Aviation Data Analysis using Hadoop and its Ecosystems

Sneha Singh Baghel*, Prof. Imran Ali Khan** and Prof. Garbita Gupta***
*-***Bansal Institute of Research and Technology, Bhopal

Abstract: In airline environments, information is collected for archival and historical purposes from a wide variety of sources. This creates a repository of data that is, perhaps, the most valuable asset for an airline. However, the challenges of managing and extracting information to aid an airline is increasingly difficult to resolve due to customer centricity. In other words, effective data analytics helps in analyzing the data of any business system. But it is the big data which helps and axial rates the process of analysis of data paving way for a success of any business intelligence system. With the expansion of the industry, the data of the industry also expands. Then, it is increasingly difficult to handle huge amount of data that gets generated no matter what's the business is like, range of fields from social media to finance, flight data. In this paper we uses an open source solution to handle such bigdata called hadoop which is used to gathered, stored, processed and analyzed it to turn the raw data information to support decision making. We can analyse the airlines data by apache hive and pig and also compare the performance of these two.

Keywords: Hadoop, aviation data, data analysis , pig, bigdata, hive.

Introduction

About ten years ago, Continental Airlines was faltering and ranked among the lowest of the major US airlines. The problems the airline faced were not unique. But, siloed, unstructured data and incomplete information made determining the root cause analysis a difficult task. Continental Airlines needed a more comprehensive view of the business in order to reduce operating expenses, increase commercial insights and improve time-to-market. Here, too, legacy systems were hampering the growth and profitability story.

In the past decade, there has been a substantial investment by airlines to mothball their legacy systems and move toward a modernization strategy. Some airlines realized substantial benefits by making investments in real-time data warehousing and analytical solutions. Notably, Continental Airlines produced an ROI of over 1,000% with $500 million in cost savings and revenue generation over a six year period.

The challenges for implementing a real-time data warehouse are not alien within an airline environment. Real-time flight data feeds from flight operating systems and either flat-file or XML data feeds from the CRS are two solid examples of velocity and variety-related challenges that real-time Data Warehouse and Business Intelligence (DWBI) systems have to surpass. For many airlines, their current infrastructure is legacy in nature and moving to a new generation of data integration and business intelligence is an added level of complexity. Scandinavian Airlines had to move from a legacy mainframe and DB2 infrastructure in order to implement a 4th generation data integration and business intelligence application.

Even though there is tremendous complexity around the large volume of data airlines produce, real-time DWBI implementation is the need of the hour. With real-time DWBI systems, historical reporting is enabled over a longer period of time, with an increased level of granularity.

Current state

Traditional systems that are deployed for airlines around the world are siloed and very ad-hoc in nature. Rarely do these systems share data in an economical manner. If they do, they are piecemeal extracts after the fact. Ideally, data generated by all operational enterprise systems and partners across an enterprise should be automatically archived and indexed. Regardless of the application or platform that created the data. Airlines should also be capable of searching the entire corporate database to retrieve the relevant data. Airlines need the right information at the right time, with the right degree of accuracy.

For the airlines at this new frontier of innovation, competition and productivity, it's about analyzing masses of unstructured or semi-structured data. This, until recently, was considered too difficult, too time consuming and / or too expensive. But, as airlines get closer to their customers, they gain insight from looking at interaction patterns throughout the customer journey.

Typical attributes of airline data can be identified by four main[12] attributes:

1. Volume – Airline data is massive. Typical tier 1 carriers and Global Distribution Systems (GDS) have Passenger Name Record (PNR) data measured in terabytes.
2. Velocity – Airline data is real-time and arrives quickly, making timely decisions difficult.
3. Variety – Ad-hoc systems traditionally have different structures and shapes. Data analysis on such data is difficult because of rigid schemas.

4. Value – Airline data on its own is low value until it is rolled up and analyzed. At this point data becomes information which, in turn, becomes knowledge for the broader market.

BIGDATA

Big data is a structured and unstructured data video, audio, pictures and information emails etc.it is very large amount of data provided by social site and daily activities of social media like news and news channels or new technology, television, mobile, and computers and industries all are big data[8]. That' s we can say it is more than thousands of information storage for the growth of the industries. Know if we have information or history of previous data than it's very easy for the next new changes for the industries or business. Today' s competitive world in this time industries and business are growing very fastly by the help of the storage of the previous data which is known as big data. Big data is very hard to process and analysis the data easily. But with the help of HADOOP [5] data is easily to process and analysis the data easily. Big data is a different-different collection of complex data sets. Big data [10] is produced by different kinds of sources like television, mobile and other sources like industries data records.

It is three characteristics of big data:

l. volume
2. velocity
3. verity.

Volume

Volume mainly defined is amount of data. Volume of data is growing exponentially megabytes, gigabytes, zeta bytes and petabytes. It' s very large amount of data and also hard to the process and storage. Like Some earlier estimates suggested by the websites that 20 petabytes of storage its very large space was used to store 260 billion Facebook photos and massages or tag. In 2010, it was some newly reported by one million photographs were processed by Facebook per second. Twitter is generate the data 12 terabytes of data per day its newly research. Now the Facebook in 2012 stated that 2.7 billion "likes" and "comments" and massages were registered per day by the peoples.

Velocity

Social media is one the major factor to provide the
data exponentially. Social sites is continuous generated a complex data unstructured and semi-structured form of data. There are currently generated 90% data in last two years. Increase the velocity of big data with help of mobile, televisions more advance technology. Internet is main factor to collecting of huge data. According to the user requirement which is save somewhere. It is known as velocity.

Verity

It is different kind of data are structured and Unstructured format. Structured data is always fixed format everywhere there is no possibility of changes of this data like tabular data, ERP, backup storage for large volume of data. But Unstructured data always no fixed format like text, audio, video, images and many social sites' data like Facebook, twitter, LinkedIn, logs file web chats etc. All companies and industries are having up to 85% of the data semi-structured and unstructured and structured format of data.

HADOOP

The Apache Hadoop[6] project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce is a software framework [8] for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Hive

After congregating the tweets into HDFS they are analyzed by queries using Hive[11]. Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. In opinion mining system, hive is used to query out interested part of the tweets which can be an opinion, comments related to a specific topic or a trending hash tag.

Twitter API loads the HDFS with tweets which are represented as JSON blobs. Processing twitter data in relational database such as SQL requires significant transformations due to nested data structures. Hive facilitates an interface that provides easy access to tweets using HiveQL that supports nested data structures. Hive compiler converts the HiveQL queries into map reduce jobs. Partition feature in hive allows tweet tables to split into different directories. By constructing queries that includes partitions, hive can determine the partition comprising the result. The location of twitter tables are explicitly specified in "Hive External Table" which are partitioned. Hive uses SerDe (Serializer- Deserializer) interface in determining record processing steps. Deserializer interface intakes string of tweets and translates it into a Java Object that Hive can manipulate on. The Serializer interface intakes a java Object that Hive has worked on and converts it into required data to be written on HDFS.

Apache Pig

Yahoo started Pig[7][4] as a research project to focus on analysis of large datasets. It was designed in the style of SQL and also MapReduce. Pig is used with Hadoop in general. Pig Latin is a procedural language used by Apache Pig. The programmers use Pig script and execute the command in the grunt shell. It runs MapReduce programs when running the pig script in grunt shell. Apache pig can execute in three modes which are explained as follows. Interactive mode: Users get the output by entering Pig Latin statements[13], Batch mode: Users run Apache Pig in single file with .pig extension and Embedded mode: User can define their own functions named User Defined Functions (UDF). The major components of Apache pig are, Parser: Checks the syntax of the script. Optimizer: Carries out the plan of the script as push down. Compiler: Compiles the plan into MapReduce job. Execution engine: Execute the MapReduce jobs and finally Hadoop produce the results.

Literature Review

In [1] the research work is carried out using Apache pig and hadoop on a crime dataset. It describes the large volume of data yielded from multiple sources and termed it as voluminous data. Crime and crime related datasets with ever growing population has raised to a higher extent and is a attention seeking subject to government for taking strict measures by prevailing law and procedure. Bigdata analytics using pig and hadoop has been applied on this crime dataset with the idea behind it as the optimal improvement for analysing some trends that needs to figure out, so among the citizens of the country there could be a feel of security and safety. Also it could help the government to furnish law and procedure and welfare among the people of the country. Analysis results shows the total number of crimes occurred in every state, crimes that took place against women, type of crime and from year 2000 to 2014 the total number of crimes that took place. Experimental setup was pseudo distributed mode of hadoop and it was concluded that scripting language Pig Latin has fewer lines of code as compared to mapreduce program but the execution time increases in pig as compared to mapreduce. In [2], the author describes that Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom form big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform ,HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

In [3], Information technology gives utmost importance to processing of data. Some petabytes of data is not sufficient for storing large amount of data. Large volume of unstructured and structured data that gets created from various sources such as Emails, web logs, social media like Twitter, Facebook etc. The major obstacles with processing Big Data include capturing, storing, searching, sharing and analysis. Hadoop enables to explore complex data. It is an open source framework written in Java which supports parallel and distributed data processing and is used for reliable storage of data. With the help of big data analytics, many enterprises are able to improve customer retention, help with product development and gain competitive advantage, speed and reduce complexity. E-commerce companies study traffic on web sites or navigation patterns to determine probable views, interests and dislikes of a person or a group as a whole depending on the previous purchases. In this paper, they compare some typically used data analytic tools.

Proposed Work

For analysing these large and complex airlines data a power tool is required, we are using hadoop which is a open source implementation of mapreduce, a processing framework designed for deep analysis and transformation of very large data.

For analysis consumer complaints datasets we need:-

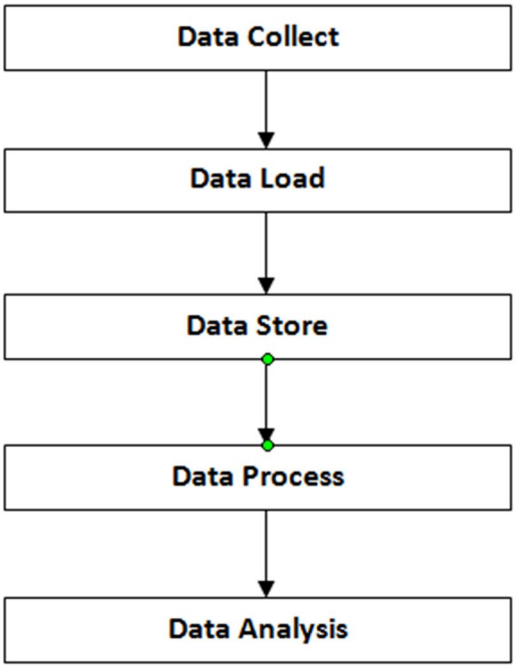
Dataset
We can collect the airlines dataset, which consist a huge amount amount of information.

Hadoop
Hadoop should be configured first as all the mapreduce job will work on hadoop framework, also hadoop comes with HDFS (hadoop distributed file system) which is used to stored such huge or large datasets and Mapreduce which is used to process these datasets.

Bigdata Analytical Tools
For analyzing these large amount of data we need efficient analytical tools[2] which work on the top of hadoop, apache hive and apache pig whereby we can analyze the airlines datasets.

Proposed Methodology

Our Steps or Algorithm Steps will follow are:
Step 1: first we collect aviation or airlines datasets from web resources.
Step 2: After collecting datasets we can load that aviation datasets using hadoop command line.
Step 3: The datasets are store into HDFS which is very reliable for storing huge or complex data size.
Step 4: the aviation datasets are processed by mapreduce which is a processing engine in the hadoop framework .
Step 5: we can analyse these aviation datasets with the help of bigdata analytical tools which can work on top of the hadoop, we can analyse the data using bigdata analytical tools and in the backend the hadoop will process the datasets.



Analysis Steps

Experimental & Result Analysis

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14 [9]. As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

- Analysis using hive
- Analysis using Pig
- Compare performance of hive and pig.

Analysis using Hive

After configuring hadoop on ubuntu we can integrate apache hive on top of the hadoop. And with the help of hadoop we can start analyzing the aviation datasets using hive , in first query we can find the total number of flights travels per year.

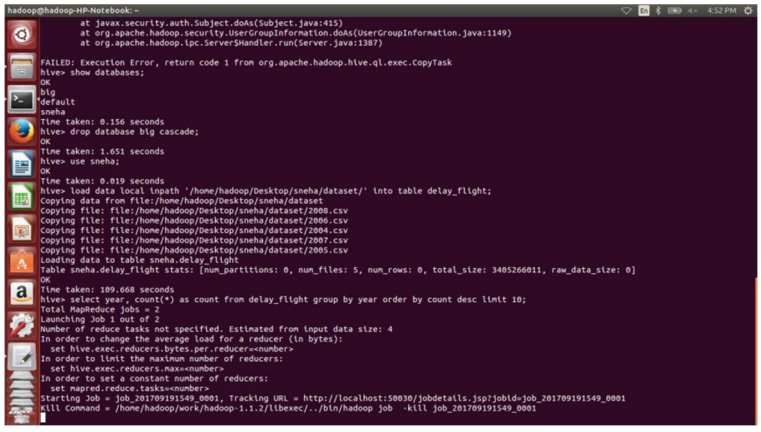


Figure-1. Query launched by hive

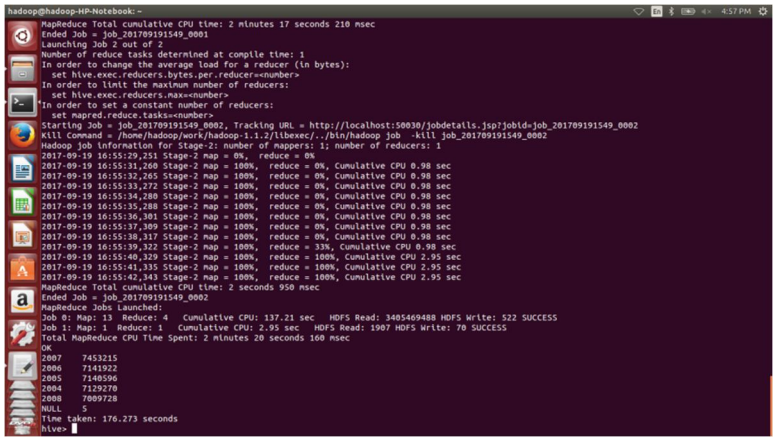


Figure-2. Output of query-1

Aftering finding these we can find the total number of flights cancel per year. By which we can againg launch the query on hive terminal and the output of that query are shown in figure-3.

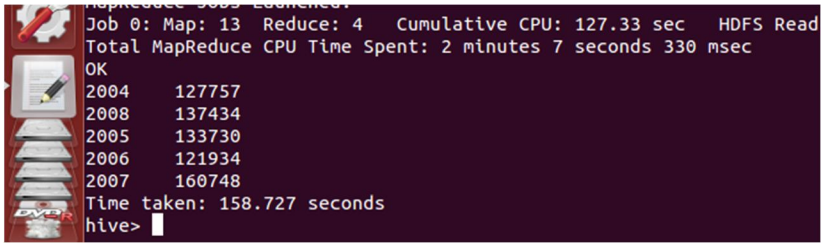
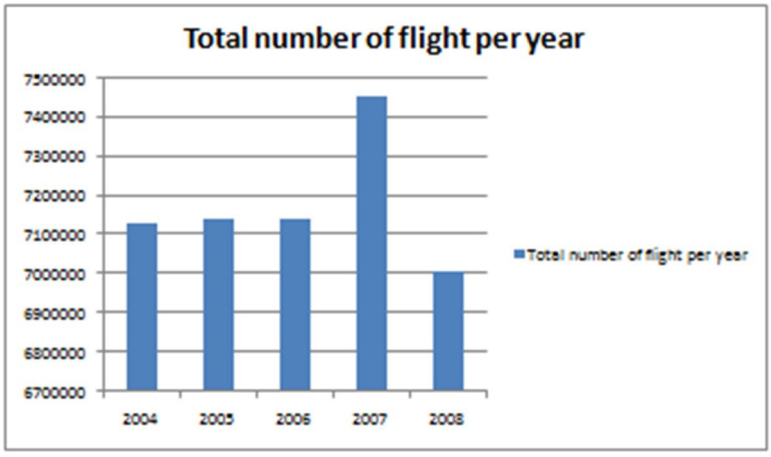
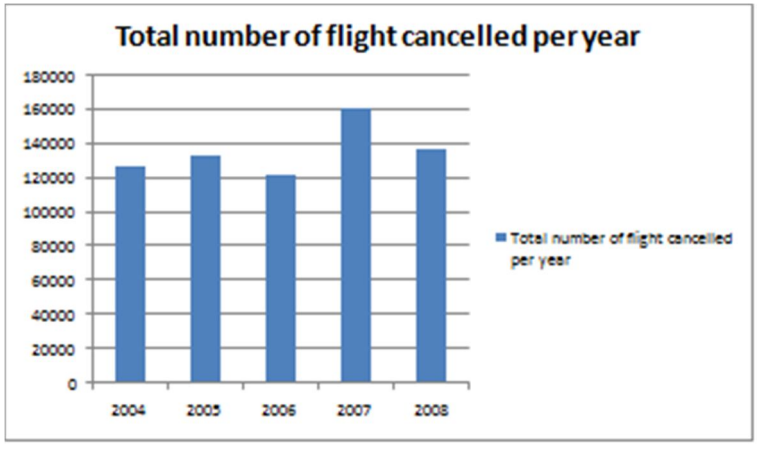


Figure-3. Total number of flights cancel per year

Analyzing using Apache Pig

After analyzing aviation dataset using hive , we can analyze same datasets using apache pig which is again an hadoop ecosystems which runs on top of the hadoop. And in these also we can find the total number of flights per year and total number of flights cancelled per year.






```
Successfully read 35874736 records (3405478137 bytes) from: "/home/hadoop/Desktop/sneha/dataset"

Output(s):
Successfully stored 5 records (60 bytes) in: "hdfs://localhost:9000/tmp/temp1782553896/tmp-892599523"

Counters:
Total records written : 5
Total bytes written : 60
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201709191708_0001   ->      job_201709191708_0002,
job_201709191708_0002   ->      job_201709191708_0003,
job_201709191708_0003   ->      job_201709191708_0004,
job_201709191708_0004


2017-09-19 17:18:47,216 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapR
DISCARDED_TYPE_CONVERSION_FAILED 5 time(s).
2017-09-19 17:18:47,216 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapR
2017-09-19 17:18:47,220 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] wa
2017-09-19 17:18:47,224 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total inp
2017-09-19 17:18:47,224 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - T
(2007,7453215)
(2006,7141922)
(2005,7140596)
(2004,7129270)
(2008,7009728)
```

Figure-4 Output of first pig script

Performance Comparison of Hive and Pig

We can also compared the performance of both the ecosystems based on its execution time, and on the basis of result which is shown in table-1 , we can say that hive is perform faster as compared to pig on large .csv files.

```
Successfully read 35874736 records (3405478137 bytes) from: "/home/hadoop/Desktop/sneha/dataset"

Output(s):
Successfully stored 5 records (60 bytes) in: "hdfs://localhost:9000/tmp/temp1782553896/tmp1794037963"

Counters:
Total records written : 5
Total bytes written : 60
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201709191708_0005   ->      job_201709191708_0006,
job_201709191708_0006   ->      job_201709191708_0007,
job_201709191708_0007   ->      job_201709191708_0008,
job_201709191708_0008


2017-09-19 17:28:23,412 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
DISCARDED_TYPE_CONVERSION_FAILED 10 time(s).
2017-09-19 17:28:23,413 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
2017-09-19 17:28:23,413 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] w
2017-09-19 17:28:23,417 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
2017-09-19 17:28:23,417 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
(2007,160748)
(2008,137434)
(2005,133730)
(2004,127757)
(2006,121934)
```

Figure-5 Output of Second pig script

Table-1 Execution time taken

Execution time in minutes	Hive	Pig
Query-1	3	7
Query-2	2.5	5.4

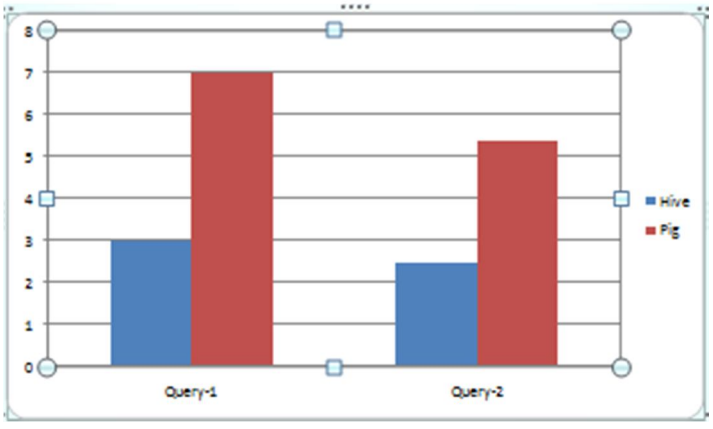


Figure-6 Execution time taken by hive and pig

Conclusion

Hadoop Mapreduce is now a popular choice for performing large-scale data analytics. Bigdata analytics using pig and hive sheds light on significant issues faced by aviation data and we can find the numbers of flight cancelled per year and based on the parameters like execution time, , lines of code it has been examined that hive holds better and efficient than pig.

References

[1] Arushi Jain, Vishal Bhatnagar, “Crime Data Analysis Using Pig with Hadoop” in International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA, in ELSEVIER 2015.

[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3] Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, “BIG DATA ANALYTICS: HADOOP AND TOOLS”, in 978-1-5090-2730-9/16, 2016 IEEE

[4] Jurmo Mehine, Satish Srirama, Pelle Jakovits “Large Scale Data Analysis Using Apache Pig”

[5] Dave Jaffe “Three Approaches to Data Analysis with Hadoop”
[6] http://hadoop.apache.org/
[7] https://pig.apache.org/
[8] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.
[9] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
[10] Sagiroglu, S., & Sinanc, D, “Big data: A review”, IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.
[11] https://hive.apache.org/
[12] Dave Jaffe “Three Approaches to Data Analysis with Hadoop”
[13] Jurmo Mehine, Satish Srirama, Pelle Jakovits “Large Scale Data Analysis Using Apache Pig”