Task 4

16 March 2023

ETL:

1. ETL stands for Extract Transform Load which is a data integration process. In short, it gathers data from different sources and loads a cleaner version of them in to a target system.
2. In extract part of it, the data is extracted from different sources. Those sources could be any thing ranging from databases to CRM or ERPs etc.
3. In the transform part of ETL, the data which is extracted from different sources, is passed through a series of transformations according to the requirements such as data cleaning,wrangling etc aso that it can be loaded in the target system.
4. In the final Load part, the transformed data is then loaded into a target system which can be a data warehouse or a data lake. From there onwards, it consumed by the data consumers such as data scientists/analysts to extract useful insights and help in making informed business decisions.
5. ETL can be manually done through programming languages such as Python (pandas) but it is now a days shifting to cloud like AWS Glue.
6. ETL is best suited for structured small data which needs to be transformed before being loaded to target system and is hosted on-premise. It can take more time to process as data first needs to be transformed and then loaded, therefore it also offers less flexibility in terms of data structure.
7. ETL is used many of the modern data driven organisations because data can be easily accessed, analyzed and visualised. It also ensures that data is consistent, reliable and ready for extracting insights and patterns.

ETL Tools:

* Microsoft SQL Server Integration Services (SSIS)
* Apache Nifi
* Informatica PowerCenter

ELT:

1. ELT stands for Extract Load Transform. It is also a data integration process but with a slight difference.
2. Although the first part is also extract, the second part is replaced by load instead of transform like in ETL. This suggests in ELT the data after being extracted from different sources is directly loaded in to the target system (like datawarehouse, data lake etc) instead of loading it in to a staging area like in ETL.
3. So after directly loading into target, transform part of the process takes place within the target source.
4. ELT is mainly used where there is huge amount of data and there is a need for direct and quicker access to data, it can be a realtime data environment like stock data and also where the data has to be processed in cloud instead of on-premise. It is also flexible in terms of data structure but the target system requires more transform power once loaded with data.
5. ELT has gained more heights after it has been given support by cloud data warehouses.

Three Tier Architecture:

1. It is a client-server architecture for data driven applications which separates those applications into 3 layers; presentation, application and data layer.
2. Presentation layer is responsible for the UI of the application. It displays the application and allow users to interact with the data shown, so it also collects inputs from user.
3. Application layer is responsible for managing and processing the application and business logic. It handles the user requests and does necessary computations such as data validation, data manipulation and busines logic on the data which it collects from presentation layer. So it is a kind of middleman which does the necessary work to make sure that correct data flows between presentation and data layer.
4. Data layer is responsible for managing and retreiving the data stored in data storage systems. It keeps check on the data validation and consistency. So it also processes and stores the information processed by application.
5. These three tiers are logical, not physical and so they all may or may not be running on same server at one time. The application layer can also work like a firewall which prevens direct SQL injections on data layer.
6. Finally it is also a client server architecture but with addition of application layer in between.