Analysis of Pentaho

BHAVESH REDDY MERUGUREDDY[1,*]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: bmerugur@umail.iu.edu*

Paper 1, May 3, 2017

Pentaho is a business analytics and data integration tool that provides a qualified open source-based platform to assist a variety of big data deployments. It enables different organizations to utilize their data which helps them in delivering their services efficiently with minimum risk and it can also be used for embedded analytics.

Keywords: Pentaho, Data Integration, Big Data, Community, ETL, MapReduce, SQL, Hadoop, OLAP

https://github.com/cloudmesh/sp17-i524/blob/master/paper1/S17-IR-2018/report.pdf

1. INTRODUCTION

Pentaho is a business intelligence suite that provides data mining, reporting, dashboarding and data integration capabilities. Generally, organizations tend to obtain meaningful relationships and useful information from the data present with them but data inconsistency and handling large amount of data are the factors that obstruct them from doing so. Pentaho addresses these obstacles [1]. The platform includes a wide range of tools that analyze, explore, visualize and predict data. It simplifies data blending, which is the process of combining data from multiple sources into a functioning dataset. Being an open and extensible source, Pentaho provides big data tools to extract, prepare and blend any data.

2. PENTAHO COMMUNITY

Pentaho provides two different editions, Community edition and Enterprise edition. As the name suggests, the Enterprise edition provides more packages to provide additional support. The Community edition enables the developers or users to create complex solutions for the problems pertaining to their business [2]. The Pentaho Community has a group of people and helps the users in becoming a part of them and benefitting from the open source contributions. The Community includes all users like developers, testers and managers. Generally, the Community edition platform enables the developers to sketch their design and develop a rough version of their product after which they can upgrade to Enterprise edition for final production. This shows the flexibility that the Community edition provides to innovate, find and experiment with the solution that the developers liked.

3. ARCHITECTURE

Pentaho architecture can be considered as a set of four components which are presentation layer, business intelligence platform, data and application integration and third party applications. Data can be provided to the presentation layer by reporting, analysis or process management. This data can then be accessed through a web service, portal or a browser [3]. The security and repository issues are dealt by the business intelligence platform. Data integration and third party applications are respectively, the integration layer and applications with database from various sources.

The architecture includes data layer, server layer and client layer. Data layer allows an application to connect to a data source. Server layer serves as a middle layer and several applications run on the server. Dashboards are provided to the end users by deploying them on the server along with the required reports. As mentioned above, a user console is provided that is used for security and configuration purposes. Client layer is of two forms, thin client and thick client. Thin client generally runs on a server. Analyzer and dashboard editor can be considered as the examples. Report designer and data integration come under thick client which act as a standalone.

4. PLUGINS AND APPLICATIONS

Pentaho provides an interactive console to its users. With a few clicks of the mouse, users are allowed to interact with new data models and data. The platform hides the database connections and underlying application server and provides access to various data sources [4]. It provides metadata management capabilities and a dashboard to allow the administrators set security levels, monitor servers and set user access. There are many server plugins and desktop applications provided by Pentaho.

4.1. Server applications

Business Intelligence platform is a basic Pentaho service that provides reports, displays dashboards, reports business rules and performs OLAP analysis. It generally runs in Apache Java application server and can be embedded in any other Java application server [1]. Pentaho analysis service is another Pentaho server application that is written in Java which primarily focuses on online analytical processing. It aggregates data into a memory cache by performing read operation from data sources like SQL. It comes with the Pentaho platform in both the editions. These are some of the server applications provided by Pentaho.

4.2. Desktop applications

Pentaho data mining is a desktop application that searches for patterns in data by performing knowledge analysis. Common data mining techniques such as classification, clustering, regression and visualization are employed by this application along with some machine learning algorithms. This helps the users in predicting the trends in future. Pentaho metadata editor is an application that is used as an abstraction layer from the underlying data sources and helps the users in creating business models which can be used by other applications in generating reports for the analytics. There are many more useful desktop applications like Pentaho report designer, design studio and aggregate designer.

4.3. Server plugins

Pentaho provides certain core services in the form of server plugins. Some of the important server plugins are community data access and data browser. Community data access is a Pentaho server plugin that provides a common layer on the business analytics server for an easy data access. It runs the server by providing a REST interface and gets back the results in various forms such as xml, csv or json. Community data browser is a plugin that helps R in performing analytics on the data. It does the job of supplying queries to R by using online analytical processing browser.

5. DATA INTEGRATION

Extract, transform and load (ETL) are the basic operations that act as a tool for transforming data from one database and placing in other database. These processes can be carried out in Pentaho with the help of a component called Pentaho Data Integration, which is also referred as kettle [5]. The most useful functions of Pentaho Data Integration include massive load of data into databases, data cleansing, migrating data between applications and integrating several applications. It is metadata oriented and can be used as a standalone application. Various input and output formats such as datasheets and text files are supported by Pentaho Data Integration.

The transformation process undergoes three steps, input step, transformation step and output step. In the input step, data is imported [6]. The data is then processed within Pentaho Data Integration and the transformed data is given out in the output step. All these steps are carried out in parallel. The throughput of transformation process is restricted to speed of the step which is slowest. The slowest step is often referred as bottleneck. To improve the performance of transformation process, two steps are run in a loop which are, identification of the bottleneck and continued improvement of bottleneck until it is no longer a bottleneck. Bottlenecks are eliminated by data conversion logic and character code-page conversion. For encoding a set of characters, a character set is used along with a set of control characters. Code page is a table of values that describes this character set.

Pentaho Data Integration has a set of components that contribute to its functionalities. They are 'Spoon', 'Kitchen', 'Pan' and 'Carte' [7]. Spoon can be considered as a desktop application that creates simple and even complex extract, transform and load (ETL) jobs without making the users write or read code. It is used for transformations and jobs with the help of editor and it is the one that is used in most of the cases such as editing, debugging or running a transformation or a job. As the transformations are created in Spoon, they can be executed with the help of a standalone command line process called Pan. It is an engine that reads data, manipulates it and loads into various data sources. Kitchen is another standalone command line process that for executing jobs. It schedules different jobs to run at regular intervals. Carte provides remote execution capabilities and a medium for setting up a remote ETL server.

6. BIG DATA USE CASES

Cyber security analysis helps the end users such as data scientists and security analysts in quickly detecting the threats. Cyber security analytics allows the users to utilize most of the staff resources via automation [8]. It enables the data scientists to perform predictive analytics with the help of machine learning tools. It also provides the automation of blending and reporting on a variety of data. Pentaho platform can be utilized for data processing, data ingestion and delivery of threat calls with minimal costs and complexity.

Pentaho optimizes data warehouse and speeds up the development and deployment processes. It employs a simplified process for offloading to Hadoop. The offloaded data is usually less frequent data. Hand coding in MapReduce jobs and SQL can be avoided by the usage of visual integration tools. It provides access to data sources ranging from relational to operational to NoSQL technologies. Pentaho MapReduce helps in achieving high performance in a cluster environment. It provides a graphical and intuitive big data integration.

Another use case identified by Pentaho is the streamlined data refinery. Pentaho data integration processes and refines different data sets by using Hadoop as its data processing platform. It provides modelled, delivered and published data sets to the users for visual analytics just by a mouse click. It can be seen as an integration process that blends huge volumes of highly diversified data. It also supplies tools for in-cluster simplified data processing and is regarded as a highly practical approach.

Pentaho's big data support extends the 360-degree view to internal and external customer related data. Customer service teams are provided with time-sensitive and blended streams of data. The presence of an adaptive big data layer relieves several organizations from evolving technologies. Customers are given access to customizable and interactive dashboards. Data scientists are provided with predictive analytics and data mining tools.

7. COMPARISON

Pentaho products compete with that of popular IT corporations like SAP, IBM and Oracle. Pentaho provides open source solutions and is considered to be much cheaper than the proprietary equivalents. Jaspersoft is an established open source rival of Pentaho. Though both Pentaho and Jaspersoft offer similar fea-

tures with similar costs, Pentaho has got wider online presence and more followers in social media [9].

8. LICENSING

Pentaho Community edition is a free open source product licensed under the General Public License version 2.0 and Mozilla Public License 1.1. The Enterprise edition is available to the users on a commercial license. It needs to be purchased under a subscription model which includes services and support [1].

9. CONCLUSION

Pentaho is an open source based platform for diverse big data deployments. It enables analytics in any environment by delivering data with availability, usability, integrity and security. It provides unified data integration and analytics components which are embeddable. The server plugins and desktop applications provided by Pentaho play a major role in enabling data analytics. Pentaho Data Integration is the Pentaho component responsible for data transformation and loading the data. It helps organizations in harnessing the value from their data in order to make their operations efficient and consistent.

REFERENCES

[1] "Pentaho," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Pentaho

[2] "Community wiki home," Webpage. [Online]. Available: http://wiki.pentaho.com/display/COM/Community+Wiki+Home

[3] "Understanding pentaho architecture," Webpage. [Online]. Available: https://www.edureka.co/blog/understanding-pentaho-architecture/

[4] "Pentaho bi suite enterprise edition," Webpage, 2006. [Online]. Available: http://searchdatamanagement.techtarget.com/review/Pentaho-BI-Suite-Enterprise-Edition

[5] M. C. Roldán, "Pentaho data integration (kettle) tutorial," Webpage, 2008. [Online]. Available: http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial

[6] R. Haces, "Pentaho data integration performance tuning," Webpage. [Online]. Available: https://support.pentaho.com/hc/en-us/articles/205715046-Best-Practice-Pentaho-Data-Integration-%20Performance-Tuning-

[7] "Pentaho data integration architecture," Webpage. [Online]. Available: https://help.pentaho.com/Documentation/5.3/0L0/0Y0/010

[8] "What is big data?" Webpage. [Online]. Available: http://www.pentaho.com/what-is-big-data#tab-3

[9] "Compare pentaho vs. jaspersoft," Webpage. [Online]. Available: https://comparisons.financesonline.com/jaspersoft-vs-pentaho