International Journal of Engineering & Technology, 7 (2.27) (2018) 1-6

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Metamorphosis of data (small to big) and the comparative study of techniques (HADOOP, HIVE and PIG) to handle big data

Rapinder Kaur [1] *, Vaishali Chauhan [1], Urvashi Mittal [1]

[1] *Department of Computer Science Engineering, Chandigarh University, Gharuan, S.A.S Nagar Mohali, Punjab (140413) India.*
**Corresponding author E-mail: rapindersaini94@gmail.com*

Abstract

Immoderate amount of data is being generated everyday across the world via miscellaneous sources or fields which create issues to the users. Due to this rapid growth, the crucial issue is to analyse the big data with the help of traditional data processing tactics. Structured data is not the peerless but moreover unstructured data and semi-structured data charge up the supplementary consequences to handle this voluminous data. As in this gigantic bulk of data highly advantageous information is hidden which can be good for what ails the individual, group or organization and for adding up to more sophisticated or valuable decisions. So in order to deal with this many new tools and techniques have been excogitated. These tools can analyse the large volume of data being generated at unprecedented speed. This paper shows the comparative study of some of the data analytics techniques which can untangle the big data analytics issues by examining it in more précised manner. The contrast study of Hadoop, Hive and Pig has been illustrated which covers the working of these techniques.

Keywords: *Big Data; Hadoop, Hive; Map Reduce; Pig.*

1. Introduction

Big data denotes the immense amount of data that has been propagated due to immeasurable reasons. The multitude of the computed data is piling up exponentially that cannot be directed, managed or organised via traditional or ancestral databases. Organisations like logistics, retail, finance, health, shipping, farming and banking, etc. are over whelming more data[1]. Marked out Big Data as an accumulation of very enormous data sets acquiring tremendous assortments of genres in order that it eventually is difficult to refine or process by using contemporary data processing onsets or traditional data processing means. Big data is being accumulated by the internet of things as the advancement of the resources used by the humans has become enormous. The US National Institute of Science and Technology (NIST) [2] epitomize big data as a collection of "extensive datasets - primarily in the characteristics of volume, variety, velocity, and/or variability - that require a scalable architecture for efficient storage, manipulation, and analysis." Nowadays the data is originating by the human, computers also generate data and moreover machines are also involved in this race. This is increasing the magnitude of data. In the historical time there were not any tactics and operations that were interconnected so the size data was small. But with the advancement in the science, the high tech mechanization has been developed, new machinery and technologies has evolved which are increasing the data size from small into big. The elevation in the approaches and techniques of big data has added those projects in the list of feasible ones which appeared to be impossible few years back [3]. [4] Provided an evidence for transforming healthcare data into useful insights for better decision making. This advancement has come not only in the single field but in every field. From a small scale organisation to a big organisation and from a small device to a big device which is connected via some network is evolving data at a rapid rate. Social media also adds to the immeasurable growth of data.

The new procreations of mobile devices are substantially powerful as the acquire gigabytes of memory and multi-core processors. These procreated devices exhibits intricate applications and sensors which are endowed of generating and accumulating enormous of megabytes of data [5]. It is the big data that originated as a technology which is proficient in assembling and transforming the colossal and divergent figures of data, providing organizations with meaningful insights for deriving improved results. Big data is accustomed to delineate technologies and techniques which are used to store, manage, distribute and analyze huge data sheets [6]. Earlier, a single computer processor can refine the data but due to large evolutions of data multiple processors are required as terrible chunk of data is over whelming the CPU. So now the inhabitants are making use of numerous multiple processors for the refining and processing of the data. Big data is refined by adopting the distributed data handling technique which comprises of multiple servers each having small components and these components carrying processor to refine the data. Figure 1 shows types of big data.

- Every single zone from government to organisations and business, trends, actions, machines, humans, sensors are giving origin to data.
- All the data has admirable aspects or features.
- Not gets accommodated into a solitary mainframe.
- Data has to be reserved in distributed appearance.
- Distributed data accumulation=Accelerated data figuring or computation.

Copyright © 2018 Rapinder Kaur et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

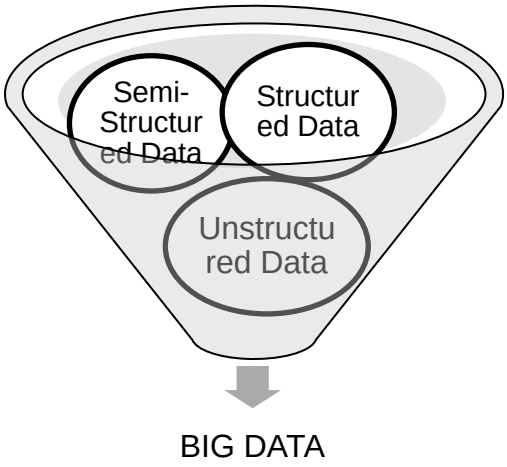


Fig. 1: Types of Data.

Characteristics of Big Data (10 V's of big data)) is shown in Figure 2.

1) Volume: The most conspicuous characteristic of the big data is volume. Majority of the data today has been commenced in the past copulated years. The data that is being progressed nowadays is over whelming to some degree [7]. The allied value of every data point declines proportionately due to increment in size of data [8].
   Exemplar: In an isolated minute, videos of almost 300 hours are uploaded on YouTube [9].
2) Variety: when we stack up with big data, it not only encompasses of structured data but also mostly comprises of unstructured data and semi-structured data. this adds up to the variety. multifarious big data appears to be unstructured data that comprises of images, videos repositories; social media updates and moreover supplementary text formats like click data, log repositories, data from sensors and machines, etc. [7][9].
3) Velocity: velocity associates to the rapidness of the engenderment, production and procreation of data.
   - Facebook affirms 600 terabytes of data is approaching every day.
   - Google take care of beyond 40,000 search interrogatories in an isolated second [7], [9].
4) Variability: Variability in big data's interest put down to lean distinct things-
   - One is data unconformity. For the productive analytics, the inconsistencies ought to be erect by anomaly and approaches of outlier detection.
   - Other is the rapidness of unconformity by virtue of which data is primed to the database.
   - Variability can also originate by virtue of aggregation of dimensions of data being connected with miscellaneous discordant data types [9].
5) Validity: validity ascribes that how précised and appropriate is the data for its clear-cut usage. It is roughly calculated that 60 percentile of data scientists spend their time in the data refining and clarifications. For the sake of positive data analytics, the data has to be governed and subjugated to guarantee steady data prediction and metadata [9].
6) Veracity: veracity ascribes to the authenticity or reliability of the data authorship, its frame of reference and how consequential it is to the interpretation stationed on it. Having insights about data, veracity contributes to the prominent awareness of the hazards related to analysis and business opinions validated on the data set [9].
7) Vulnerability: the enormous data engenders new security apprehensions. Grievously big data is being infracted [9].
8) Value: value ascribes acquiring usefulness value from data. Abundant or meaningful interpretations can be determined from large data [9].

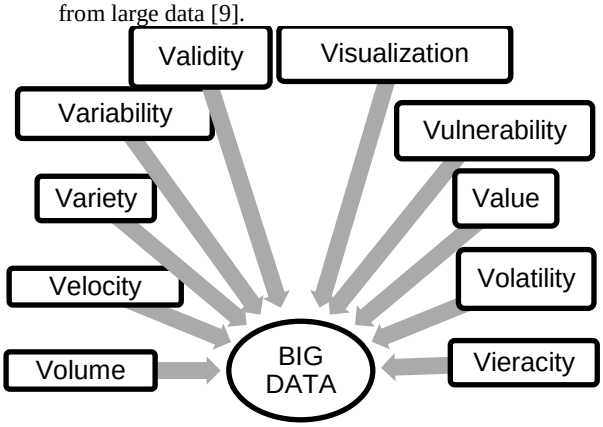


Fig. 2: 10 V is of Big Data.

9) Volatility: volatility describes the span for which data has to be clutched or retained. Also, ascribes that when the data has to be affirmed as documented or chronicle, extraneous or inappropriate? [9]
10) Visualization: this ascribes that how the data is to be featured. Distinct ways are data assembling or clustering, sunbursts, circular network sketches or cone trees [9].

2. Hadoop

Hadoop was primarily concocted by yahoo. Yahoo was the inceptive who solicited to concoct Hadoop in order to elucidate the contention of big data within petite fractions so that the refining or transformation can be brought about in aligned tenor [10]. [11] provided an AES approach for decuring the large amount of data from unauthorized access as the data.Conventionally, hadoop is an approachable antecedent, java-based programming formation that can be utilised for the processing or refining and storage of very enormous data in a distributed computing tenor. To escalate from lone servers to large multitude of machine that accords local computation and storage, hadoop is formulated. Instead of reckon on hardware to remit high-availability, the breakdowns and inadequacies are taken care of and ascertained by the library intrinsically [12].

Characteristics of hadoop are:

- High performance and scalable: the apportioned processing of data local to exclusive nodes in a clump capacitates hadoop to accumulate, govern or maintain, refine or process and analyse data at petabyte extent [13].
- Flexible: any format of data can be refined, stored or govern. In addition to structured data, unstructured data and semi-structured data is stored and then, after the process of parsing the schema can be applied to the data when read [13].
- Reliable: archetype of data in the bunch implies to reliable storage of data on the bunch of machines regardless of machine's lack of success. Even if machinery gravitates, then also data is reliable.
- Economical: apache Hadoop is not exceedingly expensive. It favours stupendous cost saving.

An apache Hadoop contributes of:

- Hadoop distributed file system (HDFS).
- Hadoop yarn.
- Hadoop map reduce.
- Hadoop common utilities.
- Apache hive.
- Hbase.
- Apache pig.
- Zookeeper.

- Spark.
- Oozie.
- Sqoop.

Hadoop distributed file system (HDFS)

hdfs was commenced via distributed file system. HDFS is eminently fault-tolerant and outlined using hardware acquiring deficient cost [14]. The hdfs allows the data of the system and applications to be stored separately [15]. Figure 3 shows Hdfs Architecture. It capacitates-

- Repository of elephantine hunk of information.
- Escalate incrementally.
- Survive the breakdown of consequential parts of the pedestal of storage without losing data.

Hdfs is file system constituent of hadoop. The approaching files are partitioned into pieces called blocks. Hdfs deposits data over various thousands of servers. Three copies of single file are stored by duplicating each piece to three distinct servers [15]. Therefore, if one fails the copy is stored safely at the other and one can access it without any loss of data. The capaciousness of block is configurable but default capacity is 64 megabytes.

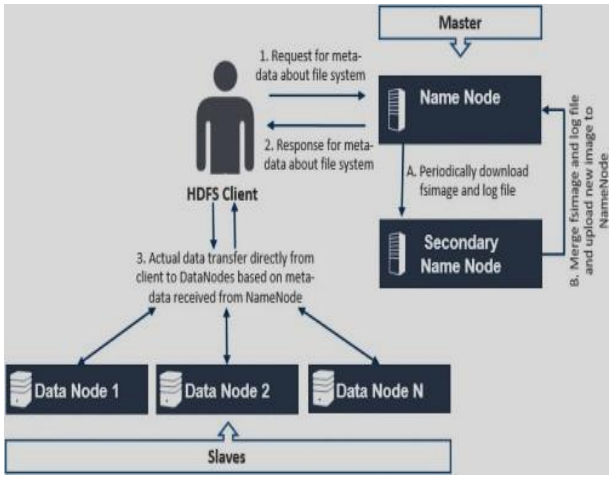


Fig. 3: Hdfs Architecture [16].

The foremost segments of hdfs are:

1) Name node: name node is the principal node that encloses apprehensions about the file system of Hadoop [15]. It is the sovereign node that perpetuates, instructs and governs the blocks that resides on the data nodes. It is assigned with the chore of inscribing metadata, aspects and explicit locations of the files and data segments in data nodes. Every respective change is inscribed that transpire in the file system metadata. Example: it will at once inscribe if a file gets deleted in hdfsTo clinch the survival of data node, name node gets hand on the regular heartbeat and block account of data nodes. Replication factor is inscribed for all the blocks by name node [17].
1) Data node: data node is the labourer node. Deriving upon sufficiency and interpretations hadoop may embrace more than lone data node. The tangible data resides in the data node [15]. Times to time the heartbeats are sent to the name node by this. Accumulating a block in hdfs and serving as a stage for running jobs are two predominant chores of data node [17].
2) Secondary name node: this node is the adherent node of the sovereign node. Rather than carrying itself as the auxiliary of the name node, it invariably reads the metadata coming out of the ram of name node and writes the same on the system of file or the hard disk [17].
3) Blocks: in hdfs, the data is dispersed over the data nodes as blocks. The minimal unending location on the hard drive at which the data reside are the blocks [17].
4) Hdfs clients: these are generally labelled as the edge nodes consistently. It operates as the interface between name node and data node [17].

2.1. Map reduce

It is a programming standard utilised for refining, processing and engendering enormous data sets. The data sets to be inputted are gathered in a cluster of segregations in a distributed file system redistributed on each lump in the cluster. Then the program is dragged into a distributed processing architecture and is then executed [13]. Map Reduce is shown in Figure 4. It has two sections:

1) Map stage: map task refines and processes data in parallelism way when the system divides the data inputs into multitude pieces. These miscellaneous pieces are assigned with map task. It interprets the inputs and generates intermediate output [18].
2) Reduce stage: after shuffling by the framework, the intermediate output is put onto the reduce task and final output is generated [18].

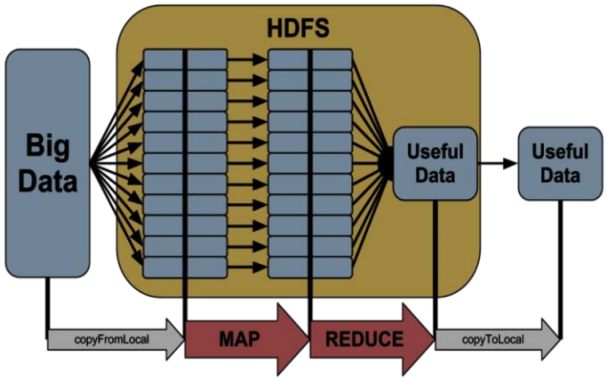


Fig. 4: Map Reduce Architecture [19].

2.2. Apache hive

Apache hive is data warehouse software put together on top of apache hadoop for proffering summarization of data, questionaries' and analysis. Hive comes across with sql like interaction [18]. Queries are run across the data stored in repositories and file systems that accommodated in hadoop. It assists the progress of reading, writing and managing enormous data sets that are resided in distributed manner. Hive is protractible query language; scalability is high and is quite fast. Figure 5 shows Apache Hive Architecture. Features:

- It endorses structured data.
- In order to run hive HADOOP has to be started.
- Rigorous to debug.
- Metadata is deposited in database.
- Endorses user-defined functions.
- Intermediate data is deposited in manually concocted tables.
- HQL query language is utilised.
- Executes quickly.

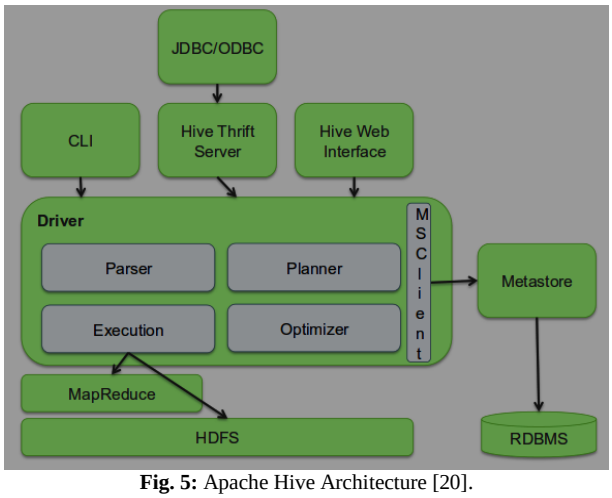


Fig. 5: Apache Hive Architecture [20].

From the name itself, the perception can be made that the user interface is something through which a user get across the system. On the whole, it comforts the user to endure their queries and bring off other undertakings to the system.

The user gets affiliated precisely to the hive drivers when it comes with cli (hive terminal) or web gui (ambari server) but the user has to affiliate through api to the hive driver when it comes with jdbc/odbc (jdbc program) [18].

Once the task of users is received by the hive driver, it sends it to the hadoop architecture. then the hadoop after receiving the task using name node, data node, task tracker and job tracker, divides the task and then send it to map reduce architecture. The compiler parses the query received from the user [21] [22].

2.3. Apache pig

Apache pig that runs on apache hadoop endorses a high level platform for creation of programs. Pig latin is the language supported by this platform [18]. The jobs or tasks of hadoop in the pig are executed in map reduce, apache tez and apache spark. It was initiated by yahoo. Any type of data can be handled or refined using pig. It requires lesser time for the mapper and reducer tasks and has more focus on the data sets analysis. More level abstractions are added by the pig in the refining of data and also it makes the data refining easy [18]. Terabytes of data can be refined using pig [7]. Pig is best suitable for the data in the form of semi-structured, for programming, can be utilized as the procedural language, it does not support apportionment and do not have inscribed metadata of database [23]. Apache pig architecture is shown in Figure 6.

Features:

- Both types of data can be refined whether structured or unstructured.
- Tables need not to be created.
- Metadata is not endorsed.
- Hadoop not required to be started.
- Functions on the client side of cluster.
- Data can be loaded conclusively and quickly.
- Commodious for programmers and developers of software.

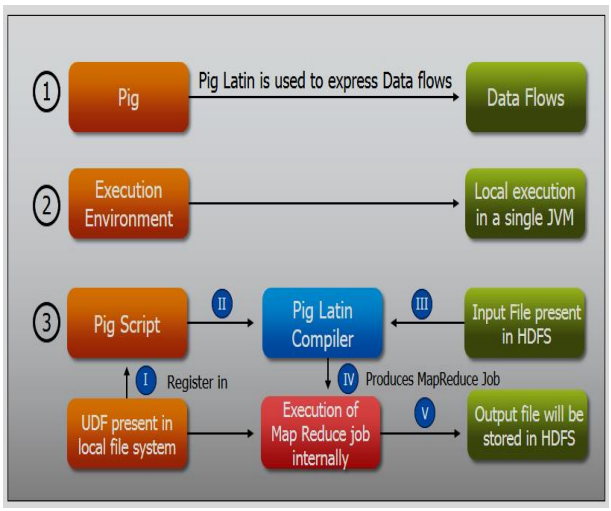


Fig. 6: Apache Pig Architecture [24].

With the help of user-defined functions (udf), pig swallows all the data from all the sources like files, streams that has been burdened in the data store once it has been commuted according to some rules [21]. Then after the data has been burdened into the pig, the next expedient is to transform that data. Various undertakings like select, iterations are accomplished. Soon after the outcome is burdened into the hadoop data file system and then to map reduce for additional execution. Pig more converges on the data than the execution.

The pig latin program is applied to the input data, which accomplishes or operates on the input to assemble output [21]. The interpretation of the data flow into executable portrayal is epitomized by these operations. This comes beneath the pig execution ambiance. Completion of these is handled by the hadoop.

3. Result and discussion

Experimentation is done using national statistics postcode lookup uk dataset. The size of this dataset is 691 mb and it consists 17, 56,258 records. This dataset contains the national statistics postcode lookup (nspl) for the United Kingdom. The nspl relates current postcodes to a range of current statutory administrative, electoral, health and other statistical geographies via 'best-fit' allocation from the 2011 census output areas. it supports the production of area based statistics from post coded data. The nspl is produced by ons geography, which provides geographic support to the office for national statistics (ons) and geographic services used by other organisations. The experimentation on the dataset has been performed 5 times in order to calculate the proper results.

3.1. Evaluation using apache hive

Create table:

The query used is-

create table postcodes_uk(postcode_1 string,postcode_2 string,postcode_3string,date_introduced string,user_type int,easting int,northing int,positional_quality int,county_code string,county_namestring,local_authority_code string,local_authority_name string,ward_code string,ward_name string,country_code string,country_name string,region_code string,region_name string,parliamentary _constituency_code string,parliamentary_constituency_name string,european_electoral_region_code string,european_electoral_region_name string,primary_care_trust_code string,primary_care_trust_name string,lower_super_output_area_code string,lower_super_output_area_name string,middle_super_output_area_code string,middle_super_output_area_name string,output_area_classification_code

string,output_area_classification_name string,longitude float,latitude float,spatial_accuracy string,last_uploaded string,location string,socrata_id int) row format delimited fields terminated by ',' ;
Create table is shown in Figure 7.

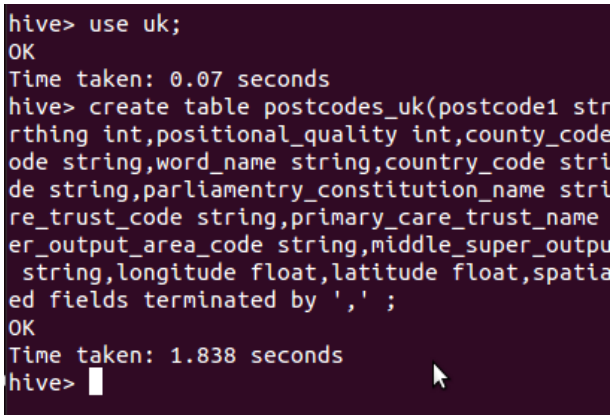


Fig. 7: Create Table.

Load data into table:
Loading of data is shown in Figure 8.
The query used is-
Load data local inpath
'/home/hduser/desktop/postcode_uk.txt' overwrite into table post-codes_uk

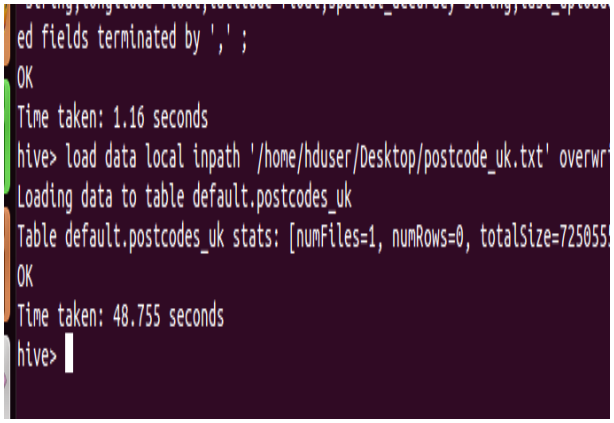


Fig. 8: Load Data.

Select operation: shown in figure 9.
The query used is-
Select * from postcodes_uk;

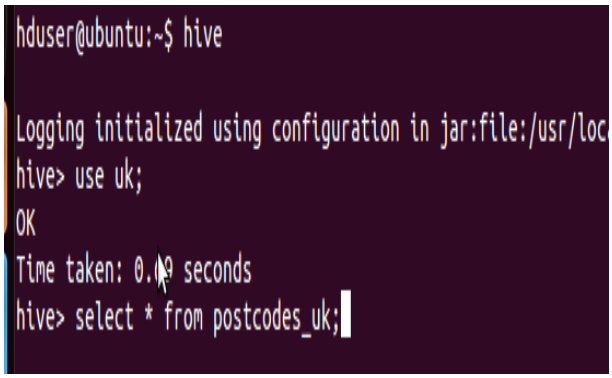


Fig. 9: Select Operation.

Output: shown in Figure 10.

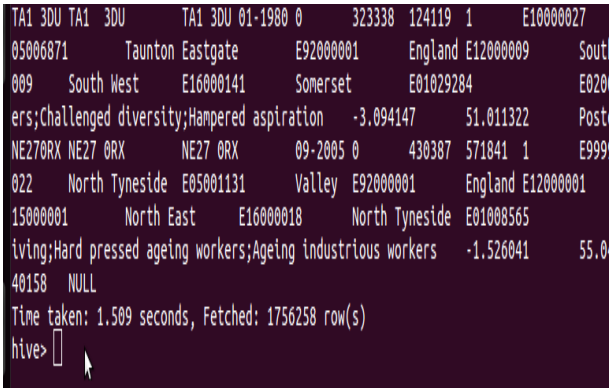


Fig. 10: Hive Result.

3.2. Evaluation using apache pig

Load data into table
The query used is-
postcode= load '/home/hduser/postcode_uk.txt' using pigstorage (',')as
(post-code_1:chararray,postcode_2:chararray,postcode_3:chararray,date_introduced:chararray,user_type:int,easting:int,northing:int,positional_quality:int,county_code:chararray,county_name:chararray,local_authority_code:chararray,local_authority_name:chararray,ward_code:chararray,ward_name:chararray,country_code:chararray,coun-try_name:chararray,region_code:chararray,region_name:chararray,parliamentary_constituency_code:chararray,parliamentary_constituen-cy_name:chararray,european_electoral_region_code:chararray,europe-an_electoral_region_name:chararray,primary_care_trust_code:charar-ray,primary_care_trust_name:chararray,lower_super_output_area_code:chararray,lower_super_output_area_name:chararray,middle_su-per_output_area_code:chararray,middle_super_output_area_name:chararray,output_area_classification_code:chararray,output_area_classifica-tion_name:chararray,longitude:float,latitude:float,spatial_accuracy:chararray,last_uploaded:chararray,location:chararray,socrata_id:int);
Figure 11 shows Loading of data.

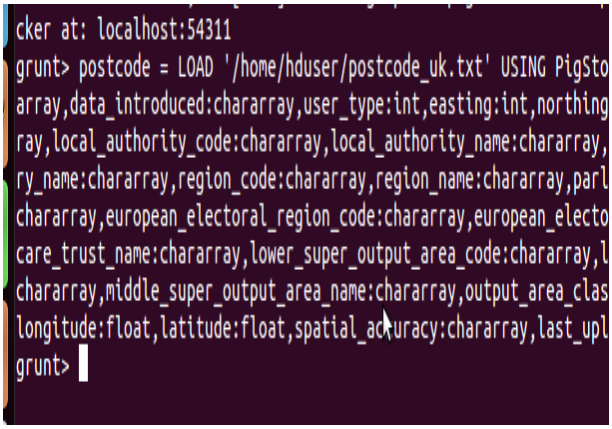


Fig. 11: Load Data.

Select operation: Shown in Figure 12
The query used is-
Dump postcode;

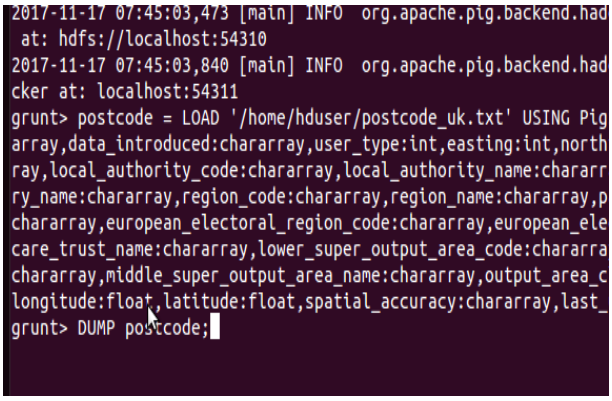


Fig. 12: Select Operation.

Output: shown in figure 13.



Fig. 13: Pig Result.

4. Performance analysis

The experimentation of hive and pig shows that the hive take less time than the pig. A comparison between hive and pig has been carried out, to see how well each platform performs during data analysis. The evaluation has been performed 5 times and the results average has been taken. The execution time of pig and hive are different in terms of seconds. It is quite evident that hive outperformed pig for loading and querying the dataset.
Hive - 1.509 seconds
Pig – 2.5 seconds

5. Conclusion

The augmentation in the capaciousness of data from cramped to copious has also prevailed to the metamorphosis in the technologies or approaches. Beginning from tiny technologies to contend the data has now reached to the large technologies to contend the enormous data that is not even in the meticulous format i.e. in the structured format. So to contend such data the technologies that has been commenced in the field are discussed which shows that how the data contending is brought about.

References

[1] C. L. Philip Chen and C.-Y. Zhang (2014), Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, vol. 275, *Information Sciences*, pp. 314–347. https://doi.org/10.1016/j.ins.2014.01.015.

[2] *National Institute of Standards and Technology (NIST) Special Publication* 1500-1, NIST Big Data Interoperability Framework: Volume 1, Definitions Final Version 1.

[3] Matthias Volk, Sascha Bosse and Klaus Turowski (2017), Providing Clarity on Big Data Technologies: A Structured Literature Review, *IEEE 19th Conference on Business Informatics*, pp. 388-397.

[4] Praveena, M., & Kameswara Rao, M. (2018). Survey on Big data analytics in Healthcare Domain. *International Journal of Engineering & Technology*, *7*(2.7), 919-925. doi:http://dx.doi.org/10.14419/ijet.v7i2.7.11097.

[5] Johnu George et al. (2014), Hadoop MapReduce for Tactical Clouds, 2014, *IEEE 3rd International Conference on Cloud Networking*.

[6] Virk, M., Chauhan, V., & Mittal, U. (2018). Analysis and Visualization of Data Assimilating Hive and COGNOS Insight 10.2.2. *International Journal of Engineering & Technology*, *7*(2.6), 318-322. doi:http://dx.doi.org/10.14419/ijet.v7i2.6.11271.

[7] Varsha B.Bobade (2016), Survey Paper on Big Data and Hadoop, Volume: 03, *International Research Journal of Engineering and Technology (IRJET)*, pp. 861-863, Issue: 01 | Jan-2016.

[8] D. Laney (2001), 3D data management: Controlling data volume, velocity and variety, vol. 6, *META Group Research Note*, p. 70.

[9] V is of Big Data, https://tdwi.org/Articles/2017/02/08/10-Vs-of-Big-Data.aspx?Page=1.

[10] Kadhar Basha J and Dr. M. Balamurugan (2017), A Review on Hive and Pig, Vol. 3, *ISSN* (Online): 2456-5717, Special Issue 39, May 2017.

[11] Hanisah Kamaruzaman, S., Nor Shuhadah Wan Nik, W., Afendee Mohamed, M., & Mohamad, Z. (2018). Design and Implementation of Data-at-Rest Encryption for Hadoop. *International Journal of Engineering & Technology*, *7*(2.15), 54-57. doi:http://dx.doi.org/10.14419/ijet.v7i2.15.11212.

[12] Apache hadoop, http://hadoop.apache.org/.

[13] Andrew Pavlo (2009), a Comparison of Approaches to Large-Scale Data Analysis, *SIGMOD*. https://doi.org/10.1145/1559845.1559865.

[14] Harshawardhan S. Bhosale and Prof. Devendra P. Gadekar (2014), A Review Paper on Big Data and Hadoop, Volume 4, *International Journal of Scientific and Research Publications*, Issue 10, October 2014.

[15] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler (2010), the Hadoop Distributed File System, Sunnyvale, California USA, *IEEE*. https://doi.org/10.1109/MSST.2010.5496972.

[16] HDFS Architecture, http://www.informit.com/articles/article.aspx?p=2460260&seqNum=2.

[17] Apache Hadoop HDFS Architecture, https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/

[18] Kadhar Basha J and Dr. M. Balamurugan (2017), A Review on Hive and Pig, Vol. 3, *ISSN* (Online): 2456-5717, Special Issue 39, May 2017.

[19] Map Reduce Architecture, http://www.glennklockwood.com/data-intensive/hadoop/overview.html.

[20] Hive architecture, https://pocfarm.wordpress.com/2016/05/09/working-of-hive/.

[21] Sunny Kumar and Eesha Goel (2016), Comparative Analysis of MapReduce, Hive and Pig, Vol. 17, *an International Journal of Engineering Sciences*, January 2016.

[22] Apache hive, https://cwiki.apache.org/confluence/display/Hive/Design.

[23] Dr. Urmila R. Pol (2016), Big Data Analysis: Comparison Of Hadoop MapReduce, Pig and Hive, volume 5, *International Journal of Innovative Research in Science, Engineering and Technology*, Issue 6, June 2016, ISSN: 2319- 8753.

[24] Apache pig architecture, https://hadoop4all.wordpress.com/.