IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ....
Email: editoriijcs@ipasj.org
Volume 2, Issue 8, August 2014
ISSN 2321-5992

Big Data Analytics OverOnline Transactional Data Set

Rohit Vaswani[1], Rahul Vaswani[2], Manish Shahani[3], Lifna Jos(Mentor)[4]

[1]B.E. Computer Engg. VES Institute of Technology, Mumbai -400074, Maharashtra, India

[2]B.E. Computer Engg. VES Institute of Technology, Mumbai -400074, Maharashtra, India

[3]B.E. Computer Engg. VES Institute of Technology, Mumbai -400074, Maharashtra, India

[4]Professor Department of Computer Engineering, VES Institute of Technology,
Mumbai -400074, Maharashtra, India.

ABSTRACT

To compete in this World, organizations need a comprehensive understanding of markets, competitors, suppliers, customer, products.. This requires the proper use and understanding of information. Most of the organization consider information as an important asset. Traditional analytics tool are not sufficient for proper analysis of data.Hence, organizations uses big data tool and framework to boost sales, improve customer service, customer retention, pattern evaluation and gain a competitive advantage.

Keywords: Big Data, Big data analytics , dataset analysis , Analysis using Hadoop.

1.INTRODUCTION

There has been massive increase in volume of data in organizations. Now, organizations are discovering new ways to compete and win – transforming themselves to take advantage of the available information support. [11] Big data is the "buzz word" that represents large amount of data that cannot be handled by the traditional systems. Big data is often described by 3 V's [5]

1. Volume: volume refers to large amount of data that is been generated every second.
2. Variety:variety refers to the different types and formats of data.
3. Velocity: velocity refers to speed with which the data is being generated [5].

Big data analytics is the discovery of the meaningful patterns from large amount of data.Big organizations usually apply these analytics to business data to describe, predict and improve the performance. Some advantages of big data analytics are customer satisfaction, increase in sales, pattern evaluation, to get meaningful from given data etc[1].To analyze large amount of the data Hadoop framework can be used.Hadoop is framework which stores and manages large amount of the data [7]. Hadoop framework along with its components are used in our project to analyze the transactional dataset generated by a website to find out top transactions.

2.IMPLEMENTATION

To develop big data applications, a platform like hadoop is required. There are many tools available in the market which provides support for these platforms. Each tool has different characteristics like fault tolerance, easy integration, hyper clustering, no ETL, zero deployment etc. Tools includes Hortonwork's HDP, Cloudera CDH, Mongo DB, Splunk, Google charts etc. Hadoop framework consists of one of the prime framework, mapReduce. The program written on this framework is mapReduce program. All processing logic is specified in that program. This program has map() and reduce() procedures, whose functionalities are given below:

- **Map() Procedure :**

  Map() procedure that performs filtering and sorting operation on data stored locally on that node by HDFS. When the mapping phase has completed, the (key, value) pairs must be exchanged between machines to send all values with the same key to a single reducer.

  Master node takes an input. Right after taking input master node divides it into smaller sub-problems with map() procedures. These sub-problems are distributed to worker nodes. A worker node processes them and does analysis. Once the worker node completes the process with this sub-problem it returns it back to master node [7].
- **Reduce() Procedure**

Reduce() procedure will perform the aggregate operation on the data. All the worker nodes return the answer to the sub-problem assigned to them to master node. The master node collects the answer and concates it to generate the single answer. The MapReduce Framework does the above Map () and Reduce () procedure in the parallel and independent to each other.

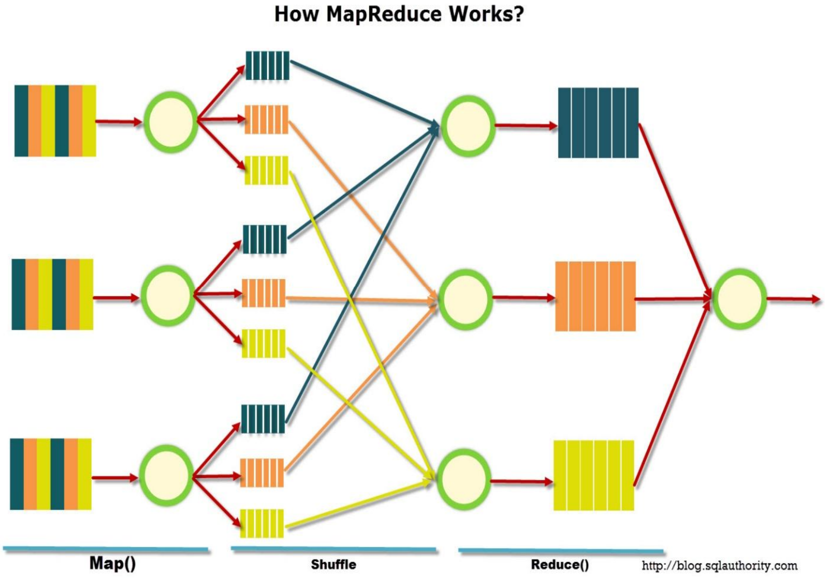


Fig: Map Reduce process

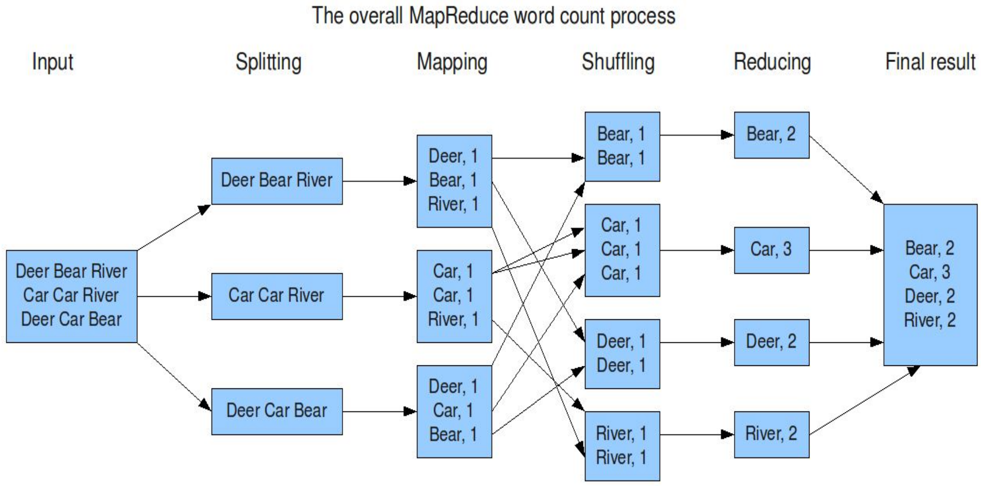
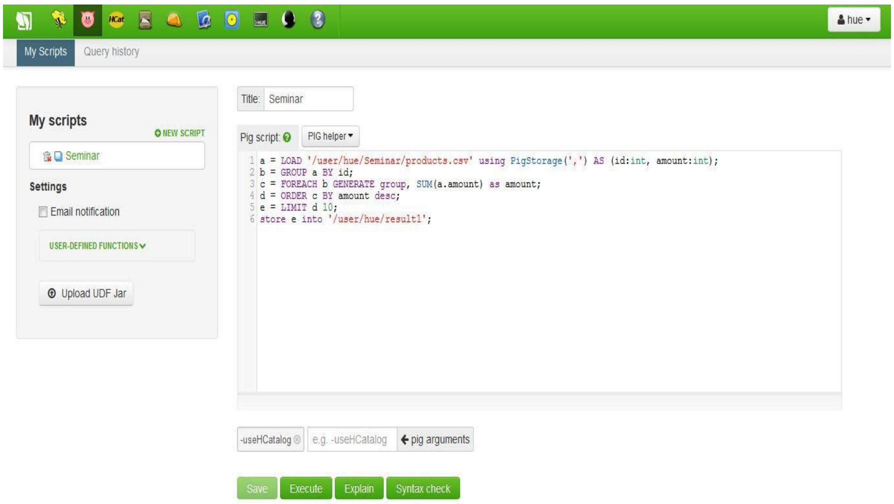


Fig: Word Count Example

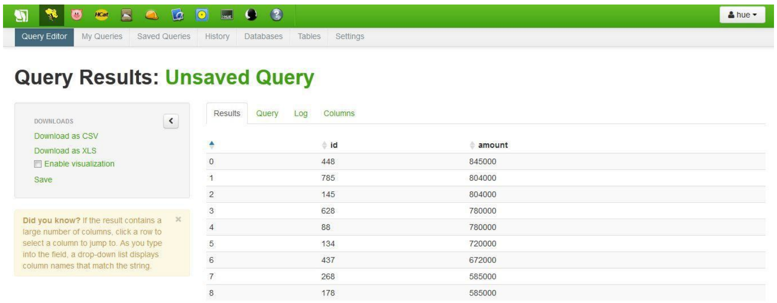
Using Hadoop framework, we have done analysis over online transactional dataset. Large database has been generated through self-developed website. The website is designed for Engineering Students to sale/purchase their products. Database obtained from website is converted into dataset using Microsoft visual tool. This dataset is then loaded into hadoop framework. An analysis is made on it to find out the dataset to find out the top 10 students who has made huge transactions. All processing logic of that analysis is written in mapReduce program using pigLatinscripting language. Resultant output is stored in record structure using hbase.



Title: Seminar

```
a = LOAD '/user/hue/Seminar/products.csv' using PigStorage(',') AS (id:int, amount:int);
b = GROUP a BY id;
c = FOREACH b GENERATE group, SUM(a.amount) as amount;
d = ORDER c BY amount desc;
e = LIMIT d 10;
store e into '/user/hue/result1';
```

Fig: MapReduce Program in Pig Scripting

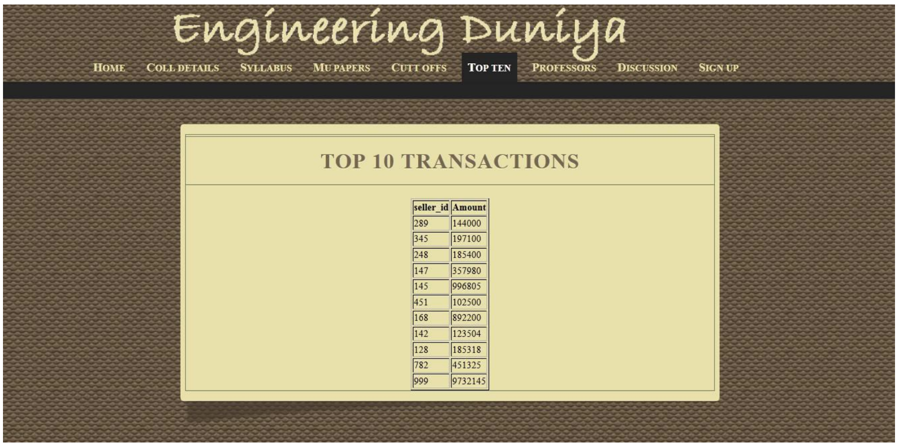


Query Results: Unsaved Query

Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string.

	id	amount
0	448	845000
1	785	804000
2	145	804000
3	628	780000
4	88	780000
5	134	720000
6	437	672000
7	268	585000
8	178	585000

Fig: Data summarization using HiveQL on Hive



TOP 10 TRANSACTIONS

seller_id	Amount
289	144000
345	197100
248	185400
147	357980
145	996805
451	102500
168	892200
142	123504
128	185318
782	451325
999	9732145

Fig: Data summarization shown on website

3. REQUIREMENTS

3.1 Hardware Requirements-

- Desktop Or Laptop with Minimum of 4GB RAM
- Minimum Free Hard-disk Space Required – 20GB

3.2 Software Requirements –

- Microsoft Access
- Eclipse IDE
- Apache Tomcat Server
- Visual Studio
- Horton-Works Sandbox

4. SOFTWARES AND FRAMEWORK

4.1 Hortonwork's HDP

Hortonworks Data Platform (HDP) is a platform for multi-workload data processing across an array of processing methods - from batch through interactive to real-time - all supported with solutions for governance, integration, security and operations [6].

4.2 HADOOP:

Apache hadoop is an open-source java basedsoftware framework for storage and large-scale processing of data-sets on clusters of commodity hardware[5]. This framework is mainly written in Java. Native code in C and utilities in shell script[6]. The Apache hadoop framework is composed of the following modules (Projects): hadoopMapReduce,hadoop DFS,hadoop common**,** Hadoop Yarn, Apache pig**,** Apache Hive, Apache Hbase**,** Apache Hquery

5. RELATED WORK

- Amazon uses big data analytics to predict client behavior.
- Yahoo has launched a tool, Genome which helps marketers to deliver more targeted campaigns [2].
- Stakeholders uses big data analytics to maximize their incentives.
- Samsung uses it to power the content recommendation engine on its newest smart TVs. Progressive Insurance relies on it to capture driving behavior, determine customer risk profiles and decide on competitive pricing [3].

6. CONCLUSION

Every organization must make use of big data tools to analyze their large set of existing data present in different formats. The existing data can be analyzed to get the proper insight in past data which help in proper decision making and to maximize profit. Access to datasets with big data tools provides greater accuracy, transparency, and predictive power. Big data analytics helps organization to boost sales, increase efficiency, improve operations, customer service, risk management, customer retention, pattern evaluation, help with product development and gain a competitive advantage.

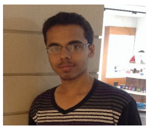
7. ACKNOWLEDGEMENT

This idea would not have been possible without noteworthy contributions of Prof. Lifna Jos who inspired us for making this project the way it is. "The Contributor" has helped us in understanding the concept of big data in detail.

REFERENCES

[1.] http://www.slideshare.net/zanorte/big-data-analytics-2013
[2.] http://www.computerworld.com/s/article/9227146/Yahoo_launches_big_data_analytics_tool_for_online_advertiser
[3.] http://www.revolutionanalytics.com/whitepaper/rise-big-data-spurs-revolution-big-analytics
[4.] http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html
[5.] http://blog.sqlauthority.com/2013/10/29/big-data-final-wrap-and-what-next-day-21-of-21/
[6.] http://hortonworks.com/hadoop/
[7.] http://en.wikipedia.org/wiki/Hortonworks
[8.] http://salsahpc.indiana.edu/ScienceCloud/pig_word_count_tutorial.htm
[9.] http://hadoop.apache.org/
[10.] http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html
[11.] RobPeglar_Introduction_Analytics _Big Data_Hadoop
[12.] Big Data Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses by Michael Minelli
[13.] Assisting developers of Big Data Analytics Applications when deploying on Hadoop clouds Weiyi Shang ; Zhen Ming Jiang ; Hemmati, H. ; Adams, B. ; Hassan, A.E. ; Martin, P. Software Engineering (ICSE), 2013 35th International Conference on DOI: 10.1109/ICSE.2013.6606586 Publication Year: 2013 , Page(s): 402 – 411
[14.] BigDataanalytics frameworks Chandarana, Parth ; Vijayalakshmi, M. Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on DOI: 10.1109/CSCITA.2014.6839299 Publication Year: 2014 , Page(s): 430 – 434

AUTHORS



RohitVaswaniis a fourth year Bachelor of engineering student in VESIT, Mumbai-University of Mumbai.

Rahul Vaswaniis a fourth year Bachelor of engineering student in VESIT, Mumbai-University of Mumbai.

Manish Shahani is a fourth year Bachelor of engineering student in VESIT, Mumbai-University of Mumbai