Research Journal of Computer and Information Technology Sciences ______________________E-ISSN 2320 – 6527
Vol. **4(9),** 5-8, September **(2016)** *Res. J. Computer and IT Sci.*

Review Paper

Analysing Big Data sets Using Descriptive Analytics

Pandey A.* and Singh R.
Department of Computer Science & Engineering, Shri Shankaracharya Group of Institutions, Chhattisgarh Swami Vivekan and Technical University, Bhilai - 490006, Chhattisgarh, India
pandeyananya92@gmail.com

Available online at: www.isca.in, www.isca.me
Received 24th Feburary 2016, revised 23rd August 2016, accepted 30th August 2016

Abstract

Big data is generally a term used to delineate massive volume of data that is difficult to process using conventional techniques of data processing. Big data, for any enterprise, refers to the data sets that exceeds its current data processing capacity. As big data is arriving from many different sources with a huge velocity, volume and variety, it is necessary to handle them and extract meaningful information that could be beneficial for an enterprise. For this various kinds of analytics are done. The objective of any analytics solution is to provide the enterprise with actionable insights for better business outcomes and smarter decisions. This paper focusses on analysing big data sets considering data from banking sector thus helping banks in various aspects like customer segmentation, sentiment analysis, transactional analysis, security and fraud management etc. Here Descriptive Analytics is done which uses business intelligence and data mining for learning from past behaviours and thus helping in decision making.

Keywords: Big Data, Descriptive Analytics, Business Intelligence, Data Mining, Customer Segmentation, Sentiment Analysis, Transactional Analysis, Fraud management.

Introduction

Problem Identification: "Big data"– which undoubtedly means many things to many people–is no longer restricted to the realm of technology. Now a day's big data is providing business solutions to cope up with the challenges in banking and financial markets companies around the world. Financial sectors are using big data to change their business process, their firms and furthermore, their entire industry[1].

Big data is proving itself to be a promising factor for growth for financial services companies. Without the need of manufacturing any physical product, data – the source of information – is one of the most important assets[2]. The business of banking and financial management is related to transactions, conducting and handling millions of transactions daily, each adding another tuple of records to the industry's database and growing as an ocean of data. So the question arises how for utilizing this information for achieving competitive advantage? The common solution is to perform analytics[3]. Big data itself does not create any value, however, until it is used for solving major challenges of business. This requires access to more than one kind of data, as well as tools skills to perform strong analytics. While examining companies like banking and finance markets which are engaged in big data activities, it is noted that these companies begin with a strong core of analytics capabilities designed to work on structured as well as unstructured data using techniques like general queries, simulations, predictive modeling and optimization. Datasets for these domains are often very vast for data analyst and business analysts to view as well as to analyze with traditional data mining and reporting tools. To analyze these data sets various tools are available. In this paper we present how big data sets are processed and analyzed using Hadoop as an analysis tool. This paper aims to describe how big data analytics is being successfully used in banking sector, with respect to following fields: i. Sentiment Analysis, ii. Transactional Analysis, iii. Customer Segmentation, iv. Spending patterns of customers[4].

Related Work: Financial services organizations around the world are experiencing drastic change. The global financial crisis of 2008 resulted in the failing of scores of banks, which also impacted incomes, jobs, and wealth. Because of this reason, financial institutions need to work hard to avoid the repetition of such financial crisis[5].

From the literature[1] it has been described how banking and financial markets extract value from uncertain data by explaining the three kinds of analysis done in big data.

Hoppermann J. and Bennett M. explains how new data sources as well as the tools and technologies that can use all of this data to provide better customer insights. This provides solution for using big data in banking sectors as well as business value insights that the solutions provide to banking sectors[6].

Financial and banking sectors believe that if they have to survive in a market that has changed so drastically, they need to be strong to cope up with inefficiencies in the operation, detect fraud accurately, manage risk, and improve customer services[7].

To accomplish this, finance and banking services companies are moving towards big data technologies as well as Hadoop for analyzing fraud patterns and reducing risk, understanding customer requirements more clearly and achieve competitive advantage[8].

Methodology

Introduction to Hadoop: For analysing data sets we have used Hadoop which is a framework written in java used for analysing big data sets. There are mainly five deamons of Hadoop through which it accomplishes its task namely- Name Node, Secondary Name Node, Data Node, Task Tracker and Job Tracker. In addition to this, it has various components for handling different varieties of data. In this project, we have used hive for query processing and data handling.

Map Reduce Methodology: The analysis focusses on Map Reduce Algorithm whose primary objective is to split the input data set into independent chunks that are processed in a completely parallel manner.

The Map Reduce framework comprises of a single master daemon called Job Tracker and a single slave daemon called Task Tracker for each node cluster. The Job Tracker is responsible for monitoring the jobs, executing the failed jobs and scheduling jobs to Task Tracker. The Task Tracker being the slave executes the tasks as directed by Job Tracker.

The Hadoop Map Reduce framework works by firstly sorting the outputs generated and secondly providing them as input to the reduce tasks. Both the input and output of the job is stored in Hadoop Distributed File System[10].

Map Reduce Algorithm works in the following manner: i. Map () input – the "Map Reduce system" assigns input to Map processors, provides a key value K1 to each processor in addition to the input data which is associated with K1. ii. Map () function – Map () executes just once for each key value pair and generates output which is organized by the other key value pair. iii. "Shuffle"– the Map Reduce algorithm uses Reduce processors and provide the next key value that each processor would work on, thus providing the reduce processor with the map generated data that is associated with that key value pair. iv. Reduce () function – Reduce () is executed just once for every key value pair generated in the second step. v. Final result – the Map Reduce algorithm gathers all the output generated through reduce operation, and sorts it by key value pair thus generating final output.

Results and Discussion

This section presents some of the snapshots which explain how big data sets are handled using Hadoop as a framework. For analysis purpose we have created a dummy data at initial that will contain customer details, transaction details of customers, card details and loan details determining the loan amount and type of loan. Below snapshots describe how operations are being carried our using Hadoop and its components.

The Figure-3 snapshot describes creation of table in Hadoop. For this, hive-a component of Hadoop is used. Tables are created so as to load the data from local system into Hadoop Distributed File System and then we can analyze those data sets.

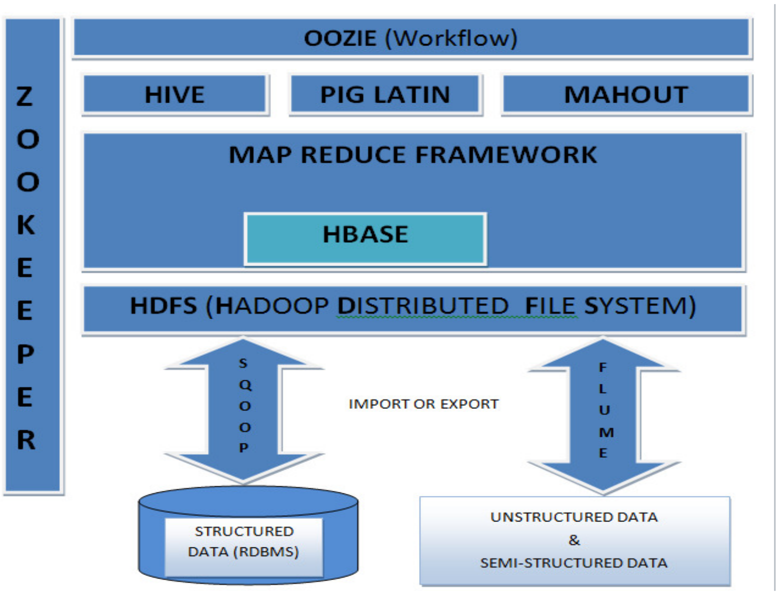


Figure-1
Components of Hadoop Ecosystem[9]

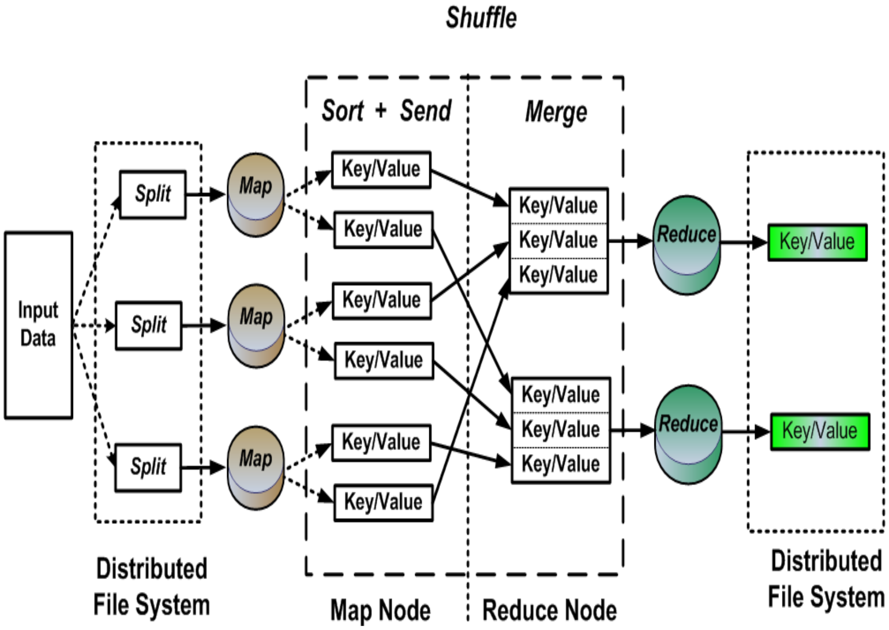
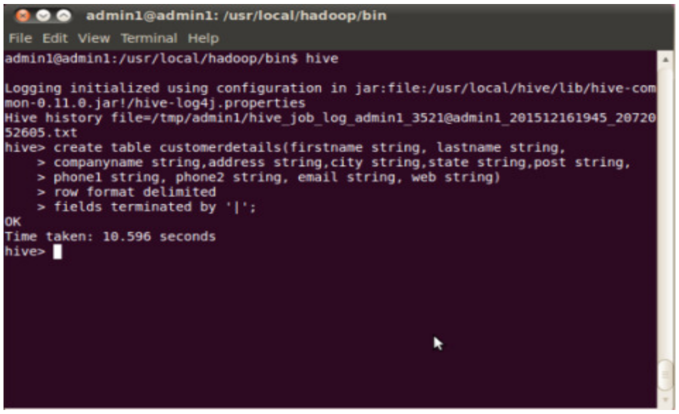


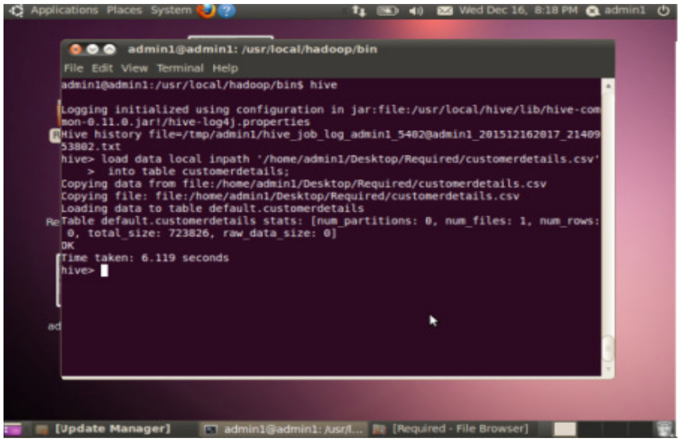
Figure-2
Workflow diagram of Map Reduce methodology[11]


admin1@admin1: /usr/local/hadoop/bin
File Edit View Terminal Help

```
admin1@admin1:/usr/local/hadoop/bin$ hive

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-0.11.0.jar!/hive-log4j.properties
Hive history file=/tmp/admin1/hive_job_log_admin1_3521@admin1_201512161945_20720
52605.txt
hive> create table customerdetails(firstname string, lastname string,
    > companyname string,address string,city string,state string,post string,
    > phone1 string, phone2 string, email string, web string)
    > row format delimited
    > fields terminated by '|';
OK
Time taken: 10.596 seconds
hive> 
```

Figure-3
Table Creation in Hadoop

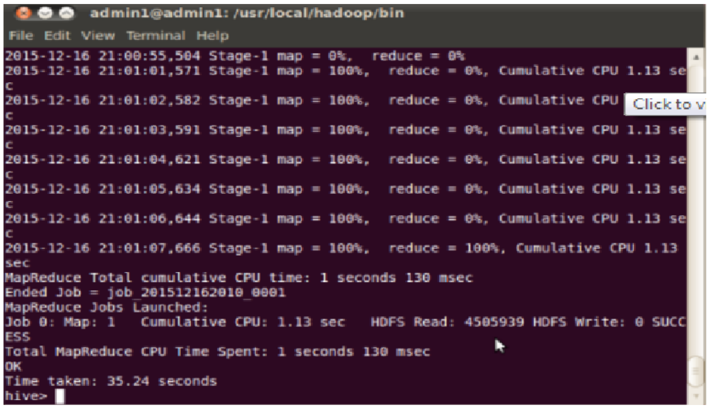

Applications Places System
Wed Dec 16, 8:18 PM
admin1
admin1@admin1: /usr/local/hadoop/bin
File Edit View Terminal Help
[Update Manager]
admin1@admin1: /usr/l...
[Required - File Browser]

```
admin1@admin1:/usr/local/hadoop/bin$ hive

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-com
mon-0.11.0.jar!/hive-log4j.properties
Hive history file=/tmp/admin1/hive_job_log_admin1_5402@admin1_201512162017_21409
53802.txt
hive> load data local inpath '/home/admin1/Desktop/Required/customerdetails.csv'
    >  into table customerdetails;
Copying data from file:/home/admin1/Desktop/Required/customerdetails.csv
Copying file: file:/home/admin1/Desktop/Required/customerdetails.csv
Loading data to table default.customerdetails
Table default.customerdetails stats: [num_partitions: 0, num_files: 1, num_rows:
 0, total_size: 723826, raw_data_size: 0]
OK
Time taken: 6.119 seconds
hive> 
```

Figure-4
Loading data into Hadoop

This snapshot describes the process of loading data from a local system to Hadoop distributed file system. In order to load data from one platform to another it is necessary that the schema of the data set to be loaded and the schema of the Table on which the dataset is to be loaded should be exactly same. This should be ensured while creation of table.


admin1@admin1: /usr/local/hadoop/bin
File Edit View Terminal Help
Click to v

```
2015-12-16 21:00:55,504 Stage-1 map = 0%,  reduce = 0%
2015-12-16 21:01:01,571 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 se
c
2015-12-16 21:01:02,582 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU
c
2015-12-16 21:01:03,591 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 se
c
2015-12-16 21:01:04,621 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 se
c
2015-12-16 21:01:05,634 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 se
c
2015-12-16 21:01:06,644 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 se
c
2015-12-16 21:01:07,666 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.13
sec
MapReduce Total cumulative CPU time: 1 seconds 130 msec
Ended Job = job_201512162010_0001
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 1.13 sec   HDFS Read: 4505939 HDFS Write: 0 SUCC
ESS
Total MapReduce CPU Time Spent: 1 seconds 130 msec
OK
Time taken: 35.24 seconds
hive> 
```

Figure-5
Map reduce execution

This snapshot represents a portion of execution of map reduce task for fetching the desired records from the Table. This execution resulted from a user entered query through Hive.

Discussion: The study was done to evaluate the influence of managing big datasets for getting business value in commercial banks and specifically to focus on extent of big data management in commercial banks, advantages of big data management, challenges in big data management and effects of big data management on business value.

This work can be extended to cover the various masking techniques that can be used to protect sensitive data. Moreover, Classification can be applied to classify various data sets to the category they belong.

Conclusion

Big data analytics is being implemented across every sphere of finance services and banking sectors thereby helping them understand customer behaviour and other aspects to improve their services for their customers and to achieve competitive advantage. This study analysed transactional analysis for the Banking Sector, and the outcomes of the same are mentioned below: i. We saw how data sets can be handled in Hadoop and how faster we can access these data sets as well as how we can perform analysis. ii. We observed transactional analysis by keeping track of transactional records of each customer and through this we observed spending patterns of customers.

References

1. Schroeck M., Schockley R., Smart J., Romero-Morales D. and Tufano P. (2012). Analytics: The Real-World Use of Big Data. IBM Institute for Business Value, http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html.22/08/16.

2. Ammu N. and Irfanuddin M. (2013). Big Data Challenges. *International Journal of Advanced Trends in Computer Science and Engineering*, 2(1), 1-4.

3. Verma R. and Mani S.R. (2012). Use of Big Data Technologies in Capital Markets. Infosys® Limited, http://www.slideshare.net/Infosys/use-of-big-data-technologies-in-capital-markets, 22/08/16.

4. Srivastava U. and Gopalkrishnan S. (2015). Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. *Procedia Computer Science*, 50, 643-652, 2nd International Symposium on Big Data and Cloud Computing, 1-10.

5. Bhosale H.S. and Gadekar D.P. (2014). A Review Paper on Big Data and Hadoop. *International Journal of Scientific and Research Publications*, 4(10), 1-3.

6. Hoppermann J. and Bennett M. (2014). Big Data In Banking: It's Time To Act. Forrester Research, http://www.pentaho.com/sites/default/files/uploads/resources/forrester-research-bigdata-in-banking.pdf.01-05

7. Siddaraju D., Soumya C.L., Rashmi K. and Rahul M. (2014). Efficient Analysis of Big Data Using Map Reduce Framework. *International Journal of Recent Development in Engineering and Technology*., 2(6), 1-3.

8. Mridul M., Khajuria A., Dutta S. and Kumar N. (2014). Analysis of Big Data using Apache Hadoop and MapReduce. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(5), 1-4.

9. Kumbhare D. (2016). Understanding the Hadoop Ecosystem. http://hadooptutorials.co.in/tutorials/hadoop/understanding-hadoop-ecosystem.html.10/08/16.

10. Patil P.S. and Phursule R.N. (2014). Survey Paper on Big Data Processing and Map Reduce Components. *International Journal of Science and Research*, 3(10), 1-6.

11. Apache Hadoop (2016). Working of Mapreduce. https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcQ7LQ02nbkgmPCbrb8_UJJcr6r1YFdAw1_3sCPhXH43urbjqSIKQ.10/08/16.