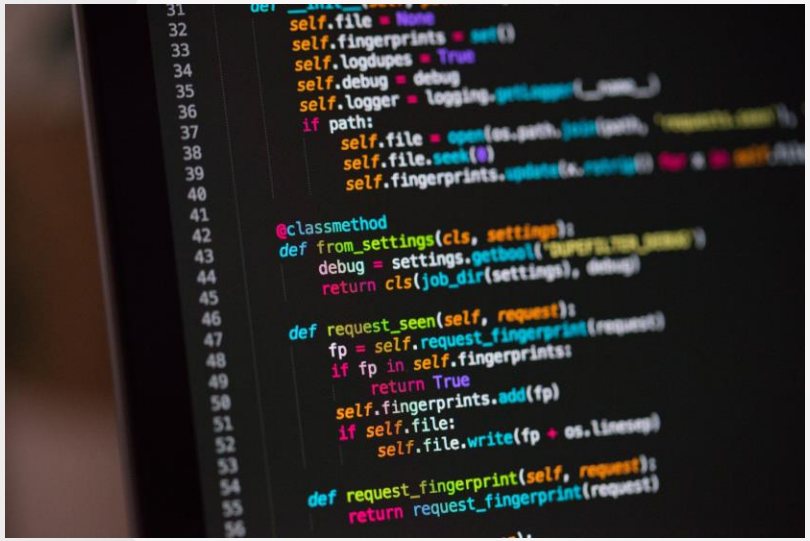
Level 5 Data Engineer
Module 6 Topic 1

Introduction to Data Collection and Ingestion



L5 Data Engineer Higher Apprenticeship
Module 6 / 12 (“**Data Collection and Ingestion pt. 1**”)
Topic 1 / 4

Ice breaker: Discussion

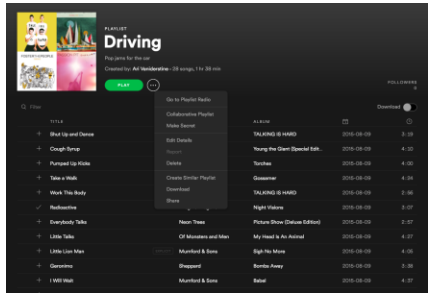
Building Careers Through Education

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

A bit of fun to start…

How do you collect data about yourself?

- Do you track your daily steps using a fitness app?
- Do you track how long you spend on social media or streaming services?
- If you could collect data on any topic in the world, what would it be and why?



Submit your responses to the chat or turn on your microphone

BPP

Case study

Amazon’s supply chain optimisation

- **Objective:** Amazon aimed to enhance inventory management and improve delivery times through data-driven decision-making.
- **Challenges:** Managing a complex supply chain with diverse products and fluctuating demand.
- **Data Collection:** Gathering data from sales, inventory, customer interactions, and logistics.
- **Implementation:** Using predictive analytics to forecast demand and optimize inventory levels.
- **Results:** Achieved significant cost savings, faster delivery times, and improved customer satisfaction.

Knowledge check poll

Your company wants to integrate data from various sources into a central data warehouse.

How would you approach the data ingestion process to ensure smooth integration and high data quality?

A. Ingest data without any preprocessing.

B. ETL (Extract, Transform, Load) processes to clean and transform data before loading it into the warehouse.

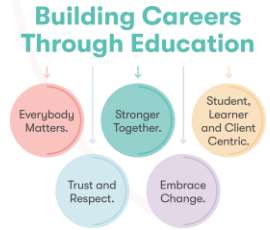
C. Load all data directly into the warehouse and clean it later.

D. Only ingest data from structured sources.

Submit your responses to the chat!

Feedback: B – ETL processes help in cleaning and transforming data, ensuring that only high-quality data is loaded into the data warehouse, which facilitates better analysis and decision-making.

Session aim and objectives

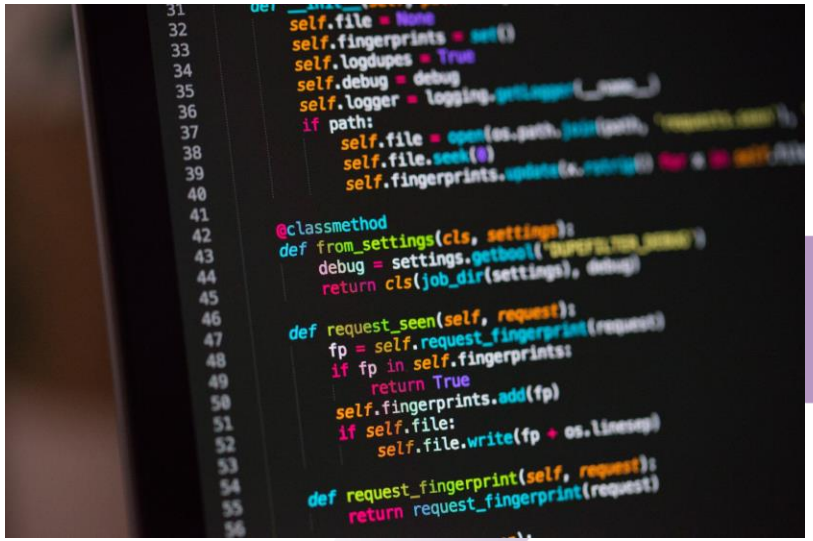


Completion of this topic supports the following outcomes:

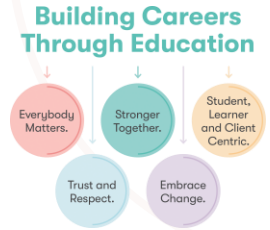
- Justify the importance of automation in data collection and ingestion
- Evaluate common data cleaning techniques
- Recognise the steps required to pre-process data for machine learning purposes
- Demonstrate practical data collection and ingestion skills

BPP

E-learning Recap



Recap discussion

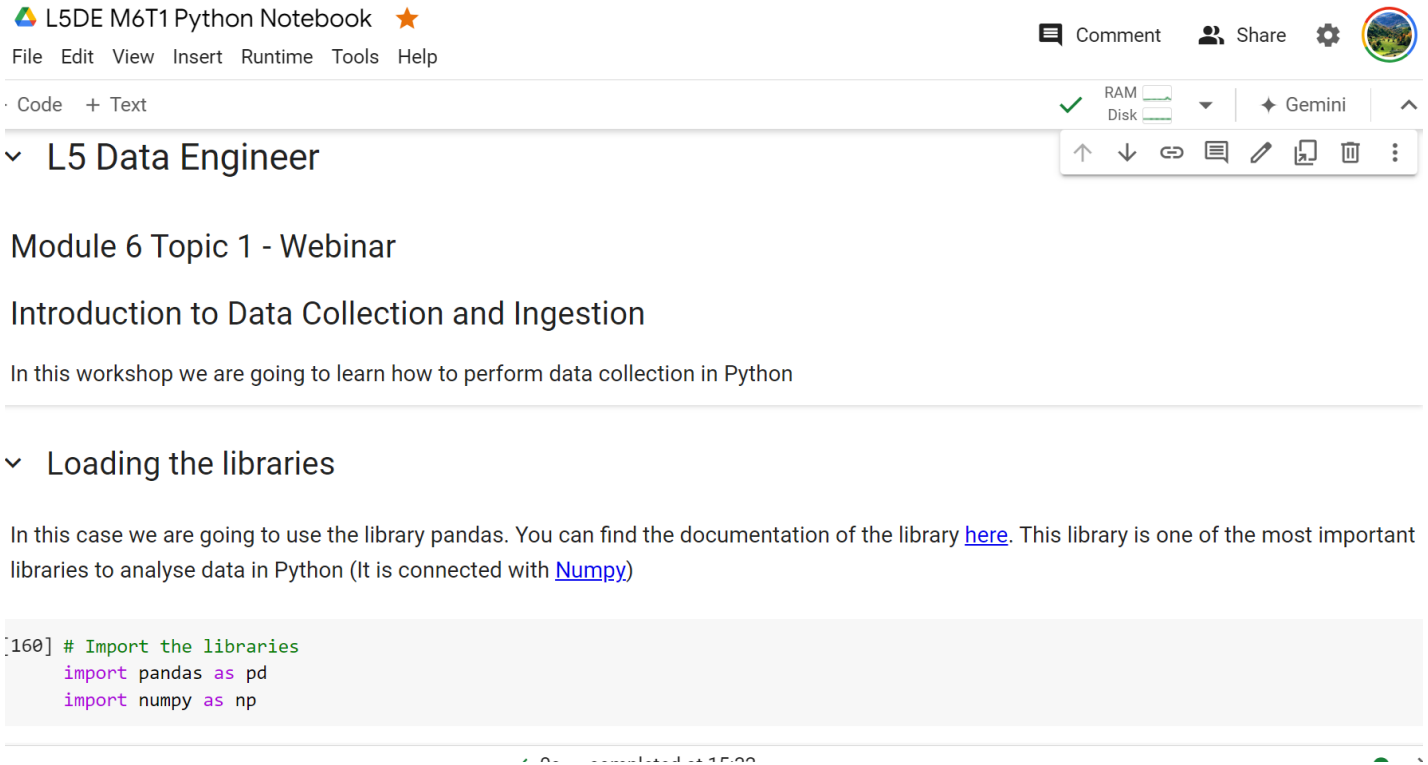


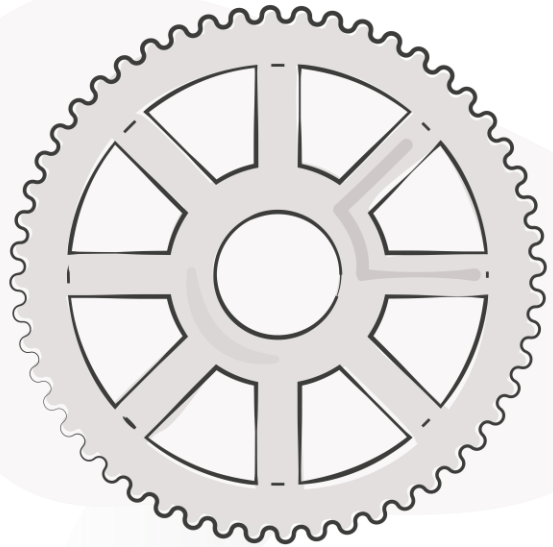
- Can you remember the difference between data collection and ingestion?
- Can you recall key approaches to collecting data.
- Can you remember the main approaches for data ingestion?
- What are the common types of dirty data?

BPP

Practical application

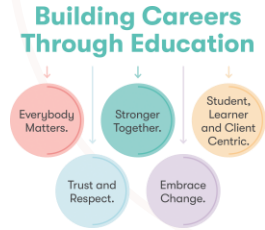
Your tutor will now walk you through the Collab notebook.





BPP

Things we learned from the code walkthrough

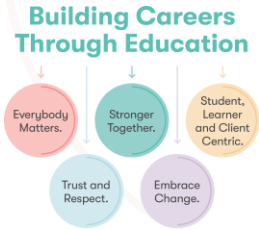


We will now recap the following functions we learned from the Python code walkthrough:

- DataFrame.T
- DataFrame.reset_index()
- DataFrame.melt()
- DataFrame.rename()
- DataFrame.to_numeric()
- DataFrame.unique()
- DataFrame.apply()
- Lambda functions

BPP

Lambda functions



Anonymous functions – function without a name

We use lambda functions when we require a nameless function for a short period of time.

Python Lambda syntax:

lambda arguments : expression

Example:

```
lambda_cube = lambda y: y*y*y
```

BPP

Using pandas documentation

pandas documentation — pandas 2.2.2 documentation

API reference

This page gives an overview of all public pandas objects, functions and methods. All classes and functions exposed in pandas.* namespace are public.

Most important webpage:

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

User Guide

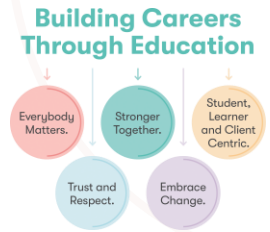
User Guide — pandas 2.2.2 documentation (pydata.org)

The User Guide covers all of pandas by topic area. Each of the subsections introduces a topic (such as “working with missing data”), and discusses how pandas approaches the problem, with many examples throughout.

BPP

Recap: Data sourcing

Practical considerations when collecting data



Identify Stakeholder Requirements:

- Engage with stakeholders from different parts of the business to understand their data needs.
- Define specific business objectives that the data should help achieve.

Data Identification and Assessment:

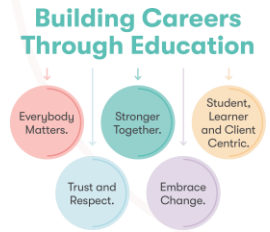
- Determine the type of data required to meet the objectives.
- Assess the volume, variety, velocity, and veracity of the data to understand the complexity of data handling.

System Selection (SQL/NoSQL/Data Lake/Warehouse/Cluster):

- **SQL Databases**: Best for structured data with complex queries and transactions.
- **NoSQL Databases**: Suitable for unstructured data, providing flexibility in terms of schema and scaling.
- **Data Lakes**: Ideal for storing massive amounts of raw data in its native format. Useful for big data applications.
- **Data Warehouses**: Optimal for analytics and business intelligence; structured for SQL queries.
- **Data Clusters**: Useful for processing large datasets across distributed computing environments.

BPP

More practical considerations



Data Acquisition Strategy:

- Plan how to acquire data (e.g., APIs, web scraping, IoT devices, third-party datasets).
- Establish data quality checks and initial processing steps.

Data Management Plan:

- Define the data governance framework (access controls, audit trails, compliance).
- Plan data maintenance, updates, and lifecycle management.
- Set up data backup and disaster recovery protocols.

Implementation and Integration:

- Integrate the chosen data systems with existing IT infrastructure.
- Develop or customise software for data ingestion, processing, and analytics.

Monitoring and Evaluation:

- Continuously monitor data quality and system performance.
- Evaluate the system’s ability to meet business needs and make adjustments as necessary.

BPP

Most Common Folder Structures

For data file imports…

- **Landing Area**: Raw data files are first ingested here; data is unprocessed and in its original format.
- **Staging Area**: Data is moved here from the landing area for initial processing and cleaning.
- **Ingestion Layer**: This is where data is transformed and loaded into more structured formats suitable for analysis.
- **Archived Data Folders**: After data has been analysed and is no longer immediately needed, it is moved to this area for long-term storage.

BPP

Case study: E-commerce

Objective: Enhance customer experience through personalised marketing and improve inventory management.

Stakeholder Requirements:

- Marketing wants customer demographic and behaviour data.
- Operations need inventory and sales data to optimise stock levels.

Data Assessment:

- Customer data includes structured data (age, location) and unstructured data (social media activity).
- Inventory data is highly structured with clear attributes (SKU, quantity, location).

System Selection:

- **Customer Data:** NoSQL database for flexibility with unstructured data and scalability.
- **Inventory Data:** SQL database for complex queries and transactional integrity.
- **Analytics:** Data warehouse to integrate various data for BI tools.

BPP

Case study continued

Data Ingestion:

Customer data through online forms, social media APIs, and loyalty programs.

Inventory data from point-of-sale systems and supplier databases.

Data Management:

Data governance policies in place to protect customer privacy.

Regular updates to inventory data with real-time processing for dynamic decision-making.

Implementation:

Deploy MongoDB for NoSQL needs; use PostgreSQL for SQL requirements.

Integrate with a cloud-based data warehouse (e.g., Amazon Redshift).

Monitoring:

Use dashboards to track key metrics like customer engagement and inventory turnover.

Regularly review system performance and stakeholder satisfaction.

BPP

Licensing and tracking data usage

- **Understand Data Licensing**: Ensure clarity on the terms of use, restrictions, and obligations associated with the data. This includes understanding whether the data can be shared, modified, or needs to be kept secure.

- **Document and Track Data Usage**: Implement a system to track how data is being used within your organisation. This helps in auditing and ensuring compliance with licensing agreements.

- **Negotiating Licenses**: Work with legal teams to negotiate data licenses that align with your business needs and risk management strategies.

- **Resolve Licensing Issues**: Establish a clear process for resolving disputes over data usage, which may include mediation or legal actions if necessary.

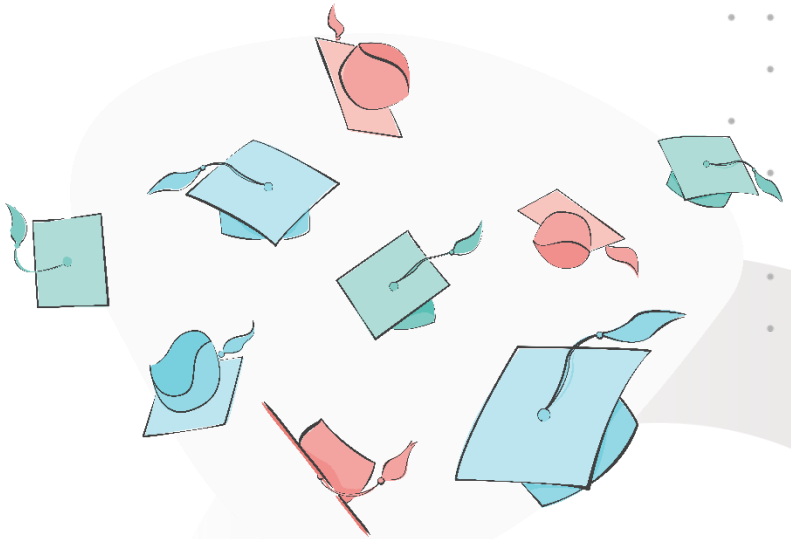
- **Training and Awareness**: Regular training sessions for employees on the importance of compliance with data licensing terms to avoid legal issues.

BPP

Key Learning Summary

The key takeaways from this session are as follows:

- **Automation in Data Collection and Ingestion:** Automation reduces errors and increases efficiency in data handling.
- **Data Cleaning Techniques:** Common techniques include removing duplicates, handling missing values, and standardising data formats.
- **Pre-Processing for Machine Learning:** Pre-processing steps include normalisation, encoding categorical variables, and feature scaling.
- **Practical Skills:** Hands-on experience with tools like Python and pandas for data manipulation and analysis is crucial.



BPP

Thank you

Do you have any questions, comments, or feedback?