An Automatic Unstructured Big-Data Analysis using Olap-Mondrian Based on Uncertain Data

D. Fernandez Raj and G. Gunasekaran

INTRODUCTION

Big Data is a collection of huge and complex datasets which cannot be processed by conventional data processing applications. There are several challenges being faced in the process of manipulating the data from the Big Datasets. A dataset is considered as a big dataset based on the organization’s capability which is in requirement of data stored the dataset and on the capabilities of the application software which is being used to manipulate the dataset. Recent analysis says that the world’s technology per capita storage capacity has been doubled every 40 months since 1980. The amount of data created per day has been increased from 2.5 Exabyte (2.5X1018) in 2012 to 2.3 Zeta bytes(2.3X1021) in 2014. A Big Data is not only a collection of large datasets, but also tends to be a moving one since the data stored in big data will mostly be characteristics or behavior of a group of people (say customers of a private firm) or technical analysis results (say the results of geological researches and so on). So the Big Data generally possesses three characteristics a) Volume-amount of data stored b) Velocity-the rate at which the data vary c) Variety-the type of data being stored. Some organizations also further classify the characteristics based on which the Big Data can be characterized such as variability, complexity, etc. Even though there are several definitions based on which a Big Data is characterized, Volume, Velocity and Variety are called three vs. of the Big Data.

The Big Data processing is considered as a complex task and it is usually performed in parallel software systems and infrastructure that can traverse through the huge amount of data without much difficulty as a personal desktop computer. The process of collecting data in digital form is believed to improve an organization’s development proportionally. The analyzing and retrieving of data will be much simpler and easier if the data are stored digitally. The challenge being faced in Big Data is not mainly about the storage, the real problem arises when the data in the dataset is manipulated or retrieved from the Big Data. The technology of Big Data is the combination of modern technology and classical technology. The Big Data management relies on the capability of the organization to manage a large volume of data at the right speed within the right time period to allow real time analysis and reaction. The organizations which are relying on Big Data are in the process of finding a solution to...
overcome the challenges faced in manipulating a Big Data. So the recent researches on Big Data concentrate on the factors like reducing the cost, reducing the time and development of new applications or products.

**Background study:**

Big Data is the data that contain several datasets which cannot be processed by conventional database systems. Data contained in a Big Data are very large, moves very fast and do not fit database architectural structures. To manipulate the Big Data an alternative processing way must be chosen (Big Data Now, 2012). McKinsey Global Institute defines Big Data as the datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze (Big Data, 2011). The majority of Big Data comes from unknown sources; therefore they are inaccurate and require correctness. Veracity is as a parameter which is used in characterizing a Big Data is defined as the credibility and suitability of the data from data source (Sathi, 2012). The data from the social networking such as LinkedIn, Netflix, Face book, Twitter, Expedia, national and political network data, such as local campaigns possess very high social, economic and political values which are also very huge in terms of storage (Davis, 2012). According to (http://bernard.lupin.pagesperso-orange.fr/) Big Data has 5 properties called 5V—Volume, Velocity, Variety, Value and Veracity. In addition to the above five properties of Big Data, a Big Data processing model should consist of methods for data specification, modeling, capture, transfer, management and procedures for their collection, transfer, analysis, storage and processing. Considering the requirements of Big Data processing techniques, there are more possibilities of techniques like MapReduce, Information Integration, Data Visualization and Recommendations to be developed in the field of data engineering. The era, we live in is the information era where big data play an important role in day to day life. So it is clear that the organizations are in need of data-driven decision making to gain advantage over others.

There are also some applications currently used for Big Data processing, amongst them Hadoop tool is the most common one. Michael Cafarella and Doug Cutting wrote the open source software framework Hadoop in Java (2014). The use of mapreduces in the Big Data processing technique allows a very high parallel and distributed execution in several nodes. But map reducing is one of the costliest techniques in terms of time consumed in the process of Big Data processing. If the Hadoop tool is used, the process of MapReducing is not necessary as there is no limit of storing and processing data in Hadoop (Jian Tan, 2013). Hadoop uses a distributed file processing system using which it can store and process a large scale of data (Yongwei, 2014). Hadoop is capable of processes in multi machines and moves data on three or more systems and creates a fault tolerant system to avoid data loss (Yongwei, 2014; Shvachko, 2010). Another middleware Java Agent Development Framework (JADE) can also be used in Big Data processing. The JADE functions based on the agent phenomenon and supports services like resource discovery, data encoding, communication support, content delivery and agents (JADE 2014; Bellifemine, 2003).

Most of the big data are taken from various resources from the world so that a lot of correctness and accuracy problems occurred in the data. Data truths, accuracy and reliability provides the credibility of the data source as well as the suitability of the data for the target developers or the end user (Sathi, 2012). Twitter, Face book, Netflix, local advertisement, LIC, various social, economical and big organizations are generating big data (Davis, 2012). Some of the examples of the big-data resources are social media text (Sathi, 2012), cell phone locations, channel-click information from set-top box, web browsing and search, product manuals, communications network events, call detail records, radio frequency identification tags, maps, traffic patterns and so on. The size of the big-data is massive and high in dimension. Genomics dataset is having 5,00,000 micro-arrays where each array consists of thousands of expression values of molecules. In bio-medical engineering, the MRI data set is available in terabytes. And each image is having more than 50,000 voxel values. Time series, e-commerce data, retail transaction records, surveillance videos are some of the big-data having high dimensional data in an unstructured manner.

**Existing approach:**

Several individual programming techniques, built-in data mining tools and private software services are used to analyze the massive amounts of data. Some of the specific tools are MapReduce, Spark, Dryad and higher level scripting languages like Pig Latin and DryadLINQ are significantly improved corresponding tasks for software developers. All the data mining tools are used and developed in terms of programming perspective and system perspective manner. Pig/Pig Latin, Hadoop are the most executable and user friendly software are used for data mining. The data passed through the trusted cloud services are used to encrypt the data and decrypt the data. Some of the predefined UDFs are used for handling cryptographic operations. In this paper, the OLAP tool is used for big data analysis, where the size of the data is increasing dynamically.

The limitation of the existing system is security. Also analyzing huge quantity of data has never been easy with Mondrian. All the features of the Mondrian software is not supported by all the vendors. But Mondrian with OLAP sourced from java can respond for all kind of queries very fast and it easily explore
the data even the size of the data is millions of record. Our contribution of the work in this paper is to apply the dynamic, growing big data into OLAP for analyzation and verify the performance and limitations. But, an OLAP vendor supports the features like parent-child, multi-level hierarchy and drill down features.

**Proposed approach:**

**OLAP – On Line Analytical Programming:**

The entire IT system can be split into OLTP and OLAP used for data transaction and data analysis respectively. OLAP tool is mainly used for analyzing the huge set of business data from the data warehouse. Figure-1 depicts the relationship between the OLAP and the data warehouse.

![OLAP for Data Warehouse](image)

**Fig. 1:** OLAP for Data Warehouse.

OLTP provides the source of data to data warehouses and OLAP helps to analyze that data. OLTP can do insert, delete and update operations on the data in online. The key feature of the OLTP is, it is fast in query processing, maintaining the data integrity in multi-access mode and it can measure accurately in terms of the number of transactions per second. Therefore, OLTP can provide in detail information about the current status of the data and schema used to persist the transaction databases. OLAP is applied for low volume of data, whereas OLTP applied for large volume of data. The key feature of the OLAP is, it is more effective in terms of time. It is mainly and widely used by data mining techniques. The database used for OLAP is aggregated and stored in multi-dimensional data schemas.

![OLTP vs. OLAP](image)

**Fig. 2:** OLTP vs. OLAP.

OLAP is a powerful tool for analyzing the data discovery, controlling the report viewing, complex analytical calculations and create plan according to the data scenario. It can be applied for business performance management, budgeting, forecasting, financial reporting, knowledge discovery and data warehouse planning and so on. Also, OLAP provides solution for the users to perform ad-hoc analysis of data in multiple dimensions for better understanding and can take better decision making. Knowledge discovery helps decision making in business intelligence, since knowledge is the basic foundation for all the successful decisions. OLAP helps in terms of cost and time in a continuous planning, report analysis. A true statistical data report can be delivered by OLAP and can make successful business decisions into shared knowledge. It helps the team leaders of the business can understand the complete data set with fast aggregation in various views very fast and they can take right decisions at the right time. OLAP is also divided into various types according to its functionality such as MOLAP, ROLAP, SOLAP and DOLAP.

All the type of business is multi-dimensional activity and it is necessary to take decisions on the multi dimensions. Tracking business is due to their activity by considering various variables related to the business. If these variables are converted into a spreadsheet, each axis [x-axis and y-axis] representing a logical grouping of variables in a category [pivotal table]. For example, in a sales business, the term sales are represented [in x-axis] in units or dollars, which is tracked periodically like over days, week, months or years are displayed in the y-axis. This grouping of variables is called as dimensions in the OLAP platform. Not like to relational databases, OLAP does not store the data transactions in rows x columns format. Instead of that, OLAP use multidimensional database structures such as pivotal table or CUBES in OLAP terminology. CUBES stores the data in a consolidated form. All the related information is stored in an optimized multidimensional database. During the analysis, the view of the cube is taken as a worksheet like view [slice]. Nowadays, all the big business environments are using many dimensions to track the data. For example, in a multinational company, accounts, locations, periods, sales-customer, purchase-customer and products are the dimensions are considered. All these dimensions are grouped according to the logical interconnections among the variables and they are used to represent the entire picture of the business, planning, analysis and reporting activities. OLAP has the capability to analyze the multi dimension of the data. Also, viewing the manipulated data along the multi dimension can improve the business model as effective.

**Security:**

In common, OLAP makes the user, can easily view and extract the data in various points of views. In terms of data storage model, OLAP uses the static as well as dynamic cubes such as local cubes and offline cubes. OLAP uses some of the Microsoft analytic services such as local cubes and power pivot.
through excel. Comparing with other software OLAP uses semi-additive measures and partitioning. Mondrian is integrated with OLAP in functionality basis. In terms of security Ess base authentication and LDAP authentication is available in OLAP. Also network encryption can be done by SSL, and the cell security, dimension security and visual totals are available in OLAP for data security. icCube is having HTTP basic and Form authentication, SSO based securities in OLAP. In terms of OS, OLAP is comfortable 100% in Windows, 80% in Linux, 70% in Unix and 50% in Z/OS.

Figure-2 represents the overview of the security model deployed in OLAP. Scripts can be developed and deployed for encrypting and decrypting the data in OLAP. This cryptographic model can be handled by different handlers and services integrated in OLAP.

**MATERIAL AND METHODS**

In this paper, the “Best Foot Forward” (http://bernard.lupin.pagesperso-orange.fr) company data are taken for manipulating using OLAP. The data has the shoe sales, monthly quantity and outlet. This data is represented as dimensions [called as axes] in OLAP. The outlet, style and the month of delivery are taken as the axis and the quantity of the shoe delivery, the customer is represented in the data measure. To deploy the entire business data of “Best Foot Forward” company put into OLAP and a sequence of steps are implemented. First, a classical database is created with tables consisting of columns, since “Best Foot Forward” company want to track the progress of its sales. The sales table is given below in Table-1.

<table>
<thead>
<tr>
<th>Month</th>
<th>Style</th>
<th>Quantity</th>
<th>The value of TE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-14</td>
<td>Boot-1</td>
<td>12</td>
<td>1 700 $F</td>
</tr>
<tr>
<td>Jan-19</td>
<td>Gum-Boot</td>
<td>435</td>
<td>65 000 $F</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

According to the sales of the company, the shoes have a style, and it is included as a key field in Table-1. But, for analyzing the data, months should be interconnected with the style which is also given in the form of rows. In order to improve the analyzation efficiently, the quantity, and the outlet are also includes and generate more tables, which are shown in Table-2.
Table 2: Sales by Outlet

<table>
<thead>
<tr>
<th>Month</th>
<th>Style</th>
<th>Outlet</th>
<th>Quantity</th>
<th>Total Value TE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr-2014</td>
<td>Gum-Boot</td>
<td>Lyon</td>
<td>10</td>
<td>1 500 F</td>
</tr>
<tr>
<td>Apr-2014</td>
<td>Sneaker</td>
<td>Paris Bastille</td>
<td>850</td>
<td>260 000 F</td>
</tr>
</tbody>
</table>

To make more efficient in analyzation, the two dimensional data is converted into three dimensional data. To do that, the month, style and outlet are treated as three dimensions for the measure quantity and the three dimensional representation is shown in Figure-3. From Figure-3, it is clear that the measure quantity is represented using the dimensions monthly, style and outlet. The data value may have millions and millions of values under same data type such as integer or float.

From Figure-3, it is very clear that the database has truly become multidimensional. Luckily the measure "Quantity" only has three dimensions. Also, it is feasible to represent the data graphically in a two dimensional screen is shown in Figure-3. In this figure, each square of the cube denotes a value. The outlines denote the dimensions of the cube. Each face of the cube denotes the values in a single direction. For example, one face of the cube is for outlet, second face is for measure and vice versa. The schema of the table and data set used by the Best Foot Forward company is:

The sales data represented in Table-1 and Table-2 are given in the form of normal relational database with the use of style and outlet monthly that is two dimensions. To make the analysis more complex, the number of dimensions is increased from two dimensions. To do that, the six possible reports can be created as:
- Prepare a report for single outlet
- Prepare reports for single or all outlets
- Prepare reports for each style or for all the style according to month and outlet
- Prepare reports for the year in terms of months with outlet and style.

If we swap the columns and rows, we get more reports for analyzing the entire data. And the formula can be applied in row wise or in column wise as shown in Figure-5.

Simulation and result:
It is assumed that only a small set of data is taken for experimenting with the OLAP tool. The entire data of the “Best Foot Forward” company is created as a table [means in terms of rows and columns] and given to OLAP tool. Initially this table data is converted into two dimensional as well as into three dimensional for analysis followed by the steps explained above. The entire data represented as a simple graph in a excel spread sheet is shown in
The data comprises of outlet, date, time, month and style with quantity. The quantity is available in the range of 0 to 3000 in numbers, the outlet type is Paris, Marseille and Lyon, weeks are 22 and 23 and the month is from January to April. Generating the reports using for Table-1 is very few and analyzing the complete data is difficult. But, generating reports using OLAP is more and it is used for analyzing the data very easily in various dimensions. From Figure-6, it is clear that, the data is represented the quantity in terms outlet.

Table 3: Best Foot Forward Data Represented using OLAP

<table>
<thead>
<tr>
<th>Outlet</th>
<th>(Units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td>1234</td>
</tr>
<tr>
<td>Marseille</td>
<td>1235</td>
</tr>
<tr>
<td>Lyon</td>
<td>1236</td>
</tr>
</tbody>
</table>

The entire data represented in multi dimension is from OLAP and table-3, is represented in graph mode is shown in Figure-7. Figure-7 shows the quantity, date, month, reference and outlet on the same graph. According to the year and month, the weekly quantity is shown separately for all the reference data. It can be filtered weekly, date wise and reference wise etc, and those kind of reports are shown in Figure-8, Figure-9 and in Figure-10.

Figure-8 shows that the data is filtered for the outlet as Lyon and the references are 1234 and 1235. From this figure, it is clear that this data available in this category is very few, whereas the reference 1234 is available in 6/1/2000 and in 6/6/2000. The reference 1235 is available in 6/3/2000 and in 6/6/2000 only. The reference 1234 is not in 6/3/2000 and the reference 1235 is not in 6/1/2000. From figure-9, it is clear that this report generated for only
the reference 1234, 1235 and 1236. The quantity of these three references is chosen under all the outlet is given here. The reference 1234 is not available in 6/3/2000 and not in 6/7/2000. Other than that, all the three references are available in both weeks of five months in the year 2000. Figure-10 shows that the report generated for week wise and it for only week 22 for all the references 1234 to 1240. In the whole data set, in week 22, there are three dates the product is produced and the quantity is measured.

From the above figures, and tables, it is concluded that OLAP can provide more reports than other data mining tools. For, less number of data, OLAP generated amore number of reports, from this it can be understood that for a huge set of data, OLAP can generate more number of data using the logical relationships among the data in various dimensions.

Conclusion:

In this paper, Big-data analysis using the OLAP tool is studied. OLAP tool simplifies the data analysis by converting the fewer dimensions of data into more dimension based data. Even though the data size is too small and it has less number of fields, OLAP can convert the data into multiple dimensions of the data and it can provide more reports to analyze the individual data entity. In future enhancement, OLAP can be applied for various size and various kinds of data to check the performance for creating a novel framework efficient than OLAP for middle sized problems.

REFERENCES


