Comparative Analysis of Apache Hive And Apache Pig on Mapreduce Engine

Sanisha[1], Divya Chauhan[2], Vikrant Bhardwaj[3], Kishori Lal Bansal[4]
[1, 2, 3, 4] Dept of Computer Science
[1, 2, 3, 4] Himachal Pradesh University, Shimla, India

Abstract- *Hadoop framework provide various components allowing Big Data to be stored and processed on commodity hardware. Big Data analytical tools are required to extract values hidden inside data. Data analysis is complicated with MapReduce. It require coding in java. Apache Hive and Apache Pig has been designed to solve the problem.by converting queries automatically into MapReduce jobs. Hive is "SQL for Hadoop" and Pig is "Scripting for Hadoop". To select the tool to meet business requirement these tools need to be compared. Hive require data to have a schema. Hence well suited for structured data only. Pig on the other hand do not need any schema definition for data and can process any type of data. Also the performance of both the tools are different. In this paper the performance of Pig and Hive has been compared on three parameters: development effort, number of operation and query execution time.*

Keywords- Apache Hive, Apache Pig, Big Data Analytics Hadoop, MapReduce

I. INTRODUCTION

lobal digitization has resulted into enormous growth of digital data. The data generated is estimated to be 2.5 quintillion bytes for each day. Approximate number of users over internet were 2.4 billion in 2014 and 3 billion in 2016. As of April 2017 this number has grown to 300 million – resulting in a total of approximately 3.7 billion users [1]. With increasing users, data is also growing at rapid rate. By 2020 data is being expected to reach 40 ZB.

Digital age has come up with very large volume of data, called Big Data. Big Data is not just the data kept on servers but the data which is very large and still increasing at high speed. Stored data by itself do not generate any value. This data can be very useful if processed to extract information. So some sort of analytics need to be applied onto it. Big Data Analytics is the process of applying advanced analytic techniques on Big Data to get value from it. Traditional RDBMS were limited with handling only few gigabytes of data. To deal with thousands of terabytes of data advanced analytic tools has been designed. These include: Apache Hive, Apache Pig, Apache Impala. The tools are capable of processing and storing very large datasets in distributed mode. Apache hive is a data warehouse tool to extract insights of data on HDFS. Apache Pig is a software providing PigLatin language to process Big Data. In this paper a comparative analysis of these tools has been done on Hadoop MapReduce.

Paper has been divided into different sections as: Section II gives a brief literature review. Section III describe the Results and Analysis. Section IV concluded the experimental work.

II. LITERATURE REVIEW

S. K. Pushpa, Manjunath T. N., Srividhya[2]: analysed the Airline data using Apache Hive. Data has been loaded using SQOOP into HDFS. Three datasets namely Airport, Airline and Route, had been created and loaded into HDFS. Hive queries were executed and result was analyzed.

Dev Naomi.G, Karthigaa.M, Keerthana.B, Janani A [3]: identified crime detecting as one of the application where huge amount of data is massively increasing. With the increasing population and crime rates, data is getting difficult to analyse by traditional way. A model has been implemented using Hive to identify areas where crime rates are very high.

Dr. E. Laxmi Lydia1, Dr. M.Ben Swarup [4]: compared MapReduce, Pig and Hive. The matrices of comparison are: Performance and Development time. MapReduce had better performance but development time is more. Hive involved SQL like queries and Pig invokes short scripts.

Jay Mehta, Jongwook Woo [5]: applied Big Data Analytics to NYSE data. Top 10 companies having highest volume of stock traded were identified. Azure had been used for storing the historical data. Hive did the analysis of data. Single table was created using HiveQL to store data on HDFS. Author has shown the possibility that Big Data Hadoop and Hive can be adopted for financial industry.

Sanjeev Dhawan, Sanjay Rathee [6]: did a comparison of two Hadoop components Pig and Hive for Big Data Analytics.

A mapreduce job was created using Hive and then Pig. The job analyzed a big database to get results. The final results has shown that the analysis performed by both of the mapreduce machines was successful and the performance of both was nearly the same.

J.Ramsingh, Dr.V.Bhuvaneswari [7]: carried out Big Data analytics using pig script using Library data set.Pig provides a scripting language to use Hadoop's MapReduce library. It has been examined that pig script run in linear fashion because the execution time is directly proportional to the size of input data. But it can handle big databases in an efficient manner.

Anjali P P and Binu A [8]: conducted a comparative study based on processing network traffic data using Hadoop Pig and MapReduce. From this it has been derived that as the input file size increases in the multiples of x, the execution time for typical MapReduce also increases proportionally. However Hadoop Pig maintained a constant time at least for x upto 5 times. Pig was tested and proved to be advantageous in this aspect with a very low computational complexity.

Krati Bansal, Priyanka Chawla [9]: conducted a research study to identify shortcomings of Hadoop and benefits of Pig on Hadoop for analyzing Big Data. Apache Pig run on Hadoop by using Map Reduce for data processing. It uses HDFS for storing data. The Analysis has revealed that Pig is one of the most suitable scripting platforms for analyzing and structuring of Big Data with lesser development time.

III. RESULTS AND ANALYSIS

A. Dataset used

A public dataset,Provider Utilization and Payment Data Physician and Other Supplier Public Use File, has been prepared by the Centers for Medicare & Medicaid Services (CMS). This dataset provide information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. The dataset contains information about followings:

i) allowed payment amount and Medicare payment amount,

ii) submitted charges organized by National Provider Identifier

iii) Healthcare Common Procedure Coding System code

iv) Place of service.

Attributes of dataset are:

Table 1:Medicare Dataset

Variable	Format	Length	Label
npi	Char	10	National Provider Identifier
nppes_provider_last_org_name	Char	70	Last Name/Organization Name of the Provider
nppes_provider_first_name	Char	20	First Name of the Provider
nppes_provider_mi	Char	1	Middle Initial of the Provider
nppes_credentials	Char	20	Credentials of the Provider
nppes_provider_gender	Char	1	Gender of the Provider
nppes_entity_code	Char	1	Entity Type of the Provider
nppes_provider_street1	Char	55	Street Address 1 of the Provider
nppes_provider_street2	Char	55	Street Address 2 of the Provider
nppes_provider_city	Char	40	City of the Provider
nppes_provider_zip	Char	20	Zip Code of the Provider
nppes_provider_state	Char	2	State Code of the Provider
nppes_provider_country	Char	2	Country Code of the Provider
provider_type	Char	43	Provider Type of the Provider
medicare_participation_indicator	Char	1	Medicare Participation Indicator
place_of_Service	Char	1	Place of Service
hcpcs_code	Char	5	HCPCS Code
hcpcs_description	Char	256	HCPCS Description
hcpcs_drug_indicator	Char	1	Identifies HCPCS As Drug Included in the ASP Drug List
line_srvc_cnt	Num	8	Number of Services
bene_unique_cnt	Num	8	Number of Medicare Beneficiaries

bene_day_srvc_cnt	Num	8	Number of Distinct Medicare Beneficiary/Per Day Services
average_Medicare_allowed_amt	Num	8	Average Medicare Allowed Amount
average_submitted_chrg_amt	Num	8	Average Submitted Charge Amount
average_Medicare_payment_amt	Num	8	Average Medicare Payment Amount
average_Medicare_standard_amt	Num	8	Average Medicare Standardized Payment Amount

B. Experimental setup

The experimental work has been divided into four tasks, A,B,C and D,to evaluate the performance of tools.

For each task to be performed, different Hive queries and Pig scripts has been designed. These queries and scripts has been placed in different tables along with corresponding execution time.

Task A: What is the maximum and minimum average submitted amount by providers in different countries?

Table 1 maximum submitted amount along with service for each country

For Hive			
	Query	**Major Operations**	**Execution_ Time(sec)**
Sub-Task 1	hive> SELECT COUNTRY,MAX(AVG_SUBMITTED_CHRG_AMT)FROM MEDICARE GROUP BY COUNTRY ORDER BY COUNTRY;	GROUP,MAX,ORDER	111.559
For Pig			
Sub-Task 1	grunt> fltr = FOREACH medicare GENERATE country,avg_sub_chrg_amt;	FILTER	
	grunt> d = DISTINCT fltr;	DISTINCT	
	grunt> grp = GROUP d BY country;	GROUP	
	grunt> out = FOREACH grp GENERATE group,MAX(d.avg_sub_chrg_amt); grunt> dump out;	MAX	158

Table 3 minimum submitted amount along with service for each country

For Hive			
	Query	**Query**	**Execution_ Time(sec)**
Sub-task 2	hive> SELECT COUNTRY,Min(AVG_SUBMITTED_CHRG_AMT)FROM MEDICARE GROUP BY COUNTRY ORDER BY COUNTRY;	GROUP,MIN, ORDER	120.568
For Pig			
Sub-Task 2	grunt> fltr = FOREACH medicare GENERATE country,avg_sub_chrg_amt;	FILTER	
	grunt> d = DISTINCT fltr;	DISTINCT	
	grunt> grp = GROUP d BY country;	GROUP	
	grunt> out = FOREACH grp GENERATE group,MIN(d.avg_sub_chrg_amt);	MAX	
	grunt> dump out;		157

Task B: What are the fields in which providers charge the highest amount in different countries?

Table 2 Country-wise list of fields for which providers have submitted highest amount

For Hive			
	Query	**Major Operations**	**Execution_Time(sec)**
Sub-task 1	hive> SELECT COUNTRY,PROVIDER_TYPE,MAX(AVG_SUBMITTED_CHRG_AMT) FROM MEDICARE GROUP BY COUNTRY,PROVIDER_TYPE ORDER BY COUNTRY;	GROUP,MAX ,ORDER	229.148
For Pig			
Sub-task 1	grunt> slct = FOREACH medicare GENERATE country,provider_type,avg_sub_chrg_amt;		
	grunt> grp = GROUP slct BY (country,provider_type);	GROUP	
	grunt> out = ORDER(FOREACH grp GENERATE group.country,group.provider_type,MAX(slct.avg_sub_chrg_amt)) BY country;	ORDER, MAX	
	grunt> dump out;		204

Task C: What is the total number of beneficiaries being served per day in different cities of India? Which services has been provided along with Speciality fields?

Table 3 Number of beneficiaries being served per day in each speciality field

For Hive			
	Query	**Major Operations**	**Execution_Time(sec)**
Sub-task 1	hive> SELECT CITY,PROVIDER_TYPE,SUM(BENE_DAY_SRVC_CNT) FROM MEDICARE WHERE COUNTRY='IN' GROUP BY CITY,PROVIDER_TYPE ORDER BY CITY;	GROUP,SUM ,ORDER	93.534
For Pig			
Sub-task 1	grunt> india = FILTER medicare BY country == 'IN';	FILTER	
	grunt> city_grp = FOREACH india GENERATE city,provider_type,bene_day_srvc_cnt;		
	grunt> out = ORDER (FOREACH (GROUP city_grp by (city,provider_type)) GENERATE group.city,group.provider_type,SUM(city_grp.bene_day_srvc_cnt)) BY city;	ORDER,GROUP,SUM	
	grunt> dump out;		152

Table 4 Services in different cities having maximum number of beneficiaries served per day

For Hive			
	Query	**Major Operations**	**Execution_Time(sec)**
Sub-Task 2	hive> SELECT HCPCS_DESCRIPTION,MAX(BENE_DAY_SRVC_CNT) AS c1 FROM MEDICARE WHERE CITY='BANGALORE' AND PROVIDER_TYPE='Internal Medicine' GROUP BY HCPCS_DESCRIPTION ORDER BY c1 DESC;	MAX,GROUP ,ORDER	121.27
	hive> SELECT HCPCS_DESCRIPTION,MAX(BENE_DAY_SRVC_CNT) AS c1 FROM MEDICARE WHERE CITY='JAIPUR' AND PROVIDER_TYPE='Neurology' GROUP BY HCPCS_DESCRIPTION ORDER BY c1 DESC;	MAX,GROUP ,ORDER	122.064
	hive> SELECT HCPCS_DESCRIPTION,MAX(BENE_DAY_SRVC_CNT) AS c1 FROM MEDICARE WHERE CITY='MUMBAI' AND PROVIDER_TYPE='Infectious Disease' GROUP BY HCPCS_DESCRIPTION ORDER BY c1 DESC;	MAX,GROUP ,ORDER	107.29

For Pig			
Sub-Task 2	grunt> prvd_srvc = FOREACH india GENERATE city,provider_type,hcpcs_desc,bene_day_srvc_cnt;		
	grunt> srvc_grp = GROUP prvd_srvc BY (city,provider_type,hcpcs_desc);	GROUP	
	grunt> rslt1 = ORDER(FOREACH srvc_grp GENERATE group.city,group.provider_type,prvd_srvc.hcpcs_desc,MAX(prvd_srvc.bene_day_srvc_cnt) AS col) BY col DESC;	ORDER,MAX	216

Task D: What is the maximum and minimum charged amount by providers? How much amount does Medicare allow for that service in India?

Table 5 Highest submitted amount and corresponding Medicare allowed amount

For Hive			
	Query	**Major Operations**	**Execution_Time(sec)**
Sub-task 1	hive> SELECT MAX(AVG_SUBMITTED_CHRG_AMT),CITY FROM MEDICARE WHERE COUNTRY='IN' GROUP BY CITY;	MAX,GROUP	55.455
	SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=4195.4545455;		51.63
	hive> SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=408.42105263;		46.253
	SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=307.0;		50.234
For Pig			
Sub-task 1	grunt>filter_in = FOREACH india GENERATE city,provider_type,hcpcs_desc,avg_sub_chrg_amt,avg_medi_allw_amt;	FILTER	
	grunt> grp = GROUP filter_in BY city;	GROUP	
	grunt> max = FOREACH grp GENERATE group,MAX(filter_in.avg_sub_chrg_amt);	MAX	
	grunt> dump max;		164
	grunt> rslt = FILTER filter_in BY city == 'BANGALORE' AND avg_sub_chrg_amt == 307.0;	FILTER	
	grunt> dump rslt;		96
	grunt> rslt1 = FILTER filter_in BY city == 'JAIPUR' AND avg_sub_chrg_amt == 4195.4546;	FILTER	
	grunt> dump rslt1;		194
	grunt> rslt = FILTER filter_in BY city == 'MUMBAI' AND avg_sub_chrg_amt == 408.42105;	FILTER	191
	grunt> dump rslt;		

Table 6 Lowest submitted amount and corresponding Medicare allowed amount

For Hive

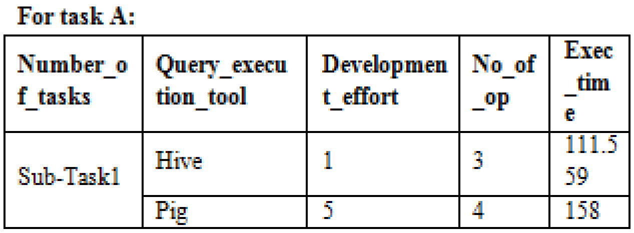
	Query	**Major Operations**	**Execution_Time(sec)**
Sub-Task 2	hive> SELECT CITY,MIN(AVG_SUBMITTED_CHRG_AMT) FROM MEDICARE WHERE COUNTRY='IN'GROUP BY CITY;	MIN,GROUP	88.385
	hive> SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=0.4686311787;		57.373
	hive> SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=204.0;		49.352
	hive> SELECT COUNTRY,STATE,CITY,PROVIDER_TYPE,HCPCS_DESCRIPTION,AVG_SUBMITTED_CHRG_AMT,AVG_MEDICARE_ALLOWED_AMT FROM MEDICARE WHERE COUNTRY='IN' AND AVG_SUBMITTED_CHRG_AMT=110.0;		48.587
For Pig			
Task 2	grunt> filter_in = FOREACH india GENERATE		
	city,provider_type,hcpcs_desc,avg_sub_chrg_amt,avg_medi_allw_amt;	FILTER	
	grunt> grp = GROUP filter_in BY city;	GROUP	
	grunt> min = FOREACH grp GENERATE group,MIN(filter_in.avg_sub_chrg_amt);	MIN	128
	grunt> dump min;		
	grunt> rslt = FILTER filter_in BY city == 'BANGALORE' AND	FILTER	189
	avg_sub_chrg_amt==204.0;		
	grunt> dump rslt;	FILTER	127
	grunt> rslt1 = FILTER filter_in BY city == 'JAIPUR' AND		
	avg_sub_chrg_amt==0.46863118;	FILTER	148
	grunt> dump rslt1;		
	grunt> rslt = FILTER filter_in BY city == 'MUMBAI' AND avg_sub_chrg_amt ==		
	110.0;		
	grunt> dump rslt;		

C. Analysis

Following is the list of parameters taken for comparison of Hive and Pig on MapReduce engine:

1) *Development effort : Number of queries/lines of script written to perform a task*
2) *Number of major operations: Operations to process data. These include grouping,filtering,aggregating the data.*
3) *Execution time: Time in seconds to perform various operations on data.*

After executing the queries to perform above mentioned tasks, the values of different parameters has been calculated. The values has been kept in the form of tables. Then the performance of Pig and Hive has been analyzed on the basis of these values.



For task A:

Number_of_tasks	**Query_execution_tool**	**Development_effort**	**No_of_op**	**Exec_time**
Sub-Task1	Hive	1	3	111.559
	Pig	5	4	158

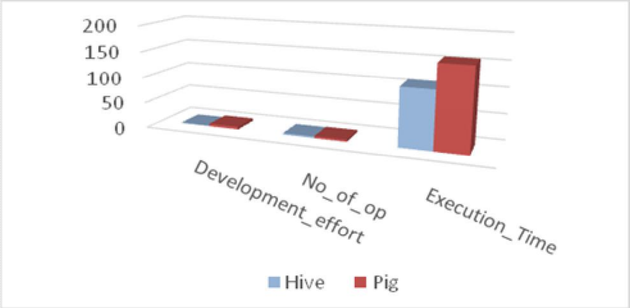


Fig 1.1 Task A (Sub-Task 1)

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	Exec_time
Sub-Task2	Hive	1	3	120.568
	Pig	5	4	157

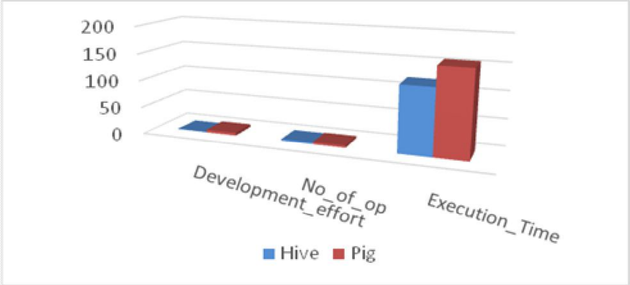


Fig 1.2 Task A (Sub-Task 2)

FromFig 1.1 and 1.2 it is observed that performance of Hive is much better than Pig. With hive only a single query needs to be written to group countries and then finding highest charged amount for service. Whereas in pig 5 pig scripts has been written to do the same task. Also the run time of query is less in hive.

For task B:

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	Execution_Time
Sub-Task 1	Hive	1	3	229.148
	Pig	4	3	204

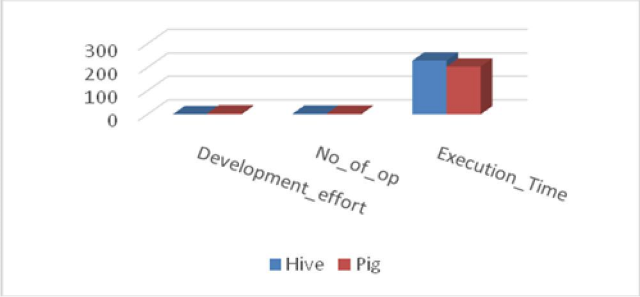


Fig 1.3 Task B(Sub-Task 1)

While performing Task B, query execution time of hive is more than pig. But hive require only a single query to be written while pig require 4 pig scripts to do the job. Number of major operations are same for both the tools.

For task C:

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	exec_time
Sub-Sub-task 1	Hive	1	3	93.534
	Pig	4	4	152

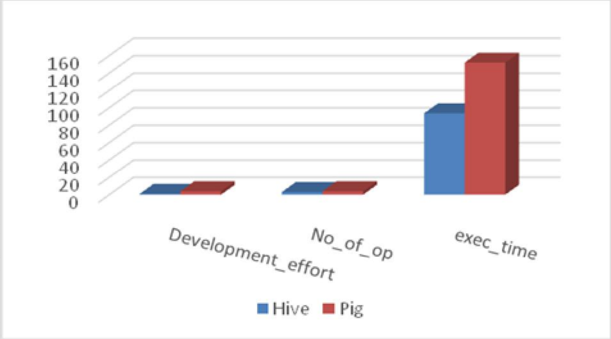


Fig 1.4 Task C (Sub-Task 1)

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	exec_time
Sub-task 2	Hive	3	9	350.624
	Pig	3	3	216

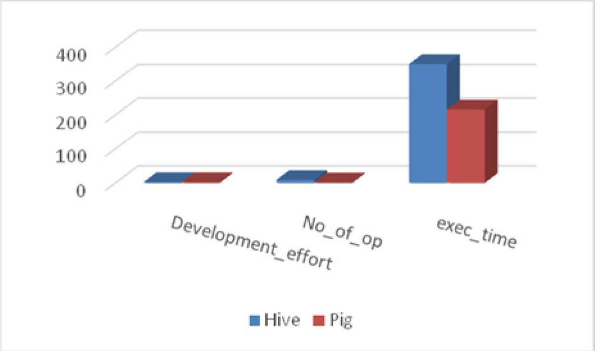
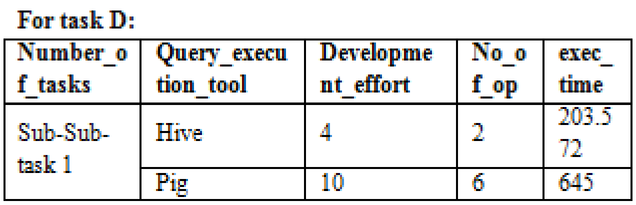


Fig 1.5 Task C (Sub-Task 2)

For doing Task C hive perform better for Sub-Task 1in terms of execution time and development effort. Execution time to run query is very low in comparison with pig. However for doing SubTask 2 pig has relatively low execution time. Also pig require only 2 operation i.e. GROUP,MAX and ORDER while hive need 9 operations which include 3 GROUP, 3 MAX and 3 ORDER operators. So performance of Pig is high for Sub-Task 2.



For task D:

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	exec_time
Sub-Sub-task 1	Hive	4	2	203.572
	Pig	10	6	645

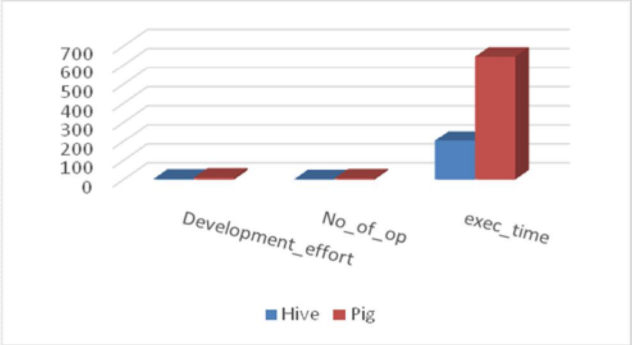


Fig 1.6 Task D (Sub-Task 1)

Number_of_tasks	Query_execution_tool	Development_effort	No_of_op	exec_time
Sub-task 2	Hive	4	2	243.697
	Pig	10	6	592

For performing Task D, performance of pig is worst. Execution time for both the sub tasks is very high in contrast to hive. 9 pig scripts has to be written whereas hive require single query for doing the same task.

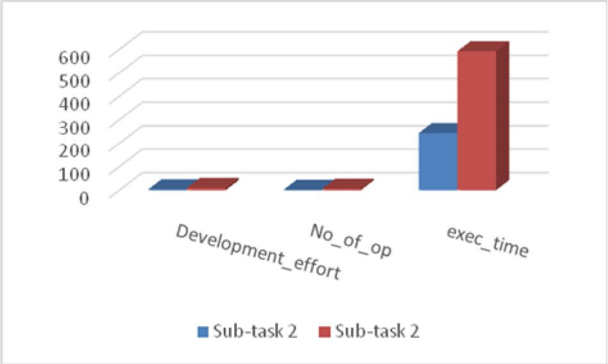


Fig 1.7 Task D (Sub-Task 2)

After performing different tasks on Medicare dataset, it can be stated that :

For Task A hive was better.
For Task B pig performed well in terms of execution time but development effort was less for hive.
For Task C performance of hive was better in Sub-Task1 while in Sub-Task 2 pig performed very well.
For Task D hive was much better than pig.

Overall performance of Apache hive is much better than Apache Pig on MapReduce engine. For performingseveral tasks, one need to write very few Hive queries in contrast to Pig scripts. For processing the data, Hive require less number of operators than Pig. Also while dumping the information, Pig takes more time than Hive.

IV. CONCLUSION

Large amount of digital data is being generated every second. The bulk of data itself cannot create any value. So it need to be processed by applying some analytical techniques. Apache Pig and Apache Hive are the tools to analyse the data. Each tool have its own advantages over another. Hive require the data to be in tabular format. However Pig do not require any schema to be defined for the data. So it can work with any type of data. From the experimental work done in this paper, it can be stated that HiveQL is similar to SQL and do not require strong hand in programming. PigLatin also do not need programming background but is bit difficult to manage. For each operation in PigLatin a temporary table is created. With increasing number of operations, tables increase as well. Parameters considered for comparison of these tools are: development effort, number of operations and execution time for each query. From the experimental results it can be concluded that Hive engine takes less time to run queries. Also few queries needed to be written to do a task in Hive .So the performance of Hive is much better than Pig on MapReduce engine.

REFERENCES

[1] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[2] S. K. Pushpa, Manjunath T. N., Srividhya,"Analysis of Airport Data using Hadoop-Hive: A Case Study", International Journal of Computer Applications, 2016

[3] Dev Naomi.G, Karthigaa.M, Keerthana.B, Janani A ,"Big Data Prediction on Crime Detection", Global Research and Development Journal for Engineering | National Conference on Computational Intelligence Systems, Mar 2017

[4] Dr. E. Laxmi Lydia,Dr. M.Ben Swarup," Analysis of Big data through Hadoop Ecosystem Components like Flume, MapReduce, Pig and Hive", International Journal of Computer Science Engineering (IJCSE), Vol. 5 No.01 Jan 2016,

[5] Jay Mehta, Jongwook Woo," Big Data Analysis of Historical Stock Data Using HIVE", ARPN Journal of Systems and Software, Vol. 5, No. 2, Aug 2015.

[6] Sanjeev Dhawan, Sanjay Rathee, "Big Data Analytics using Hadoop Component Like Hive and Pig", American

International Journal of Researching Science, Technology, Engineering & Mathematics, Mar-May 2013

[7] J.Ramsingh , Dr.V.Bhuvaneswari,” An Insight on Big Data Analytics Using Pig Script”, International Journal of Emerging Trends & Technology in Computer Science, Vol 4, Issue 6, Nov - Dec 2015

[8] Anjali P P and Binu A, “A Comparative Survey Based on Processing Network Traffic Data Using Hadoop Pig and Typical Mapreduce”, International Journal of Computer Science & Engineering Survey,Vol 5, No.1, Feb 2014

[9] Krati Bansal, Priyanka Chawla,” A Study of Big Data Analysis Using Apache Pig”, International Journal of Control Theory and Applications, 2016, pp. 8665-8672