International Journal of Advanced Information Science and Technology (IJAIST)ISSN: 2319:2682
Vol.5, No.3, march 2016 **DOI:10.15693/ijaist.2016.v5.i3.74-81**

Implementation of Business Intelligence For Sales Management

Bouzekri MOUSTAID
Laboratory of informations processing and decision support,
Faculty of Sciences and Technics
Sultan Moulay Slimane University, Beni-Mellal, Morocco

Mohamed FAKIR
Laboratory of informations processing and decision support,
Faculty of Sciences and Technics
Sultan Moulay Slimane University, Beni-Mellal, Morocco

***Abstract*—Today's company operates in a socio-economic environment increasingly demanding. In such a context, it is obliged to adopt a competitive approach by exploiting at best the information that it possesses for developing appropriate action plans and taking effective decisions. The decision support systems provide to the enterprise the tools that help it for decision-making based on techniques and methodologies coming from domain of applied mathematics such as optimization, statistics and theory of the decision. The decision support systems are composed of various components such as data warehouses, ETL tools and reporting and analysis tools.**

Index terms -business intelligence, Extract, Transform, Load, Data warehouse, Data mining, Talend Open Studio, Pentaho Data Integration.

I. INTRODUCTION

The decision-making systems are based on ETL (Extract-Transform-Load) tools, whose main role is to extract data from one or more source systems (operational databases, files), to clean them, transform and load them into a data warehouse enhancing the coherence and quality of data. Therefore, the ETL system constitutes the interface between the data sources and the data warehouse.

Decision making is the fundamental goal of any organization and any management. One of the main problems is to determine relevant information for decision making. It is therefore essential to use Interactive Systems Decision Support, denoted DSS (DSS English: Decision Support Systems), which provide tools for assessing various alternatives and their impacts for optimal decision making.

The decision is defined as a choice between several alternative actions at a given moment in time [1]. It is assimilated to an act, action or process of solving problem facing to the individual or organization. In general, we call decision making any mantel process after which everyone, in front of several alternatives, choose one of them.

Decision aiding can be defined as the activity of the person who, through the use of explicit but not necessarily completely formalized models, helps obtain elements of responses to the questions posed by a stakeholder in a decision process. These elements work towards clarifying the decision and usually towards recommending, or simply favoring, a behavior that will increase the consistency between the evolution of the process and this stakeholder's objectives and value system [2].

To support this decision support in the most efficient way, the development of computer systems is necessary and inevitable.

II. Decision Support Systems

Keen and Scott-Morton [3] present the Systems Decision Support (DSS) as systems designed to solve decision problems little or poorly structured. The SIAD incorporate the statistics, the operations research, the optimization algorithms and the numerical computations and manage information (databases, file management and information flow within the company).

The decision information system is a set of data organized in specific way, easily accessible and appropriate for the decision making or an intelligent representation of these data through specialized tools [4]. The main interest of a decision support system is to provide the decision maker a transversal vision of the company in all of its dimensions.

Two main functions are designed for decision support tools:

- Collecting, Storing and Transforming: ETL, Datawarehouse, Datamart, Dataweb.
- Extracting and Presenting: Data mining, OLAP.

The different components of a decisional system:

Datawarehouse: is a collection of thematic data, integrated, non-volatile and historiated organized for decision making [5].

Datamart: This is a departmental solution of Datawarehouse supporting a portion of the data and business functions. It is a subset of a Datawarehouse that contains only data of a company's craft.

ETL: is an inter-software technology to extract data from multiple sources, transform it and load it in one or more destinations.

OLAP (On Line Analytical Processing): Online analytical processing is the technology that can produce descriptive syntheses online (or views) of data contained in Datawarehouses. OLAP is based on a data structure especially adapted to the crossings and extractions: hypercube (or cube).

MOLAP: systems whose type is MOLAP constitute an approach which allows representing data of Datawarehouse as a multidimensional array with n dimensions, where each dimension of the array is associated with a dimension of the hyper cube of data.

ROLAP: Systems whose type is ROLAP use a relational representation of the data cube. Every fact is a table called fact table and each dimension corresponds to a table called dimension table.

Data mining: Data mining is the set of methods and techniques for exploring and analyzing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies; special systems output the essentials of the useful information while reducing the quantity of data [6].

There are two types of Data mining's techniques:

- The descriptive (or exploratory) techniques are designed to bring out information that is present but buried in a mass of data (as in the case of automatic clustering of individuals and searches for associations between products or medicines).
- The predictive (or explanatory) techniques are designed to extrapolate new information based on the present information, this new information being qualitative (in the form of classification or scoring) or quantitative (regression).

III. ETL (EXTRACT/TRANSFORM/LOAD)

A. ETL processes

BI applications are based on data coming from different data sources, which can be managed by different operating systems. The ELT process provides the fusion of data coming from these heterogeneous platforms and transforms it into a standard format for the target databases in the environment of decision support.

The ETL process is composed of the following:

- **Reformatting**: data source coming from different databases and different files must be formatted in a common format.
- **Conciliation**: Redundancies cause inconsistencies. They must be found and reconciled during the ETL process.
- **Cleaning**: The goal is to clear the erroneous data that were found during the analysis.

B. Design of the extraction programs

The extraction process can be done in two ways: duplicate data sources and give this data to ETL developers in order to exploit them or work directly on the source data by querying the operational system.

The first method has the advantage of avoiding the congestion of the operational system by the massive querying to perform data extraction. However, its disadvantage is the increase and complexity of the task of the team developing the ETL process.

Regarding the second method, it allows the development team of the ETL process to query directly data sources and thus target the subset that they need. However, in some cases, this method can overload the operational system and prevent proper operation.

C. Design of the transformation programs

Transformation is the major part of ETL process. During this phase, the main problems of data sources are:

- Inconsistent primary keys
- Inconsistent data values
- Different data formats
- Inaccurate or missing data values
- Synonyms and homonyms
- Embedded process logic

The operations of transformations most encountered are as follows:

- Part of the data must be renamed according to the standards of naming decision project.
- Some elements of source data should be merged into a single data element.
- Translation of certain data elements in mnemonics.

D. Design of the load programs

The last step of ETL processes is loading data after the previous two steps in the decision-making target databases, this can be done in two ways: insert new rows in tables or use the load utility DBMS. However, it is necessary to study referential integrity and indexing.

IV. OPEN SOURCE ETL: TALEND OPEN STUDIO / PENTAHO DATA INTEGRATION

The field of Business Intelligence saw the appearance of free software covering all areas of decision: reporting, multidimensional analysis, data mining and of course the ETL.

TOS and PDI compete effectively the owners ETL, and they have a real alternative. Both tools derive their reputations of their abilities and their performances. Moreover, these two products occupy an important place in the Magic Quadrant of the Gartner Group published in July 2013.

A. Presentation of Talend Open Studio

Talend Open Studio is developed by the French company Talend. The first version of "Talend Open Studio" came into being in 2006, and the current version is 5.4. TOS is an ETL whose type is "code generator". It provides a graphical interface, the "Job Designer" (based on Eclipse RCP), which allows the creation of process of data manipulation.

Characteristics:

- Compatibility with multiple operating systems
- Prerequisites: 3GB of memory (4GB recommended), 3GB of disk space for installation and over 3GB for use.
- Traces and statistics of performance in real time.
- Enrichment of treatments by adding specific code (in Java or Perl).
- Integration with large number of DBMS.

Environment of design under TOS:

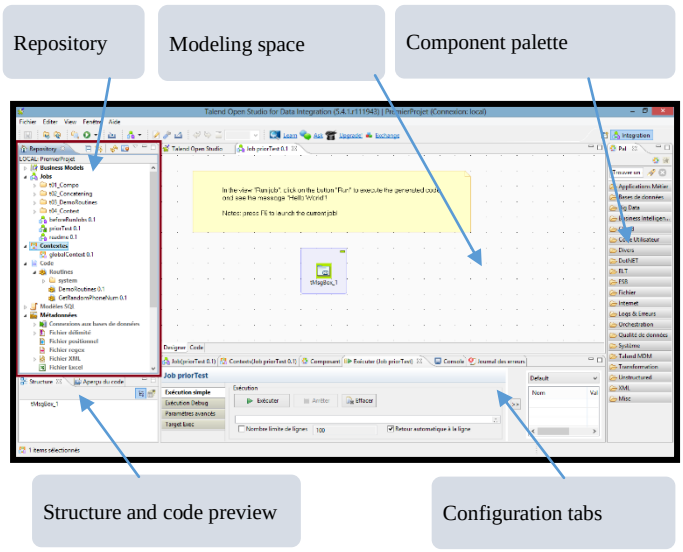

Figure 1. Components of environment design under TOS

Modeling space: where developers place and configure the components to build a data integration task. It is the key window of development.

Component palette: Contains components that can be used in data integration tasks.

Configuration tabs: shows properties of task or specific components that are selected in the design space.

Structure: lists the components and allows quick access to standard variables for each component.

Code preview: displays a preview of the code associated with each component.

B. Presentation of Pentaho Data Integration

Pentaho Data Integration (PDI) is an ETL whose type is "transformation engine". PDI originally called kettle, it is acquired by Pentaho Corporation in April 2006. Similarly Matt Casters, the founder of the kettle, also joined the Pentaho team. Pentaho Data Integration has the "Spoon" GUI based on SWT (Standard Widget Toolkit), enabling the creation of two types of treatments: Transformations and Tasks (Jobs) using the version 4.4.0.

Characteristics:

- Compatibility with multiple operating systems
- It is easy to install, it comes to decompressing a file containing the tool, available at: http://www.community.pentaho.com/.
- It allows the preview of the data streams processed.
- It allows the execution of processes on the local machine, a remote server, or set of servers.
- It fits perfectly with the business intelligence platform Pentaho.
- Very flexible and easy to customize.

Environment of design under PDI:

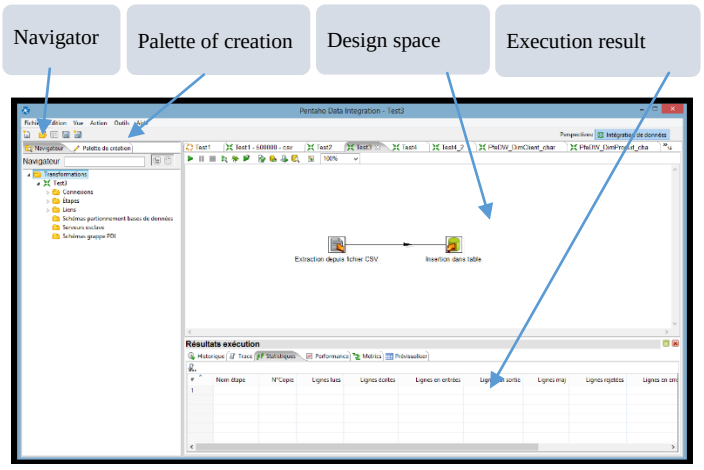

Figure 2. Components of environment design under PDI

Navigator: includes objects which are in association with a particular transformation.

Palette of creation: Contains components that can be used in creating data transformations.

Design space: where developers place and configure the steps to build a processing or data integration task, it is the key window of development.

Execution Result: shows the properties of execution results of a transformation or a task.

C. Functionalities Comparison

A comparison of the functionalities [7] and the processing times was made by the following versions:

- Pentaho Data Integration 4.4.0.
- Talend Open Studio v5.4

Access to relational databases

	Pentaho Data Integration	Talend Open Studio
read full table	Yes	Yes
read full view	Yes	Yes
calling stored procedures	Yes	Yes
add clause where / order by	Yes	Yes
query execution	Yes	Yes
query design tool	No	Yes
reading / writing of all the simple types of data	Yes	Yes
reading / writing of complex data types	No	cartographic data

Both tools have the ability to access databases implemented within different DBMS.

Triggering processes by message

	Pentaho Data Integration	Talend Open Studio
CORBA	No	Yes
XML RPC	No	Yes

JMS	Yes	Yes
MOMS	No	Yes

TOS stands out regarding the triggering process by message in comparison with PDI. Note that both tools do not support the triggering by the CORBA protocol.

Transformations and calculations default

	Pentaho Data Integration	Talend Open Studio
Transformation functions of dates and numbers	Yes	Yes
Statistical functions of quality	Yes	Yes
Allows transcoding with a reference table	No	No
Heterogeneous joins	No	No
Join modes supported (BD)	Yes	Only join of Flows
Management of nested queries	No	No

TOS and PDI provide the basic functions for modeling elementary transformations that is the functions of transformation of dates, strings and numbers.

Manual transformations

	Pentaho Data Integration	Talend Open Studio
possibility of processing by a programming language	Yes	Yes
adding new transformations and business processes	Yes	Yes

In addition to their transformation functions default TOS and PDI make available to developers the means to add new features to meet their business needs.

Flat files

	Pentaho Data Integration	Talend Open Studio
CSV	Yes	Yes
fixed / limited	Yes	Yes
XML	Yes	Yes
Excel	Yes	Yes
validate flat files	No	Yes
validate XML files	Yes	Yes

TOS and PDI allow easy access to data in files.

Triggering by polling

	Pentaho Data Integration	Talend Open Studio
Folder	Yes	Yes
POP	Yes	Yes
Socket	No	Yes

TOS and PDI make available the means to wait for specific events, such as the appearance of a file in a directory, to orchestrate data integration treatments.

Advanced development

	Pentaho Data Integration	Talend Open Studio
Presence of an API	Yes	Yes
Integration of external functions	Yes	Yes
Crash recovery mechanism	No	Entreprise Edition
Parameterization of buffers / indexes / caches	Yes	Yes
Management team development	Yes	Yes, but paying
Versioning	No	Yes

Among the features offered by TOS and PDI, we find API for supporting the development of advanced data integration process. However, these tools do not offer error recovery.

Processing data

	Pentaho Data Integration	Talend Open Studio
Graphical mapping	Yes	Yes
Drag and Drop	Yes	Yes
Graphical representation of flows	Yes	Yes
Data visualization in development	Yes	Yes
Impact analysis tools	Yes	Entreprise Edition
Debugging tools	Yes	Yes
Management of technical documentation	No	Yes
Management of functional documentation	No	Yes
Management of documentation through the web	Yes	Yes
Management of integration errors	certain steps	Yes

TOS and PDI offer mechanisms graphical Mapping and Drag and Drop which makes them relatively easy to take in hand to develop treatments for data integration.

Deployment

	Pentaho Data Integration	Talend Open Studio
Compilation treatments	No	Yes for JAVA No For PERL
Type start of production	Command line windows or unix	

PDI is 'transformation engine'. Thus each transformation and each task are stored as meta-language and which may be stored either in XML or in a database. Therefore, treatments designed under PDI cannot be compiled. Conversely, TOS is 'code generator'. So it generates a code for each job either in Java or Perl. Therefore, treatments designed under TOS can be compiled for the case of the Java language.

Connectors

	Pentaho Data Integration	Talend Open Studio
Connectors	OpenERP, SalesForce, SAP (Read)	Connectors CRM (SugarCRM, SalesForce, ...) Connectors ERP

The application connectors allow interoperability between ETL tool and applications. In this context, we note that TOS offers more possibilities than PDI.

Security

	Pentaho Data Integration	Talend Open Studio
Use of rights of a directory	No	No
Security type	Security DBMS that contains the repository	Owner
Security scenario creation	Yes	Yes
Security update scenario	Yes	Yes
Security access to metadata	Yes	Yes
Security on the administration console	Yes	Yes
Security on the manual launch of tasks	Yes	No

TOS and PDI are equipped with security mechanisms. The security under PDI is based on the security of DBMS. While TOS has its own scenarios.

Others

	Pentaho Data Integration	Talend Open Studio
Web Services	Yes	Yes
OLAP Cubes (Mondrian)	Yes	Yes
Various	LDAP, RSS	RSS, LDAP, MOM, SCP, XMLRPC

Both TOS tools and PDI support the web service and OLAP.

D. Comparison of processing times

The comparison of the processing time [7] is made being varied source files and files destinations. So, the following four tests were realized and are graphically presented expressing the processing time according to number of lines treaties.

TEST No.1

This test involves extracting data from a CSV file and load them into another CSV file while changing the separator ';' of the source file by ',' in the target file. The source file has a structure that has seven fields: sequence; now; first; second; third; fourth; fifth. Here is an excerpt of this file:

```
001;2013/09/0510:44:43.014;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
002;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
003;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
004;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
005;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
006;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
007;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
008;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
009;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
010;2013/09/0510:44:43.029;12.345;undeuxtroisquatrecinq;0304/12/0500:00:00.000;Y;12345
```

The target file is constructed according to the same structure with the only difference which involves changing the separator.

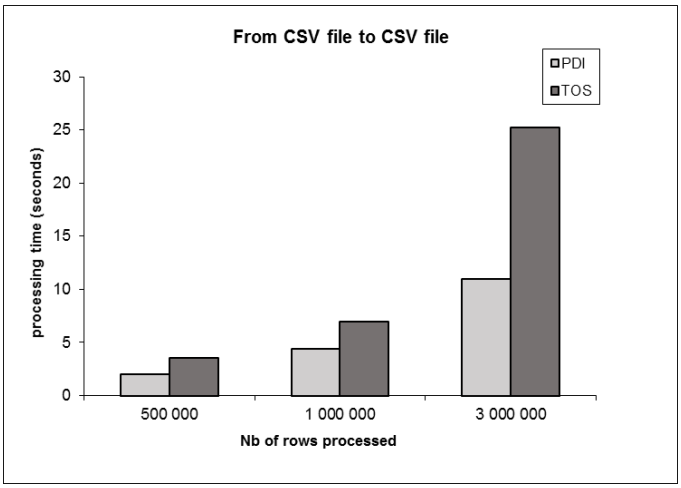

Figure 3. Results of test No.1

Based on this test, TOS has taken double more than PDI in terms of the execution time for the extraction of data from a CSV file and loading them into another CSV file.

TEST No. 2

The test consists of extracting data from a CSV file and loads it into an XML file. The source file has the same structure as that of the previous test. The target file has a structure that maps each element of the file to a XML tag. Below an illustrative extract from the file structure:

```
<root>
  <row>
    <sequence>0000000001</sequence>
      <maintenant>2013/09/0510:44:43.014</maintenant >
      <premier>12.345</premier>
      <second>undeuxtroisquatrecinq</second>
     <troisieme >0304/12/0500:00:00.000</ troisieme >
     <quatrieme >Y</ quatrieme >
     <cinquieme >12345</ cinquieme >
  </row>
<root>
```

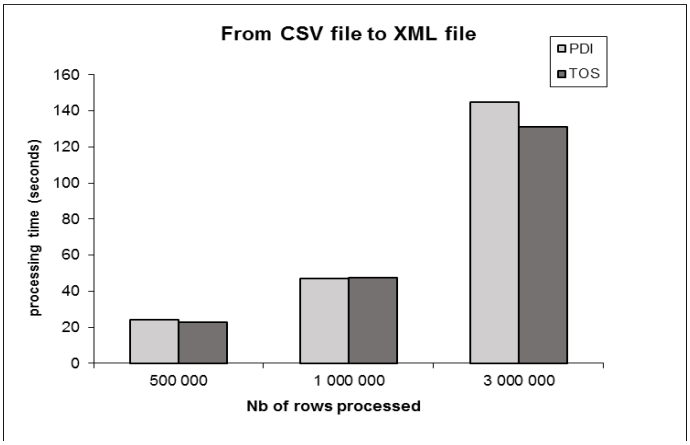

Figure 4. Result of test No. 2

The results of this test are similar for the two tools, they consume almost the same execution time to extract source data from a CSV file and load it into an XML file. Thus, the two tools have the same performances for this test.

TEST No. 3

Here, we perform the extraction and loading of data from a CSV file into a table managed by the MySQL DBMS. The source file has the same structure as in the previous test. Each column of the table is associated with an element of the source file:

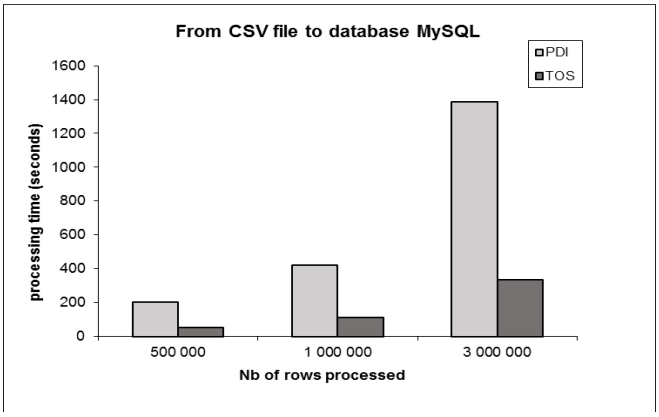

Figure 5. Result of test No. 3

On this test, TOS has much more interesting performance against PDI. TOS is three times faster than PDI to extract source data from CSV files and load them into a table managed by MySQL.

TEST No. 4

This test involves extracting data from a CSV file and loads them into another CSV file. Between the extraction and loading is carried out a transformation of dates. In the case of TOS we use the powerful tMap, while for the case of PDI, we use Rhino.

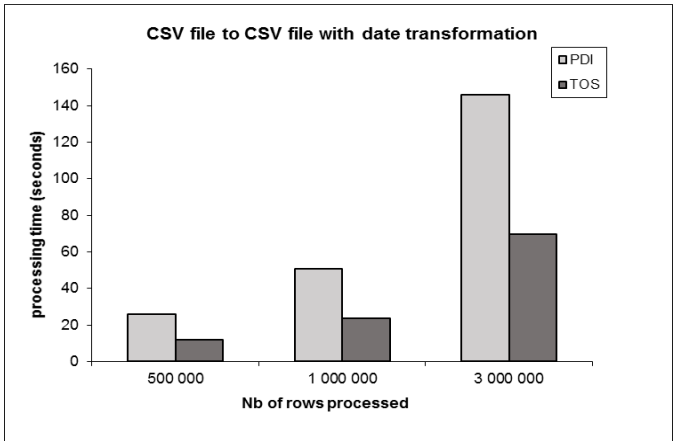

Figure 6. Result of test No. 4

TOS is twice as fast as PDI to extract source data from CSV files and ensure the transformation and loading in other CSV files.

V. IMPLEMENTATION OF BI SYSTEM

A. Data warehouse design

The design of the Data Warehouse schema for commercial management according to the snowflake approach produced a diagram consists of seven tables:

- A fact table: FaitFacture.
- Six dimensions tables: DimProduit, DimClient, DimPays, DimFormeJuridique, DimEffectif and DimTemps.

These tables and their relationships are illustrated in the Figure 7.

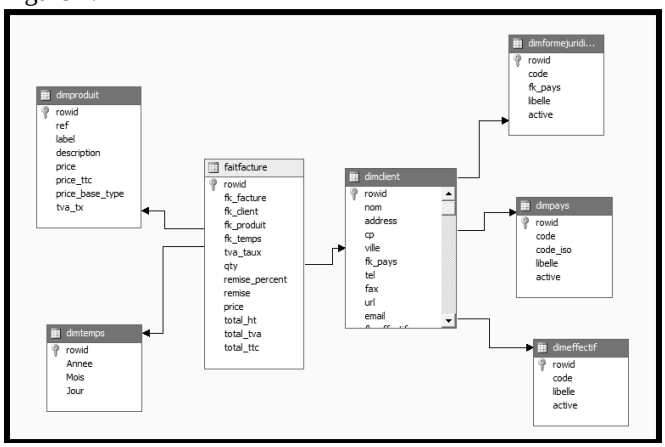

Figure 7. Data Warehouse schema

B. ETL process under PDI

The figure 8 illustrates the ETL design for the fact table FaitFacture:

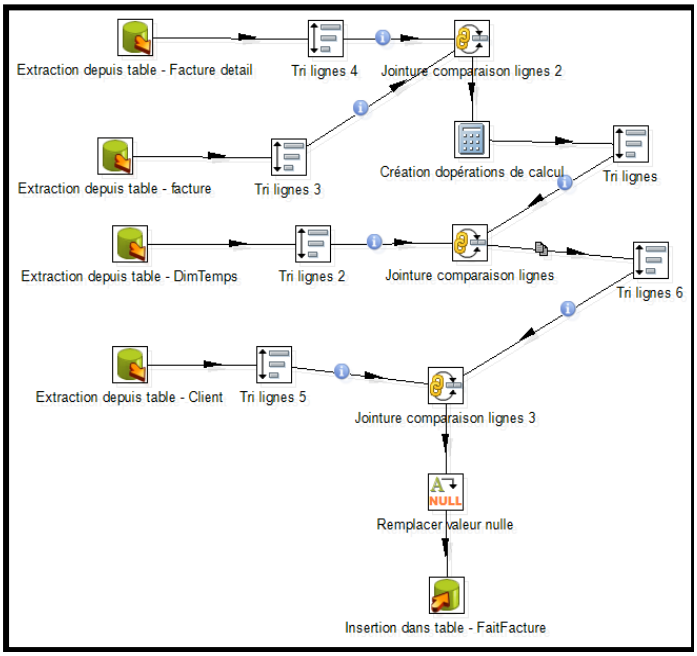

Figure 8. ETL process for the fact table "FaitFacture" under PDI

C. ETL process under TOS

ETL modeling for the fact table FaitFacture with Talend Open Studio is given in the Figure 9.

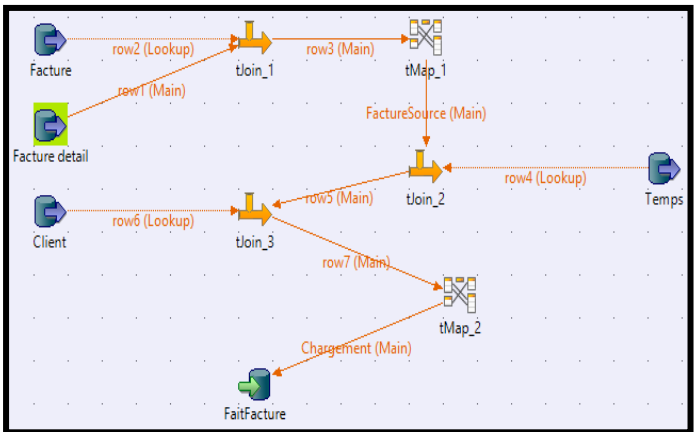

Figure 9. ETL process schema for the fact table "FaitFacture" under TOS

D. Association rules

In order to proceed to the analysis of data from the data warehouse implemented, we used the Weka software and web interface developed in J2EE.

The results obtained using the Apriori algorithm with Weka is illustrated in figure 10.

The same results are obtained using the Apriori algorithm showed on the web interface developed in J2EE (Figure 11).

```
Best rules found:

 1. JOINT CYLINDRE=t JOINT GROUPE d=t 14 ==> JOINT CARTER=t 14    conf:(1)
 2. JOINT CARTER=t JOINT CYLINDRE=t 14 ==> JOINT GROUPE d=t 14    conf:(1)
 3. JOINT CARTER=t JOINT CULASSE=t 13 ==> JOINT CYLINDRE=t 13    conf:(1)
 4. JOINT CARTER=t JOINT CULASSE=t 13 ==> JOINT GROUPE d=t 13    conf:(1)
 5. JOINT CULASSE=t JOINT CYLINDRE=t JOINT GROUPE d=t 13 ==> JOINT CARTER=t 13    conf:(1)
 6. JOINT CARTER=t JOINT CULASSE=t JOINT GROUPE d=t 13 ==> JOINT CYLINDRE=t 13    conf:(1)
 7. JOINT CARTER=t JOINT CULASSE=t JOINT CYLINDRE=t 13 ==> JOINT GROUPE d=t 13    conf:(1)
 8. JOINT CARTER=t JOINT CULASSE=t 13 ==> JOINT CYLINDRE=t JOINT GROUPE d=t 13    conf:(1)
 9. KICK DE MANIVELLE=t CYLINBLOC SCOOTER=t 11 ==> CYLINBLOC TREK=t 11    conf:(1)
10. CARBURATEUR=t POIGNET DROIT=t 10 ==> POIGNEE GAUCHE FLY=t 10    conf:(1)
11. CARBURATEUR=t POIGNEE GAUCHE FLY=t 10 ==> POIGNET DROIT=t 10    conf:(1)
```

Figure 10. Result of Apriori algorithm using Weka

Support minimum : 0.01
Nombre de règles : 30
Confiance minimum : 0.0
Envoyer

Liste des règles d'association entre produits

Antécédants	Conséquents	Support	Confiance
1. JOINT CYLINDRE; JOINT GROUPE d	JOINT CARTER	14.0	1.0
2. JOINT CARTER; JOINT CYLINDRE	JOINT GROUPE d	14.0	1.0
3. JOINT CARTER; JOINT CULASSE	JOINT CYLINDRE	13.0	1.0
4. JOINT CARTER; JOINT CULASSE	JOINT GROUPE d	13.0	1.0
5. JOINT CULASSE; JOINT CYLINDRE; JOINT GROUPE d	JOINT CARTER	13.0	1.0
6. JOINT CARTER; JOINT CULASSE; JOINT GROUPE d	JOINT CYLINDRE	13.0	1.0
7. JOINT CARTER; JOINT CULASSE; JOINT CYLINDRE	JOINT GROUPE d	13.0	1.0
8. JOINT CARTER; JOINT CULASSE	JOINT CYLINDRE ; JOINT GROUPE d	13.0	1.0
9. KICK DE MANIVELLE; CYLINBLOC SCOOTER	CYLINBLOC TREK	11.0	1.0
10. CARBURATEUR; POIGNET DROIT	POIGNEE GAUCHE FLY	10.0	1.0

Figure 11. Result of Apriori algorithm using J2EE application

The application of Apriori algorithm allows the extraction of knowledge in the form of association rules between products of consequent and the products listed in the antecedent of the rule. For example, consider the association rule:

(JOINT CYLINDRE, JOINT GROUPE) ➔ *(JOINT CARTER)*

This rule has a 100% confidence. This result is very important for decision making for procurement. Indeed, it will be more beneficial to order quantities of the product "JOINT CARTER" in proportion to the quantities ordered for products "JOINT CYLINDRE" and "JOINT GROUPE", and to include the product "JOINT CARTER" in any promotional offers including products "JOINT CYLINDRE" and "JOINT GROUPE".

VI. CONCLUSION

This work consisted in the construction of a decision support system for the management of sales. For that purpose, we presented the notion of the Decision, the notion of Decision-making support and that of the Decision support system as well as their components of extract, transform and load (ETL), storage of data, and the presentation tool layer such as querying, analysis (Data Mining) and reporting.

So during this work, we presented both ETL Talend Open Studio and Pentaho Data Integration and then their features were compared between them. Both ETL is of OPEN SOURCE types, are complementary and establish real

alternatives to one ETL owners as Informatica Power Center, Oracle Warehouse Builder, Cognos Decision Stream, etc.

At the end of this work, it was preceded to the writing of the query on the data of data warehouse by using the extension of SQL for OLAP (SQL3). So and to extract from the knowledge in the form of rules of associations between products, the algorithm Apriori of data mining was used via the software WEKA. Also, a Web interface was developed in J2EE to facilitate the use of this algorithm.

REFERENCES

[1] Schneider, D. K. (1994). Modélisation de la démarche du décideur politique dans la perspective de l'intelligence artificielle. Thèse de l'Université de Genève, Suisse.
[2] ROY B., BOUYSSOU D., 1993. Aide multicritère à la décision: méthodes et cas, Economica, Paris.
[3] Keen, P., & M. Scott-Morton (1978): "Decision Support Systems: an organizational perspective", Addison-Wesley Publishing.
[4] Goglin J.-F. (2001, 1998). La construction du datawarehouse : du datamart au dataweb. Hermes, 2ème édition.
[5] INMON W.H. Building the Data Warehouse. 2nd Ed. New York: Wiley, 1996, 401 p.
[6] Stéphane TUFFERY, 2007. Data mining et statistique décisionnelle: l'intelligence des données.
[7] LIVRE BLANC Les ETL Open Source Une réelle alternative aux solutions propriétaires Sylvain DECLOIX - Responsable Pôle OSBI Atol Conseils et Développements

Authors Profile

Bouzekri MOUSTAID is a PHD research student in the Computer Sciences Department Sultan Moulay Slimane University. He received a degree in chemical engineering and energy from the national school of the Mineral industry, Rabat Morocco in 1990, and a Master degree of Multimedia databases and integration from the University of Nice Sophia Antipolis in 2003. His research interest includes big data, datamining.

M.FAKIR received a master degree from Nagaoka university of Technology in 1991 and a PHD degree from University Cadi Ayyad in 2001. He is currently with the Faculty of Scinences and Technics, University Sultan Moulay Slimane, Morocco. His research interest includes image processing, pattern recognition and datamining.