Efficiency Comparaison and Evaluation between Two ETL Extraction Tools

Abdellah Amine*, Rachid Ait Daoud, Belaid Bouikhalene
Sultan Moulay Slimane University, Faculty of Science and Technology
Po Box 523, Beni Mellal, Morocco, Phone: + 212 (0) 661748520 Fax: +212 (0) 523 481351
*Corresponding author, email: a.amine@usms.ma

Abstract

In the prospects of making an array of onboard decision support in a public university, we present a comparison between two ETL extraction tools from a production database containing student information. For the implementation we use Pentaho and Sql Server tools and we illustrate the application on the case of Sultan Moulay Slimane University in Beni Mellal, Morocco.

Keywords*: pentaho, sql server, data warehouse, business intelligence*

1. Introduction

Data Warehouse (DWs) is defined as “subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process” [1]. Data warehousing emphasizes the capture of data from diverse sources for useful analysis.

The heart of DWs is the Extraction-Transformation-Loading (ETL) process. ETL is a process, which is used to extract data from various sources, transform that data to desired state by cleaning it and loading it to a target database. The result of this is used to create reports and analyze it. ETL consume up to 70% of ressources [2-5].

In the professional field before choosing an ETL tool, it is mainly to proofs of concept. However it is almost impossible to do proofs of concept of all tools ETL available on the market. Then it proceeds to a preliminary choice in such a way as to keep two ETL suites to test. This preliminary choice is in general based on criteria which are summarized as follows: the category of the tool, the cost, the nature of the project ETL and the proof of concepts.

This whitepaper will only cover the use of two ETL tools (Microsoft SQL Server Integration Services SSIS and Pentaho Kettle) [6] based on the generalized criteria for selection of better tool.

2. Related work

Various approaches for designing, optimizing, and automating ETL processes have been proposed in the last few years. In this section we briefly review these different approaches [7]. Some of the leading data integration providers are: IBM, Informatica, Oracle, Microsoft, Talend, Pentaho, Information Builders, etc.

There are many research papers that provide a comparative analysis of the market leading ETL tools, such as [8-9]. They analyze in depth the functionalities and capabilities that these tools offer, and from that can be derived that all of them provide support to all features that define data integration tools

Different varieties of approaches for the integration of ETL tool in data warehouse have been proposed. Shaker H. Ali ElSappagh tries to navigate through the efforts done to conceptualize abbreviations for ETL, DW, DM, OLAP, ion-line analytical processing.A data warehouse gives a set of numeric values that are based on a set of input values in the form dimensions [10]. Li, Jain, conquered the weak points of traditional Extract, Transform and Load tool’s architecture and processed a three layers architecture based on metadata. That built ETL process more flexible, multipurpose 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

Received February 2, 2016; Revised May 25, 2016; Accepted June 10, 2016

identify, extract and analyze the main proposals on modeling conceptual ETL process for Data warehouse. The main proposals were identified and compared based on the features, activities and notation of ETL processes and concluded the study by reflecting on the approaches being studied and providing an update skeleton for future study.

2.1. MicrosoftSQL Server Integration Services: SSIS

SSIS is packaged with Microsoft SQL Server and requires a SQL Server License to use it. Microsoft also offers a full business intelligence suite. Additionally, SSIS may be used with a number of database servers through OLE and ADO.NET drivers. Microsoft does not offer the source code as part of the product, meaning the developer cannot make modifications to the product to suit the project needs. Also, there is no avenue for a developer to contribute to the future version of the product other than requesting the functionality to Microsoft [12].

2.2. Pentaho Data Integration Overview

Pentaho Data Integration (PDI), long known as the Kettle, is an open source ETL that allows to design and implement handling and data transformation.It is a comprehensive tool with advanced features such as "clustering" of ETL processing. These features are available from the open source version of PDIand are found only in commercial versions of ETLS competitors.

Pentaho Data Integration provides a graphical interface "Spoon" (based on SWT), from which you can create two types of treatment: transformations and tasks (jobs).Jobs and transformations are stored in a meta language, which can either be stored in XML format or ina database [13].

3. 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.

3.1. Access to Data

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.

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	✓	✓
ERP,Web Services	✓	✓
Validity flat files	x	✓
Validity of XML files	✓	✓
SAP	reading	✓
Cube OLAP	✓	✓
Others	LDAP	RSS, LDAP, POP

3.2. Triggering pocess

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 this JMC implementation, it is usually possible to find some sort of messaging bridge that will transform and forward messages from one JMC implementation to another.

Table 2. Triggering Process

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

3.3. 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	✓

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

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.

3.4. Advanced Development and Deployment/Production Start (Table 4)

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.

3.5. Administration and Security management

Table 5. Administration and Ssecurity 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	✓	✓

4. Comparative Treatment Times

4.1. Tests Realization Methodology

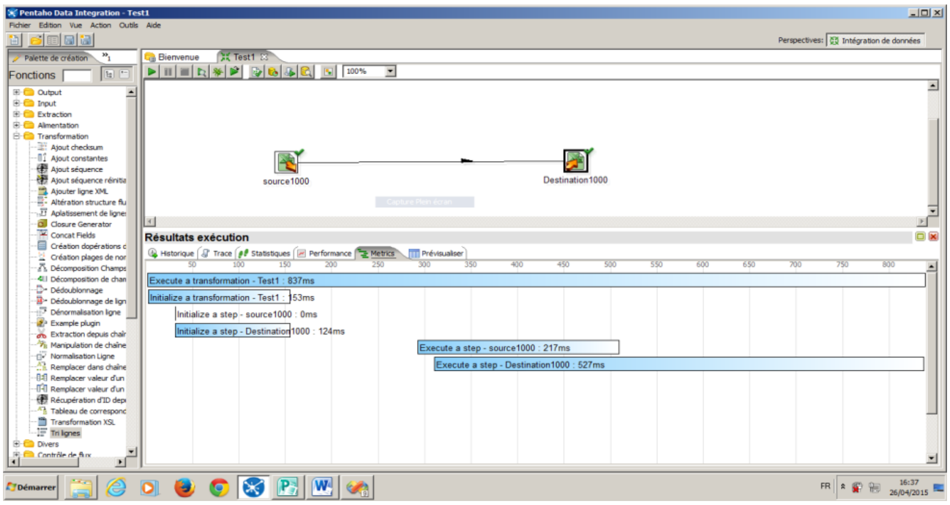
Test n°1

Descriptive

1. Extracting data from an Excel file
2. Loading data into another Excel file

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)

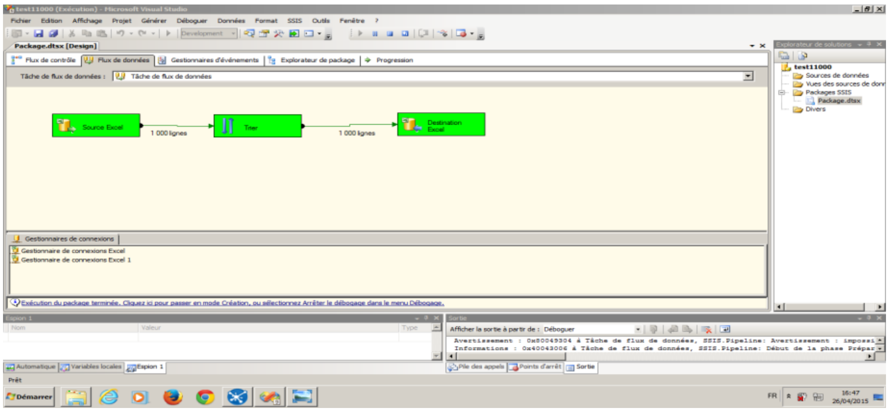
4.2. Modeling in Pentaho Data Integration (PDI)



```
Execute a transformation - Test1 : 837ms
Initialize a transformation - Test1 : 153ms
Initialize a step - source1000 : 0ms
Initialize a step - Destination1000 : 124ms
Execute a step - source1000 : 217ms
Execute a step - Destination1000 : 527ms
```

Figure1. Extraction of 1000 Rows with PDI

4.3. Modeling in SQL Server Integration Services (SSIS)



Exécution du package terminée. Cliquez ici pour passer en mode Création, ou sélectionnez Arrêter le débogage dans le menu Débogage.

```
Avertissement : 0x80049304 à Tâche de flux de données, SSIS.Pipeline: Avertissement : impossi
Informations : 0x40043006 à Tâche de flux de données, SSIS.Pipeline: Début de la phase Prépar
```

Figure 2. Extraction of 1000 Rows with SSIS

We performed the same work for 5000 and 10000 rows.

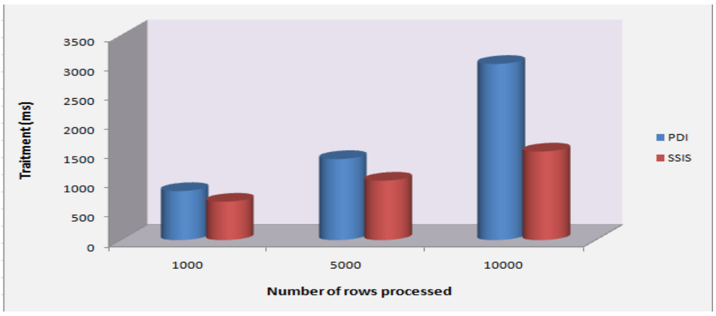


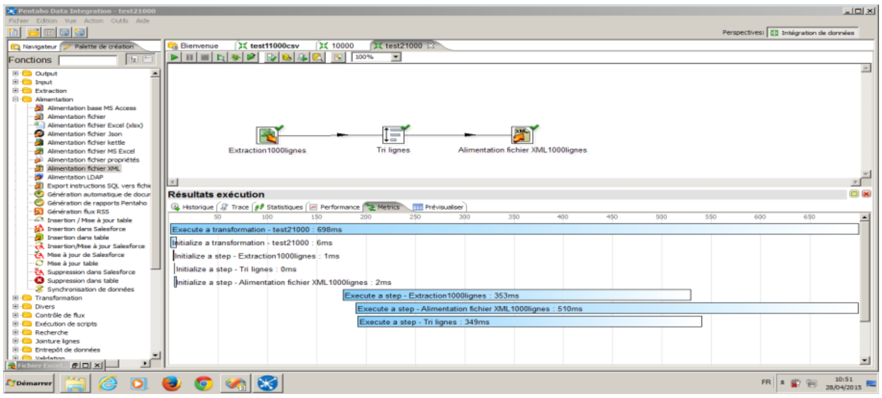
Figure 3. Comparison of the Results Obtained for the Two Tools

Test n°2

Descriptive

1. Extracting data from an Excel file
2. Loading data into an XML file

4.4. Modeling in Pentaho Data Integration (PDI)

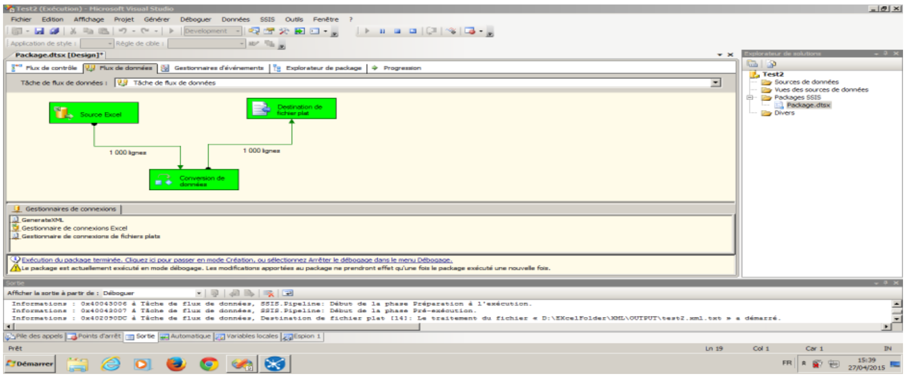

Pentaho Data Integration - test21000
Fichier Edition Vue Action Outils Aide
Perspectives: Intégration de données
Navigateur
Palette de création
Fonctions
Output
Input
Extraction
Alimentation
Alimentation base MS Access
Alimentation fichier
Alimentation fichier Excel (xlsx)
Alimentation fichier Json
Alimentation fichier kettle
Alimentation fichier MS Excel
Alimentation fichier propriétés
Alimentation fichier XML
Alimentation LDAP
Export instructions SQL vers fichie
Génération automatique de docur
Génération de rapports Pentaho
Génération flux RSS
Insertion / Mise à jour table
Insertion dans Salesforce
Insertion dans table
Insertion/Mise à jour Salesforce
Mise à jour de Salesforce
Mise à jour table
Suppression dans Salesforce
Suppression dans table
Synchronisation de données
Transformation
Divers
Contrôle de flux
Exécution de scripts
Recherche
Jointure lignes
Entrepôt de données
Validation
Bienvenue
test11000csv
10000
test21000
100%
Extraction1000lignes
Tri lignes
Alimentation fichier XML1000lignes
Résultats exécution
Historique
Trace
Statistiques
Performance
Metrics
Prévisualiser
50
100
150
200
250
300
350
400
450
500
550
600
650
Démarrer
FR
10:51
28/04/2015

```
Execute a transformation - test21000 : 698ms
Initialize a transformation - test21000 : 6ms
Initialize a step - Extraction1000lignes : 1ms
Initialize a step - Tri lignes : 0ms
Initialize a step - Alimentation fichier XML1000lignes : 2ms
Execute a step - Extraction1000lignes : 353ms
Execute a step - Alimentation fichier XML1000lignes : 510ms
Execute a step - Tri lignes : 349ms
```

Figure 4. Extraction of 1000 Rows with PDI

We performed the same work for 5000 and 10000 rows.

4.5. Modeling in SQL Server Integration Services (SSIS)



Tâche de flux de données : Tâche de flux de données

GenerateXML
Gestionnaire de connexions Excel
Gestionnaire de connexions de fichiers plats

Exécution du package terminée. Cliquez ici pour passer en mode Création, ou sélectionnez Arrêter le débogage dans le menu Débogage.
Le package est actuellement exécuté en mode débogage. Les modifications apportées au package ne prendront effet qu'une fois le package exécuté une nouvelle fois.

Afficher la sortie à partir de : Déboguer

```
Informations : 0x40043006 à Tâche de flux de données, SSIS.Pipeline: Début de la phase Préparation à l'exécution.
Informations : 0x40043007 à Tâche de flux de données, SSIS.Pipeline: Début de la phase Pré-exécution.
Informations : 0x402090DC à Tâche de flux de données, Destination de fichier plat [14]: Le traitement du fichier « D:\EXcelFolder\XML\OUTPUT\test2.xml.txt » a démarré.
```

Figure 5. Extraction of 1000 Rows with SSIS

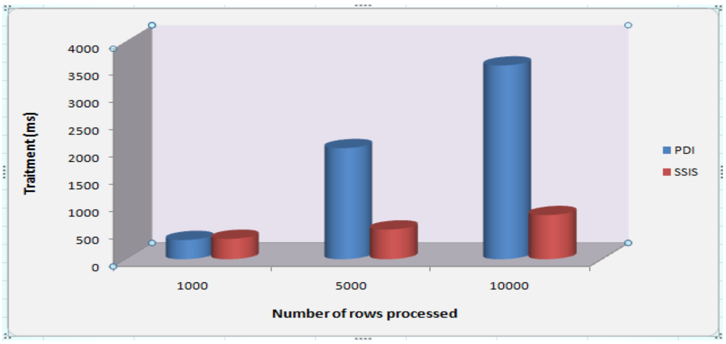


Figure 6. Comparison of the Results Obtained for the Two Tools

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.

4.6. Achievement of an Example of Extraction of the Data with the SSIS Tool

The data are obtained from the application known as APOGEE (Application for the organization and the management of students and teachers) of Beni Mellal University.

4.6.1. The Implementation of the Data Warehouse

For the realization of our warehouse of data, we have used the tools of the SQL server. As data, it has used the following tables: 1. Student 2. Pathway 3. Region 4.Institution 5. Baccalaureate 6. Time

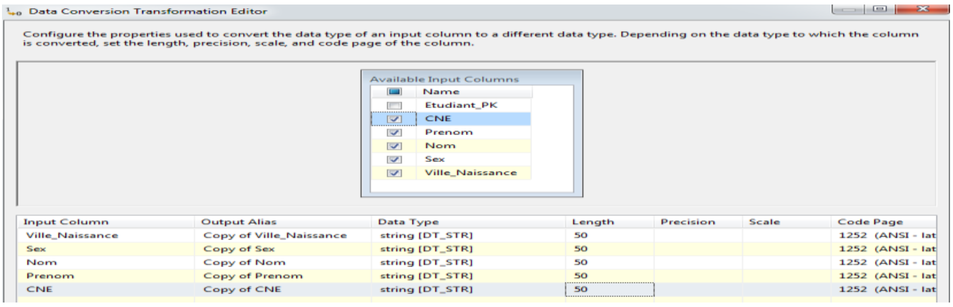
Step1: **Loading the data using the SQL Server Integration Services (SSIS)**

In this Part a package will be automatically created by default;it is named PACKAGE.dtsx. Afterthis creation we will achieve a connection between our package of ETL and the servers of sources and destination.

Step2: **ETL**

Before beginning the steps of the ETL, it was first necessary to create an empty database and a connection between the Microsoft Visual Studio 2010 software (Environment business intelligence of SQL Server 2012) and servers of sources and destination.

We will illustrate the steps for the data from the Student Table, and the same thing will bedone for the other tables.And then we are going to view the columns of thesource file; after that we will transform the tables by converting the data and we will achieve a mapping between the converted data and those of the destination.The figures above show these steps.


Data Conversion Transformation Editor

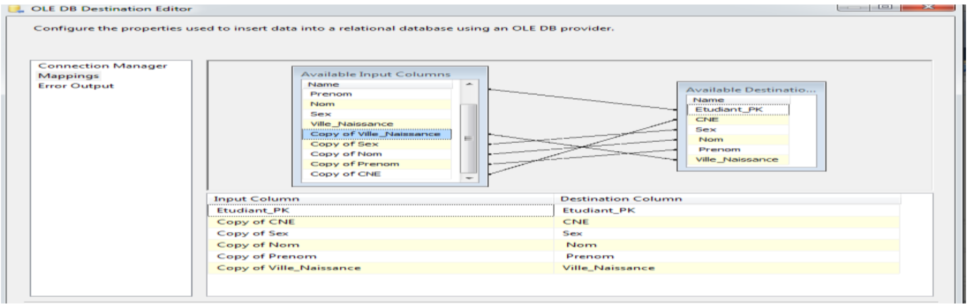
Configure the properties used to convert the data type of an input column to a different data type. Depending on the data type to which the column is converted, set the length, precision, scale, and code page of the column.

Available Input Columns

	Name
□	Etudiant_PK
☑	CNE
☑	Prenom
☑	Nom
☑	Sex
☑	Ville_Naissance

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
Ville_Naissance	Copy of Ville_Naissance	string [DT_STR]	50			1252 (ANSI - lat
Sex	Copy of Sex	string [DT_STR]	50			1252 (ANSI - lat
Nom	Copy of Nom	string [DT_STR]	50			1252 (ANSI - lat
Prenom	Copy of Prenom	string [DT_STR]	50			1252 (ANSI - lat
CNE	Copy of CNE	string [DT_STR]	50			1252 (ANSI - lat

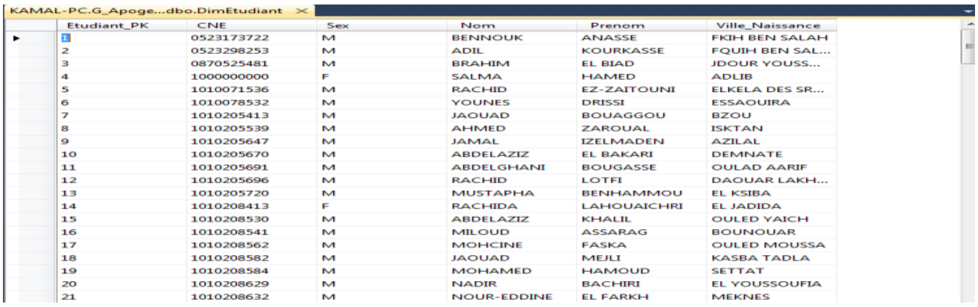
Figure 7. Data Conversion


OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Input Column	Destination Column
Etudiant_PK	Etudiant_PK
Copy of CNE	CNE
Copy of Sex	Sex
Copy of Nom	Nom
Copy of Prenom	Prenom
Copy of Ville_Naissance	Ville_Naissance

Figure 8. The Mapping Between the Data


KAMAL-PC.G_Apoge...dbo.DimEtudiant

Etudiant_PK	CNE	Sex	Nom	Prenom	Ville_Naissance
1	0523173722	M	BENNOUK	ANASSE	FKIH BEN SALAH
2	0523298253	M	ADIL	KOURKASSE	FQUIH BEN SAL...
3	0870525481	M	BRAHIM	EL BIAD	JDOUR YOUSS...
4	1000000000	F	SALMA	HAMED	ADLIB
5	1010071536	M	RACHID	EZ-ZAITOUNI	ELKELA DES SR...
6	1010078532	M	YOUNES	DRISSI	ESSAOUIRA
7	1010205413	M	JAOUAD	BOUAGGOU	BZOU
8	1010205539	M	AHMED	ZAROUAL	ISKTAN
9	1010205647	M	JAMAL	IZELMADEN	AZILAL
10	1010205670	M	ABDELAZIZ	EL BAKARI	DEMNATE
11	1010205691	M	ABDELGHANI	BOUGASSE	OULAD AARIF
12	1010205696	M	RACHID	LOTFI	DAOUAR LAKH...
13	1010205720	M	MUSTAPHA	BENHAMMOU	EL KSIBA
14	1010208413	F	RACHIDA	LAHOUAICHRI	EL JADIDA
15	1010208530	M	ABDELAZIZ	KHALIL	OULED YAICH
16	1010208541	M	MILOUD	ASSARAG	BOUNOUAR
17	1010208562	M	MOHCINE	FASKA	OULED MOUSSA
18	1010208582	M	JAOUAD	MEJLI	KASBA TADLA
19	1010208584	M	MOHAMED	HAMOUD	SETTAT
20	1010208629	M	NADIR	BACHIRI	EL YOUSSOUFIA
21	1010208632	M	NOUR-EDDINE	EL FARKH	MEKNES

Figure 9. Loading Data

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.

ETL tools are designed and used to save time and cost when a new data mart or data warehouse is developed, we find that Microsoft SQL Server Integration Services (SSIS) are mostly satisfied the needs of large organizations, as it can manage the large database. In case of freeware or open sources ETL tools, Pentaho Data Integration (Kettle) is mostly used for small enterprises, as it limits the speed and having limited debugging facility.

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.

[12] Coutaud R, Jehl F, Harel P. *Editors*. SQL Server 2012 Integration Services (SSIS) Broché. France: ENI Édition. 2012.

[13] Carina MR. Pentaho Data Integration Beginner's Guide. Buenos Aires: Packt Publishing Limited. 2013.