09/09/2020 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 2

Aim: To study about ETL process & its tools.

Theory:

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load. It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process.

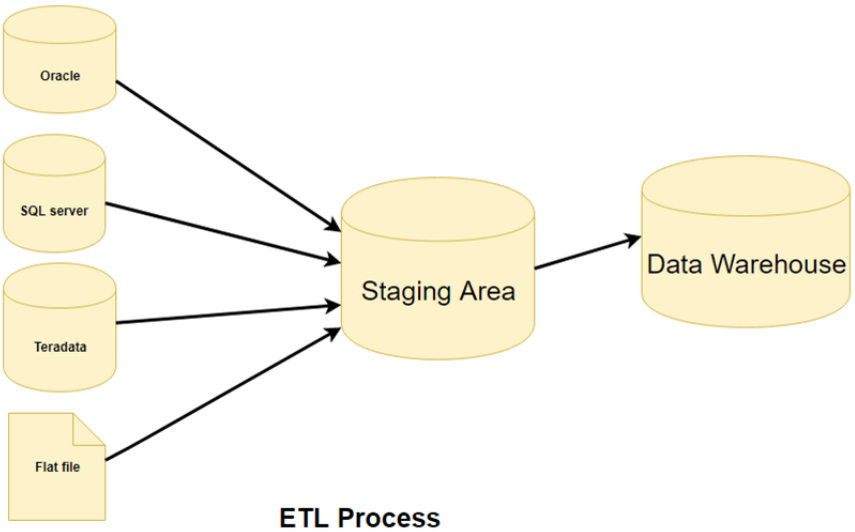
Need for ETL:

- There are many reasons for adopting ETL in the organization:
- It helps companies to analyze their business data for taking critical business decisions.
- Transactional databases cannot answer complex business questions that can be answered by ETL.
- A Data Warehouse provides a common data repository
- ETL provides a method of moving the data from various sources into a data warehouse.
- As data sources change, the Data Warehouse will automatically update.
- Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.
- Allow verification of data transformation, aggregation and calculations rules.
- ETL process allows sample data comparison between the source and the target system.
- ETL process can perform complex transformations and requires the extra area to store the data.

ETL Process in Data Warehouses:

ETL is a 3-step process:

1. **Extraction:**



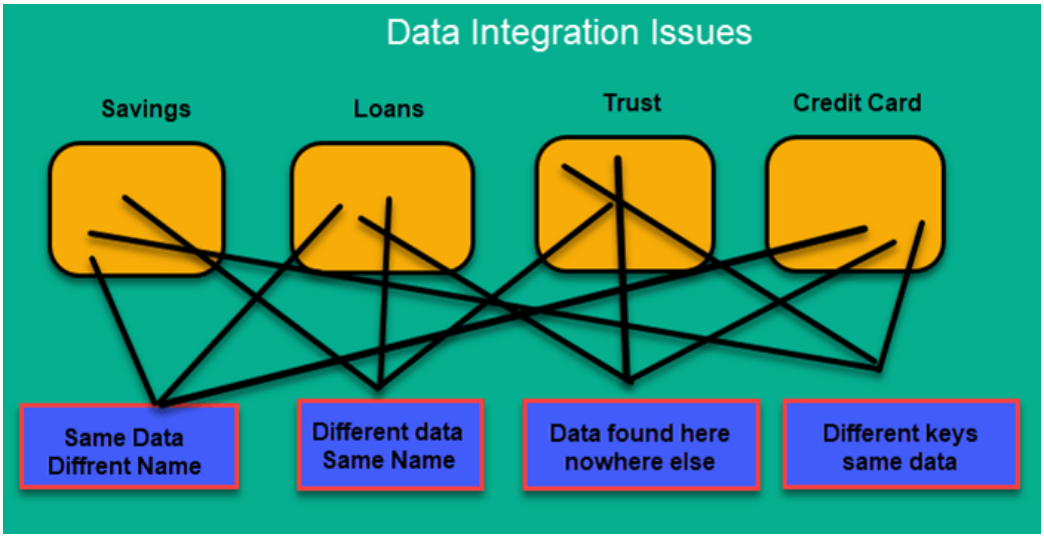
ETL Process

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

Three Data Extraction methods:

1. Full Extraction
2. Partial Extraction- without update notification.
3. Partial Extraction- with update notification.

2. Transformation:



Data extracted from source server is raw and not usable in its original form. Therefore, it needs to be cleansed, mapped and transformed. In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as direct move or pass-through data. Following are Data Integrity Problems:

1. Different spelling of the same person like Jon, John, etc.
2. There are multiple ways to denote company name like Google, Google Inc.
3. Use of different names like Cleave land, Cleveland.
4. There may be a case that different account numbers are generated by various applications for the same customer.
5. In some data required files remains blank
6. Invalid product collected at POS as manual entry can lead to mistakes.

3. Loading:

Loading data into the target data warehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

Types of Loading:

- Initial Load — populating all the Data Warehouse tables
- Incremental Load — applying ongoing changes as when needed periodically.
- Full Refresh —erasing the contents of one or more tables and reloading with fresh data.

Load verification

- Ensure that the key field data is neither missing nor null.
- Test modeling views based on the target tables.
- Check that combined values and calculated measures.
- Data checks in dimension table as well as history table.
- Check the BI reports on the loaded fact and dimension table.

ETL tools:

There are many Data Warehousing tools are available in the market. Here, are some most prominent one:

Mark Logic:

Mark Logic is a data warehousing solution which makes data integration easier and faster using an array of enterprise features. It can query different types of data like documents, relationships, and metadata.

https://developer.marklogic.com/products/

Oracle:

Oracle is the industry-leading database. It offers a wide range of choice of Data Warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

Features:

- Distributes data in the same way across disks to offer uniform performance
- Works for single-instance and real application clusters
- Offers real application testing

Amazon RedShift:

Amazon Redshift is Datawarehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.

Features:

- No Up-Front Costs for its installation
- It allows automating most of the common administrative tasks to monitor, manage, and scale your data warehouse
- Possible to change the number or type of nodes.