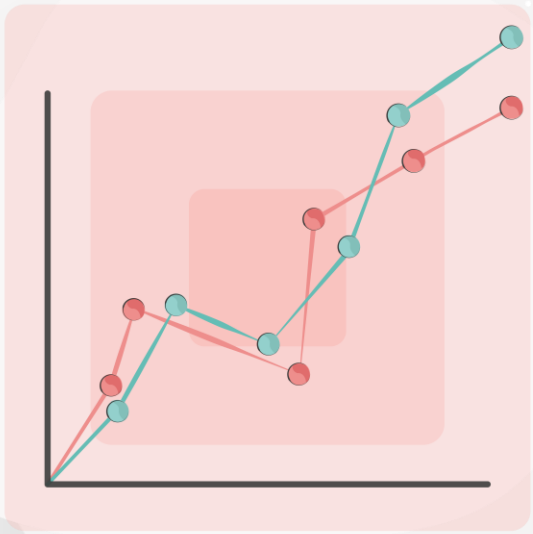
Data engineering

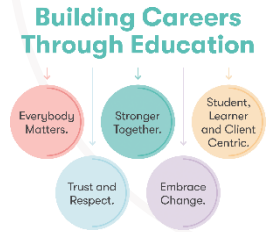
Module: Data pipelines

Topic: Workflow management for data pipelines



Welcome to today's webinar.

Ice breaker



Discussion…

- How are you feeling today? Motivated, happy etc?
- What is your key takeaway from the e-learning topic?
- What is one key skill or insight you hope to gain from today's session?

Submit your responses to the chat or turn on your microphone

BPP

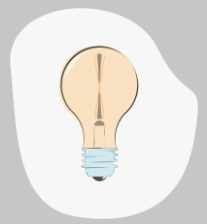
Why workflow management matters

Building Careers Through Education

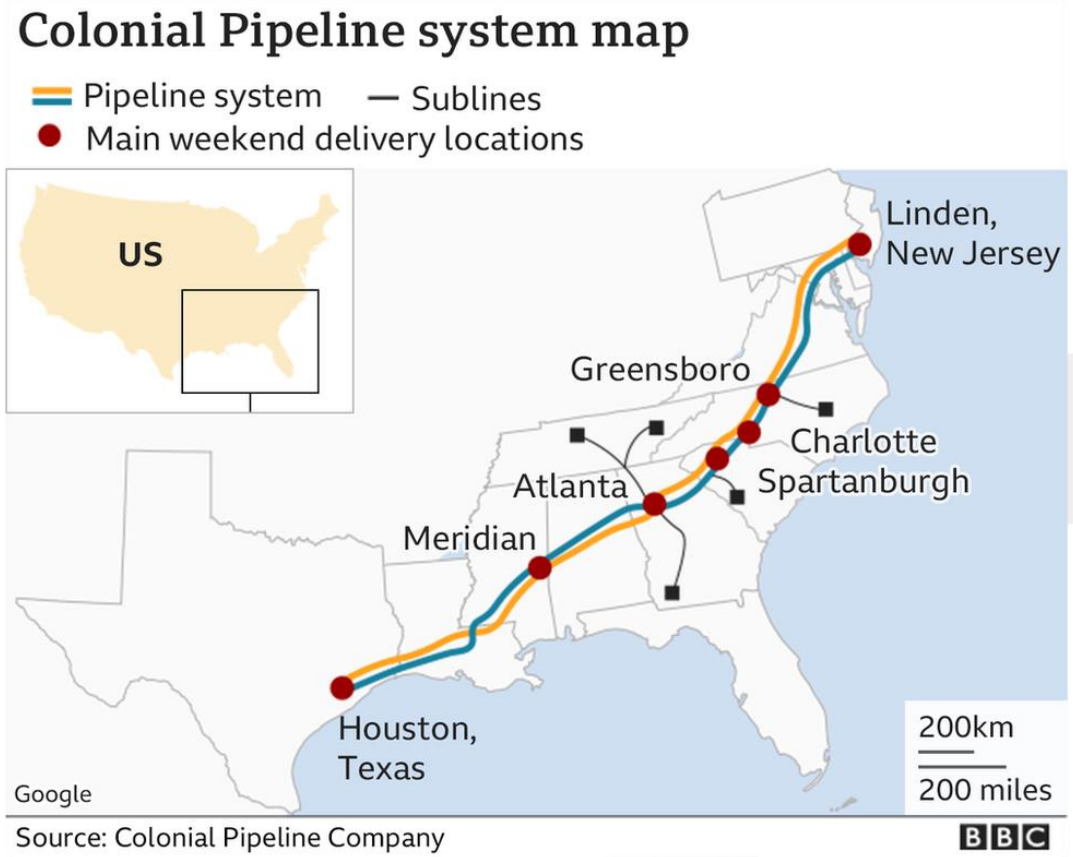
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Case study: The cost of broken pipelines…

- Missed runs = outdated reports → misleading decisions
- Failed jobs = blocked processes → operational delays
- Poor scheduling = system overload → performance bottlenecks



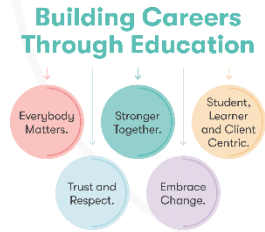
Real-world example: Colonial Pipeline outage → $4.4M loss due to lack of workflow resilience



The colonial hack, How did cyber-attackers shut off pipeline? ***Image source:*** *BBC*

BPP

e-learning recap



Reflecting on your learning…

The e-learning for this topic, covered the following areas:

- **Workflows** are structured sequences of tasks that move and transform data reliably.
- **Workflow management** ensures control, visibility, and resilience in data pipelines.
- Core principles include **clear task boundaries**, **dependency-driven execution**, and **modular design.**
- **Workflow orchestration tools** (e.g., Airflow, Prefect) automate, schedule, and monitor pipelines.
- **Optimisation techniques** like parallelism, caching, and resource scaling improve efficiency.
- **Automation** reduces manual effort using schedules, triggers, and reusable templates.
- **Monitoring and maintenance** ensure workflows remain reliable, observable, and scalable.

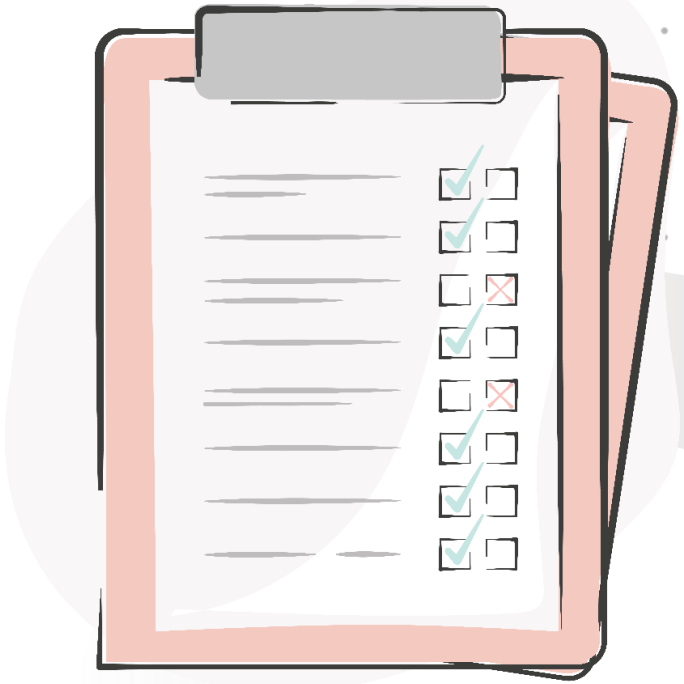
- Do you have any questions about any of these areas?

BPP

Webinar Agenda

Today, we will cover the following:

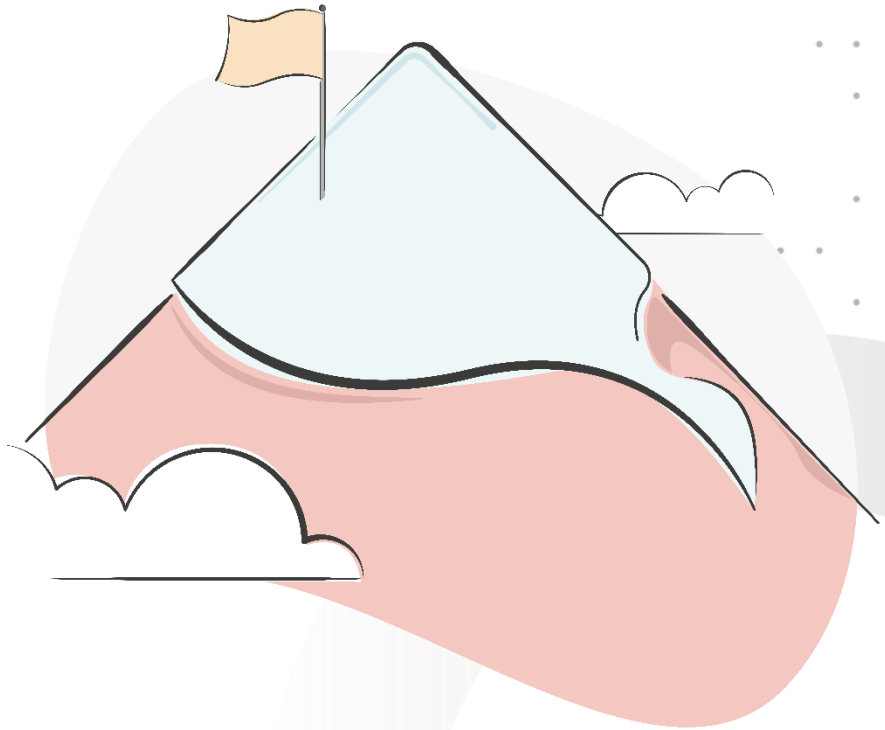
1. Recap of core concepts, including:
    - Workflow definition
    - Core principles of workflow management
    - From script to pipeline
    - Workflow patterns
    - Orchestration tools
    - Optimising workflows
    - Automating repetitive tasks
    - Monitoring and maintenance
2. Practical lab
3. Q&A and Wrap-Up



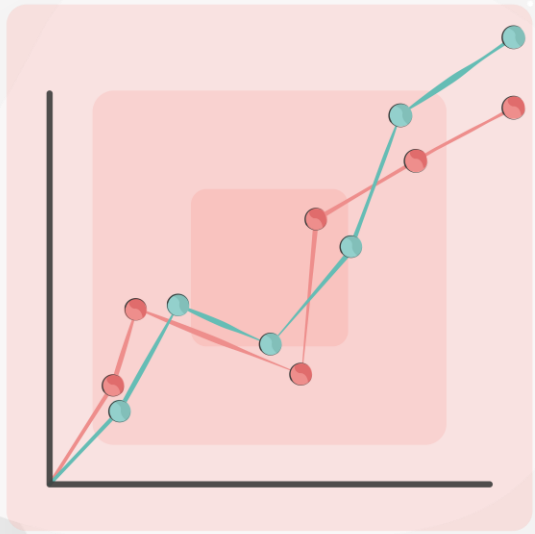
Session aim and objectives

By the end of this session, you should be able to:

- Design and automate robust workflows that power data pipelines from start to finish.
- Understand the role of orchestration tools in managing dependencies, schedules, and error handling across complex data systems.
- Apply optimisation and automation strategies to ensure workflows run efficiently and scale reliably over time.



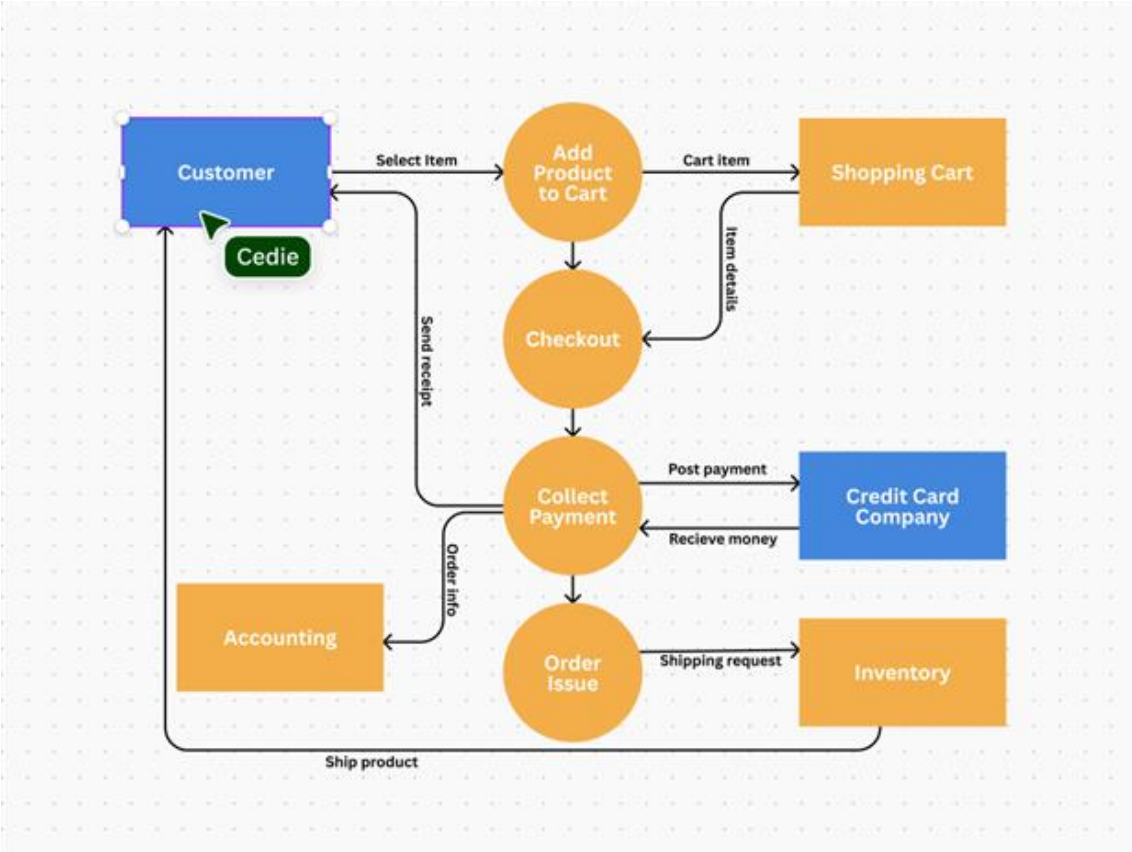
Recap of core concepts



What is a data workflow?

More than just a script

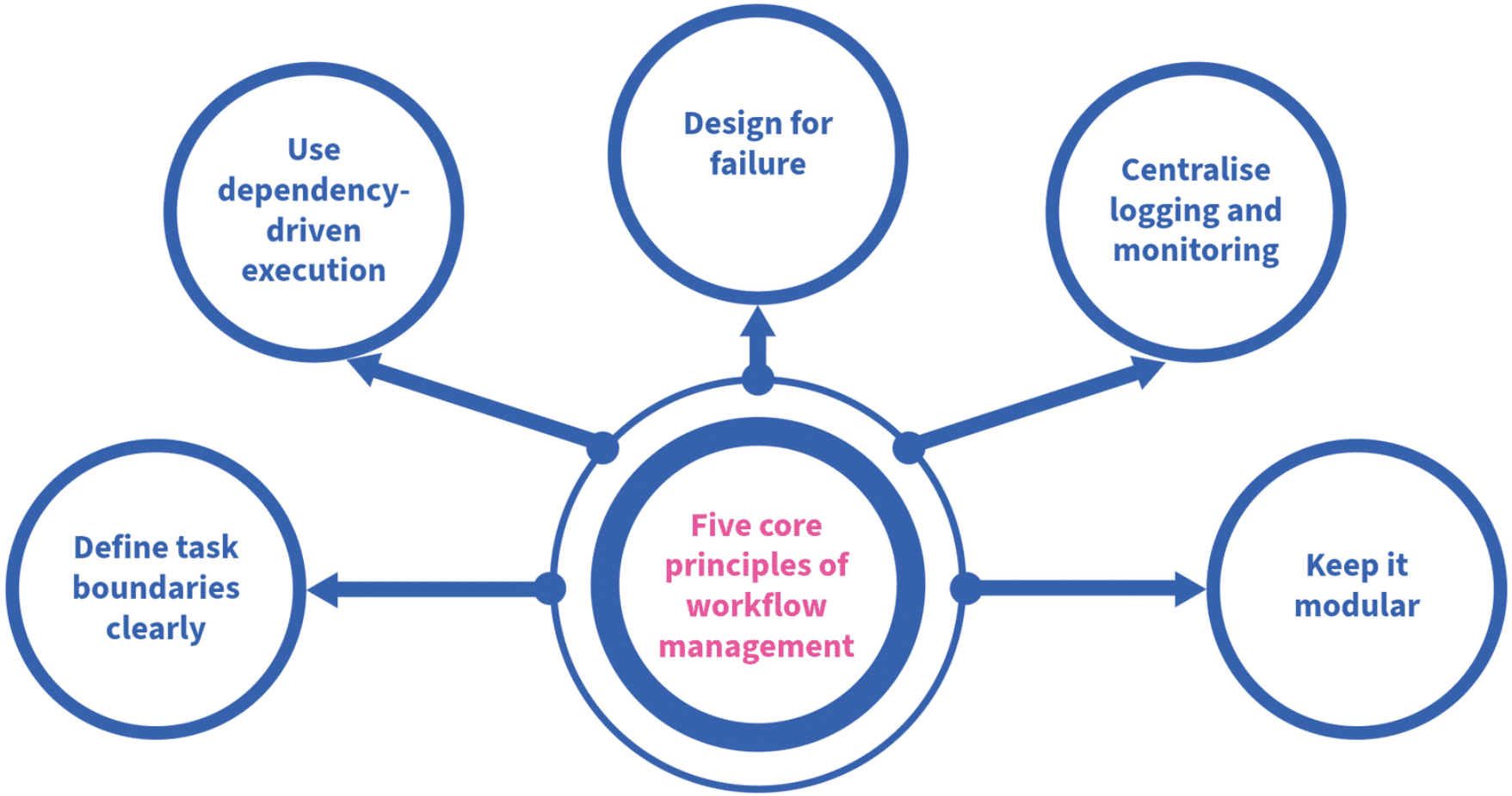
- Sequence of dependent tasks
- Moves and transforms data
- Requires control and visibility



A data flow diagram visualises the flow of data within a system or process, image source: Canva

Core principles of workflow management

Achieving good workflow management



The five core principles of workflow management

From script to pipeline

Coordinating tasks

- A workflow is more than just code—it's about task coordination.
- Each task has dependencies that define execution order.
- These dependencies form the structure of the workflow.
- Managing workflows means defining tasks and their relationships

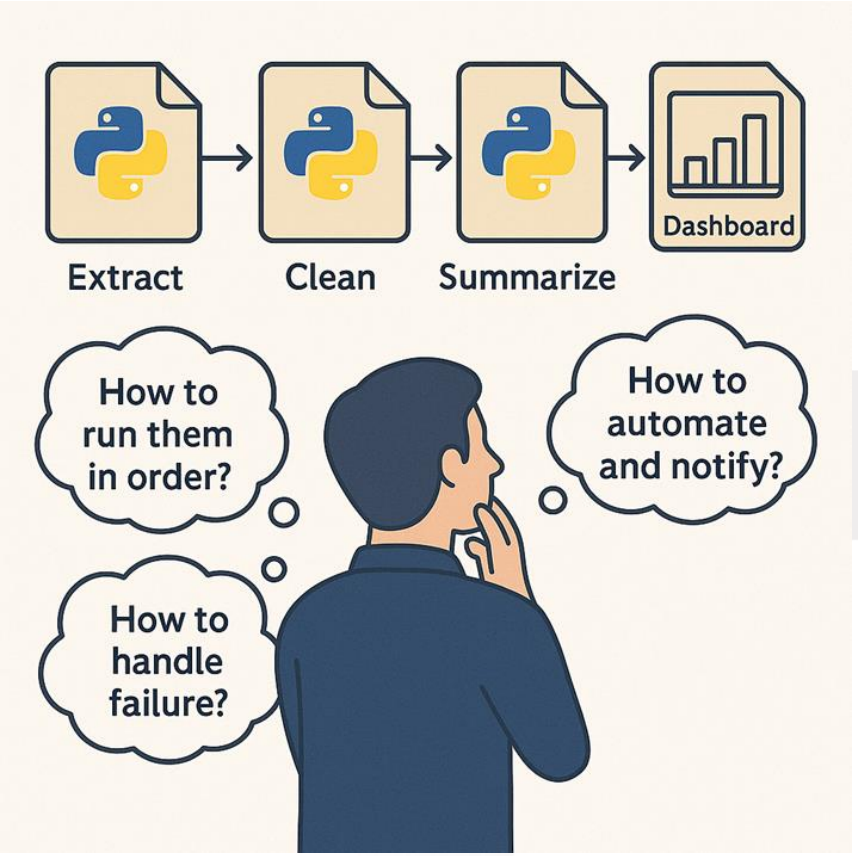


Figure: From script to pipeline

Workflow patterns

Recognise and reuse

- Linear Chains – Simple A → B → C sequences. Easy to follow, common in batch processing.
- Fan-out/Fan-in – A single task splits into parallel tasks (e.g., processing multiple datasets), which then merge back into a single task.
- Conditional Branching – The path changes based on task outputs (e.g., only clean data if the file exists).
- Sensor-Driven – Workflows that wait for an external event, like a new file or a message.

```
<!DOCTYPE html>
<html lang="en">
        <head>
                <meta charset="utf-8">
                <title>title</title>
                <link rel="stylesheet" href="style.css">
                <script src="script.js"></script>
        </head>
        <body>
                <!-- page content -->
        </body>
</html>
```

Computer code showcasing a simple Airflow code with four stages: extract, clean, summarise, and load.

BPP

Orchestration tools

Airflow vs Prefect

Feature	Apache Airflow	Prefect
Adoption	Widely adopted, industry standard	Gaining popularity, especially with developers
Workflow Definition	Uses DAGs (Directed Acyclic Graphs) in Python	Python-native syntax, more intuitive
Best For	Batch processing, ETL pipelines	Dynamic workflows, real-time triggers
Ease of Use	More boilerplate, steeper learning curve	Simpler, more developer-friendly
Monitoring	Built-in UI for tracking and retries	Cloud dashboard with rich observability
Deployment Options	Self-hosted, Kubernetes, cloud integrations	Cloud-native or self-hosted
Key Strength	Mature ecosystem, strong community support	Flexibility, modern developer experience

Apache Airflow

PREFECT BPP

Optimising workflows

Speed and efficiency

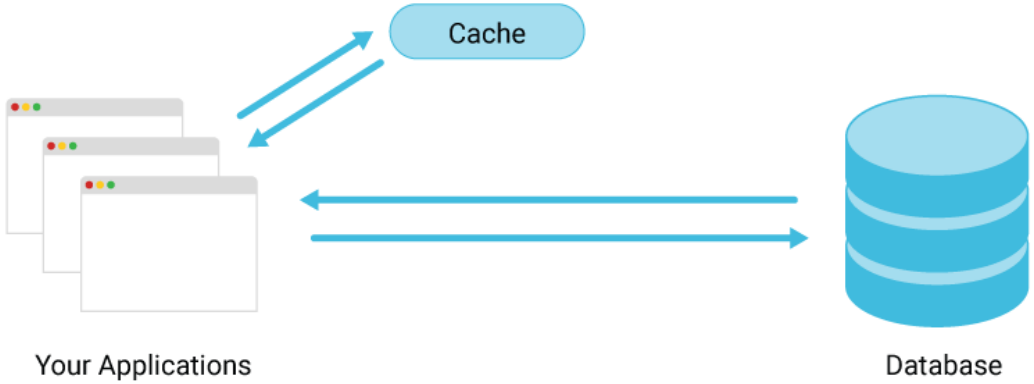
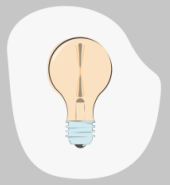


Figure: caching can help save time and resources
Source: ScyllaDB



Three techniques:

- Parallelism and concurrency
- Caching intermediate results
- Resource scaling

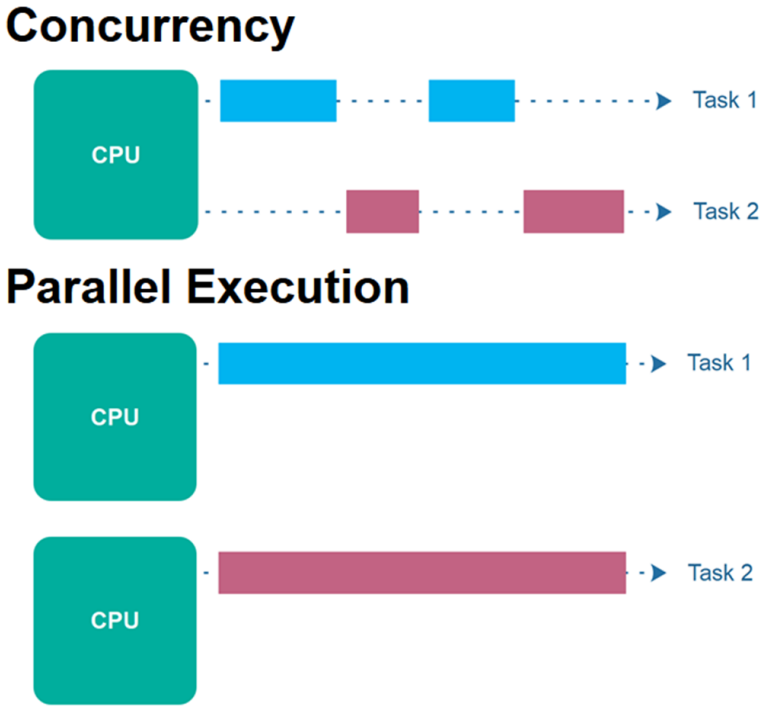
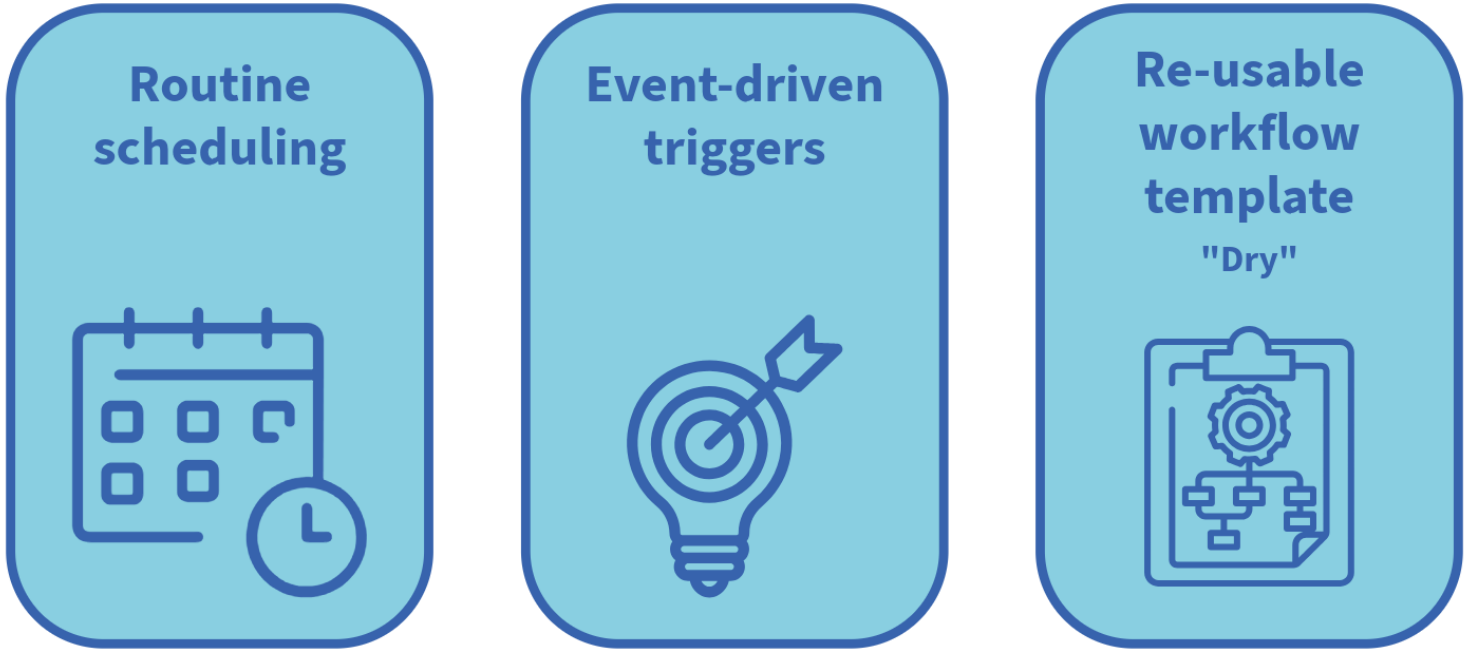


Figure: Concurrency vs Parallel Execution, Source: JENKOV

Automating repetitive tasks

Save time, reduce errors

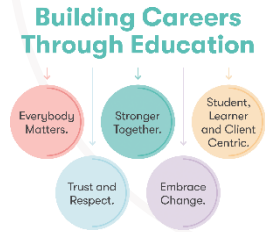


Task categories that are strong candidates for automation

Good practice:

- Version controlled (e.g., in Git)
- Documented with clear instructions
- Modular enough to be reused and extended

Monitoring and maintenance



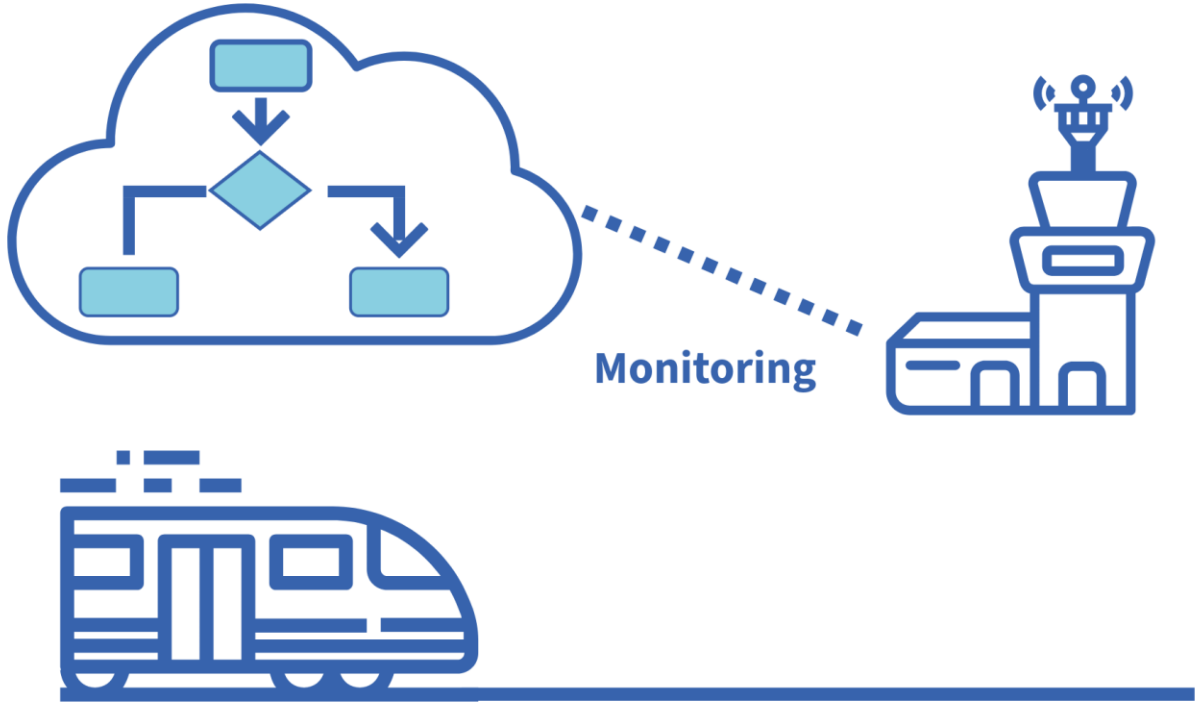
For workflow efficiency

At a minimum, good workflow monitoring tracks:

- Task Success/Failure Rates
- Execution Times
- Resource Usage
- Queue Backlogs.

Think of your data pipeline like a commuter train system.

Each task is a train, and the orchestrator is the control tower.



Monitoring data pipelines are like monitoring trains.

BPP

Maintaining workflow health

A proactive approach

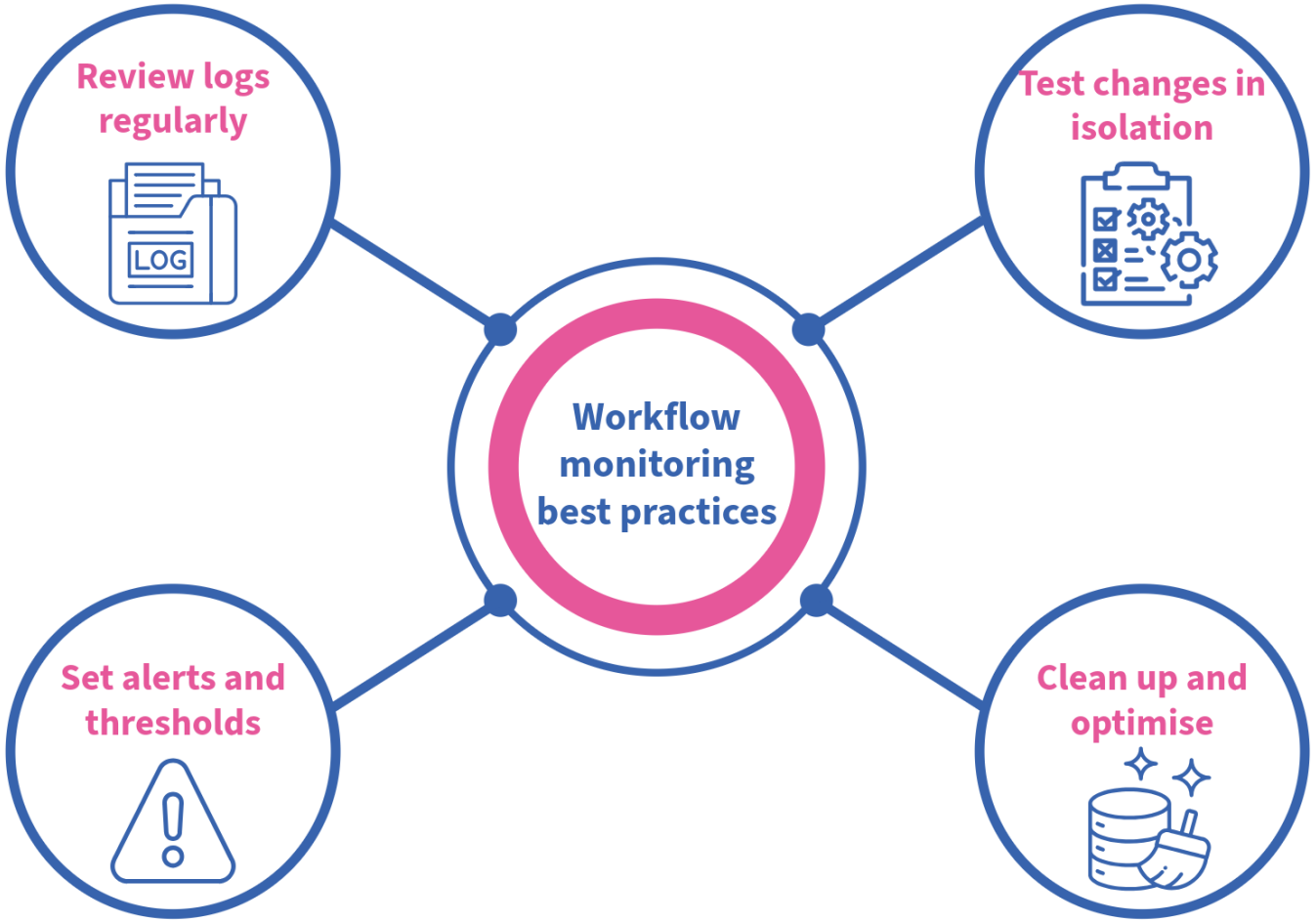
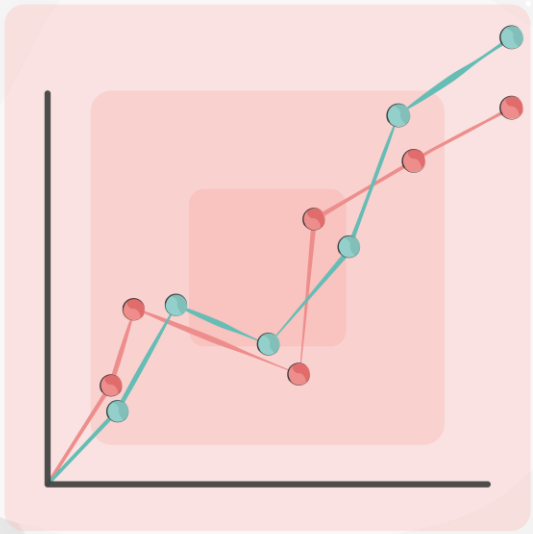


Figure: Best practices when it comes to monitoring workflows.

Practical lab



Exercise part 2

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Next phase of design based on FR data

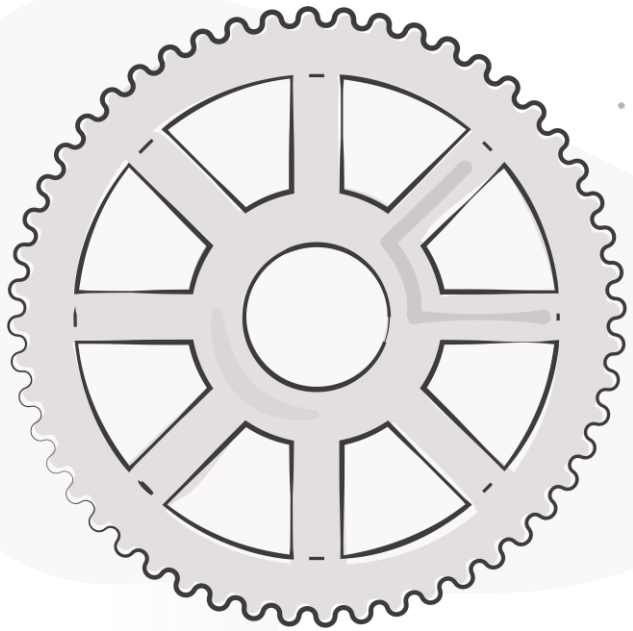
1.Update Database Schema: Modify the schema to accommodate data from the French version of the software application.

2.Re-implement Data Pipeline: Develop a pipeline to clean, validate, and load both UK and French data into the revised schema.

3.Use Test Data: Utilise provided test data for 10 French users and their login timestamps.

4.Future Phases: Prepare to integrate data from additional countries with varying formats.

5.Documentation: Establish comprehensive documentation for the schema and pipeline for future implementation by other teams.



Practical challenge

Files Provided in the Hub:

- FR User Data.csv: Contains 10 sample records.
- FR-User-LoginTS.csv: Contains login timestamps for January 2025.

.

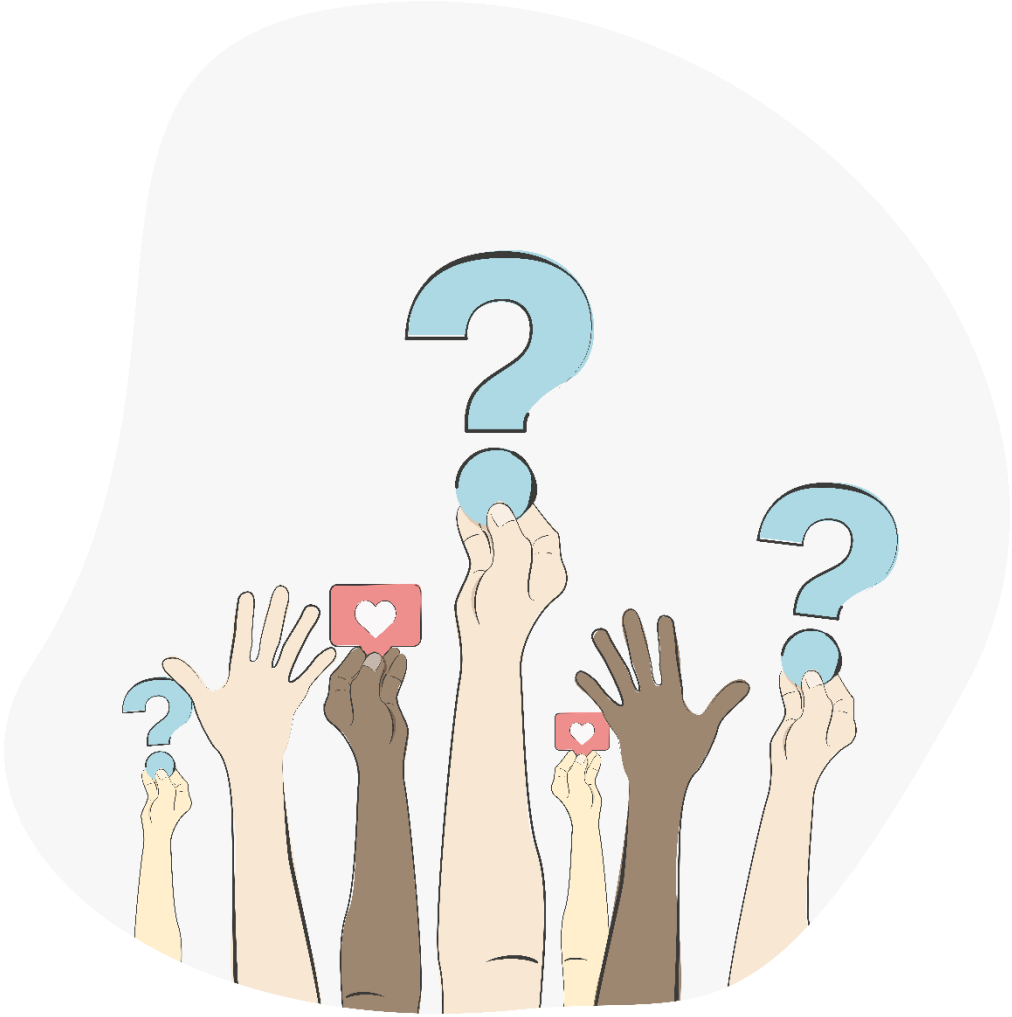
BPP

Key Learning Summary

- **Workflows** in data engineering are structured sequences of dependent tasks that manage how data is moved and transformed.
- Effective **workflow design** involves clear task boundaries, defined dependencies, and scheduled execution using tools or code.
- **Workflow orchestration tools** like Apache Airflow and Prefect allow engineers to schedule, monitor, and automate complex pipelines.
- **Optimisation techniques** such as parallelism, caching, and resource tuning improve the speed and efficiency of data workflows.
- **Automation** reduces manual repetition by using time-based or event-driven triggers and reusable workflow templates.
- **Monitoring and maintenance** involve tracking workflow performance, setting alerts, and regularly reviewing logs to ensure reliability.

BPP

Any questions or feedback?



Building Careers Through Education

Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.

BPP

Thank you