HCTL Open International Journal of Technology Innovations and Research (IJTIR)
http://ijtir.hctl.org
Volume 14, April 2015
e-ISSN: 2321-1814, ISBN (Print): 978-1-62951-946-3

Big Data: Study in Structured and Unstructured Data

Motashim Rasool[1], **Wasim Khan**[2]

mail2motashim@gmail.com, khanwasim051@gmail.com

Abstract

With the overlay of digital world, Information is available in the form of massive data. This Huge amount of data is generated as a rapid rate which may exceed the capacity of online storage. There are many sources that generate such large and unstructured data such as Social Media sites, Scientific Data, Engineering and Technologies, Government Data and many other activities which advances the scientific and technical world. Earlier we work on Structured Data and are manage easily by traditional databases. There are many algorithms and procedures to operate on structured data but at present structured data are overlaid by unstructured data. Information or data are originated at a very high speed which is surpassing the limit of online data storage. Big data is the answer to this giant data problem. In this paper, first we study the privacy concern of big data, Big Data consistencies and inconsistencies and approach to address Big Data Problem using Hadoop File System.

Keywords

Unstructured data, Big Data, privacy issues, Inconsistencies, HDFS, Map Reduce.

Introduction

Big Data is a new challenge to electronic industry. As the evolution of organizations, technologies, and other social networking sites results into a big size databases that cannot be handled by traditional databases. Data can be transferred in a range of terabytes to peta bytes and peta bytes to exabytes in a single data set. To extract meaningful information from huge amount of data is very difficult but there are many domains and economic sectors [1] get advantages from the big data such as physical sciences, medicine education, healthcare, banking and insurance sector, government, financial services etc. There are two main sources of big data first one is the data generated inside the organization like emails, mainframe logs, PDF Documents, structured data, unstructured data, semi structured data, etc and second one is the public data available outside the organization. Some data is available free of cost or some may be under paid subscription and some data is available for specific business partners. Big data is a gigantic collection of structured and unstructured data and to work on structured data there are many algorithms and procedures are available but for unstructured data, big data analytics can be required.

Characteristics of Big Data

Big data can be characterized by Dough Laney with 3 Vs [2]

- Volume: represents the dimensions of data. The size of available data has been growing at an increasing rate. This applies to companies and to individuals. A text file is in a Kilo Bytes, a sound file is in a mega bytes while a full length movie is in a Giga Bytes [3] due to which size of databases also increases. According to George Lee 90% of world's data has been created in last two years [4] .
- Velocity: represents the rate of data generation. It has two aspects first is throughput of data that means data moving in pipes and second is latency that refers data in motion or measure of velocity.
- Variety: represents a type of data or category of data such as text, audio, video, Adobe pdf documents, etc. Big data analytics done on mining of data and transform that data into data warehouse.

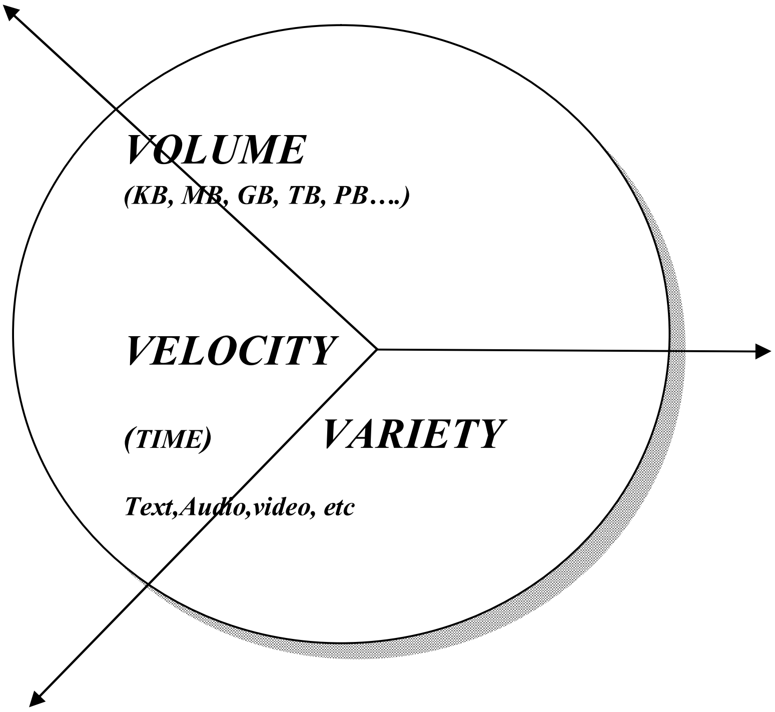

Fig: The Three vs. of Big Data

Now there are two more Vs [5]

- Variability: represents the data interpretation cause by change in structure of data.
- Value: represents the business values extract only required data from unstructured data which may cause faster transformation of data.

Big Data Problem

To work on big data using relational databases is very difficult because we are working in faster data transformation world. Since big data is a large or very large amount of data, we need 100s or 1000s of servers to run parallel to achieve such faster data transfer speed. As data transfers in peta bytes to Exabytes to scale, understand and extract meaningful information or quality information from such huge data is also a big challenge. Variability of data or types of data such as data in text, audio, video, images, etc also differs from industry to industry.

Big Data issues

There are many issues with data like privacy issues [6] , management issue, storage issue, Inconsistencies issues [7] .

- Privacy issues: We are working on the various online portals such as social media sites where we post our personnel information, photos, likes and dislikes which may leads to disclosure of personnel information. Some privacy issues are- Individual privacy, business privacy, log storage privacy, privileged privacy, and Geo Information privacy.
- Inconsistencies issues: inconsistencies occur due to variance in data sets. Data inconsistencies generate ambiguous form of data which affects the integrity of data. To handle data consistencies, levels of big data play important role.

Levels of Big data:

- Data: Data is basic level of big data. It is a unstructured form of text, audio, video, images and other type of data.
- Information: It is a partially structured form and aided some meaning to data which is easily understood to its user and may remove inconsistency.
- Knowledge: represent the specialization form of information in a specific domain. It leads to accurate decision making.

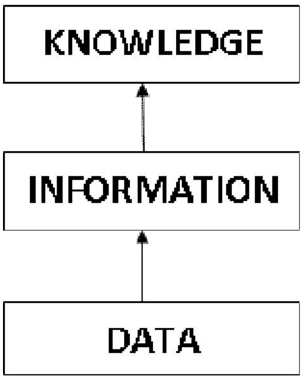

Figure: Levels of Data

Types of inconsistencies [8] : The most common inconsistencies are-

S/no	Type of inconsistency	Description
1	Text Inconsistency	When two or more texts referring to single entity or event in positive or negative way then this type of consistency may occur. The sources of text inconsistency are blogs, forums, emails, social networking sites, etc. Example: John is very honest towards his job. John is careless towards his job. From the above two contradiction its very hard to decide which statement is correct for john.
2	Functional Dependency Inconsistency(FDI)[8]	Functional dependency occurs due to the violation of integrity constraints. There are three types of FDI- i-Single FD ii- multiple FD iii- conditional FD.
3	Temporal Inconsistency	Temporal inconsistency occurs with reference to time. It may be partial temporal or complete temporal inconsistency. Data generated on daily basis can overlapped the old data values which may cause this type of inconsistence.
4	Spatial inconsistency	Spatial inconsistency occurs due to the geometric representation of objects. Conflicts geometrical attributes may violate the spatial constraints.

Big Data Techniques: There many organizations like Facebook , Twitter, Linkedin working on big data. Apache introduces a Hadoop Project develops open source software. Hadoop provides a distributed file system to process large data sets on a multiple servers. Hadoop was derived from Google's map reduce and Google file system [9] .

Hadoop Distributed File system: To store data, Hadoop utilizes its own distributed file system, HDFS, which makes data available to multiple computing nodes. A typical Hadoop usage pattern involves three stages [10] :

- loading data into HDFS,
- Map Reduce operations, and
- retrieving results from HDFS.

HDFS provides high throughput access to application data and is suitable for applications that havelarge data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster [11] .

Conclusion

In this paper, we study the origin of big data, big data problems. Privacy issues and inconsistencies effects big data analytics. In recent trends big data is required in various fields such as health care, banking and insurance sector, economic sector. To deal with big data, addressing of big data issues is mandatory. To overcome big data inconsistencies Hadoop plays an important role.

References

[1] Du Zhang, “Inconsistencies in big data”, Proc. 12th IEEE International conference on cognitive informatics & cognitive computing [ICCI*CC’13] , pp.61-67.

[2] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

[3] http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

[4] Goldman Sachs: CIO of Investment Banking Division.

[5] Wei Fan and Albert Bifet: Mining Big Data: Current Status, and Forecast to the Future

[6] Arul Murugan, Anguraj, Boopathi: BIG DATA: PRIVACY AND INCONSISTENCY ISSUES: IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[7] Du Zhang, “Inconsistencies in big data”, Proc. 12th IEEE International conference on cognitive informatics & cognitive computing [ICCI*CC’13] , pp.61-67.

[8] W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, Conditional functional dependencies for capturing data inconsistencies, ACM Transactions on Database Systems, Vol. 33, Issue 2, June 2008.

[9] Apache Software Foundation. Official apache Hadoop website, http://hadoop.apache.org/, Aug, 2012.

[10] Big Data Now: 2012 Edition by O’Reilly Media, Inc.

[11] Aditya B. Patel, Manashvi Birla, Ushma Nair: Addressing Big Data Problem Using Hadoop and Map Reduce: 2012 Nirma University International Conference On Engineering, NUiCONE-2012, 06-08DECEMBER, 2012.