**Week 2 – Data warehousing, BigQuery, SQL**

Topics to review:

Data warehouse

BigQuery

SQL

Be prepared to write SQL queries.

Be prepared to work in your BigQuery environment.

Be prepared to answer questions regarding data warehousing, including architecture, schema, types of data.

**What is OLTP?**

OnLine Transaction Processing.

RDBMS (Relational Database Management Systems).

Contain homogeneous data, usually normalized till 3NF.

Databases fill up quickly, data needs to be sent somewhere else (Data Warehouses)

They contain current, transactional data.

Examples: Oracle, MySQL, Postgres, DB2, SQL Server, IBM Informix"

**What is OLAP?**

OnLine Analytical Processing

A form of decision support statement (DSS).

Generates a pre-prepared report based on data from DWH

Report can be written summary, or visuals can be sent like charts, dashboard, etc.

Examples: Sap Business Objects, IBM Cognos, MicroStrategy, QlikView, Tableau

**What is a Data Warehouse?**

Stores historical data from OLTP databases periodically

Stores for Reports, Queries, Analysis, Business decisions, Analytics

Data in DWH is stored within a slowly changing dimension (SCD) [ETL should cleanse and summarize data before coming in here).

Examples: Teradata (NCR), Exadata (Oracle), Vertica, Netezza, Greenplum

**What is the standard structure of tables in a Data Warehouse?**

Star Schema. Should have a center Fact Table which contains Key Performance Indicators (KPI) and foreign keys to Dimension tables (specific summarized data holders).

This helps reporting tools to perform reports in a fast manner without complex queries.

Snowflake Schema

**What is a KPI (Key Performance Indicator)?**

It's a calculated value which can quickly answer business questions. An example would be: Total Sales Last Year. It is used to avoid querying the Dimension Tables when it is not needed.

**A Data Warehouse should follow which properties?**

SINTA

Subject-oriented

  Data is based around main objective of company using data

Integrated

  Data is integrated, summarized, compact

Non-volatile

  Data will persist and won’t change. Write once, read many times.

Time-variant

Data should have a time stamp to keep track of period of data retrieval. Data like ""Age"" will become stale.

Archive

Data is archived to remember location of data in DWH and high volumes.

**What is ETL?**

Extract-Transform-Load where data from the OLTP databases are extracted, transformed into either a summary of the data or the summary and an inclusion of more detailed information regarding the data, then loaded into the DWH.

Examples: Informatica, DataStage, SSIS, Abnitio, Oracle Data Integrator (ODI)

**What is a Staging Database?**

Before entering DWH, data is quality checked in the “Staging Area”, otherwise known as Operational Data Store (ODS)

Quality checking also known as “Data Cleansing”

Includes removing trailing spaces, data validation, filtering out unnecessary data (“Mr.” from “Mr. Ram”)

**What is a Data Mart?**

A data mart is a simple form of a data warehouse that is focused on a single subject or line of business, such as sales, finance, or marketing. Given their focus, data marts draw data from fewer sources than data warehouses.

**What is Big Data?**

Big Data is a terminology used to describe a problem. Data has some specific properties defined by Gartner:

  Volume: How much data are companies currently holding? A lot. And where do we store it?

 Velocity: How fast can we process so much growing data?

 Variety: Data comes in different kinds, they can go from flat files to relational tables that actually have a schema.

**What kind of data do we have available?**

Structured Data

  Defined Format (like RDBMS)

Semi-structured Data

 XML, Email, Excel Spreadsheet

  Has an apparent pattern, enabling analysis

Quasi-structured Data

  Erratic format that can be formatted with tools

  Clickstream Data

Unstructured Data

  No inherent structure

  Text Documents, PDFs, image, videos

**Extra questions to consider**

**Define Big Query.**

**Explain the Architecture of Big Query.**

**How can data be loaded into Big Query?**

**What is Big Query Storage?**

**What are Views in BigQuery?**

**What are the features of Big Query?**

**Is BigQuery OLTP or OLAP?**

**Why is BigQuery Faster than SQL?**

**What are window functions in Big Query?**

**What are the different data types in Big Query?**

**How are Arrays used? JSON?**

**What is sharding?**

**Write down a command to fetch each record between two dates.**

**What is the difference between Legacy SQL and Standard SQL?**