Data Warehousing - ETL

Agenda

Here, we will ocver:

- ETL Definition
- Framework of ETL
- Data Extraction - Overview
- Extraction Methods
- Data Transformation Activities
- Aggregation Techniques
- Data Loading
- Types of Load
- Data Extraction vs. Data Loading
- ETL Process Flow

ETL - Definition

- ETL stands for Extract, Transform, and Load, which refers to a process used to integrate data from multiple sources into a target system, typically a data warehouse or a data mart.
- The process involves extracting data from the source systems, transforming the data to fit the target system’s data model and business rules, and loading the data into the target system.
- The process plays a key role in ensuring the accuracy and consistency of the data in the target systems.

Framework of ETL

- Planning and analysis
  - Identify business requirements
  - Define data sources and destinations
  - Create a project plan and schedule
- Data Extraction
  - Extract Data from various sources
  - Validate and clean data
  - Convert data into a common format

Framework of ETL

- Data Transformation
    - Convert data into a desired format
    - Apply business rules and logic
    - Merge and consolidate data
- Data loading
    - Load transformed data into target system
    - Validate and reconcile data
    - Monitor and optimize performance

Data Extraction - Overview

- Data extraction is a process of retrieving data from one or more source systems, which can be structured or unstructured, and transforming it into a common format that can be used for data loading (moving from operational systems to data warehouses, data marts, or other target systems) or analysis.
- The data can come from multiple sources, including databases, flat files, web services, APIs
- The data extraction process can be performed in various ways – including batch processing, incremental processing, or real-time streaming.

Extraction methods

- Full extraction – Involves extracting all the data from the source systems and bringing it into a data warehouse or other centralized location
- Real-time extraction – Involves continuously capturing the data from source systems as soon as it is generated or updated using APIs or web services. Used in scenarios where data needs to be made available for decision-making as quickly as possible, such as financial or fraud detection systems.
- Incremental extraction – Involves extracting only the data that has changed since the last extraction.
- Delta extraction – involves extracting all the records and then filtering out the records that have been already processed – that is, only the deltas or changes since the last extraction.
- Change Data Capture (CDC) – capturing only changes made to data from the source system using approaches like triggers, log files, or database replication.

CDC vs. Delta extraction

- CDC Captures only the changed data in the source system since the last data extraction, while the data extraction method extracts all the records from the source system and filters out the record that has been processed in the previous data extraction.
- CDC is faster compared to Delta extraction, while Delta extraction allows for more efficient and targeted updates to the data warehouse.
- CDC is used when you need to capture every change that happens to a source system’s data in near real-time. In other words when the requirement is when the data warehouse needs to reflect the most recent changes made in the source system.
- Delta extraction is useful when dealing with very large datasets where capturing every change would be inefficient and time-consuming.

Data Transformation Activities

- It is the process of converting data from one format, structure, or type to another in order to prepare it for further analysis or use in downstream applications. The topic covers the following activities -
  - Data Cleaning – Removing for fixing invalid, missing, or inconsistent data values.
  - Data Aggregation – Combining multiple data sources or records to create new or summary records.
  - Data Conversion – Changing data types and formats for compatibility with downstream applications.
  - Data Enrichment – Adding additional data elements to existing data providing more context or insights.
  - Data Splitting – Dividing data into smaller subsets for quick and easier analysis or processing.

Aggregation Techniques

- Can refer to a broader range of techniques to **summarize** or **group** data . The goal is to reduce the amount of data and simplify its analysis. Methods include -
  - Roll-up: this technique involves summarising data from a lower level to a higher level of granularity. For example, data can be rolled up to weekly, monthly, quarterly, or yearly levels.
  - Drill-down: This technique is the opposite of roll-up which involves breaking data from a higher level to a lower level of granularity.
  - Slice and dice: The technique involves selecting a subset of data and analysing it from a different perspective by selecting specific dimensions and measures to slice and dice.

Aggregation Techniques

- Pivot tables: This technique involves summarising and analyzing data in a tabular format, allowing users for quicker analysis and manipulation of large data.
- Grouping: This technique involves grouping data based on certain criteria or attributes for analysis. Examples of such attributes include by categories such as product categories, customer segments, and geographic regions.

Benefits of Aggregation Techniques

- Pre-calculated and pre-summarised data at different levels such as product, region, time, and customer can drive benefits which include -
  - Improved query performance
  - Reduce storage requirements
  - Simplified query development
  - Improved data quality by detecting errors and inconsistencies in data

Data Loading

- It is the process of moving extracted data into the data warehouse after transformation.
- The process includes data validation, data cleaning, and data transformation.
- The purpose of Data Loading is to ensure that the data is in a usable format.
- The data extraction process can be performed in various ways – including full load, incremental load, real-time load, and change data capture (CDC).

Types of load

- Full Load – All data are extracted from source systems and loaded into the target system. Typically used when the amount of data is small, or data is relatively static.
- Incremental Load – Only new or changed data are extracted from the source system and loaded into the target system. Typically used when the amount of data is large or is frequently changing
- Real-time Load – Data is extracted from the source system and loaded into the target system as it becomes available. Used when data needs to be updated in real-time such as in OLTP.
- Delta Load – Loading only the changes or deltas that have occurred since the last load.
- Slowly Changing Dimension (SCD) Load – only the changes to a specific dimension in the source systems are loaded into the target system. Typically used for dimensions that change frequently.

Data Extraction vs Data Loading

Data Extraction

- Process of retrieving data from source systems and making it available for further processing.
- Involves identifying relevant data, filtering and transforming it to fit the target system.
- Can involve multiple sources and data formats.

Data Loading

- Process of loading data into the target system, typically a data warehouse.
- Involves inserting, updating, or deleting data in the target systems.
- Typically involves a single target system and a standard data format.

ETL Process Flow

- Extract – Raw data is extracted from source systems and bought into a staging area or temporary location for further processing.
- Clean – Cleaning the data to remove any errors, inconsistencies, or duplicates. This stage ensures data is reliable and accurate.
- Transformation – Structured and convert the data to the format that matches the target repository. Consistency, accuracy, and completeness are ensured by applying business rules, data quality checks, and formatting changes.
- Load – Loads the cleansed, transformed, and structured data into the data warehouse.
- Validate – in this stage, the data is validated to ensure that it has been correctly loaded and meets the expected quality standards. The stage involves profiling, data quality checks, and reconciliation of data between the source and target systems.

Summary

A brief recap:

- ETL stands for Extraction, Transformation, and Load, which is the process of extracting data from various sources, transforming it to meet the desired format, and loading it into the data warehouse.
- Data Extraction is the process of extracting data from various sources. The main techniques include Full Load, Incremental Load, and CDC.
- Data Transformation involves cleaning, structuring, and consolidating the extracted data to fit the desired data model.
- Aggregation Techniques, like Roll-up, Drill-Down, Grouping, Slice and Dice, are used in ETL process to summarise large volumes of data to provide insights.

Summary

- Data Loading involves loading the transformed data into the data warehouse. Types of load include Full Load, Incremental Load and CDC
- ETL Process Flow involves five stages – Extraction, Cleaning, Transformation, Loading and Validation. The process helps ensure that the data warehouse is accurate, and have latest information.