining, and on-line analytical processing."[6]

Currently there are five major dimensions to BI, number likely to change due to the constant evolution of the BI Industry. The dimensions are Reporting, Analysis, Planning, Monitoring and Advanced Analytics.[7]

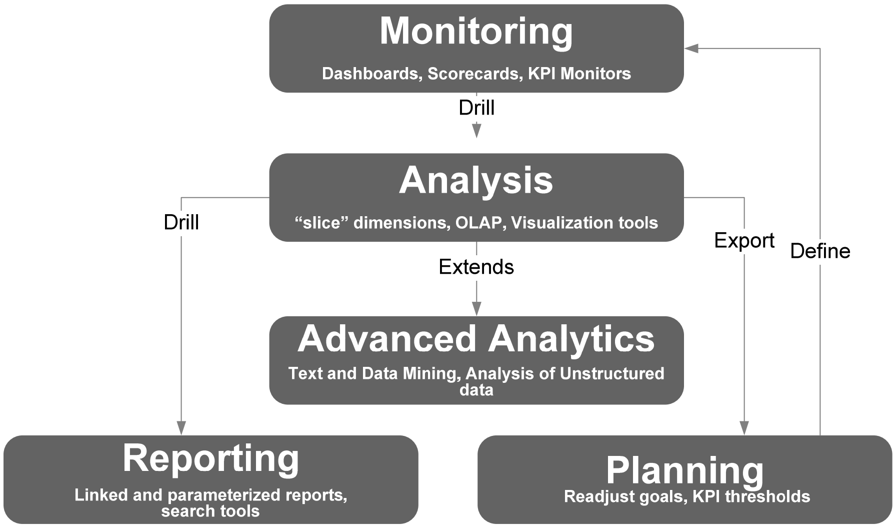

Monitoring
Dashboards, Scorecards, KPI Monitors
Drill
Analysis
"slice" dimensions, OLAP, Visualization tools
Drill
Extends
Export
Define
Advanced Analytics
Text and Data Mining, Analysis of Unstructured data
Reporting
Linked and parameterized reports, search tools
Planning
Readjust goals, KPI thresholds

Figure 1 – Business Intelligence five major dimensions and relationships

[6] *About Us - Business Intelligence Institute. [Online] [Cited: 2 November 2007] http://www.bii.be/belgium/indexaboutus.jsp.*

[7] **Eckerson, Wayne.** *Business Intelligence 2006 - Only the Beginning! TDWI. [Online] 21 May 2006. [Cited: 29 November 2007.] http://www.tdwi.org/Publications/WhatWorks/display.aspx?id=7977*

The image above (Figure 1.) illustrates the five major dimensions and how they relate to each other. From a user perspective the first dimension is monitoring. Most users don't want or have the knowledge to create ad-hoc queries or slice the multidimensional data; they simply want the right data delivered when there is some critical event that they need to examine. In other words, users prefer a customized delivery than a self-service BI. This customized delivery appears as a set of dashboards or other graphic displays with the different key metrics.

If a critical event is raised in the monitoring dimension, the user can drill into the analytical tools and therefore determine the root cause by "slicing" the multidimensional data.

After using the analytical tools, user can drill do the reporting dimension, in order to have a detailed examination of the data. With the reporting tools, the user can better determine what actions to take. In the past the reports were paper-based and reported past activity or events, informing the user of what had already happened. Nowadays with the emergence of Web and Rich Internet Applications (RIA), reporting became more interactive and dynamic, allowing access to linked and parameterized reports.

Finally, the user may export any results to third-party tools to plan, forecast and model activities and to readjust goals or establish new thresholds to the different performance indicators (shown at the monitoring dimension). This dimension combines the Business Performance Management (BPM), a methodology that helps the alignment between the users and processes to organizational strategic objectives, with the BI capabilities to measure and monitor the performance.

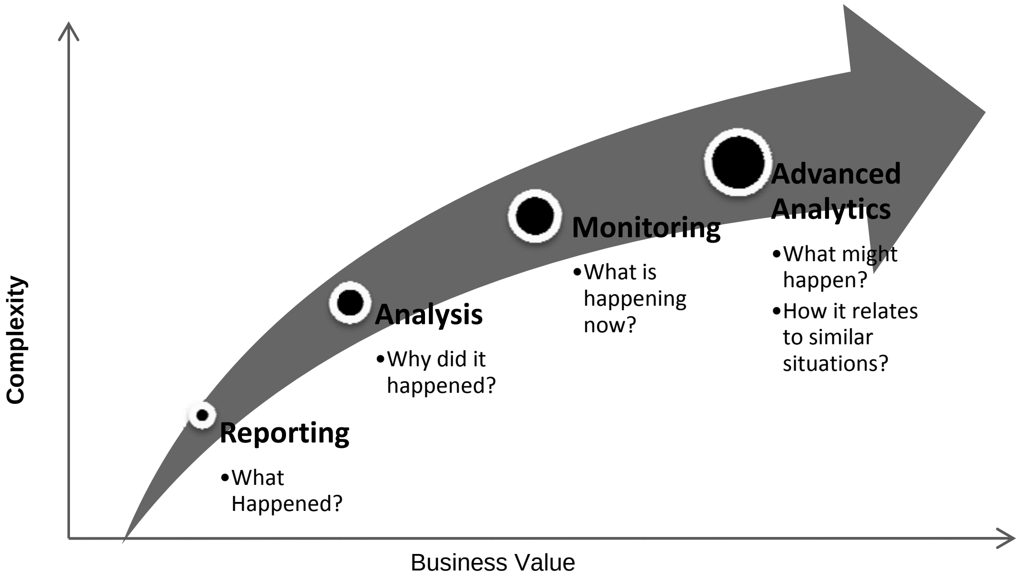


Figure 2 – Business Intelligence dimensions Complexity vs Business Value

The most recent extension to this model is the advanced analytics dimension, it is the layer that provides the most business value but is also the most complex (see Figure 2). It includes data mining, text mining and advanced visualization. These tools help the company in achieving their business targets without requiring statistical experts. Currently, some BI Vendors are also allowing analyzing of unstructured data, like images, video, web pages, documents, etc. that usually constitutes the majority of information within organizations.

This dimension also includes predictive analytics, a set of tools that uncovers relationships and patterns within large volumes of data that can be used to predict behavioural events.

2.1.2. Business Intelligence Goals

The goal of BI is to transform data into useful information and to help not only, companies to become more knowledgeable of the factors that affect their business, but also to make better competitive analysis and business decisions. In other words, BI goal is to improve visibility into business so that business users can react faster to any given situation and to provide a quick, easy, self-service and intuitive access to information [(4)].

In detail, the main objectives of BI, is to help organizations to:

- Manage their business, through metrics that provide insight on productivity, profitability and compliance.
- Grow, by providing useful information to make better decisions.
- Optimize business, through metrics that can help streamline and improve the effectiveness and responsiveness of processes and people.
- Predict problems and opportunities, in order to gain competitive advantage in the company's marketplace.

Many companies already benefit from applying BI solutions and the BI tools market has experienced rapid growth over the past years. According to a survey from CIO[8,] the majority of companies identified the following benefits in BI applications:

- Improved competitiveness
- Improved customer service
- Improved profitability
- Improved revenue generation
- Improved capacity

[8] ***Allan Alter.*** *Does BI Really Help Businesses Get Smarter? CIO Insight. [Online] 15 October 2007. [Cited: 02 November 2007.] http://www.cioinsight.com/c/a/Research/Does-BI-Really-Help-Businesses-Get-Smarter/.*

2.2. Impact of Open Source on Business Intelligence

For a long time, Open Source Solutions (OSS) has been a part of IT organizations, but one area in which there have been few projects, is the area of Business Intelligence. However this has dramatically changed in the last few years and nowadays the number of organizations that are adopting OSS (Open Source Solutions) for BI and Data Warehouse (DW) is increasing – a trend that is expected to continue with the introduction and growth of several open source projects – By choosing Open Source, organizations can take advantage of inexpensive commodity hardware and reduced licensing costs.

Early adopters had a limited variety of open source projects to choose from, there were few for Extract, Transform and Load (ETL) tool, one for reporting, and another one for analysis, but nothing too comprehensive. The most matured projects were on Relational Database Management Systems (RDBMS), but none were optimized for very large databases. Today a company can choose from over twenty open source projects that support every aspect of BI.

According to a survey from CIO Insight[9] [10] [11] , in 2005 more than ninety percent of IT executives at midsize firms were planning to apply BI in their business. Two years later, a new survey was conducted and it showed that these firms are avid BI users. These firms invested in business intelligence systems for three main reasons: to improve business processes, compile financial results and gather and analyze customer information.

However in a CIO Insight survey about open source software, the majority of companies report that cost was one of the initial drivers that made them explore open source, eleven percent already use open source software in BI and nine percent are planning to deploy it in twelve months. (Figure 3.)

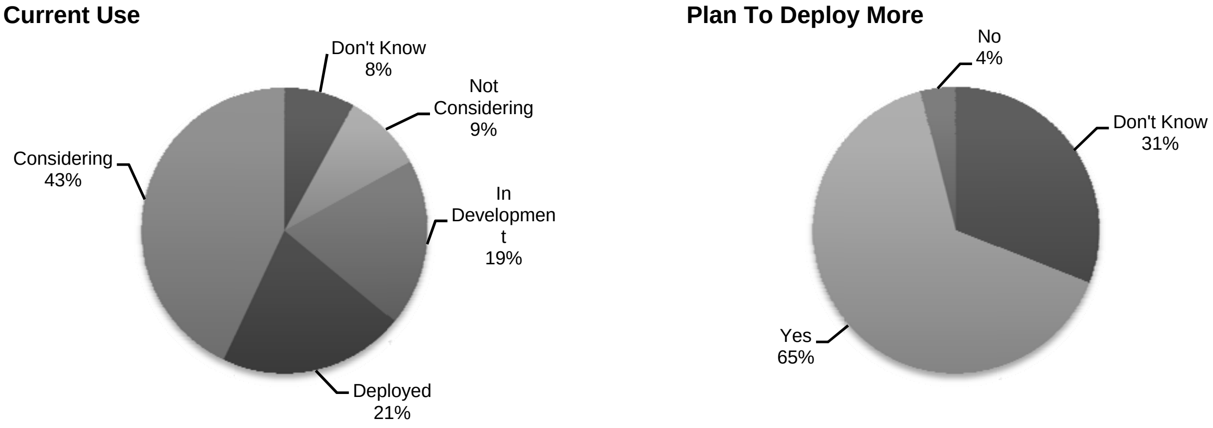


Figure 3 – Open Source BI Market Adoption - Source Ventana Research

[9] *CIO Insight – http://www.cioinsight.com/*

[10] ***Allan Alter.*** *Does BI Really Help Businesses Get Smarter? CIO Insight. [Online] 15 October 2007. [Cited: 02 November 2007.] http://www.cioinsight.com/c/a/Research/Does-BI-Really-Help-Businesses-Get-Smarter/.*

[11] *How Valuable is Business Intelligence to the Enterprise? CIO Insight. [Online] 11 October 2007. [Cited: 02 November 2007.] http://www.cioinsight.com/c/a/Research/How-Valuable-is-Business-Intelligence-to-the-Enterprise/.*

Another survey confirms, that Open Source BI as an emerging market is the one conducted by the Ventana Research[12], in late 2006. (5)

By analyzing the graphics available in the studies, it is possible to say that the market adoption was higher than anticipated and new deployments have been successful.

Combining these surveys I can conclude that open source software can bi an increasingly viable alternative to commercial software in the BI market. Its impacts are being felt by both customers and vendors.

2.2.1. Vendors and Open Source Software

Commercial BI/DW vendors are taking advantage of open source projects, reducing the total cost of a BI implementation and maintenance, making it easier to justify and fund. Some vendors use commodity servers with an open source operating systems, while others benefit by replacing product components with open source alternatives, such as application servers, *Tomcat* or *JBoss*. *Business Objects* goes further and allows organizations to use *MySQL* database for metadata store as well as data warehouse[13].

However, this availability of OSS alternatives can become a drawback. Even if this OSS projects don't offer the same features or level of scalability, some are good enough in some areas to pose some pressure or even challenge commercial BI vendors.

One of the advantages that OSS can offer against commercial software is a faster rate of incremental improvement. While commercial vendors are driven to provide features that will differentiate them from the others, OSS often gives priority to implement features that costumers want or to improve those which are most used. This means that open source focus on the functionality needed by the user.

These factors combined with the current growing number of companies that are already using OSS BI tools, will help to urge OSS adoption and over time, OSS will certainly mature.

2.2.2. Companies and Open Source Software

In a theoretic manner, the open source business model creates unique advantages for IT.

Everyone has the same resources and tools available as the creators of the software, which results in an increased support options. Since multiple parties can provide services, whereas in the commercial development model, only the vendor can provide support, the open source community (those who use OSS) can become, in the future, self-supporting. This also means that the project source code can be changed according to any organizational demands, offering a level of flexibility that can't be matched by commercial vendors.

Another difference between Open Source and Commercial software lies on vendors imposed upgrade cycles. Because of the modular architecture of data warehousing, an upgrade can be a major headache for customers who are forced to upgrade, since that changing one product may mean changing several.

[12] *Ventana Research – http://www.ventanaresearch.com/*

[13] *MySQL and Business Objects Announce Partnership: http://www.businessobjects.com/news/press/press2005/20050418_mysql_part.asp*

Open Source can give an advantage to organizations, because the upgrade cycle is not on the vendor's frame but on the company's.

Data Warehouse is abroad range of technologies, and the adoption of OSS only depends on the maturity of the open source project. The modular architecture of data warehousing makes easy the integration of OSS with commercial products, offering features that commercial BI software won't provide. Today, for almost every data warehouse component there is an open source project. This high availability allows companies to explore several areas of BI, without worrying about trial licenses or vendor's agreements. This ability to test and deploy different approaches for solving business problems is one of the biggest impacts OSS can have on companies. (6)

2.3. Business Intelligence Architecture

Data Warehousing is a term commonly associated with Business Intelligence. In fact, it is the infrastructural component on which all successful BI initiatives are built upon. For organizations who seek a long term sustainable analytics as a competitive advantage, building a data warehouse is an absolute necessity.

The process of data warehousing consists on the integration of the enterprise-wide corporate data into a single repository that will be used to support a variety of decision analysis or strategic functions. As explained in the next sections, this data can be originated from different types of data sources, and then transformed and loaded into one or more repositories to facilitate the data analysis. [7]

Implementing the data warehouse architecture includes some direct benefits as:

- Contains pre-processed, directly usable and easily available data
- Single repository for multiple data sources
- Data is cleaned, it avoids “multiple versions”

Figure 4. shows a multi-tiered architecture. [8]

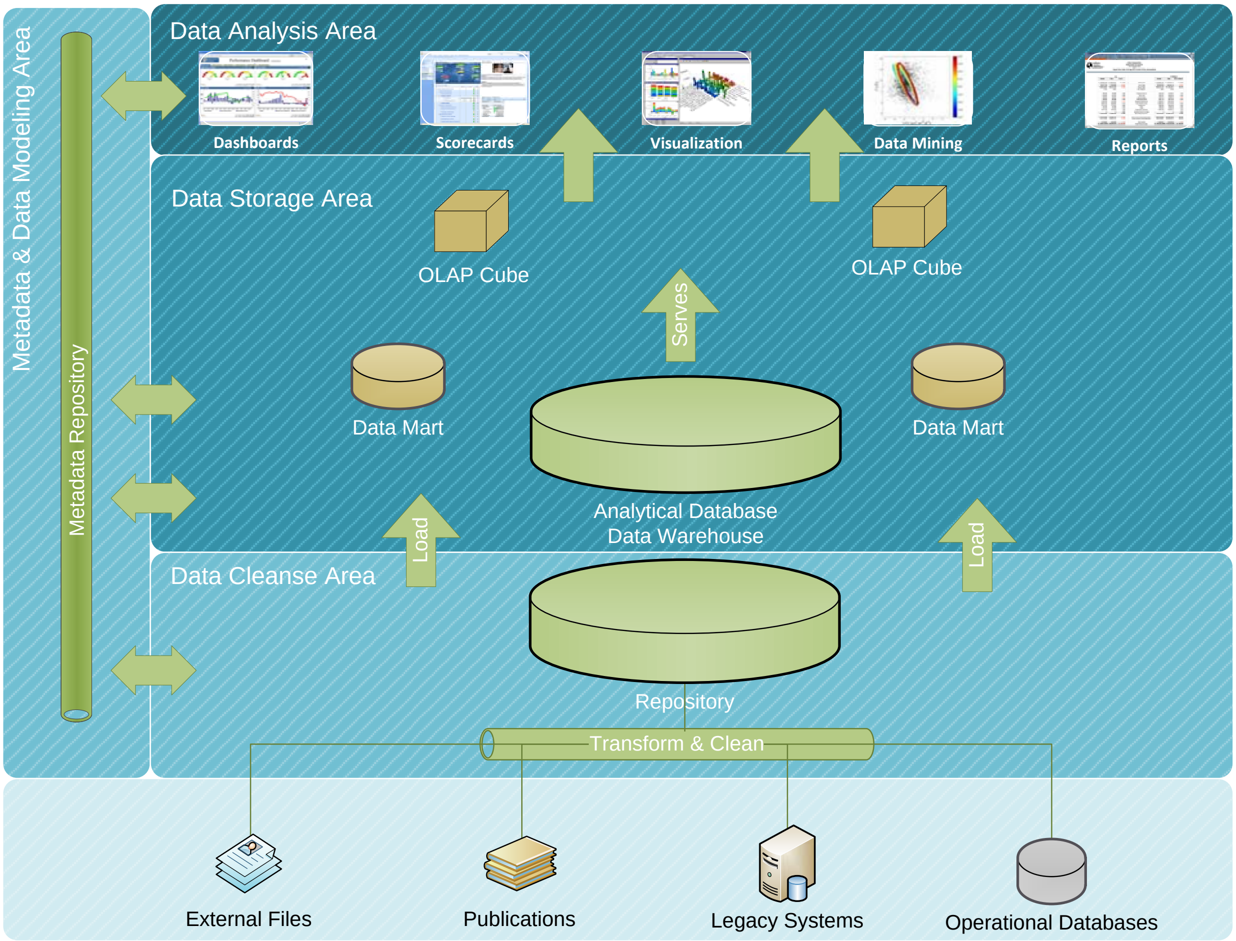


Figure 4 – Data Warehouse Multi-tiered Architecture

As illustrated, there are five distinct components that need to be considered when analyzing a data warehouse environment – operational source systems: data cleanse; data storage; data analysis; and data modelling;

2.3.1. Operational Source Systems

The operational databases and other external systems constitute the starting point in the overall data flow. They capture all the transactions that occur within the organization and are considered 'external sources' because it is assumed that there is no control over them. These systems can be tables from relational database management systems, flat files from mainframes, spreadsheets and/or other unstructured data. This data is what is typically stored, retrieved, and updated by the Online Transactional Processing (OLTP) system.

2.3.2. Data Cleanse Area

Since that a data warehouse is used for decision making and there are multiple sources with large volumes of data, it is important to keep the information without any anomalies and inconsistence. Briefly, the data staging area ensures that this discrepant data won't be loaded into the data warehouse. This is guaranteed by a set of processes commonly referred to as Extract, Transformation and Load (ETL) and Data Cleaning.

The first step is to extract data from a remote data source either through automatically scheduled events or via manual intervention. This extraction implies the understanding of the data structure so it can be manipulated in the next stage.

Once the copying of the data is finished, it is needed to apply to it a transformation to remove any errors and derive the data to be loaded into the warehouse. Some of the data extracted will require very little or none manipulation, while others must be filtered thru one or more transformations in order to meet the required business needs.

Finally, and after extracting, cleaning and transforming, the data is ready to be loaded into the data warehouse.

The process of loading to a data warehouse has to deal with much larger data volumes than for operational databases, and there is only a small time window (usually at night) when the warehouse can be taken offline to refresh it. Since sequential loads can take a very long time – try to imagine load a terabyte of data, it would take weeks or months – the solution is to use pipelined and partitioned parallelism.

2.3.3. Data Storage Area

The data storage area is where the active data of business value is stored, organized and published for direct querying by users or analytical applications.

Data Warehouse

A data warehouse can be described as a decision support database that is maintained separately from the organization's operational database and that supports information processing by providing a solid platform of consolidated, historical data for analysis.
The term of Data Warehouse was created by W.H Inmon, which he defined as:

"A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process." [9]

W.H. Inmon

It is subject-oriented because gives information about a particular subject instead of about company's daily operations. It means that it provides a simple and concise view around particular subjects, such as customer, products or sales, by excluding data that it is not useful in the decision support process.
Since that data is gathered from a variety of different types of sources and merged into a coherent location, it means that a data warehouse is an integrated collection of data.
All data in the data warehouse is only accurate and valid at some point in time. While operational databases can provide the current value of data, a data warehouse can provide information from a historical perspective. The time-variance is also shown on the implicit or explicit association of time with all data, in other words, every key structure contains an element of time.
A data warehouse is considered non-volatile because the data is not updated in real-time but refreshed from the operational systems on a regular basis, and the new data is added as a supplement of the data warehouse, rather than replacement. This enables management to gain a consistent picture of business.
Comparing to traditional heterogeneous database integration, a data warehouse can provide a higher performance. Traditional database integration requires a query-driven approach, which means that when a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for each individual heterogeneous sites involved. Once all the sites answer, the results are integrated into a global answer set. This approach can return inconsistent results and consumes much more time than an update-driven approach. This method is also used in a data warehouse, however the query is much simpler and with faster executions, because all the information from the heterogeneous sources are integrated in advance and stored. Then when a query is posed, all the information required to answer is on one specific location and with no inconsistencies.
Besides the differences above mentioned, it is enumerated on the table below, the main differences of a relational db system and a data warehouse systems.

Relational DB Systems	Data Warehouse Systems
Holds up-to-date data.	Holds historical data.
Stores detailed data.	Stores detailed and summarized data.
Data is dynamic, flat and relational.	Data is largely static and multidimensional.
Supports day-to-day decisions.	Supports strategic decisions.
Serves operational users.	Serves knowledge worker.
Has an application-oriented db design.	Has a subject-oriented db design.
Repetitive usage.	Ad-hoc, unstructured and heuristic processing usage.
Allows short or simple transactions.	Allows complex queries
Transaction-driven (OLTP).	Analysis-Driven (OLAP).

Table 1 – Differences between Relational Database and DW.

As explained before a Data Warehouse is much more efficient than a relational database model for data analysis, but nowadays more and more systems perform analysis on a relational databases, so why should we need to separate a data warehouse?

- **Missing data:** For an accurate decision support it is required historical data which operational DB's do not typically maintain.
- **Data consolidation:** Decision support requires the aggregation and summarization (consolidation) of data from heterogeneous sources.
- **Data Quality:** different sources typically use inconsistent data representation, codes and formats that need to be reconciled.

While operational systems are optimized for simplicity and speed of modification the data warehouse is tuned for complex queries, multidimensional view and consolidation, in other words, tunes for OLAP.

The term on-line Analytical Processing (OLAP) was defined in a white paper written for Arbor Software Corp in 1993 (10), as an interactive process of creating, managing, analyzing and reporting on data used for analyzing large quantities of data in real-time. In order to do this, data is perceived and manipulated as though it were stored in a multi-dimensional array.

Conceptual Modelling of Data Warehouses

Commonly used in data warehousing systems, multidimensional modelling is a design technique based on dimensions & measures instead of a relational model, designed to solve complex queries in real time.

Facts and Dimensions tables

Dimensional data modelling starts with a fact table. It is where is recorded what happened (contains the measures of interest). For example, a customer bought a bottle of wine in London, what should be in the fact table? It should be facts about it, ideally facts that are numeric and that can be continuously added (sums or averages) because a fact table can grow to billions rows or more and a user prefers to see sums or averages then details of a transaction. Another important decision to make is the granularity of the fact table. If a company doesn't care about whether or not a bottle of wine was sold at 12:03 am, recording each transaction in the fact table is too granular, and the result will be an unnecessary consumption of CPU time and disk space. So it can be useful to aggregate all the sales of any product in one specific store on a per-day basis, this choice implies that it will only be one row in the fact table instead of hundreds of them that represented every bottle of wine that was sold on a specific date in London.

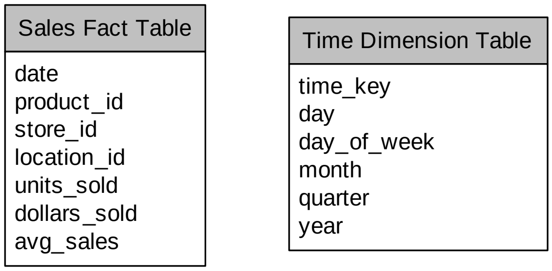


Figure 5 – Example of a Fact Table and Dimension Table

As illustrated in Figure 5, the fact table contains facts as units sold, the average of sales and the dollar value of a sale of a particularly product. In a dimensional data warehouse there will always be just one of these and all other tables will define the dimensions. A dimension contains a set of unique values that identify and categorize data.

A dimension provides the context of a fact and can be used to group related facts together. Dimensions are often hierarchical, e.g. the location might include the building, state, and country.

Considering the definition of the time dimension on figure 5:

Why is it useful to define a time dimension? If we keep the date of the sales fact as a date column, it is “hard” to ask for holiday versus non-holiday sales or winter versus summer sales.

Every dimension table is described by a set of attributes, which are related via a hierarchy of relationships. For example, the dimension "time" consists of four attributes: day, day of week, month, quarter, and year. These attributes form a chronological hierarchy (see illustration below). The function of data hierarchy is to allow data to be handled at varying levels of abstraction. This is achieved by operations such as pivoting, drill - down, roll - up, and slice & dice. These operations supported by the data multidimensionality, is discussed at a further stage.

With this dimension table we must redefine the fact table as follows:

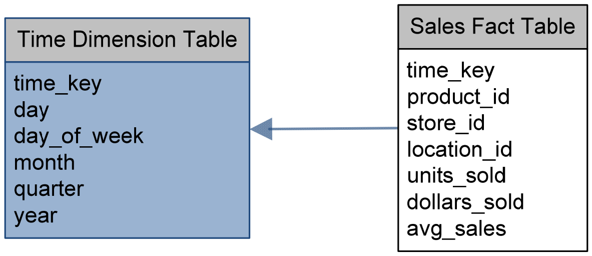


Figure 6 – Example of Fact Table and Dimension Table Relationship

Instead of storing a date in the fact table, it is used an integer key pointing to an entry in the time dimension. With this structure it is possible to easily generate complex queries such as “get a report of sales by fiscal year or sales by day of week”.

It is possible to add as many as dimensions as the fact table require.

Conceptual Models

The relational implementation of the dimensional data model is typically a star schema, a snowflake schema or a fact constellation.

A Star schema is the simplest style of dimensional schema, and is characterized for having a fact table in the middle connected to a set of dimension tables.

The following figure illustrates the relationship of the fact and dimensions tables within a simple star schema with a single fact table and four dimension tables. The fact table has a primary key composed of four foreign keys, time_key, item_key, branch_key and location_key, each of which is the primary key in a dimension table. The other columns in a fact table are referred as measures, in a dimension table, they are referred as attributes.

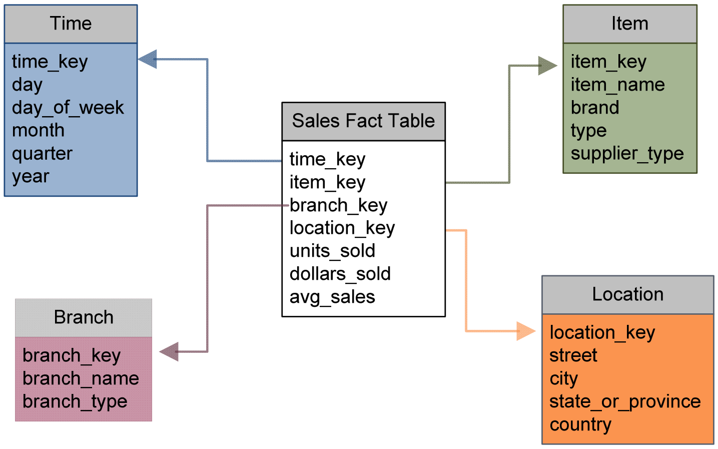


Figure 7 – Example of a Star Schema

The snowflake schema is a more complex conceptual model than a star schema and is a refinement of star schema. It consists of one Fact table connected to many dimension tables, which can be connected to other dimension tables.

Snowflake schemas normalize dimensions to eliminate redundancy. In other words, the dimension data has been grouped into multiple tables instead of one large table. For example, an item dimension table in a star schema might be normalizes into an item table and an item supplier table in a snowflake schema. While this saves space, it increases the number of dimensions tables and requires more foreign key joins. The result is more complex queries and reduced query performance.

Snowflake schema is appropriated to use when a dimension is very sparse and/or a dimension has a very long list of attributes which may be used in a query. This schema can reflect the way in which users think about data.

The illustration below (Figure 8) represents a graphical representation of a snowflake schema.

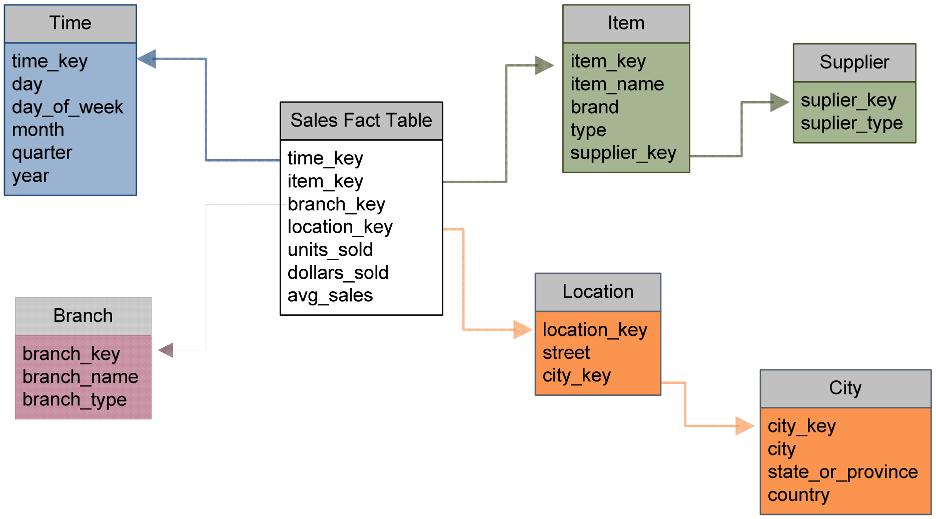


Figure 8 – Example of Snowflake Schema

For each star schema or snowflake schema it is possible to construct a fact constellation schema. This schema is more complex than star or snowflake architecture, because it contains multiple fact tables. They can be viewed as a collection of stars and therefore called galaxy schema or fact constellation.

That solution is very flexible, however it may be hard to manage and support because of the many variants of aggregation that must be considered.

An example of constellation schema model is shown in the illustration below (Figure 9).

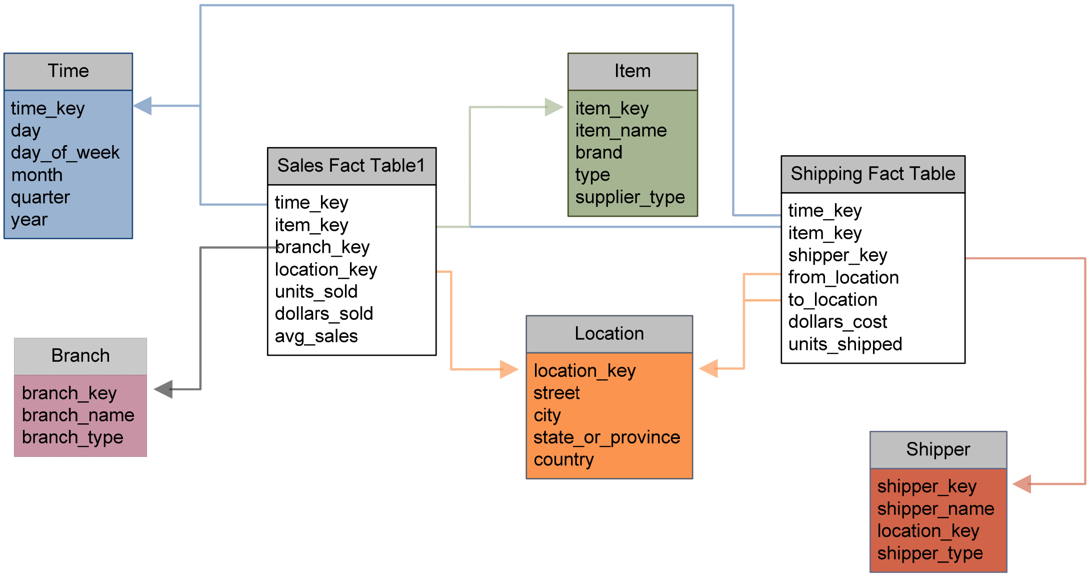


Figure 9 – Example of Fact Constellation Schema

Multidimensional Data Model

A popular data model that influences this area is the multidimensional model, which views data in form of a cube. A cube can be thought of as an extension to the two-dimensional spreadsheets. It allows different views of the data to be quickly displayed. In other words, a cube is a data abstraction that allows viewing aggregated data from a number of perspectives.

Figure 10 shows a small data cube example from the automotive industry. This particular data cube has three dimensions; manufacturer, year and region, and a single measure attribute – sales. By selecting cells, planes, or sub cubes from the base cuboids, we can analyze sales figures at varying granularities. Such queries form the basis of OLAP functions like roll-up and drill-down.

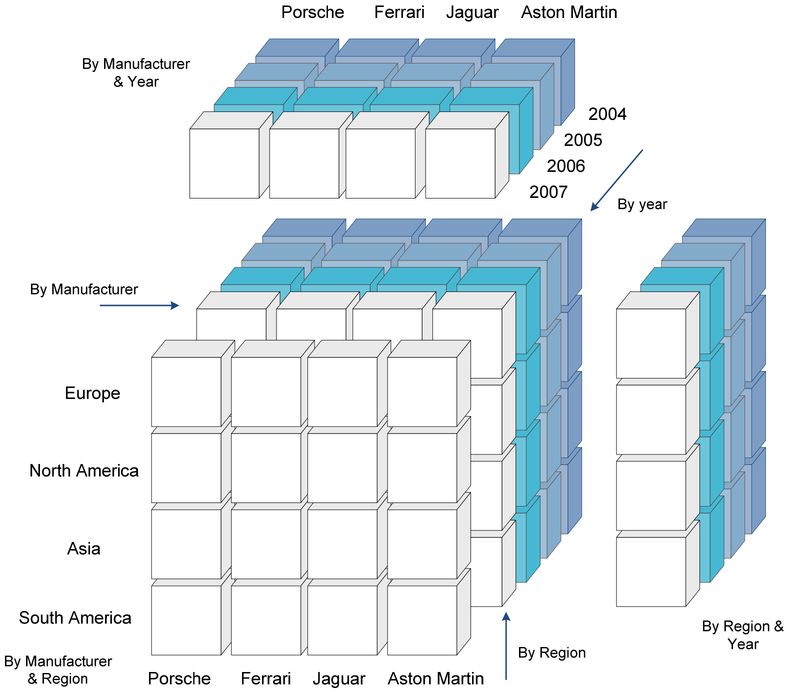


Figure 10 – Example of a Cube

2.3.4. Data Modelling or Metadata Area

The name Metadata suggests some high-level concept, but it is simple as “data about data”. A good description of the data is essential to the operation of a data warehouse, since it can assist companies during the data acquisition, data transformation and data access phases. During the data acquisition phase, metadata helps in the translation of information from the operational to the analytical systems.

Since that Metadata describes the business data, it helps business user’s navigation with the warehouse content by providing an easy and intuitive interface. (11)

2.3.5. Data Analysis Tools

The last and most critical component of the BI Architecture is the Data Analysis Tools that help the interaction between business users and the data stored in the repositories. The variety of data access tools range from the simple ad-hoc query tool to a complex data mining application.

The following sections explain the most known tools available in this component.

Data Mining

Data mining relates to the discovery of useful and/or previously unknown patterns within a data set. Since it uses some of the most advanced computational techniques, such as neural networks, predictive modelling or genetic programming, Data Mining is normally used for prediction analysis and classifications. Therefore, this process extracts knowledge stored or predictive information from the DW without requiring specific queries. – E.g. what is the likelihood that a costumer will terminate contract? Or migrate to a competitor.

Data Visualization Tools

Data Visualization tools are used to display data from the repository, providing live interactivity, this means that the user is allowed to manipulate data to show relevancy and patterns.

In this section we can categorize the available tools in three different areas: Analytic Applications, Dashboards and Scorecards.

A dashboard is a business management tool similar to car dashboards. They provide multiple indicators or reports based on the company's key performance indicators on a highly visual way. As illustrated on Figure 11, a dashboard can be comprised of devices, normally set in a portal-like environment, such as red/green/yellow lights, alerts and graphics such as bar charts, gauges or others.

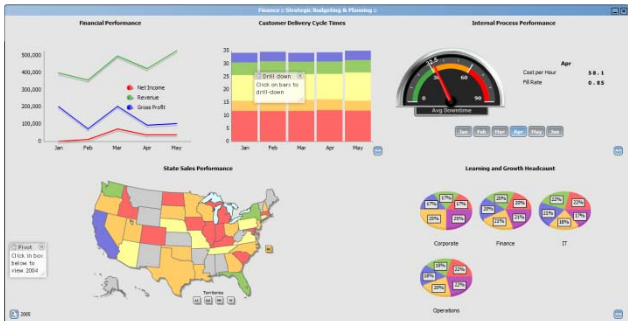


Figure 11 – Example of a Finance – Strategic Budgeting Dashboard (Source: iDashboards[14])

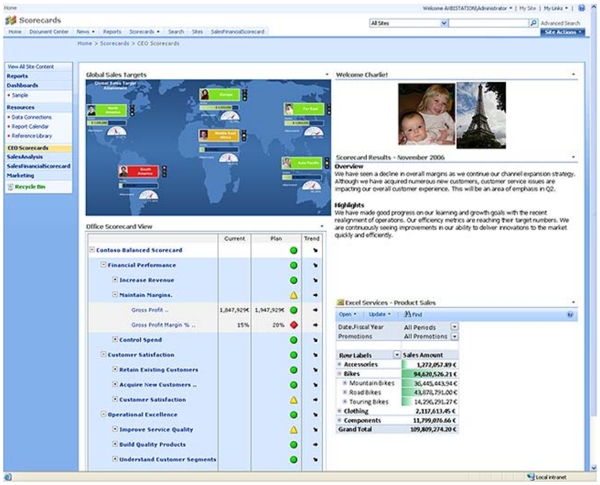
[14] *http://www.idashboards.com/*

There is no limitation to what a dashboard can track. Most companies uses it to assess the progress of departments like human resources, recruiting, sales, operations, customer relationship management and many others.

In conclusion, dashboards, through a visual presentation of performance measures, will help a company to realign resources and strategies more efficiently.

Whereas dashboards present multiple numbers in different ways, a scorecard focuses on a given metric and compares is to a target.

A scorecard is the specific methodology associated with the Kaplan and Norton model [(12)], and it is a custom interface that helps organizations to optimize their performance.



Office Scorecard View

	Current	Plan		Trend
Contoso Balanced Scorecard			●	↘
Financial Performance			●	↘
Increase Revenue			●	↘
Maintain Margins.			△	→
Gross Profit ..	1,847,929€	1,947,929€	●	→
Gross Profit Margin % ..	15%	20%	◆	→
Control Spend			●	↘
Customer Satisfaction			●	↘
Retain Existing Customers			●	↘
Acquire New Customers ..			●	↘
Customer Satisfaction			△	→
Operational Excellence			●	↘
Improve Service Quality			△	↘
Build Quality Products			●	→
Understand Customer Segments			●	↘

Scorecard Results - November 2006

Overview

We have seen a decline in overall margins as we continue our channel expansion strategy. Although we have acquired numerous new customers, customer service issues are impacting our overall customer experience. This will be an area of emphasis in Q2.

Highlights

We have made good progress on our learning and growth goals with the recent realignment of operations. Our efficiency metrics are reaching their target numbers. We are continuously seeing improvements in our ability to deliver innovations to the market quickly and efficiently.

Date.Fiscal Year	All Periods
Promotions	All Promotions

Row Labels	Sales Amount
Accessories	1,272,057.89 €
Bikes	94,620,526.21 €
Mountain Bikes	36,445,443.94 €
Road Bikes	43,878,791.00 €
Touring Bikes	14,296,291.27 €
Clothing	2,117,613.45 €
Components	11,799,076.66 €
Grand Total	109,809,274.20 €

Figure 12 – Example of a Scorecard (Source: Microsoft)

Reporting Tools

These tools provide an easy access to reports on the business relevant data and help users to keep track of the company's key performance indicators. There are two different types of reporting style, the production and the management style.

Production style is the process of querying an OLTP database, then formatting it to create a document, like an invoice or a statement. Since the information needs of these reports rarely change, they are normally done via custom programming,

Management style query and reporting is intended for users who want to author their own reports. They are less concerned with the precise layout but do want charts and tables quickly and intuitively.

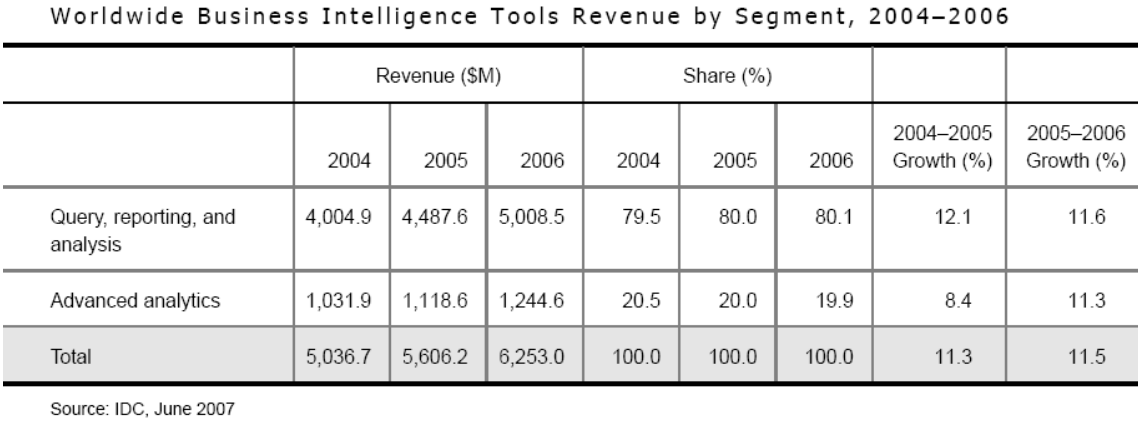
2.4. Business Intelligence Market Overview

In the last two years (2006 and 2007), the customer demand for real-time reporting, the penetration of BI into operational level and the proliferation of data sources has earned to BI the top spot in Gartner's "Top 10 Technology Priorities" [(13)].

2.4.1. Market Segmentation

The International Data Corporation (IDC) suggests that, in the software taxonomy, the BI Tools are part of a broader market called business analytics. [(14)] [(15)]

The IDC also suggests uneven market segmentation composed by two market segments: Query, Reporting and Analysis (QRA) and advanced analytics. QRA software includes ad-hoc query and multidimensional analysis tools as well as dashboards and production reporting tools while the advanced analytics segment includes data mining and statistical software and uses technologies such as neural networks, rule induction, clustering to discover relationships in data and make predictions too complex to be obtained using query or reporting tools. The evolution of this market in the years 2004-2006 is presented below (Figure 13).



Worldwide Business Intelligence Tools Revenue by Segment, 2004–2006

	Revenue ($M)			Share (%)				
	2004	2005	2006	2004	2005	2006	2004–2005 Growth (%)	2005–2006 Growth (%)
Query, reporting, and analysis	4,004.9	4,487.6	5,008.5	79.5	80.0	80.1	12.1	11.6
Advanced analytics	1,031.9	1,118.6	1,244.6	20.5	20.0	19.9	8.4	11.3
Total	5,036.7	5,606.2	6,253.0	100.0	100.0	100.0	11.3	11.5

Source: IDC, June 2007

Figure 13 – BI Market evolution in the years 2004-2006 (Source: IDC)

In 2006, the market grew by 11.5% to reach worldwide revenue of $6.25 billion. This result means that here was no significant consolidation in the BI market, since there was only a 0.2% improvement from the last year.

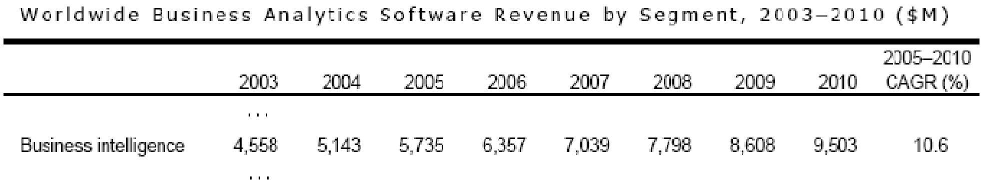
The Americas region continues to be the largest segment of the market, followed by Europe, Middle East and Africa (EMEA) and Asia/Pacific.

Windows continues to dominate the market, followed by UNIX and mainframe Platforms. Although Linux represents a small fraction of the worldwide OS share, it is by far the fastest-growing platform, and new open source initiatives are likely to sustain or accelerate this trend.

2.4.2. Market Trends

The BI market is driven by the need of improvement of management performance. This management performance, can take on the form of various decision-support systems to improve revenue, decrease costs, uncover opportunities, and others. Since that organizations are in need to find relevant information, in order to detect future trends or conduct what-if scenarios the BI tools are becoming an important ally to companies.

According to IDC, the business analytics software market reached $19.3 billion in 2006, representing a growth rate of 11.2 percent, from which $6.25 billion (32% of revenue) belongs to BI tools. The authors expect the worldwide Business Intelligence market to continue to grow at a healthy compound annual growth rate of 10.6 percent until 2010, as depicted above on Figure 14



Worldwide Business Analytics Software Revenue by Segment, 2003–2010 ($M)

	2003	2004	2005	2006	2007	2008	2009	2010	2005–2010 CAGR (%)
	. . .								
Business intelligence	4,558	5,143	5,735	6,357	7,039	7,798	8,608	9,503	10.6
	. . .								

Figure 14 – Market forecast 2005-2010 (Source: IDC)

Furthermore, IDC has made the following predictions about the future of business analytics:

- A broader set of organizations is beginning to look at business intelligence not just as a set of reporting functions, but as a means to gain competitive advantage through better decision management and process optimization.
- Business analytics solutions will increasingly incorporate functionality for unified access and analysis of structured data and unstructured content, business process management, collaboration and workflow management functionality.

Below, are mentioned the trends that are the most expected to happen in a near future, or are already emerging.

Convergence of Structured and Unstructured Data

As far as the data analysis process and user interface are concerned, it s expected that Structured data remains just that, structured. However, if complemented with unstructured data, it can become more valuable.

For example, today, when companies analyze their costumer segmentation, they take in consideration data such as buying patterns or behaviours, customer demographics, and other types of analytical data. Not forgetting that 80% of the company's information is unstructured, data such as, hidden comments in customer emails, customers complain or requests, combined with the existence structured data would contribute for a more effective analysis. This combination can become a powerful differentiator for all types of information-intensive applications.

The idea behind this trend is that, in the future, the OLAP engine will automatically match the unstructured data to one or several dimension, quantify and qualify them into star schema facts, as if they were structured data.

Today, there are several text mining tools that help analyzing unstructured data, and the majority of them enable search, categorization, pattern recognition, and information extraction. It is expected that this tools became available in most of the BI Suites in a near future.

Dashboards Evolution

Dashboards are no longer static graphs and charts, the best ones implement components of Advanced Data Visualization: Visual/Actionable Interface, Visual Query, Dynamic Data Content, Multiple Linked Visualizations, Animations and Geospatial Representations. However these components aren't enough for turning information into decisions and actions, therefore is expected a marriage between Dashboards and Business Activity Monitoring (BAM) that would allow the combination of data and business processes.

An added benefit is "actionable BI". BI has been traditionally used to report the current state of business, but when the processes and data converge it is possible to upgrade to an actionable BI. This means that when an alert or event is triggered by a data condition, it would activate a business process so that the event can be followed up and acted on.

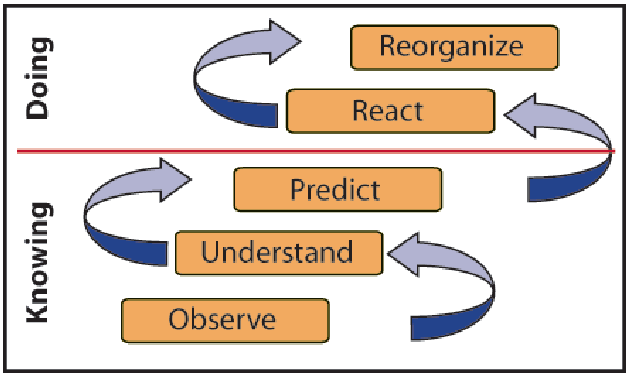


Figure 15 – Actionable BI - How it works. (Source: DMReview)

2.4.3. Commercial Vendors

To date, the BI market is dominated by proprietary vendors. Below is a graphical representation of the marketplace for the first quarter of 2008. It depicts Gartner's analysis of the major vendors.

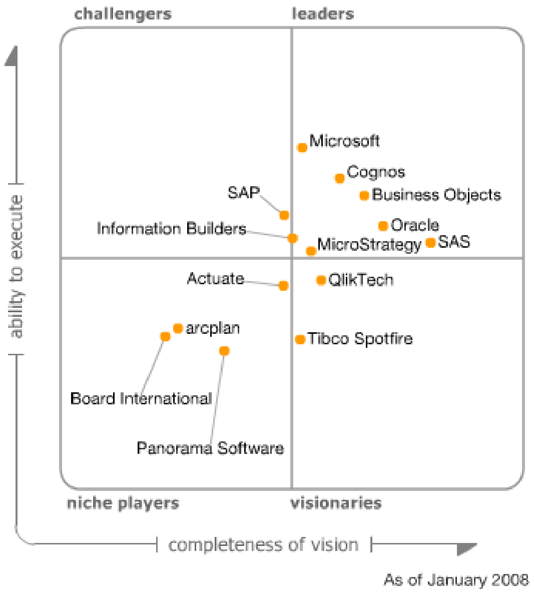


Figure 16 – Magic Quadrant First Quarter – 2008 (Source: Gartner)

A *revolution* is what better sum up what is happening in the market of Business Intelligence.

It all started in late 2005, with the acquisition of privately-held *Infommersion*, a leading vendor of interactive visual analytics, by the world's leading provider of BI solutions at that time, *Business Objects*. Since then and until today (early 2008), *Oracle* has acquired more than 40 Companies, including *Hyperion*, a leading provider of Enterprise Performance Management solutions. In this last two years, *Cognos* acquired Celequest, Applix and in late 2007 was bought by *IBM*, SAP acquires *OutlookSoft* and *Business Objects,* and there are some rumors that *Microsoft* may acquire *SAP.*

So, now we have *Microsoft, Oracle, SAP,* and *IBM* fighting side by side for supremacy. If these large corporations are spending major amounts in acquiring companies is probably because BI is expected to be a major asset to help achieve leadership in the Information Technologies. For this reason, it is possible that some other large companies, such as HP or Google (web can be a platform for future BI), appear in the market, and have a more significant role.

Open-source BI has come a long way, but its vendors do not yet generate enough revenue to be included in the Magic Quadrant.

2.4.4. Open Source Vendors

Although, the impact of open source BI tools is forecasted to be limited in the next five years, the market had evolved and became more matured, mainly thanks to a thriving ecosystem of open source BI Vendors. It is expected that after this period, open source BI develop into a stronger competitive force.

BI is dominated by giants like IBM and Oracle, but now, small open source challengers are appearing. The most well known open source BI tools vendors are *Pentaho, JasperSoft, Jedox* and *Greenplum*. However these companies aren't alone, there are many communities and universities that offer open source BI solutions, such as the University of Waikato, Neuseeland, SpagoBI, and the Eclipse Foundation.

Although I've already mentioned in the chapter two, it is important to refer some of the advantages of using open source instead of commercial products:

- Usually open source software is free, and it doesn't have any dependence with proprietary software. It can run on any operating systems, like Mac OSX, Linux, Solaris or Windows, therefore companies can benefit by eliminating the license costs.
- Normally, commercial vendors offers a complete solution that includes reporting, analysis, data mining and many other functions, that are sometimes unnecessary for some clients. Using open source, the organization can focus on which components do they really need.
- Organizations can access to open source code and adapt or embed it to any existing application.

Open Source Business Intelligent Suites

Today, we have a large variety of choices to build BI solutions from open source components. It is easy to find an open source projects that support every feature required to build a BI solution.

In 2004, there were 14 open source projects that delivered a BI related function and one year later, that number had grown to almost the double, and is still growing today.[15] However, there are a few applications that cover all the units of a BI Solution and since there isn't a strict definition for what is a BI Suite (should it be a comprehensive or end-to-end solution?) I considered that if a project unites more than one tool for creating a BI solution then it is a Suite.

Considering this definition we can find four full featured BI Suites, but JasperSoft, Pentaho e SpagoBi seem to be the most active in end of 2007, beginning of 2008, providing the most features and innovation. Above is small overview of each BI Suite and their capabilities. A deeper analysis can be found in the following chapter.

The Jasper BI Suite provides not only full OLAP and ETL capabilities but also a framework capable of automatic or ad-hoc reporting. It is based on JasperIntelligence JasperServer Framework and Mondrian-based JasperAnalysis that provides a Web-based environment for Reporting, data analysis (OLAP) and

[15] ***Mortensen, Rick.*** *The Open Source BI Trend Will Grow - Here's Why. DMReview. [Online] March 2006. [Cited: 20 December 2007.] http://www.dmreview.com/dmdirect/20060317/1050215-1.html.*

data integration (via JasperETL). It also includes a JasperReports and iReports component for ad-hoc report creation.[16]

Openi provides a web-driven Business Intelligence application, based on J2EE that acts as an interface to OLAP, statistical and data mining sources. It can be integrated with SQL 2005.

It is an out-of-the-box solution for building and publishing reports from different OLAP data sources, such as Microsoft Analysis Services or Mondrian. [17]

The Pentaho BI Suite provides a complete business intelligence platform that includes reporting, analysis, dashboards, data mining, and data integration. It is the most known Open Source BI Suite and is currently used by leading organizations including MySQL, Motorola, Terra Industries, DivX and more. [18]

SpagoBI is a platform focused on the BI needs at the enterprise level. It offers a complete analytical layer with OLAP, data mining, dashboards and visual data inquiring. It also allows a mix approach, with open source and proprietary products. Relatively to other BI Suites, it provides a higher level of abstraction. [19]

The next sections will detail both open source BI market leaders (*Pentaho* and *Jaspersoft*) and their solutions.

Pentaho

Pentaho was founded in 2004 by an experienced team (former consultants of leading commercial vendors such as *SAS*, *Oracle*, *IBM* and *Hyperion*) to provide Business Intelligence under a professional open source business model. This model is used by leading open source companies such as, *MySQL, JBoss* and *Zimbra* and can be applied to companies that are the main source code contributor.

The company not only develops components for the open source community but also improves those that already exist and were built by other individuals. Therefore, Pentaho integrates components into a flexible “puzzle” that developers can use to easily assemble a custom solution to be applied in a specific area (Reporting, Data Cleanse, Data Mining, etc), or to create a comprehensive BI Platform.

Pentaho also provides technical support, release management, quality and commercial extensions to open source products.

The Pentaho mission is to provide an Open Source alternative and exceed all commercial offerings in terms of features, functions and benefits.

“Pentaho manages, facilitates, supports, and takes the lead development role in the Pentaho BI Project - a pioneering initiative by the Open Source development community to provide organizations with a comprehensive set of Business Intelligence (BI) capabilities that enable them to radically improve business performance, efficiency, and effectiveness.” [20]

[16] http://www.jaspersoft.com/

[17] *http://openi.sourceforge.net/*

[18] *http://www.pentaho.com/*

[19] *http://spagobi.eng.it/*

[20] *Pentaho About – http://www.pentaho.com/about/*

Pentaho provides a full spectrum of BI products that cover the areas of reporting, analysis, dashboards, data mining and data integration.

Pentaho Reporting

In January 2006, Pentaho adopted *JFreeReport* (now Pentaho Reporting), a free java reporting library for embedded solutions. It allows organizations to easily access, format, and distribute securely information to employees, customers and partners via corporate portals, e-mail or custom applications. It also provides access to information in relational, OLAP or XML data sources, and delivers output in different formats including PDF, HTML, Spreadsheets, Rich Text Format or Plain Text. Offers a flexible deployment, the user can choose from a standalone desktop or an interactive web-based reporting tool.

It also has a graphical designer that provides full control of data access, layout, grouping, charting and formatting for reports.

Pentaho Analysis

Pentaho Analysis is a powerful tool that helps knowledge workers to operate with maximum effectiveness allowing them to explore business information by drilling into and cross-tabulating data. In other words, it helps users to make optimal decisions. It provides extensive analysis capabilities, including pivot table viewers (thought *JPivot*), advanced graphical displays using SGV or flash, data mining and workflow integration.

To perform analysis on relational database it is used the Pentaho Analysis Services, which is based on *Mondrian OLAP*.

In addition, Pentaho Spreadsheet Services (paid service) allows users to interactively explore (browse, drill, pivot and chart) and analyze data directly within Excel and against Pentaho Analysis Services.

Dashboards

Pentaho Dashboards provide an immediate insight into departmental or enterprise performance. It provides metrics management capabilities and appropriate visualization, which helps users to immediately see the value of each key performance indicator. It can be integrated with Pentaho Reporting and Analysis, allowing users to understand what factors are influencing the performance, and with external content, such as, web pages, RSS feeds or map applications.

Data Integration

Pentaho Data Integration is based on the open source project *Kettle*, and provides an easy-to-use graphical, drag-and-drop environment and includes:

- Large rich transformation library with mapping objects;
- Database Exports to text-files or other databases;
- Import data from text-files to style sheets;

- Data migration between database applications;
- Exploration of data in existing databases (tables, views, etc.);
- Enterprise-class performance and scalability;
- SAP and AS400 Adaptor;

Data Mining

Pentaho Data Mining uses *Weka* for data analysis. It has a collection of machine learning algorithms such as clustering, segmentations, decision trees, neural networks and other that are combined with OLAP. This tool provides machine-intelligent data analysis to end users. Therefore, Pentaho Data Mining allows users to analyze historical data and to create predictive models. It can be combined with Pentaho Reporting and Analysis, to ease the information flow to the appropriate people.

Features and Benefits of Pentaho Data Mining:

- Interactive output, with visualization of the results.
- Provides role-based security and business rules.
- Filters for normalization, re-sampling, attribute selection, and transforming and combining attributes.

Pentaho BI Platform

All the capabilities above described can be integrated in the Pentaho BI Platform.

The Pentaho BI Platform is some way different from any traditional BI offerings. It is a solution-oriented and process-centric platform that enables companies to build complete solutions to any given BI problem.

The platform is considered process-centric because it includes an embedded workflow engine as the central controller, thus can be easily integrated into business processes.

Logging, auditing and security are also built in the core and are used automatically, to ensure that there is always an accurate audit trail available for both maintenance and performance monitoring.

The framework also includes a solution engine that integrates the former presented components (reporting, analysis, dashboards, etc) to form a comprehensive BI Platform.

It is built upon a foundation of servers, engines and components. These provide the J2EE server, security, portal, workflow, rules engine, charting, collaboration, content management, data integration, analysis and modelling features. Since that many of these components comply with many associated specifications, they can be replaced with other products.

To deliver this solution, the Pentaho BI Platform contains the Pentaho Server and an Eclipse-Based Design studio.

The Pentaho Server contains not only the engines and components for reporting, analysis, business rules, email and notifications, and workflow, but also the infrastructure that provides advanced system administration, such as, system monitoring services, usage reports, web service support, and diagnostic tools. It also supports Enterprise Application Integration (EAI) as well as ETL capabilities.

Some parts of the architecture use a combination of technologies that can be swapped for equivalent ones. For example, the J2EE server is JBoss, but any 1.4JDK compliant application server can be used. The OLAP Engine (Mondrian) can also be changed for any MDX-compliant OLAP server like Microsoft OLAP services or Hyperion Essbase.

Some of the benefits of using Pentaho BI Solution are:

- Java developers can use the components for rapidly assemble custom BI solutions.
- Independent software vendors can enhance the value and capability of their solutions by embedding BI functionality.
- End users can benefit from applying lower cost BI tools.

JasperSoft

JasperSoft was founded in 2001, as a vendor of commercial open source business intelligence solutions. According to their website, there are more than 70,000 worldwide deployments of JasperSoft products. It is also mentioned that they have more than 7,000 paying customers, dispersed in 96 countries.

The business model used by Jaspersoft is slightly different from the one used by Pentaho. For both companies, their products are freely available as open source and can be downloaded at sourceforge.com, however JasperSoft product are also available as Commercial Licenses.

JasperSoft also provides a complete set of support and services options for any size organization.

During the last years, Jaspersoft established partnerships with some of the major Open Source Software vendors, such as, Sun, MySQL, Red Hat/JBoss, Novell, Ingres, Unisys and Talend.

Jaspersoft mission is to make BI more accessible to everyone.

"JasperSoft offers the most widely used open source business intelligence software in the world. Designed for both developers and businesses, the JasperSoft Business Intelligence Suite is comprised of an interactive reporting server, graphical and ad hoc report design interfaces, OLAP analysis, an ETL tool for data integration, and a Java reporting library." [21]

Products

The JasperSoft Business intelligence Suite is build on JasperReports, and provides a robust production reporting, interactive reporting, data analysis and data integration capabilities. Similar to Pentaho, all these capabilities are available as stand-alone products or as part of an integrated BI suite.

Besides the components described below, JasperSoft provides the Jasper4Solutions. It joins the JasperSoft BI Suit to known business and applications. It simply provides reporting, OLAP analysis, dashboards and data integration for CRM, ERP and other applications. The existent "adaptors" are Jasper4SalesForce, Jasper4MySQL, Jasper4Oracle and Jasper4Sugar.

[21] *Jaspersoft - http://www.jaspersoft.com/*

JasperReports – Report Library

JasperReports is one of the most known open source reporting tools. Is designed for developers and can be easily embedded into any Java application to give to it advanced reporting capabilities. Has the ability to deliver rich content to the printer or into PDF, HTML, XLS, CSV and XML files.

The data used to populate a report can be defined within JasperReports. It includes JDBC-wrapped data providers for relational databases, JavaBeans (EJB, Hibernate), plain old Java Objects (POJO) and xml data sources. Custom data source providers can also be easily added.

It can also support an extensive array of chart types, and create OLAP-style drill-down reports.

Similarly to Pentaho Reporting, JasperSoft provides a graphical report designer to create Pixel-Perfect reports.

JasperServer

Designed for business users and developers, provides a standalone server or a web service reporting engine, that can deliver mission critical information on a real-time (Dashboards) or scheduled basis (Reports) to the web, printer or to a variety of file formats.

It reduces the time required to build and deploy server applications that need reports, analytics and dashboards, giving to decision-makers easily access to critical data.

JasperAnalysis

JasperAnalysis provides data analysis capabilities for business users. It can be used to explore trends, patterns, anomalies. It also allows users to “slice and dice”, pivot, filter and drill down data in real-time.

It also Include an OLAP server and a web-based user interface. The JasperAnalysis Server is a secure relational OLAP (ROLAP) server, which allows database administrators to manage the ROLAP data store. The JasperAnalysis User Interface is an easy to use web-based interface application for non-technical users and analysts. It provides institutive slice-dice interface to drill, pivot and visualize data.

JasperETL

JasperETL provides data integration capabilities. In other words, it can be used to merge and transform information from multiple data sources into a single consistent store, so it can be analyzed.

The component supports a wide range of over 100 data sources and targets, including JDBC, delimited text, XML, etc. Custom components can be added to provide access to unusual or legacy data sources.

Additionally and as Pentaho Data Integration, JasperSoft provides an Eclipse graphical designer to ease the process of extract, transform and load the data into the warehouse.

2.4.5. Open Source Business Intelligent Suites Comparison

There are several Business Intelligence tools. Therefore it is important to understand which of them are more suitable for specific needs.

This chapter provides a table (in appendices) that shows the comparison of Pentaho BI Suite and JasperSoft BI Platform. The table is sectioned by six core BI capabilities: ETL, Metadata, Reporting, Analysis, Monitoring and Administration.

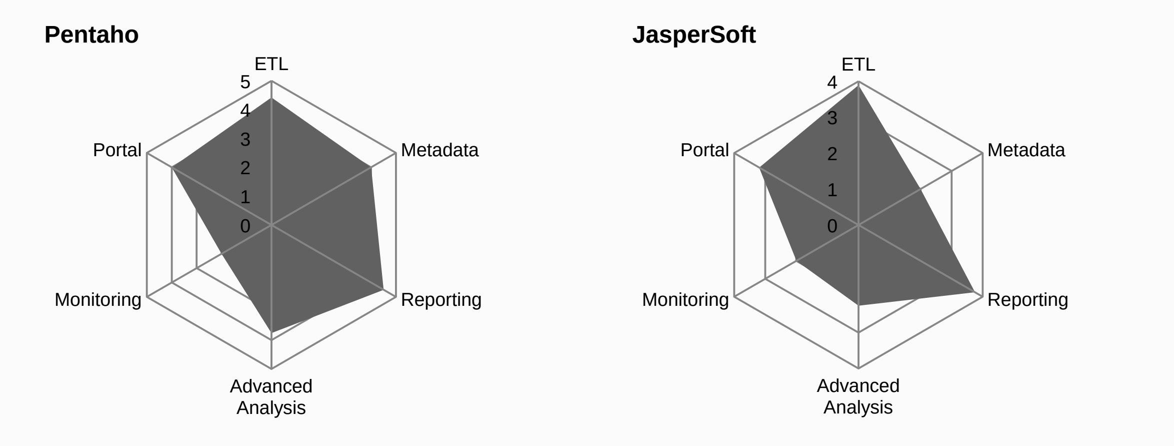


Figure 17 – Pentaho and Jaspersoft Overview Result

The above charts (Figure 17) were generated with the data collected during the evaluation and comparison of the two business intelligence solutions.

2.5. Conclusion

As a result of an intense competition and strict regulatory requirements, organizations face new challenges. As explained on the paper, to surpass them, they need to increase their customer profitability, reduce operational costs and improve service offerings. These solutions are heavily dependent on the gathering and analysis of detailed enterprise information. Due to the high-priced and risk, these solutions are only available for some organizations.

Although IDC predicted that the use of open source BI would be limited in the next years, open source vendors are developing into a stronger competitive force, mainly because of the lower costs, reduced dependence on proprietary tools, and flexibility offered.

By using open source, organizations, in special, small and midsize business can reduce their project risks and increase the Return of Investment (ROI) expected for a project of this kind, helping them to become even more competitive. Therefore, in the future, open source business intelligence software can become a must-have solution for any organization.

Chapter 3 – Case Study

In this chapter, it is give an overview of the transports industry and is introduced the *Incident Management Standard Platform* that was suggested to proof that Open Source Business Intelligence is a viable solution to use in current BI Projects.

3.1. Transports Sector Overview

The transports sector plays a major role in a country's economy because it enables the mobility of goods, services and people as efficiently as possible. This sector can be characterized by a strong heterogeneity, not only because of the wide range of modes available (including aerospace, automotive, marine and rail), but also because of the wide variety of resources used, from traffic infrastructure (routes, networks, node etc) and vehicles or containers, to workforces and energy supplies.

Moreover, the transport has been a particular sector open to innovation, whether regarding the modernization of processes, or through the emergence of new technologies.

It can be broken down into the following categories:

- Surface transport, which includes road and rail transport.
- Air transport
- Sea transport.

For this case study, I will focus in public land passengers transport, a category of road transportation.

Land passenger transport is a generic term used to describe both public and private modes of transports. It covers all passenger movements from a short distance urban ride to long distance or inter-city transport and can have different scheduling, from a regular to an occasional service.

Urban public transports provide a wide range of potential benefits to a city (16).

- One that should be taken in account is Environmental Sustainability. The energy consumption and emission per passenger/per kilometre is lower than of the automobiles (17) and that some operators are replacing diesel for other greener energy, helps providing a cleaner environment.
- The availability of alternative transports to the automobiles tends to relief the congestion in cities, allowing passengers/travellers a way to bypass traffic and achieve higher levels of mobility.
- It is a safer mode of transport (accidents per km), in comparison with private transportation.
- The cost of a car ownership is a barrier for many households, especially the poor. By contrast, the cost of public transports is low. In this way the public transport as a vital role as provider of low-cost mobility for the car-less travellers.

Urban Transport operators usually operate in a limited geographic area, however some companies that are established in a common region, have joint ventures and created an association to became more effective and efficient.

A person choose the mean of transport depending on the availability, speed, convenience and safety of each one, therefore, to achieve success, a transport operator must implement a regular and efficient network. As a result of this offer, the operator must manage effectively their resources and achieve scale economies.

Theoretically, a transport operator could achieve scale economies by raising the number of passenger transported (larger vehicles, reducing the labour costs) or by increasing frequencies of buses (smaller vehicles, higher labour costs). Either of the choices implies that the transport company should maximize each resource, to make any profit.

As an additional infrastructure to Intelligent Transport Systems, Business intelligence can help companies to better understand their data and their resources in order to make better decisions about the daily operations and strategies.

3.2. Incident Management Standard Platform

For most of public transport companies, the most pressing problems associated with this increasing transport activity are traffic and operation incidents that will affect directly the quality of service provided.

In order to provide a high level of mobility to travellers, the public transport companies need to operate as efficiently as possible, and therefore manage effectively the traffic incidents. Business intelligence can help improving this effort, by developing a better way of capturing incident data and operational analysis.

Improvement in Data Analysis

Although there are some efforts from the Portuguese companies to analyze their incident data, they are unable to compare or relate it with other organizations in an efficiently or effectively manner. The existence of an occurrences standard platform will help professionals to improve occurrences management by better understanding what happened in the past and by the comparison of events from one company to another.

Improvement in Mobility

The impact of the analysis of the incident data is huge on the traveller’s mobility. Taking in account the “2007 Relatório de Sustentabilidade” [22]published by CARRIS, 40% to 50% of the incidents are related with traffic congestion or vehicle breakdown and therefore result in lost time. It is also mentioned in the report that almost 15% of the incidents are related to vandalism and that half of them results in broken windows, this requires the end the service and therefore lost time and poor quality of service. With a better analysis on this incidents, routes may be changed during specific time periods to avoid congestions or places, vehicles may be changed for more effective ones etc. In short, any activity that can reduce incidents results in significant benefits for the passenger and for the company.

[22] ***CARRIS.*** *Relatorio de Sustentabilidade. [Online] 2007. www.carris.pt/downloads/relatorio_sustentabilidade_2007.pdf.*

Improvement in Data Sharing

Considering for example OTLIS, a group of transport companies from the region of Lisbon, whose mission is:

"(..) to create value for Transport Operators, Grouped and Joiners through efficient management of the central systems and shared resources, seeking help to improve the quality of services of Operators to the citizens of the Metropolitan Area of Lisbon and investing in technological development of innovative products and services."

A common incident management platform would make immediate sharing of incident data within Lisboa, and between the different transport operators. Therefore, it will improve the regional awareness and would provide better knowledge of the incidents data.

3.3. Purpose and Scope

The purpose of this case-study is to proof that is possible to build an end-to-end open source Business Intelligence solution and to evaluate the current open source offer.

The scope is to pilot a Business Intelligence solution for transports focused on Incident Management and Quality of Service, and therefore it focuses on a limited set of functional content rather than providing an exhaustive coverage of the business functionality. In particular the objectives of the project are:

1. Determining the best integration architecture
2. Determining the most suitable open source technology to use within the project.
3. Validating that it is possible to embed BI artefacts within any external application.

The expected deliverables from this project are:

A prototype that illustrates the connection between SmartCITIES™ Operations Management and a set of BI Dashboards and Reports with the functionality here described and that illustrates that OSBI technologies are emerging, and that can be a competitive choice for any SMC. The prototype shall be comprised of:

- A design for the most appropriate data warehouse schema (facts and dimensions) to support the functional requirement
- A BI server with schema and meta-data to support the dashboard and the reports.
- A set of ETL scripts to move the data from the different operational data sources to the BI server.
- A set of reports, dashboard and analysis views that would help a transport operator to analyze incident and passenger information data.

3.4. SmartCITIES™

The incident management standard platform mentioned above and developed for this thesis uses as informational source systems some of the SmartCITIES™ offer, in particular: SmartCITIES™ Interoperable Fare Management and SmartCITIES™ Operations Management Integrated System. It is important to refer that the project isn't limited to this data sources, due to the staging data model designed it s possible to integrate with any other data system.

SmartCITIES™ offers a range of solutions and services based on Public Transport and with an extensive use of innovative technologies, such as smartcards, RFID and image processing. [23] It is divided into two main vertical areas:

- **SmartCITIES™ Interoperable Fare Management:** an integrated and modular system that implements a full and customizable electronic ticketing/smartcard system.
- **SmartCITIES™ Operations Management Integrated System:** Designed to transport operators or companies whose business requires the management of fleets or remote equipment in real time. The system helps the operational management of the fleet in real-time, supporting multiple services such as expedition, monitoring and maintenance (of vehicles or equipment) and management of events that occur during activity, such as accidents, breakdowns or security problems. The system supports not only the daily operations of public transport operators but is also used by operators of flexible mobility services, such as renting, car-sharing or security transport companies. Can be integrated with surveillance cameras to help detect bottlenecks of mobility in urban areas, such as public transport corridors (BUS) infringements and parking in the second row.

Both these areas are supported by SmartCITIES™ Embedded Framework, which helps the interaction with all kinds of equipment and hardware, thereby ensuring the independence of any equipment manufacturer.

3.5. Methodology

Business Intelligence projects pose a set of challenges that are unlike to be found during any development projects of operational systems. This happens because the main core of BI system is to condense data into useful and usable information that helps companies to develop business knowledge. A successful BI implementation should be a result of repeatable development processes that involve not only the project team but also the client and end users. Following a development process is crucial to any project.

[23] Link SmartCITIES - http://www.link.pt/conteudos/artigos/detalhe_artigo.aspx?idc=636&idsc=637&lang_id=1

Considering the following beliefs from the Agile Manifest[24] it seems that a BI project should follow an agile methodology:

- **Individuals and interactions over processes and tools:** It doesn't real matter what is used to build the software, what is really important is to understand the requirements and develop a system that fulfils them.
- **Working software over comprehensive documentation:** A software that actually provides the output that the client need is much more important than providing manuals.
- **Customer collaboration over contract negotiation:** What is really important is to understand what costumer need and not what is going to be built.
- **Responding to change over following a plan:** It is preferred to develop a system that will be in constant mutation rather than one with that has no flexibility.

Therefore it was chosen to use a methodology based on an iterative development.

Before explaining the methodology used it is important to refer that was not possible to contact directly the target end-users of the system. Therefore, it was considered that LINK consultants were the end-users of the pilot, a situation that it is not the ideal and that obviously limits the potential adaptation of the work done within a client.

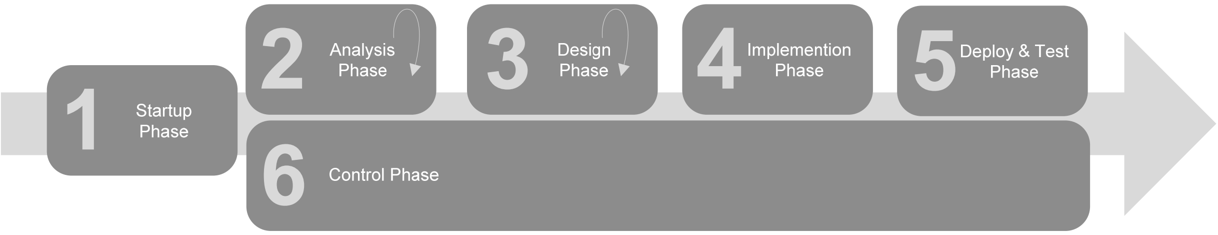


Figure 18 – Pilot Development Phases

As depicted on Figure 18, the pilot can be breakdown in six phases:

1. **Startup:** During this phase, a sketch of the initial architecture and pilot goals are drawn. It is also when business requirements and pilot end-users are identified. The objective of this step is to do some initial architecture "thinking".
2. **Analysis:** The result of this step should be an analysis of each 'Track' (see above). It means that this step includes the processes of collecting and analyzing the business requirements, data and data sources, technical requirements and metadata. It is an iterative step because each analysis depends on each other, and therefore one change can affect all analysis. During this stage, it is usual to **Control and Validate** the work done with the Link Consultants.
3. **Design:** During this phase, the data warehouse architecture and the technological environment are specified based on the requirements defined in the previous step. Like the previous step, it is usual to **Control and Validate** the work done with the Link Consultants.

[24] *Agile Manifest - http://agilemanifesto.org/*

4. **Implementation:** It consists in the development of the system planned in previous stages. In order to reduce risk, this phase was breakdown by project tracks.
5. **Deploy & Test:** In this stage each component build is deployed and tested. Minor bugs were solved "just-in-time".
6. **Control:** This stage occurs in parallel with the previous stages and can be characterized by being the "simulation" of end-users interaction. As mentioned before, it is related to the validation and feedback of the work done.

As mentioned on the Implementation phase, it was chosen to breakdown the BI project in more functional areas. This breakdown in parallel "Modules" or "Tracks" helps coordinate various activities, and optimize the resource utilization. It also results in more timely implementations because there is no need to be waiting for the end-user (link consultants) approval before hitting the next iteration or phase.

This project was breakdown in four tracks:

- **Project:** is focused on orientation and management of the pilot.
- **User Interface:** is focused on business requirements and the user interaction with the system.
- **Business Data:** is focused on data related tasks, like data source analysis, data model design, data warehouse design etc.
- **Technical:** is focused on technical component design, evaluation, selection and integration.
- **Metadata:** is focused on business data and metadata access and design.

In Appendix C there is a detailed diagram of the methodology followed in this project.

Chapter 4 – Architecture

In this chapter, it is presented the solution architectures that were designed during the startup and analysis phase. The first one is a high-level architecture that was a result of the initial meetings with the LINK consultants during the startup phase and that guided the creation of the others two during the analysis phase.

Since the solution and the system functionality drives the technology and not the other way around, it was chosen a top-down approach, starting in a high-level architecture and finishing in a technical-detail level.

4.1. Introduction

Before starting design any solution model it's important to have at least a general idea of the final system. Having a high-level architecture model helps development and communication between developer and end-user, because it gives an overview of what could be built. What is modelled in this architecture doesn't mean that it has to be developed. It just gives to the project parties an advantage of having a guiding vision of the system.

Therefore a meeting was arranged with the end-users (LINK consultants). During this meeting some key factors were taking in account for the high-level architecture:

- **Multiple data sources from multiple data providers:** One of the goals of this pilot is to offer a solution that isn't just focused in data from one client. It should allow several levels of entities, such as, company stations, company and companies associations. Today transport companies tend to have stations with independent management. However companies are also forming associations in order to get better insight and a more efficient management of their business. Therefore is important to have a hierarchical level of data analysis.
- **Different end-users:** Since that the system could aggregate data from several transport operators or related companies, and since that today the question of quality of service is getting more and more relevance, the system must have the ability to allow the access by different type of users, such as: normal passenger; companies; metropolitan transport authorities; etc.
- **Different means of interaction:** Due to the wide variety of end-users, the system must be proactive and customizable. This means that the system must allow users to create reports and schedule them to be sent by email, SMS or be available on a web portal.

Based on the above mentioned factors, and with the focus on the BI Architecture (explained in Chapter 2), the above high level architecture was built.

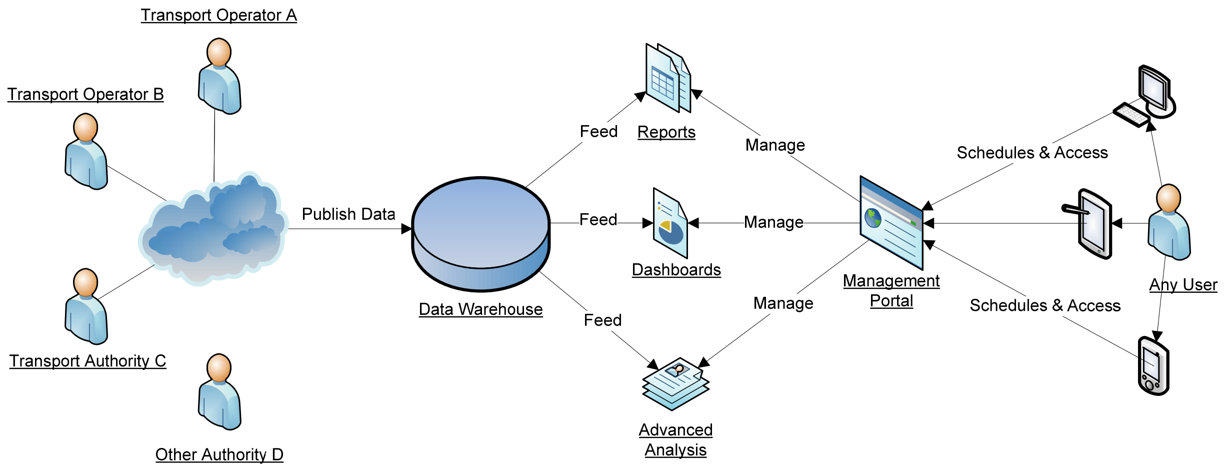


Figure 19 – High-Level Architecture Diagram

As depicted on Figure 19, transport operators or other authorities publish data in the data warehouse. This data will be transformed and integrated. After this the consolidated data will be used to feed the reports that the users created through the Management Portal. Besides the creation and edition of reports, dashboards and advanced analysis, the users should be able to schedule tasks like "generate report and send to email every first day of month".

4.2. Conceptual Architecture

The conceptual Architecture of the Incident Management Platform was the result of several iterations of design and analysis, which had as starting point the High-Level and the Data Warehouse models previously explained.

As illustrated below, there are three operational data sources:

- **Incident System**: Data source that contains the information about the incidents, such as accidents, interruptions and others.
- **Passenger Information System:** These types of systems are responsible for manage and deliver service schedule to the passenger.
- **Ticketing:** Data source system that contains the info about each ticket validated in the bus transport network.

This architecture relies on an extraction, transformation and load phase to store and manage the data from the several operational databases into a staging area and then into dimensional data marts that contain both atomic and summary data. This stage is characterized by:

- Contains a non-volatile repository of information, which contains the consolidation of the data that was published by the operational data sources.
- Works as an interface to the multidimensional data model, therefore is independent from the data sources.

After the data is stored into the staging area, it is ready to be loaded into dimensional data marts that contain both atomic and summary data. In this stage, information is organized and stored according to a dimensional model and which is available to be accessed. Typically this area is organized by date marts, usually organized by areas of information. In this solution is considered three data marts.

- **Incidents;**
- **Passenger Information;**
- **Ticketing;**

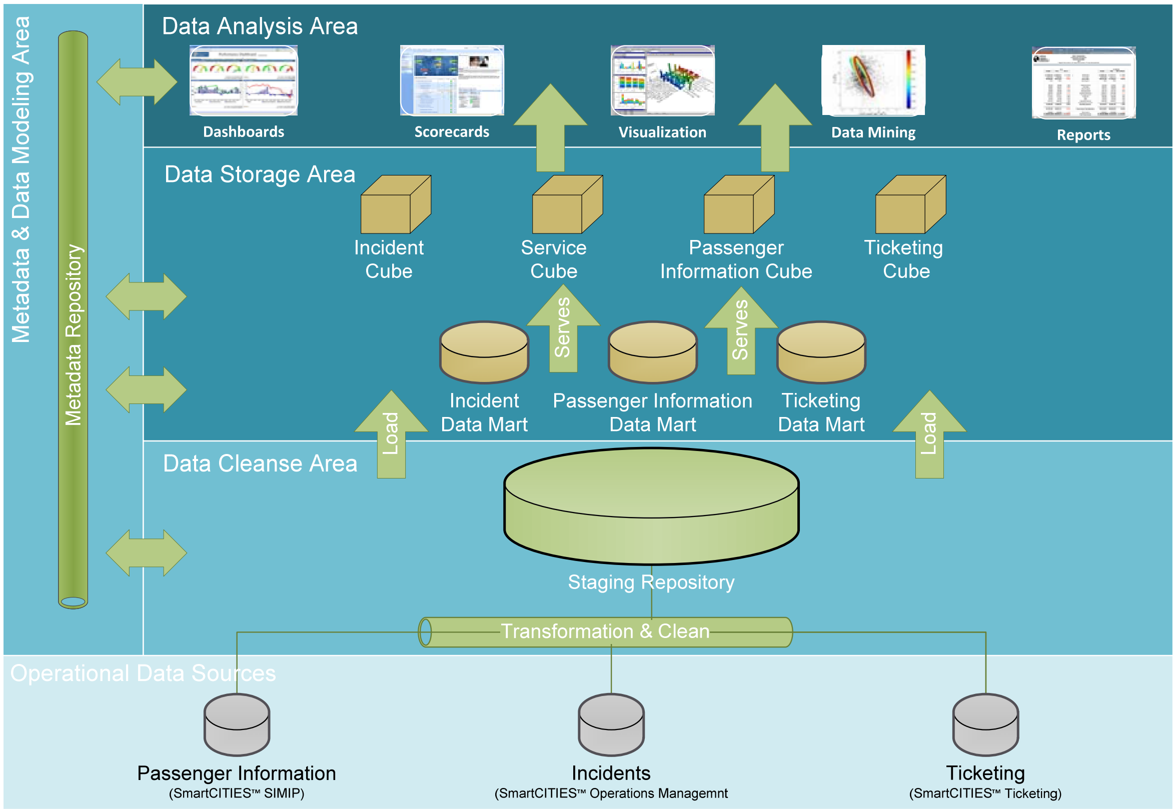


Figure 20 – Conceptual Architecture Diagram

4.2. Technical Architecture

This section details the architecture that is used to support the architecture described in the previous section. Only open source tools were considered.

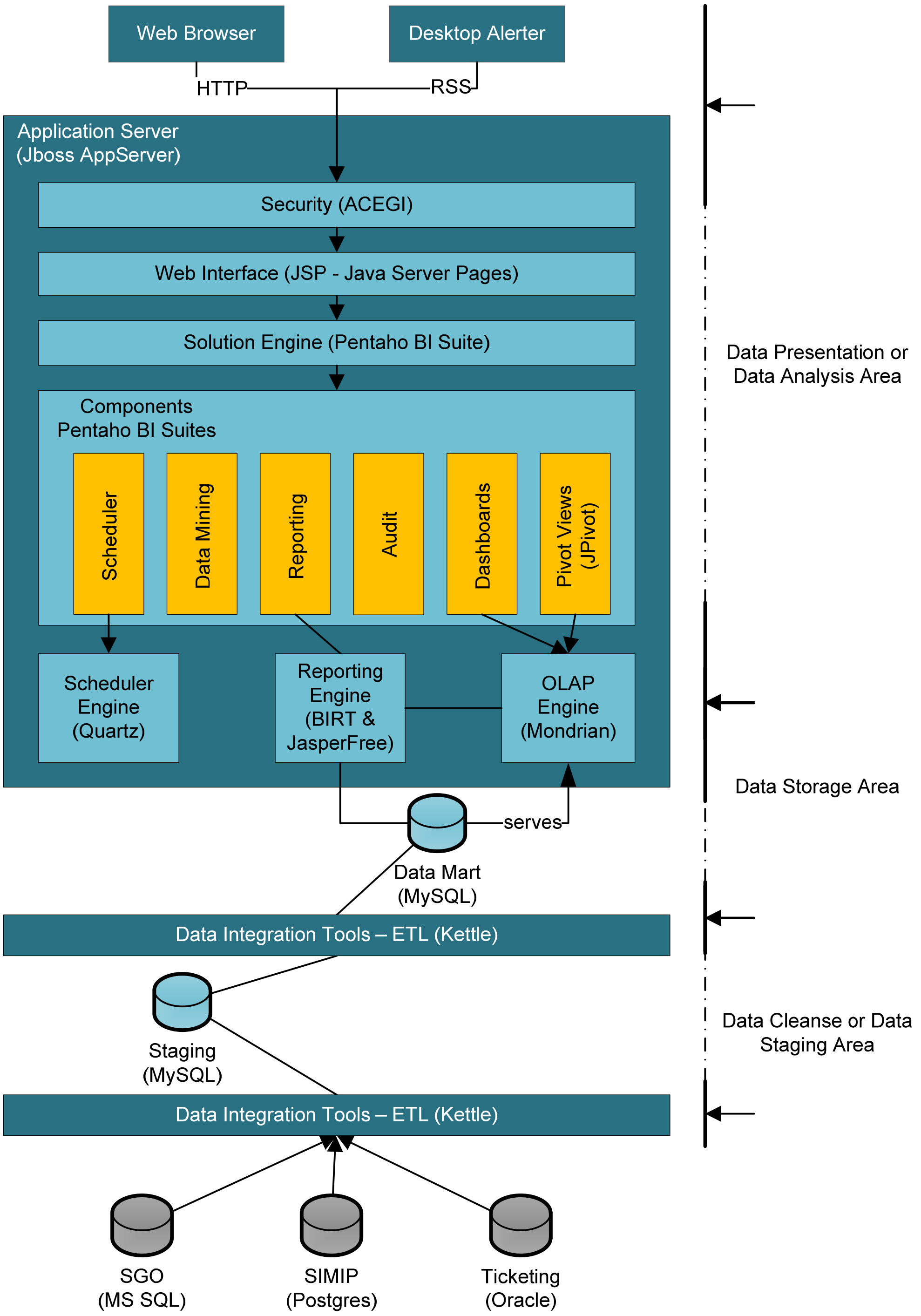


Figure 21 – Technical Architecture Diagram

As illustrated on Figure 21, the system has three stages (Operational Data Stage is not included in technical architecture):

The first layer (Data Cleanse) is supported by MySQL Server, the most known and used open source database engine. In this stage, ETL Tools are used for extracting and transforming the data from the operational data sources. From the list of ETL solutions I mentioned in the second chapter, two of them were more mature then the others: Kettle and Talend. The first one is a metadata driven framework (and is part of the Pentaho Offer), while the other is a code generator. I've chosen Kettle and the main reason for that choice was the active Kettle community. However a few months ago, Talend and JasperSoft joint efforts, and the Talend community has improved enormously since then.

The second layer is similar to the first one. In addition to the MySQL database server and the Kettle as the ETL tool used to load the data from the staging repository to the data mart and multidimensional model, this stage contains an OLAP Engine. The chosen engine was Mondrian (Palo was an alternative), mainly because it is part of Pentaho offer, and allows an easier integration with the Pentaho Suite.

The last layer is supported by JBoss, an open source application server that allows the build, deploy and orchestration of web application. Inside JBoss there are four applications running as a service. The first one is ACEGI. ACEGI is an open source security framework, which is used to manage users and permissions. The second application is a web portal that allows the authenticated user to navigate, schedule and create reports, dashboards or analysis views, in order words, is the user interface with the system. This last application is supported by Pentaho BI Solution, which integrates the web application with the different BI components (Reporting, Data Mining, Analysis, etc). The last application running on JBoss Server is a java scheduler, Quartz. This application allows the user to schedule tasks along time, per example, “Send the monthly passenger information report to all administrators on every first day of each month”.

Chapter 5 – Design

In this chapter, it is presented the result of the design phase. It will be shown the results of the Data Staging, Data Access and ETL design phase. In data staging and multidimensional model design was depicted the several data models, while the other two design steps resulted in a detailed specification of dashboards and ETL processes.

Like in other phases, the meetings with LINK consultants were vital to understand the problems and the needs of the transports operations and incidents management.

It was decided that the ticketing data source, would be used to test the ETL tool, given the number of validations received each day – approximately one million.

5.1. Staging Data Model

Data Staging (or Data Cleanse) Area is the place where data is merged from different operational source systems. This stage handles different formats of data sources, from flat files to databases and spreadsheets, which need to be cleansed, transformed and prepared to be loaded into the data warehouse.

Not all business requires a data warehouse staging area, for the passenger information data source used in this project it is feasible to use ETL to copy data directly into the data warehouse due to the light weight information. However, such approach could result in data loss due if a heavy processing needed. Also due to the varying business cycles, it is not feasible to extract all the data from all the operational data sources at the same time. For example, in this project, it is reasonable to extract passenger information data on a weekly basis, however, weekly extracts may not be suitable for ticketing or incidents that requires a daily extraction process.

For this model it was considered that the data that was extracted from the data source was based in a timestamp. This means that, only the data that have been changed since a specific timestamp will be extracted and loaded into the staging repository. Therefore every row as a timestamp of last update and a field that mentions which ETL process made the changes.

The staging data model was designed with the intention of building a data model that could be set as a framework or reference to the transport industry, that’s the main reason why it was created so many generic entities such as environment conditions and incident types. The first one contains the details about the road state and visibility while the second one contains all the incident types organizes in a hierarchical level.

Figure 22 represents the data model diagram for the Event Domain. The full data model can be found in the Appendix D.

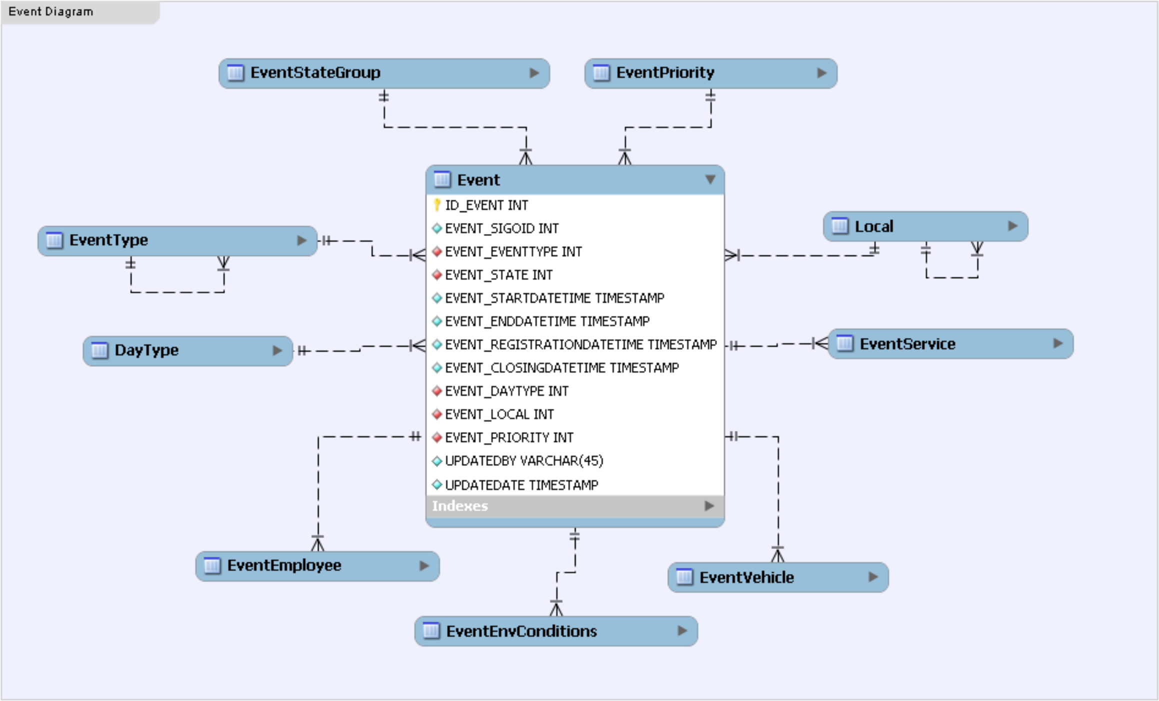


Figure 22 – Staging Data Model - Domain Event

Since that one of the objectives of the data staging area is to provide a place where the data can be checked for cleanliness and correctness, data profiling and data quality tools should be considered in this stage. These tools can help companies to fix data quality issues and to understand the problem and where it originates. However, only in middle/late 2008 the open source solutions for this type of tool started emerging [25], and therefore none was used.

5.2. Multidimensional Data Model

The multidimensional data model is an important part of OLAP, and must be designed to solve complex queries in real-time. As Ralph Kimball states, the design must be as simple as possible.

"The central attraction of the dimensional model of a business is its simplicity (…) that simplicity is the fundamental key that allows users to understand databases, and allows software to navigate databases efficiently." [(18)]

The model created for this prototype is composed of four logical cubes (organized for functional area), a set of measures and dimensions. It was design as simple as possible to allow an easy a meaningful

[25] ***DMReview.*** *Tallend Announces First Open Source Data Profiler. DMReview.com. [Online] 3 July 2008. [Cited: 28 July 2008.] http://www.dmreview.com/news/10001630-1.html .*

navigation for the business users. The next sections will explain these components of the multidimensional data model in detail.

5.2.1. Dimensions and Measures

The dimensions and measures detailed in this section came as result of several meetings with LINK consultants, and from the analysis of several documents provided by different transport operators.
This specification is the basis for the analysis and design of the multidimensional data model.
An overview of the dimensions is made below.

Dimension	Description
Company Station	Contains the details of the Company Station and Company.
Date	Contains information related with date.
Day Type	Contains the different types of days.
Employee	Contains information about an employee, such as function, name, sex and age.
Employee Final State	Dimension that contains de final situation of an Employee after an incident.
Employee Involvement	Contains information about which role an employee can have in an incident.
Environment	Contains information about the visibility, road state and weather conditions.
Event State	Has the information about the states of an incident (Closed, In Treatment etc)
Event Type	Have all the different types and sub-types of an incident.
Event Type Category	This information groups several event type's in categories.
Notified Entity	Contains information of who is notified of an incident.
Place	Has the information about the region and place of the incident.
Priority	Contains Information about the level of priority of an incident.
Season	Contains information about the Season (School Day, Holidays etc)
Service	Has the information of all services (line and service type)
Time	Contains details related to time (hour and minute)
Vehicle	Has all the details about a vehicle (license plate, age, vehicle class, vehicle type and category, etc)
Vehicle Final State	Dimension that contains de final situation of a vehicle after an incident.
Mobile Operator	Dimension that contains the different Mobile operators' details.
Request Type	Dimension that contains all the possible requests that can be made by the passenger
Response Type	Contains all the possible response that can be delivered to the passenger

Table 2 – Dimensions Overview

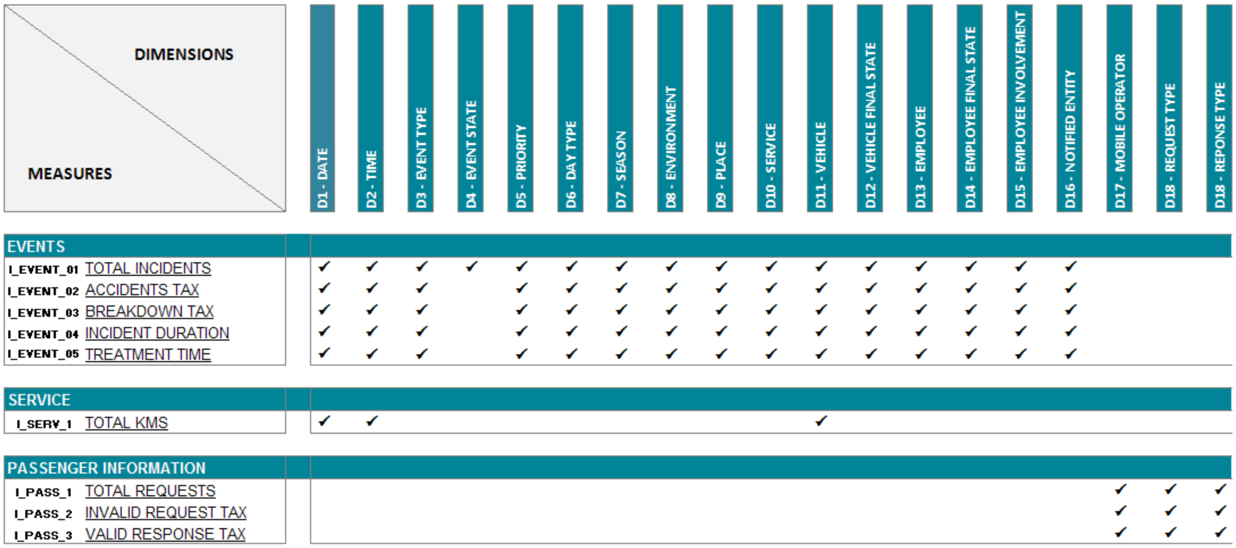
This design is limited, because it was only possible to take in account data from Incidents and Passenger Information. The ideal would be to cross information with other data sources such as service operation or maintenance. Therefore there were a limited number of measures.

Measure	Description
Total Incidents	Total Incidents
Accidents Tax	Total of Accidents per 1 000 000kms of service
Breakdown Tax	Total of Breakdowns per 100 000kms of service
Incident Duration	Total duration of an incident
Treatment Time	Time taken by an entity to finish their action.
Total Requests	Total requests from passengers.
Invalid Request Tax	Total of invalid requests per total of requests.
Valid Response Tax	Total of Valid response per total of requests
Total Service kms	Total of covered distance during a service.

Table 3 – Measures Overview

Based the analysis made above, it was prepared a matrix of measures and dimensions (see image 23). This matrix illustrates the matching between measures and the involved dimensions and helps modelling the data marts.

The measures are divided in three areas of analysis (pre-data mart): Incident, Passenger Information; Service.



DIMENSIONS MEASURES		D1 - DATE	D2 - TIME	D3 - EVENT TYPE	D4 - EVENT STATE	D5 - PRIORITY	D6 - DAY TYPE	D7 - SEASON	D8 - ENVIRONMENT	D9 - PLACE	D10 - SERVICE	D11 - VEHICLE	D12 - VEHICLE FINAL STATE	D13 - EMPLOYEE	D14 - EMPLOYEE FINAL STATE	D15 - EMPLOYEE INVOLVEMENT	D16 - NOTIFIED ENTITY	D17 - MOBILE OPERATOR	D18 - REQUEST TYPE	D18 - REPONSE TYPE
EVENTS																				
I_EVENT_01	TOTAL INCIDENTS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
I_EVENT_02	ACCIDENTS TAX	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
I_EVENT_03	BREAKDOWN TAX	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
I_EVENT_04	INCIDENT DURATION	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
I_EVENT_05	TREATMENT TIME	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
SERVICE																				
I_SERV_1	TOTAL KMS	✓	✓									✓								
PASSENGER INFORMATION																				
I_PASS_1	TOTAL REQUESTS																	✓	✓	✓
I_PASS_2	INVALID REQUEST TAX																	✓	✓	✓
I_PASS_3	VALID RESPONSE TAX																	✓	✓	✓

Figure 23 – Measures vs Dimensions Matrix

5.2.2. Data Mart Model

Based on the measures vs. dimensions matrix, it was possible to design the project data marts.

There are three data marts (Incidents, Service and Passenger Information) and due to size of each diagram, they can be found on Appendix E.

5.3. Data Access Design

Before the analysis and design of dashboards, reports and analysis views it was necessary to define and understand the end-user roles. During the analysis phase it was decided that it will be three different roles for users:

- **Administrator:** Normally this role is given to members of the board. Thus the available indicators must be coupled with quality and reliability of service.
- **Maintenance:** Users that are associated with the maintenance profile are responsible for managing company’s vehicles and equipments. As a result it is important to show indicators related to the resources, and to incidents that have direct impact on them, such as breakdowns or accidents. It is also significant to focus on indicators that can evaluate the level of quality of maintenance service, this can be measured by the number of vehicles that have the same breakdown repeatedly during a specific period.
- **Operations:** The role of operations is used by those who are responsible for managing the operation (or services) of the transport operator. For that reason, indicators should focus on a more operational and detailed level providing a way to examine the services by security, liability or reliability of services and usage, availability and performance of the drivers and vehicles.

5.3.1. Dashboards

As mentioned on Chapter 2, Dashboards are the first layer of interaction with user. They usually contain data summarized, which consequently enables users to quickly understand and examine critical events that are affecting the company. It's important that the dashboard display information in an obvious and intuitive style and that provide the ability to "drill down" to other hierarchical levels as required.

Figure 24 represents a diagram that shows the dashboards designed for each user role.

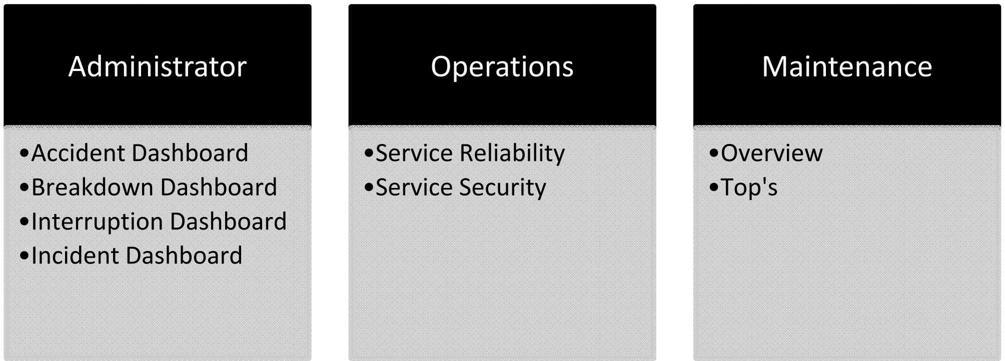


Figure 24 – Dashboards per User Role Diagram

As depicted in figure above, eight dashboards were designed, distributed across the three roles considered.

For the administrator role, there were designed four dashboards. Each one of them gives an overview of the events during the month in analysis.

During the design phase, four concepts were associated with the operations management: Reliability, Security, Usage and Availability. The first two are related to the service and allows the user to understand and evaluate the quality of service offered to the passenger, while the other two are coupled with the company's resources, vehicle and driver, and help the end-user on the selection of which resource should be used for a specific service. However, only the first two concepts, reliability and security, were possible to design and implement, because there isn't available information on the services and resources.

Finally, the maintenance dashboards focus on the breakdowns and accidents by vehicle and places. These dashboards provide to end-user information on the places and vehicles with most events. It also provides details about the quality of maintenance service by comparing the number of repeated events for the same vehicle during a specific time.

5.3.2. Reporting

When the information available on the dashboards is not enough to understand a critical event that affects the organization, the end-user will resort to the reporting layer. Therefore, the reports should focus on a more detailed level of information that the one displayed in the dashboard.

For this project, there were design four reports that show information about the different categories of events (Breakdowns, Incidents, Accidents and Interruptions) and one report to detail information on the company fleet. There also designed a report on Passenger Information, which allows an analysis focused on quality of service and on clients, allowing the end-user to understand information like the first-use tax or the number of request by service and bus stop.

If the designed reports don't meet the end-user needs there is an ad-hoc query reporting page that allow the user to generate custom reports.

5.3.3. Analysis Views

Analysis Views is the more detailed level available to the end-user to explore the data warehouse. With an analysis view the user may explore, ‘drill’ or ‘slice’ the multidimensional cubes.

The only analysis views designed are the ones associated with the Administration dashboards that allow the user to analyze the information of each event type during a specific period. With this analysis, the user may cross the event type with other dimension of the cube.

Similar to the ad-hoc reporting tool, there is available a page that allows the user to create analysis views on the existent cubes.

5.4. ETL Design

Normally on a data warehousing project the ETL design and development take most of the development work spent on. The success of a data warehouse project is dependent on solid data preparation, and if the ETL is poorly design, few knowledge can be taken from the information and in result the system will be worthless.

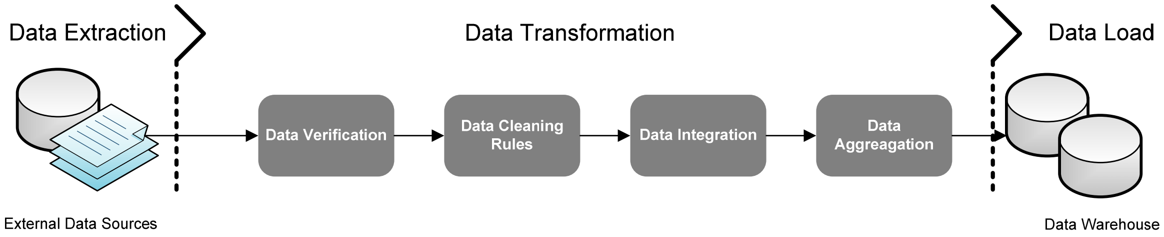


Figure 25 – Typical ETL Process

Figure 25represents a typical ETL process. It consists in three steps. Extract data from one or several data sources, transform it to be compliant with the destination schema and load to the destination data source. While extraction and load steps are straightforward, data transformation is the vital step in the process, because is where it can be performed a variety of transformations that ensures data quality. These transformation options include data verification, data cleaning, data integration and data aggregation.

Since there is the possibility that poor-quality data will be generated in external data sources, it is important to implement a data verification stage to reject invalid data before it can affect the remaining ETL process and the data warehouse. This validation of data consists of several data checks, including some basic concepts such as:

- Check if the value is valid within the domain.
- Check is the foreign keys are valid.
- Check if the row data is valid.

The data cleaning corresponds to the process of making data more meaningful and precise. For example, it is where data is merged from different sources or converted to other data types.

Data integration is a process of consolidating multiple data sources into a dimension or a fact table that are used for data analysis, and that are easy for users to explore.

Data aggregation is related to the process in which information is gathered in a summary table. It means that an attribute value is derived from the aggregation of two or more characteristics.

The objective of this ETL design is to study the interfaces from the external data sources and staging area, the respective data flow and transformations needed.

As stated on the previous chapter, there are two ETL stages in this project. One to extract data from external data sources, aggregate them and load into the data staging and another to load into the data warehouse.

In Appendix F there is a table with the analysis made for each ETL transformation. It shows the source and destination and also the required transformation steps.

Chapter 6 – Implementation

In this chapter is presented an overview of the prototype developed, as well as, an analysis of each tool that was used to develop and support the designed architecture.

6.1. ETL

To implement the ETL processes it was chosen Pentaho Data Integration, also known as Kettle, which is a free, open source ETL tool. The product name is an acronym for "Kettle Extraction, Transport, Transformation and Loading Environment".

Kettle is built in java, and consists of five separate applications:

- **Spoon:** is a graphical user interface tool to model the data flow or in other words to design the transformation.
- **Pan:** is a tool that 'executes the transformations modelled with 'the spoon'.
- **Chef:** is a GUI tool to model jobs. A job consists of a set of steps that allow the user to execute and control several transformations. Some of the steps are, download files through FTP, execute transformations files, send email, etc.
- **Kitchen:** is where the Chef produces his recipes, for that reason is a tool that executes the jobs created with Chef.
- **Carte:** Carte is a web server that allows the execution of transformations remotely.

The models designed in Spoon or Chef can be saved to a XML file, or they can be stored into a repository. When handling many models a repository can be a major advantage, however I don't recommend it when designing a model, because the application response becomes slow.

Below is an example of an ETL processed modelled in spoon.

Although the transformation shown below is very simple, it does illustrate the two most important 'ingredients' of an ETL model: steps and hops:

- The steps are the square icons showed in the graphical view. A step denotes a kind of action that is performed on data. It is possible to parameterize the step in order to specify his exact behaviour. Kettle offers a wide range of steps, spread through a wide range of categories.
- Hops are the lines that connect the steps, and specify the data flow.

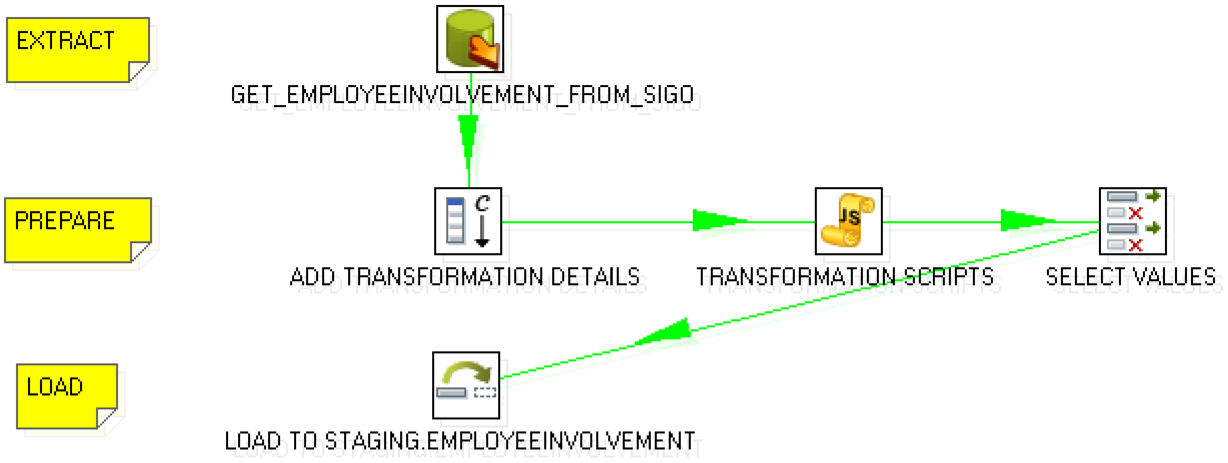

EXTRACT
PREPARE
LOAD
GET_EMPLOYEEINVOLVEMENT_FROM_SIGO
ADD TRANSFORMATION DETAILS
TRANSFORMATION SCRIPTS
SELECT VALUES
LOAD TO STAGING.EMPLOYEEINVOLVEMENT

Figure 26 – Employee Involvement ETL Process

By just looking at the graphical representation of the process (figure 26), it is easy to imply what this ETL process does. First, in the step named 'GET_EMPLOYEE_FROM_SIGO', data is read from a MSSQL Database Table. From there, the data then flows into two steps named 'ADD TRANSFORMATION DETAILS' and 'TRANSFORMATION SCRIPTS', the first one just add control variables to the row while the second one makes transformation on the data retrieved. Before loading to the data staging, the values to insert are selected in the step 'SELECT VALUES'. Finally the data is ready to be load into the staging database, to perform this action is used the step named 'LOAD TO STAGING EMPLOYEEINVOLVEMENT'.

Due to the number of processes designed, the above one was chosen because it is similar to the majority of processes.

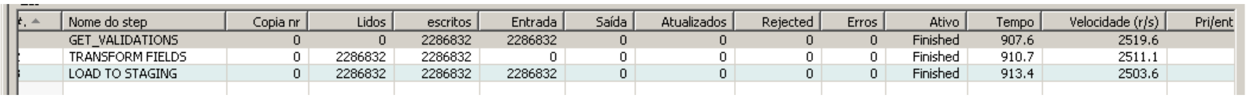
One advantage of using kettle as data integration tool, is the possibility of execute the transformations in the web portal. In this way, the end-user may refresh the data warehouse data by executing or scheduling kettle transformations.

To test the performance of Kettle, it was created a process that extracts data from the Ticketing data source and loads to the data staging.

The ticketing system generates nearly one million data transactions per day. It translates to a nightmare for any data integration tool. Therefore it is a great example to test Kettle capabilities.

The ticketing process load the values from the data source on the step named 'GET_VALIDATIONS' then the data flow to a step called 'TRANSFORMATION FIELDS' where data types are changed and attributes formatted, then on the step 'LOAD TO STAGING' the transformed data is loaded into the staging repository.

The results obtained are shown below.



#.	Nome do step	Copia nr	Lidos	escritos	Entrada	Saída	Atualizados	Rejected	Erros	Ativo	Tempo	Velocidade (r/s)	Pri/ent
	GET_VALIDATIONS	0	0	2286832	2286832	0	0	0	0	Finished	907.6	2519.6	
	TRANSFORM FIELDS	0	2286832	2286832	0	0	0	0	0	Finished	910.7	2511.1	
	LOAD TO STAGING	0	2286832	2286832	2286832	0	0	0	0	Finished	913.4	2503.6	

Figure 27 – Kettle Performance: Overview Ticketing Process

Over 2 million results were extracted from the ticketing data source (Oracle). This process had an execution time of 913 seconds, which is equivalent to process approximately 2505 lines per second.

The figures below show the speed (rows per second) of extraction the data from ticketing data source and the speed of writing results to the staging database throughout the process, with an interval of 30 seconds between each snapshot.

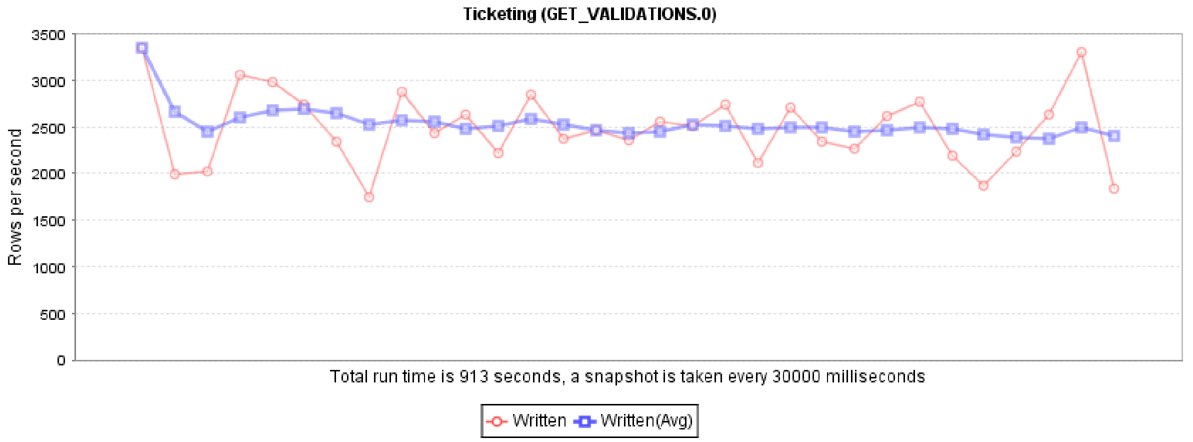


Figure 28 – Kettle Performance: Read from Ticketing data source

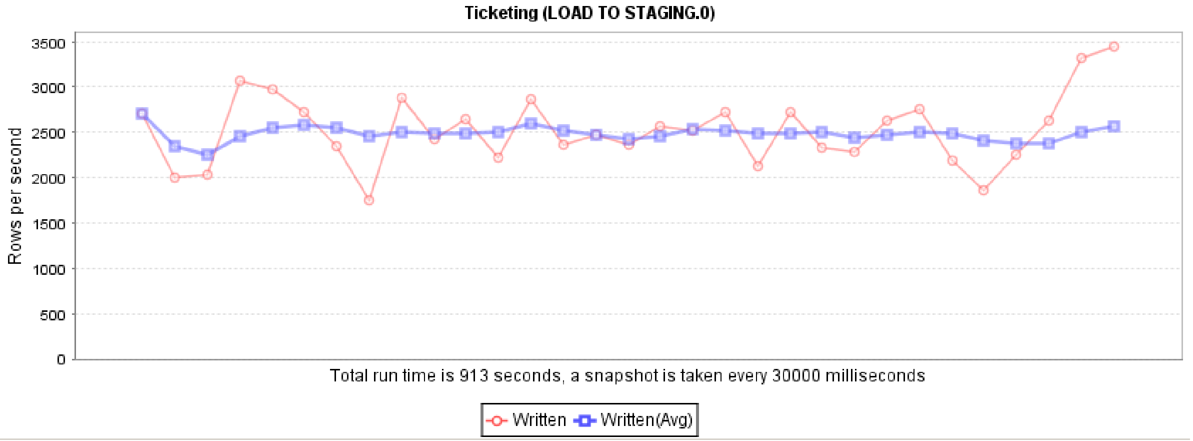


Figure 29 – Kettle Performance: Write to Data Staging Repository

During the process the tool remained stable – there weren't big oscillations in the number of rows read/written per second. Kettle also allows the execution of processes in clustering, thus easing the burden of a machine, and treat cases with data load much higher than the above tested.

6.2. OLAP Cubes

Mondrian was the OLAP engine chosen for this project. It is written in Java and it executes queries written in the MDX language, from data stored in a relational database, and presents the results in a multidimensional format via a Java API.

The Mondrian architecture consists of four layers:

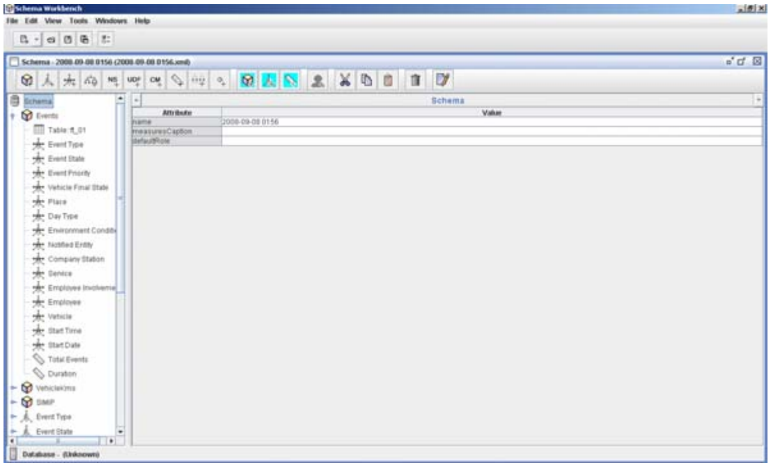
- **Presentation layer:** The presentation layer is responsible for the data presentation and the interaction with the end-user. There are many ways to present multidimensional datasets, including pivot tables, charts and other advanced visualization.
- **Dimensional layer:** This layer is responsible for parsing, validation and execution of MDX queries.
- **Star layer:** It is responsible for maintaining a cache of the aggregate data.
- **Storage layer:** Is the layer that contains the relational database and is responsible for aggregating the data and dimensions tables.

To define a multi-dimensional database in Mondrian, it is necessary to define a schema. This schema contains the specification of a logical model, which includes cubes, hierarchies, levels, members and the mapping onto a physical model. These schemas are represented in a XML file.

The schema used in this project was design through Mondrian Schema Workbench, which is a GUI that allows the user to create and test Mondrian OLAP cubes schemas.

The Mondrian Schema Workbench is a designer interface that allows you to create and test Mondrian OLAP cube schemas visually.

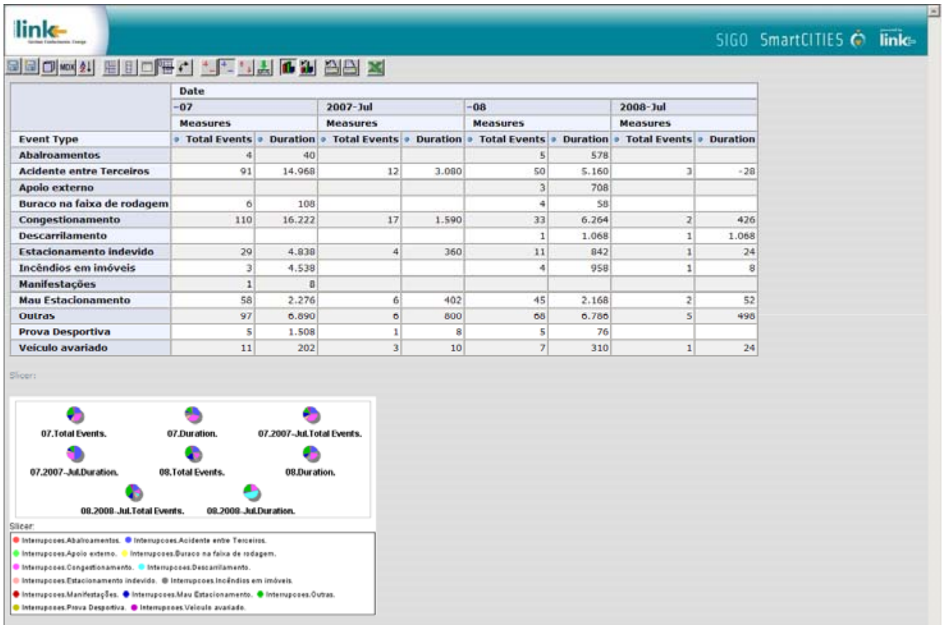
Figure 30 is a screenshot of the workbench interface.



Attribute	Value
name	2008-09-08 0156
measuresCaption	
defaultRole	

Figure 30 – Mondrian Workbench GUI

The advantage of using Mondrian as the OLAP engine for this project is that it can be easily integrated in the web portal. Thanks to JPivot, an open source project it is possible to have a web interface to explore the data cubes. It allows users to perform typical OLAP operations like slice and dice, drill down and roll up. JPivot also supports XMLA data source access, which means that it can be connected with Microsoft Analysis Services. Figure 31 is an example of the integration of Mondrian and JPivot on the project web portal. The remaining analysis views are depicted in Appendix I.



	Date							
	-07		2007-Jul		-08		2008-Jul	
	Measures		Measures		Measures		Measures	
Event Type	Total Events	Duration	Total Events	Duration	Total Events	Duration	Total Events	Duration
Abalroamentos	4	40			5	578		
Acidente entre Terceiros	91	14.968	12	3.080	50	5.160	3	-28
Apoio externo					3	708		
Buraco na faixa de rodagem	6	108			4	58		
Congestionamento	110	16.222	17	1.590	33	6.264	2	426
Descarrilamento					1	1.068	1	1.068
Estacionamento indevido	29	4.838	4	360	11	842	1	24
Incêndios em imóveis	3	4.538			4	958	1	8
Manifestações	1	8						
Mau Estacionamento	58	2.276	6	402	45	2.168	2	52
Outras	97	6.890	6	800	68	6.786	5	498
Prova Desportiva	5	1.508	1	8	5	76		
Veiculo avariado	11	202	3	10	7	310	1	24

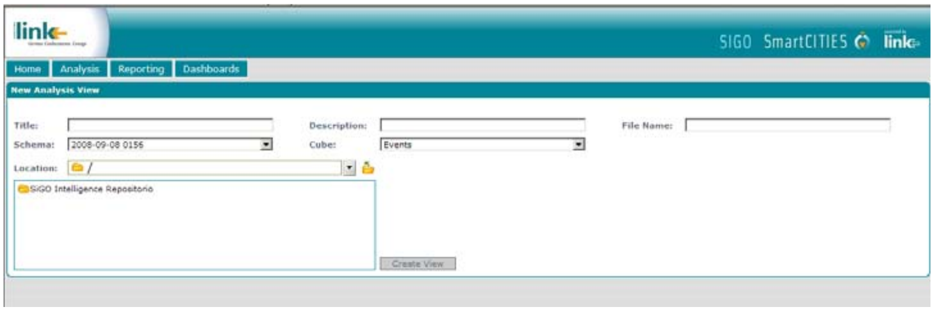
Slicer:

Slicer:

● Interrupcoes.Abalroamentos. ● Interrupcoes.Acidente entre Terceiros.
● Interrupcoes.Apoio externo. ● Interrupcoes.Buraco na faixa de rodagem.
● Interrupcoes.Congestionamento. ● Interrupcoes.Descarrilamento.
● Interrupcoes.Estacionamento indevido. ● Interrupcoes.Incêndios em imóveis.
● Interrupcoes.Manifestações. ● Interrupcoes.Mau Estacionamento. ● Interrupcoes.Outras.
● Interrupcoes.Prova Desportiva. ● Interrupcoes.Veiculo avariado.

Figure 31 – Screenshot – Analysis View 'Interrupções'

Mondrian can also be integrated with Pentaho BI Suite, and the result is a wizard to create a new Analysis View. The user just needs to select the data cube he wants to explore.



Title:
Description:
File Name:
Schema: 2008-09-08 0156
Cube: Events
Location: /
SIGO Intelligence Repositorio

Figure 32 – Screenshot – New Analysis View Wizard

6.3. Metadata

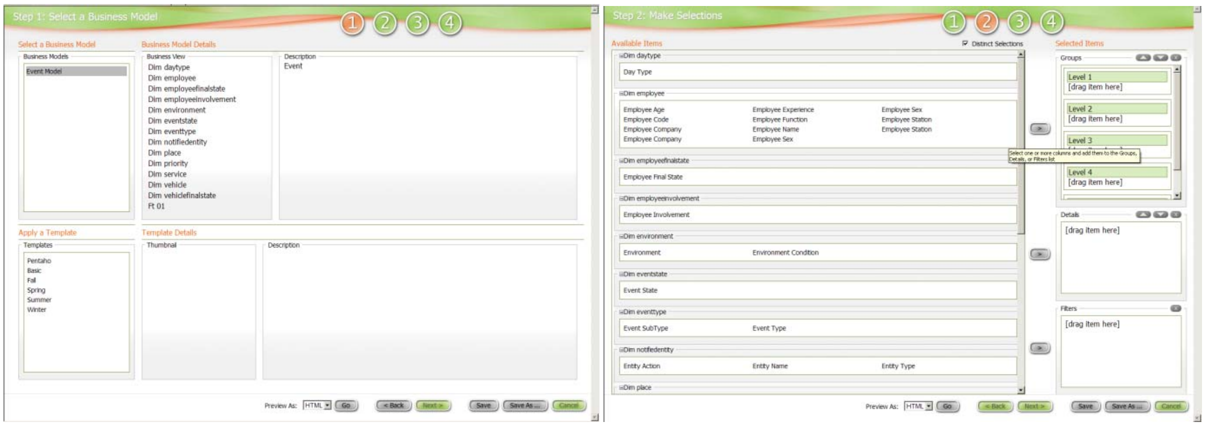
In a project of this nature, that involves a range of end-users with different perspectives on the data and information, it is difficult to build and maintain a common model that would be valid in all the different perspectives. For example, an end user would like to consult a model that is targeted to his business, a developer would be more interested in a model focused on the physical data, and if the end-user is a vendor, what matters to him is to have a model that is an abstract vision of the business and can be used for any company of a given industry. Therefore, it would be needed a model that could do the mapping between the business model, with an abstract view of the business and the physical model data.

Metadata appears as a solution to the problem mentioned above. For this thesis was chosen to use the specifications of Common Warehouse Meta (CWM). This model helps the interchange of data among warehouse, business intelligence and portal. It is based on three standards:

- **Unified Modelling Language (UML):** It is an object-oriented modelling language that supports the visualization, specification, construction and the documentation of artefacts of any software system.[26]
- **Meta Object Facility (MOF):** Is a set of interfaces that defines and manipulates meta-models and their matching models. [27]
- **XML Metadata Interchange (XMI):** The objective of this standard is to enable the interchange of metadata between the above mentioned standards.[28]

Besides the main BI vendors, such as SAS, IBM, Oracle, Informatica, Hyperion and Cognos, also Pentaho has active projects supported by this metadata model.

To build metadata models Pentaho developed a tool named Pentaho Metadata Editor (PME) that maps a physical database into a business model resulting in business views. These business views are the part of the business model that applications will operate against, and end-users will interact with. The integration of these business views in the portal resulted in the Pentaho Ad hoc Reporting Interface (see figure 33), which allows users to create simple reports from the available business views.



Select one or more columns and add them to the Groups, Details, or Filters list

Figure 33 – Screenshot – Ad-Hoc Query Reporting

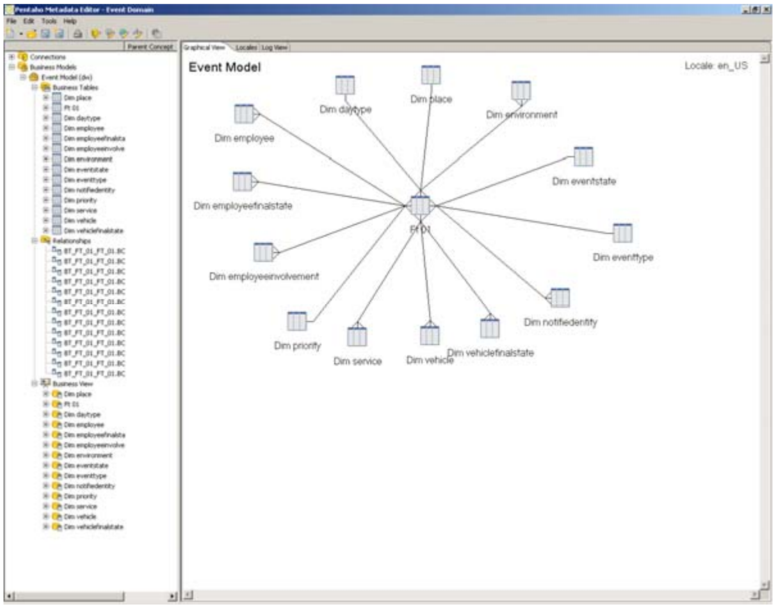
For this project it was designed and implemented a simple metadata model for the event domain. It was used the Pentaho Metadata Editor explained above and the result was an XMI file.

Depicted in Figure 34 are the domain implemented and the interface of the PME.

[26] *http://www.service-architecture.com/web-services/articles/unified_modeling_language_uml.html*

[27] *http://www.service-architecture.com/web-services/articles/meta-object_facility_mof.html*

[28] *http://www.service-architecture.com/web-services/articles/xml_metadata_interchange_xmi.html*



..

Figure 34 – Event Domain Metadata Model

6.3. Web Portal

A web portal allows the sharing of business information across departmental limits, integrating the company business intelligence and consequently sharing a common view of the business between all users. However how can be this accomplished?

First it is necessary that all of the BI components, such as reporting, OLAP and dashboards, and features like scheduling can be available over the web browser. Secondly it is important to have a common repository where all the BI artefacts are stored, and at last, there must be security mechanisms that guarantee user authentication.

6.3.1. Security

In the Pentaho BI Platform security is based on the infrastructure provided by the ACEGI Security Systems.

ACEGI is an open source security framework that implements security within a web application. This framework can be integrated with JA-SIG's open source Central Authentication Service (CAS)[29] to provide enterprise single sign on.

With the integration of Pentaho and ACEGI, it is possible to control de access of each user to each report or other BI artefact.

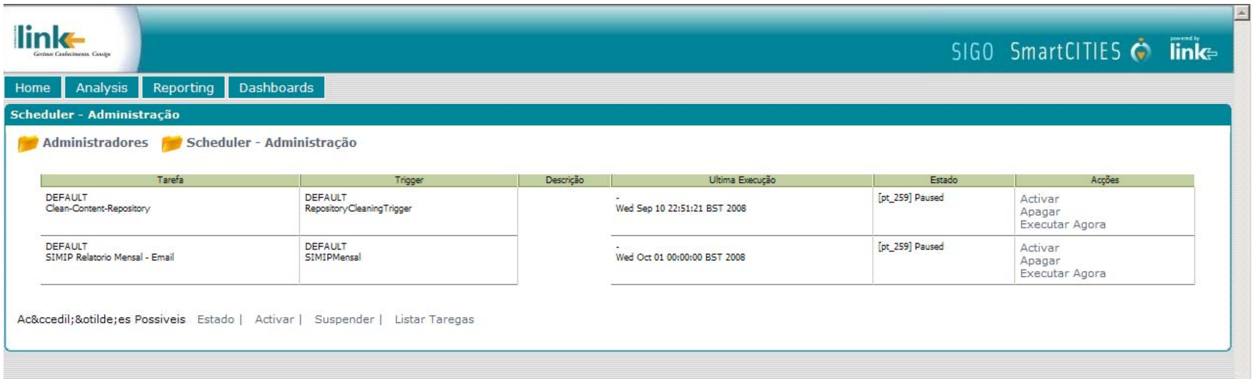
For this pilot there were created three users with full administration privileges and full access to all folders of the repository. Below is a list o the username and passwords created:

[29] *http://www.ja-sig.org/products/cas/*

- admin/admin
- exploração/exploração
- manutenção/manutenção

6.3.2. Scheduler

The Pentaho BI platform currently uses Quartz as its job scheduler. Quartz is an open source job scheduling system that can be integrated with any J2EE. It is fault tolerant and fault recoverable and consequently the misfires jobs can be handled.



Scheduler - Administração

Administradores Scheduler - Administração

Tarefa	Trigger	Descrição	Ultima Execução	Estado	Acções
DEFAULT Clean-Content-Repository	DEFAULT RepositoryCleaningTrigger		- Wed Sep 10 22:51:21 BST 2008	[pt_259] Paused	Activar Apagar Executar Agora
DEFAULT SIMIP Relatorio Mensal - Email	DEFAULT SIMIPMensal		- Wed Oct 01 00:00:00 BST 2008	[pt_259] Paused	Activar Apagar Executar Agora

Ac&ccedil;&otilde;es Possiveis Estado | Activar | Suspender | Listar Taregas

Figure 35 – Screenshot – Scheduler Administration

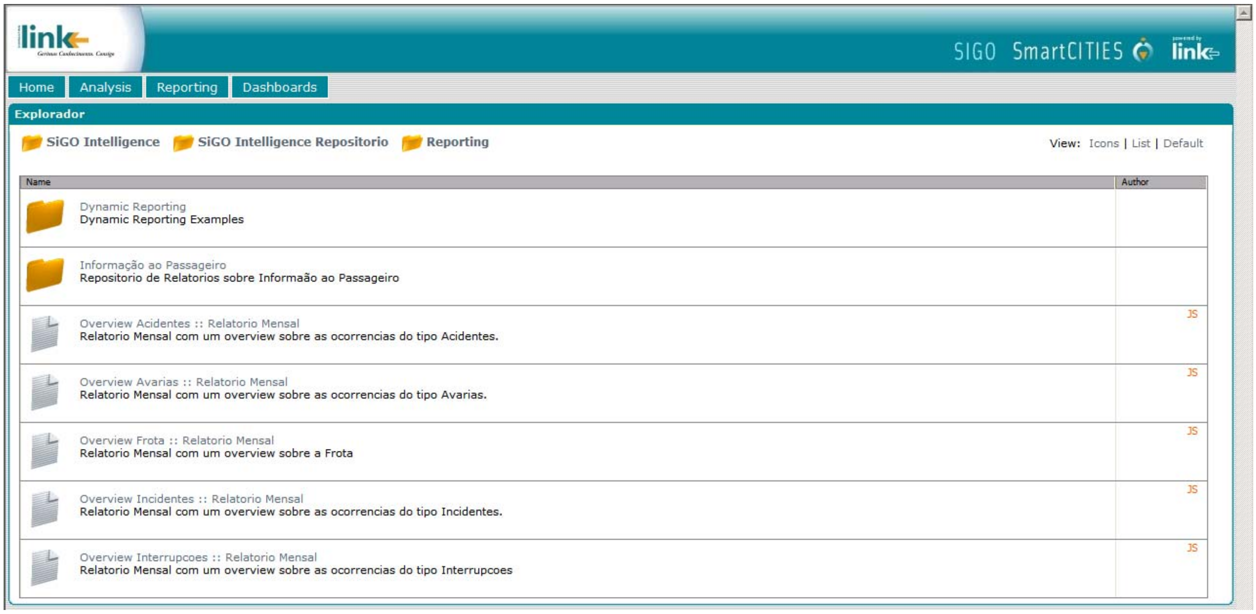
Picture 35 shows the scheduler administration integrated with the BI Portal. There are four different actions available to the end-user: start, suspend, resume and delete a job. It also shows the scheduler created to test the mechanism. The task is set to run every first day of each month and it generates the SIMIP Report and sends it to a list of emails.

To create the above mentioned task it was necessary to create two action files. One action file invokes the BIRT Engine component to generate the Passenger Information report and sends it to a list of emails. While the other one invokes the Scheduler component with the specific periodicity (Execute every first day of the month) and the action to be executed (the previously created).

6.3.3. Solution Engine

The Solution Engine is the crucial component within Pentaho BI Platform. It operates as an orchestrator, since that all the requests are forward to the solution engine and then are routed to the appropriate component or components for execution. These requests are known as actions, and they are XML documents that specify the parameters, resources and settings required to execute one task. They are normally associated to one component, which is an interface to another application. These Action definitions are stored within a solution. To help design and publish this action files, Pentaho provides a graphical environment for designing and testing of Action documents.

For this project it was created one solution that contains the action files for execute the dashboards, reports and analysis views. It is possible to navigate thru this solution and action files on the web portal (see figure 36).



Name	Author
Dynamic Reporting Dynamic Reporting Examples	
Informação ao Passageiro Repositorio de Relatorios sobre Informaão ao Passageiro	
Overview Acidentes :: Relatorio Mensal Relatorio Mensal com um overview sobre as ocorrencias do tipo Acidentes.	JS
Overview Avarias :: Relatorio Mensal Relatorio Mensal com um overview sobre as ocorrencias do tipo Avarias.	JS
Overview Frota :: Relatorio Mensal Relatorio Mensal com um overview sobre a Frota	JS
Overview Incidentes :: Relatorio Mensal Relatorio Mensal com um overview sobre as ocorrencias do tipo Incidentes.	JS
Overview Interrupcoes :: Relatorio Mensal Relatorio Mensal com um overview sobre as ocorrencias do tipo Interrupcoes	JS

Figure 36 – Screenshot – Solutions Explorer

One advantage of this solution engine is that it is independent of the Web Portal. This means that if I have an external application and I would like to embed any report or dashboard, it will only be necessary to create an action, deploy it on the solution repository, and open the link to the action. For example considering the action of generating the Passenger Information Monthly Report (SIMIP_REPORT_Mensal.xaction), if I would like to embed it on another application I would just need to open the link[30] and the report would be embed.

6.4. Dashboards

As mentioned in the previous chapter, eight dashboards were designed for three different user roles.

The dashboards implementation didn't follow the guidelines suggested by Pentaho. Instead it was used a dashboard framework developed by the Pentaho community[31]. In comparison with the Pentaho approach, this framework provides a more generic and agile way to implement and publish dashboards mainly because, there is no need to create additional pages in order to deliver a new dashboard, and adding a new dashboard is easy as adding a new folder containing the action files required for each artefacts on the solution repository to the solution repository.

[30] */ViewAction?solution=samplesMSC&path=reporting/SIMIP&action=SIMIP_REPORT_Mensal.xaction*

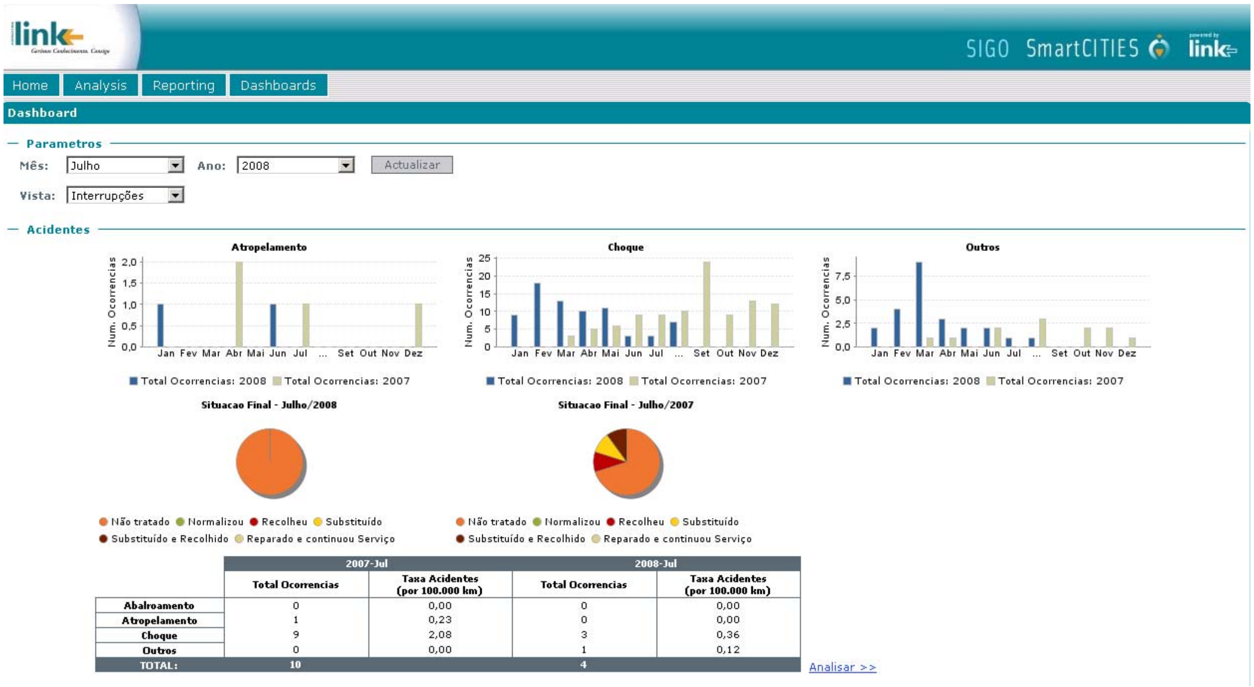
[31] *http://wiki.pentaho.com/display/COM/Community+Dashboard+Framework.*

The Community Dashboard Framework is already integrated with wide range of components, from simple date pickers and text boxes, to Google Maps or OpenStreetMaps.

To implement a dashboard with this framework, it is needed:

- An action file for each artefact used (charts, reports etc).
- An HTML file that contains the content layout and the definition of each artefact.

One of the dashboards created was the Accidents overview (see figure 37). This dashboard was built with a mixture of Charts created with Pentaho Reporting and a Pivot Table created with BIRT.


link Gerimos Conhecimento. Consigo
SIGO SmartCITIES
powered by link
Home
Analysis
Reporting
Dashboards
Dashboard
Parametros
Actualizar
Acidentes
Atropelamento
Num. Ocorrencias
2,0
1,5
1,0
0,5
0,0
Jan Fev Mar Abr Mai Jun Jul ... Set Out Nov Dez
■ Total Ocorrencias: 2008 ■ Total Ocorrencias: 2007
Choque
Num. Ocorrencias
25
20
15
10
5
0
Jan Fev Mar Abr Mai Jun Jul ... Set Out Nov Dez
■ Total Ocorrencias: 2008 ■ Total Ocorrencias: 2007
Outros
Num. Ocorrencias
7,5
5,0
2,5
0,0
Jan Fev Mar Abr Mai Jun Jul ... Set Out Nov Dez
■ Total Ocorrencias: 2008 ■ Total Ocorrencias: 2007
Situacao Final - Julho/2008
● Não tratado ● Normalizou ● Recolheu ● Substituído
● Substituído e Recolhido ● Reparado e continuou Serviço
Situacao Final - Julho/2007
● Não tratado ● Normalizou ● Recolheu ● Substituído
● Substituído e Recolhido ● Reparado e continuou Serviço
Analisar >>

Mês: Julho Ano: 2008

Vista: Interrupções

	2007-Jul		2008-Jul	
	Total Ocorrencias	Taxa Acidentes (por 100.000 km)	Total Ocorrencias	Taxa Acidentes (por 100.000 km)
Abalroamento	0	0,00	0	0,00
Atropelamento	1	0,23	0	0,00
Choque	9	2,08	3	0,36
Outros	0	0,00	1	0,12
TOTAL:	10		4	

Figure 37 – Accidents Overview Dashboard

To build this dashboard it was needed to:

- Design and implement an html page that defines the layout of the page.
- Implement a “listeners” system that would allow the refresh of variables that are used to generate the graphs and consequently to refresh the charts and pivot tables. For this prototypes the listeners considered was Month and Year.
- Design the Charts in XML files that are invoked in an Action file that is responsible to output the results to the html page created.
- Create an action file for each of the charts.
- Create a pivot table report in BIRT and an action file to generate it and insert in the html.

The remaining dashboards implemented can be found in the Appendix G.

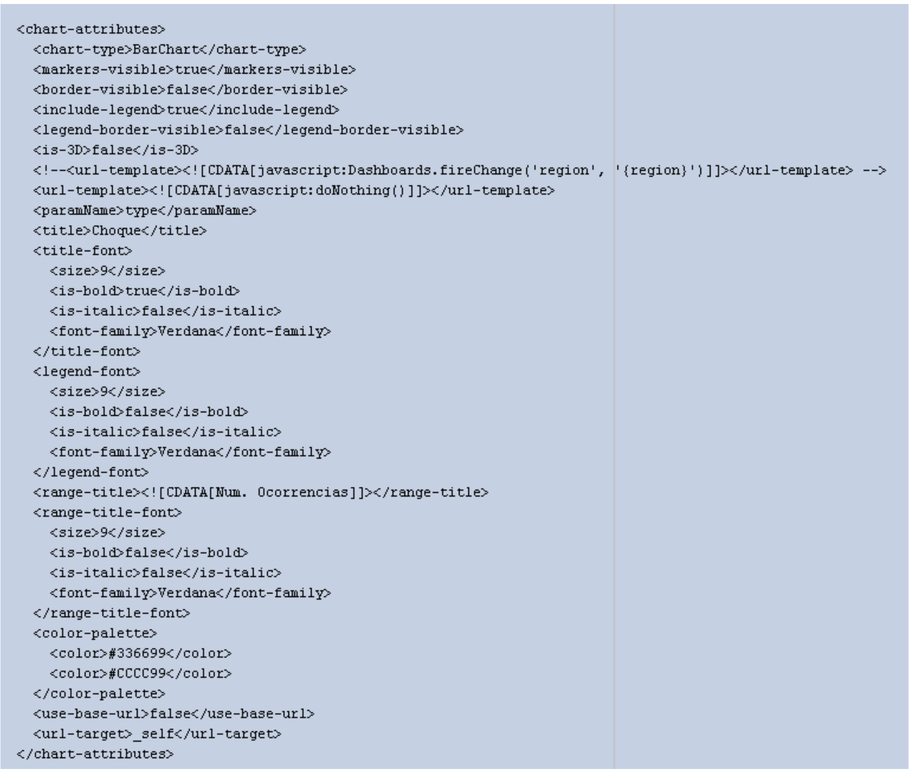
6.5. Reporting

During the implementation of this prototype, two reporting engines were used. One is the Pentaho Reporting, also known as JFreeReport, and the other is BIRT that is the acronym for Business Intelligence and Reporting Tools.

The merger of Pentaho Reporting and JFreeReport resulted in a new version of JFreeReport (0.9). This new engine is driven by the report layout instead of focusing on the data used. This approach is similar to Document Object Model (DOM) used to describe HTML pages or XML Files. Accordingly to Pentaho, this version is a complete turnover in terms of reporting engine, and is considered the cornerstone of the Pentaho Reporting, maybe that's the reason why this version is so unstable. In comparison with the other open source projects, Pentaho Reporting offers lesser reporting features

For the charts implemented for the Administration dashboard it was used Pentaho Reporting as main engine.

Below is an example of the implementation of one of the charts made with Pentaho Reporting. These types of charts are defined in a XML file. The data that used to populate the chart is defined with the action that calls the render of this chart.



```
<chart-attributes>
  <chart-type>BarChart</chart-type>
  <markers-visible>true</markers-visible>
  <border-visible>false</border-visible>
  <include-legend>true</include-legend>
  <legend-border-visible>false</legend-border-visible>
  <is-3D>false</is-3D>
  <!--<url-template><![CDATA[javascript:Dashboards.fireChange('region', '{region}')]]></url-template> -->
  <url-template><![CDATA[javascript:doNothing()]]></url-template>
  <paramName>type</paramName>
  <title>Choque</title>
  <title-font>
    <size>9</size>
    <is-bold>true</is-bold>
    <is-italic>false</is-italic>
    <font-family>Verdana</font-family>
  </title-font>
  <legend-font>
    <size>9</size>
    <is-bold>false</is-bold>
    <is-italic>false</is-italic>
    <font-family>Verdana</font-family>
  </legend-font>
  <range-title><![CDATA[Num. Ocorrencias]]></range-title>
  <range-title-font>
    <size>9</size>
    <is-bold>false</is-bold>
    <is-italic>false</is-italic>
    <font-family>Verdana</font-family>
  </range-title-font>
  <color-palette>
    <color>#336699</color>
    <color>#CCCC99</color>
  </color-palette>
  <use-base-url>false</use-base-url>
  <url-target>_self</url-target>
</chart-attributes>
```

Figure 38 – 'Choques' Chart definition using Pentaho Reporting

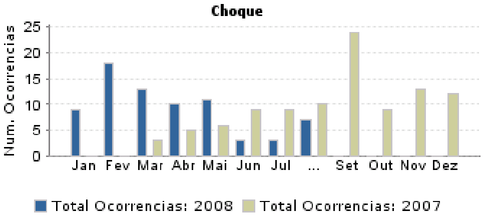


Figure 39 – 'Choques' Chart

However, during the development of the remaining dashboards it was necessary to use other types of graphics, but the JFreeReport engine wasn't compatible with the data input or it didn't generate the needed type. So I've decided to use the BIRT as an alternative.

BIRT is an open source reporting system that can be integrated with various data sources such as databases, xml files or javabeans. Like Pentaho Reporting, BIRT has two main components, a report designer based on Eclipse and a runtime reporting engine. Thanks to this last component, it is easy to deploy the engine in any application or web server.

The charts and pivot tables on the Exploration and Maintenance dashboards and the reports available via portal were designed with BIRT.

Below are depicted (Figure 40-41) the report layout made in BIRT Designer and the final output. The BIRT Report implementation method is much straightforward and simple than using Pentaho Reporting.

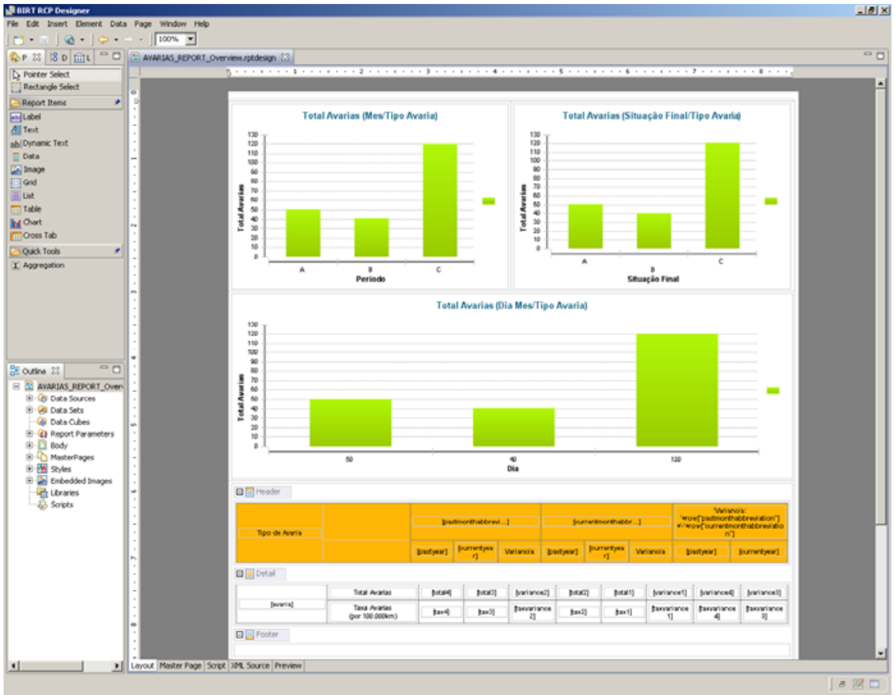
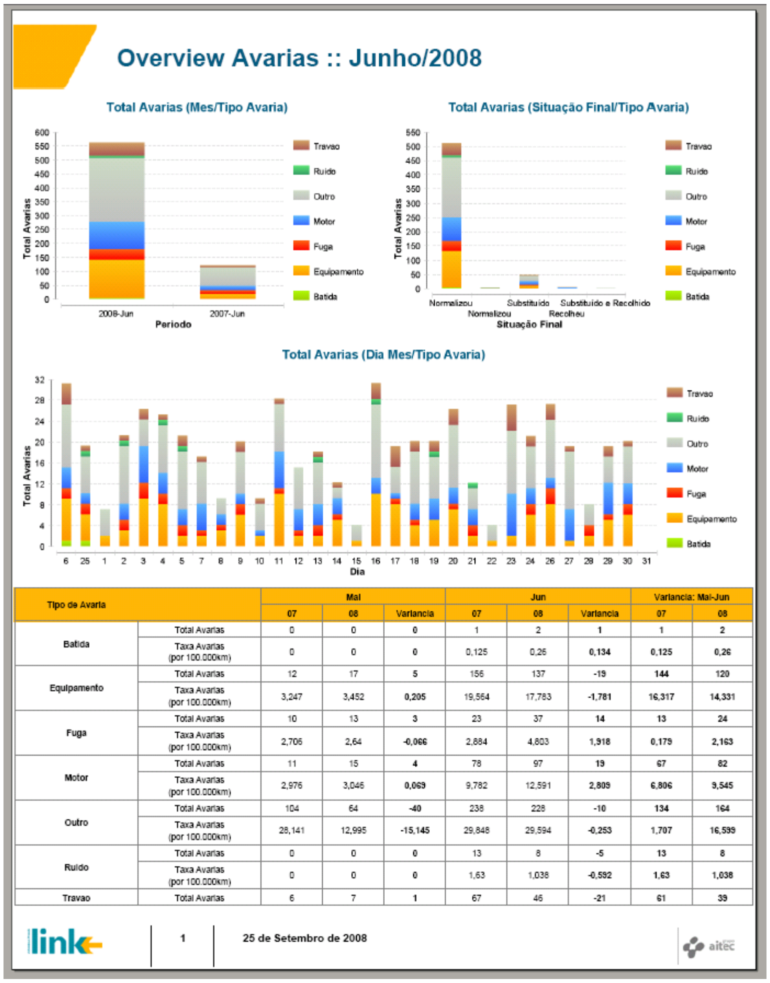


Figure 40 – 'Avarias' Report – Design via BIRT



Overview Avarias :: Junho/2008

Total Avarias (Mes/Tipo Avaria)

Total Avarias (Situação Final/Tipo Avaria)

Total Avarias (Dia Mes/Tipo Avaria)

Tipo de Avaria		Mai			Jun			Variancia: Mai-Jun	
		07	08	Variancia	07	08	Variancia	07	08
Batida	Total Avarias	0	0	0	1	2	1	1	2
	Taxa Avarias (por 100.000km)	0	0	0	0,125	0,26	0,134	0,125	0,26
Equipamento	Total Avarias	12	17	5	156	137	-19	144	120
	Taxa Avarias (por 100.000km)	3,247	3,452	0,205	19,564	17,783	-1,781	16,317	14,331
Fuga	Total Avarias	10	13	3	23	37	14	13	24
	Taxa Avarias (por 100.000km)	2,706	2,64	-0,066	2,884	4,803	1,918	0,179	2,163
Motor	Total Avarias	11	15	4	78	97	19	67	82
	Taxa Avarias (por 100.000km)	2,976	3,046	0,069	9,782	12,591	2,809	6,806	9,545
Outro	Total Avarias	104	64	-40	238	228	-10	134	164
	Taxa Avarias (por 100.000km)	28,141	12,995	-15,145	29,848	29,594	-0,253	1,707	16,599
Ruido	Total Avarias	0	0	0	13	8	-5	13	8
	Taxa Avarias (por 100.000km)	0	0	0	1,63	1,038	-0,592	1,63	1,038
Travao	Total Avarias	6	7	1	67	46	-21	61	39

link 1 25 de Setembro de 2008 aitec

Figure 41 – 'Avarias' Report

As mentioned in previous chapter, there is an ad-hoc query reporting application that allows the user to create simple reports via a web interface.

The reports created are in Appendix H.

Chapter 7 – Conclusion

This chapter contains the final conclusions of this Dissertation, with special emphasis on the goals achieved and on future work.

7.1. Accomplishments

The first goal of this thesis was to identify any relevant open source projects in the Business Intelligence area and select those that were the best in class. In order to accomplish this task, it was necessary to understand the concept of Business Intelligence and all its underlying components. This study was done by reading articles and books on this theme and the result can be found on the Chapter 2 – Literature Review. Once fully introduced to the BI concept, an initial survey was made. This first iteration resulted in an extensive list of more than fifty projects. Since there are a large variety of choices it was necessary to focus only on projects that cover more than one component of a BI architecture – BI Suites.

During the evaluation of BI suites, the Pentaho BI Suite was the one that offered the most comprehensive and mature solution, however during the last months the JasperSoft project has merged with other open source projects and can now provide a more flexible application, and even in some components such as data integration. This poses a serious threat to Pentaho. It is also important to add that one suite that during the evaluation phase wasn’t mature enough, is now offering a solution as good as the Pentaho one. This solution is SpagoBI, however this solution fails to deliver a critical “feature” of any open source project, the community. Comparing with Pentaho the SpagoBi community is almost inexistent. This constant change in the OSBI market let me to conclude that OSBI is becoming an attractive market and in the near future these tools may present a serious threat to the vendors of classical commercial business intelligence tools.

The main objective that I proposed to fulfill was to prove that it is possible to create a business intelligence solution comprised of only open source tools and that could be a viable alternative to the commercial tools. There were two possible approaches to accomplish this goal. The first one was to evaluate each tool in a generic environment and compare them with the commercial tools. The other approach was to implement a business intelligence solution that could bring an extra value to a company or industry. Both approaches would give a contribution to the scientific community. However the result of the second one would be more realistic since it would be tested in a real environment. The second approach was selected, to build a business intelligence solution for an industry where a solution like this would add value to those using it. More in detail, it was decided to build the prototype for the public transportation area, mainly because only in the last years this industry has made efforts to modernize their information and operational systems, for example only recently operators migrated from paper tickets to electronic smartcards.

The developed prototype consists in a secure web application, accessible from any device with an Internet connection and a Web browser, which allows the end-users to create and visualize reports and dashboards. It was built as a modular system so that any component can be changed to another one or serve as a basis for future work.

In addition to the typical advantages that outcome from any open source project, such as reduced total project cost due to the inexistent software licensing or high flexibility in component customization to an organization because of the availability of the source code, the main advantage that this prototype offers to any company is a fully modular architecture, in other words, the open source projects used to build it are independent of each other. Therefore it is possible to change the database server or the ETL tool without compromising the remaining prototype. For example, it is possible to use Microsoft Analysis Services as main Analysis Engine instead of Mondrian, because JPivot, the web interface that allows multidimensional analysis can be integrated with Microsoft Analysis Services or any other XMLA compliance data source.

Besides the achievement of the proposed goals for this thesis, it is important to mention the commercial value of this prototype. In an overview the prototype allows to transport operators to analyze their incident data and compare it with other organizations that operate in the same region. It provides a tool that helps them to improve their incident management by providing better information about what has happened in the past. Consequently the number of incidents can be reduced, resulting in significant benefits for both the passenger and company and increasing the awareness of a company,

7.2. Future Work

In terms of future developments there are at least four recommendations.

The first one is to explore other areas of BI that were not considered for this thesis, such as the Data Mining and Data Profiling. It would be considered as future work, not only the survey and analysis of open source projects in the two aforementioned areas as well as the integration of these projects with the BI solution designed to prove that it is possible to deliver a full open source BI solution.

Another one would be the analysis and integration of advanced visualization tools, such as maps. This type of visualization could have particular interest when considering the case study scope: Public transportation.

Another interesting future work is related with the commercial value of the developed prototype and his positioning on the SmartCITIES™ offer, in order words, on moving the produced prototype towards a product. Therefore it would be important to expand the scope to cover other functional areas of a transport operator such as Ticketing, Maintenance or Operations Management.

Considering also the commercial value of this prototype it would be interesting to create dashboards (integrated with maps) and reports to analyze the passenger information data, and in this way understand the needs of the transport operator’s client.

Appendices

Appendix A: Open Source Solutions

A.1. ETL Open Source Software Solutions

Below it is shown a table describing the most active Open Source ETL tools existing on the market.

Product	Description
Apatar https://sourceforge.net/projects/apatar/	It is an open source leader in data mashup integration. It helps non-developers to easily perform any transformation. It can be integrated not only with data sources such as Oracle, MySQL, Microsoft SQL 2005 but also with Salesforce.com, Amazon, SugarCRM, Flickr or even XML, CSV and RSS.
Celtix http://forge.objectweb.org/projects/celtix/	It is a java-based Enterprise Service Bus (ESB) that simplifies the construction, integration and reuse of business components.
Clover ETL http://www.cloveretl.org/	It is a Java-based framework for ETL applications used to transform structured data. There is also a version that allows users to create and modify data transformations in a graphical environment.
CpluSQL https://sourceforge.net/projects/cplusql	It is an ETL tool for extraction and transformation of data from databases and flat files. It is used for a terabyte scale data warehouse. It is written in C++.
Enhydra Octopus http://www.enhydra.org/tech/octopus/index.html	It is Java-based ETL tool. It connects to any JDBC data sources and performs transformations previously defined in an XML file.
JasperETL http://jasperforge.org/sf/projects/jasperetl	It is an ETL Tool powered by Talend. It can be used as a stand-alone application but it can also be integrated in the JasperSoft BI Suite.
KETL http://www.ketl.org/	KETL is an ETL tool that allows companies to manage the process on high volume transactions.
KETTLE http://kettle.pentaho.org/	K.E.T.T.L.E (Kettle ETTL Environment) is a meta-data driven ETTL tool (Extraction, Transformation, Transportation and Loading).
Mule http://mule.mulesource.org/display/MULE/Home	It is an enterprise service bus (ESB) for Service-oriented Architecture (SOA) scenarios. Mule 1.3 further supports XFire [32], a next-generation SOAP framework that makes service-oriented development approachable through an easy to use API and support for common standards. Besides XFire, developers can also interoperate between Apache Axis, WebMethods Glue and .Net Web Services.
Pequel ETL https://sourceforge.net/projects/pequel	According to their SourceForge description: "A comprehensive and high performance data processing/transform system. It features a simple, user-friendly event driven scripting interface that transparently generates & executes highly efficient Perl/C code"
Service Mix http://servicemix.apache.org/home.html	According to their website:"It is an Open Source ESB (Enterprise Service Bus) that combines the functionality of a Service Oriented Architecture (SOA) and an Event Driven Architecture (EDA)..."
Spagic http://www.spagic.org/ecm/faces/public/guest/home/solutions/spagic	It is a solution composed by a set of visual tools and back-end applications oriented towards design, realization, deploy and monitoring of ESB infrastructures adherent to the SOA

[32] XFire – http://xfire.codehaus.org/

	paradigm.
Talend Open Studio http://www.talend.com/	It is the first provider of open source data integration software and probably is the most enterprise oriented of the open source data integration vendors.

Table 4 – Appendix A: ETL Open Source Software Solutions

A.2. Data Warehouse and Database Open Source Software Solutions

Currently there are big offer of Open Source Relational Database Model System (RDMBS), however few are optimized to handle very large databases (VLDB). Below it is a description of the most important open source relational databases.

Product	Description
BerkeleyDB http://www.oracle.com/database/berkeley-db/index.html	According to their website: “Oracle Berkeley DB is a family of open source, embeddable databases that allows developers to incorporate within their applications a fast, scalable, transactional database engine with industrial grade reliability and availability”.
Bizgres http://www.bizgres.org/home.php	This open source project aims to make PostgreSQL the database for Business Intelligence.
Derby http://db.apache.org/derby/	It is an open source relational database implemented in Java and available under Apache.
EnterpriseDB http://www.enterprisedb.com/	It is a powerful combination of PostgreSQL and Oracle PL/SQL.
Firebird http://www.firebirdsql.org/	It is a relational database that runs on Linux, Windows, and a variety of Unix platforms.
GreenPlum http://www.greenplum.com/	It is an open source database that can scale to support multi-terabyte data warehousing demands.
Ingres 2006 http://www.ingres.com/products/ingres-2006.php	It is an open source database that provides enterprise-level functionality, services and support. There is also a BI Appliance.
LucidDB for DW http://www.luciddb.org/	It is an open source RDBMS purpose-built entirely for data warehousing and business intelligence.
MySQL http://www.mysql.com/	“The world's most popular open source database”. It has proven suitable for VLDB architectures.
MonetDB http://monetdb.cwi.nl/	It is an open source database system for high-performance applications in data mining, OLAP, GIS, XML Query, text and multimedia retrieval.
Perst http://www.mcobject.com/perst/	It is an Embedded Database.
PostgreSQL http://www.postgresql.org/	As defined on their website: “a highly scalable, SQL compliant, open source object-relational database management system.”

Table 5 – Appendix A: Open Source Databases

A.3. OLAP Open Source Software Solutions

There are many On-Line Analytical processing tools to choose from, Multidimensional Databases OLAP (MDDB OLAP or MOLAP), Relational OLAP (ROLAP) and Hybrid OLPA (HOLAP). And for each one of these tools there are open source software projects.

This list below includes not only engines or servers for OLAP or MDDB use, but also front-end tools for performing typical OLAP functions such as slice-and-dice, drill-down and roll-up.

Product	Description
Cubulus http://sourceforge.net/projects/cubulus/	Recently new engine that uses *hierarchical range-clustering of keys* (1) The main difference to other OLAP Engines is the fact that Cubulus does all the aggregations inside the relational database.
jPivot http://jpivot.sourceforge.net/	It is a JSP library that provides a front-end OLAP table to the Mondrian OLAP engine, allowing typical OLAP functions such as slice-and-dice, drill-down and roll-up.
Mondrian http://mondrian.pentaho.org/	It is a java-based OLAP Server used in open source software BI Suite projects such as JasperAnalysis and PentahoAnalysis.
OpenOlap for MySQL http://forge.mysql.com/projects/view.php?id=13	An OLAP tool for MySQL. It has a Model Designer for building cubes, a Report Designer for creating reports and "Viewer" to analyze data. Currently it is only available a Japanese Version. There is also a version for PostgreSQL.
Palo http://www.jedox.com/en/enterprise-spreadsheet-server/excel-olap-server/palo-server.html	IT provides a MDDB for Excel, with future plans to allow access through other APIs as well.

Table 6 – Appendix A: OLAP Open Source Tools

A.4. Data Mining Open Source Solutions

This type of open source projects is taking a very important role in Open Source BI Suites. The most known data mining open source projects are explained below.

Product	Description
CPAS https://www.labkey.org/Project/home/begin.view	Helps scientists to manage, analyze and share complex datasets. It supports some complex calculations with web-based collaboration systems.
Rattle http://rattle.togaware.com/	Provides a simple and logical interface for quick and easy data mining.
Weka http://www.cs.waikato.ac.nz/~ml/weka/index.html	It is a collection of machine learning algorithms for solving real-world data mining problems. It is integrated in Pentaho Suite. Contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Table 7 – Appendix A: Data Mini.ng Open Source Solutions

A.5. Data Visualization Open Source Solutions

Usually these tools are associated to a BI Suite, but it is easy to find some vendors that are focused on a single component. Below it is represented some of the open source products.

Product	Vendor	Tool Type	Description
BreadBoardBI http://www.cs.waikato.ac.nz/~ml/weka/index.html	BreadBoardBI	Analytical Application	Application that extracts data from logs, loads databases tables and present the information in dashboards, cubes or reports for business users.
DASH http://www.marvelit.com	MarvelIT	Dashboards	According to SourceForge.com description: “It is an open source Business Intelligence solution based on Apache Jetspeed. DASH attempts to simplify the process of creating and maintaining a web-based business intelligence dashboard and centralized reporting for companies of any size.”

Table 8 – Appendix A: Data Visualization Open Source Solutions

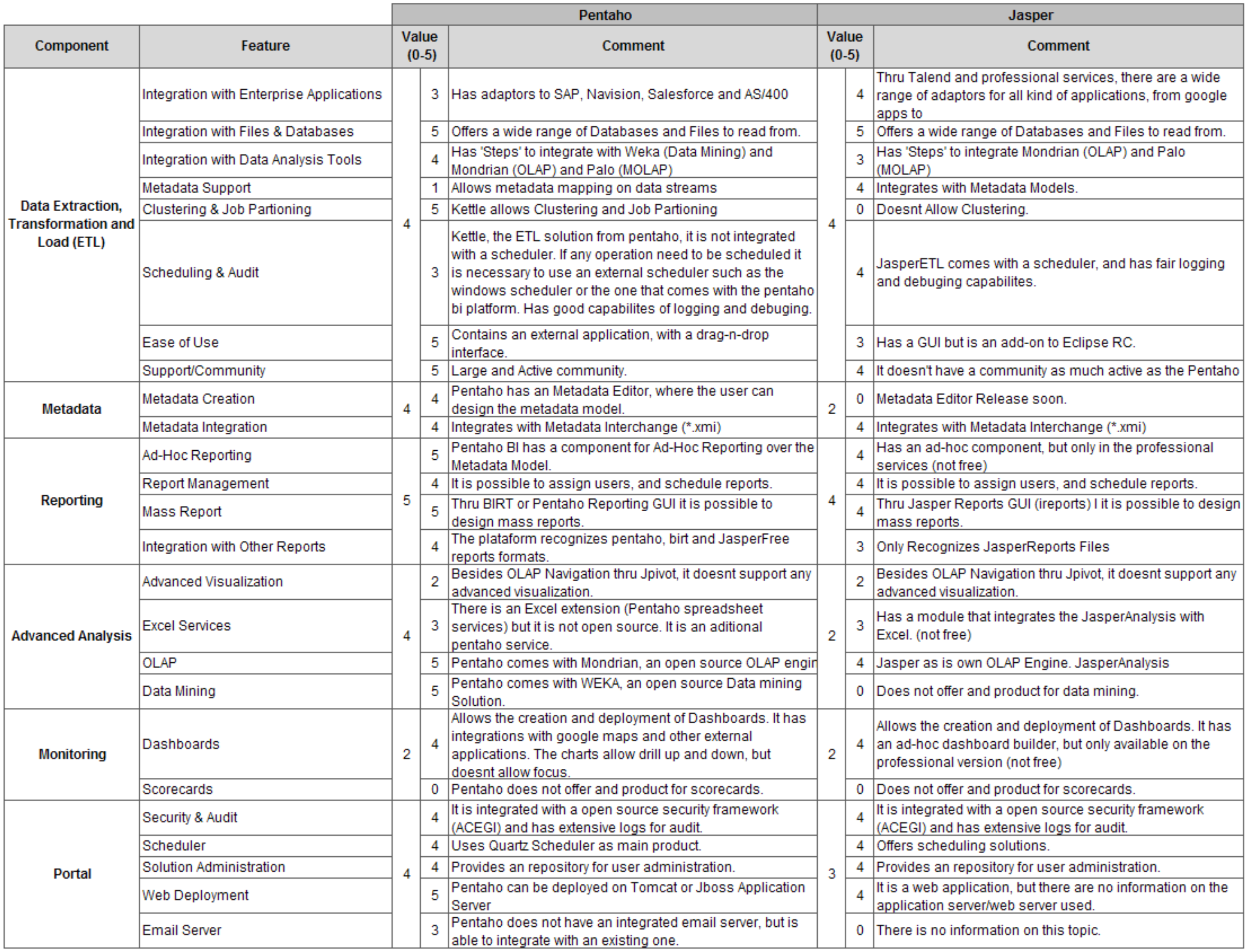
A.6. Reporting Open Source Software Solutions

There is a large variety of reporting tools; some are simple, while others are extremely complex. Most of them are web-based, and others have report design in.cluded. Below is the list of them.

Product	Description
Agata Report http://www.agata.org.br/	It is a Database Reporting Tool and EIS tool similar to Crystal Reports. It generates graphs. It can be connected to several databases (PostgreSQL, MySQL, Oracle, SyBase, MsSql, FrontBase, DB2, Informix and InterBase).
DataVision http://datavision.sourceforge.net/	It allows drag-and-drop report design through its GUI. It is written in Java and can connect to any database supporting JDBC.
iReport http://jasperforge.org/sf/projects/ireport	It is a report builder/designer for JasperReports.
JasperReports http://jasperforge.org/sf/projects/jasperreports	Is the leading open source reporting engine.
JFreeReport http://www.jfree.org/jfreereport/index.php	Started as a standalone Java report library, in January, 2006, became a part of the Pentaho suite.
OpenReports http://www.oreports.com/	Is described in their website as "a flexible open source web reporting solution that allows users to generate dynamic reports in a browser”
OpenRPT http://www.xtuple.com/openrpt/	It is a full featured, cross-platform SQL report writer that stores its report definitions as XML

Table 9 – Appendix A: Reporting Open Source Software Solutions

Appendix B: Open Source Solutions



		Pentaho			Jasper		
Component	**Feature**	**Value (0-5)**		**Comment**	**Value (0-5)**		**Comment**
Data Extraction, Transformation and Load (ETL)	Integration with Enterprise Applications	4	3	Has adaptors to SAP, Navision, Salesforce and AS/400	4	4	Thru Talend and professional services, there are a wide range of adaptors for all kind of applications, from google apps to
	Integration with Files & Databases		5	Offers a wide range of Databases and Files to read from.		5	Offers a wide range of Databases and Files to read from.
	Integration with Data Analysis Tools		4	Has 'Steps' to integrate with Weka (Data Mining) and Mondrian (OLAP) and Palo (MOLAP)		3	Has 'Steps' to integrate Mondrian (OLAP) and Palo (MOLAP)
	Metadata Support		1	Allows metadata mapping on data streams		4	Integrates with Metadata Models.
	Clustering & Job Partioning		5	Kettle allows Clustering and Job Partioning		0	Doesnt Allow Clustering.
	Scheduling & Audit		3	Kettle, the ETL solution from pentaho, it is not integrated with a scheduler. If any operation need to be scheduled it is necessary to use an external scheduler such as the windows scheduler or the one that comes with the pentaho bi platform. Has good capabilites of logging and debuging.		4	JasperETL comes with a scheduler, and has fair logging and debuging capabilites.
	Ease of Use		5	Contains an external application, with a drag-n-drop interface.		3	Has a GUI but is an add-on to Eclipse RC.
	Support/Community		5	Large and Active community.		4	It doesn't have a community as much active as the Pentaho
Metadata	Metadata Creation	4	4	Pentaho has an Metadata Editor, where the user can design the metadata model.	2	0	Metadata Editor Release soon.
	Metadata Integration		4	Integrates with Metadata Interchange (*.xmi)		4	Integrates with Metadata Interchange (*.xmi)
Reporting	Ad-Hoc Reporting	5	5	Pentaho BI has a component for Ad-Hoc Reporting over the Metadata Model.	4	4	Has an ad-hoc component, but only in the professional services (not free)
	Report Management		4	It is possible to assign users, and schedule reports.		4	It is possible to assign users, and schedule reports.
	Mass Report		5	Thru BIRT or Pentaho Reporting GUI it is possible to design mass reports.		4	Thru Jasper Reports GUI (ireports) I it is possible to design mass reports.
	Integration with Other Reports		4	The plataform recognizes pentaho, birt and JasperFree reports formats.		3	Only Recognizes JasperReports Files
Advanced Analysis	Advanced Visualization	4	2	Besides OLAP Navigation thru Jpivot, it doesnt support any advanced visualization.	2	2	Besides OLAP Navigation thru Jpivot, it doesnt support any advanced visualization.
	Excel Services		3	There is an Excel extension (Pentaho spreadsheet services) but it is not open source. It is an aditional pentaho service.		3	Has a module that integrates the JasperAnalysis with Excel. (not free)
	OLAP		5	Pentaho comes with Mondrian, an open source OLAP engir		4	Jasper as is own OLAP Engine. JasperAnalysis
	Data Mining		5	Pentaho comes with WEKA, an open source Data mining Solution.		0	Does not offer and product for data mining.
Monitoring	Dashboards	2	4	Allows the creation and deployment of Dashboards. It has integrations with google maps and other external applications. The charts allow drill up and down, but doesnt allow focus.	2	4	Allows the creation and deployment of Dashboards. It has an ad-hoc dashboard builder, but only available on the professional version (not free)
	Scorecards		0	Pentaho does not offer and product for scorecards.		0	Does not offer and product for scorecards.
Portal	Security & Audit	4	4	It is integrated with a open source security framework (ACEGI) and has extensive logs for audit.	3	4	It is integrated with a open source security framework (ACEGI) and has extensive logs for audit.
	Scheduler		4	Uses Quartz Scheduler as main product.		4	Offers scheduling solutions.
	Solution Administration		4	Provides an repository for user administration.		4	Provides an repository for user administration.
	Web Deployment		5	Pentaho can be deployed on Tomcat or Jboss Application Server		4	It is a web application, but there are no information on the application server/web server used.
	Email Server		3	Pentaho does not have an integrated email server, but is able to integrate with an existing one.		0	There is no information on this topic.

Table 10 – Appendix B: Open Source BI Suites Evaluation

Appendix C: Pilot Methodology Diagram

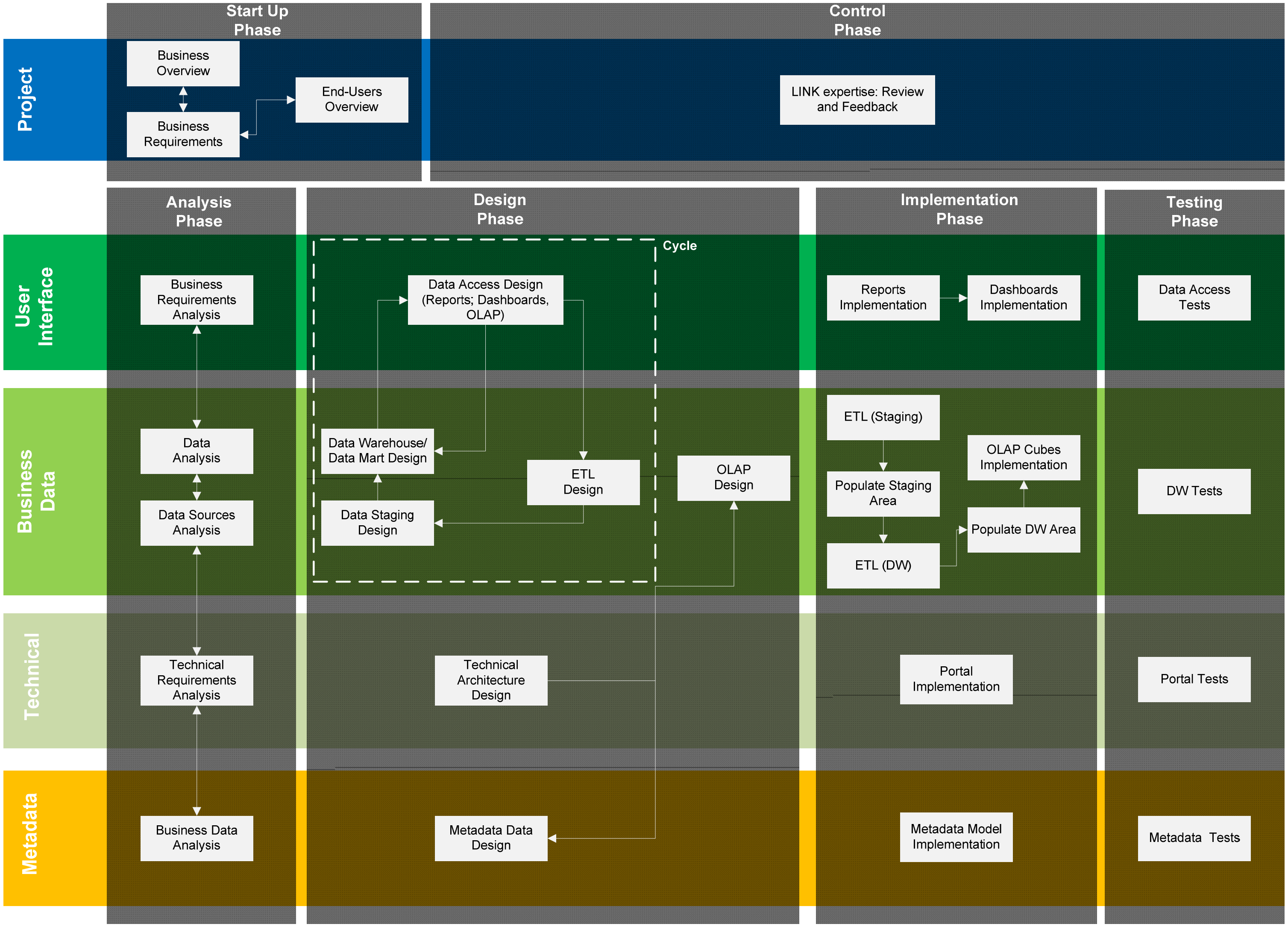


Figure 42 – Pilot Methodology Diagram

Appendix D: Staging Data Model

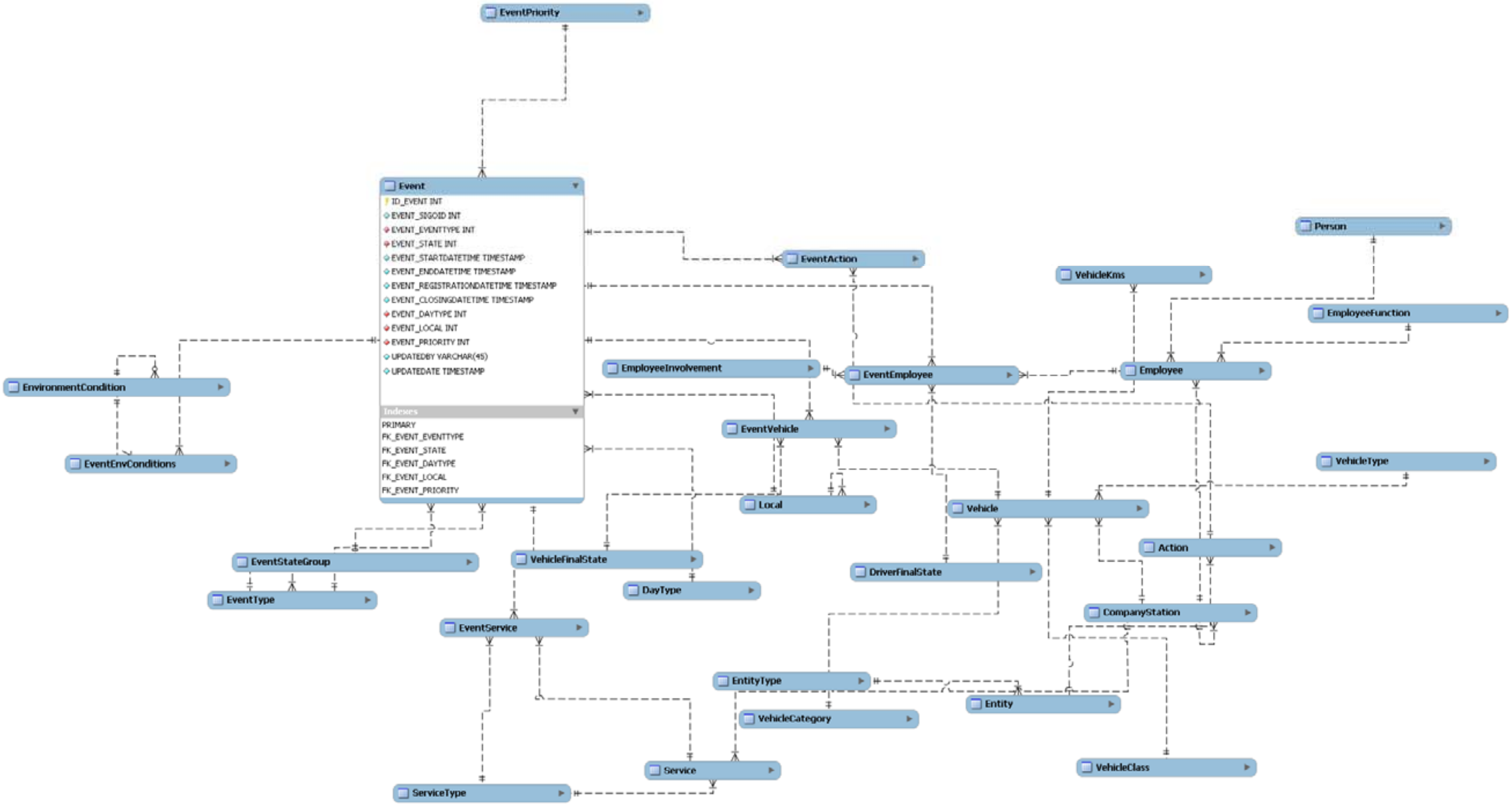


Figure 43 – Staging Data Model

Appendix E: Multidimensional Data Marts

E.1. Incident Data Mart

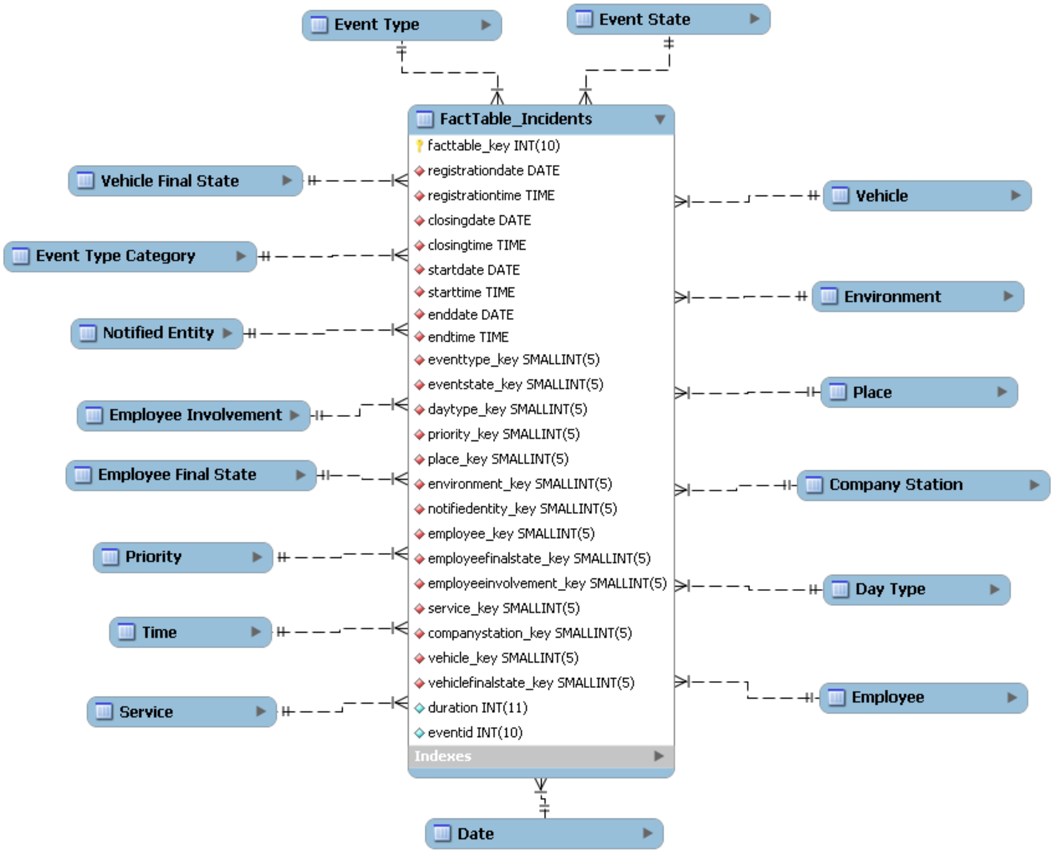


Figure 44 – Incident Data Mart Diagram

E.2. Service Data Mart

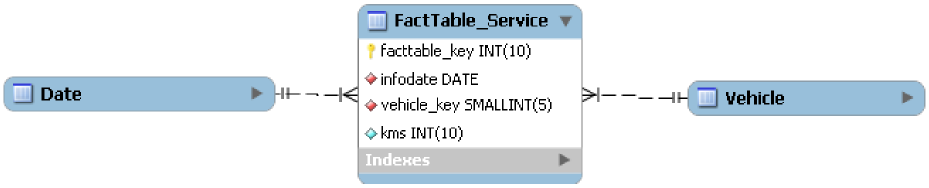


Figure 45 – Service Data Mart Diagram

E.3. Passenger Information Data Mart

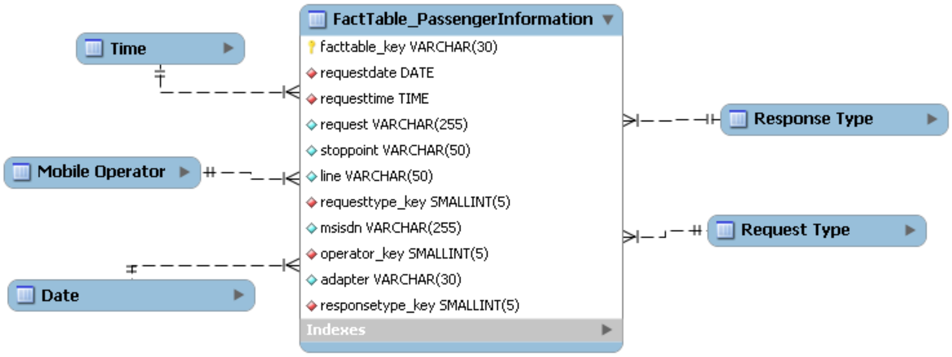
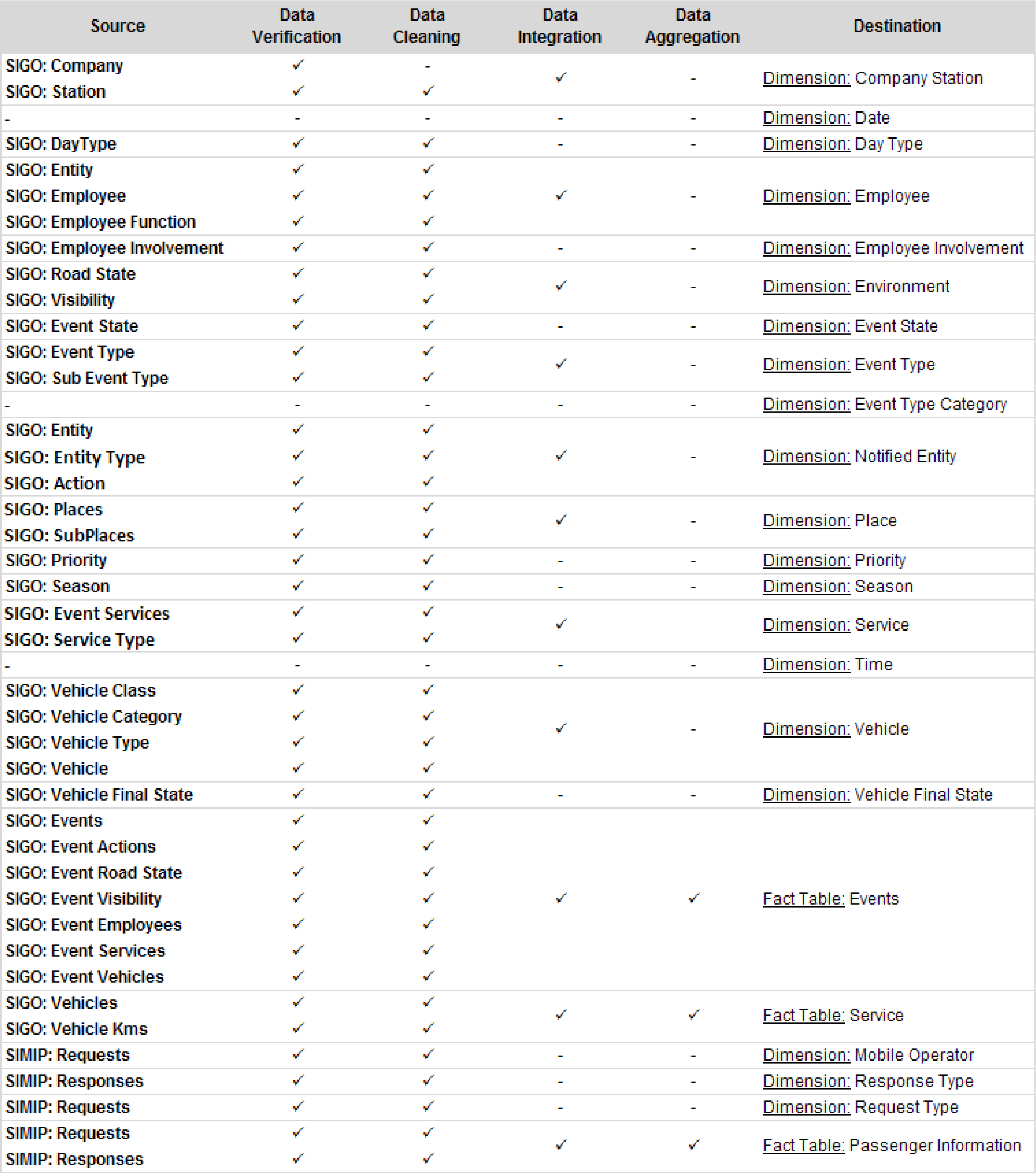


Figure 46 – Passenger Information Data Mart Diagram

Appendix F: ETL Process Design

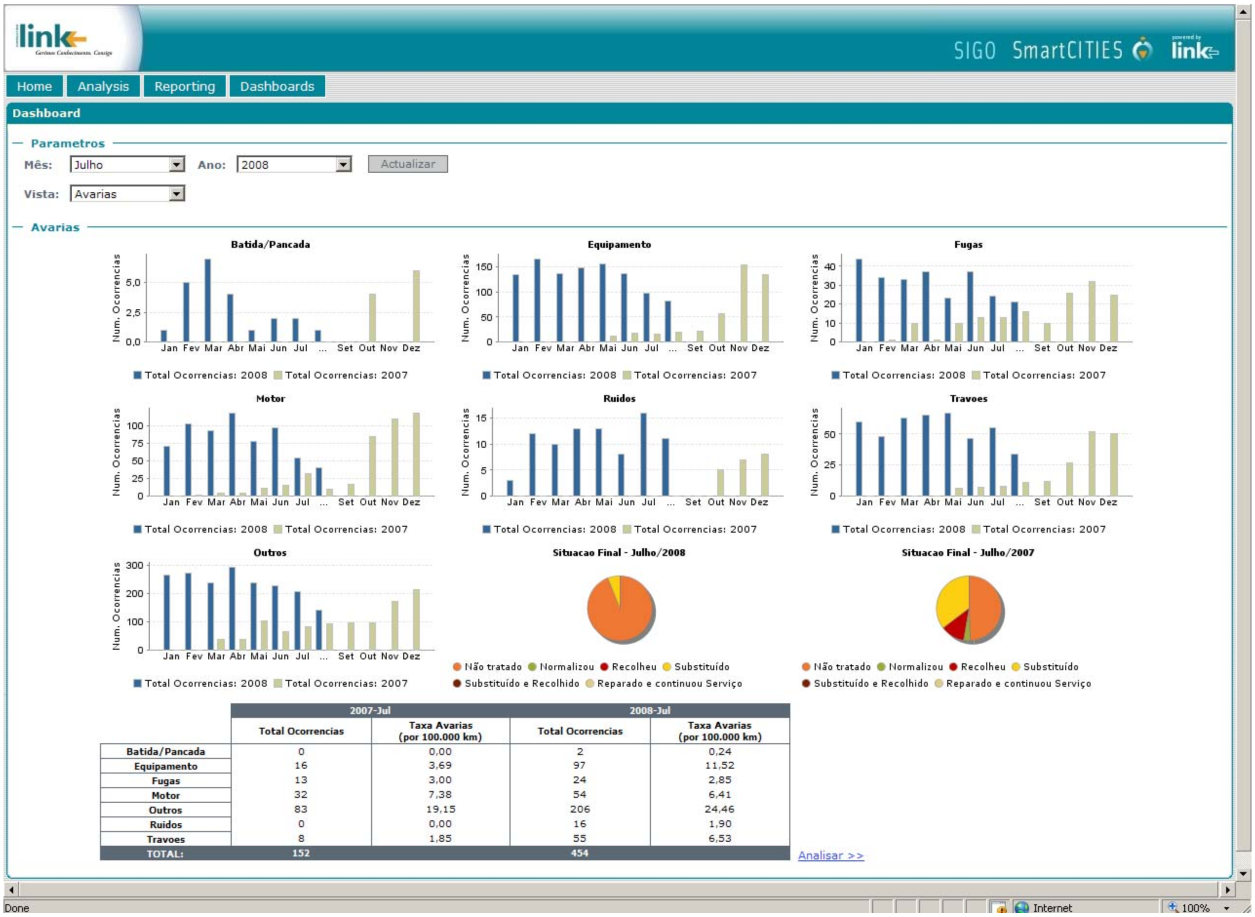


Source	Data Verification	Data Cleaning	Data Integration	Data Aggregation	Destination
SIGO: Company	✓	-	✓	-	Dimension: Company Station
SIGO: Station	✓	✓			
-	-	-	-	-	Dimension: Date
SIGO: DayType	✓	✓	-	-	Dimension: Day Type
SIGO: Entity	✓	✓	✓	-	Dimension: Employee
SIGO: Employee	✓	✓			
SIGO: Employee Function	✓	✓			
SIGO: Employee Involvement	✓	✓	-	-	Dimension: Employee Involvement
SIGO: Road State	✓	✓	✓	-	Dimension: Environment
SIGO: Visibility	✓	✓			
SIGO: Event State	✓	✓	-	-	Dimension: Event State
SIGO: Event Type	✓	✓	✓	-	Dimension: Event Type
SIGO: Sub Event Type	✓	✓			
-	-	-	-	-	Dimension: Event Type Category
SIGO: Entity	✓	✓	✓	-	Dimension: Notified Entity
SIGO: Entity Type	✓	✓			
SIGO: Action	✓	✓			
SIGO: Places	✓	✓	✓	-	Dimension: Place
SIGO: SubPlaces	✓	✓			
SIGO: Priority	✓	✓	-	-	Dimension: Priority
SIGO: Season	✓	✓	-	-	Dimension: Season
SIGO: Event Services	✓	✓	✓		Dimension: Service
SIGO: Service Type	✓	✓			
-	-	-	-	-	Dimension: Time
SIGO: Vehicle Class	✓	✓	✓	-	Dimension: Vehicle
SIGO: Vehicle Category	✓	✓			
SIGO: Vehicle Type	✓	✓			
SIGO: Vehicle	✓	✓			
SIGO: Vehicle Final State	✓	✓	-	-	Dimension: Vehicle Final State
SIGO: Events	✓	✓	✓	✓	Fact Table: Events
SIGO: Event Actions	✓	✓			
SIGO: Event Road State	✓	✓			
SIGO: Event Visibility	✓	✓			
SIGO: Event Employees	✓	✓			
SIGO: Event Services	✓	✓			
SIGO: Event Vehicles	✓	✓			
SIGO: Vehicles	✓	✓	✓	✓	Fact Table: Service
SIGO: Vehicle Kms	✓	✓			
SIMIP: Requests	✓	✓	-	-	Dimension: Mobile Operator
SIMIP: Responses	✓	✓	-	-	Dimension: Response Type
SIMIP: Requests	✓	✓	-	-	Dimension: Request Type
SIMIP: Requests	✓	✓	✓	✓	Fact Table: Passenger Information
SIMIP: Responses	✓	✓			

Table 11 – Appendix I: ETL Process Design Analysis

Appendix G: Dashboards

G.1. Administration Dashboard: Avarias



Mês: Julho

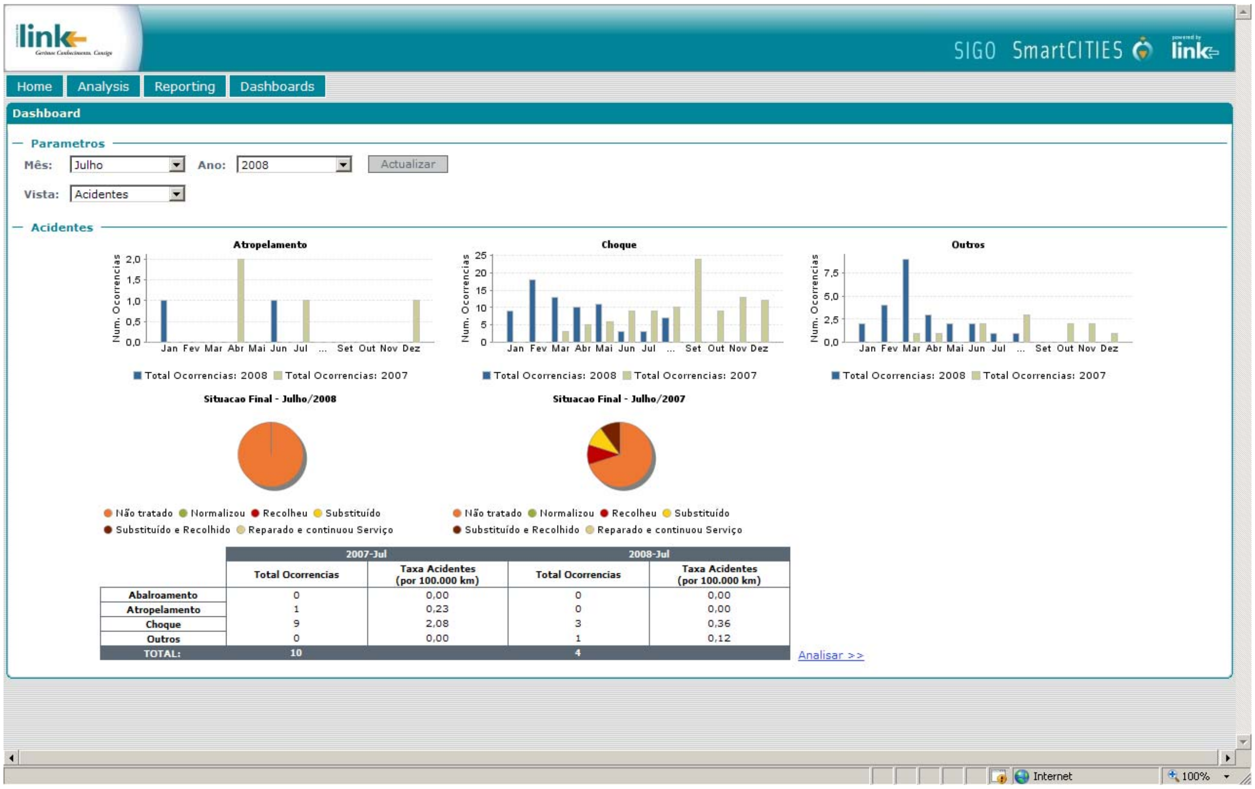
Ano: 2008

Vista: Avarias

	2007-Jul		2008-Jul	
	Total Ocorrencias	Taxa Avarias (por 100.000 km)	Total Ocorrencias	Taxa Avarias (por 100.000 km)
Batida/Pancada	0	0,00	2	0,24
Equipamento	16	3,69	97	11,52
Fugas	13	3,00	24	2,85
Motor	32	7,38	54	6,41
Outros	83	19,15	206	24,46
Ruidos	0	0,00	16	1,90
Travoes	8	1,85	55	6,53
TOTAL:	152		454	

Figure 47 – Administration Dashboard 'Avarias'

G.2. Administration Dashboard: Acidentes


link Gerimos Conhecimento. Consigo
SIGO SmartCITIES
powered by link
Home
Analysis
Reporting
Dashboards
Dashboard
Actualizar
Atropelamento
Choque
Outros
Num. Ocorrencias
Jan Fev Mar Abr Mai Jun Jul ... Set Out Nov Dez
■ Total Ocorrencias: 2008 ■ Total Ocorrencias: 2007
Situacao Final - Julho/2008
Situacao Final - Julho/2007
● Não tratado ● Normalizou ● Recolheu ● Substituído
● Substituído e Recolhido ● Reparado e continuou Serviço
Analisar >>
Internet
100%

Parametros

Mês: Julho Ano: 2008

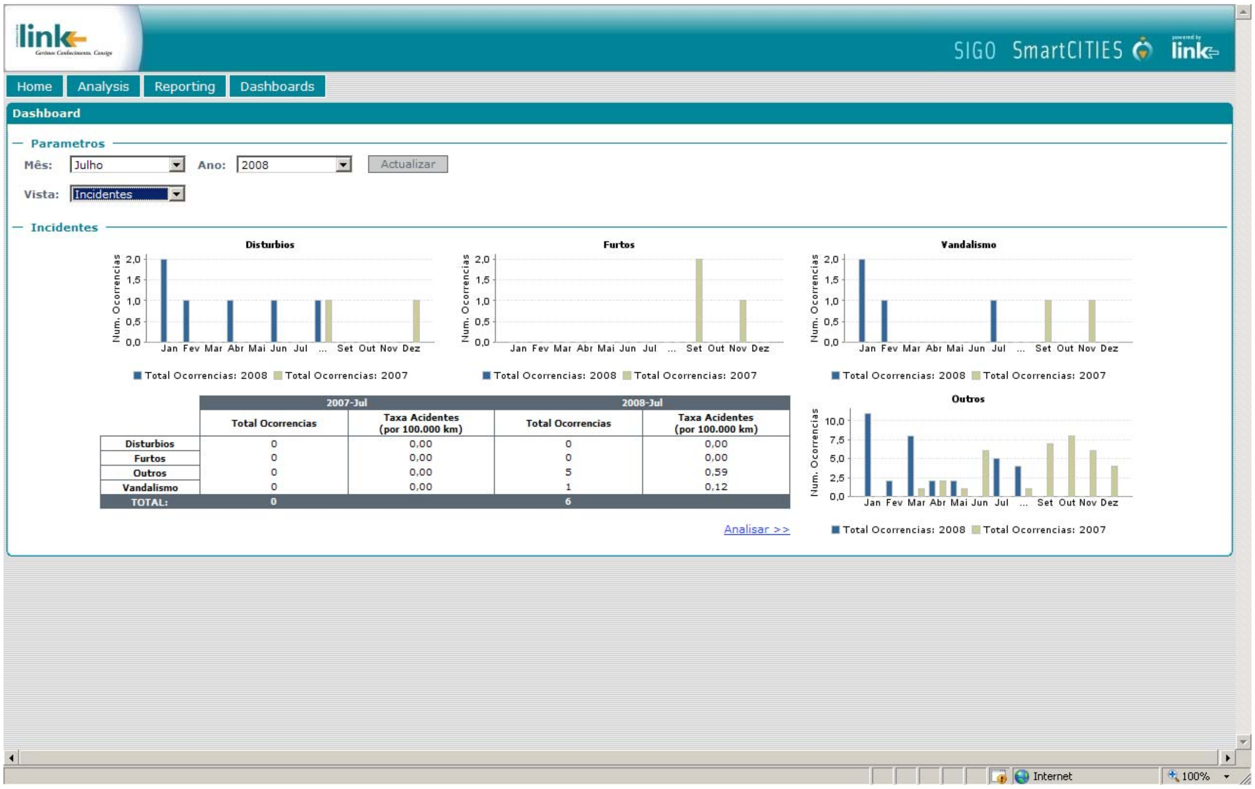
Vista: Acidentes

Acidentes

	2007-Jul		2008-Jul	
	Total Ocorrencias	Taxa Acidentes (por 100.000 km)	Total Ocorrencias	Taxa Acidentes (por 100.000 km)
Abalroamento	0	0,00	0	0,00
Atropelamento	1	0,23	0	0,00
Choque	9	2,08	3	0,36
Outros	0	0,00	1	0,12
TOTAL:	10		4	

Figure 48 – Administration Dashboard 'Acidentes'

G.3. Administration Dashboard: Incidentes



Mês: Julho

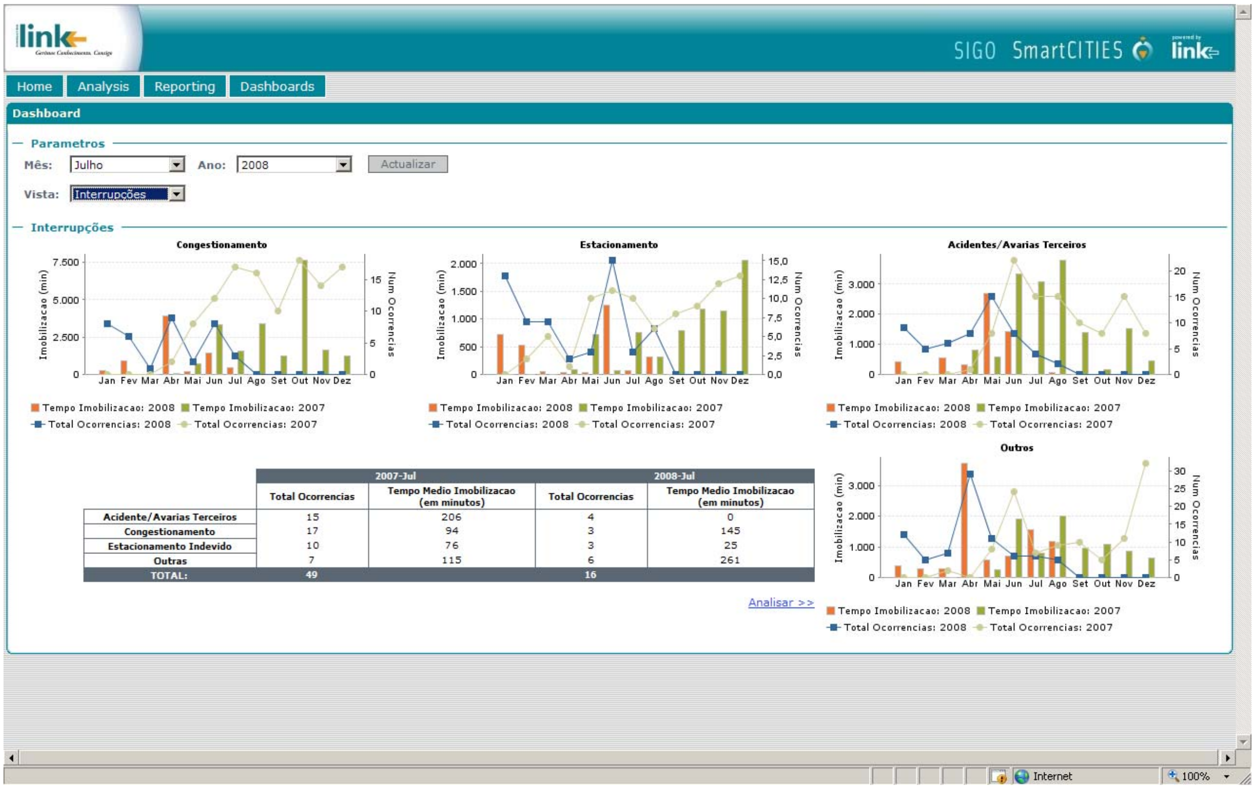
Ano: 2008

Vista: Incidentes

	2007-Jul		2008-Jul	
	Total Ocorrencias	Taxa Acidentes (por 100.000 km)	Total Ocorrencias	Taxa Acidentes (por 100.000 km)
Disturbios	0	0,00	0	0,00
Furtos	0	0,00	0	0,00
Outros	0	0,00	5	0,59
Vandalismo	0	0,00	1	0,12
TOTAL:	0		6	

Figure 49 – Administration Dashboard 'Incidentes'

G.4. Administration Dashboard: Interrupções



Parametros

Mês: Julho Ano: 2008

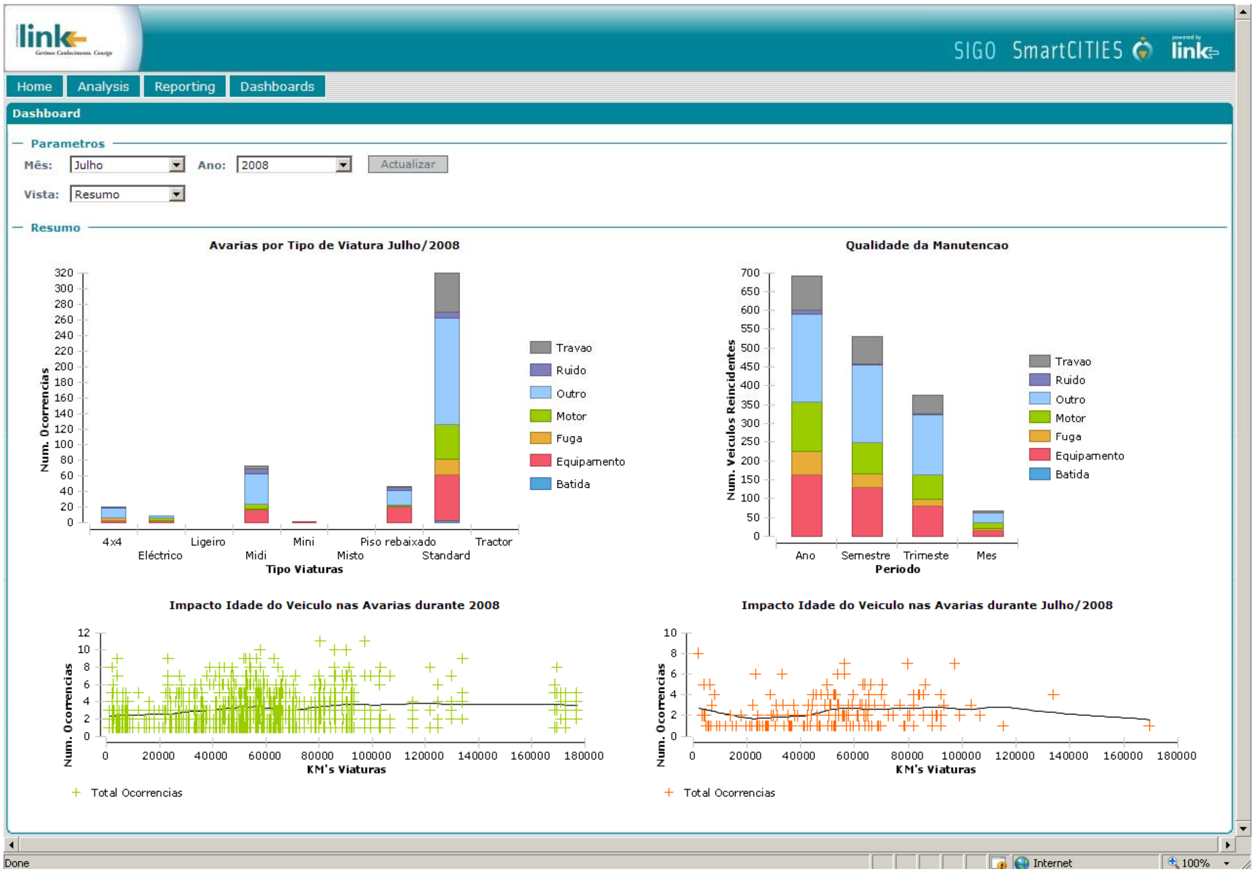
Vista: Interrupções

Interrupções

	2007-Jul		2008-Jul	
	Total Ocorrencias	Tempo Medio Imobilizacao (em minutos)	Total Ocorrencias	Tempo Medio Imobilizacao (em minutos)
Acidente/Avarias Terceiros	15	206	4	0
Congestionamento	17	94	3	145
Estacionamento Indevido	10	76	3	25
Outras	7	115	6	261
TOTAL:	49		16	

Figure 50 – Administration Dashboard 'Interrupções'

G.5. Maintenance Dashboard: Resumo

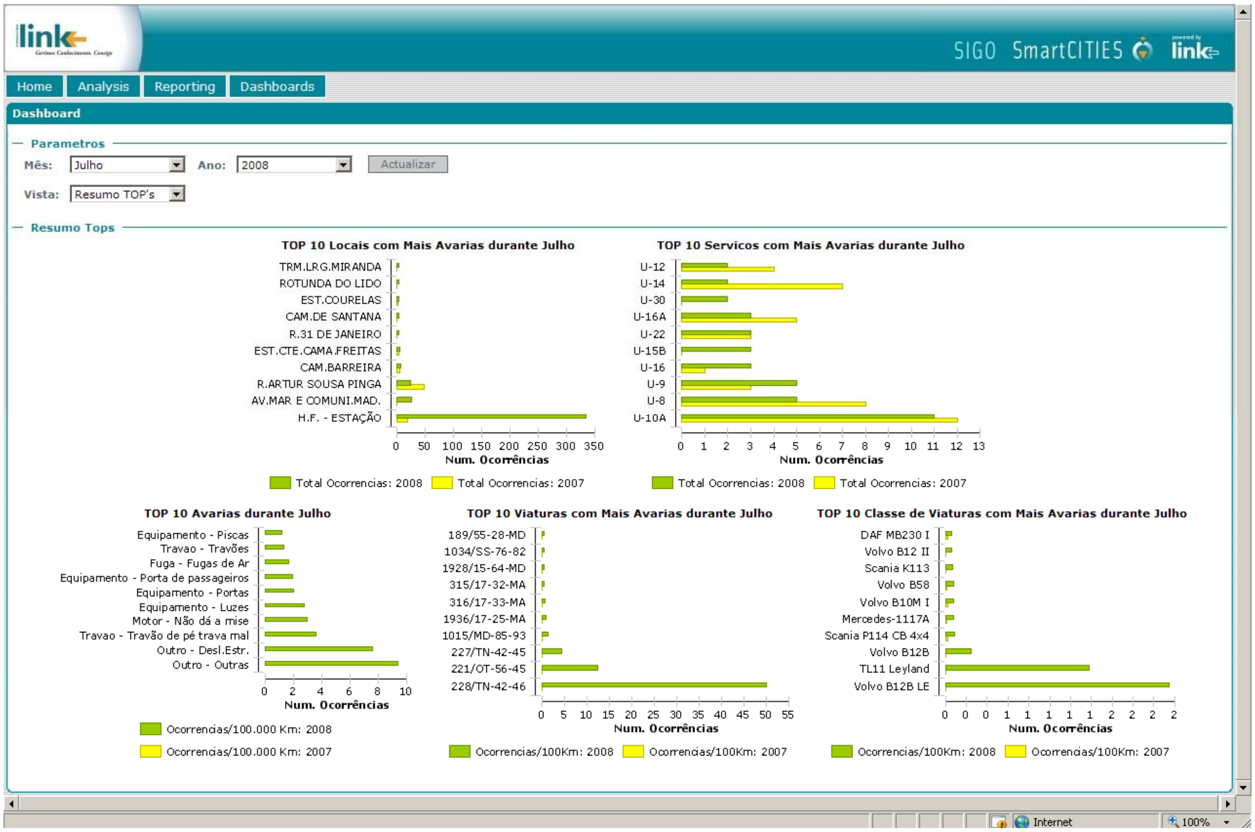


Mês: Julho Ano: 2008

Vista: Resumo

Figure 51 – Maintenance Dashboard 'Resumo'

G.6. Maintenance Dashboard: Resumo Top's



Dashboard

Parametros

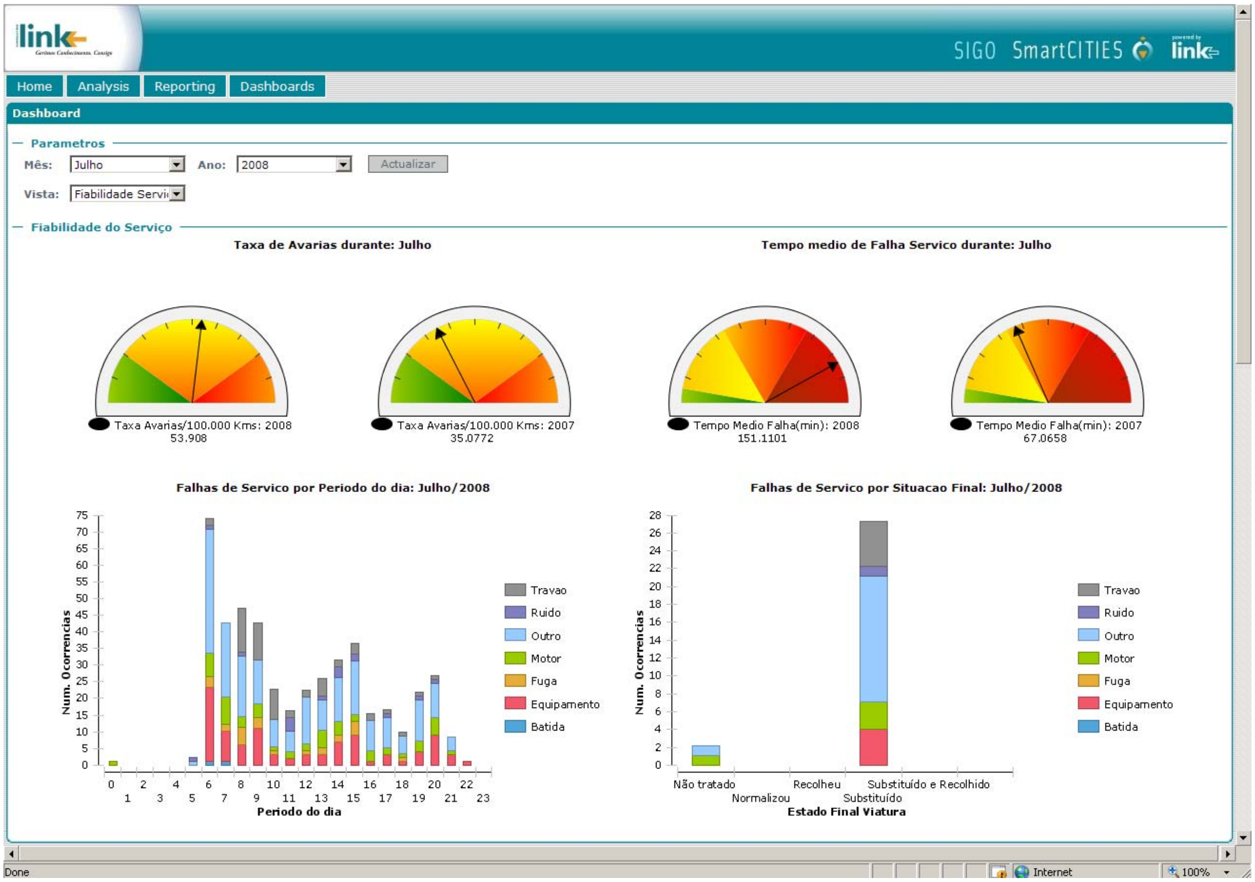
Mês: Julho Ano: 2008 Actualizar

Vista: Resumo TOP's

Resumo Tops

Figure 52 – Maintenance Dashboard 'Resumo Top's'

G.7. Operations Dashboard: Fiabilidade Serviço I



Dashboard

Parametros

Mês: Julho

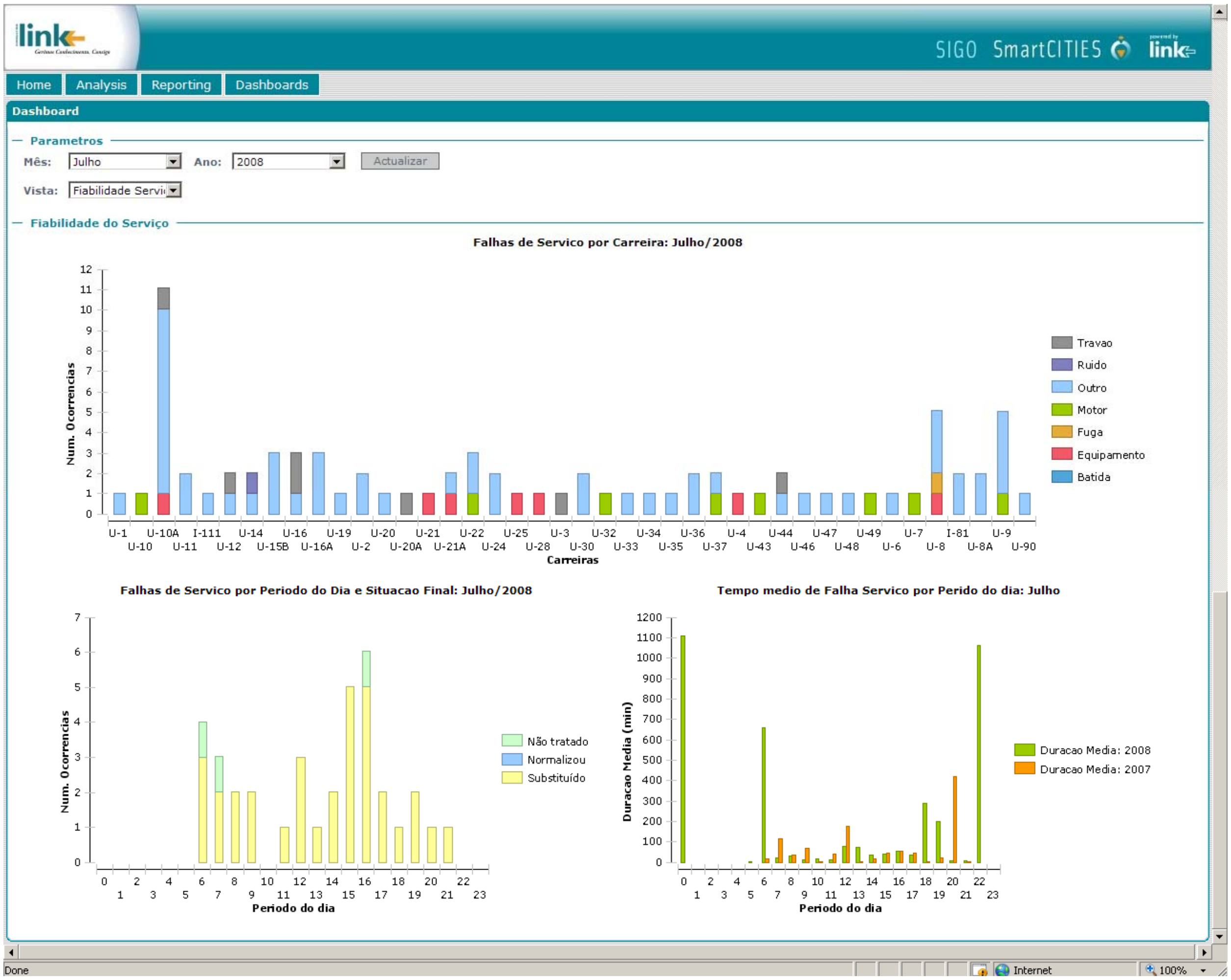
Ano: 2008

Vista: Fiabilidade Servi

Fiabilidade do Serviço

Figure 53 – Operations Dashboard ‘Fiabilidade Serviço I’

G.8. Operations Dashboard: Fiabilidade Serviço II



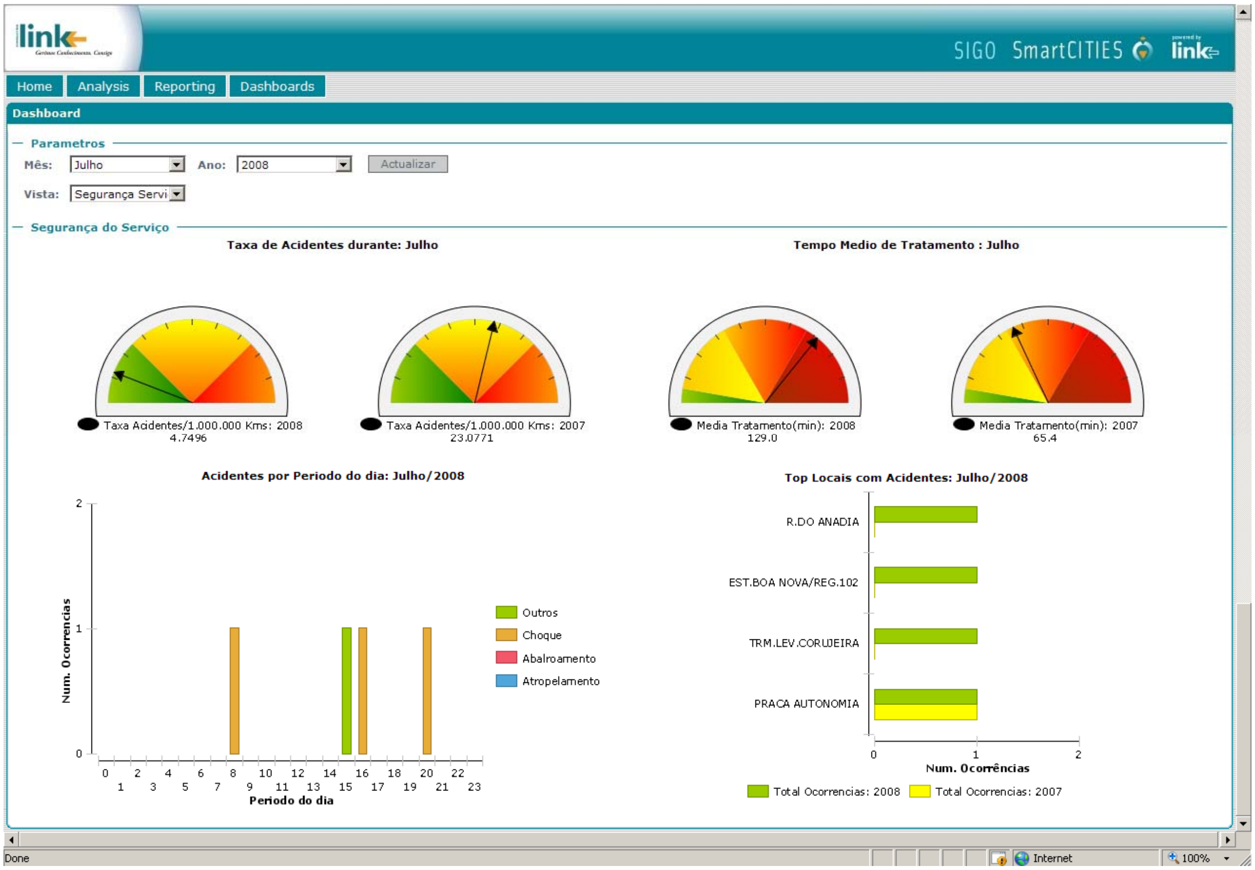
Mês: Julho

Ano: 2008

Vista: Fiabilidade Servi

Figure 54 – Operations Dashboard 'Fiabilidade Serviço II'

G.9. Operations Dashboard: Segurança Serviço I



Parametros

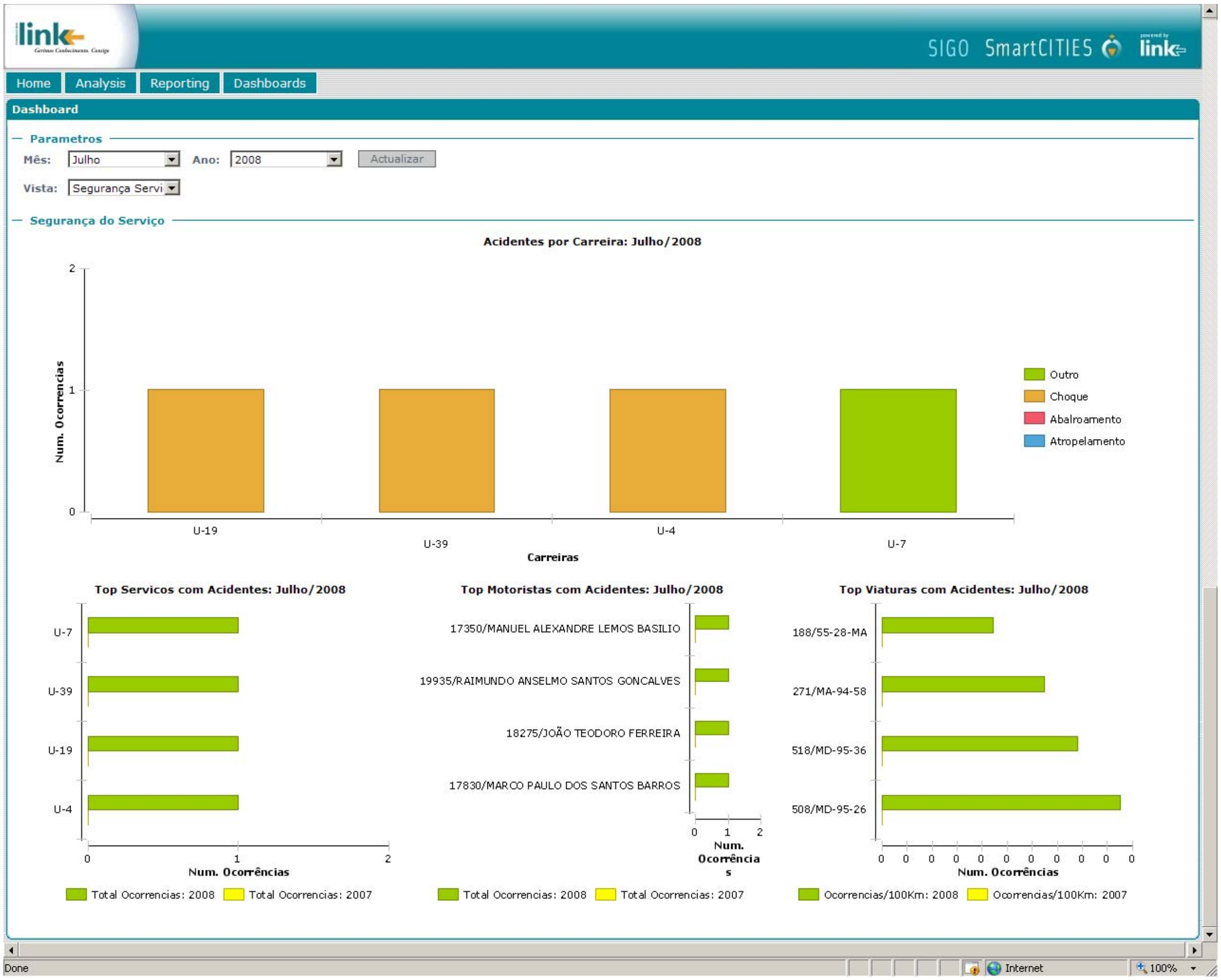
Mês: Julho

Ano: 2008

Vista: Segurança Servi

Figure 55 – Operations Dashboard ‘Segurança Serviço I’

G.10. Operations Dashboard: Segurança Serviço II



Dashboard

Parametros

Mês: Julho

Ano: 2008

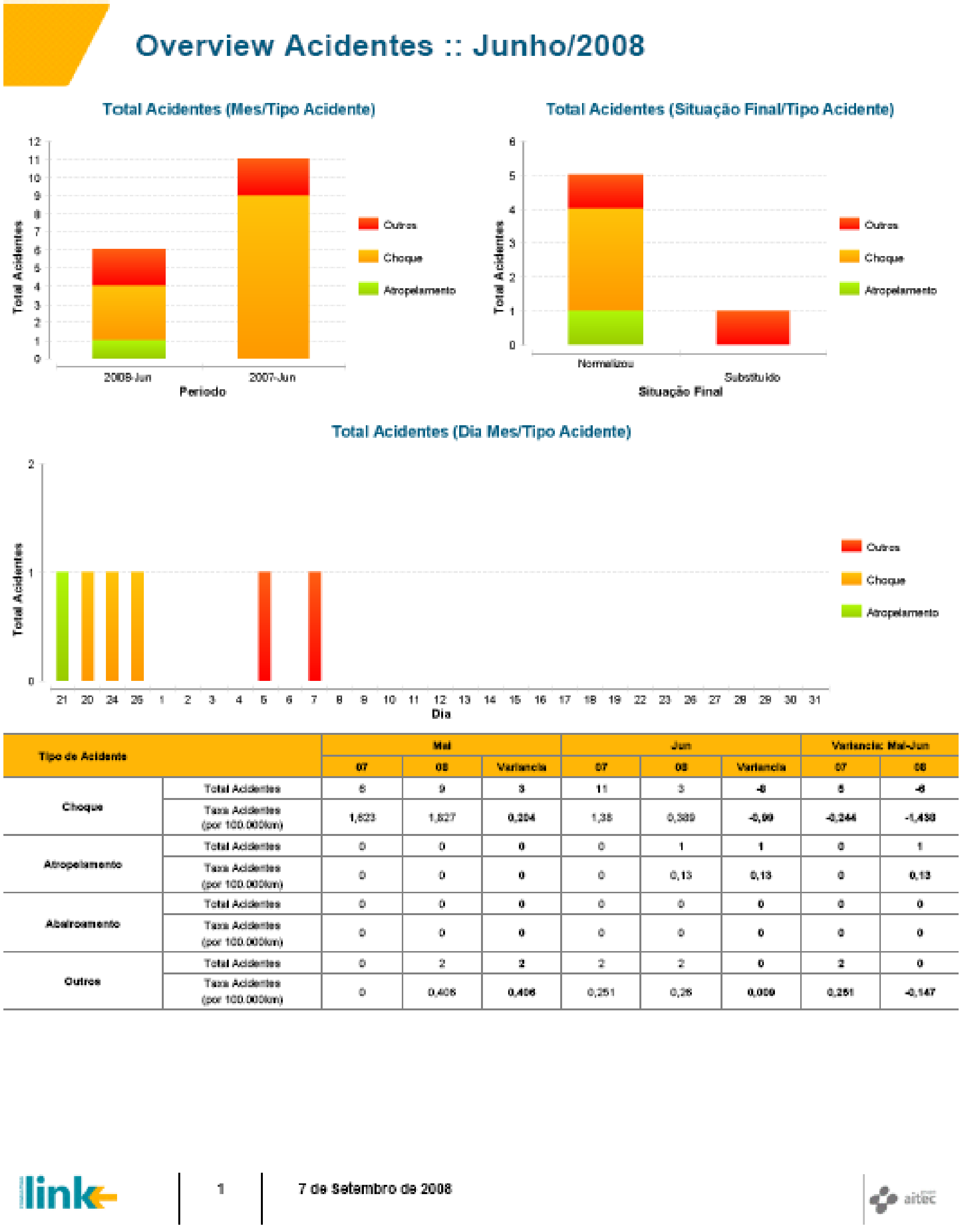
Vista: Segurança Servi

Segurança do Serviço

Figure 56 – Operations Dashboard 'Segurança Serviço II'

Appendix H: Reports

H.1. Overview Acidentes Report



Overview Acidentes :: Junho/2008

Total Acidentes (Mes/Tipo Acidente)

Total Acidentes (Situação Final/Tipo Acidente)

Total Acidentes (Dia Mes/Tipo Acidente)

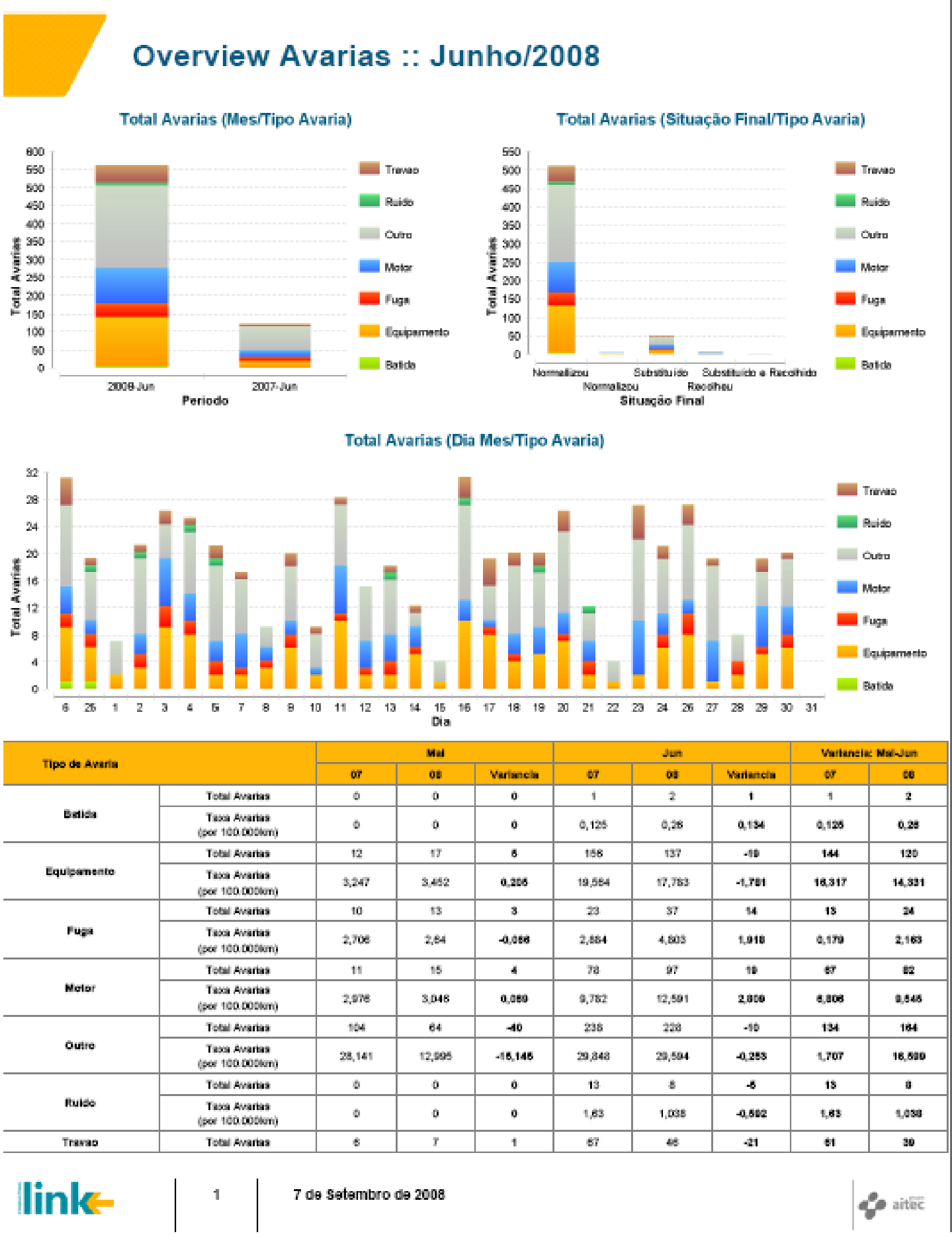
Tipo de Acidente		Mai			Jun			Variancia: Mai-Jun	
		07	08	Variancia	07	08	Variancia	07	08
Choque	Total Acidentes	6	9	3	11	3	-8	5	-6
	Taxa Acidentes (por 100.000km)	1,623	1,827	0,204	1,38	0,389	-0,99	-0,244	-1,438
Atropelamento	Total Acidentes	0	0	0	0	1	1	0	1
	Taxa Acidentes (por 100.000km)	0	0	0	0	0,13	0,13	0	0,13
Abalroamento	Total Acidentes	0	0	0	0	0	0	0	0
	Taxa Acidentes (por 100.000km)	0	0	0	0	0	0	0	0
Outros	Total Acidentes	0	2	2	2	2	0	2	0
	Taxa Acidentes (por 100.000km)	0	0,406	0,406	0,251	0,26	0,009	0,251	-0,147

link | 1 | 7 de Setembro de 2008

aitec

Figure 57 –Overview ‘Acidentes’ Report

H.2. Overview Avarias Report



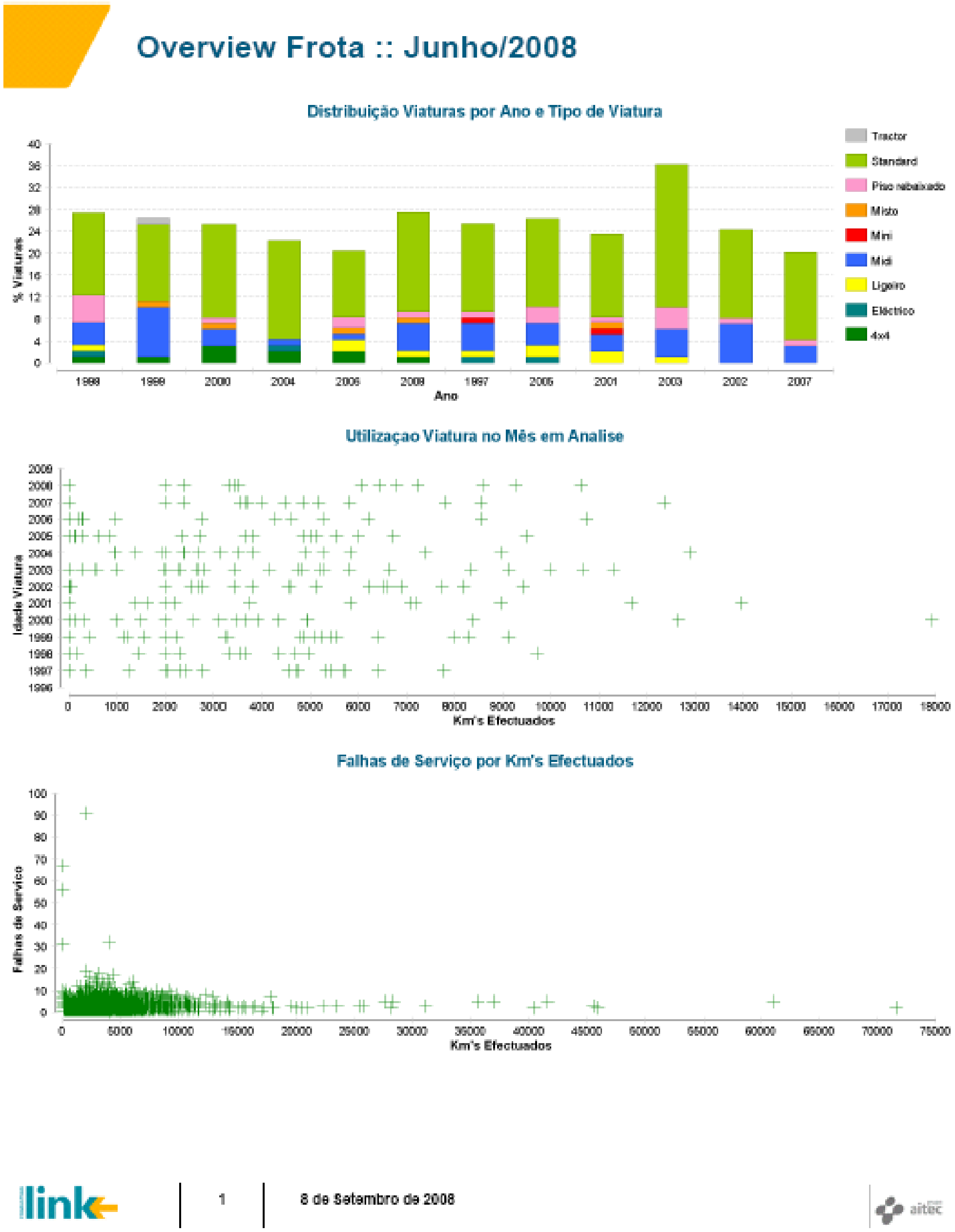
Overview Avarias :: Junho/2008

Tipo de Avaria		Mai			Jun			Variancia: Mai-Jun	
		07	08	Variancia	07	08	Variancia	07	08
Batida	Total Avarias	0	0	0	1	2	1	1	2
	Taxa Avarias (por 100.000km)	0	0	0	0,125	0,26	0,134	0,125	0,26
Equipamento	Total Avarias	12	17	5	156	137	-19	144	120
	Taxa Avarias (por 100.000km)	3,247	3,452	0,205	19,564	17,783	-1,781	16,317	14,331
Fuga	Total Avarias	10	13	3	23	37	14	13	24
	Taxa Avarias (por 100.000km)	2,706	2,64	-0,066	2,884	4,803	1,918	0,179	2,163
Motor	Total Avarias	11	15	4	78	97	19	67	82
	Taxa Avarias (por 100.000km)	2,976	3,046	0,069	9,782	12,591	2,809	6,806	9,545
Outro	Total Avarias	104	64	-40	238	228	-10	134	164
	Taxa Avarias (por 100.000km)	28,141	12,995	-15,146	29,848	29,594	-0,253	1,707	16,599
Ruido	Total Avarias	0	0	0	13	8	-5	13	8
	Taxa Avarias (por 100.000km)	0	0	0	1,63	1,038	-0,592	1,63	1,038
Travao	Total Avarias	6	7	1	67	46	-21	61	39

link | 1 | 7 de Setembro de 2008

Figure 58 – Overview 'Avarias' Report

H.3. Overview Frota Report



Overview Frota :: Junho/2008

Distribuição Viaturas por Ano e Tipo de Viatura

Utilizaçao Viatura no Mês em Analise

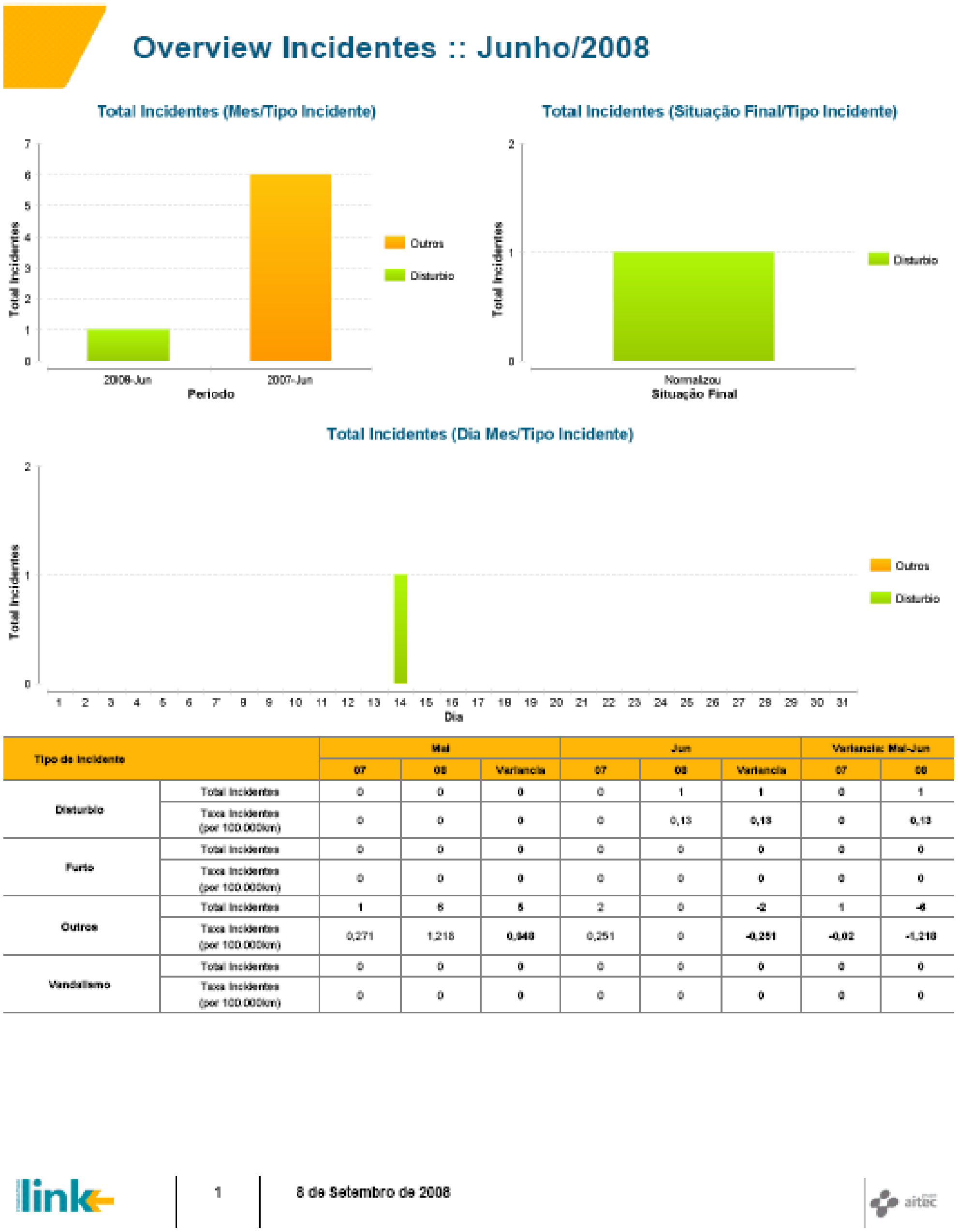
Falhas de Serviço por Km's Efectuados

1

8 de Setembro de 2008

Figure 59 – Overview 'Frota' Report

H.4. Overview Incidentes Report



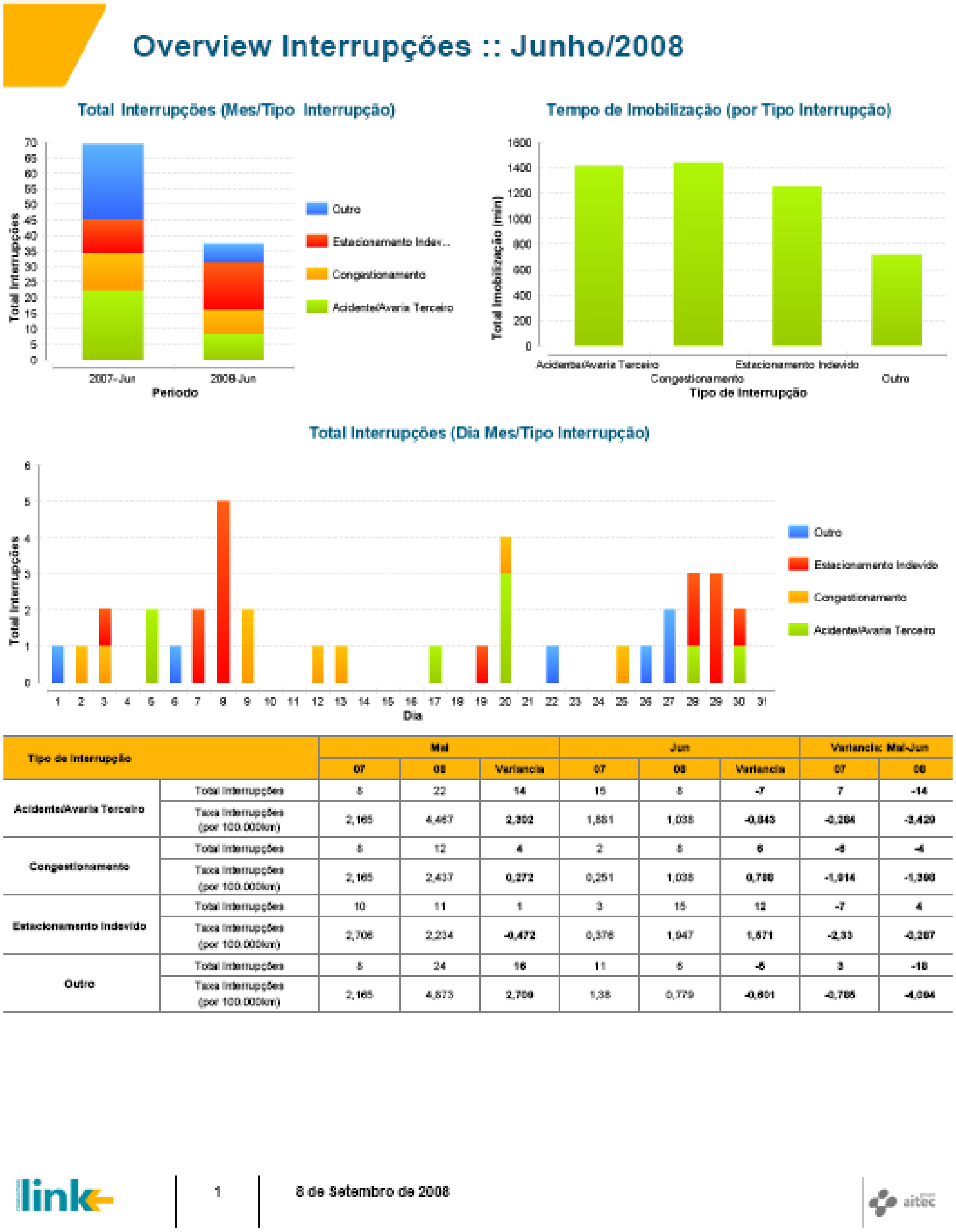
Tipo de Incidente		Mai			Jun			Variancia: Mai-Jun	
		07	08	Variancia	07	08	Variancia	07	08
Disturbio	Total Incidentes	0	0	0	0	1	1	0	1
	Taxa Incidentes (por 100.000km)	0	0	0	0	0,13	0,13	0	0,13
Furto	Total Incidentes	0	0	0	0	0	0	0	0
	Taxa Incidentes (por 100.000km)	0	0	0	0	0	0	0	0
Outros	Total Incidentes	1	6	5	2	0	-2	1	-6
	Taxa Incidentes (por 100.000km)	0,271	1,218	0,948	0,251	0	-0,251	-0,02	-1,218
Vandalismo	Total Incidentes	0	0	0	0	0	0	0	0
	Taxa Incidentes (por 100.000km)	0	0	0	0	0	0	0	0

1

8 de Setembro de 2008

Figure 60 – Overview ‘Incidentes' Report

H.5. Overview Interrupções Report



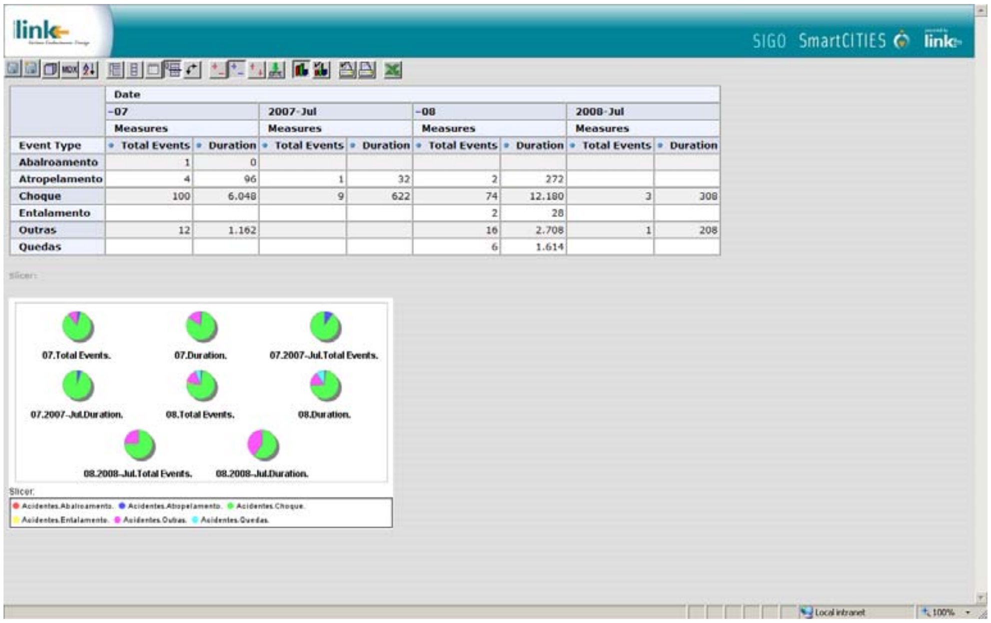
Tipo de Interrupção		Mai			Jun			Variancia: Mai-Jun	
		07	08	Variancia	07	08	Variancia	07	08
Acidente/Avaria Terceiro	Total Interrupções	8	22	14	15	8	-7	7	-14
	Taxa Interrupções (por 100.000km)	2,165	4,467	2,302	1,881	1,038	-0,843	-0,284	-3,429
Congestionamento	Total Interrupções	8	12	4	2	8	6	-6	-4
	Taxa Interrupções (por 100.000km)	2,165	2,437	0,272	0,251	1,038	0,788	-1,914	-1,399
Estacionamento Indevido	Total Interrupções	10	11	1	3	15	12	-7	4
	Taxa Interrupções (por 100.000km)	2,706	2,234	-0,472	0,376	1,947	1,571	-2,33	-0,287
Outro	Total Interrupções	8	24	16	11	6	-5	3	-18
	Taxa Interrupções (por 100.000km)	2,165	4,873	2,709	1,38	0,779	-0,601	-0,785	-4,094

link | 1 | 8 de Setembro de 2008 airtec

Figure 61 – Overview ‘Interrupções' Report

Appendix I: Analysis View

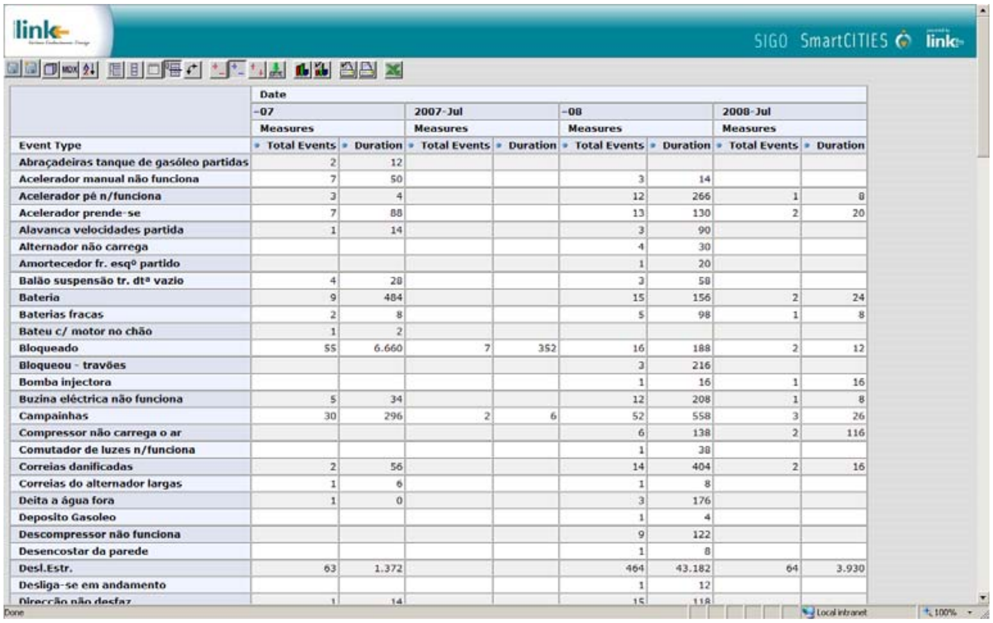
I.1. Overview Acidentes Analysis View


link
SIGO SmartCITIES
link
Slicer:
07.Total Events.
07.Duration.
07.2007-Jul.Total Events.
07.2007-Jul.Duration.
08.Total Events.
08.Duration.
08.2008-Jul.Total Events.
08.2008-Jul.Duration.
Slicer:
● Acidentes.Abalroamento. ● Acidentes.Atropelamento. ● Acidentes.Choque.
● Acidentes.Entalamento. ● Acidentes.Outras. ● Acidentes.Quedas.
Local intranet
100%

	Date							
	-07		2007-Jul		-08		2008-Jul	
	Measures		Measures		Measures		Measures	
Event Type	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration
Abalroamento	1	0						
Atropelamento	4	96	1	32	2	272		
Choque	100	6.048	9	622	74	12.180	3	308
Entalamento					2	28		
Outras	12	1.162			16	2.708	1	208
Quedas					6	1.614		

Figure 62 – Overview 'Acidentes' Analysis View

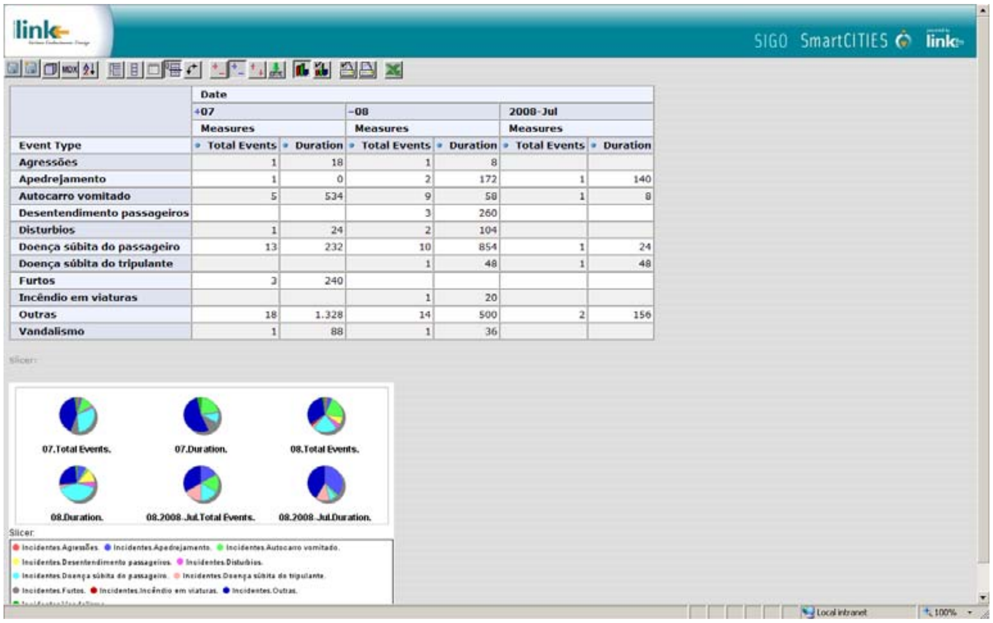
I.2. Overview Avarias Analysis View



	Date							
	-07		2007-Jul		-08		2008-Jul	
	Measures		Measures		Measures		Measures	
Event Type	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration
Abraçadeiras tanque de gasóleo partidas	2	12						
Acelerador manual não funciona	7	50			3	14		
Acelerador pé n/funciona	3	4			12	266	1	8
Acelerador prende-se	7	88			13	130	2	20
Alavanca velocidades partida	1	14			3	90		
Alternador não carrega					4	30		
Amortecedor fr. esqº partido					1	20		
Balão suspensão tr. dtª vazio	4	28			3	58		
Bateria	9	484			15	156	2	24
Baterias fracas	2	8			5	98	1	8
Bateu c/ motor no chão	1	2						
Bloqueado	55	6.660	7	352	16	188	2	12
Bloqueou - travões					3	216		
Bomba injectora					1	16	1	16
Buzina eléctrica não funciona	5	34			12	208	1	8
Campainhas	30	296	2	6	52	558	3	26
Compressor não carrega o ar					6	138	2	116
Comutador de luzes n/funciona					1	38		
Correias danificadas	2	56			14	404	2	16
Correias do alternador largas	1	6			1	8		
Deita a água fora	1	0			3	176		
Deposito Gasoleo					1	4		
Descompressor não funciona					9	122		
Desencostar da parede					1	8		
Desl.Estr.	63	1.372			464	43.182	64	3.930
Desliga-se em andamento					1	12		
Direcção não desfaz	1	14			15	118		

Figure 63 – Overview 'Avarias' Analysis View

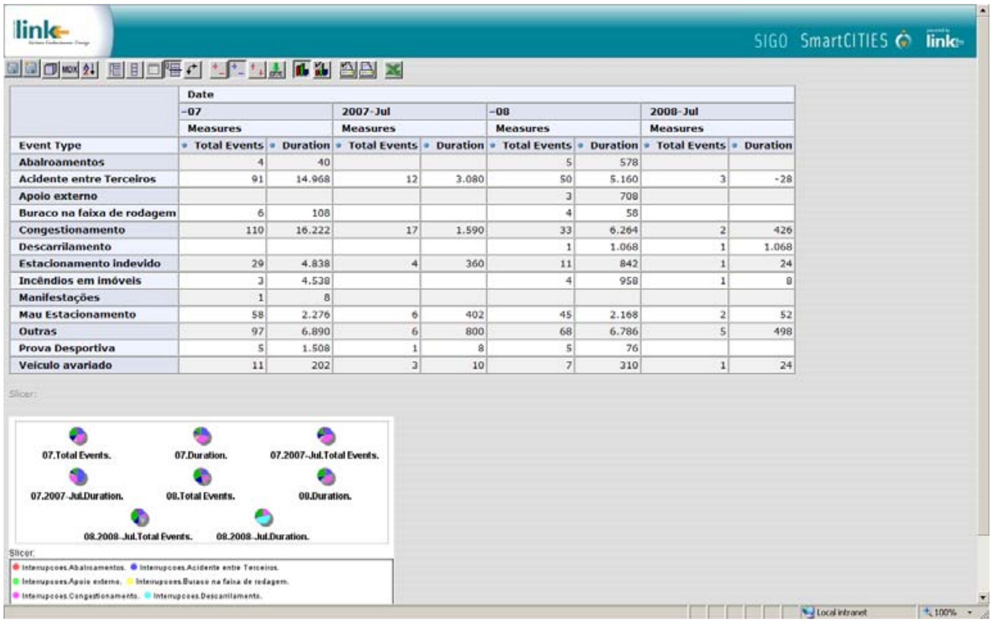
I.3. Overview Incidentes Analysis View


link
SIGO SmartCITIES
link
Slicer:
07.Total Events.
07.Duration.
08.Total Events.
08.Duration.
08.2008-Jul.Total Events.
08.2008-Jul.Duration.
Slicer:
● Incidentes.Agressões. ● Incidentes.Apedrejamento. ● Incidentes.Autocarro vomitado.
● Incidentes.Desentendimento passageiros. ● Incidentes.Disturbios.
● Incidentes.Doença súbita do passageiro. ● Incidentes.Doença súbita do tripulante.
● Incidentes.Furtos. ● Incidentes.Incêndio em viaturas. ● Incidentes.Outras.
Local intranet
100%

	Date					
	+07		-08		2008-Jul	
	Measures		Measures		Measures	
Event Type	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration
Agressões	1	18	1	8		
Apedrejamento	1	0	2	172	1	140
Autocarro vomitado	5	534	9	58	1	8
Desentendimento passageiros			3	260		
Disturbios	1	24	2	104		
Doença súbita do passageiro	13	232	10	854	1	24
Doença súbita do tripulante			1	48	1	48
Furtos	3	240				
Incêndio em viaturas			1	20		
Outras	18	1.328	14	500	2	156
Vandalismo	1	88	1	36		

Figure 64 – Overview 'Incidentes' Analysis View

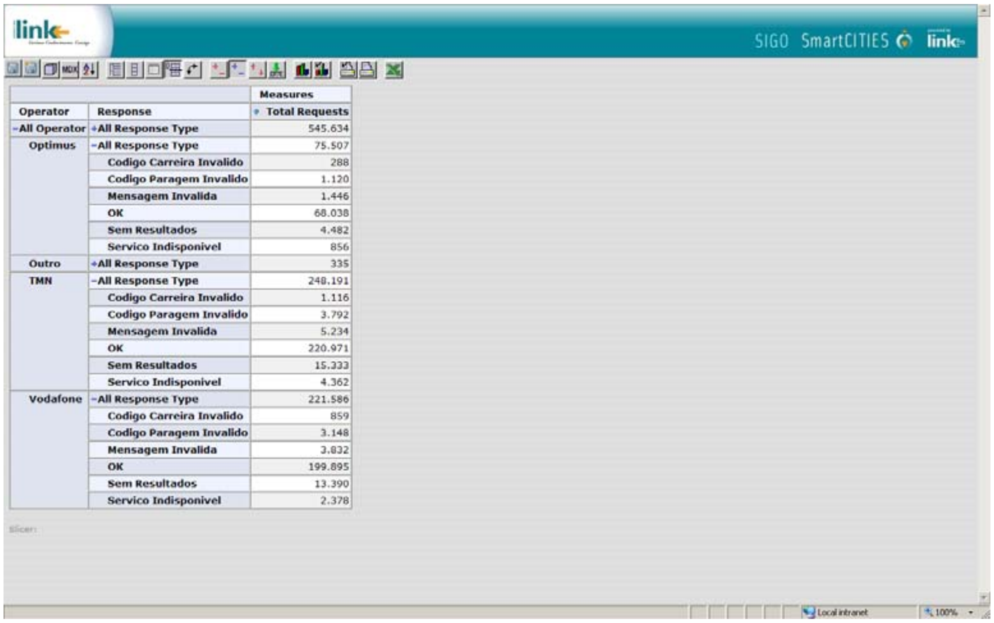
I.4. Overview Interrupções Analysis View


link
SIGO SmartCITIES
link
Slicer:
07.Total Events.
07.Duration.
07.2007-Jul.Total Events.
07.2007-Jul.Duration.
08.Total Events.
08.Duration.
08.2008-Jul.Total Events.
08.2008-Jul.Duration.
Slicer:
● Interrupcoes.Abalroamentos. ● Interrupcoes.Acidente entre Terceiros.
● Interrupcoes.Apoio externo. ● Interrupcoes.Buraco na faixa de rodagem.
● Interrupcoes.Congestionamento. ● Interrupcoes.Descarrilamento.
Local intranet
100%

	Date							
	-07		2007-Jul		-08		2008-Jul	
	Measures		Measures		Measures		Measures	
Event Type	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration	• Total Events	• Duration
Abalroamentos	4	40			5	578		
Acidente entre Terceiros	91	14.968	12	3.080	50	5.160	3	-28
Apoio externo					3	708		
Buraco na faixa de rodagem	6	108			4	58		
Congestionamento	110	16.222	17	1.590	33	6.264	2	426
Descarrilamento					1	1.068	1	1.068
Estacionamento indevido	29	4.838	4	360	11	842	1	24
Incêndios em imóveis	3	4.538			4	958	1	8
Manifestações	1	8						
Mau Estacionamento	58	2.276	6	402	45	2.168	2	52
Outras	97	6.890	6	800	68	6.786	5	498
Prova Desportiva	5	1.508	1	8	5	76		
Veiculo avariado	11	202	3	10	7	310	1	24

Figure 65 – Overview 'Interrupções' Analysis View

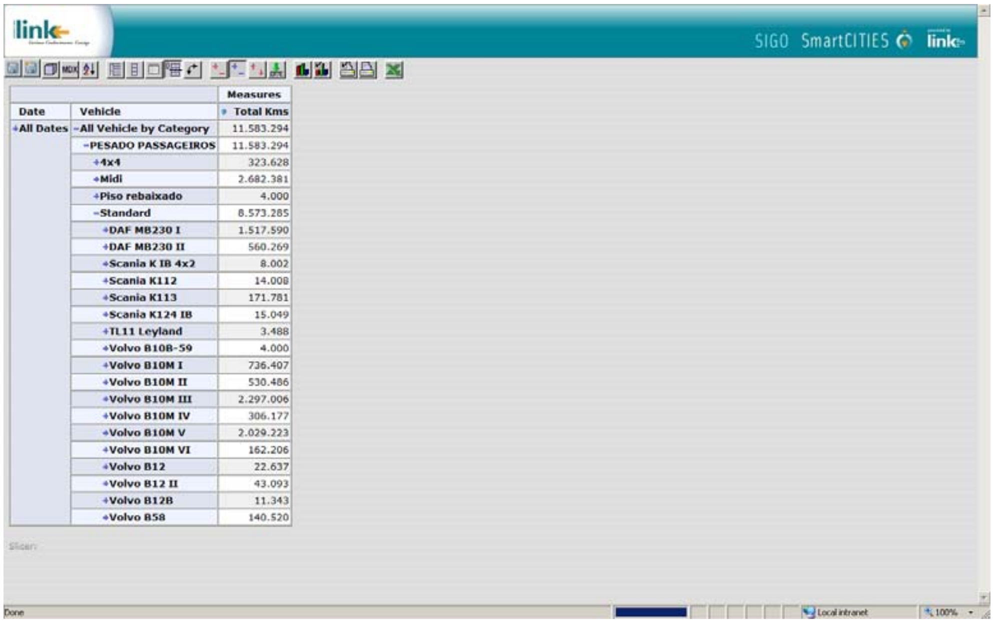
I.5. Overview SIMIP Analysis View



		Measures
Operator	Response	Total Requests
-All Operator	+All Response Type	545.634
Optimus	-All Response Type	75.507
	Codigo Carreira Invalido	288
	Codigo Paragem Invalido	1.120
	Mensagem Invalida	1.446
	OK	68.038
	Sem Resultados	4.482
	Servico Indisponivel	856
Outro	+All Response Type	335
TMN	-All Response Type	248.191
	Codigo Carreira Invalido	1.116
	Codigo Paragem Invalido	3.792
	Mensagem Invalida	5.234
	OK	220.971
	Sem Resultados	15.333
	Servico Indisponivel	4.362
Vodafone	-All Response Type	221.586
	Codigo Carreira Invalido	859
	Codigo Paragem Invalido	3.148
	Mensagem Invalida	3.832
	OK	199.895
	Sem Resultados	13.390
	Servico Indisponivel	2.378

Figure 66 – Overview 'SIMIP' Analysis View

I.6. Overview Frota Analysis View



		Measures
Date	Vehicle	Total Kms
+All Dates	-All Vehicle by Category	11.583.294
	-PESADO PASSAGEIROS	11.583.294
	+4x4	323.628
	+Midi	2.682.381
	+Piso rebaixado	4.000
	-Standard	8.573.285
	+DAF MB230 I	1.517.590
	+DAF MB230 II	560.269
	+Scania K IB 4x2	8.002
	+Scania K112	14.008
	+Scania K113	171.781
	+Scania K124 IB	15.049
	+TL11 Leyland	3.488
	+Volvo B10B-59	4.000
	+Volvo B10M I	736.407
	+Volvo B10M II	530.486
	+Volvo B10M III	2.297.006
	+Volvo B10M IV	306.177
	+Volvo B10M V	2.029.223
	+Volvo B10M VI	162.206
	+Volvo B12	22.637
	+Volvo B12 II	43.093
	+Volvo B12B	11.343
	+Volvo B58	140.520

Figure 67 – Overview 'Frota' Analysis View

Bibliography

1. **Luhn, H. P.** A Business Intelligence System. IBM Journal. October 1958.

2. **Kaniclides, T. and Kimble, C.** Executive Information Systems: A framework for their development and use. s.l. : University of York - Department of Computer Science, 1994.

3. **Podolecheva, Monika.** Open Source meets Business Intelligence. s.l. : University of Konstanz, 2006.

4. **R.F.Braams.** Benefits of Business Intelligence. Amsterdam : Vrije Universiteit, 2004.

5. **Smith, Mark.** Demystifying Open Source BI. s.l. : Ventana, 2007.

6. **Rawat, Rajeev.** The Ins and Outs of Open Source Business Intelligence. s.l. : TDWI Research, 2007.

7. **Surajit Chaudhuri, Umeshwar Dayal.** An Overview of Data Warehousing and OLAP Technology.

8. **Inmon, W.H.** Building the Data Warehouse. s.l. : John Wiley, 1992.

9. **Inmon, W.H.** *What is a Data Warehouse?* Number 1, s.l. : Prism, 1995, Vol. Volume 1.

10. **Codd, E.F., Codd, S.B. and Salley, C.T.** Providing OLAP to User-Analysts: An IT Mandate. s.l. : Arbor Software Corporation, 1993.

11. **Kimball, Ralph.** Making a List of Data About Metadata and Exploring Information Cataloging Tools. *Data Warehouse Architec.* 1998.

12. **Kaplan, Robert S. and Norton, David P.** Using the Balanced Scorecard as a Strategic Management System . *Harvard Business Review.* January/February 1996.

13. **Gartner.** *Gartner EXP Survey of More than 1,400 CIOs Shows CIOs Must Create Leverage to Remain Relevant to the Business.* 2007.

14. **Dan Vesset, Brian McDonough.** *Worldwide Business Intelligence Tools 2006 Vendors Share.* s.l. : IDC, 2006.

15. **Dan Vesset, Brian McDonough, Kathleen Wilhide.** *Worldwide Business Analytics Software 2006-2010 Forecast and 2005 Vendor Shares.* s.l. : IDC, 2005.

16. **Antos, Justin David.** *The optimal, the actual, And the possible.* s.l. : Massachusetts Institute of Technology, 2003.

17. **Shapiro, R., Hasset, K. and Arnold, F.** *Conserving Energy and Preserving the Environmnet: The Role of Public Trasnportation.* Washington DC : American Public Transportation Association, 2002.

18. **Kimball, Ralph.** *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses .*

19. **Volker Markl, Rudolf Bayer, Frank Ramsak.** Improving OLAP Performance by Multidimensional Hierarchical Clustering*.* 1999.

20. **Boston Corporate Finance.** *Enterprise Integration Software - Industry Spotlight.* s.l. : Boston Corporate Finance, 2006.

21. **Lyman, Peter and Varian, Hal R.** How Much Information. s.l. : Berkeley, 2003.