Big data mining: In-database Oracle data mining over hadoop

Zlatinka Kovacheva, Ina Naydenova, Kalinka Kaloyanova, and Krasimir Markov

Citation: AIP Conference Proceedings **1863**, 040003 (2017); doi: 10.1063/1.4992195
View online: http://dx.doi.org/10.1063/1.4992195
View Table of Contents: http://aip.scitation.org/toc/apc/1863/1
Published by the American Institute of Physics

Big Data Mining: In-Database Oracle Data Mining over Hadoop

Zlatinka Kovacheva[1, a)], Ina Naydenova[2, b)], Kalinka Kaloyanova [3, c)] and Krasimir Markov[4, d)]

[1]*Middle East College, Department of Mathematics and Applied Sciences, Muscat, Oman*
[2]*Technologica, Department of Software Development, Sofia, Bulgaria*
[3]*University of Sofia, FMI, Department of Computer Informatics, Sofia, Bulgaria*
[4]*Bulgarian Academy of Sciences - IMI, Sofia, Bulgaria*

[a)] Corresponding author: zkovacheva@hotmail.com
[b)] inaydenova@technologica.com
[c)] kkaloyanova@fmi.uni-sofia.bg
[d)] markov@folbg.com

Abstract. Big data challenges different aspects of storing, processing and managing data, as well as analyzing and using data for business purposes. Applying Data Mining methods over Big Data is another challenge because of huge data volumes, variety of information, and the dynamic of the sources. Different applications are made in this area, but their successful usage depends on understanding many specific parameters. In this paper we present several opportunities for using Data Mining techniques provided by the analytical engine of RDBMS Oracle over data stored in Hadoop Distributed File System (HDFS). Some experimental results are given and they are discussed.

INTRODUCTION

Big data becomes common sense these days and poses new challenges to the methods of data storage and usage. Defined by Gartner (META group) as high volume, high velocity, and/or high variety information assets Big data requires new forms of processing to enable enhanced decision making, insight discovery and process optimization. This forces the most of the companies that provide traditional database management systems (DBMS) - like relational DBMS, to add new functionalities to handle with Big data.

In this paper we explore decisions and possibilities of Oracle Big data solutions that are used for advanced analytics - Data Mining methods.

Following the Introduction, Section "Dm Methods and Big Data: Theory and Practice" briefly presents the classical Data Mining methods, provided by Oracle. As it is challenging to apply these methods for Big data, in Section "The Technology Stack" a short description of the used technology stack is given. The possibilities for dynamical data extraction from HDFS using Oracle are outlined. Section "Association Rules Analysis over Hadoop" presents experimental results from the usage of Association Rule Analysis on data stored in RDBMS Oracle and data stored in HDFS. The experiments are based on the usage of SQL Connectors for text files. Conclusions and directions for future investigations are given in last Section.

DM METHODS AND BIG DATA: THEORY AND PRACTICE

Data Mining incorporates methods and technologies from different areas like artificial intelligence, statistics, machine learning, etc. These enhanced mathematical methods are used to search data and to discover hidden patterns and trends [1].

International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2016)
AIP Conf. Proc. 1863, 040003-1–040003-4; doi: 10.1063/1.4992195
Published by AIP Publishing. 978-0-7354-1538-6/$30.00

Data Mining in Oracle Environment

Oracle Data Mining allows data analysts and data miners to mine star schemes, transactional data and unstructured data stored inside the database. Oracle Data Mining contains a suite of data mining algorithms that are embedded in the database that allows performing advanced analytics on the data.

In fact, Oracle provides the main DM methods, whose realizations are based on different algorithms:

- Associations rules
- Classification
- Regression
- Clustering

The data mining algorithms are integrated into the Oracle Database kernel and operate natively on data stored in the tables in the database. This integration removes the need for extraction or transfer of data into stand-alone mining/analytic servers, as is typical with most data mining applications. This can significantly reduce the time frame of data mining projects by requiring nearly no data movement [2].

The advantage of the usage of in-database Data Mining decreases if the application of the method requires moving of the big data stored in Hadoop and loading them into the corporate Data Warehouse systems. That's why, in this paper, we are focusing on the possibilities of dynamical data extraction from Hadoop for the purposes of building of an in-database data mining model. The results of the data mining method can be stored in the corporate Data Warehouse (DW) to ensure the easy integration with the existing analytical platforms of the enterprise, the defined security model for authentication and authorization, etc.

The Data Mining methods provided by Oracle are using tables or views defined in the Oracle RDBMS as a data source for the building and application of the analytical models. Using so called SQL Connectors for HDFS is possible to define views in RDBMS Oracle, which dynamically read external for the database server data, which is stored in Hadoop. Another possible approach for usage of Oracle Advanced Analytics Engine over data stored in Hadoop is by defining a transparent gateway (using ODBC connections to HDFS - e.g. by Cloudera ODBC Driver for Impala).

All Data Mining algorithms and methods provided in Oracle can use data stored in Hadoop as a source. The experimental results in this paper are based on the Association Rule method because it is particularly suitable for application to big transaction data which can be stored in Hadoop. The described approaches, however, can be applied to an arbitrary data mining method. The Association Rule Analysis is one of the most suitable methods for HDFS data because it doesn't require a specific data customizing before the application of the method. Yet, the necessity of preliminary customizing the data is not an unavoidable restriction for the approaches presented in this paper.

Oracle Data Mining provides a very convenient option which pre-processes the data according to the data requirements for a particular data mining method (ADP or Automatic Data Prepare). However, using ADP, the data should be moved to the RDBMS Oracle for the application of the procedures for data customizing, whereas in case of manual preparation of data these operations could be performed on a "lower level", without the necessity of moving all the source data from Hadoop file system to the Oracle RDBMS.

THE TECHNOLOGY STACK

Apache Hadoop and SQL-on-Hadoop solutions

Apache Hadoop is a popular solution to big data problems. It is of an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System - HDFS) and a processing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process [3]. The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples [4].

In this paper we focus on the opportunities of the application of Oracle Data Mining component over Hadoop data. Oracle provides special tools for R algorithms over data stored in Hadoop. This implies that in many cases the

preferred technology for implementation of Data Mining methods over Hadoop would be R. The aim of this paper and the presented practical experiment is to show that the Oracle Data Mining users who already have experience with this software component, can also apply this technology over the data stored in HDFS. A short description of the Oracle's implementation of R technology over Hadoop is included for completeness of the statement.

ASSOCIATION RULES ANALYSIS OVER HADOOP

Association Rule Analysis is an unsupervised data mining technique that looks for frequent item sets in the corporate data. This data mining technique is frequently used in the retail sector to discover what products are frequently purchased together. This type of data mining is very common in the retail sector and is sometimes referred to as Market Basket Analysis. By analyzing what products or services previous customers have consumed, the company can then prompt a new customer with products they might be interested in buying. For example every time somebody looks at a product on Amazon.com, the site also presents him with a list of other products that previous customers bought in addition to the product he/she is looking at.

Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.

An association rule can be defined as follows [5]:

Given a set of transactions, where each transaction is a set of items, an association rule is an implication A->B, where A and B are sets of items.

In terms of database data it means that if a transaction contains the items in A, there is a probability to contain also the items in B. A is called the antecedent of rule and B - its consequent.

Association Rule Analysis is the only data mining method in Oracle which can operate on native transactional data and it is particularly suitable for data extraction form HDFS, as it is reasonable for a big organization to store its transaction data in Hadoop.

Statement of the Experiments

The experiments are based on transaction data of products purchases of the clients of chain stores. Every transaction contains data for the client, purchased product, sales channel, geographical location of the store, paid sum etc.

For the experimental purposes, four identical Data Mining scenarios have been created – the first one was based on the sales data stored in Hadoop as a text file (the data is extracted in Oracle via External Tables from Delimited Text Files). The second and the third ones were based on the same sales data but stored in Hive table (in the second scenario the data have been stored in an internal hive table, while in the third scenario – in an external hive table). In the fourth scenario, the sales data have been read from an internal Oracle table. The sales data set contains over 6 million transactions. All Data Mining workflows have been performed using the same hardware and the same loading of the system.

Experimental results

On Fig.1 the average time of performance of each scenario is presented (in seconds).

It is shown that the execution time when the data is stored in HDFS as a text file is compatible with the time when the data is stored internally in an Oracle table (1 min 59 sec versus 1 min 41 sec). The most inefficient method is the storage and data access to Internal Hive Table, and it is not recommendable to use this method in data mining workflows building, using in-database data mining methods.

The results show that it is possible to create scenarios when the data is stored in Hadoop and it is used dynamically in data mining workflows on the base of the Data mining component of Oracle Advanced Analytics. These scenarios can be used successfully in practice.

The dynamical extraction of data from text files stored in Hadoop using SQL Connectors for HDFS is performed in a quite acceptable time, though the productivity in the process of building the data mining model can be slightly improved if the data has been transferred to Oracle tables.

Using such scenarios, the process of the building and usage of a data mining method can remain almost transparent for data scientists acquainted with the data mining component of Oracle Advanced Analytics.

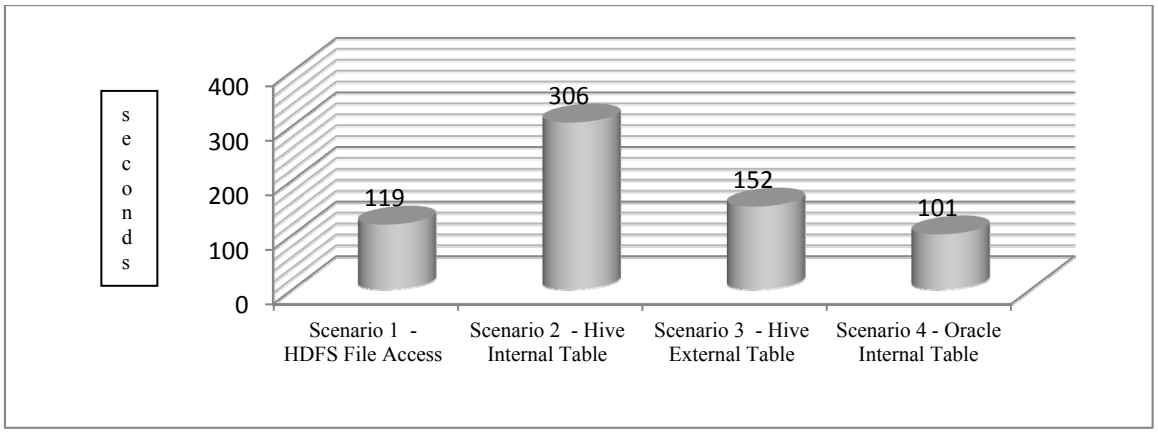

FIGURE 1. Average time of the scenarios

The analyzers can choose an Oracle table as a data source, not necessarily knowing if it is an external table pointing at data stored in Hadoop or it is a usual Oracle table of the Data Warehouse of the company. When working with Data Source component on the base of an Oracle external table for Hadoop data access, an insignificant delay of several seconds is observed.

CONCLUSION

During the last decades, the big producers of relational DBMS integrated a lot of methods as a part of their data management systems. They added powerful analytical engines and options to their database core functionalities. These producers provide also frameworks for data transformation and preparation in order to adopt them to the specific requirements of the different data mining methods.

On the other hand, growing volumes of electronic data over the last decade posed new challenges to the methods of data storage. There were created new concepts and platforms for storing large volumes of data (e.g. Apache Hadoop project).

In this paper we discussed the implementation of Associative rule analysis over data using a set of Oracle technologies. We described methods and technologies briefly and provided some experiments. The results obtained were presented and discussed. The paper aims to evaluate different approaches for extraction of data over Hadoop for the needs of Oracle Data Mining models building and to give a practical direction for using Oracle Big data decisions.

The obtained results show that integrated analytical engines and frameworks in relational DBMS can continue effectively be used for data analysis of big data. Because of their volume, it is more appropriate for them to be stored in a file format on specialized external platforms instead of relational tables. The results also show that the data sources of the currently established / operating data mining projects with minimal changes can be redirected to external file data sources.

The performed experiments and the comparison analysis can be extended including implementations using Oracle Transparent Gateways and/or R scripts.

REFERENCES

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann, San Francisco, 2nd edition), 2006.
2. Br. Tierney, *Predictive Analytics Using Oracle Data Miner* (Oracle Press, March 2015).
3. *What is the Hadoop Distributed File System (HDFS),* ibm.com, (IBM Retrieved 2014-10-30) http://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/
4. *What is Map Reduce,* ibm.com, (IBM Retrieved 2014-10-30) https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/
5. Agrawal, T. Imielinski, and A. Swami, *Mining Association Rules Between Sets of items in large databases*, Proceedings of the ACM SIGMOD Conference on Management of data, 1993, pp. 207-216.