Open source BI tools for SMEs: Comparative analysis of Pentaho CE and Knowage CE stacks

Igor Jakimovski
Faculty of Computer Science and Engineering
Skopje, Republic of North Macedonia
jakimovski.igor.1@students.finki.ukim.mk

***Abstract*—It is well known that Business Intelligence (BI) plays an important role in helping companies in dialing with competition and gain better success in business as many published successful world companies stories show. But how many companies in the domestic market have implemented it? BI is a discipline that transforms data into information that can be used to accelerate the organization's business success. In other words, BI uses data with a bigger historical diapason for comparison with the present ones, for responsible companies people, be able to draw conclusions on what actions to take in the near future or which steps to choose for making better business results or keeping business on the wanted level. This paper makes a proposal while evaluating of two different concrete implemented BI stacks, that could be implemented for small and medium-sized organizations. Evaluation is based on the experience while implementation, testing, and of course manuals of the developers as of the community support.**

Keywords—BI, DW, ETL, PDI, TOS, OLAP

I. Introduction

The importance of information is well known. But how do to collect really important information that can help do better business? Finally, how can do use that information at all? According to the parts of [1] and [2], BI covers all processes beginning with collecting data from the sources to the final visual presentation of the information, that can be used for decision support or others.

Although in more developed business environments around the world, BI has been implemented for years, and the positive experiences can be read, our domestic business environment does not emphasize the BI at the level it deserves to be, nor much has been researched for technology comparisons.

The purpose of this paper is as a manual that is based on personal experience to present with and know the domestic small and medium-sized companies and their Managements through IT sectors with the BI, and trigger them to try to implement it. For this purpose, two free-of-charge and enough serious BI stacks (their Community Edition (CE) versions can be used in production very well), were used. Let me repeat, the evaluation was made on Pentaho CE and Knowage CE (with Talent Open Studio for ETL) with all processes from ETL to visualization. As in [3] and following the processes, four different Data Warehouses were created, two using the PDI tool with loading data in MySQL and PostgreSQL databases and another two using the TOS tool also loading the data in MySQL and PostgreSQL databases. A comparative analysis is made for each performed BI process which highlights the more specific features or shortcomings of the respective tools if any. As a final result, the paper proposes the best combinations of tools, from several aspects, that could help the future BI implementers in selection.

The paper begins with a brief theoretical description of the topic, after which goes through all the BI processes with theoretical and practical explanation, to move to a comparative analysis of technologies, after which it ends with a concrete conclusion.

II. Implementation

Before starting the implementation, first have to learn which kind of data the users like to have available in the Data Warehouse DW, or the users' requirements as in [2][5]. As a trading company [3], had financial and commercial data available, but generally would be very good the DW contains manufacturing data too. Next, to discover all tables and find what data from the existing tables would fit in the DW tables, a very precise analysis of the transaction operational database should be made. If there is a need for data from other sources, should be found and analyzed too as well. First, the solution to the dimension tables should be found. For example, every transaction operational database of an organization contains some kinds of tables with the data for articles, clients, and so on. Every one of those tables fits in some dimension would be needed should be analyzed, and minimum possible fields to be chosen. A similar should be done for the fact tables later too. In this case [3], dimension tables for articles, clients (including suppliers in the same dimension), accounting codes, company organization units, document type, and calendar as a date dimension. Once the tables and the fields had been chosen, the ETL process can start organizing.

A. Architecture

As in [3], the architecture of implemented BI system contains two layers, the bottom one that houses the DW connected to the data sources through the ETL process. The upper layer houses the user applications on the upper part and OLAP servers on the lower one, connected with the DW databases. DW is practically a database, multidimensional or here relational, optimized for reading, aggregating, and searching big amount of data that is updated regularly in certain terms, usually when the sources from which it is charged are free and which is designed to store a long-range of historical data. As in [1] the tables of the DW database are connected in specific schematic structures, from which the most known are the star, snowflakes, and constellation schemas. As the name says, the star schema looks like a star, with a fact table in the middle, and dimension tables on the angles, connected directly to the fact table forming composite primary keys using their foreign. The snowflake schema looks like the star one but allows connections between dimensions or has a higher normalization level, and the constellations schema allows more than one fact table and different connections. The point of a DW database is faster return results of the queries, so the database should be as less normalized. As for the tables, the fact ones contain the measures and the dimension tables contain attributes that describe the measures.

B. ETL process

As described in [1], ETL – extract, transform, loading means that we have to choose the data from the sources, extract them on the stage area where will do cleansing and transformation before loading the ready data in the DW database. The ETL process is very important and statistically can spend 70-80% of the time needed to develop a BI. After the tables and the fields of the tables that would be used for the dimension tables are found, analyzing table by table to search for dirty data (refers to data that contains erroneous information) could start. The data can be dirty because for different reasons, writing data in wrong fields, wrong names, dates formats and etc, but all those have to be corrected. How this is done in practice? As in [3], PDI and TOS too, contain different steps for different actions. The job in TOS or transformation in PDI for ETL data in dimension tables (except calendar one) starts in the same way, with reading (extract) data from the original tables from the basic data source. As a row by row is read, the row flows to the first of the cleansing steps, then the second, third, then some other purposed steps, and so on until finished all programmed for cleansing and transforming the data. The last step is loading in the DW database, in the given dimension table. As in [4], specific for the dimension tables is that they should contain information on the changes of the data that could occur through time and the way to handle them. That is the part managed by the slowly changing dimension (SCDI) named principle that offers seven ways how we can handle the changes, starting from SCDI 0 as ignoring the changing, SCDI 1 as overwriting the changing with deleting the history, to keeping history details in new created database's tables. The most used is SCDI 2 which keeps history with the versions and dates of change added in newly created columns of the existing tables. So for the final loading of the data in the dimension table, the SCDI loading step with set the type for all columns from rows would be used.

According to [5] and [6], both PDI and TOS handle all steps from reading to loading very well, in their own way. They have different steps for different actions according to their working principles. The creation of the end tables, in PDI is going by pressing a button while setting the transformation while TOS does it automatically. For reading the sources, both tools have steps for it, for working with strings, PDI has separate steps while TOS mainly uses the tMAP step for different actions including the mentioned one. Both tools allow using regular expressions but with their own syntax, also java code expressions. The steps in the dimension in both tools are starting with reading the original tables from the source databases listed by the years we set, sorted from the oldest ascending, followed by one or more cleansing steps as strings operations depending on a cleansing procedure, ending with SCDI loading steps. The exception of the normal dimensions is the calendar one that has to be created from the scratch setting the start date as the attribute for both tools, starting with generate rows step, then adding date sequence and calculate dates steps followed by select values and calculate additional fields steps, ending with the loading of the record as the last step in PDI version. The TOS version starts with the tGenerateRows step followed by the tMap step (that is doing all other PDI steps job), ending with the step that loads the data in the table. That means this transformation or job is not dependent on source data. In that case, the loading step should not be SCDI as not depends on the changes, and the normal loading step works very well in TOS too. Once the dimension tables loading has finished, the same with the facts tables should be done too. The principle is the same with the extra steps for joining the dimensions with the facts, calculating final measures, and the last step which is normal loading in the database table. The source tables should be searched between the ones containing input and output transactions as for receipt and sale of goods for the commercial part, same for input and output of accounting document for the financial part. In this case, the commercial part transformations in PDI or jobs in TOS begin with the reading of two source tables which continue in two separate streams, then a few filters on both streams followed by joining of the two streams, then again filtering after which coming joining with the dimensions, then calculating measurements step and at the end loading in the table step. As to the financial fact tables transformations or jobs, the same starts with a reading source table followed by four filters according to dates and accountant codes, then dimension joining followed by calculating measurements step and at last the step for loading data in the table.

According the [3], the only problem that could be found while executing was a bad speed performance for the SCDI loading step using TOS. That step loads so slow that is not worth using. For comparison, PDI executes all dimensions transformations connected serial for seven to eight minutes while TOS executes only the smallest one for a similar time. According to the explanation, the recommendation is to do the complete ETL process with the PDI tool, or if any want to do a mixture can do with the TOS tool only the ETL process for the facts.

C. Data Warehouse

As the transformations and the jobs have been created and executed in the order dimensions then facts, four loaded databases for four DW have been created. The last step of the DW part is the scheduling of updating new data. Depending on the usual data amount, the ETL process of only fiscal year data can be scheduled to run in some therm out of the working hours once a week or oftener, as in [7]. With all this previous, the DW part has finished and the next BI parts can be worked on.

D. OLAP

According to [1], DWs that are low-level in BI architecture are complemented by a set of top-level technologies that provide specialized functionality for different BI scenarios through user applications and tools that lie on top of them. OLAP is just one of such technologies which is the background of many BI applications. The OLAP server is a core component of OLAP technology that sits between the DW and front-end client applications and provides functionality for aggregating, calculating, and analyzing historical data stored in the DW. It optimizes the questionnaires it downloads from client applications, redirects them to the DW database, and sends the data received from the DW to the same client applications. As a result, the end-users receive the required information presented in an easily recognizable form in seconds.

One of the most popular OLAP servers is Mondrian which supports Relational OLAP. OLAP offers a wide range of operations.

Drill down operation refers to dis-aggregation, viewing data at the level of magnified detail. The roll-up operation performs aggregation of the data cube by climbing the hierarchical tree, Slice and Dice operations are performed so that a large segment of data is cut into smaller pieces and this process is repeated until the right level of detail for analysis is reached. Pivot or rotation is a visualization operation that rotates the data axes to provide an alternative presentation of the data. It may contain replacing rows and columns or moving one of the row dimensions to the column dimensions. As in [1], The most common OLAP architectures encountered are ROLAP - Relational OLAP, applications based on relational databases, MOLAP - Multidimensional OLAP, applications based on multidimensional databases, and HOLAP - Hybrid OLAP, applications use both relational and multidimensional techniques.

For practical implementation, the first step is creating an OLAP schema [8] that represents an XML file containing a description of the OLAP environment with the dimensions and facts the cube contains. The file can be written in some text editor or through an application built for it as SchemaWorkbench SW is [8], developed to build Mondrian schemas. In the implementation, the SW application has been used to create the schemas and a text editor to create virtual cubes in the schemas as the application does not do it. The SW application allows checking the schemas using MDX statements writing and executing through the executor the application contains. Once the OLAP schema has been created, should be loaded into the Mondrian servers. In the Pentaho CE environment that operation is going through the Pentaho console choosing "add a new source" and importing the existing schema in the repository, while in the Knowage environment the schema should be uploaded through the "Mondrian schemas catalog" sub-menu from the administration's menu of the server. Now, that OLAP schemas have been uploaded, the transformations of the data into the information using visualization tools can be started

E. Visualization

As in [2], the upper end of the BI architecture is the part where business users use client tools for business intelligence. The responsibility of these tools is to perform clear transmission and visualization of data from data sources through DW to end-users. This level is sometimes called the presentation of knowledge because it is responsible not only for presenting the data but also for presenting it in a format that is easy and understandable to consume. Typical BI implementation does not use a single type of presentation software but includes more specific tools for different purposes. For example, if the CEO of a company prefers a high-level business overview for which he sees data in high visual formats, such as dashboards, a well-trained financial analyst may prefer a powerful spreadsheet or spreadsheet-like format instead of simple graphical representations. This is an example of why most BI software implementations provide a mixed set of tools tailored not only to specific functionalities but also to user types. Some of the more used and standard BI tools are spreadsheets, querying and reporting software, OLAP analyzers, digital boards and scorecards, business performance management, ETL tools, and others.

According to [9] and [10], using both technologies, the next visualization documents have been created:

Reports - two tabular reports, the first one shows the top buyers for a chosen year, and the second gives an analytic view of the buyers also for a chosen year, and two reports with charts, the first one shows the perceptual sales for all cities for a chosen year presented with a pie chart, and the second sales per cities presented with bar chart also for a chosen year. Should be mentioned that the created DW contains 20 years of history and the reports are parameterized with the year parameter that has to be chosen when the report is run. With that, the reports have some kind of interactivity.

Dashboards - two types static and interactive. The statics dashboards are for use in management (showing present and historic financial data for comparison) and for use in commerce (showing buy/sale data). Both contain different elements like charts, tables, and maps. The interactive dashboards are for use in commerce and contain elements like charts, tables, and Pivot tables.

OLAP analytic documents - for use in the commerce department and show Pivot tables and charts (in Pentaho CE).

The procedure of creating is very long for this paper so will not be explained, but the conclusions of the creation and use will be summarized in comparative analyses of the stacks.

III. Comparative Analyses of the BI stacks

A. *Functionalities of the Business Intelligence stacks*

As in [11], Hitachi Ventura as the owner of Pentaho, as time goes by mentions the CE version less and less. Unlike the Pentaho EE stack, which automatically integrates all modern BI functionalities, with the installation of the Pentaho CE stack you get only: Pentaho BA Platform - Server as a base software part that houses content created on the server itself and includes functions for managing security, executing reports, displaying whiteboards, bursting reports, scripted business rules, OLAP analysis, and scheduling. Additional functionalities are upgraded through commercial plug-ins from the Pentaho market. As standalone software tools that can be installed and used for free are: Pentaho Data Integration PDI with Pentaho for Big Data as a plugin, Pentaho Report Designer PRD, Pentaho Data Mining Weka, Pentaho Metadata Editor PME, Pentaho Aggregate Designer PAD, Pentaho Schema Workbench PSW and Pentaho Design Studio PDS, as from Pentaho market plugins which are numerous, the relevant ones for the case are: CTools with Community Dashboard Editor CDE, Community Dashboard Framework CDF, and Community Graphic Generator CGG. Ivy CDE connectors complete as CTools plugins, Pivot4J Analytic, Datafor Visualizer, and BTable.

With all summarised [3], Pentaho CE has the following functionalities: ETL - with PDI, Reports - creating reports with PRD, publishing on Pentaho server and report viewer from the same, Dashboards - Creating with CDE and Datafor Visualizer and display on Pentaho Server, OLAP Analytics - built-in Mondrian Server with Pivot4J Analytic, PAD, PSW, Datafor Visualizer, Metadata - Creating Business Models with PME and Predictive Analytics - with Weka.

Summary: It does not have all the functionalities that Knowage CE has but what it has is enough to use in production. The main components are updated regularly with new versions, the auxiliary ones are less frequent or are stalled, but all our work is based on the basic ones.

According to [10], Knowage, on the other hand, has different principles for limiting CE from EE versions. Namely, it has all the necessary modules but there is a limitation of functionalities in them except ETL which we work with Talend and has the ability to physically add as an external module in the stack. Otherwise, the architecture of the Knowage stack is presented through modules in which all the functions of the stack are arranged and which should be grouped to obtain the desired functionalities, which are called sub-products and with which specific actions are performed. During the installation, can be chosen which modules to install, but ER (Enterprise Reporting) and SI (Smart Intelligence) are considered as initial and can be used alone or in a combined way, sharing all the common features. The SI module can be enriched with four different plugins such as SD (Smart Data), LI (Location Intelligence), PM (Performance Management), and CA (Custom Analytics) which provide additional features. As in [3], functionalities of the Knowage CE stack here according to the modules are: Knowage CORE - ETL with Talend, the module that contains the functions of menu management, document search, profiling and security, drilling everywhere (for all documents), and other very important for server administration. Some of the essentials are also missing such as map navigation, tracking and monitoring, repository management, line data connection, and others. As second is Knowage ER - ER reporting tool which is part of the module of the same name, and is part of the stack with the stand-alone tool Knowage Report Designer KRD which is used for creating reports, publishing on the Knowage server, and so on. As the third one is the Knowage SI - SI product which is part of the module of the same name and which covers: 1. Dashboards - the part for creating interactive Cockpits, 2. OLAP analytics, 3. Metadata - creating business models, 4. As a plugin of this module should mention LI which serves to link business data with spatial or geographical information, data used in analytical documents. Not available functions: advanced graphical editing with pixel options, Creating a data set from federated, template designer for OLAP documents, wizard for creating OLAP analysis template, etc. The fourth one is Knowage PM - a plugin that enables performance management work. Unavailable features: Scorecards and Alarms. As the last one is Knowage CA - a plugin that enables Predictive Analytics with what-if scenarios. Unavailable features: Pre-built analytics features, Pre-built R / Python advanced visualizations.

Summary: It has more functionality than Pentaho CE but lacks features. What it offers is still enough to function in production. Updating versions is done regularly.

B. *Scalability of views*

Pentaho CE – according to the tools [9], PRD and CTools (CDE Dashboards) is known for the so-called pixel editing of their components. Pentaho OLAP analyzers offer template/background editing - stylish according to the tools themselves, that is enough. The first two offer a large amount for the selection of components such as spreadsheets, charts, other elements, and the third one small number of charts. Regarding data sources, PRD and CDE Dashboards offer the use of a lot of ones, based on SQL and MDX statements, using GUI too. Pentaho OLAP analyzers use MDX statements or drag and drop of cubes and dimensions using GUI. PRD make export in all standard export documents, CDE Dashboards only on the basis of individual components in standard forms, and Pentaho OLAP analyzers in xls, xlsx, and pdf. Automate filtering offer only CDE Dashboards for the charts in the case of series and labels. Automatic drilling is not available for all.

Knowage CE – as for the tools [10], KRD offers template editing - pixel according to certain rules, The Cockpits according to the capabilities of the template designer himself, and OLAP Analyzers, manually making a template in XML format and editing options as such (if any) without giving an example in the tutorial. The first and second offer a large number of components such as spreadsheets, charts, cross-tabulations, and others even with quite big restrictions for the second one. The third one offers only Pivot table. For the data sources, KRD and Cockpits offer sufficient, with SQL statements as well as sub-statements directly or with SQL Query Designer, Business Model-Based Knowage Metadata (QbE) Statements, Dynamic Statement Styling. The OLAP Analyzers accepts only MDX statements manually entered in the template. For an export preview, the first one offers export in all standard documents, the second allows export of the whole cockpit in pdf and separate components in Excel spreadsheets, and the third one only classic print can also be in pdf. Automating filtering offers the Cockpit only for the charts with series and labels same as Automatic drilling - only for some charts with series and labels.

C. *Smart visualization*

As in [3], Pentaho CE has two-way interactivity with all tools except the standalone OLAP analyzer which has a one-way (has two-way included in the CDE Dashboard) with settings using script encoding.

As in [3], Knowage CE has two-way for all its components with setups using GUI.

D. Integration with Existing Transaction Processing Systems

As in [13], Pentaho CE offers direct embedding / merging with existing Web applications - the various components of the BI Platform can be combined to be added to an external application with more or less optional functionality, which is up to 100% of the platform's capabilities. The platform can also be expanded with its own components and services.

As in [10], Knowage CE offers Direct embedding/ integration with existing Web applications - a Web services called the BI Service and the BI API, which are both ways of communicating with the Knowage delivery layer, allow stack components to communicate with external applications or integrate with custom business applications with or without a graphical user interface.

E. Possibility to use mobile devices

As in [9], Pentaho CE - While the EE version has a module for the mobile version, the CDE component offers the size of window panels for viewing on mobile devices depending on the size of the devices, which means the size of the elements is adjustable.

As in [10], Knowage CE - With the help of BI Mobile also one of the ways to communicate with the delivery layer and thanks to the interaction between the Knowage server and the client interface remotely, reports, cockpits, or other useful documents can be accessed and displayed on mobile devices.

F. Security, managed ability to distribute documents by users

According to [9], Pentaho CE allows the management of user authentication and authorization types by the role as well as permissions to access folders and documents in the repository. Authentication types can be standard on Pentaho and LDAP which are managed directly from the Pentaho User Console (PUC), CAS for which settings must be added directly to the security files in the Pentaho settings folder as well as a username and password from another application that is sent with a URL when called. A user with administrator privileges can manage existing and create new roles and users directly from the PUC as well as set up access to documents and folders directly from the repository browser.

As in [10], Knowage CE also allows for the management of types of authentication, authorization of users by roles and permissions to access folders and documents in the repository as well as the use of analytical drivers and value lists. Types of authentication can be standard on Knowage, via CAS, LDAP, Google Sign-In, Azure Sign-In, and with the help of Javascript code authentication through a user application, but the authorization must be pre-set by the administrator and used only for login.

G. Ease of use of the tools

As in [3], Pentaho Reporting Design PRD - for developers, medium-difficult to create, SQL can be entered manually or you can work with a designer but all others manually and each individual action must be programmed. Pentaho CDE Dashboard - for developers, medium-difficult to create, all data sources are entered manually and each individual action needs to be programmed. Pentaho OLAP Analyzer - for developers with Pivot4J and Datafor quite easy, set upping via GUI with drag and drop facts and dimensions from the existing ones, except for interactivity that has to be programmed. Users - easy for all tools. Knowage Report Designer KRD - for developers, medium-difficult to create, SQL can be entered manually or with a designer but all else manually. Base interactivity is set via the GUI, and any more advanced action needs to be programmed. Knowage Cockpits / Dashboards - for developers, quite easy, all data sources are selected from previously entered data sets that are created manually or with GUI through a business model. Interactivity through GUI other than advanced things to be programmed. Pentaho OLAP analyzer - for developers, medium easy with inserting a hand-crafted template as the designer in CE does not work, but even if it works it is not easy to use. The rest with all the interactivity is realized through GUI. Users - easy for all tools. Infrastructure, both stacks work on the most of existing operating systems and servers.

H. Customer support

Pentaho CE - has pretty good user support from developers, users who write by experience but also a lot of third parties who have described many use cases. Knowage CE - has limited user support mainly from partially responsive developers.

IV. Conclusion

No doubt is that is useful for organizations to implement BI. If there is any decision or interest, as in [3], here are a few words as conclusions to be considered before.

Regarding databases, PostgreSQL is giving better performances in the speed and stability used in all BI processes in comparison with MySQL. On the other side, MySQL has a well know syntax and manipulation tools, but recommend PostgreSQL.

Regarding the ETL tools, as already mentioned, the TOS problem is a slow SCDI step but TOS has problems also with the MySQL database in some situations too. So the recommendation is to use PDI.

For the BI stacks, if one prefers more OLAP processing better chose Pentaho CE as offers more OLAP-like options. All his tools use MDX language even the ones that use OLAP with drag and drop in GUI environment. Of course, Pentaho CE offers SQL language too between all other options and works very well and stable.

On the other hand, if you prefer easier development choose Knowage CE. Creating Dashboards and setting the interaction is easier in Knowage CE than in Pentaho CE where all actions should be coded. Knowage CE also works very well and is stable.

For the user's support, Pentaho CE offers better.

For maturity, Knowage is a younger project than Pentaho, so should expect better things in the future.

References

[1] F.Almeida, "Concepts and Fundaments of Data Warehousing and OLAP," INESC TEC and University of Porto, 2017

[2] Cindi Hovson, “Successful BUSINESS INTELLIGENCE Secrets to Making BI a Killer App,” The McGraw-Hill Companies, 2008

[3] I. Jakimovski, “Sistemi za delovna inteligencija/razuznavanje za mali i sredni pretprijatija – analiza na arhitekturata i finkcionalnite karakteristiki,” unpublished

[4] R. Kimball and M. Ross, “Kimball dimensional modeling techniques” in The Data Warehouse Toolkit, The Kimball Group, 2013

[5] R. Holowczak, "Building ETL Transformations in Pentaho Data Integration (Kettle)." holowczak.com http://holowczak.com/building-etl-transformations-in-pentaho-data-integration-kettle/ (accessed March 12, 2022)

[6] https://help.talend.com\ (accessed March 12, 2022)

[7] V. Poe, P. Klauer and S. Brobst, “Building a Data Warehouse for Decision Support,” 2nd Edition, Prentice Hall, 1998

[8] https://mondrian.pentaho.com/ (accessed March 11, 2022)

[9] https://help.hitachivantara.com/Documentation/Pentaho/8.2 (accessed March 11, 2022)

[10] https://knowage-suite.readthedocs.io/en/7.2(accessed March 11, 2022)

[11] https://www.predictiveanalyticstoday.com/pentaho-community-edition/ (accessed March 12, 2022)

[12] https://www.knowage-suite.com/site/licensing/knowage-editions/ (accessed March 13, 2022)

[13] Integrating Pentaho Software and Content, Pentaho Corporation, USA, 2011