Includes Redshift,Synapse,Lambda,Airflow

**Day 25(11/12/23)**

* Data modelling
* MPP-massively parallel processing, dividing tasks into small processes for faster speeds, strategy followed by Redshift
* Data is collected from various sources and stored in different different sources, it needs to be integrated and processed.
* Data Warehouse is –

1. Subject Oriented-data is stored on the basis of subject rather than business application.
2. Integrated- data collected from various sources and stored together
3. Time-variant- data stored as a series of snapshots
4. Non-Volatile- data is not updated or deleted in a warehouse as it is archival data

* Data mart is a small version of Data WareHouse which has only one subject and hence occupy very less area.3 types- Dependent,Independent and Hybrid
* Data Warehouse schemas-

1. Star Schema

Contains of a FACT table with single table for each dimension

1. Snowflake Schema

Type of star schema where it has multiple level of dimension tables

1. Constellation Schema

Known as galaxy schema as one dimension table has multiple FACT tables

* Types of clusters

1. Tightly Couple-storage and compute connected when we increase storage we need to increase the computing power.

Dc2

Storage is better than computing power

Ds2

Computing power is better than storage

1. Decouple

RA3

* Data distribution in Redshift

1. Even- distribute data evenly across all nodes without any specific column values,good for small tables
2. Key- distribution based on some specific column,key should be known and key should be used to join the tables
3. All- a full copy of the entire data is maintained in each node

* Always make table in dev and not sample\_data\_dev

DAY 27

* Docker and Airflow
* DAG is a Directed Acylic Graph which is a collection of all the tasks you want to run organized in a way that shows the relationships.

1. Directed- The tasks have a defined order.One task follows another.
2. Acyclic- There are no loops in the task order.A task doesn’t run more than once for a run.
3. Graph- It visualizes the order and relationships between tasks.

* DAG in Airflow

1. Airflow uses DAG to define workflows and each task is represented as a node in the graph
2. Operator is a single atomic task that we want to execute ,there are multiple operators like:
3. PythonOperator-Calls a Python function
4. MySQLOperator-executes an SQL command
5. EmailOperator-Sends an email

* Docker helps us in hosting applications in virtual env that is not VM or on cloud.

Day 28

* MPP vs PP- Massively Parallel Processing data is divided into multiple systems for higher speeds of processing. Parallel Processing is dividing data into various partitions of a single systems memory
* ADLS-Azure Data Lake Storage

Day 29

* Dedicated SQL pool is like a SQL DB in Synapse analytics
* Indexing in SQL should be based around the queries