Day 7 – Workflow & Job Orchestration

Databricks Notebooks vs Databricks Jobs

📓 *Databricks Notebook*

•*Used for development & exploration*
•*You run cells manually*

Good for:
-*Learning*
-*Testing transformations*
-*Debugging*

Example:
•*Create bronze table*
•*Try transformations*
•*Visualize data*

⚙️ *Databricks Job*

•*Used for automation & production*
•*Runs automatically*

Can be:
-*Scheduled (daily/hourly)*
-*Triggered manually*

Supports:
-*One notebook OR*
-*Multiple notebooks (workflow)*

What is a Workflow?

A Workflow is a pipeline of tasks

Tasks can be:
- *Notebooks*
- *SQL*
- *Python scripts*

Example:
Bronze → Silver → Gold

Each step:
- *Runs only if the previous one succeeds*
- *Ensures data quality & order*

Multi-Task Workflow (Medallion Flow)

Typical structure 👇

Task 1 (Bronze) -Load raw data
↓
Task 2 (Silver) -Clean & transform
↓
Task 3 (Gold)- Aggregate for analytics

This is exactly Medallion Architecture in action 🚀

Parameters & Widgets

Why Use Parameters?

•Avoid hard-coding values

•Reuse same notebook for multiple tasks

Makes pipelines dynamic

Examples:

-Layer name

-Run date

-File path

Databricks Widgets

Widgets Example

- *Widgets accept user input*
- *Values passed from jobs*

Benefits:

- *Same notebook → multiple layers*
- *Easy configuration from UI*

✔ *Reusable*

✔ *Scalable*

Error Handling

Why Error Handling is Important?

- *Jobs run automatically*
- *No manual supervision*

Basic strategy:

- *Validate data*
- *Fail fast if something is wrong*

Result:

- *Job stops*
- *Downstream tasks protected*

Scheduling Jobs

Job Scheduling

- *Daily / Hourly / Weekly*
- *Time-based execution*
- *No manual run required*

Example:

- *Run pipeline every day at 6 AM*

✔ *Production-ready pipelines*

Real-World Pipeline Flow

End-to-End Automation

1.Raw data arrives
2.Bronze loads data
3.Silver cleans data
4.Gold prepares analytics
5.Dashboard consumes Gold data

➡️ *Fully automated system* 🚀