**TASK 5 DATA ENGINEERING BASICS 16-03-23**

**TYPES OF ETL LOADS**

Data Load is the process that involves taking the transformed data and loading it where the users can access it.

There are 3 types of ETL Loads

**1. HISTORICAL LOAD**

The Data Warehouse/Mart is expected to house historical data. Based on the duration for how long the end users want to perform the analysis, we keep the data for that long. In other words, we would have users want to compare the Stores Monthly sales and compare it with the monthly sales of for the last 3 years. Here we would have to keep at least 3 years of data so that end users can perform their analysis. When we build and implement data warehouse/mart, it is empty. We would not want to start building the history from the day it is implemented. In this case the end user would have to wait for 3 years from the day of implementation to perform this particular analysis. Hence we identify the source where we can find the history data, and perform a onetime ETL to extract the required history data and load it to the warehouse.

**2. FULL LOAD**

With a full load, the entire dataset is dumped, or loaded, and is then completely replaced (i.e. deleted and replaced) with the new, updated dataset. No additional information, such as timestamps, is required.

For example, take a store that uploads all of its sales through the ETL process in data warehouse at the end of each day. Let’s say 5 sales were made on a Monday, so that on Monday night a table of 5 records would be uploaded. Then, on Tuesday, another 3 sales were made which need to be added. So on Tuesday night, assuming a full load, Monday’s 5 records as well as Tuesday’s 3 records are uploaded – an inefficient system, although relatively easy to set up and maintain. While this example is overly simplified, the principle is the same.

**3. INCREAMENTAL LOAD**

Only the difference between the target and source data is loaded through the ETL process in data warehouse. There are 2 types of incremental loads, depending on the volume of data you’re loading; streaming incremental load and batch incremental load.

Following the previous example, the store that made 3 sales on Tuesday will load only the additional 3 records to the sales table, instead of reloading all records. This has the advantage of saving time and resources, but increases complexity.

Incremental loading is of course much faster than a full load. The main drawback to this type of loading is maintainability. Unlike a full load, with an incremental load you can’t re-run the entire load if there’s an error. In addition to this, files need to be loaded in order, so errors will compound the issue as other data queues up.