**Discussion Blog**

**1) Different tools for data platform, ETL/ELT, and cloud**

We can use multiple tools such as SSIS/Informatica/DataStage/Talend for data extraction and also we can apply business transformation logic and load it into the data warehouse. Coming to Cloud Data Platforms we have Informatica cloud services which extract data from unstructured, semi-structured, and relational data sources. Also, AWS Data lake is used to store all formats (unstructured, semi-structured, and relational sources) of data

**2) Scheduling the process**

External/third-party tools are used to schedule all ETL, SQL, Unix/Bat jobs Eg: Tidal, Autosys, Main Frame application ESP

**3) Incremental updates to data sources**

After the Initial full load of data from source to target. We identify the delta/incremental data based on the timestamp that we have loaded from a source that gives updates/inserts from the last run date. Another Technique to identify the Incremental updates is by calculating the MD5 checksum between source and target. If the source to target checksum is changed we can pull all Incremental loads.

**4) Data Modeling techniques**

It is created in 3 stages -

**1) Conceptual Model** - It contains the self-sufficient entities and establishes a high level of relationship b/w entities. (Complete set of rows, columns, and all values in a table are collectively called entities)

**2) Logical Model** –Describes attributes of entities. > Identify keys > achieve complete normalization> eliminate many too many relationships. (Normalization - clubbing of two tables into a single table that is related )

**3) Physical Model** - Ready to implement data > creation of tables > Metadata of columns (establishing datatype for attributes)> database file size allocation > creation of PK/FK relation

Dimensional Modeling: Dimensional model contains facts and dimensions.

**Facts**- Facts are tables that contain measures. It always contains measures. Any transactional data which changes rapidly is part of facts. Facts are mainly numeric like bank withdrawals, deposits, etc. But it can be non-numeric too like customer feedback that changes.

**Dimensions** - This table is an entity that contains self-sufficient blocks. E.g. employee data, customer data, etc. Dimensions contain attributes.

**5) Data Quality**

Data Quality checks can be done in two phases

**Phase 1: Preload Checks**

a) Before loading data into our stage/target tables We can perform the following techniques

a1) Balancing & Reconciliation by comparing source trailer file count with target counts

a2) Formatting checks based on record types like fixed or delimited and length of records

**Phase 2: Post Load Checks**

B) After Loading data into the target table following techniques can be performed

B1) Threshold Checks: If we receive more records compared than expected we can have

Threshold checks to avoid duplicated data getting loaded into Facts/Dimension Tables.

B2) Aggregating Check: We can have data aggregation checks on transactions that we receive from

Source to Target.

Other than these checks we can have monitoring of historical load.

**5) Testing the URL in the command line**

We can use curl to get the responses for an API URL. We can also use the wget utility to test URLs from Linux servers.