Heterogeneous Data Integration with ELT and Analytical MPP Database for Data Analysis Application

Sivabalan S
Computer Science and Engineering
SRM Institute of science and technology
Chennai, India
ss4699@srmist.edu.in

Minu R I
Computer Science and Engineering
SRM Institute of science and technology
Chennai, India
minur@srmist.edu.in

2021 Innovations in Power and Advanced Computing Technologies (i-PACT) | 978-1-6654-2691-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/i-PACT52855.2021.9696841

***Abstract*—In a data analysis project, data integration is performed on a large volume of data collection from heterogeneous sources. Projects need to collect data from Web API, nonrelational, Relational databases, Edge AI, Connected Vehicles, Structured and Unstructured Files, and system logs. Before storing, data will be collected, Validated, and Transformed. This process is created in Extract Load Transform (ELT) and scheduled. Traditional Extract Transform and Load (ETL) systems process the data before loading it into the database, and the data transformation needs dedicated resources in the ETL server. Projects often encounter performance issues and high hardware and licensing costs while processing extensive data set, and this is handled efficiently in Extract Load and Transform (ELT) and Analytical Massively Parallel Processing (MPP). Data transform will take place in MPP Databases, and the resources are highly scalable in the cloud. ELT with Analytical MPP makes data analysis much faster and more efficient. We will discuss the roles of ETL and ELT tools in data integration for a successful data science application.**

Keywords—Data Integration, Data Cleansing, MPP Database, Data warehouse, ETL and ELT

INTRODUCTION

Business intelligence projects start from Collect, Process and store data into a single target database to find insight for the business. Before conducting data analysis, it is essential to clean, remove data redundancy and enrich the data to achieve the best results. Integration involves moving the data between different applications to keep them synchronized. The application integration is ideal for powering operational applications. For example, duplicate customer records have a customer support system. Typically, every application has a particular way to transmit and accept data, and these data move in smaller quantities. Traditional ETL tools need high configuration as the data transformation is done before loading the data [1]. The conventional method requires high hardware and license cost to process a larger dataset. ELT tools and MPP database are adequate to handle a high volume, Variety, and velocity data. The MPP data bases are available in cloud as a service and the cost is depends on the utilization of space and resource used for computation. The migration of data from legacy database to cloud-based database can be migrated using the ELT programs.

Every organization transitioning significant data developments from on-premises to the cloud can profit greatly. They make it possible to save data as business expected; business intelligence (BI) application retrieves the stored data and represents the data with visualizations. Database handles Complex logic, and they also result in improved performance.

ELT gives the impression to be the future for data integration, which fixes the performance concerns that ETL encounters. The organization's data volumes have risen dramatically. ETL systems are unable to integrate all this data efficiently into a database. ELT provides greater agility and requires fewer admin activities which save cost for enterprises.

I. DATA WAREHOUSE

A Data warehouse is a multidimensional database that stores a large volume of data. Data warehouse stores data in Dimension and fact data, Categorizing Fact and dimensions reduces data redundancy and improves the data retrieval for visualization [2]. ELT works to load the data from source to Stage, Operational data source, and Data warehouse tables. The Data presentation layer is the final stage layer is the Data analytics helps gain insight into future corporate performance, which is crucial in the decision-making process. The most renowned implementation of data integration is the likely data warehouse building for a company. The benefit of data storage enables data to be analyzed based on the data in data storage. Therefore, ELT plays a critical role to load the business data into the Data warehouse [4].

Traditional ETL architectures are monolithic, frequently connected exclusively to schema-based data sources, and have little or no room to process data streaming at rapid speeds. As a result, it's nearly complicated for ETL tools to collect the complete or a portion of the source data into memory, execute transformations, and then load it to the warehouse when businesses deal with high velocity and veracity of data. The high volume of data could cause delay in data loading. The ETL programs makes the works easier by filtering the records that are less significant. Data further aggregated and the data redundancy is reduced in the database.

978-1-6654-2691-6/21/$31.00 © 2021 IEEE

II. DATA INTEGRATION

Data integration simplifies business intelligence (BI) analysis processes by providing a unified view of data from various sources. It combines multiple data sources into a data warehouse and is responsible for loading data into multiple layers such as Stage, Operation data store (ODS), and Data warehouse (DWH). Therefore, data integration must be collaborative and unified across the enterprise to improve collaboration and Integration [3].

This process converts Data Integration to perform Data Extract, Transform and Load actions to move data to the destination with the applied logic. This process works to combine multiple data sources to carry out the analytical activities.

A. Extract

Extract reads data from different source systems. Structured/semi-structured data can be the source. The primary goal of the Extract Part is to receive and load data from various sources into RDBMS. Only new or changeddata can be saved in the staging table in Incremental Load. This reduces over and over the reprocessing of the data. Various sources can be integrated into the Stage layer of the Data warehouse Application. The data is transformed and loaded into final DWH tables [9].

B. Transform

The actual business rules on data are applied in this layer. The Dimension and Fact data mapping logic will be established on this layer.

- Decision-making conditions like If, Decode, Null checks will be done on source data.
- The data conversion process is carried out, such as data format modifications, math functions, aggregation operations, and data joints.
- Duplicate removal, Dimension, and fact data mapping, data transform and data normalization are carried out.

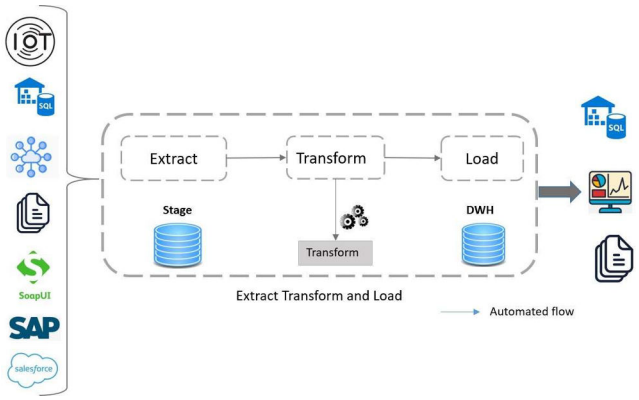

Fig. 1. ETL Data Integration

ETL has the connecters to integrate the various upstream systems and the services works from ETL to pull the records from the source systems.

Data cleansing is important in the data science domain as this might cause inaccuracy due to Data Redundancy, Noise in data, Data fields that are no longer required for analysis. ETL tools are equipped with a pre-Written program that comes in handy to enrich the source data for analysis. This also reduces storage as only the required data is stored without any duplication.

- Remove the redundancy of data and store in one location
- Ensure that all corresponding information is stored together, i.e. data dependency
- The Data segregated as Dimension and Fact data.
- The data set is transformed as per the business rules.
- The data is checked for discrepancies and missing values.
- The Data segregated as Dimension and Fact data.

Normalization is required to store the Master and Trans action data in separate tables. This helps to redundancy and improves the performance of the system. The tables are related to foreign key and surrogate key relationships. The most significant phase in the ETL process is transformation, which is usually regarded as the most important. Data transformation improves data integrity and guarantees that data arrives at its new place fully compliant and ready to use.

C. Load

Fig 1 shows the stages in the ETL process, the newly transformed data is imported into the target. Data can be loadedin bulk (full load) or at predefined intervals called as incremental load. In an ETL full loading scenario, everything that comes off the transformation assembly line is turned into new, unique entries in the data warehouse. Though this is sometimes usefulfor study, comprehensive loading results in data sets that grow exponentially and become difficult to manage [7].

Incremental loading is an approach that is less comprehensive but more manageable. When new data is compared to existing records, incremental loading creates new records only when new and unique data is detected. Smaller, less expensive data warehouses might keep and manage corporate intelligence using this technique. Load is the layer to get the data loaded into final DWH tables and the summary is created [5]. DWH tables undergo Slowly changing Dimension (SCD) options, and this is also called reporting layers, where BI applications access the data from this layer.

III. ELT

The ELT is a different approach to the classic ETL. ELT has the advantage of leveraging databases to drive transformation for improved performance. This ability is useful for the processing of the massive BI and Big Data analysis data sets. ETL is time-consuming and requires additional ETL- specific resources like CPU, RAM, and separate licenses for data processing.

Fig. 2. ELT Data Flow

Fig 2 shows that ELT Tools could overcome this by loading the data set and then applying the transformation logic in the MPPdatabase. Before a data model is developed, ETL can affect your business data in ways that ELT cannot. This intermediate phase of the ETL process allows for more extensive datatransformation before the data is loaded.

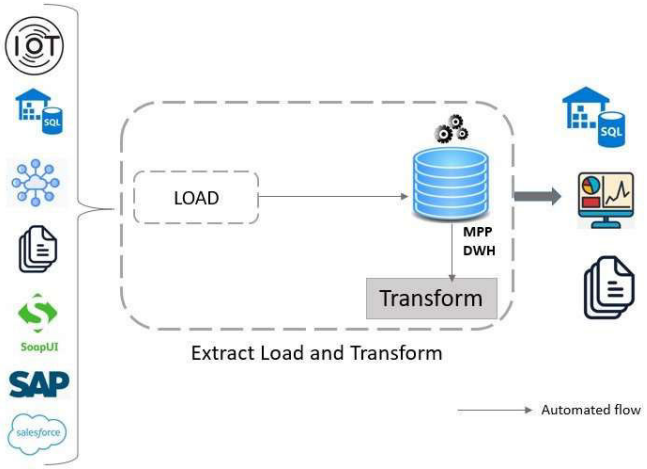

Fig. 3. ELT Data Integration

A. ELT with Analytical MPP Database

Analytical Massively Parallel Processing (MPP) databases have the efficiency in performing high analytical processing. They could carry out data transformation and processing for large datasets. A large volume of data is distributed in the cloud and performs data operations. The highly elastic cloud architecture utilizes dynamic resources and brings down the processing time considerably when compared with on-premiseETL tools.

The dataset is distributed across many nodes to process a high volume of data. The data transformation logic is distributed to the nodes, and processing is done in parallel. Asymmetric multiprocessing assigns each unique processor task. One large processor, known as the master, oversees the operations of the other subsidiary processors.

All processors are available to all individual processes in symmetric multiprocessing. As a result, performance is improved because the task is distributed more or less evenly among the processors. Symmetric multiprocessing is also known as tightly coupled multiprocessing because the multiple processors share the computer's memory and I/O resources while using only one instance of the operating system.

- Grid computing This architecture makes use of resources as needed, based on their availability. This architecture saves money on server space, but it also restricts band- width and capacity at peak periods or when there are many requests.
- Database clustering is the process of connecting a single database with several servers or instances. When one server is insufficient to handle the volume of data or the number of requests, a Data Cluster is required.

Massive parallel processing (MPP) is a method of crunching vast amounts of data by dividing the task among hundreds or thousands of processors, which may be housed in the same box or on separate, remote machines. Each processor in an MPP system has its memory, storage, programs, and operating system. Thus, the problem is divided into several components, each of which is handled by a different system at the same time.

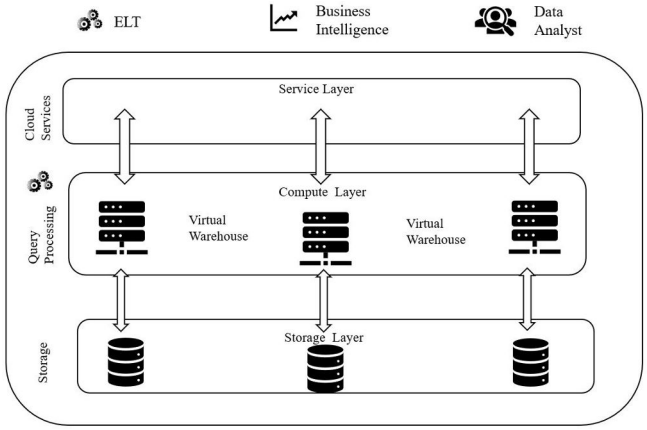

Fig. 4. ELT Data Integration

Fig 4 shows the MPP analytical databases are typical as data stores, central transaction sales, tracking, marketing information, personalized data services, inventory/logistical data, HR, and system recording data for organizations. Since numerous data are processable via MPP databases, such databases can be easily used by an organization to store data and support analytical workloads from the various business functions.

B. ETL and ELT Critical Differences

The processing differs between ETL and ELT while loading the data. It has been observed that ELT has high performance than the traditional ETL tool. Below is the list of the critical difference between ETL and ELT [6].

- Transformations are handled by the database, which reduces the importance of the staging layer.
- In ETL, All the source data is processed in the ETL server: this Required additional CPU, RAM, and storage in Giga Bytes. When the data grows exponentially, then Additional CPUand RAM have to be procured along with additionallicenses.
- Data loading is faster because there are no transformations to wait for, and the data is just loaded once into the destination data system. However, data analysis takeslonger than ETL.
- ELT tools are more suitable for cloud-based data warehouses like Snowflake, Microsoft Azure, and Redshift Databases. Cloud databases accommodate rapid volume growth in lesser time than traditional on-premises databases.
- Transformations are faster since they happen after load-

ing, on an as-needed basis—and you transform only the data you need to examine at the time [8]. The necessity to transform data on a regular basis, on the other hand, slowsdown the total time required for querying and analysis.

C. ETL vs ELT Data Migration

In an ETL load, If the Business wants to join employee details from multiple systems into one database. The ELT will get all source data and perform data transformation to get a single record with Insert, Update, or deletion against the target Database as shown in Fig 3.

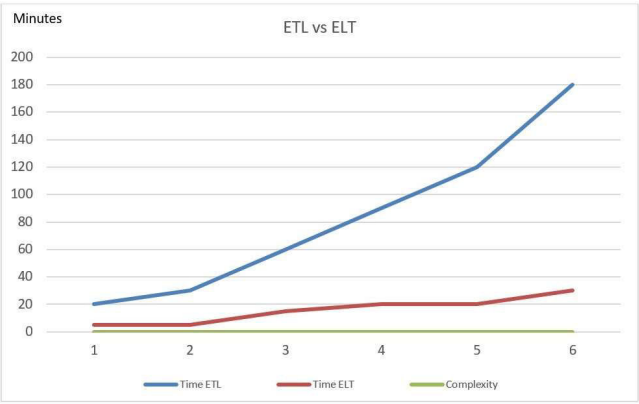

Fig. 5. ELT Data Integration

The data from multiple source systems is loaded into the target database. In this model, the data join and business logic are applied to the target database instead. This eliminates the data transformation while reading the data. The data connection can be closed as soon as the data is loaded to the target database. Fig 5 shows the data load comparisons. Once the data is stored in the database,the transformation is applied to the target database.

Many ELT/ETL tools are used in projects, and below are some from the list • IBM Data stage • Informatica Powercenter • Microsoft SSIS • SAP HANA • SAS • Talend Data integration studio • Teradata

IV. CASE STUDY

In the Shipping industry, Cargo ships carry nearly twenty thousand containers and make many transits. Each container could contain multiple deliverables, which would come to more than 2 Million deliverables. One cargo company could operate with 500 cargo ships, and the complete deliverables would come around 1000 million transactions in a month. Following are the steps to build an ELT job to load data into DWH tables. The Files are transferred to the ELT server through FTP.

The ELT jobs first load data into the Stage (STA), and then Operation Data Warehouse (ODS), later to Datawarehouse (DWH) layers.

A. Data modeling

Data modelling needs to be done after understanding the data and its relationships. Then, the modelling tools are usedto establish the relationship between the tables. There is a feature to get the DDL of the tables as an outcome of data modelling. Tables included Dimensions and Fact tables were created before building the ELT flow, and the foreign key relationships were established. ELT tools handle the load inall the tables.

B. Stage (STA)

This Stage layer is where the data enters the Datawarehouse. The data from the files will be imported into database tables. The data contains essentially small data changes and is loaded incrementally. Example: Conversion Date formats, Number formats, etc. Most of the time, the stage is truncated andthen loaded. The stage layer could feature History tables to store history data. In addition to the data conversion stage also handles incremental load. This assists with data duplication at the start of the data process. This reduces the workload of ELT results in load optimization. Incremental Processing:

The ELT job will reject the false records because they are processed in previous loads. The stage jobs will be more in numbers as for each file, one stage table will be created. The stage table will be having dedicated jobs for each file load. The jobs also stored the source file name in the staging table to help understand the data source [10].

- Loads Calculated columns like load date, updated by, source file name
- Performs incremental load.
- All the file and source extracts are placed in a sequence.

C. Operational Data Store (ODS)

The business logic is handled by the ODS, which is the critical layer. The ODS layer will run data lookup, aggregations, joins, and data filtering. Dimensions and Facts were created from the source data in the ODS layer. This lowers data redundancy and reduces data size. Example: Currency Conversion from USD to EUR. Joins the data with country-specific tables. Calculating Profit and loss for each transaction. Calculating Number of days taken to deliver the shipment, etc.

D. Datawarehouse layer (DWH)

Data warehouse tables containing the final version of Dimension and Fact data. The table data is organized using a data model. The two types of tables are described below. Dimension: The dimensions that characterize the data in the fact table are stored in the dimension tables. In the Dimension table, primary keys are defined. For example, country, Date, Cargo Ship, and Forex data are all kept in Dimension tables in a transaction, and each row has a foreign key relationship with Fact tables. The transactional data is storedin the fact table as numeric. In addition, the dimension data is linked to the foreign key in the Fact table. The ELT performs the following data processing.

- Maintaining Slowly changing Dimension data (SCD Types).

- Loading Fact and Dimension data.
- Insert / Update the dimension data.
- Transforms the data.

The same load process will be time-consuming due to the complete data is processed in the ETL server. The Stage first read the data and loads it into stage tables. The business logic such as Joins, Sorting needs cache memory, and the ETLserver memory is utilized. The time to take the load in ELT is less as all the data operations are performed at the DB level.

V. CONCLUSION

In Data Science, ETL or ELT tools perform data integration from various sources. These tools can handle both Structured and unstructured data for analysis purposes. The data is cleansed, enriched, and loaded for analysis. ETL operates with on-site and cloud-based data stores. Therefore, a related or structured data format is necessary. ETL is best for handling smaller data sets requiring complex transformations. ELT is the best way to handle large amounts of structured and unstructured information. In addition, ELT is the speedier solution when it comes to data availability.

REFERENCES

[1] L. Munoz, J. Mazon and J. Trujillo, ”ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study,” in IEEE Latin America Transactions, vol. 9, no. 3, pp. 358-363, June 2011.

[2] Rifaie, Mohammad, Keivan Kianmehr, Reda Alhajj, and Mick J. Ridley. "Data warehouse architecture and design," IEEE International Conference on Information Reuse and Integration, pp. 58-63, August 2008.

[3] Anureet Kaur. “Big data : A review of challenges, tools and techniques,” International journal of scientific research in science, engineering and technology, vol. 2, pp. 1090–1093, 2016.

[4] Waas F, Wrembel R, Freudenreich T, Thiele M, Koncilia C, Furtado P. “On-demand ELT architecture for right-time BI: extending the vision,” International Journal of Data Warehousing and Mining (IJDWM), vol. 9, pp. 21-38, Apr 2013.

[5] Mandeep Kaur Sandhu, Amanjot Kaur and Ramandeep Kaur, “Data Warehouse Schemas,” (IJIRAE) International Journal of Innovative Research in Advanced Engineering, vol. 2, pp. 47-51, April 2015.

[6] D. M. Tayade, “Comparative Study of ETL and E-LT in Data Warehousing,” Int. Res. J. Eng. Technol, vol. 6, pp. 2803– 2807, 2019.

[7] Sreemathy, J., S. Nisha, and Gokula Priya RM. "Data integration in ETL using Talend," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1444-1448, 2020.

[8] Li, Zehai, Jigui Sun, Haihong Yu, and Jian Zhang. "Common cube-based conceptual modeling of ETL processes," In 2005 International Conference on Control and Automation, IEEE, vol. 1, pp. 131-136, 2005.

[9] R Elmasri and S Navathe “Fundamentals of Database Systems,” Addison-Wesley Publishing, vol. 4, no.7, pp. 321-329, Nov 2012.

[10] P. S. Diouf, A. Boly and S. Ndiaye, ”Variety of data in the ETL processes in the cloud migration and validation : State of the art,” 2018 IEEE International Conference on Innovative Research and Development(ICIRD), Bangkok, pp. 1-5, 2018.