Performance and analyses using two ETL extraction software solutions

Abdellah AMINE

Sultan Moulay Slimane University, Beni Mellal, Morocco

a.amine@usms.ma

Rachid AIT DAOUD

Sultan Moulay Slimane University, Beni Mellal, Morocco

daoud.rachid@gmail.com

Belaid BOUIKHALENE

Sultan Moulay Slimane University, Beni Mellal, Morocco

b.bouikhalene@usms.ma

Abstract—

In the prospect of doing a set of decision support onboards in a public university, we will present a comparison of two ETL extraction based in a production databases of students' information. For the deployment, we use Pentaho and Sql Server Tools and we demonstrate the application on the case of Sultan Moulay Slimane University in Beni Mellal, Morocco

Keywords—*:* Pentaho; Sql Server; Data Warehouse; Business Intelligence

I. INTRODUCTION

Data warehouse (DWs) is delineated as "subject-oriented, more integrated, timely-variant, and non-volatile collection of data to support the management decision process" [1]. Data warehouse emphasizes the collection of data from multiple sources for useful analysis.

At the center of DWs is the extraction-transformation-loading (ETL) process. ETL is a process utilized to extract data from multiple sources, transform that data to the desired state through cleansing, and load it into a target database. The deliverable is used to generate reports and for analysis. ETL consumes up to 70% of all the resources [2-5].

In the most professional field, the main approach before selecting an ETL tool is to perform proofs of concept. However, it is almost impossible to perform proofs of concept of all ETL tools available on the market. Then a pre-selection is made in the way that two ETL suites are kept for testing. This pre-selection is generally based on criteria summarized as follows: the category of the tool, the cost, the type of ETL project, and the proof of concepts.

In this white paper, we will only look at the use of two ETL tools (Microsoft SQL Server Integration Services SSIS and Pentaho Kettle) [6] based on the generalized criteria for selecting the better tool.

II. RELATED WORK

In the recent years, a number of different approaches have been suggested for the design, optimization, and automation of ETL operations. In this section, we present a brief overview of these several approaches [7]. Some of the leading data integration vendors are IBM, Informatica, Oracle, Microsoft, Talend, Pentaho, Information Builders, etc.

There are many available research papers that offer a comparative view of the leading ETL tools in the market, such as [8-9]. They analyze in details the functionalities and features offered by these tools, and it can be deduced that all of them provide support for all the features that define data integration tools.

Different variants of some approaches for integration of ETL tools with data warehouses have been proposes. Shaker H. Ali ElSappagh tries to navigate through the efforts that have been made to use acronyms for ETL, DW, DM, OLAP, Ion-line analytical processing.A data warehouse gives a set of numerical values based on a set of input values in the form of dimensions [10]. Li, Jain, overcame the limitations of the traditional architecture of Extract, Transform, Load tools, and developed a three-layer architecture based on metadata. This made the ETL process more flexible, versatile and efficient, and finally they designed and implemented a new ETL tool for the drilling data warehouse [11]. A systematic review method was proposed to identify, extract, and analyze the main proposals for modeling the conceptual ETL process for data warehouse. The main proposals were identified and compared based on the characteristics, activities, and notation of ETL processes, and the study was concluded by reflecting on the studied approaches and providing an update skeleton for future studies.

III. Feature Comparison beween PDI and SSIS

In this section, we are going to do a comparative study of the features for the two extraction tools, especially the Pentaho Data Integration and the Microsoft SQL Server Integration Services

A. *Access to data*

Table 1. Access to data

features	PDI	SSIS
Read the full table	✓	✓
Complete view of reading	✓	✓
Calling stored procedure	✓	✓
Uploading clause where/order by	✓	✓
Query	✓	✓
Query Builder	✓	✓
Reading / writing all simple and complex data types	✓	✓
Read the full table	✓	✓
CSV	✓	✓
Fixed / Limited	✓	✓
XML	✓	✓
Excel	✓	✓
Validity flat files	x	✓
Validity of XML files	✓	✓

For the access to relational data, flat files and applications of connectors, PDI and SSIS are good solutions for thesefeatures.The two tools allow the analysis of data from various sources to determine the transformations necessary to perform aggregations, data deletions, automatic corrections of errors, etc.But for the validation of the flat files, the SSIS tool is more robust in comparaison to PDI.

B. *Triggering pocess*

Table 2. Triggering process

features	PDI	SSIS
CORBA	x	✓
XML RPC	x	✓
JMC	x	x
MOMS	x	✓
Index	✓	✓
POP	✓	✓

We note for the triggering process by message, the PDI tool is not suitable for this procedure, whereas for the trigger by type of polling the two tools are robust.
Oracle is the only database that supports JMS natively in the form of Oracle Advanced Queueing. If the message receiver is not tookeen on thisJMS implementation, it is usually possible to find some sort of messaging bridge that will transform and forward messages from one JMS implementation to another.

C. *Data processing*

Table 3. Data processing

Features	PDI	SSIS
Transformation functions of dates and numbers	✓	✓
Statistical functions qualities	x	✓
Allows transcoding with a reference table	x	✓
Heterogeneous joints	x	✓
Supported modes of joint	external	✓
Management of nested queries	x	✓
Treatment options for a programming language	✓	✓
Added new transformations and business processes	✓	✓
Mapping graphics	✓	✓
Drag and Drop	✓	✓
Graphical representation of flow	✓	✓
Viewing under development data	x	✓
Impact analyses tools	✓	✓
Debugging Tools	✓	✓
Generation of technical and functional documentation	x	✓
Viewing documentation through the web	x	✓
Management of integration errors	For some steps	✓

The two tools provide a mechanism of query directly in SQL which allows to make all modes of joint and nested queries.It ispossible with SQL Server to join data from an active directory to data in a SQL Server and create a view of the joined data. For the treatment of the data, the two tools are not compatible for the transformations and calculations by default, they are recommended for the manual transformations except for the generation of technical and functional documents.

D. Advanced development and deployment/production start

Table 4.Advanced development and deployment/production start

Features	PDI	SSIS
Application Programming Interface	✓	✓
Integration of external functions	✓	✓
Crash recovery mechanism	x	x
Setting buffers / indexes / caches	✓	✓
Team Development Management	✓	✓
Versioning	x	✓
Compilation treatments	x	Yes for C#
Type into production	Windows or Unix command line	Windows command line
History visualization into production	x	x

It was found that the two tools are not compatible for the recovery mechanism on incident and for the history visualization into production,but generally they are used for the other properties of the advanced development and deployment of production setting.

E. Administration and security management

Table 5.Administration and security management

Features	PDI	SSIS
Administration Console	✓	✓
Automated log management	✓	✓
Specific log generation	x	✓
Interfacing with monitoring tools	x	✓
Integrated treatment planning tool	x	✓
Use of rights of a directory	x	x
Security type	DBMS security which contains the repository	✓
Security scenario creation	✓	✓
Security access to metadata	✓	✓
Safety manual task launch	✓	✓
Security Administration Console	✓	✓

We note that the PDI is not compatible for the generation of specific log, the interfacage with Tools of Supervision, the planning of integrated treatment and for the security of the database management system that does not contain the repository.

IV. COMPARATIVE TREATMENT TIMES

A. Test realization methodology

Test n°

Descriptive

1. Extracting data from an Excel file
2. Loading data into another Excelfile
3. The input file contains 5 typed fields:
   - COD_IND [NUMBER] (Student Code)
   - COD_NNE_IND [NUMBER] (National ID of the student)
   - DATE_NAI_IND [DATE] (Date of birth of the student)
   - LIB_NOM_PAT_IND [String] (Family name of student)
   - LIB_PR_IND [String] (Student's first name)

B. Modeling in Pentaho Data Integration (PDI) [8]. [9].

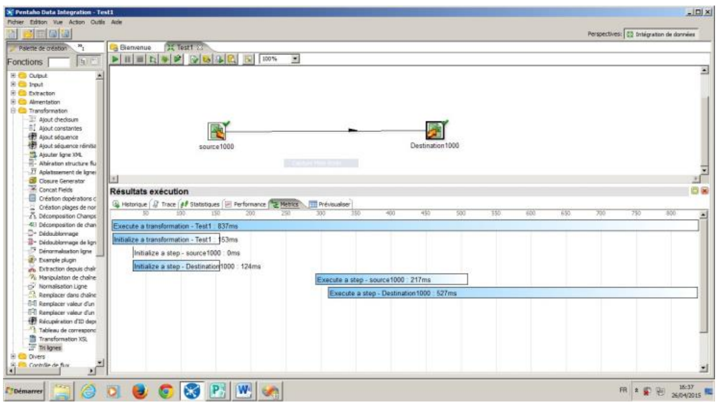

Fig. 1: Extraction of 1000 rows with PDI

C. Modeling in SQL Server Integration Services (SSIS)[10].

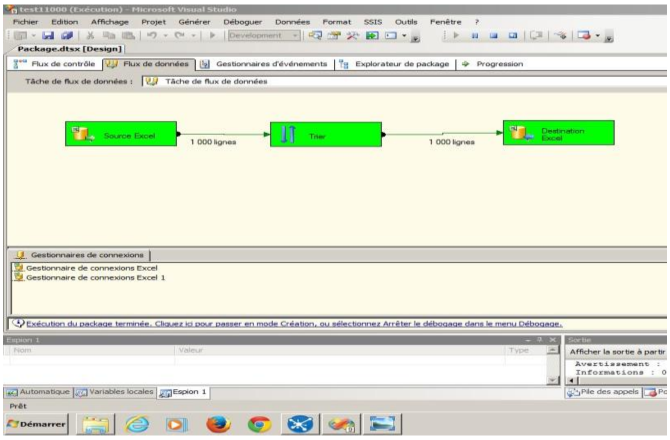

Fig. 2: Extraction of 1000 rows with SSIS

We performed the same work for 5000 and 10000 rows.

Table 6.Processing time for both tools

Number of rows	PDI	SSIS
1000	837	655
5000	1384	1014
10000	3009	1513

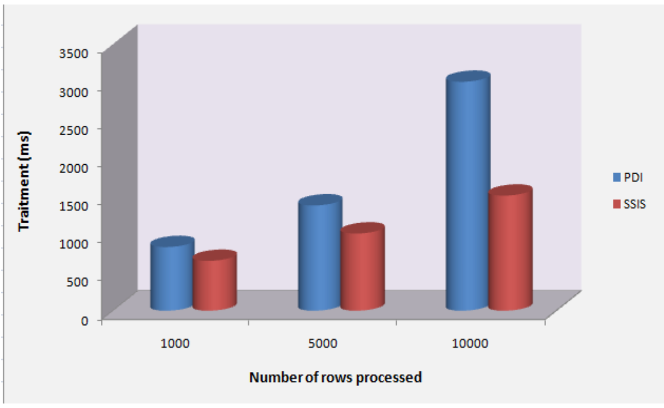

Fig3: Comparison of the results obtained for the two tools

The performance of the treatment of time is an important criterion in the choice of an ETL, but from these results we cannot prejudge the actual performance in a production environment, since time of execution variesfollowing the typology of treatments.

At the end of our comparative study, we can conclude that SSIS and PDI are two tools of ETL with their own specificities.These are real alternatives to the ETL owners as Informatica Power Centeror Oracle Warehouse Builder.These two tools offer all the features necessary for an ETL.

V. 5. Conclusion

Both SSIS and PDI are robust solutions to perform ETL in a data warehouse. SSIS emphasizes configuration over coding; however, because of the limited amount of available transformation objects, coding will be required to process complex data. SSIS's strength comes from its control flow, data flow and event driven architecture. It allows great flexibility to the developer to design the structure and flow the ETL process.On the other side, PDI includes many more options to access outside data such as a Google Analytics and several options to access Web services. It can be used on either Windows or Linux operating systems.

The choice between the SSIS ETL and PDI thus depends essentially on the typology of the project it leads.

References

[1] Inmon W, Strauss D, Neushloss G. DW 2.0 the Architecture for the next generation of Data Warehousing. Morgan Kaufman. 2007.

[2] Simitisis A, Vassiliadis P, Skiadopoulos S, Sellis T. Data Warehouse Refreshment. Data Warehouses and OLAP: Concepts, Architectures and Solutions. IRM Press. 2007: 111-134.

[3] Kimball R, Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing, Inc. 2004.

[4] Kabiri A, Wadjinny F, Chiadmi D. Towards a Framework for Conceptual Modelling of ETL Processes. Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science. Heidelberg. 2011; 241: 146-160.

[5] Vassiliadis P, Simitsis A. Extraction-Transformation-loading, (ETL) Processes in Data Warehouse Environments. In Encyclopedia of Database Technologies and Applications, Idea Group. 2005.

[6] Golfarelli M, Rizzi S. Data Warehouse Design, Modern Principles and Methodologies. McGraw Hill Companies, srl Publishing Group Italia. 2009.

[7] Jovanovic P, Theodorou V, Abelló A, Nakuçi E. Data generator for evaluating ETL process Quality (Science direct). In Press. 2016.

[8] Thoo E, Friedman T, Beyer Mark A. Magic Quadrant for Data Integration Tools. Gartner RAS Core Research Note G. 2013.

[9] Pall AS, Khaira JS. A comparative review of extraction, transformation and loading tools. Database Systems Journal BOARD. 2013; 4(2): 42-51.

[10] Simitsis A, Vassiliadis P, Sellis T. State space optimization of ETL workflow. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(10): 1404-1419.

[11] Jian L, Bihua X. ETL tool research and implementation based on drilling data warehouse. Seventh International Conference on Fuzzy Systems and Knowledge Discovery. 2010; 6: 2567-2569.