Analysis and Visualization of Data Assimilating Hive and COGNOS Insight 10.2.2

Mandeep Virk*, Vaishali Chauhan, Urvashi Mittal

Department of Computer Science Engineering, Chandigarh University, Gharuain, S.A.S Nagar Mohali, Punjab (140413) India.

*Email: mannvirk.virk27@gmail.com

Abstract

Data analysis is the most grueling tasks in the coinciding world. The size of data is increasing at a very high rate because of the procreation of peripatetic gadgets and sensors attached. To make that data readable is another challenging task. Effectual visualization provides users with better analysis capabilities and helps in deriving evidence about data. Many techniques and tools have been invented to deal with such problems but to make these tools amendable is the main mystification. It is the big data that originated as a technology which is proficient in assembling and transforming the colossal and divergent figures of data, providing organizations with meaningful insights for deriving improved results. Big data is accustomed to delineate technologies and techniques which are used to store, manage, distribute and analyze huge data sheets. The existent of administrating this research is to make the data readable in a more suitable form with less comprehend. Mainly the research emphasizes on the fabrication of using COGNOS insight 10.2.2 for visualizing data and implementing the analyzed results derived from the hive. The assimilation between tools has also been reformed in this research.

Keywords: Big Data; Data Analysis; Data Visualization; Hadoop.

1. Introduction

As the volume of data grows the requirement for deriving more accurate insights from data also raises that will help the business organizations in better decision making. The resulting amount of data available is in different forms and this data is still increasing at a very high rate. The rate of increase of such data is high to such an extent that it becomes impossible to cope with the computational performance of data. Therefore, the problem of processing and storage for such huge volumes of data is faced. Processing and analyzing the data creates new data which also needs to be handled. And visualizing such data is a cumbersome task. Effectual visualization provides users with better analysis capabilities and helps in deriving evidence about data. Big data is a hitch that relates to extremely large and varied data volume which is difficult for storing and processing. The conventional database technologies are not competent to deal with the elevation of data. The action of big data is fuzzy which demands substantial ways to distinguish and restate the data into novel and meaningful insights [10]. Several researchers have narrated big data in different manners in the earlier literature. For instance, [2] mentioned huge data as the hefty figure of data which can be utilized for visualization in scientific areas. There are different definitions of big data. For example, [3] mentioned it as the volume of data that is on the far side of technology's capability to handle. Meanwhile, [4] and [5] mentioned it as described by three Vs: variety, volume, and velocity. These terms (variety, volume, and velocity) were initially presented by Gartner as a starting point to define the reasons for different challenges in big data. IDC also stated big data technologies as “a birth of novel technologies which are designed to collect, process and investigate the gigantic bulk of data” [1] interpreted big data as a methodology that demands substitutive forms of consolidation to unveil values from massive data sets which are complex, multiple and hefty. [6] Defines big data in five distinct phases (generally referred to 5 Vs) that are variety, volume, velocity, value, veracity.

(1) The volume deals with the expansion of data which is getting generated from several sources and is not easy to handle.

(2) Variety concerns with the distinct structures of data originated from varied sources like web, enterprises, customer relations, weather, farm etc. [7].

(3) Velocity concerns with the speed at which data transfers. [8].

(4) Value is the indispensable facets of big data; it deals with discovering concealed assessments from extensive datasets [9].

(5) Veracity indicates the deformity of data. Veracity of data sources results in the accuracy of the analysis.

The problem with the gigantic bulk of data hinges on these three factors: Volume, Velocity, Variety. When the enormous bulk of
data structured/unstructured gets generated at a faster rate that becomes incapable of handlingthe traditional systems contributes to a big data problem [11]. More reliable tools are necessary to perform such task applied on bulky data collections. New models and computing paradigms were intended to brace them using hardware resources in form of clusters and other distributed computing architectures. Hadoop framework and MapReduce paradigm are the supported technologies. Map Reduce [13] is dominant technique among the other techniques for managing huge data in the cloud environment; it provides an environment for large datasets which are stored in the cluster to get processed. The foundation of cloud framework can act as a suitable environment for executing big data analysis by contending the required data storage [14]. The pretension of Hadoop was the space utilization.[13] Hadoop is a framework that is open source and is developed by using JAVA platform. Hadoop utilizes a DFS to store and contend the bulky data by providing the fault-tolerance environment. The data is analyzed by the substantial usage of technologies like ‘HIVE’. A query language ‘HQL’ is accustomed for summarizing data and querying.

Hadoop is constructed of:

- Hadoop File System (HDFS)
- Programming Paradigm (Map Reduce).

HDFS is a platform-independent file system formed using Java. In this largely sized queries are split into blocks of certain size to elevate the data processing rate. It provides well-organized data storage. MapReduce is consociated with dealing bulky data sets on a cluster by using different algorithms. The analyzing process of sizeable datasets that get accumulate in HDFS is braced by technologies like Hive using hql. Constituents of HDFS are:

NameNode: Name node is the principal node that encloses apprehensions about file system of Hadoop [13]. It is the sovereign node that perpetuates, instructs and governs the blocks that reside on data nodes. It is assigned with the chore of inscribing metadata, aspects and explicit locations of files and data segments in data nodes. Every respective change is inscribed that transpire in the file system metadata. Example: It will at once inscribe if a file gets deleted in HDFS.

To clinch the survival of data node, NameNode gets a hand on the regular heartbeat and block account of data nodes. Replication factor is inscribed for all the blocks by NameNode.

DataNode: Data node is the laborer node. Deriving upon sufficiency and interpretations Hadoop may embrace more than one data node. The tangible data resides in the data node. Times to time the heartbeats are sent to the NameNode by this. Accumulating a block in HDFS and serving as a stage for running jobs are two predominant chores of the datanode.

Secondary NameNode: This node is the adherent node of the sovereign node. Rather than carrying itself as the auxiliary of the name node, it invariably reads the metadata proximately out of the RAM. NameNode and writes the same on the system of the file or the hard disk.

Blocks: In HDFS, the data is dispersed over the data nodes as blocks. The minimal unending locus over the hard drive at which the data reside are the blocks.

HDFS Client: These are generally labeled as the edge nodes consistently. It operates as the interface between NameNode and data node.

MapReduce: It is a programming standard utilized for refining, processing and engendering enormous datasets [13]. The datasets to be inputted are gathered in a cluster of segregations in a DFS/redistributed on each lump in the cluster. Then the program is dragged into a distributed processing architecture and is then executed. It has two sections:

Map Stage: Map task refines and procedures data in parallelism way when the system divides the data inputs into multitude pieces. These miscellaneous pieces are assigned with map task. It interprets the inputs and generates intermediate output.

Reduce Stage: After shuffling by the framework, the intermediate output is put to the reduce task and final output is generated.

The IBM (COGNOS INSIGHT ten.2.2) that may be a significant tool to investigate and visualize information [15]. IBM COGNOS Insight acts as a significant tool which helps the users by making the procedure of importing, analyzing and sharing the personal over and above corporate data faster. Not solely will its accustomed analyze and make “what if” situations by attempting out totally different arrangements of your information, it also can be accustomed produce and publish plans, forecasts, and budgets and supply visualizations with less comprehend and moreover results are enforced to indicate the prognostic results of the information. It makes advanced information a lot of accessible, perceivable and usable.

The reason behind conducting this research is to fabricate the data easily interpretable. As the data collected is in very large and is not easily understandable without a meaningful visualization. In this research, COGNOS has been accustomed to fabricate the visualizations for the data which is being analyzed. The idea behind using this tool is that the tool requires analytical skills and knowledge of stats instead of long queries that are required in HIVE. Using COGNOS will make the data analysis process easier. The reason behind using COGNOS for visualizations and the relativestudy of data visualizations have been discussed in this research after performing various data related tasks using the tools.

The objective deals with the analyzing data using ‘HIVE’ and visualizing it with COGNOS 10.2.2. Which is to form the data simply explainable with less comprehends. Further, this paper is arranged as follows: Section II presents the analysis of data and workflow. The results and discussions over square measure explained in Section III. Section IV presents the conclusions.

2. Analysis of data and workflow

The planned technique is created by considering following situation under consideration: Criminology has a vast quantity of knowledge concerning a gamut of crimes, date and time of happening of crim. No. of crimes worn out every location. The matter they faced until currently, and they need the competence to research restricted information from databases. The planned model meaning is to unfold a model for the crime information to produce a platform for brand spanking new analytics supported the subsequent queries.

Dataset used: Crime from 2001-2016[16]
Table 1: Crime dataset

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Shows the identity of each record.</td>
</tr>
<tr>
<td>Case number</td>
<td>A unique number given to a particular case.</td>
</tr>
<tr>
<td>Date</td>
<td>Date and time of crime.</td>
</tr>
<tr>
<td>IUCR</td>
<td>Illinois uniform crime reporting code. It is linked to type and description of crime.</td>
</tr>
<tr>
<td>Primary type</td>
<td>Type of crime.</td>
</tr>
<tr>
<td>Description</td>
<td>Describes the crime</td>
</tr>
<tr>
<td>Location description</td>
<td>Where the crime has happened.</td>
</tr>
<tr>
<td>Arrest</td>
<td>Shows the person is arrested or not.</td>
</tr>
<tr>
<td>Domestic</td>
<td>True/false</td>
</tr>
<tr>
<td>Beat</td>
<td>Time and territory at police patrols.</td>
</tr>
<tr>
<td>District</td>
<td>Region associated with area.</td>
</tr>
<tr>
<td>Ward</td>
<td>Particular street.</td>
</tr>
<tr>
<td>Community area</td>
<td>Related to old and needy people where crime happened.</td>
</tr>
<tr>
<td>Fbi code</td>
<td>Code given to each crime.</td>
</tr>
<tr>
<td>X coordinate</td>
<td>Region allotted to police for patrolling.</td>
</tr>
<tr>
<td>Y coordinate</td>
<td>Region associated with patrolling.</td>
</tr>
<tr>
<td>Year</td>
<td>Years in which crimes has happened.</td>
</tr>
<tr>
<td>Updated on</td>
<td>Time of updating of data</td>
</tr>
<tr>
<td>Longitude</td>
<td>Range of area associated with police.</td>
</tr>
<tr>
<td>Location</td>
<td>Complete area on mapping region.</td>
</tr>
</tbody>
</table>

A. Workflow

Analysis using Hive

The process goes as:

1. Create tables with required attributes.
2. Extract semi-structured data into the table employing the load command.
3. Analyze data for the following queries.

Where IUCR=820 and primary type='theft'

Where arrest='true' and beat=2011

Where location = 'restaurant'.

This displays the result with the number of cases of 'theft' which are governed and criminals are accused accordingly.

Visualization using COGNOS insight

The data is large and complex and the results are not easily understandable if they are not in a visual format. The hive doesn’t give any visualization of data. Therefore, COGNOS insight is employed to provide visualizations to the crime dataset to by extracting information making it more understandable with less comprehend.

The progress work of the research flows by defining the issues related to the earlier analysis reports. The workflows as follow:

Problem definition

Literature review

Requirement specification

Gathering data

Importing data into the tool

Data processing

Results

The workflow goes as:

- The reason behind conducting the research and need to carry the research –explained in the introduction section.
- Data requirement: The dataset included in the research is crime data (size=1.52 GB). It boasts with 65,0000 records or more than that. This dataset reflects reportable incidents of crime happened within the town of Chicago from 2001-2016.
- In the first, the data will get loaded into the hive database through creating a table. Furthermore, the analysis will be performed on the data using HQL for following queries: a. A number of criminal cases where the cardinal cause of crime is theft with IUCR_NO = 820.
  B. how many of them got arrested having ‘restaurant’ as the location where theft has occurred.
- Next, data will be imported to the COGNOS 10.2.2 for analysis and visualization. This will help the data to understand easily for better decision making.

3. Results and discussion

This paper accentuates on the analysis of crime data. The usage of a contemporary analytical tool like a hive on big data has been reformed in this paper that emphasizes on the basic requirements of any crime department. Visualization of the data following the same queries has been done to presage the number of criminal records accordingly. It is found that the cardinal cases of theft happened in restaurant is 1 and the criminals are accused for conducting theft in such locations accordingly. Also, the visualized results accordingly are defined using the figures. Certain instances are highlighted below with the sample snapshots shown in Figure two to six. Figure a pair of shows the produce table and cargo information commands for HDFS system. It additionally offers a range of Map and scales backт that are internally taken care of the underlying tools of Hadoop System. Figure 3, four shows sample queries that are executed with Hive on Hadoop. Fig 5,6 shows the envisioned results of the queries victimization in COGNOS. Figure 4 shows the results with the primary type of crime theft with particular IUCR number while the records according to a particular location are visualized in figure 5 COGNOS insight requires analytical skills instead of long queries that are being used in other tools which makes the visualization tasks easier with less comprehend. As the data
displayed with the help of images is readable with less comprehend.

Fig.2: Loading data into table

Fig.3: List of records where crime is related to theft

Fig.4: List of crime records according to specific location

Data visualization using cognos 10.2.2

- Where primary_type= theft and IUCR=820

Fig.5: Visualized results with primary type theft.

- Where arrest= true and location description= restaurant

Fig.6: List of records according to location

4. Conclusion

The size of information at this time is big and continues to extend daily. The variability of information being generated is additionally increasing. The rate of data generation and growth is increasing as a result of the acceleration of devices connected to the internet. This information gives opportunities that permit businesses over several industries to realize the period of time business insights. The analysis provides the broad road of the info from totally different aspects. The issues concerning shrinking attributes and pattern analysis can be reformed using these techniques. The visualization makes the data easily understandable. The work can be done by integrating the COGNOS with hive tool hereafter to derive the results more expressively and quickly.

References


