International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Survey Paper on Data Lake

Surabhi D Hegde[1], Ravinarayana B[2]

[1]M. Tech Student, Department of Computer Science and Engineering, Mangalore Institute of Technology and Engineering, Moodabidre -574225, Mangaluru, Karnataka, India

[2]Associate Professor, Department of Computer Science and Engineering, Mangalore Institute of Technology and Engineering, Moodabidre -574225, Mangaluru, Karnataka, India

Abstract: ***One of the key driving forces behind the problem of Big Data is the rapid growth of unstructured data, which constitutes huge percentage of overall data [1]. The Big Data is not only about massive data capture and storage, but intelligently combining the past data that already exists inside an organization with the unstructured data. For an organization to be really successful to meet the latent benefits of Big Data, it needs the perfect technology in place to acquire the data, store it, combine it and enrich huge volumes of unstructured data in raw format. It should also have the ability to perform analytics, real-time, near-real-time analysis, batch processing on these huge volumes of data. To address these businesses needs efficiently, the concept of Data Lake is proposed. It is one of the empowering data capture and processing capability for Big Data analysis. Data Lake makes it possible to store all types of data irrespective of their schema and the formats. Data Lake is a massive, easily accessible, flexible enough and scalable large data repository.***

Keywords: Big Data, Big Data analytics, Data Warehouse, Data Lake.

1. Introduction

Big Data is the data that exceeds the processing capability of the conventional database systems. It can be said that the data is too big, and moves too furious that it do not fit the structure of the relational database architecture [4]. Big Data is set to offer organizations with effective and reliable insights. Big data is a data set that exceeds the sizes beyond generally used hardware components and software tools to store, maintain, and analysis it within an acceptable period of time for its user group. Big data analysis can said to be as the analysis of a special sort of data which comprise of structured, semi structured and unstructured data. Big data is most often produced due to the sensors, log files, audio messages, video messages, social media websites, network packets and web. Since data is being produced in enormous volumes with greater velocity in all the formats including, structured data, unstructured data and semi-structured data for multiple sources. It is necessary to concentrate on how to store this data efficiently without any loss, process it and then further analyze it efficiently so that it can be productive. Big data is set to offer companies useful insight. But with terabytes and petabytes of data pouring in to organizations today, traditional system architectures and their infrastructures are not up to the challenge level.

Big data analysis is a continuous, and not an isolated set of activities. Thus you need a perfect set of solutions for big data analysis, from acquiring the data from the source and finally discovering new insights to makedecisions and for ongoing analysis. The increasing data in the big data era brings about huge challenges on data acquisition, storage, management and analysis. Traditional data management and analysis systems are based on the relational database management system (RDBMS). However, such RDBMSs only apply to structured data, and not on semi-structured or unstructured data. In addition, RDBMSs are increasingly utilizing more and more expensive hardware. It is apparently that the traditional RDBMSs could not handle the huge volume and heterogeneity of big data.
Big Data has the following characteristics:

- **Volume**: It is about the about the size of the data, mainly considered starting from petabytes. The data may be in form of videos, audios and large images on in the social media. It is very common to have Terabytes and Petabytes of the storage system for enterprises.
- **Variety**: It refers to all the diverse data and file types that are used for data analysis efficiently. Data can be stored in multiple formats and schemas. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file.
- **Velocity**: is the speed at which the data are being created or processed for the analysis result.
- **Veracity**: represents both the correctness of the data source and in addition the suitability and the issue of the data for the intended group.

The key goal of big data analytics is to assist to settle on more educated choices by making use of large volumes of data for analysis by professional analytics, predictor and data scientists, additionally also to analyze distinct separate forms of data that may be undiscovered by conventional programs.

2. Data Warehouse

Data Warehouse is a large storage for data collected from a wide range of data sources. They store currently and past data and are used for creating analytical reports throughout the enterprise or organizations as per requirement. It restructures the data [2]. So that it delivers excellent performance, even for the complex analytical queries without impacting the any of the internal systems.

In data warehouse, input data is transformed and processed to a pre-defined schema and saved to data warehouse. This method is known as ETL (extract, transfer, and load)

processing. Before we can load data into a data warehouse, we first need to give it some format and structure i.e., we need to model it. That's called schema-on-write. This process of ETL, in most cases, results into data loss due to fixed schemas. There are no data processors required for the data as the data is already in a pre-defined schema ready to be consumed by data analysts.

Few reasons for the traditional approaches to fail are:

1) The conventional Data warehouse systems are not designed to integrate measure and handle this exponential growth of multi-structured data. But with the emergence of Big Data, there is a need to combine together data from various sources and to generate a powerful meaning of it.
2) The traditional systems lack the ability to integrate data from various sources. This leads to proliferation of data silos, due to which, business users view the data in various perspectives, which eventually stops them from making precise and useful decisions.
3) The schema-on-write method followed by conventional systems mandate the data model and analytic model to be designed before any data is loaded.
4) With traditional approaches, optimization for analytics is too much time consuming and incurs huge cost. These methods fails when there are new requirements.
5) It is difficult to identify what type data is available and to integrate the data to answer any questions. Manual recreation of data is error-prone and consumes lot of time which is a big problem.

3. Data Lake

Data Lake is a massive, easily accessible, flexible and scalable large data repository or large storage. Data Lake is a place to store practically unlimited amounts of data of any type, schema and format that is relatively inexpensive and massively scalable [3]. Hadoop ideally suits this use case and no other technology is so well suited as much Hadoop does. Hadoop implements a scalable and parallel processing framework that will process exceedingly large amounts of data in a smooth way, and makes it almost impossible to lose any kind of data, as it is replicated across the cluster. As organizations rush to take advantage of huge and diverse data sets, it's difficult to manage increase in the volume, velocity and variety of information today.

Data is coming in at such an overwhelming rate that organizations with traditional approaches cannot hope to capture the data and process the data efficiently. Some of the most valuable information particularly unstructured data remains without much processing [5]. And organizations have no way of knowing how much critical information and insight is being lost. To meet this challenge, the Data Lake was introduced which is a new approach that not only manages the velocity, volume and variety of data, but actually becomes more powerful as all three aspects increase. What makes this possible is a transformative shift from "schema-on-write" to "schema-on-read".

Data Lake Capabilities:

1) Captures and stores raw data from various source at low cost.
2) It can store various types of data in the same directory.
3) It performs various modifications, and on the data.
4) It will define the structure of the data when it will be used.

Key Attributes of Data Lake [6]:

1) Collect everything. A Data Lake contains all data, both raw data from various sources as well as any processed data.
2) Dive in anywhere. A Data Lake enables users across multiple business units to refine, explore and process the data efficiently.
3) Flexible access. A Data Lake can access the data in various forms: batch, interactive, online, in-memory and other processing engines.

Key benefits of Data Lake [7]:

1) Scale as much as you can: HDFS based storage in Hadoop gives the flexibility to support large clusters while maintain efficient performance. The Hadoop for underlying storage makes the Data Lake more scalable than Data warehouse by any order of magnitude.
2) Plug in disparate data sources: unlike Data Warehouse that can ingest only structured data, Hadoop supported Data Lake has an ability to ingest multi-structured and massive data sets from variant sources. This is one huge benefit and enables quick integration of data sets.
3) Store in native format: In Data warehouse the data is pre prepared into some format during ingestion phase. But Data Lake skips this phase, and provides iterative and immediate access to raw data. This provides the analytical insights.
4) Do not worry about schema: Traditional Data warehouses support schema for storing the data. But the Data Lake uses Hadoop's simplicity in storing data based on schema less write and schema based read modes.
5) Administrative resources: The Data Lake works better than Data warehouse in reducing the resources necessary for pulling, transforming, aggregating and analyzing the data in an efficient way.

4. Conclusion

Big Data is the data that exceeds the processing capability of the conventional database systems. It can be said that the data is too big, and moves too furious that it do not fit the structure of the relational database architecture. Big Data is offers the organizations with tremendous insights. But with large data like terabytes and petabytes and even more that pours into the organizations every day, traditional architectures are not up to the level to take up the challenges. Hence the concept of Data Lake was introduced. Data Lake is a huge repository which is not just limited to one type or source of data but all the data belonging to an organization in variant schemas and formats. Storing all this data at one place will increase availability and reusability of data among different departments and business units. Large and growing volume of data is stored in the Data Lake. It is used as a multi-tenant service and stores sensitive data and it mainly works on schema on read.

References

[1] Tom White, Hadoop: The Definitive Guide, O'Reilly Media / Yahoo Press 3rd Edition, March 2015
[2] http://hortonworks.com/blog/enterprise-hadoop-journey-data-lake
[3] http://info.hortonworks.com/rs/h2source/images/Hadoop-Data-Lake-white-paper.pdf
[4] http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf
[5] http://info.hortonworks.com/rs/h2source/images/Hadoop-Data-Lake-white-paper.pdf
[6] http://hortonworks.com/wp/content/uploads/2014/05/TeradataHortonworks_Datalake_WhitePaper20140410.pdf
[7] https://www.bluegranite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses

Author Profile

Mrs. Surabhi D Hegdecompleted theBachelor's Degree in Information Science & Engineering at Alva's Institute of Engineering & Technology, Mijar, Moodbidrifrom Visvesvaraya Technological University (VTU). Currently pursuing M. Tech degree in Computer Network Engineering at Mangalore Institute of Technology and Engineering Mijar, MangalorefromVisvesvaraya Technological University (VTU).

Mr. Ravinarayana Bis currently working as Associate Professor with 11 years of experience in the department of Computer Science & Engineering, Mangalore Institute of Technology and Engineering, Badaga Mijar, Mangalore.