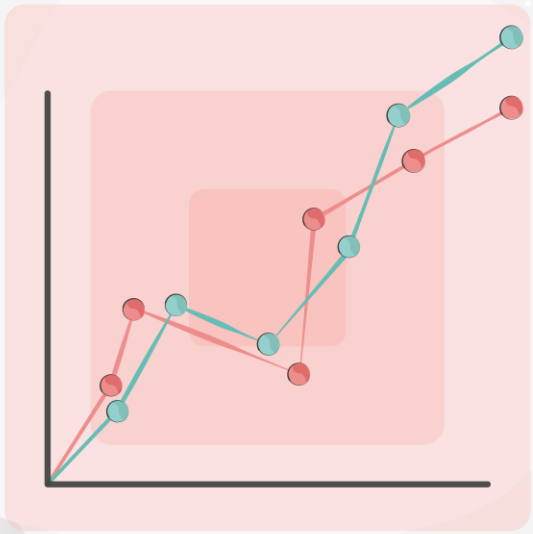
Data engineering

Module: Data pipelines

Topic: Data cleansing and enrichment in data engineering workflows



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 on data cleansing and enrichment?

Submit your responses to the chat or turn on your microphone

BPP

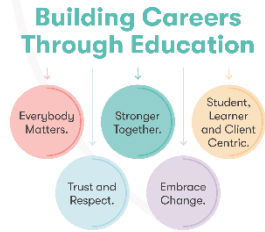
Airline loyalty program failure

How dirty data grounded a global campaign

- 1.2M loyalty records across 4 systems
- 18% had invalid or missing email addresses
- 12% had duplicate frequent flyer IDs
- Campaign ROI dropped by 40% due to delivery failures

Grounded planes: Image source: reuters.com

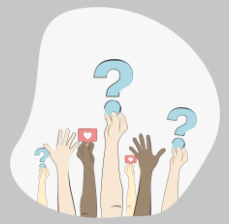
e-learning recap



Reflecting on your learning…

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

- Identifying and Correcting Data Quality Issues
- Standardisation, Deduplication, and Validation
- Using Tools for Data Cleansing
- Techniques for Data Enrichment
- Practical Exercises and Pipeline Integration



- Do you have any questions about any of these areas?
- Did everything in the e-learning make sense?

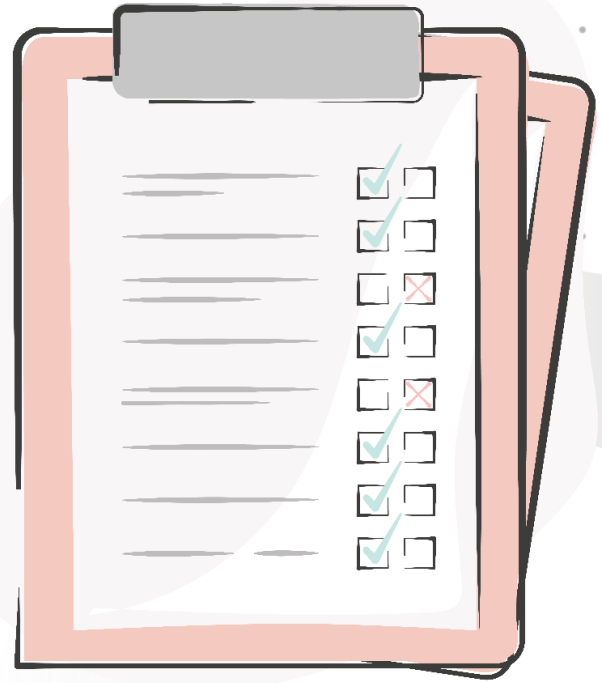
Q&A discussion

BPP

Webinar Agenda

Today, we will cover the following:

1. Identifying Data Quality Issues
2. Standardisation, Deduplication & Validation
3. Tools for Cleansing
4. Data Enrichment Techniques
5. Integrating External Data
6. Hackathon
7. Summary
8. Q&A



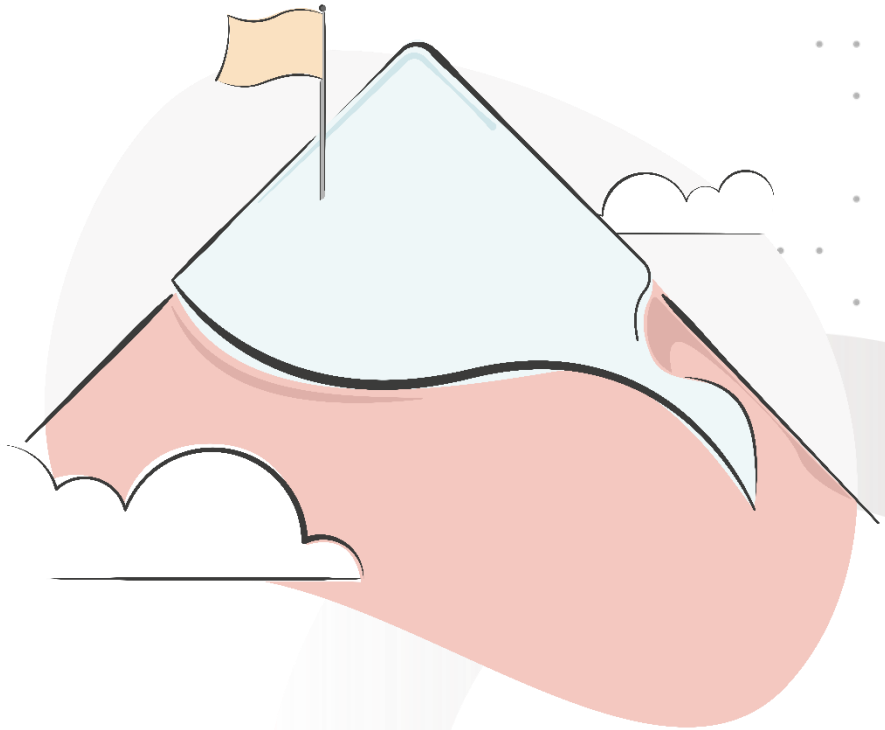
Session aim and objectives

Building Careers Through Education

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

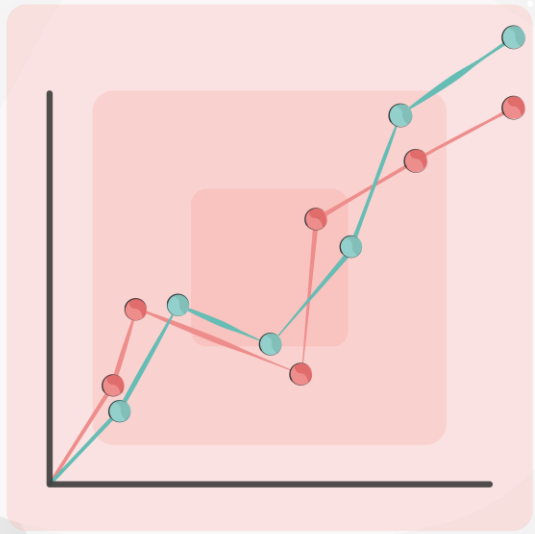
By the end of this session, you should be able to:

- Identify common data quality issues and explain their impact on data pipelines.
- Apply standard data cleansing techniques including standardisation, deduplication, and validation.
- Describe how to enrich datasets using additional internal or external information.



BPP

Core concept recap



Knowledge check poll

Which of the following best describes the purpose of data enrichment?

A. Removing duplicate records from a dataset

B. Converting inconsistent formats into a standard structure

C. Adding new, meaningful context to existing data

D. Validating data types and ranges

Submit your responses to the chat!

Feedback: C – Adding new, meaningful context to existing data

Introduction

Why clean data is the foundation of reliable analytics

- Dirty data leads to broken insights
- Cleansing and enrichment are essential steps
- Today’s goal: make your data trustworthy

customers.csv

C	Customer ID	Name	Country	Email	Phone
1	1001	John Doe	United States	john.doe@	(555) 123-4567
2	1001	John Doe	United States	john.doe@	(555) 123-4567
3	Jane Smith	Jane Smith	Fmcil	john.doeexm	555-987-6543
4	Alice Johnson	Frnce	E	jane.smith	+33 1 2345678
5	Bob Brown	UK	—	alice.johnson	—
6	Charlie Williams	Minc	—	—	0 200 7946 012
7		United	Exited	bob.brown@	(020)7946 0123
8	United Kingdom	Uhıted	Kingdom	charlie.williamseem:ail.com	

Figure: *An example of a bad CSV file*

What is “Bad Data”?

Building Careers Through Education

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

Understanding the enemy…

- Incomplete, inconsistent, or inaccurate values
- Duplicates and formatting issues
- Subtle errors can break entire pipelines

Figure: Good data vs. Bad Data

BPP

Common data quality issues

Know what to look for…

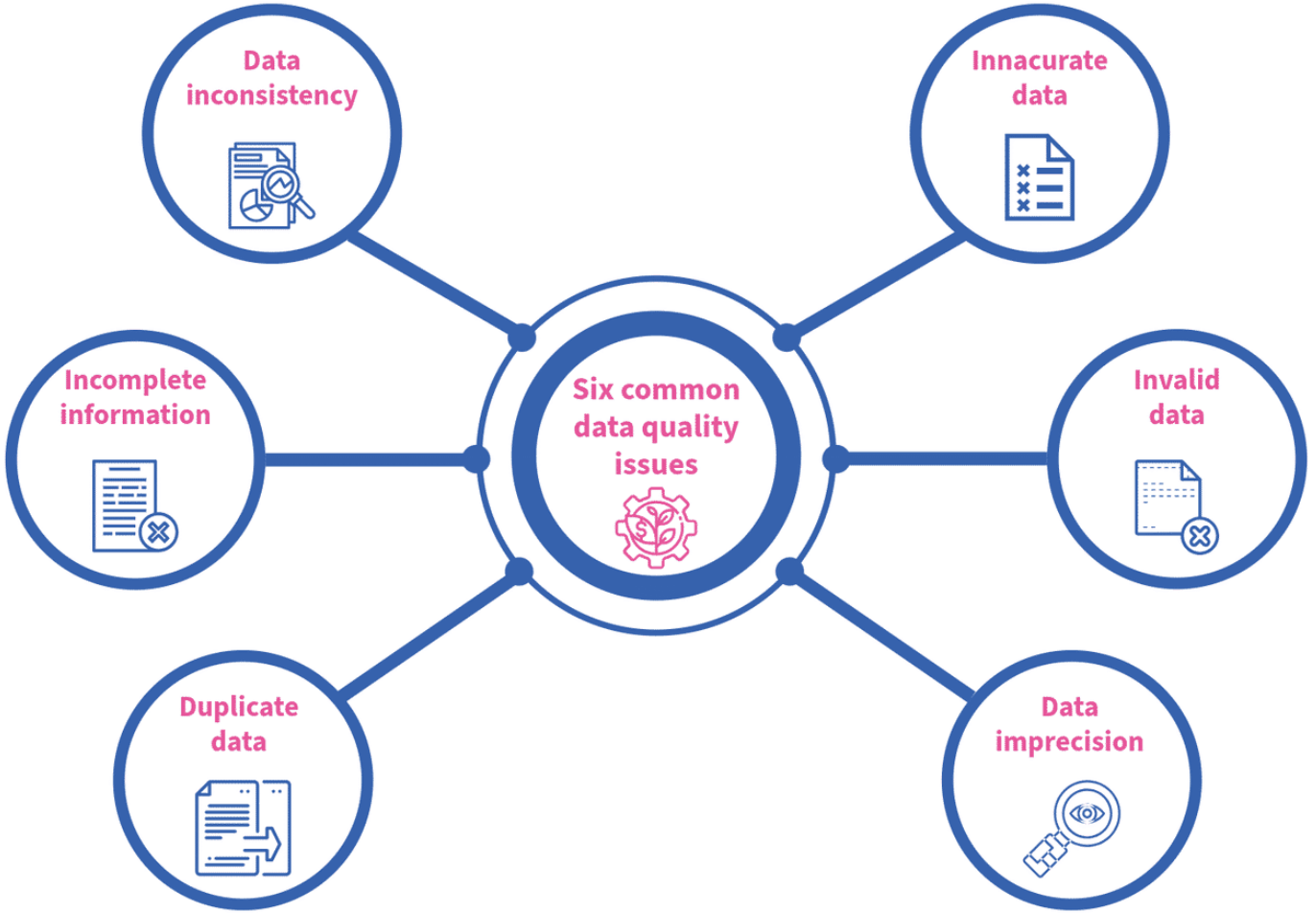


Figure: The main types of data quality issues

Standardisation techniques

Bringing consistency to chaos…

- Align formats (dates, units, text case)
- Normalise categories (e.g., “UK” vs. “United Kingdom”)
- Use automation where possible

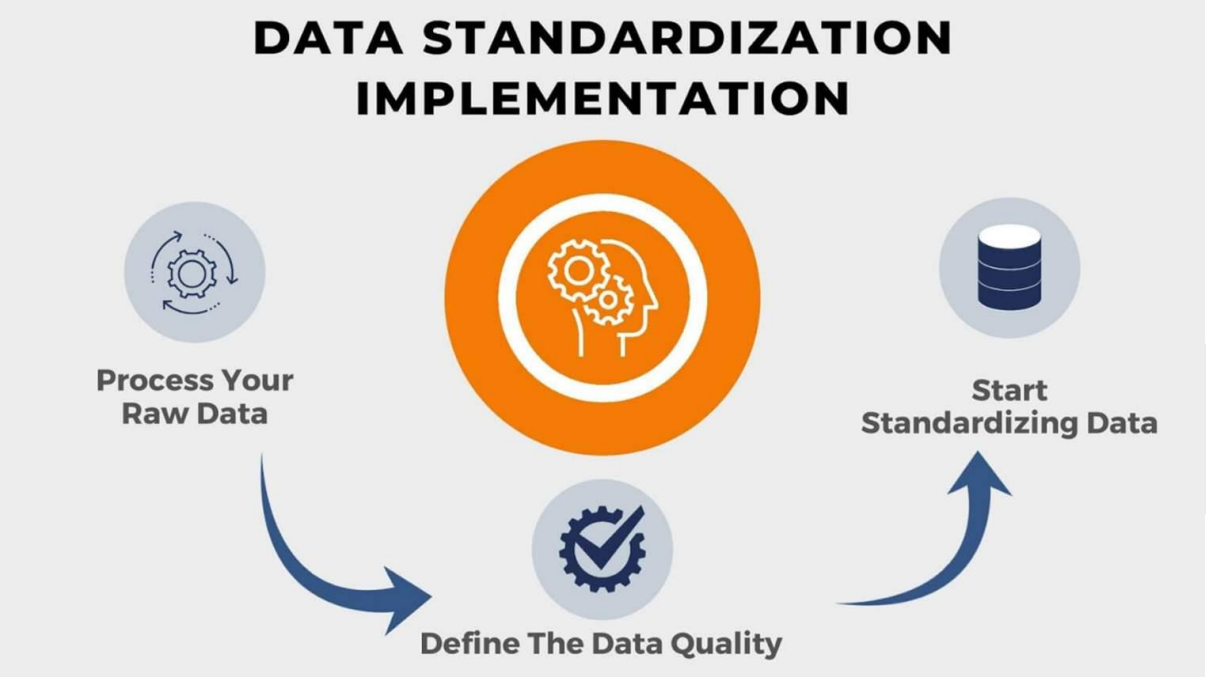


Figure: Data Standardisation Implementation

Deduplication strategies

One record, one truth…

- Exact vs. fuzzy matching
- Use unique identifiers
- Merge or flag near-duplicates

Example
Two records for “Chris Taylor” have slightly different addresses:

- chris.taylor@company.com, “123 Main St”
- chris.taylor@company.com, “123 Main Street”

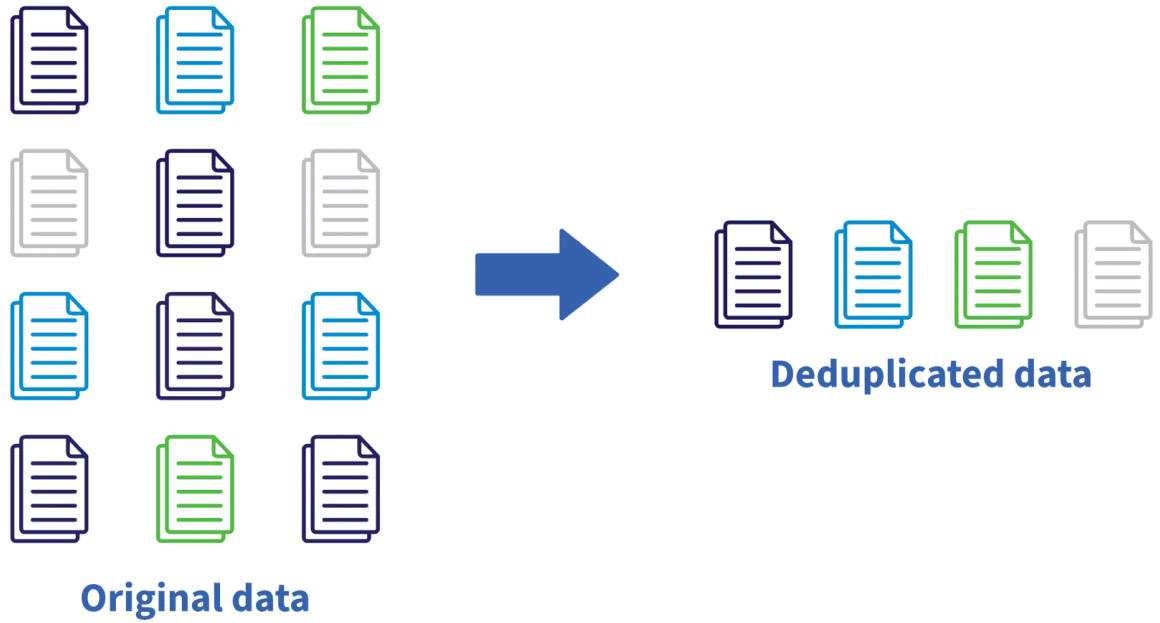
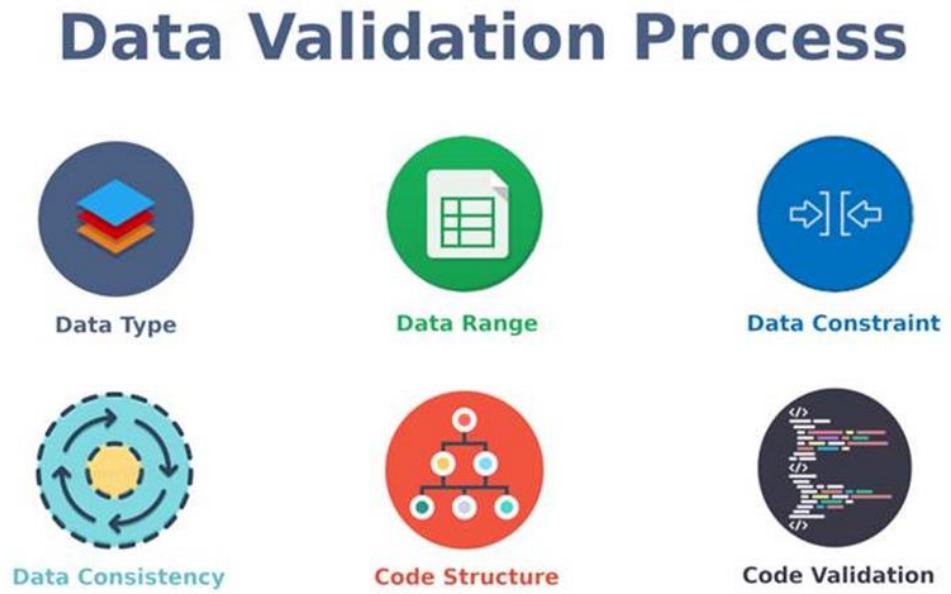


Figure: Data deduplication

Validation techniques

Catching errors before they spread…

- Data type, range, and constraint checks
- Consistency across fields
- System-level validation



The six types of data validation,
Source: LinkedIn

Tools for cleansing

Code or no-code – choose your weapon

- **OpenRefine:** faceting, clustering, transformations
- Trifacta: intelligent suggestions, visual profiling

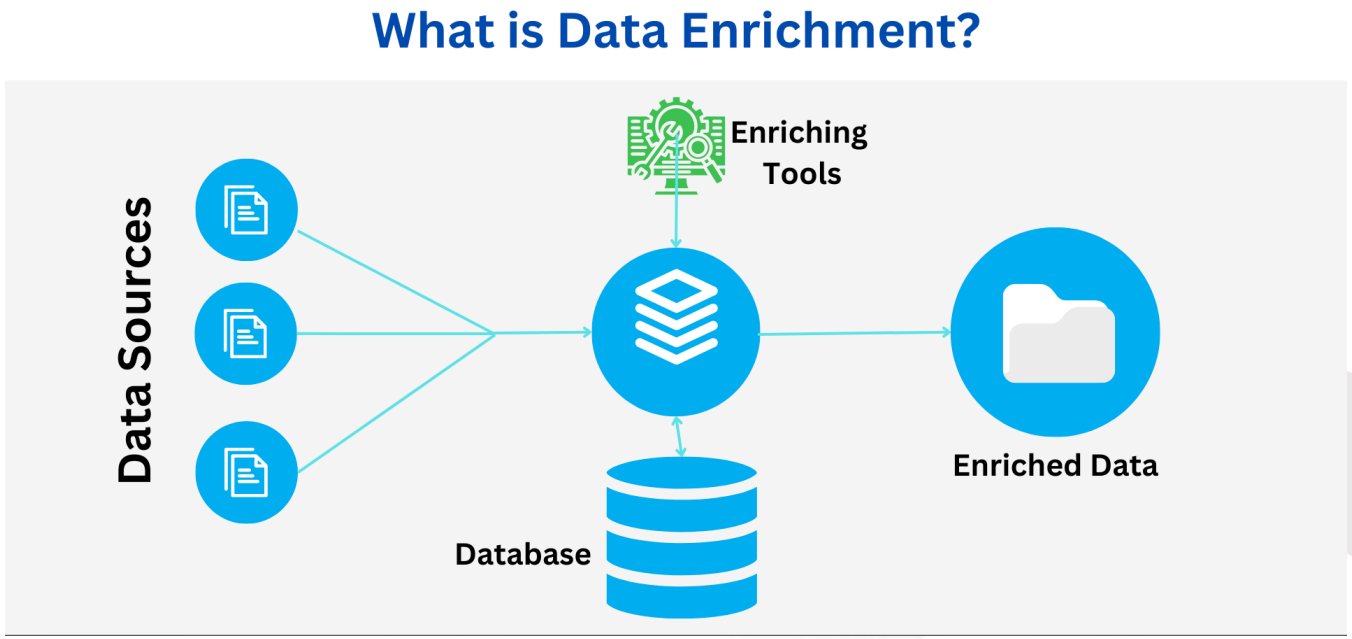
Figure: OpenRefine tool

Figure: Trifacta tool

Validation techniques

Adding depth and context…

- Internal joins (e.g., loyalty status)
- External APIs (e.g., weather, demographics)
- Derived fields (e.g., age, spend per night)



Transforming raw data into valuable insights through data enrichment, image source: credencys.com

Integrating external data

APIs, files, and streaming feeds…

- API integration (real-time, flexible)
- Static files (scheduled updates)
- Streaming data (real-time insights)

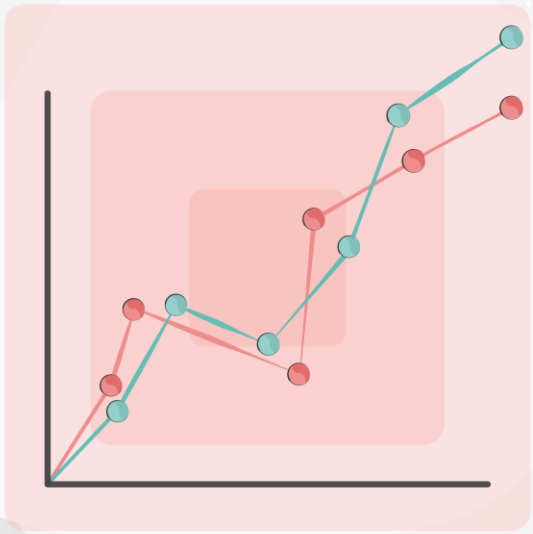
API integration

Static file ingestion

Streaming data feeds

BPP

Practical lab



Hackathon

Part A and B…

Your step-by-step tasks are as follows:

Part A – Pipeline Design

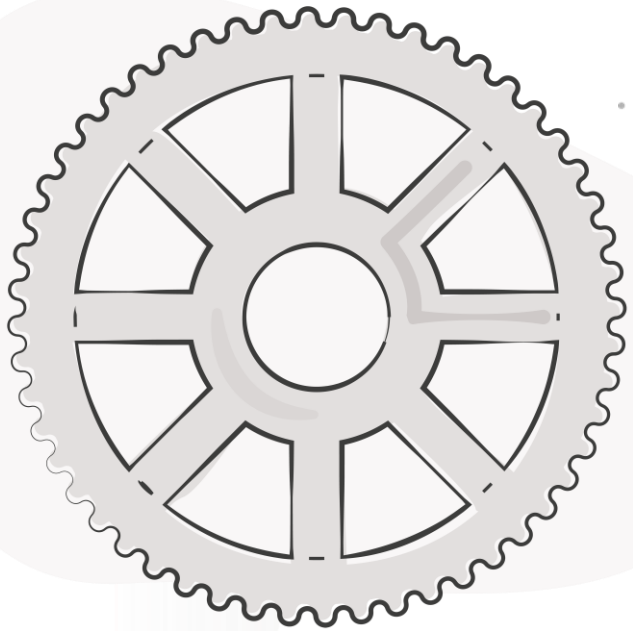
- Design, implement, and test a data pipeline
- Merge sunrise/sunset and weather data for Edinburgh
- Generate minute-level temperature estimates for 2012

Part B – Pipeline Reuse & Adaptation

- Execute a peer's pipeline on new data (London & Hertfordshire)
- Assess documentation, adapt for schema differences
- Validate output and provide feedback

Files Provided in the Hub:

- Hackathon brief



Practical challenge

Hackathon extension

Part C and D…

Part C – Big Data Model Building (Extension)

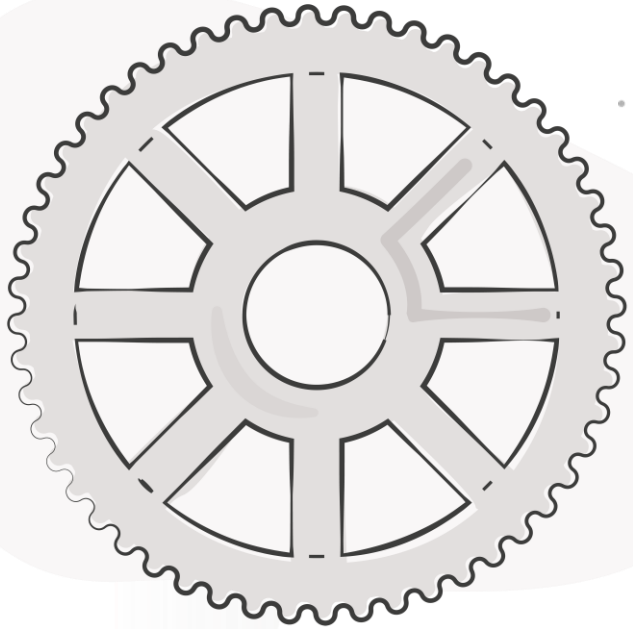
- Model relationship: temperature ↔ equipment performance
- Link model to customer complaints
- Simulate global warming impact on mast reliability
- Present findings to a non-technical audience
- Bonus: Identify peak usage times (weekday vs weekend)

Part D – Compare Big Data Sets (Extension)

- Apply & adapt model to 8 new masts (Hertfordshire)
- Adjust for local temperature differences
- Identify unusual mast behavior
- Evaluate significance of anomalies
- Present conclusions clearly to a lay audience

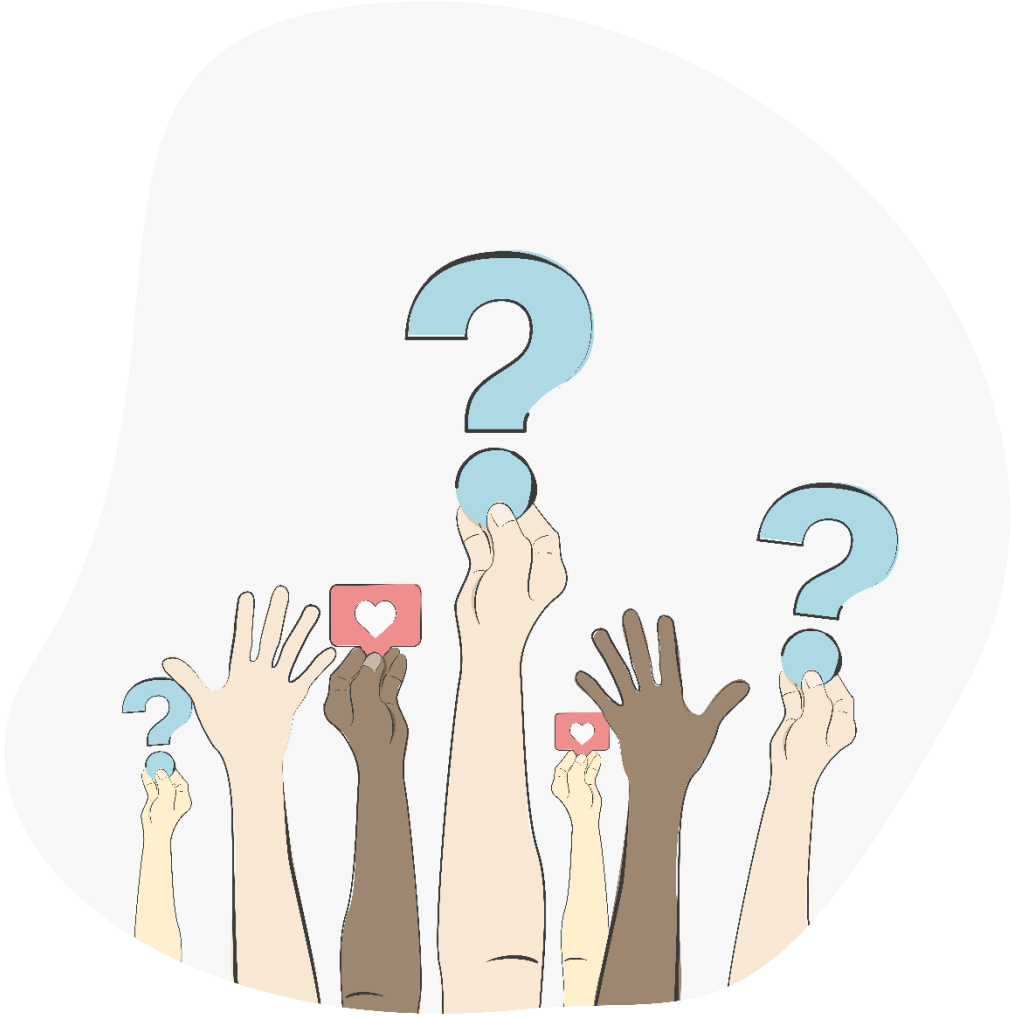
Files Provided in the Hub:

- Hackathon brief



Practical challenge

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