CSE408 –DATA MINING PROJECT

ANALYZING ENTERPRISE DATA USING MAPREDUCE ON HADOOP

Project Report

Slot – F1

Faculty-

Dr. Selvakumar K

SCOPE

Submitted by –

14BCE0703 Pulkit Mittal

14BCE0733 Nikita Gill

ABSTRACT

Big data is presently a buzzword throughout the software industry. Many IT giants like IBM, Google and Oracle have invested billions of dollars into the research to develop frameworks that can handle the big data efficiently. In this project, we make an attempt to analyze enterprise data that cannot be analyzed locally due to size and computation limitations. We analyze these datasets and try to extract the hidden insights from the same. The project aims to design an algorithm that can group and analyze the datasets as per the requirement. Based on the results of the program, we try to identify certain patterns with the help of data visualization and predict information.

LITERATURE SURVEY

Early 2000s — Google stumbled across an obstacle while carrying out its mission — to organize information from all across the globe — which meant that it was crawling, copying, and indexing the entire Internet continuously. Back then, no software could handle the excessively large volume of data to be processed. Even Google's own custom infrastructure couldn't.

To deal with the situation, Google's engineers designed and built a new data processing infrastructure that comprised of two core components —the Google File System, or GFS, which provided fault-tolerant, reliable, and scalable storage, and MapReduce, a data processing system that allowed work to be split among large numbers of servers and carried out in parallel. Google published an academic paper [1] in 2004 describing its work.

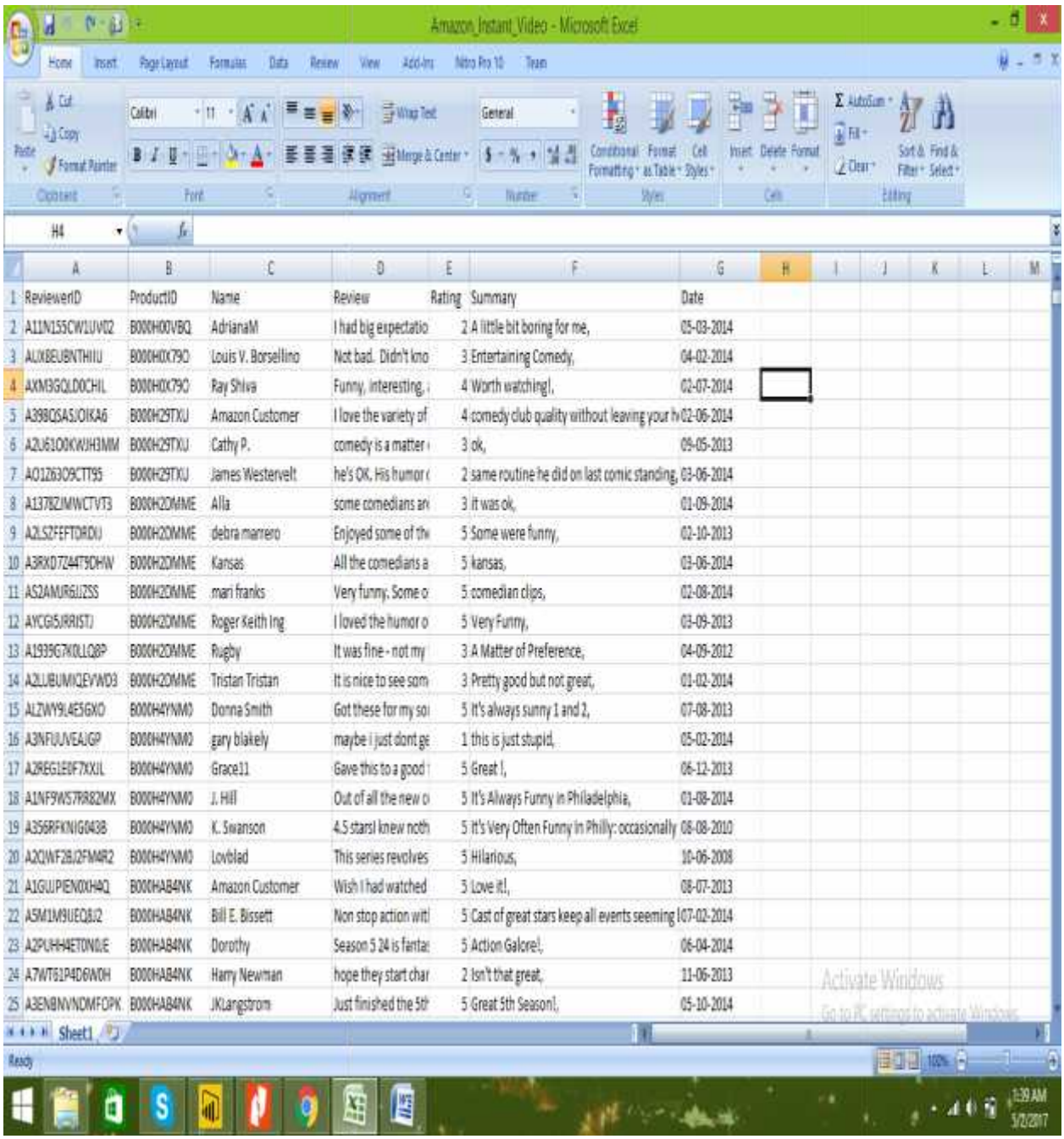
Doug Cutting, a well-known open source software developer, thereafter decided to use the technique Google's paper described. He was working on a web crawler called Nutch [2] and was having the same problems with data volumes and indexing speed that had driven Google to develop MapReduce. He replaced the data collection and processing infrastructure behind the crawler, basing his new implementation on MapReduce. He named the new software Hadoop, after a toy stuffed elephant that belonged to his young son.

Hadoop is an open source project [3] and operates under the Apache Software Foundation today. Hadoop has become a household name and is one of the most popular technologies today to handle big data. It is a data storage and analysis system which is scalable, incredibly flexible and works under the assumption that hardware failures are common occurrences and should be automatically handled by the framework [4] – an assumption that directly leads to its fault tolerant nature.

Hadoop can be deployed in a traditional on-site datacenter as well as in the cloud. Microsoft offers its cloud services via the Microsoft Azure Cloud Service platform which includes HDinsight – the service which shall be used in this project to create and deploy clusters as well as run

mapreduce jobs on the data that needs to be analyzed. HDinsight offers efficient, reliable andperformance centric results [5] with a pay-per-use model which is perfect for a project like the one we are aiming for.

Dataset



	A	B	C	D	E	F	G
1	ReviewerID	ProductID	Name	Review	Rating	Summary	Date
2	A11N155CW1UV02	B000H00VBQ	AdrianaM	I had big expectatio	2	A little bit boring for me,	05-03-2014
3	AUXBEUBNTHIIU	B000H0X79O	Louis V. Borsellino	Not bad. Didn't kno	3	Entertaining Comedy,	04-02-2014
4	AXM3GQLD0CHIL	B000H0X79O	Ray Shiva	Funny, interesting,	4	Worth watching!,	02-07-2014
5	A398QSA5JOIKA6	B000H29TXU	Amazon Customer	I love the variety of	4	comedy club quality without leaving your h	02-06-2014
6	A2U61O0KWJH3MM	B000H29TXU	Cathy P.	comedy is a matter	3	ok,	09-05-2013
7	AO1Z63O9CTT95	B000H29TXU	James Westervelt	he's OK. His humor	2	same routine he did on last comic standing,	03-06-2014
8	A1378ZJMWCTVT3	B000H2DMME	Alla	some comedians an	3	it was ok,	01-09-2014
9	A2LSZFEFTDRDU	B000H2DMME	debra marrero	Enjoyed some of th	5	Some were funny,	02-10-2013
10	A3RXD7Z44T9DHW	B000H2DMME	Kansas	All the comedians a	5	kansas,	03-06-2014
11	A52AMUR6UZSS	B000H2DMME	mari franks	Very funny. Some o	5	comedian clips,	02-08-2014
12	AYCG95JRRIST7	B000H2DMME	Roger Keith Ing	I loved the humor o	5	Very Funny,	03-09-2013
13	A1939G7K0LLQ8P	B000H2DMME	Rugby	It was fine - not my	3	A Matter of Preference,	04-09-2012
14	A2LJBUMXQEVWD3	B000H2DMME	Tristan Tristan	It is nice to see som	3	Pretty good but not great,	01-02-2014
15	ALZWY9L4E5GXO	B000H4YNM0	Donna Smith	Got these for my so	5	It's always sunny 1 and 2,	07-08-2013
16	A3NFUUVEAJGP	B000H4YNM0	gary blakely	maybe i just dont ge	1	this is just stupid,	05-02-2014
17	A2REG1E0F7XXJL	B000H4YNM0	Grace11	Gave this to a good	5	Great !,	06-12-2013
18	A1NF9WS7RR82MX	B000H4YNM0	J. Hill	Out of all the new o	5	It's Always Funny in Philadelphia,	03-08-2014
19	A356RFKNIG043B	B000H4YNM0	K. Swanson	4.5 stars! knew noth	5	It's Very Often Funny in Philly: occasionally	08-08-2010
20	A2QWF2BJ2FM4R2	B000H4YNM0	Lovblad	This series revolves	5	Hilarious,	10-06-2008
21	A1GUJPIEN0XHAQ	B000HAB4NK	Amazon Customer	Wish I had watched	5	Love it!,	08-07-2013
22	A5M1M3UEQ8J2	B000HAB4NK	Bill E. Bissett	Non stop action wit	5	Cast of great stars keep all events seeming	07-02-2014
23	A2PUHH4ET0N0JE	B000HAB4NK	Dorothy	Season 5 24 is fanta	5	Action Galore!,	06-04-2014
24	A7WT61P4D6W0H	B000HAB4NK	Harry Newman	hope they start char	2	Isn't that great,	11-06-2013
25	A3ENBNVNDMFOPK	B000HAB4NK	JKLangstrom	Just finished the 5th	5	Great 5th Season!,	05-10-2014

HARDWARE AND SOFTWARE REQUIREMENTS

Since this project is dependent on cloud services for the analysis, requirements are:

Hardware Requirements – Any system that can run Microsoft's Azure services and have enough storage space to handle the data being analyzed can be used.

Software Requirements – Microsoft Azure's HDinsight platform for Cluster creation and Deployment, Apache Hadoop for analyzing the data and handling data using HDFS, Tools such as Microsoft Power BI for data visualization.

SOFTWARE ARCHITECTURE AND DESIGN

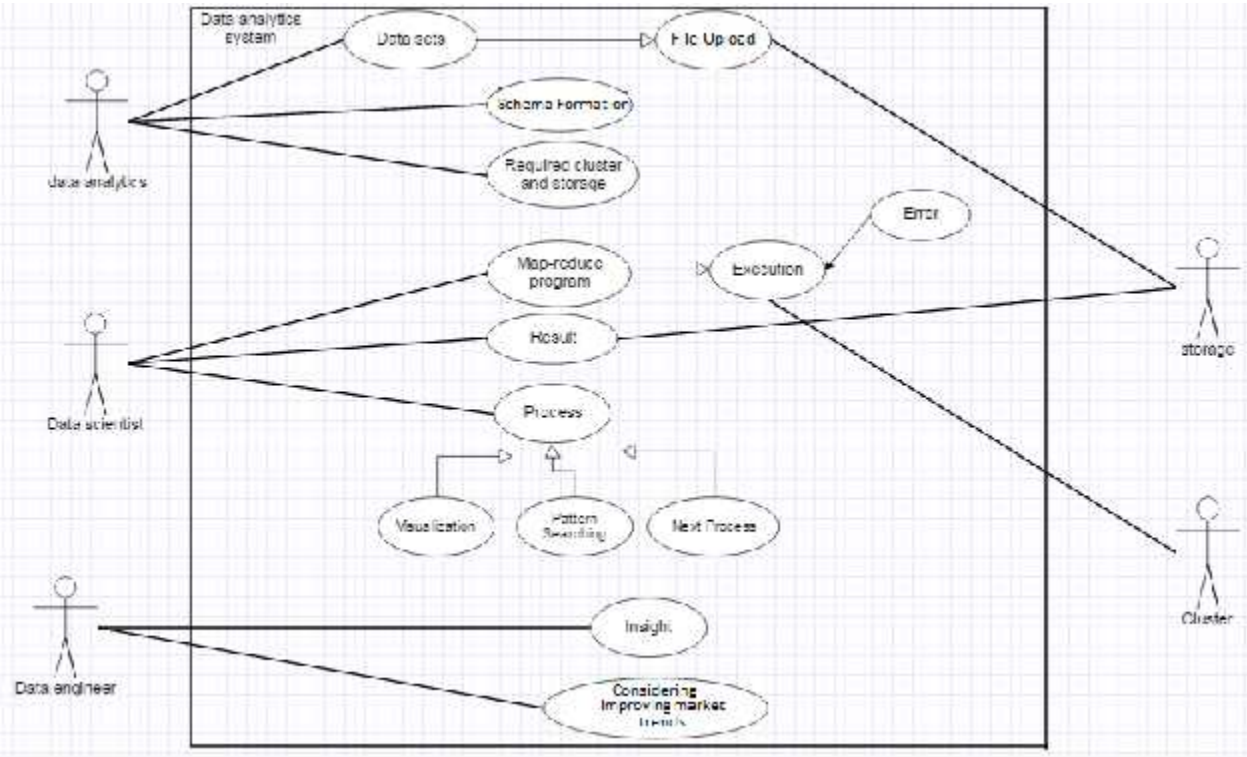


Figure 1. A Use-Case Diagram that illustrates the actors and various use-cases.

This project uses Microsoft's Azure Cloud Service which provides us with a multitude of services. HDinsight, specifically, is the service that is being used in this project.

Azure HDInsight is a service that deploys Hadoop on Microsoft Azure. HDInsight uses Hortonworks HDP and was jointly developed for HDI with Hortonworks. HDInsight also supports creation of Hadoop clusters using Linux with Ubuntu. HDinsight is very flexible in its usage and we can, at any point of time, scale up the cluster by increasing the

amount or type of worker/head nodes that exist in the cluster.

HDinsight is used to set up and deploy clusters oncloud. These clusters are comprised of head as well as worker nodes. Once the cluster is set up along with the storage container,the cluster can be used to upload/download data as well as to run programs that, on a fundamental level, utilize mapreduce to perform analytical tasks.

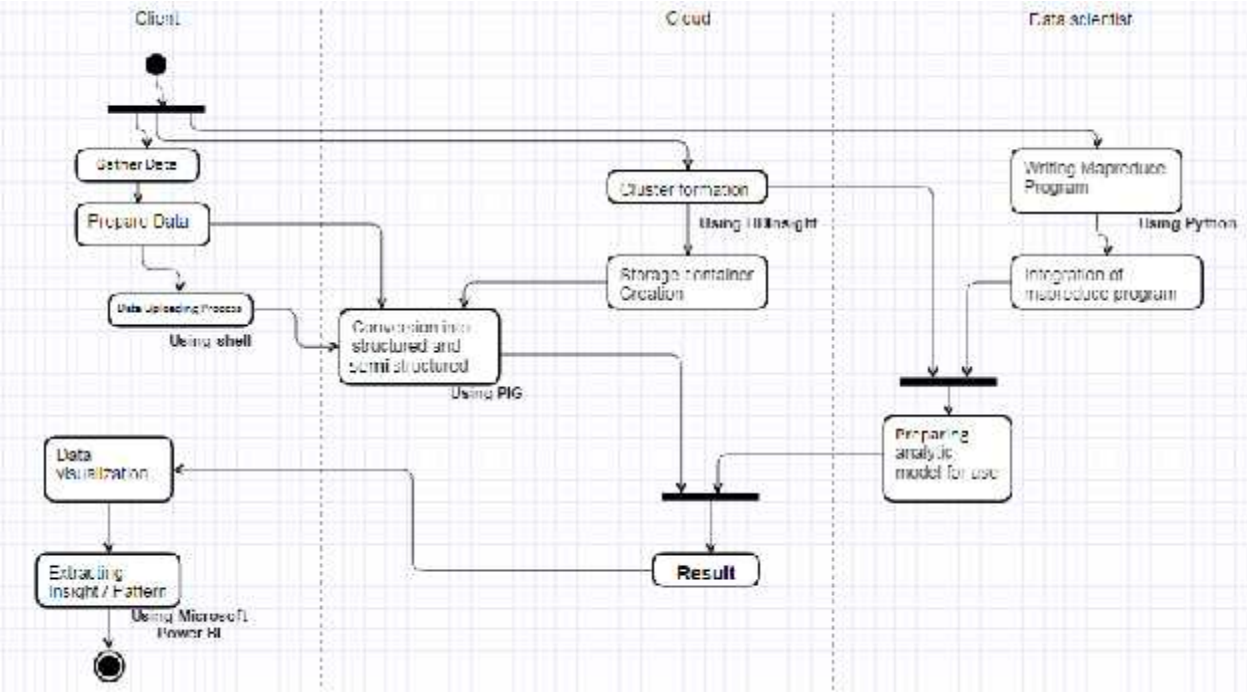


Figure 2. Activity Diagram that covers the entire workflow – from setting up a cloud-based cluster, to managing data on the storage, its analysis and visualization of the results obtained from analyzing it.

As shown in Figure 2, the initial step comprises of gathering data that needs to be analyzed. Depending on the size of the data and scope of the analytical process which determines the processing power needed, a cluster is deployed using HDinsight on the cloud. The Hadoop cluster, depending on the user's need, can be deployed as a cluster utilizing Linux-based machines or Windows-based machines. And depending on the choice made, HDinsight offers different services unique to the selected OS. For example, HDinsight offers SSH access to the Linux clusters while the Windows clusters have exclusive access to the Remote Desktop Access functionality which is otherwise missing from the Linux-based clusters.Since Microsoft's toolset for a Big Data Analyst comprises of a variety of tools, the entire process of cluster creation and deployment can be automated using a script by utilizing Microsoft's Powershell service. The underlying Hadoop framework handles distribution of data for storage using its HDFS

and distribution of analytical work with MapReduce framework.

Once the cluster has been deployed, data to be analyzed is uploaded to the cluster via Azure Command Line Interface. This data is then analyzed by running a mapreduce program on the various machines present on the cluster. The result is then either analyzed further for more efficiency or downloaded locally for it to be used for data visualization. Power BI is a free tool that is then used on the output of the analysis to obtain the results in a graphical format ie, pie-chart, bar graph etc.

MODULE DESCRIPTION

The modules comprise of the following aspects –

ANALYSIS

DATA PROCESSING

VISUALIZATION

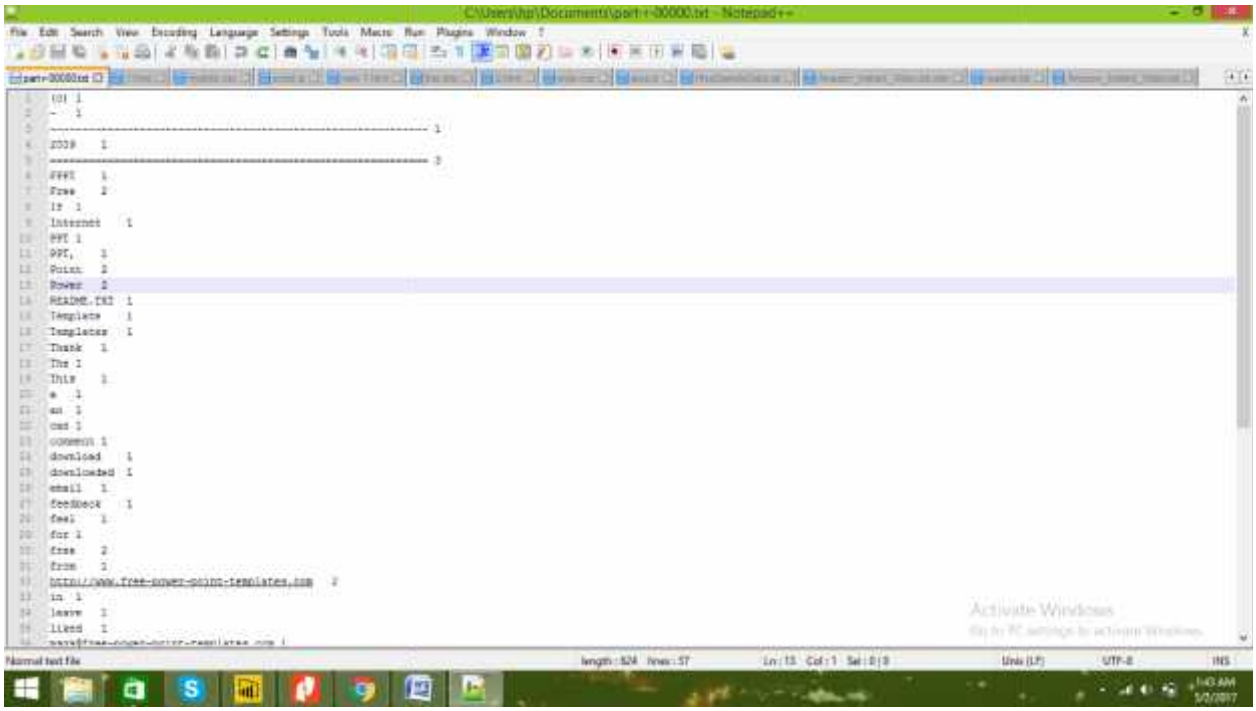
They are discussed below:

Data Processing (using MapReduce)

This process forms the core of data analysis module. It includes both the MapReduce programs written in order to be used with the Tez engine that perform the Map & Reduce jobs when analyzing data using pig scripts or hive queries as well as the mapreduce programs that can be used to perform specific tasks given structured data.

The former forms a part of the package that our environment provides us along with Hive/Pig. The Hive/Pig scripts are run and processing is carried out by carrying out Map and Reduce jobs with the help of these programs. At no point of time does the data analyst explicitly evokes these programs.

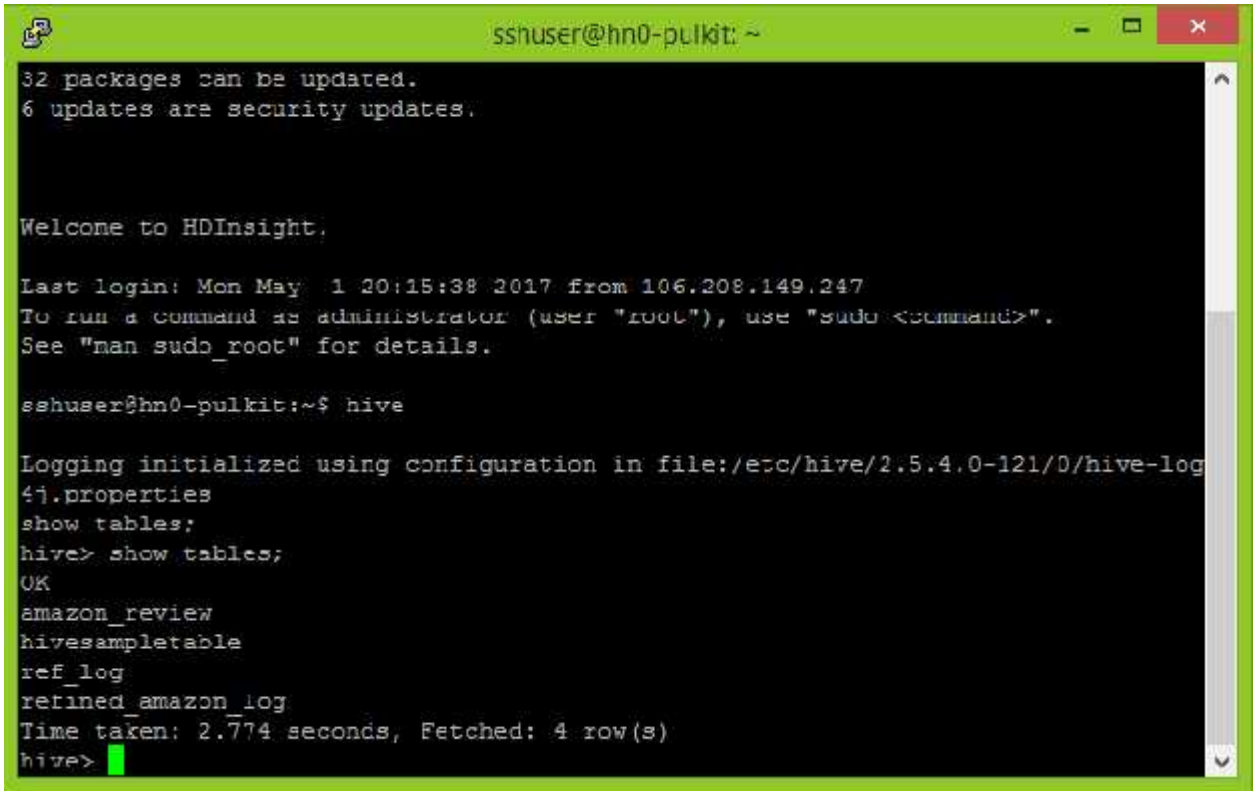
The latter includes programs written explicitly in java like wordcount that form a part of specific mapreduce programs that the package provides analysts. These programs have to be explicitly evoked and data passed to them for them to process it and return useful data.



Analysis (using Hive)

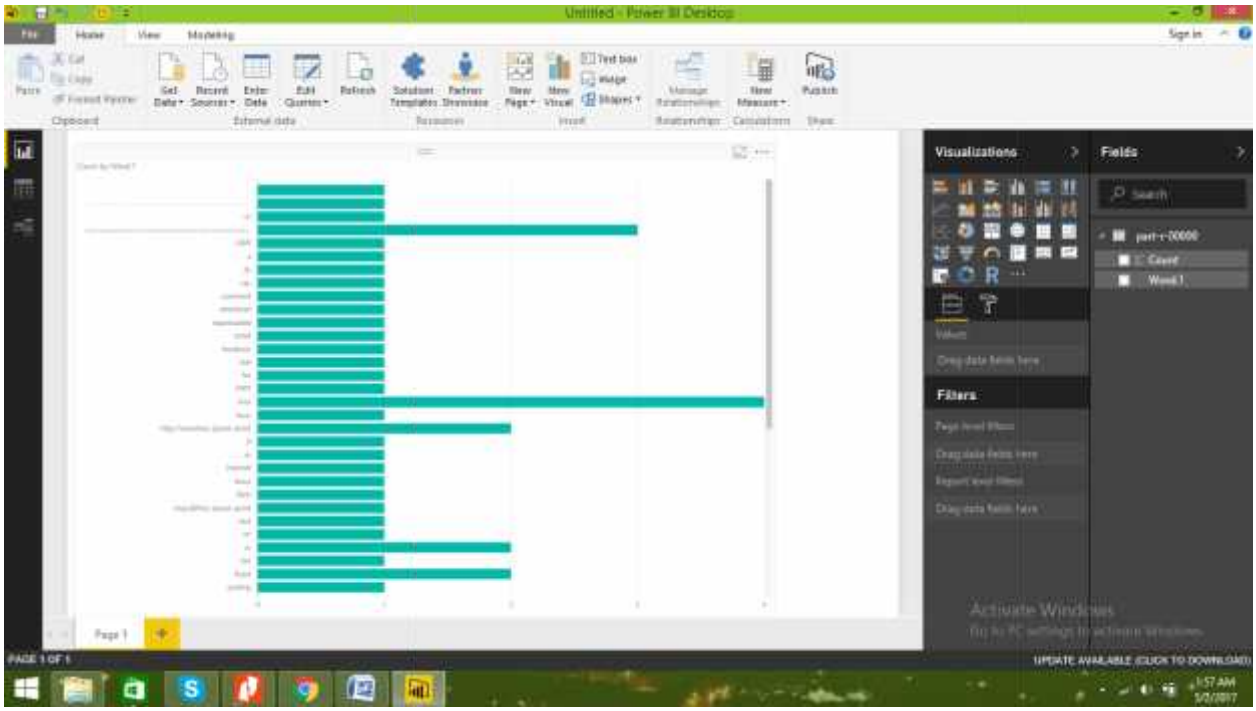
Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

Hive projects a database schema on top of existing directories and the data it contains. It can be used to conveniently import data from a text file, store the resultant data in one or more directories (partitioning to improve performance) and perform SQL-like queries on the said directories. It can be used to work with structured and semi-structured data and is a very powerful tool for data analysis. The SQL-like query language that it uses is called HiveQL.



Visualization (using PowerBi)

PowerBi is a data visualization tool that is provided free of cost by Microsoft. It can be used to visualize structured data such as text files with tab delimited data or data in csv formats. It is a very powerful tool that gives the user a lot of freedom when it comes to the kind of data visualization that can be used.It also enables drilling down and rolling up operations that can be used to abstract data or get into the details of the visualized data. As a result, PowerBi is the software of choice for the data visualization aspect of our project.



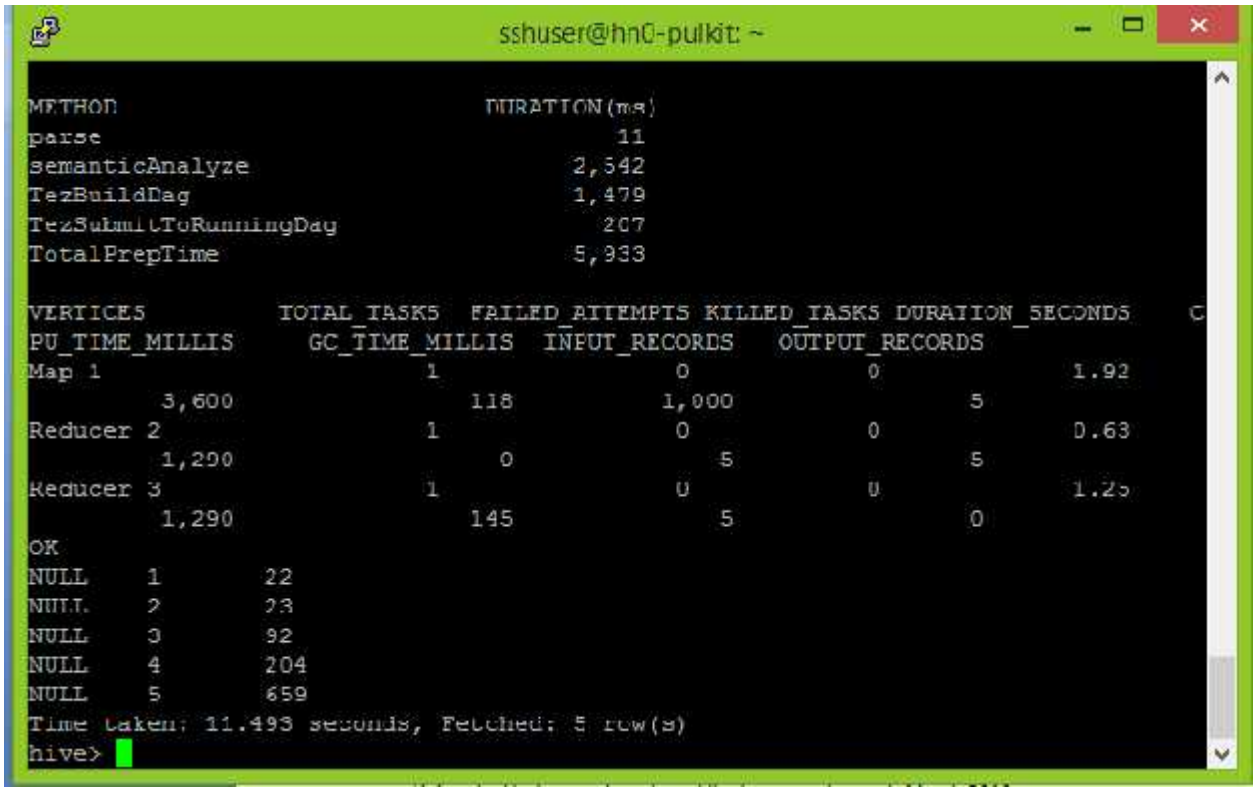
SAMPLE CODE

The hive script below has been used to convert an amazon dataset that was originally in the json encoded format into a tab delimited structured format that can now be easily used for data visualization in PowerBi. It is basically performing the task of dataset refinement. Apart from that, it also uses this staged data to perform the task of counting the total number of 1 – 5 star ratings present in the video reviews dataset.

Hive Script(s) –

```
CREATE EXTERNAL TABLE Amazon_Review

(ReviewerID STRING,

 ProductID STRING,

 Name STRING,

 Review STRING,

 Rating INT,

 Summary STRING,
```

```
 Period STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/data/Amazon_Review';

LOAD DATA INPATH '/data/Amazon_Instant_Video.txt' INTO TABLE Amazon_Review;

CREATE TABLE Refined_Amazon_Log
(ReviewerID STRING,
 ProductID STRING,
 Name STRING,
 Review STRING,
 Rating INT,
 Summary STRING,
 Period STRING);

INSERT INTO TABLE Refined_Amazon_Log
SELECT * FROM Amazon_Review
WHERE Rating IS NOT NULL;

CREATE VIEW Ref_Log
AS
SELECT from_unixtime(unix_timestamp(Period, 'dd/MM/yyyy hh:mm:ss')) AS Period, Review,
Rating, Summary
FROM Refined_Amazon_Log;

SELECT CAST(SUBSTR(Period, 1, 10) AS date) AS Period, Rating, COUNT(*) AS Rating
FROM Ref_Log
GROUP BY CAST(SUBSTR(Period, 1, 10) AS date), Rating
ORDER BY Period, Rating;
```



```
sshuser@hn0-pulkit: ~

METHOD                          DURATION(ms)
parse                                     11
semanticAnalyze                        2,542
TezBuildDag                            1,479
TezSubmitToRunningDag                    207
TotalPrepTime                          5,933

VERTICES          TOTAL_TASKS  FAILED_ATTEMPTS KILLED_TASKS DURATION_SECONDS    C
PU_TIME_MILLIS     GC_TIME_MILLIS  INPUT_RECORDS   OUTPUT_RECORDS
Map 1                        1               0            0              1.92
         3,600                118           1,000                5
Reducer 2                    1               0            0              0.63
         1,290                  0               5                5
Reducer 3                    1               0            0              1.25
         1,290                145               5                0
OK
NULL    1       22
NULL    2       23
NULL    3       92
NULL    4       204
NULL    5       659
Time taken: 11.493 seconds, Fetched: 5 row(s)
hive>
```

References:

1. Dean, J. and Ghemawat, S., “MapReduce: Simplified Data Processing on Large Clusters.” Appeared in Proceedings of the SixthSymposium on Operating System Design and Implementation, San Francisco, CA, December, 2004. Available online at http://labs.google.com/papers/mapreduce.html, March 2010

2.Apache Nutch project, http://lucene.apache.org/nutch/, March 2010

3.Apache Hadoop project, http://hadoop.apache.org/, March 2010.

4. Mike Olson, “HADOOP: Scalable, Flexible Data Storage and Analysis”. Appeared in IQT Quarterly, Spring 2010. Available Online at https://blog.cloudera.com/wp-content/uploads/2010/05/Olson_IQT_Quarterly_Spring_2010.pdf

5. Bhardwaj, Aditya et al, “Analyzing BigData with Hadoop cluster in HDInsight azure Cloud”. Appeared in 2015 Annual IEEE India Conference (INDICON), December 2015