Glossary

Airflow Hooks: Hooks provide a reusable interface to external systems and databases. With hooks, you don't have to worry about how and where to store the connection strings and secrets in your code.

Data Lineage: The data lineage of a dataset describes the discrete steps involved in the creation, movement, and calculation of that dataset.

Data Validation: Data Validation is the process of ensuring that data is present, correct & meaningful. Ensuring the quality of your data through automated validation checks is a critical step in building data pipelines at any organization.

Database (Components of Airflow): Saves credentials, connections, history, and configuration

Directed Acyclic Graphs (DAGs): DAGs are a special subset of graphs in which the edges between nodes have a specific direction, and no cycles exist. When we say "no cycles exist" what we mean is the nodes cant create a path back to themselves.

Edges: The dependencies or relationships other between nodes.

End Date: Airflow pipelines can also have end dates. You can use an end_date with your pipeline to let Airflow know when to stop running the pipeline. End_dates can also be useful when you want to perform an overhaul or redesign of an existing pipeline. Update the old pipeline with an end_date and then have the new pipeline start on the end date of the old pipeline.

Logical partitioning: Conceptually related data can be partitioned into discrete segments and processed separately. This process of separating data based on its conceptual relationship is called logical partitioning. With logical partitioning, unrelated things belong in separate steps. Consider your dependencies and separate processing around those boundaries.

Nodes: A step in the data pipeline process.

Operators: Operators define the atomic steps of work that make up a DAG. Airflow comes with many Operators that can perform common operations.

Schedule partitioning: Not only are schedules great for reducing the amount of data our pipelines have to process, but they also help us guarantee that we can meet timing guarantees that our data consumers may need.

Scheduler (Components of Airflow:) orchestrates the execution of jobs on a trigger or schedule

Schedules: Data pipelines are often driven by schedules which determine what data should be analyzed and when.

Size Partitioning: Size partitioning separates data for processing based on desired or required storage limits. This essentially sets the amount of data included in a data pipeline run. Size partitioning is critical to understand when working with large datasets, especially with Airflow.

Start Date: Airflow will begin running pipelines on the start date selected. Whenever the start date of a DAG is in the past, and the time difference between the start date and now includes more than one schedule intervals, Airflow will automatically schedule and execute a DAG run to satisfy each one of those intervals.

Task Dependencies: Describe the tasks or nodes in the Airflow DAGs in the order in which they should execute

Web Interface (Components of Airflow): provides a control dashboard for users and maintainers

Work Queue (Components of Airflow): holds the state of running DAGS and Tasks

Worker processes (Components of Airflow): that execute the operations defined in each DAG