Modern Data warehouse

In this modern world, data is everything for modern business analysis and data is being generated from several sources in different formats. It’s very hard to identify and streamline the data flow to make any sense out of it. As we have been growing every minute to minute in a web world, it is quite common that data is generating from heterogenous sources such as devices, sensors, machines, social media, etc. So, it’s been a challenge to storage and maintain for large amount of structured and unstructured data for analysis and archival process.

Big data is a modern solution for enterprise data warehousing system where we can store large amount of data on a specially designed file system which is Hadoop distributed file system. And data would be retrieved very quickly using map process. MapReduce process is a process where a piece of software will go to the data location and do the required work such as aggregations, calculations, etc. So that we don’t need to transfer and store the data for transformations. Hadoop is a best suited platform to store huge amount of data and processing without any lag.

**A typical characteristics of a modern data warehouse would follow below key features:**

1. Ability and capability of handling verity of data sources and subject areas.
2. Capability of handling large amount of data.
3. Expansion from traditional DWH to data lake which is specifically maintained by Hadoop, NoSQL databases.
4. Multi-platform architecture which balances scalability and performance.
5. Ability to facilitate near real time data using specially designed data streamline processes such as Kafka. These specially designed API’s will follow Lambda architecture.
6. DWH would be built with Agility for quicker delivery life cycles.
7. Hybrid integration with cloud.
8. Master data management for curation of reference Data.
9. Should support for all users rather than high level (business) people.
10. Support for self-service BI to augment corporate BI.
11. Delivery of data discovery and data exploration, in addition to reports and dashboards.
12. Ability to certify and promote self-service solutions to the corporate BI/analytics environment.

**A few best data warehouse techniques to build robust DWH:**

1. **S**witch to ELT (Extract, Load, Transform) rather than ETL wherever we needed to transform the data from source.(this tip depend on the type of data and source ).
2. Limit the unnecessary aggregations if not required.
3. Always summarise the data rather than storing entire raw data.
4. Always identify the Business Key’s before opt a data model.
5. Considering partition for larger amount of data tables. Usually suggest to create a separate partition for every 20million rows.

Create clustered index on date key on fact table.

Below is the typical architecture to store large amount of data using Hadoop file system. As shown in the below diagram, we will be having huge data from the several sources in the form of flat files, databases such as ERP, CRM and from social media such Facebook, LinkedIn, Twitter. So, store all these data would be a tedious and very expensive. To maintain a data warehouse for storing this data we must spending huge revenue along with expensive hardware.

Hadoop made it easy and very cost effective as it uses commodity hardware which is very cheap and still it is scalable and robust. It also provides high availability by providing fault tolerance. Hadoop provides its own file systems which is HDFS. And it supports variety of API’s where we can easily integrate with Hadoop, such as Kafka. And data can be retrieved using Hive query statements using Hive engine. Hive is similar to SQL but very powerful to retrieve data from HDFS file system. And we can use the data to create and maintain DWH for data analysis and data science/mining purpose. From the OLAP system to data mart and then to data visualizers.

Reports

Data Marts

Web Data

Flat files

d

DWH

Hadoop

Data marts

Direct Connect - Database

HDFS File System

Reports

Devices, Sensors, other source

ERP, CRM

Streaming application.

Example: Kafka

Data marts

Hive