International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

Life Cycle of Big Data Analysis by using MapReduce algorithm

Dr. Bharti kalra[1], Dr. D.K.Chauhan[2], Anupam Kumar Yadav [3]

[1]*Assistant Prof., Dept. of CSE, Institute of Technology Gopeshwar, Uttarakhand, India*
[2]*Director Technical, Noida International University, Gautam Budh Nagar, India*
[3]*Assistant Prof., Dept. of ECE,* Noida International University, Gautam Budh Nagar, India

---

Abstract - *Big data is defined as the large set of data which may be structured, Semi-structured and unstructured having 3 properties variety, volume and Velocity. This paper define the Life cycle of big data followed by we analyze the data using hadoop 2.3.0 and mapreduce. Furthermore, this analysis is explained by the trend analysis.*

Key Words: Data Analytics, Algorithm, Mapreduce, Big Data

1. INTRODUCTION

Big Data refers to data sets that describe any voluminous amount of structured, semi-structured and unstructured data that has to be mined for valuable information. Big data is not defined by a specific quantity; the term is often used when talking about petabytes and exabytes, zettabyte of data. Big data define by its 3 properties that is Variety, velocity and volume, furthermore two additional dimensions also specify by the SAS when thinking about the big Data i.e. variability and complexity[1].

1.1 Life cycle of big Data

BIg data life cycle can be explained as the three primary stages, with the data governance that helps to manage these:

1. Data Generation
2. Data Processing
3. Data Storage

Data generation phase define the privacy protection, access restriction and falsifying data techniques. Data can be generated from multiple resources in large volume and also in different variety. Data can be generated by human as well as by machines. IBM , Every day, create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. Data comes from everywhere: they use sensors to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few [3].

In the data storage phase privacy protection is based on the encryption techniques.

Encryption techniques classified into 3 techniques:

1. Storage path encryption (SPE)
2. Attribute based encryption (ABE)
3. Identity based encryption (IBE)

To protect against sensitive information hybrid clouds are used and sensitive data are stored in private cloud [5]. This phase defining the storage as well as managing of the large data set. The data processing stage define two techniques

1. Privacy preserving data publishing(PPDP)
2. Knowledge extraction from the data

To protect the data privacy anonymization techniques for example suppression and generalization are used. These approaches further classified into classification, clustering and association rules. Clustering and classification based upon the dividing data into parts, and association rule mining based upon input data trends and relationship.

As big data comes from various sources, so there is a need of efficient and effective way to process this data and handle measurements of big data in terms of volume, variety and velocity. So big data require experience to multiple phases during its life cycle.

2. PROBLEM DEFINITION AND METHODOLOGY

This case study uses a data set provided by the Google developers (Freebase API Deprecated) available online [2].This data set provides a dump of triples that have been deleted from Freebase over time. This is a one-time dump through March 2013. In the future, we might consider providing periodic updates of recently deleted triples, but at the moment we have no specific timeframe for doing so, and are only providing this one-time dump.

The dump is distributed as a .tar.gz file (2.1 GB compressed, 7.7 GB uncompressed). It contains 63,036,271 deleted triples in 20 files (there is no particular meaning to the individual files; it is just easier to manipulate several smaller files than one huge file).

Thanks to Chun How Tan and John Giannandrea for making this data release possible [2].

The columns in the dataset are defined as:
• creation_timestamp (Unix epoch time in milliseconds)

- creator
- deletion_timestamp (Unix epoch time in milliseconds)
- deletor
- subject (MID)
- predicate (MID)
- object (MID/Literal)
- language_code

This Paper proposes an implementation of Analytics of Big Data using Hadoop and Mapreduce.

3. SIMULATION BACKGROUND

The implementation done on windows 8 with the hadoop 2.3.0, Map reduce and the JAVA version 6. The data is to be analyse using the word count algorithm available with the hadoop 2.3.0, map reduce is the core component of hadoop which is available to analysis data in parallel fashion, in two phase 1) Map 2) Reduce. Each map and reduce phase has a key-value pairs for input and output, Where map phase runs a map per function that generates a set of intermediate key-value pairs. Reduce phase reduce the number of data using the key-value pairs. The reducer class display the final output by file name following the part-r-00000[4].

First of all, we analysis the last column of this data set that is Language code, the describes the language for the particular data set. The data set describes the total number of language used by the different dataset in 20 different set. The analysis is done for the different 6 Languages that is English(en), Hindi(hi), French(fr), German(de), Japanese(ja) and for Italian(it). All these languages mentioned through its short code that is en, hi, fr, de, ja and it. The analysis of data defined through it trend analysis.

4. TREND ANALYSIS

The analysis the frequency of different language code in different 20 files. These 20 file consider as the different object. In this analysis, we will incorporate the regression based prediction.

Regression based prediction formula
Equation for forecasting (regression based prediction)
$Y = a + bX$
Where $a = \bar{y} - b\bar{x}$ and $b = \Sigma(x - \bar{x})(y - \bar{y}) / \Sigma(x - \bar{x})2$
$\bar{x}$ and $\bar{y}$ are means of x and y respectively.

First analysis is for the English language which language code is en and also searches by the value en. It is the most used language for the data set. Fig 1 shows the frequency of English language. By the analysis we get more than 2 crore data set using the English language in this data and also in third file 1034621 data set using the English language.

Second analysis is for the Hindi language which language code is hi and also searches by the value hi. Fig 2 shows the frequency of Hindi language. By the analysis we get few data sets written in hindi, as also shown by figure.

Third analysis is for the French language which language code is fr and also searches by the value fr. Fig 3 shows the frequency of French language.

Fourth analysis is for the german language which language code is de and also searches by the value de. Fig 4 shows the frequency of german language.

Fifth analysis is for the Japanese language which language code is ja and also searches by the value ja. Fig 5 shows the frequency of Japanese language.

Sixth analysis is for the Italian language which language code is it and also searches by the value it Fig 6 shows the frequency of Italian language.

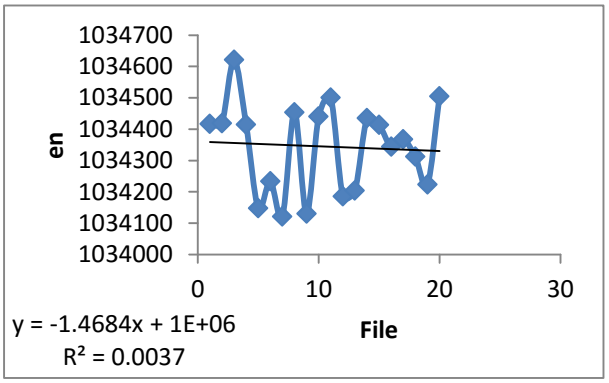

Fig -1: Frequency of en (english)

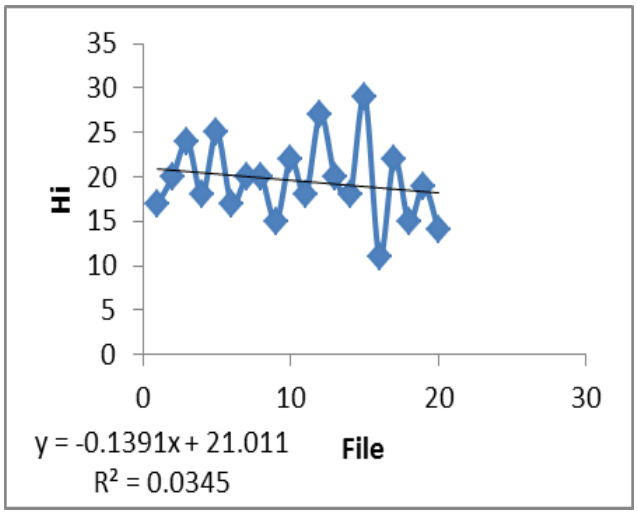

Fig -2: Frequency of Hi (Hindi)

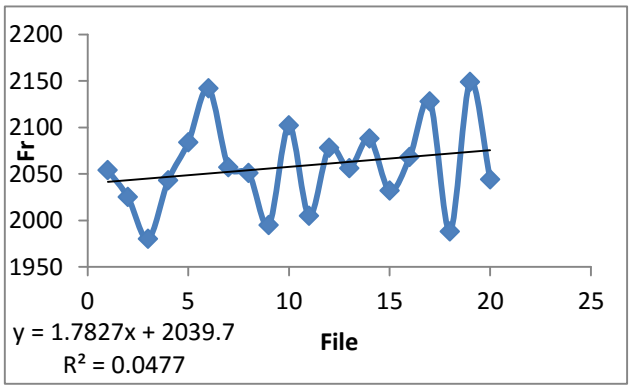

Fig -3: Frequency of Fr (French)

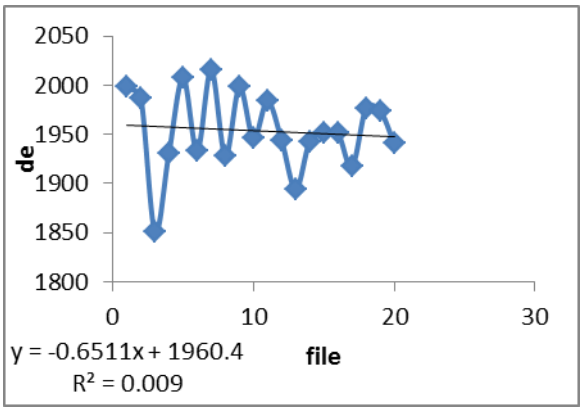

Fig -4: Frequency of de (German)

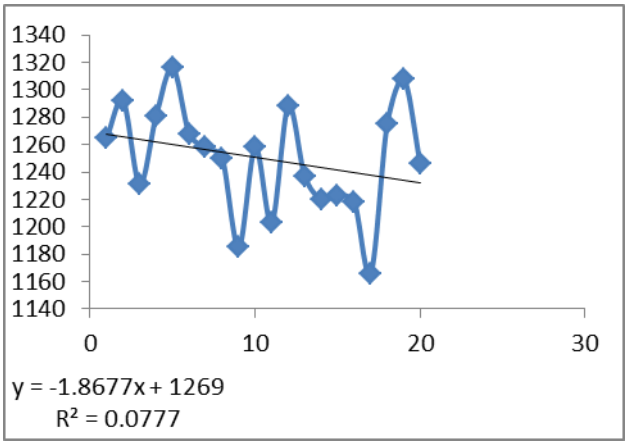

Fig -5: Frequency of ja (Japanese)

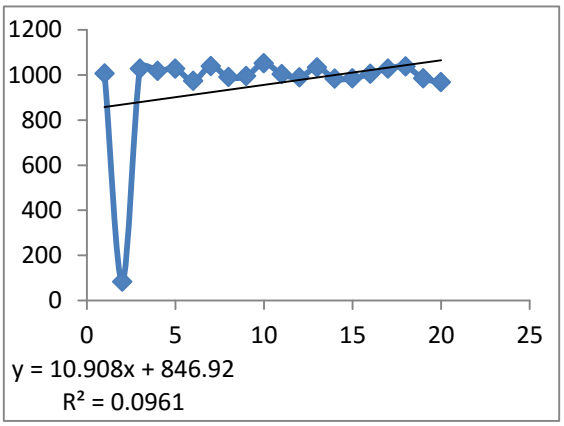

Fig -6: Frequency of it (Italian)

5. CONCLUSIONS

Our objective to analysis the freebase decrypted data through the hadoop and mapreduce Programming model and summarizes the result with the help of trend analysis of some aspects of multiples columns.

REFERENCES

[1] Kalra Bharti, "A Review of Issues and Challenges with Big Data", International Journal of Computer Science and Information Technology Research(IJCSITR) Vol. 2, Issue 4, pp: (97-101), Month: October - December 2014, Available at: www.researchpublish.com, ISSN 2348-1196 (print), ISSN 2348-120X (online).

[2] Google freebase data dumps. (n.d.). Retrieved November 02, 2015, from https://developers.google.com/freebase/data, December, 05, 2015.

[3] https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[4] Kalra Bharti, "Analysis of Data Using Hadoop and Mapreduce", International Conference on Emerging Trends in Engineering and Technology (ICET-15) Organized by IFERP on 27 December 2015, Ahmedabad, ISBN: 9788192958048. This paper is also published in International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015 Doi : IJERCSE-IFERP-DOI-21206, pp:20-22.

[5] Jain et al. J Big Data (2016), Big data privacy: a technological perspective and review, Journal of Big Data, DOI 10.1186/s40537-016-0059-y.