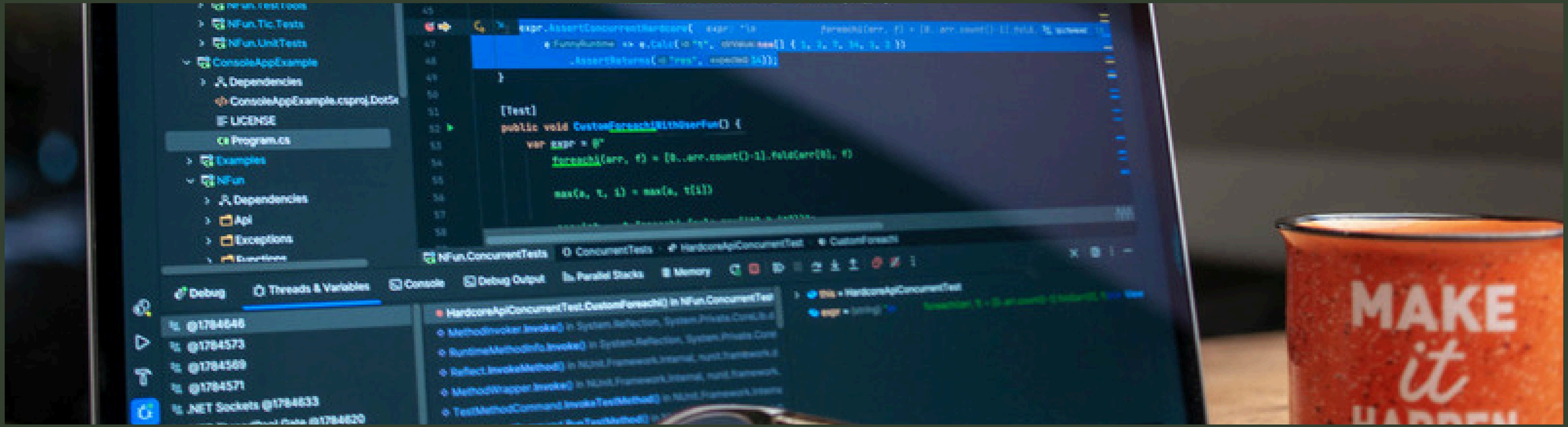
Understanding Airflow, Prefect, and Pentaho



Presented by :

Pavani Korada

Introduction

Modern data projects involve many steps like collecting, cleaning, and processing data. To manage these steps efficiently, we use tools that help automate and monitor data pipelines.

Apache Airflow and **Prefect** are used for workflow orchestration, meaning they control when and how tasks run. **Pentaho** is mainly used for ETL, meaning it focuses on extracting, transforming, and loading data using a visual interface. This presentation explains these three tools and compares them.

Apache Airflow

What is Apache Airflow?

- Apache Airflow is an open-source workflow orchestration tool.
- It is used to schedule, monitor, and manage data pipelines.
- Workflows are written in Python and defined using DAGs (Directed Acyclic Graphs).
- It is mainly used for large, complex, and batch-oriented pipelines in production systems.

“Pipeline as a code.”

Challenges with Apache Airflow

- Complex and heavy installation process
- Steep learning curve because of DAG concepts
- Requires good Python and system knowledge
- Debugging errors can be time-consuming
- Too much setup for small or simple workflows

Purpose of Using Apache Airflow

- To schedule and automate complex data workflows
- To manage task dependencies in large pipelines
- To monitor workflows using a web interface
- To handle retries, failures, and logging in one place
- To run production-grade data pipelines reliably

PREFECT

What is Prefect?

- Prefect is a modern workflow orchestration tool.
- It is Python-first and easier to use compared to Airflow.
- It is mainly used for data pipelines, automation, and machine learning workflows.
- It provides better monitoring, error handling, and easier setup

"Build workflows that don't break."

✳ Challenges with Prefect

- Requires good understanding of Python programming
- New concepts like deployments, agents, and work pools take time to learn
- Smaller community compared to Airflow
- Some features depend on cloud or server setup
- Still evolving, so frequent updates and changes

✳ Purpose of Using Prefect

- To build workflows using simple Python code
- To get better error handling, retries, and logging automatically
- To monitor workflows easily using a modern UI
- To reduce setup and infrastructure complexity
- To manage data and ML pipelines in a flexible way

PENTAHO

What is Pentaho?

- Pentaho is a data integration and ETL tool.
- It is mainly used for Extract, Transform, and Load (ETL) processes.
- It provides a graphical drag-and-drop interface.
- It is widely used in data warehousing and business intelligence projects.

"Design data pipelines the visual way."

Challenges with Pentaho

- Not suitable for complex workflow orchestration
- Visual pipelines become hard to manage when they grow big
- Less flexible for dynamic or conditional workflows
- Limited automation compared to Airflow and Prefect
- Not ideal for advanced scheduling use cases

Purpose of Using Pentaho

- To build ETL pipelines using drag-and-drop interface
- To integrate data from multiple sources easily
- To clean, transform, and load data into data warehouses
- To reduce coding effort for data integration tasks
- To support reporting and business intelligence workflows

Core Purpose

✓

Airflow

Airflow is used when you have big, complex, batch jobs and many steps that must run in a fixed order.

✓

Prefect

Prefect does the same job but in a more modern and flexible way, especially for ML and dynamic workflows.

✓

Pentaho

Pentaho, on the other hand, is not mainly a workflow scheduler — it is a complete ETL tool used to move, clean, and transform data.

→

So, while **Airflow** and **Prefect** decide when and in what order things run, **Pentaho** focuses on how data is transformed.

Workflow Design Style

How Pipelines Are Created

Tool 1: Apache Airflow

Airflow uses Python code and DAG structure, which is strict and structured.

Tool 2: Prefect

Prefect uses normal Python functions, which is more flexible and easier to write.

Tool 3: Pentaho

Pentaho uses a visual drag-and-drop interface, with very little coding.

Target Users

Who Uses These Tools?

Airflow is mainly used by data engineers and platform teams.

Prefect is used by data scientists and developers who want things to be easier and faster.

Pentaho is mainly used by business users, analysts, and ETL teams.

Installation & Setup

Airflow

Airflow needs multiple components like scheduler, database, and web server, so setup is complex.

Prefect

Prefect is much easier to set up and works well locally or in cloud.

Pentaho

Pentaho is like installing normal software and opening it.

Error Handling & Monitoring

Airflow can handle errors well, but you must define everything manually.

Prefect is very smart with failures, retries, logs, and monitoring built-in.

Pentaho handles errors inside the visual ETL flow, but it’s not very advanced.

Conclusion

Airflow and **Prefect** are workflow orchestration tools, where Airflow is powerful but complex and Prefect is modern and easier to use.

Pentaho is mainly a visual ETL and data integration tool. The choice of tool depends on project complexity, team skills, and the type of data work required.

Thank You So Much!

any questions?