***ecommerce-data-warehouse***

Data Warehouse

what is Business Intelligence?

Business Intelligence means analysing legacy data to get meaningful insight for business development.

**For instance,**

**analysing the last ten years of sales data to understand the buying pattern of customers.**

**Importance of data warehouse**

In an organization, there are a lot of front-end and back-end activities or business processes.

These processes generate tremendous amount of data, and each data group needs to be

managed efficiently to carry out day to day business operations.

The data generated by these business processes are stored in a database which is known as the Operational or Transactional database.

**Transactional databases :**

**These databases are normalised to take care of daily transactions involving**

**operations like INSERT, UPDATE and DELETE.**

Transactional database known as **OLTP**, facilitate day to day business activities.

Data warehouse : These are the de-normalised relational databases optimised for query

processing.

**Data warehouses collect and store from all the relevant datastores that are operational in the enterprise's ecosystem and provide the combined data for business intelligence.**

Data warehouse uses **OLAP** systems to analyse and summarize large amount of data.

Helps in the decision making process.

**Facts & Dimensions:**

Facts - are generally numeric data associated with a business process. **For example, 'quantity' and 'total amount sold' would be the numeric quantities associated with the sales transaction business process and are stored in a fact table.**

Dimensions - are used for grouping or categorising records that are stored in data warehouses.

**For example, if a business analyst wants to know which product is trending in the 18-25 age group, he/she may group the data with the help of the Age dimension stored in the Customer dimension table. And then, he/she can categorise the filtered (grouped) data with the help of the Product dimension table.**

Data Warehouse design can be broadly described as the following steps:

1. Determine the business process

2. Define the granularity

3. Identify the facts

4. Find the dimensions

**About Adveture works dataset**

Adventure Works database supports standard online transaction processing scenarios for a fictitious bicycle manufacturer - **Adventure Works Cycles**. Scenarios include Manufacturing, Sales, Purchasing, Product Management, Contact Management, and Human Resources.

Contains around 71 Tables under various departments.

Note : in this case of our e-commerce company “**Adventure Works Cycles”**, Manufacturing, Sales, Purchasing, Product Management, Contact Management, and Human Resources all comes under the category of OLTP systems.

Refer Link for better understanding : **https://dataedo.com/samples/html/AdventureWorks/index.html**

<< Complete Adventure works Dataset - Refer AW\_backup.sql >>

**We have customer , sales records for customersID between 11000 to 11300 for the purpose of this project, complete dataset is available in the file.**

<< Refer - 01\_partial\_dataset\_creation.sql>>

**Problem Statement or Business Objective:**

To Understand the purchase pattern of the customer based on gender, education and salary

To identify high performing stores so that allocation of goods can be planned.

To perform the above analytics, we need to load and maintain the repository of data , further run analytics on top of that using tool like Hive.

Let us try to Analyse for the below statements :

-- Which salary range group of customers contribute to more sales ?

-- To find the upper and lower discount limits offered for any product

-- Sales contributions by customer

-- To Understand customer persona purchasing pattern based on gender, education and yearly income

-- To find the sales contribution by customers on the overall year to date sales

belong to categorised by same gender, yearly income.

-To identify the top perfoming territory based on sales

-- To find the territory-wise sales and their adherence to the defined sales quota.

Identify the associations between the tables

v\_customer - (customer,individual,salesterritory)

v\_salesorderheader - (salesorderheader, salesperson sp , salesterritory st)

v\_salesorderdetails - (salesorderdetail, specialoffer)

v\_stores – stores , salesperson

v\_creditcard

Refer << 02.Views Creation.sql >>

**Data Ingestion**

Choosing a data ingestion tool depends on the type of data to be ingested. For examples, if you want to ingest structured data like data from RDBMS then Apache Sqoop is the right tool to use. Similarly, Apache Flume is suitable for ingesting unstructured data.

Apache Sqoop :

Sqoop only deals with structured data. Most organisations such as banks and e-commerce businesses maintain data in the form of tables, which Sqoop can deal with. However, unstructured data such as music, PDFs, web dumps, etc. can’t be handled by Sqoop.

**Sqoop jobs are basically used to copy/transfer data from rom a relational database to the Hadoop cluster/HDFS. No aggregation work is done using Sqoop jobs. Therefore, these jobs are map-only.**

<< Refer 03\_Sqoop-import.txt >>

**--split-by :** It is used to specify the column of the table used to generate splits for imports. **This means that it specifies which column will be used to create the split while importing the data into your cluster. It can be used to enhance the import performance by achieving greater parallelism.**

target-dir : Sqoop creates a new directory with the name provided in the command and imports the data.

Incremental Import:

1. --incremental, - To achieve this incremental import, you use the parameter --incremental, with the value as append.

2. --check-column - The parameter –check-column can be used to tell sqoop that this is the column to be checked to find the row is new.

3. – last value parameter indicates the value of this column for the last inserted row.

sqoop job --create myjob

sqoop job –list

sqoop job --show myjob

sqoop job --exec myjob

Just For Knowledge

How long a sqoop job will take to fetch a 500 GB ?

**1. Depends on what is the networking layer**

**2. Where are the data is sitting. If cloud is sitting in India and RDBMS is on USA then it will take time.Purely depends on the latency of the network layer.**

**Loading Data into a Warehouse tool - HIVE**

HIVE

- Hive is a data warehouse software

- Hive - Write once, Read many times ; RDBMS - Read and Write many times.

- Uses of Hive-Metastore - It stores all the information about Hive tables, in a central repository.

- HIVE Data Types - Primitive Data type , Complex Data type

- Primitive Data type - String,Int,float,double,Timestamp,Binary

- Complex Data type – STRUCT(Object),MAP(Key-Value),ARRAY(Indices)

- Internal vs External tables

- External table - External tables is used when data to remain stored on the HDFS even after dropping tables because Hive does not delete the data stored outside (of the Hive database).

- Internal table - In internal table, the data is temporary. So, when the Hive table is dropped, the data stored in the internal table is deleted along with the metadata.

Refer << 04. hive tables.txt >>

customer\_test

sales\_order\_header

sales\_order\_details

store\_details

creditcard

Creating and loading HIVE tables for performing analytics.

-- **Which salary range group of customers contribute to more sales ?**

-- **To find the upper and lower discount limits offered for any product 0.00 to 0.02 only 10 product ID’s**

**PID Max Discount offered**

**716 0.02**

**779 0.02**

**781 0.02**

**784 0.02**

**860 0.02**

**873 0.02**

**874 0.02**

**961 0.02**

**965 0.02**

**999 0.02**

-- **Sales contributions by customer**

**-- To Understand customer persona purchasing pattern based on gender, education and yearly income**

**-- To find the sales contribution by customers on the overall year to date sales**

**belong to categorised by same gender, yearly income.**

-- **To identify the top performing territory based on sales**

**-- To find the territory-wise sales and their adherence to the defined sales quota.**

Customers table

Sales table

Stores table

Understand the facts and dimesions available .

Mining demographics XML to get valuable facts

Table : customer\_demo

Totalpurchaseytd, datefirstpurchase, birthdate, maritalstatus, yearlyincome, gender,

Totalchildren, numberchildrenathome, education, occupation, homeownerflag, numbercarsowned ,commutedistance

<< >>