Read Write Ser-De for JSON data in MapReduce Abstraction

B. Nandan[1], K. Sai Kiran[2], O. Sai Krishna[3], K. Pradeep[4], K. Vishnu Vardhan[5]

[1]Associate Professor, Department of CSE, Gurunanak Institutions, Ibrahimpatnam, Hyderabad, India
[2,3,4,5]B.Tech student, Department of CSE, Gurunanak Institutions, Ibrahimpatnam, Hyderabad, India

---***---

Abstract - *In this Advanced Generation of Technology data is generated in huge amounts. Data generated is generally in Structured and Unstructured format. Preserving and Analysis of such data is very essential. There are different methods to store and analyse data but analysis of unstructured data is one of the major problems faced by many Tech Giants. Unstructured data is not organised, text heavy and has many irregularities. About 70-80% of the data generated by companies is generally Unstructured. Hence in this project the focus is on Analysis of Unstructured data. The unstructured data is first converted into semi-structured data using Hive and SerDe after which analysis is performed using HIVEQL. Analysis of the data is carried out in a cluster, Serialisation and Deserialisation of data is done so that it can be easily transferred from the Master to all the Slaves in the cluster. Finally the converted Semi-Structured data is analysed by the slaves.*

KeyWords: Unstructured data, Semi-Structured data, HIVEQL, Cluster, Serialisation, Deserialisation

1. INTRODUCTION

Data on analysis gives results which are of huge value and essential to run any organisation. Storing such huge amounts of data and analysing them is a tedious task but in recent times technology has evolved so much that processing speeds have exceeded human expectations. Currently Unstructured data holds about 70-80% in the data generation process. It is not preferred to store this data which is difficult to analyse and store as well. This project will analyse Unstructured JSON data which will bring about results which have never been expected before.

1.1 Big data

Big data is referred to large data sets which are too complex to be analysed using traditional formats of data processing. Predictive, behavioural and many other advanced data analytics methods are used to retrieve data from the large data sets and convert it into a size suitable to use. Big data as the name suggests is based on its size which is generally ranging from petabyte to yottabyte. The data is too complex to be understood and is too diverse to find a pattern. For big data to be actually used, right queries must be posed and also the data should be easily analysed and cleaned. Big data can be better understood by these characteristics - volume, velocity, variety, variability and veracity.

1.2 Hadoop

Hadoop has been a revolution in the field of big data analytics. Hadoop was able to analyse, clean and present large data sets in a proper manner with simple queries. Hadoop uses MapReduce algorithm to smaller units and then carried onto the slaves in the cluster which follows a Master-Slave architecture. The data is stored in a separate filesystem called HDFS (Hadoop Distributed File System) which helps with aggregating bandwidth across cluster.

1.3 Hive

Apache hive is one of the data warehouses of Hadoop used for data analysis, query writing and summarisation. Hive provides an interface similar to SQL to analyse large data sets. Hive also has its own query language almost similar to SQL which is very easy to run and implement queries on. Hive provides the necessary abstraction required to integrate HIVEQL queries into low level Java API. Hive also has its own built in user defined functions to manipulate strings, date and similar mining tools.

1.4 SerDe

SerDe is expanded as Serialization and Deserialization. Serialization is the conversion of objects into byte streams and Deserialization is the exact opposite of that. Data is transferred the fastest when it is in a byte stream format and hence SerDe comes into application. Data objects are seldom complex and are difficult to convert to byte stream and hence Serialization will help convert these complex data objects into byte stream to transfer to all the slaves in the cluster and then deserialization is performed and the byte stream is converted back to data objects to work on.

1.5 JSON data

JSON also called JavaScript Object Notation is a lightweight data interchange model. It is the type of data which is easiest for humans to read and understand and also generating JSON data is quite simple. JSON data is said to be the subset for JavaScript programming language. The property that this data doesn't depend on the language at all but uses protocols that are similar to that of the C language family makes it the best data interchange model. JSON is basically built on two structures name/value pairs collection and ordered list of values. These can be considered as universal data structures and almost all programming languages support them.

2. EXISTING SYSTEM

RDBMS (Relational Database Management System) and DFS (Distributed File Systems) were the goto means for storage of Big data before Hadoop came into picture. But in RDBMS and DFS storing large amounts of data and retrieving it takes a lot of time. It gets complicated when we use RDBMS in a cluster model to retrieve data, as data has to be retrieved from all the slaves database where it is stored.

JSON data is particularly a new concept to many individuals but many Tech Giants prefer using JSON data as it is a lightweight data interchange format. Parsing and generation of JSON data is relatively easy. But since it is a new concept, analysis of this data is considered quite complex, time consuming and not preferred.

Analyzing Unstructured Data is out of the question for most Tech Giants as the data is too messed up and writing queries to analyse data is simply put very difficult.

3. PROPOSED SYSTEM

Hive is a data warehouse in Hadoop that specializes in coordinating and executing tasks that contain HDFS. Unstructured data is converted to Semi-Structured or Structured data for best analysis of the data. The data is then Serialized from objects into a stream of bytes to allow easy transmission of data from Master to all the Slaves in the network. HIVEQL is used to create queries to analyse the Semi-Structured or Structured data. This process will result in faster data analysis, data retrieval and also unstructured data can be analysed. Since Hive works in a Master-Slave structure, efficient management of the data analysis is carried out. JSON data analysis is also possible through Hive and is the best means to analyse JSON data.

4. SYSTEM ARCHITECTURE

In this system we mainly discuss on conversion of data from Unstructured format to SemiStructured or Structured data, after which analysis on the data is performed using HiveQL. The System architecture is depicted below.

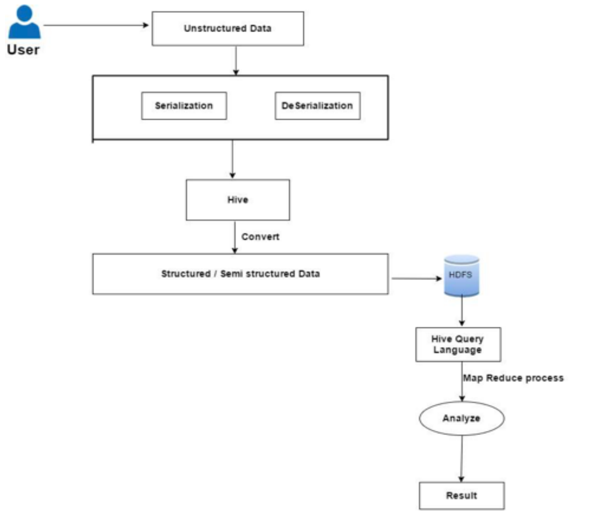

Fig -1: System Architecture.

5. IMPLEMENTATION

Fig -2.1: Unstructured JSON data example.

Fig -2.2: Unstructured JSON data example.

Description: Unstructured JSON data is text heavy, has no layout, variable content and hence it is converted to SemiStructured or Structured data using Hive and SerDe.

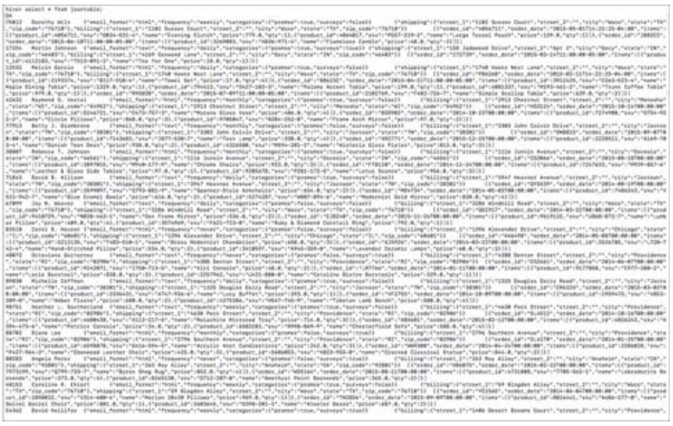

Fig -3.1: Semi Structured JSON data example.

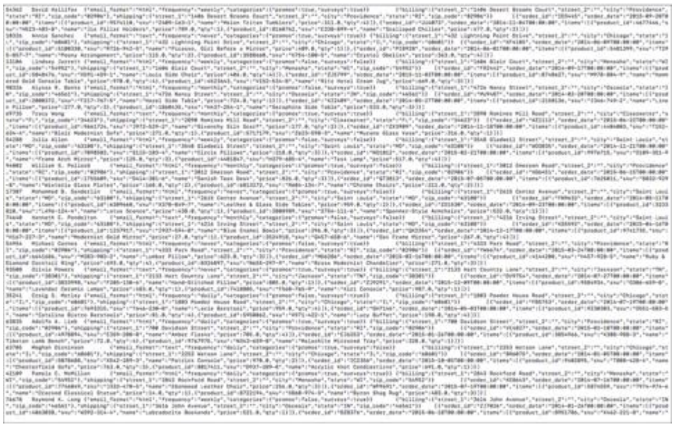

Fig -3.2: Semi Structured JSON data example.

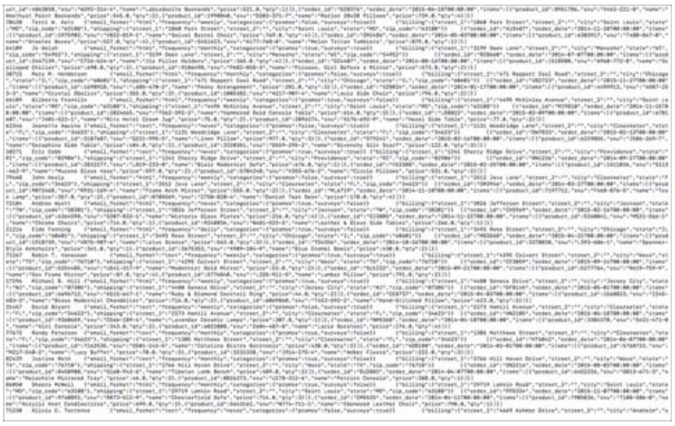

Fig -3.3: Semi Structured JSON data example.

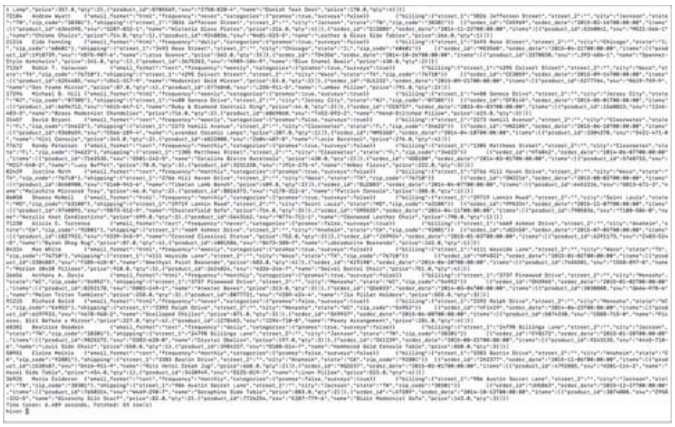

Fig -3.4: Semi Structured JSON data example.

Description: Structured/SemiStructured data is better when compared to Unstructured data. This type of data has definite content and better layout and format over Unstructured data.

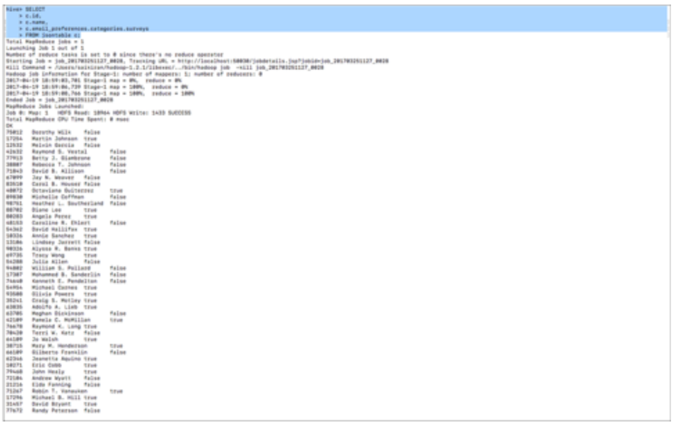

Fig -4.1: Analysis of Semi Structured JSON data example.

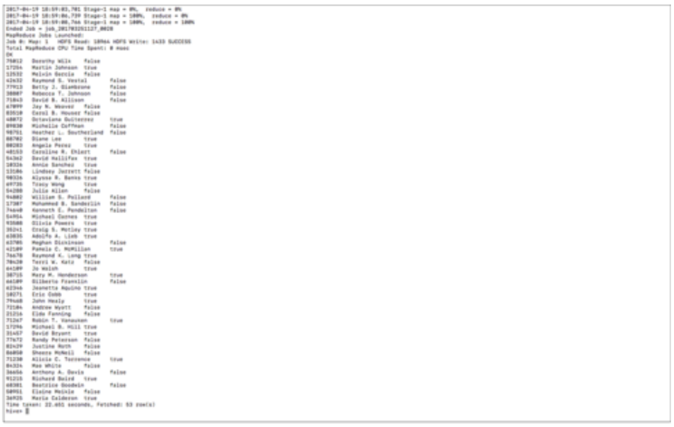

Fig -4.2: Analysis of Semi Structured JSON data example.

Description: The converted data is analyzed based on few understandable values in the JSON SemiStructured data. This ends with faster analysis of data and efficient output.

6. CONCLUSIONS

Hadoop is an emerging technology and further development is being continued. To be able to analyze Unstructured data is itself a great achievement in the field of data analysis. Data processing and storage has reached speeds which have never been thought possible. RDBMS and DFS have been replaced with Hadoop for data storage and analysis. Structuring of JSON data so that data will be organized and data analysis is also carried out through this project. Finally a new insight to the existing process of data analysis in a Cluster and also Unstructured data is brought about through this project.

7. FUTURE ENHANCEMENT

Further development is necessary as Hadoop has already become the best platform for Big Data processing and also Hadoop is capable of multi structure data processing. Hence development can be carried out by making sure HDFS file management and relational DBMS coexist on the same storage nodes. Apache Hive and Pig can come together to prosper in the field of Big Data analytics.

REFERENCES

[1] K. Balakrishna ,Smt. S. Jessica Saritha, C. Penchalaiah, (2015), “Extracting Structured Data From UnStructured Data Through HiveQL”

[2] Harpreet Singh Padda, Gulabchand K. Gupta, (2015), “Analysing Impact of Delimiters on the Size of JSON Data Interchange Format”

[3] Harpreet Singh Padda, Gulabchand K. Gupta, (2016), “Compressing JSON using Data Maps”

[4] https://en.wikipedia.org/wiki/Apache_Hive#cite_note-26

[5] Bakshi, K.,(2012),” Considerations for big data: Architecture and approach”

[6] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , “Shared disk big data analytics with Apache Hadoop”

[7] Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),“Addressing Big Data Problem Using Hadoop and Map Reduce”

[8] Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),”A Big Data implementation based on Grid Computing”, Grid Computing

[9] Zhu, X. ; Wu, G. ; Ding, W.,(26 June,2013),” Data Mining with Big Data”

[10] http://www-01.ibm.com/software/in/data/bigdata/

[11] https://en.wikipedia.org/wiki/Serialization

[12] https://en.wikipedia.org/wiki/Apache_Hadoop