A PUBLICATION OF MOTHERHOOD UNIVERSITY, ROORKEE

(Recognized by the UGC with the right to award degrees u/s 22(1) of the UGC act 1956 andestablished under Uttarakhand Government Act No. 05 of 2015)

Motherhood International Journal of Multidisciplinary Research & Development

A Peer Reviewed Refereed International Research Journal

Volume I, Issue IV, May 2017, pp. 01-09

ONLINE ISSN-2456-2831

Data Lake: A Next Generation Data Storage System

[1]Dr.Anil Kumar Kapil & [2]Ankit Kumar
[1]Faculty of Mathematical & Computer Sciences
[2]Faculty of Engineering & Technology
Motherhood University, Roorkee

Abstract

The traditional big data analytics makes it compulsory to store data in some structured format before analysis is done. Even storing data in warehouse require numerous preprocessing activities. This creates a limitation in an environment where new data are added frequently, like social networking analytics where big data analytics need to be performed on frequently changing data. But storing the data in warehouse or in any particular schema becomes a tedious job. Data Lake provides a solution for this problem.

Keywords: *Data Lake, Data Warehouse, Data Analytics, Big Data.*

Introduction

A data lake is a storage repository that stores vast data in flat architecture. It stores data in its native format. There is no preprocessing of the data before storing it in Data Lake. Data is stored in raw format in Data Lake; each data element is assigned a unique identifier and given a tag with related metatags. Queries can be fired on these data lakes, it results in small data sets which are later on analyzed and structured accordingly. This makes it possible to stores various format of data under one single repository. The content of the Data Lake need not to be converted in a particular schema, it can be done when they are queried. The data lake performs the extract, load and transforms (ELT) methods to accumulate and integrate data instead of traditional ETL approach. It follows a "Schema on Read" approach means when data is fetched at that time it is structured or transformed. The data lake allows us to store structured data from relational databases (tables, rows, columns), semi-structured data (CSV, XML, JSON, logs).It

also includes unstructured data (PDF, documents, emails) and images, audio, video etc., hence creating a centralized data store including all forms of data. Following are the key aspects of Data Lake [1]:

- Harness raw data at low cost.
- Multiple type of data (text, audio, video, feeds, doc, xml etc.) under one infrastructure.
- No need of data transformation at load time.
- To handle single subject analytics
- Perform real-time big data analytics efficiently

Comparison of Data Lake with Data Warehouse

In today's ever evolving operational environment where advanced analytics is employed, data warehousing is facing the challenges in dealing with the velocity of the data that arrives. Data Lake has come up with the solution for this. Following are the difference between Data Warehouse and Data Lake [2].

1. Schema: In Data Warehouse schema is defined before data is stored and in Data Lake schema is defined after data is stored. Hence Data Lake provides Agility and can work properly even if some data is unavailable.

2. Scale: Data Warehouse, if scaled then cost will increase (for preprocessing activities like cleaning, transforming etc). But Data Lakes provides scalable data repository with low cost.

3. Access Methods: Data Warehouse can be accessed with standard BI tools or by standard SQL. While Data Lake can be accessed by user defined programs.

4. Form of Data: In Data Warehouse data is cleaned i.e. all the activities like cleaning, smoothing, clustering etc. are performed and then data is stored. While in Data Lake data is raw.

5. Cost and Efficiency: Data Warehouses are costlier to implement while Data Lake is a low cost storage.

Building a Data Lake Infrastructure

Companies and Organizations can build Data Lake as per their requirements. Following stages represents how Data lake infrastructure is build:

Stage 1: Creating a place where data is gathered on a large scale.Probably Hadoop provides the solution for this.

Stage 2: Designing a tool that can retrieve data and transform the required data i.e. building analytic environment.

Stage 3: Delivering the data and analytics to more people as possible.

Stage 4: Providing enterprise features like governance, auditing and security

Factors to be considered for designing a Data Lake

Creating data lakes does not merely involve loading data into a repository but requires many factors to be considered. While designing or creating a data lake we have to consider following things [3]:

1. Indexing: Whatever data we are loading in the Data Lake has to be given a unique identifier. The index of these identifiers has to be maintained as a centralized index.

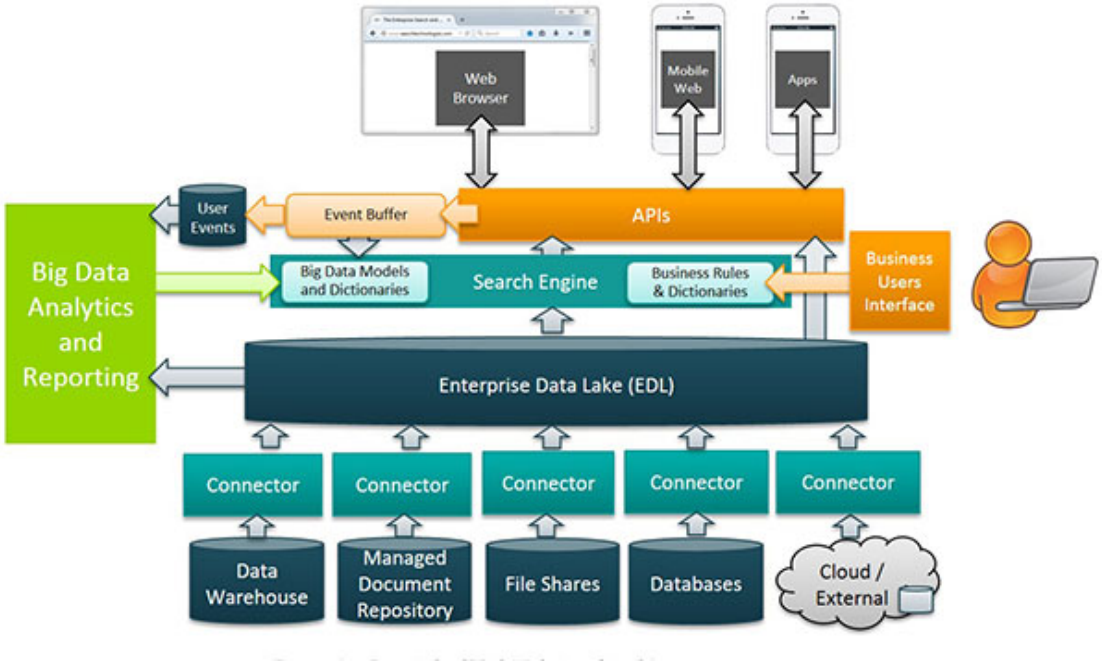
2. Authorization: While designing Data Lake, the grant access to the data has to be maintained so as to prevent unauthorized access.

3. Data Protection: The Data Lake designer should incorporate data security and data availability under circumstances like system failure or any other disaster.

4. Agile Analytics: Data Lake should allow multiple analytical models and approaches.

A data Lake Architecture

All content will be ingested into the data lake or staging repository (based on Cloudera) and then searched (using a search engine such as Cloudera Search or Elasticsearch). Where necessary, content will be analyzed and results will be fed back to users via search to a multitude of UIs across various platforms [4].



Searching the Data Lake

Data lakes will have tens of thousands of tables/files and billions of records. Even worse, this data is unstructured and widely varying.In this environment, search is a necessary tool:

- To find tables that you need - based on table schema and table content
- To extract sub-sets of records for further processing
- To work with unstructured (or unknown-structured) data sets
- And most importantly, to handle analytics at scale
- Only search engines can perform real-time analytics at billion-record scale with reasonable cost

Search engines are the ideal tool for managing the enterprise data lake because:

Search engines are easy to use – Everyone knows how to use a search engine.

Search engines are schema-free – Schemas do not need to be pre-defined.

Search engines can handle records with varying schemas in the same index.

Search engines naturally scale to billions of records.

Search can sift through wholly unstructured content.

Advantages of Data Lake

The Data Lake allows us to integrate and store all useful data under single infrastructure. It is useful in environment where dynamic data has to be analyzed. It is useful in real time analysis of streaming data. It is one of the smart methods of data analysis which give a quick insight of the analytics. It enhances the ability to analyze data in real time environment. It can be widely and efficiently utilized in social networking big data analytics. Data Lakes allows us to perform big data analysis on different types and kinds of data. Following are the advantages of Data Lake.

1. Low Cost
2. Fidelity (Unchanged Data)
3. Accurate Results since updated data is available
4. Easily Accessible
5. Runtime Binding

Conclusion

Data Lake gives a new way to manage new types of data and use the data as and when required. It will be very beneficial when faster results of analytics are expected. Nowadays faster result are expected, Data Lake will provide a platform where accelerating result can be expected. More insight of data can be gained using Data Lake.

Future Work

Archival process of out dated data • Implementing the variety of retrieval processes and a measure their efficiencies. • Scaling the storage of data in to various file systems • Compare and contracts against Hadoop based Data Lake with this plutonic implementation of Data Lake • Implementation of Data Lake clusters to mimic the data marts • Use HDFS and Map Reduce on content of the Data Lake.

References:

1. Hassan Alrehamy, Coral Walker (2015), Data Lake with Data Gravity Pull, IEEE Fifth International Conference on Big Data and Cloud Computing.

2. Forget data warehousing, it's 'data lakes' now, By Digital News Asia.

3. White Paper: How to Design a Successful Data Lake, by www.knowledgent.com.

4. Carlos Maroto, A Data Lake Architecture with Hadoop and Open Source Search Engines https://www.searchtechnologies.com/blog/search-data-lake-with-big-data