Report 1

Data Lake Architecture Overview

Le Nhu Chu Hiep

Big Data

The big data is a lage, diverse data sets of information that grow at ever-increasing rate. Both volume, velocity and variety of big data expand in furious speed. The big data come from various sources: mobie app, website, social networks, v.v. with many different types which can split to 3 main types: structured data (data from organization database, spreadsheets, v.v.), semi-structured data (XML, JSON, v.v.) and unstructured data (image, sound record, log file, v.v.). Since our world is reshaped by data (data-driven), we must save almost any valuable data for future analysing. And also because the data is too enormous and keep increasing day by day, we can not refine the meaning of all data to decide which data will be stored. Therefore, we try to collect as much as possible all recordable data and this data sets called big data. In other word, the big data produce the valuable information and better insight for the organization than the traditional pre-refine data.

4 charecteristics of big data:

- Volume: size of the data, mainly considered beginning from petabytes.
- Variety: refers to all the diverse data and types that are used for data analysis efficiently. Data can be stored in multiple formats and schemas.
- Velocity: the growing speed of the data.
- Veracity: represents both the correctness of the data source and in addition the suitablility and the issue of the data for the intended group.

Data Warehouse

Data Warehouse is a large repository for data collected from multiple of the data source (mainly from OLTP). It store currently and past of data. It serve the structured and can also for semi-structured data but not unstructured data. The DW mostly based on RDBS to store pre-refined data with the schema is designed supporting data analysis process (OLAP). This is old fasion technology to integrate data and support BI activity.

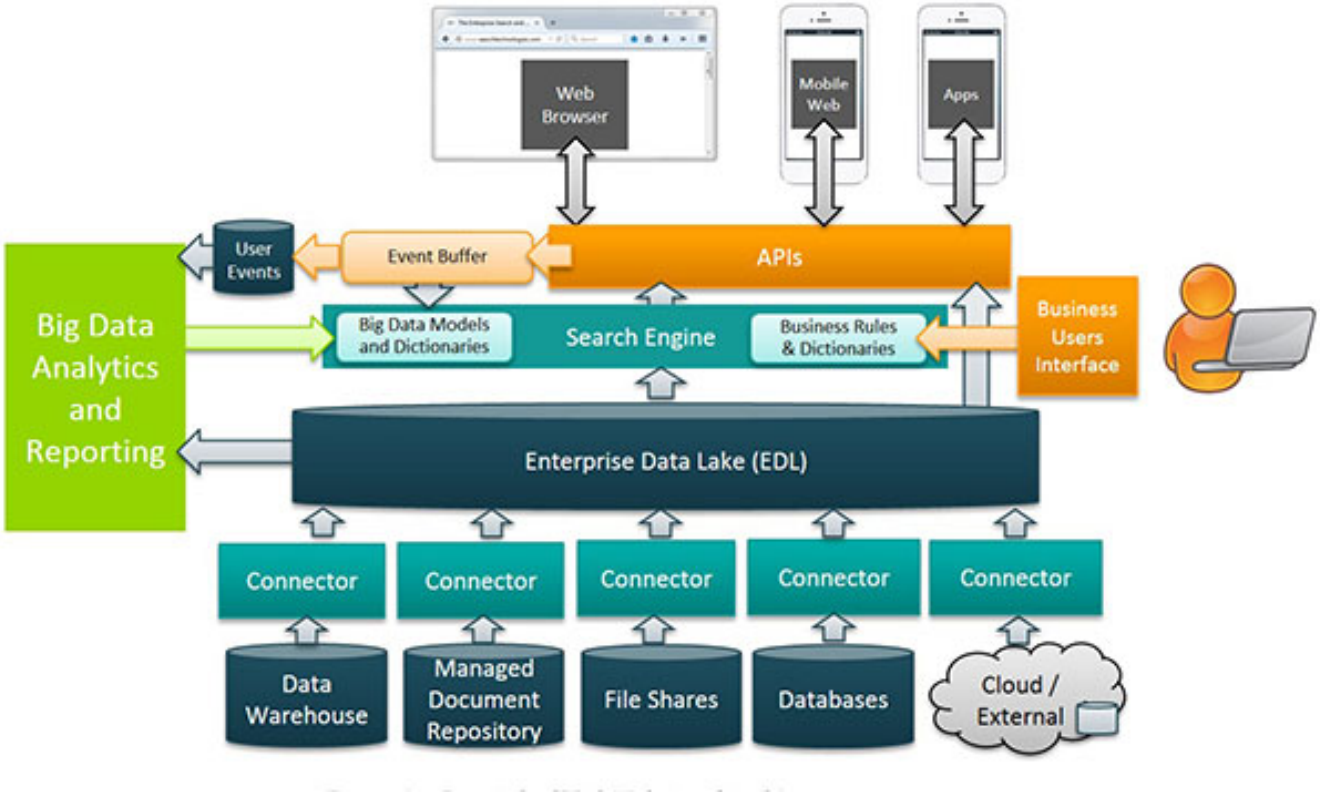
Data Lake

A data lake is a centralized repository that allows you to store all types and formats of data at any scale. It accpet both structured, semi-structured and unstructured data and store them as their natural/raw format. Because the data lake store data in raw and apply ELT process instead of ETL which mean schema on read, it solve the first problem of a big data - remains data without lossing the valuable data. Moreover, schema on read technique avoids overhead problem of ETL activity since the ingest data is large and fast. The data lake uses the low cost and already exit storage system so that the volume scaling is inexpensive. Then the data lake can hanlde not only large velocity, volume and variety of data but also the growth up of all 3 aspect.

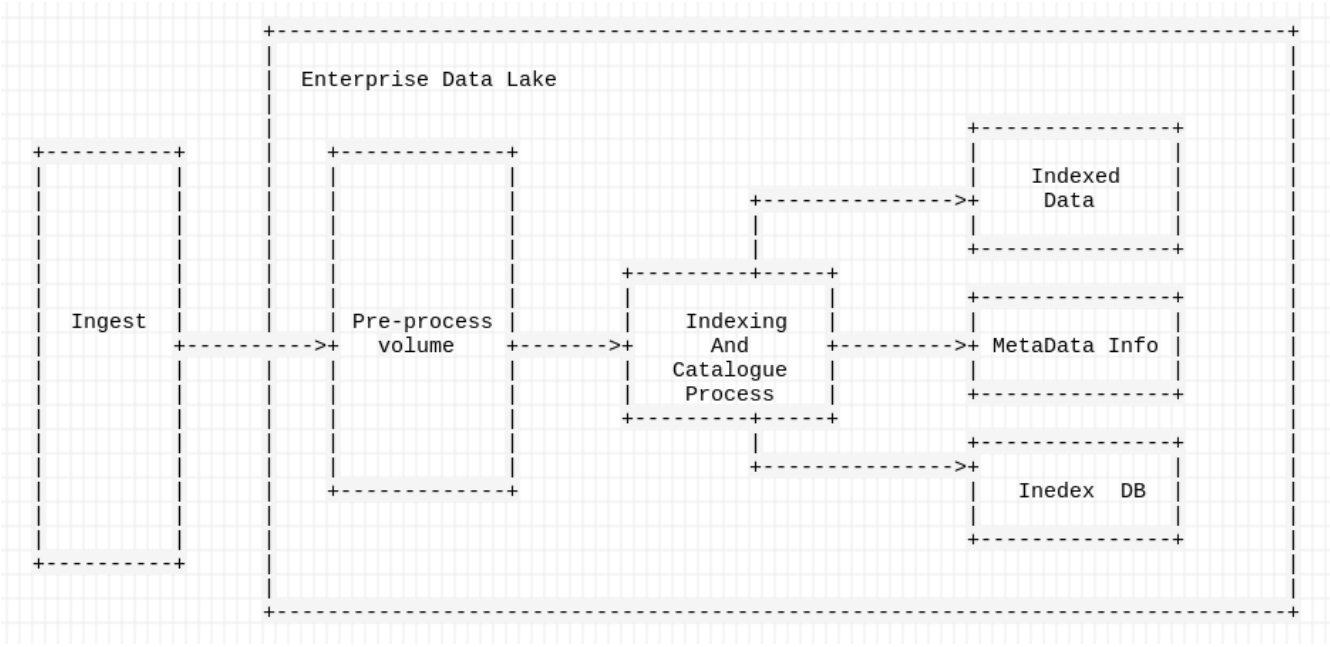
Key Difference

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Data Lake Prototype



EDL Prototype Recommendation v0.1



Reference

1. Surabhi D Hegde, Ravinarayana B: Survey Paper on Data Lake

2. Moh'd Alsqour, Kamal Matouk, Mieczyslaw L.Owoc: A survey of data warehouse architectures: preliminary results
3. https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/