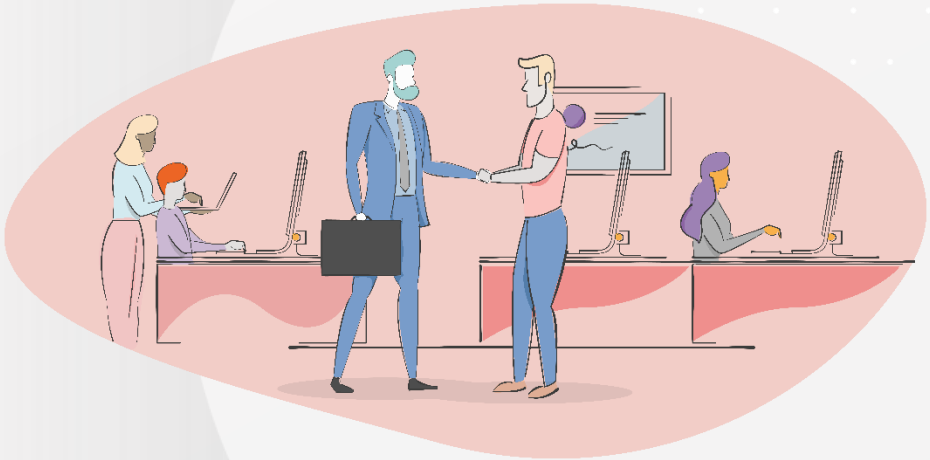
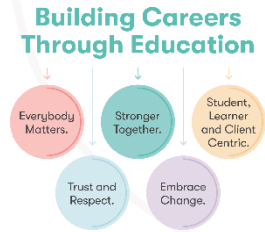
Introduction to Data Quality



L5 Data Engineer Higher Apprenticeship
Module 1 / 12 (“**Data Fundamentals**”)
Topic 2 / 4

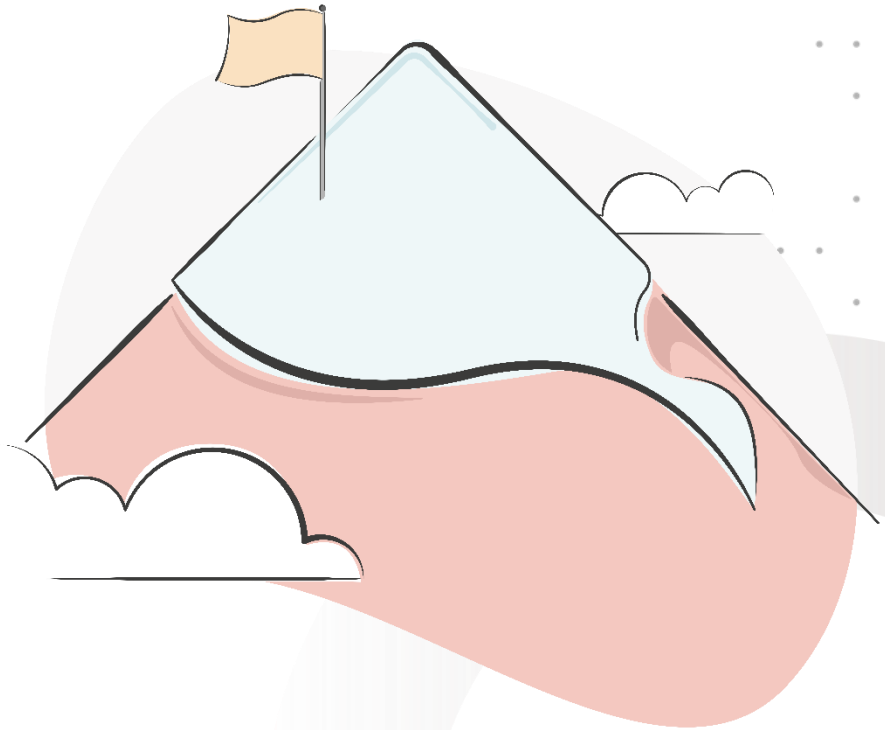
rev. 1 (2024)

Learning outcomes



This webinar is designed to support the following learning outcomes:

- Define data quality and its significance
- Demonstrate an understanding of the importance of data standards in data engineering
- Apply data quality management strategies proficiently to improve the accuracy and reliability of data
- Analyse sustainable data practices and governance principles to advocate for long-term data quality and integrity



BPP

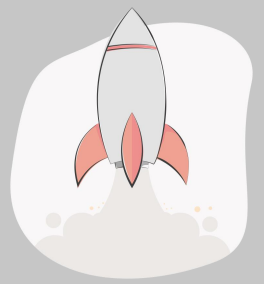
The value of data quality

Case study example: Novartis

The problem - Novartis faced data management and integration challenges across various sources

The solution:

- Adopted FAIR Data Principles and implemented data governance framework
- Implemented Novartis Data Catalog, a metadata management solution like OpenMetadata



The benefits:

- Improved data findability, accessibility, interoperability, and reusability
- Enhanced data quality and consistency
- Increased collaboration and data sharing
- Efficient data discovery and access
- Better regulatory compliance

Novartis Pharmaceutical Corporation

Image source: *Reuters*

Data quality challenges

Building Careers Through Education

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

Discussing your previous experience…

Let's discuss your prior knowledge and experiences with data quality challenges.

Share any instances where you or your organisation faced issues related to data accuracy, completeness, consistency, or timeliness.

Consider the following questions:

- How did these data quality challenges impact your work or decision-making processes?
- What measures, if any, were taken to address these challenges?

Group discussion

BPP

Knowledge Check Poll

Let's see what you can remember from the e-learning!

Which of the following is NOT one of the four main data quality metrics discussed in the e-learning?

A) Accuracy

B) Integrity

C) Consistency

D) Relevance

Submit your responses to the chat!

Feedback: D – Relevance is not one of the four main data quality metrics discussed in the e-learning.

The four main metrics covered were accuracy, integrity, consistency, and timeliness (also known as recency or availability).

Data quality metrics

An introduction

Data quality encompasses various dimensions (or metrics).

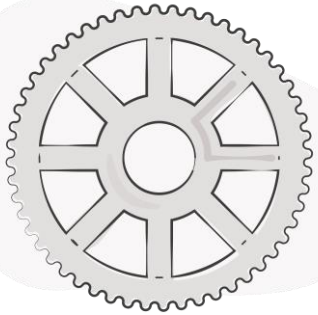
These dimensions or metrics are broken down below:

Data quality metric	Alternative names	Answers the questions
Accuracy	Correctness, validity	How well does our data reflect the real world? Are false items likely?
Integrity	Completeness	How well does our data cover the real world? Are missing items likely?
Consistency	Uniformity	How well does our data conform to a single standard? Are discrepancies likely?
Timeliness	Recency, availability	How up-to-date is our data? Are stale items likely?
Reliability	N/A	Is the data consistently accurate over time? Can the data be depended upon to make critical decisions?

Introducing open standards

Definition: *guidelines for technology that anyone can use and contribute to*

What they offer: *Compatibility, flexibility, innovation and cost-effectiveness*



Their importance: *Ensure tools and systems are compatible and adaptable for seamless integration and evolution*

Examples: *HL7 in healthcare, HTTP, OpenAPI*

BPP

The FAIR data standard

An introduction

FAIR data principles include:

- Findability
- Accessibility
- Interoperability
- Reusability

Defined in a March 2016 paper in Scientific Data by a consortium of scientists and organisations.

Components of the FAIR standards

Applying the FAIR data standard

Group discussion

Credit Bank Corporation's HR department encountered data issues due to inconsistent data practices across the enterprise.

This resulted in inaccuracies in key HR metrics.

Applying FAIR data principles can resolve this challenge.

Consider the following questions:

- How can the bank ensure that HR data is "Findable" by authorised users?
- How can the bank achieve "Interoperability" between HR data and other datasets or systems?

Case study: *Credit Bank Corporation*

Group discussion

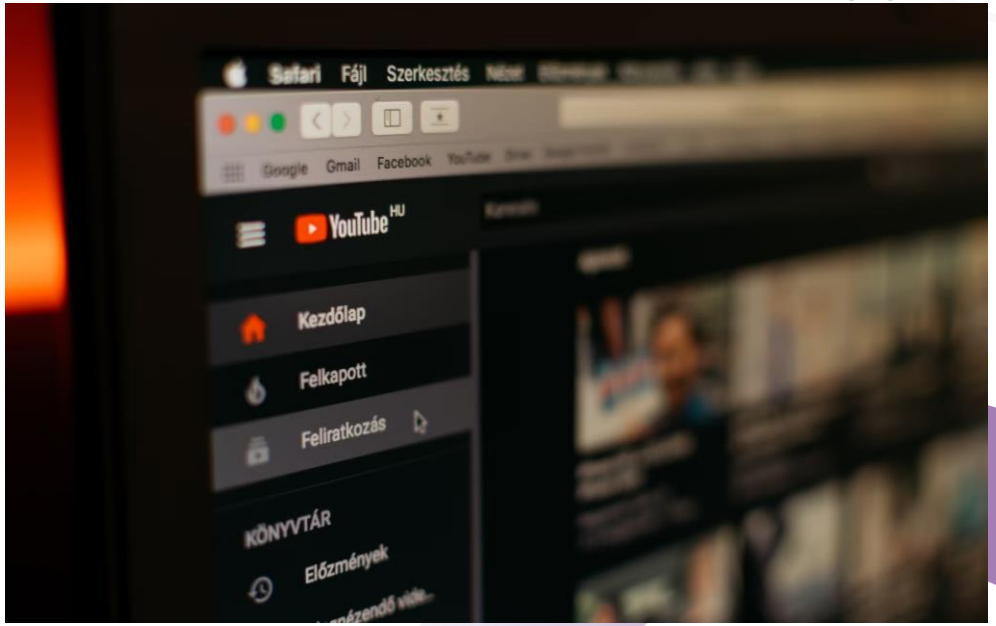
Possible answers

- Metadata and Data Cataloging
- Unique Identifiers
- Search Tools and Indexing
- Standardization of Naming Conventions
- Access Control and Security
- Data Formats and Standards
- APIs for Data Exchange
- Semantic Interoperability
- Data Integration Tools
- Collaboration and Training

The Dublin Core Metadata Initiative

Introducing DCMI

The Dublin Core Metadata Initiative (DCMI) offers a set of metadata standards used to describe resources in various domains.



Dublin Core™ Metadata Initiative

Its core elements provide a basic framework for describing digital resources such as documents, images, videos, and web pages.

BPP

The Dublin Core Metadata Initiative

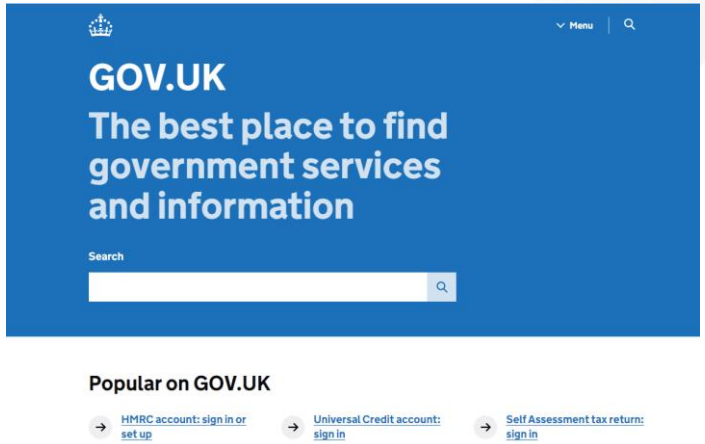
Use cases and implementation

Digital libraries

Archives and museums

Educational resources

Digital publishing



Government portals

Unstandardised data

The challenges

In today's abundance of data, organisations face a myriad of challenges when dealing with unstandardised data.

These include the following:

Data silos

Inconsistencies

Integration issues

Unstandardised data

Overcoming the challenges

Standards play a pivotal role in overcoming issues of data siloing, inconsistencies and integration.

Navigating data quality issues

Section Introduction

Data formats and quality issues

We will now explore the common issues with the quality of widely-used data formats and how to overcome them, including:

- XML
- CSV
- JSON

Addressing data quality issues with these formats is crucial for enabling seamless data exchange and integration.

Challenges and quality issues

XML

While XML offers flexibility and interoperability, it also presents quality issues. Here is some guidance to help you:

Issue	Description	How can it be mitigated?
Parsing errors	XML parsing errors stem from improper document structure or syntax	Validate, use parsers, implement error handling, sanitise data, and conduct testing
Character encoding	Stem from poorly coded or inconsistent encoding schemes	Standardise character encoding, encode special characters, validate encoding
Schema validation	Verifying that XML documents adhere to a predefined schema or structure	Adhering to defined schema

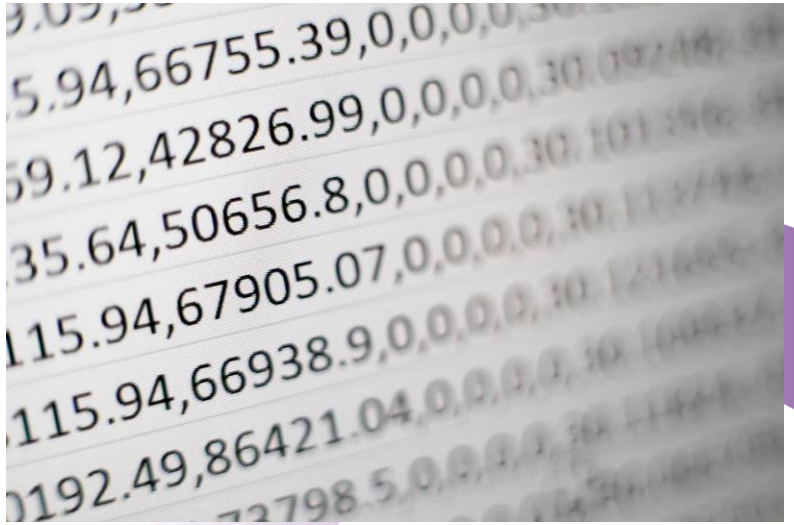
The XML data format

Challenges and quality issues

CSV

While CSV files offer simplicity and portability, they also present several quality challenges. Here is some guidance to help you:

Issue	Description	How can it be mitigated?
Inconsistent formats	Inconsistencies in delimiters, encapsulators, or line terminators	Standardise the use of delimiters, encapsulators, and line terminators, Implement robust parsing algorithms
Data type mismatch	The absence of explicit data types in CSV files	Provide metadata or schema information, perform data validation and type casting during parsing
Header misalignment	Missing or misaligned headers	Ensure consistent headers and alignment, implement error handling mechanisms



The CSV data format

Challenges and quality issues

JSON

While JSON offers several advantages in terms of its lightweight and human-readable format, it also presents a set of challenges.

Here is some guidance to help you:

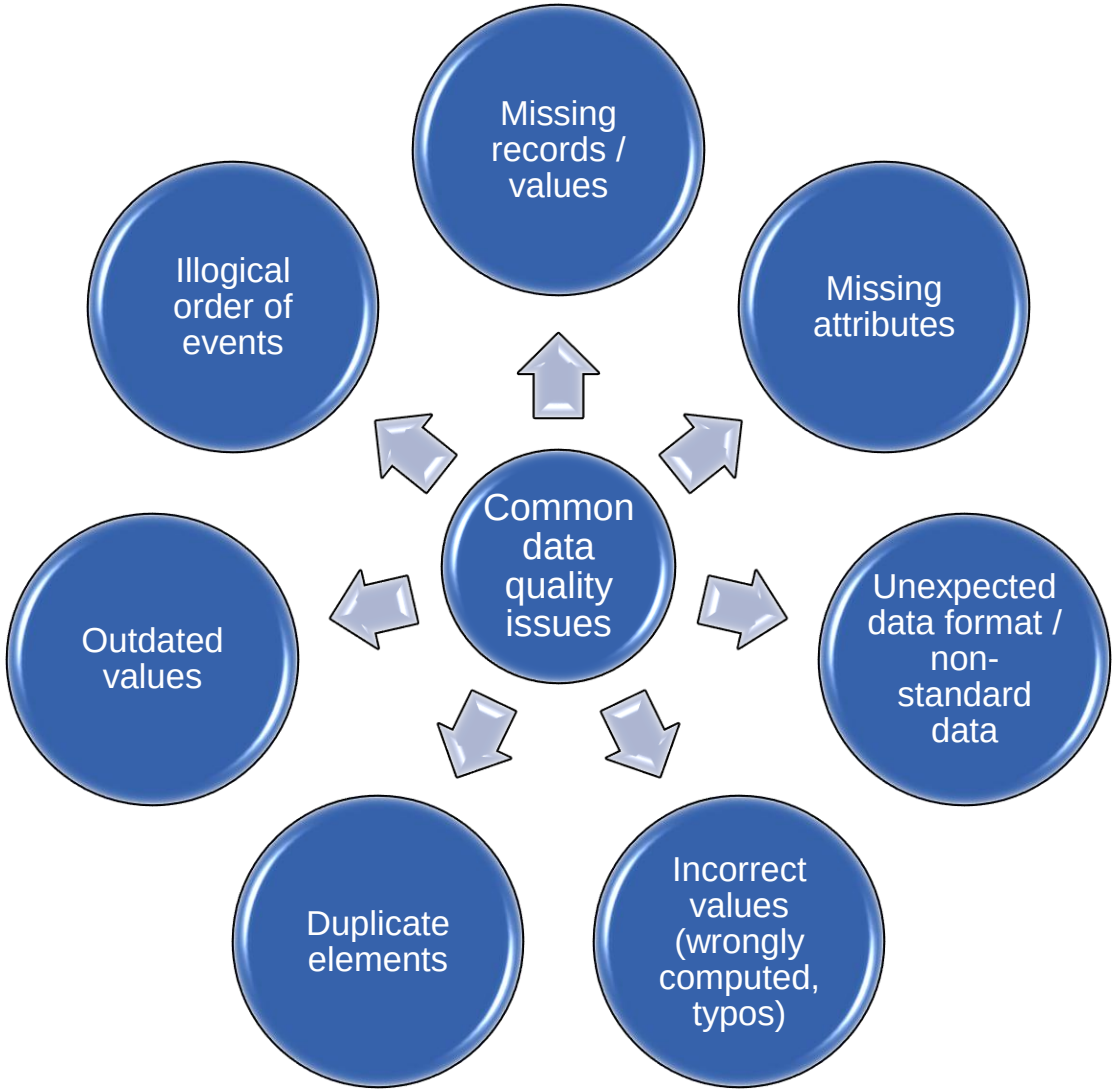
Issue	Description	How can it be mitigated?
Nesting complexity	Nested structures can complicate parsing and increase processing time	Design JSON to balance complexity and efficiency
Key-value pair integrity	Inconsistent key naming or missing/extra keys can lead to errors	Establish naming conventions and validation rules
Data volume	Large files can present challenges for memory usage, processing time, and network bandwidth	Data streaming, pagination, compression, and distributed processing

The JSON data format

BPP

Other common data quality issues

Further data quality issues you may encounter include the following:



	id	first_name	last_name	email
▶	1	Carine	Schmitt	carine.schmitt@verizon.net
	4	Janine	Labrune	janine.labrune@aol.com
	6	Janine	Labrune	janine.labrune@aol.com
	2	Jean	King	jean.king@me.com
	12	Jean	King	jean.king@me.com
	5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
	10	Julie	Murphy	julie.murphy@yahoo.com
	11	Kwai	Lee	kwai.lee@google.com
	3	Peter	Ferguson	peter.ferguson@google.com
	9	Roland	Keitel	roland.keitel@yahoo.com
	14	Roland	Keitel	roland.keitel@yahoo.com
	7	Susan	Nelson	susan.nelson@comcast.net
	13	Susan	Nelson	susan.nelson@comcast.net
	8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Duplicate elements example

Applying your learning

Group discussion

The team at Credit Bank Corporation will have to import data from multiple systems to build their HR dashboard, including:

- Enterprise Management Cloud (XML)
- Payroll software (export as CSV)
- 360 Feedback Tool for employee evaluations (exports as CSV)

Consider the following questions:

- What potential XML quality issues may arise when importing from the Enterprise Management Cloud? How can they be mitigated?

- What potential CSV issues may be found when importing payroll and feedback data? How can they be mitigated?

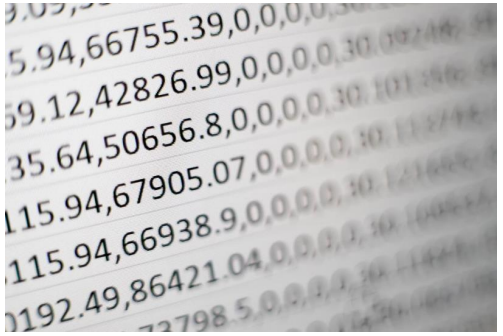
Case study: *Credit Bank Corporation*

Group discussion

Importing HR data

Practical top tips!

Follow these tips to ensure data quality and seamless integration when importing HR data in different formats:

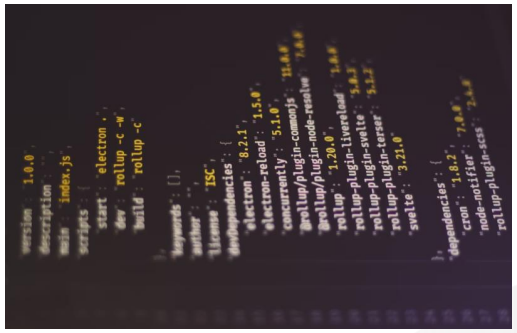


The CSV data format

- Watch for data type misinterpretations, define columns explicitly
- Check date/time formats, convert to standard corporate timezone
- Support Unicode for non-Latin text fields like names

The XML data format

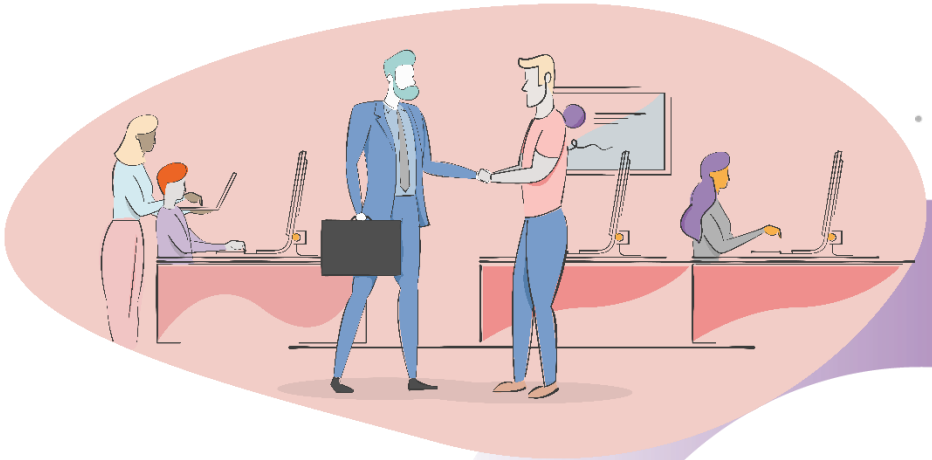
- Validate structure/syntax to prevent parsing errors
- Standardize encoding, escape special characters
- Adhere to defined schemas for conformity



The JSON data format

- Design nested structures for complexity/efficiency balance
- Implement strict naming/validation for key-value pairs
- For large files, use streaming, pagination, compression

Common strategies for ensuring data quality

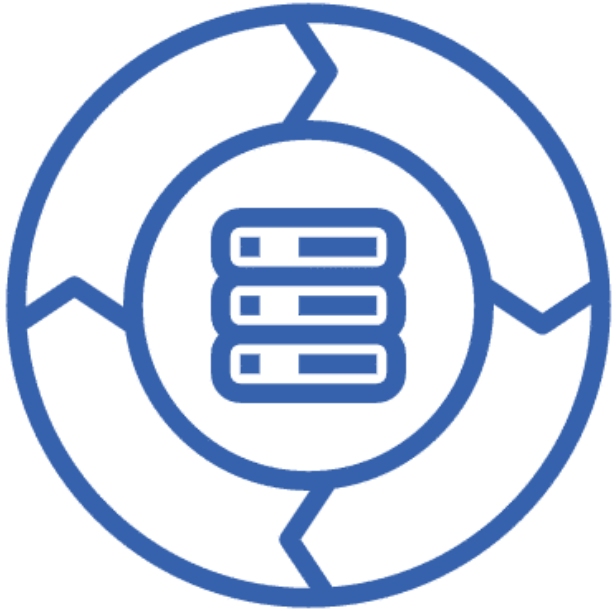


Section Introduction

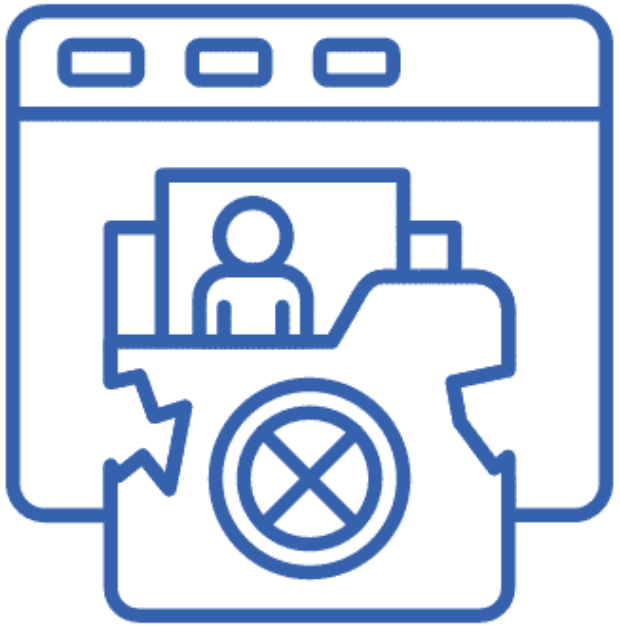
Data lineage and unique identifiers

Let’s now take a look at data lineage and unique identifiers.

These are pivotal for transparent and reliable data management, ensuring integrity and compliance.



Data lineage: the lifecycle of data, encompassing its origin, usage, and movement over time



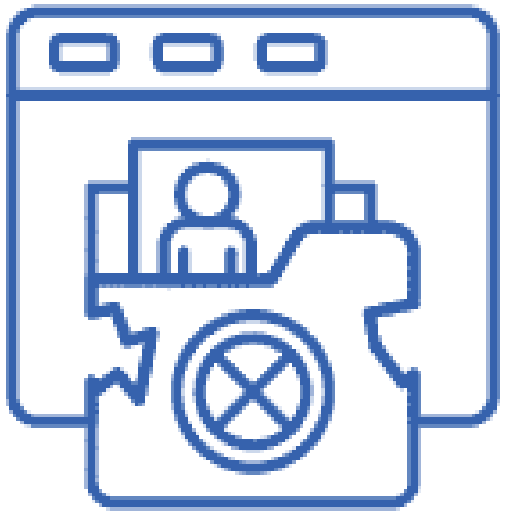
Unique Identifiers: UIDs ensure each data item is distinct, enabling effective data management and retrieval

UIDs vs UUIDs

A comparison

So, what's the difference between Unique Identifiers (UIDs) and Universally Unique Identifiers (UUIDs)?

Here's how to remember them:



Vs

UIDs: Unique within their individual system

UUIDs: UUIDs are globally unique across different systems

Data lineage and unique identifiers

Application in the context of our case study

Credit Bank Corporation planned to build a comprehensive analytics dashboard to track key HR metrics

- In HR dashboards, UIDs can differentiate employees, track their performance reviews, and manage payroll data securely and efficiently
- UIDs vs. UUIDs: While both serve as unique identifiers, UUIDs provide a higher level of uniqueness, generated through standardised algorithms

An analytics dashboard

In the case of the HR dashboard, it is crucial that we use a single, unique, and standardised identifier for every employee!

Data wrangling and cleansing

An introduction

Data wrangling and cleansing are essential processes in data engineering, where raw data is transformed into a structured and usable format for analysis and decision-making.

In the context of our case study, a HR dataset for Credit Bank Corporation could contain data quality issues such as

- Missing values
- Incorrect data types
- Duplicate entries

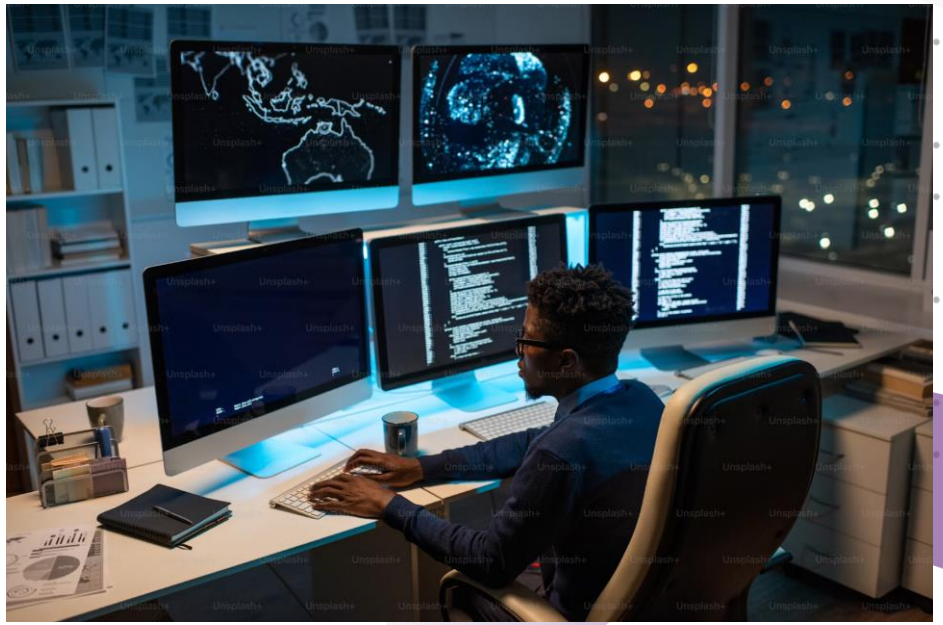
Case study: *Credit Bank Corporation*

Data wrangling and cleansing

For data engineers

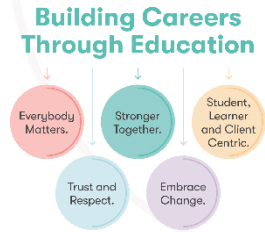
Using a data wrangling tool like Pandas in Python, a data engineer could:

- Remove duplicate records
- Handle missing values (imputation or deletion)
- Convert data types (e.g. converting strings to dates)
- Standardise formatting (e.g ensuring consistent date formats)
- Direct and remove outliers or anomalies



A data engineer at work

Knowledge Check Poll



Which of the following is NOT a benefit of understanding data lineage?

A) Ensures transparency in data usage and movement

B) Aids in understanding patterns of data usage

C) Enhances trust in data for strategic decision-making

D) Eliminates the need for data quality checks

Feedback: D – While data lineage provides valuable insights into the lifecycle of data, it does not eliminate the need for data quality checks.

Submit your responses to the chat!

BPP

Focus on Data Stewardship

Section Introduction

Data stewardship

- Data stewardship: Assigns roles for ensuring data accuracy, completeness, reliability
- Importance: Crucial for data quality management
- Best practices: Implementing effective stewardship within organisation

What would data stewardship mean in the context of our case study?

Data stewards are the guardians **of accuracy**, **integrity**, **consistency** and **timeliness** metrics.

Credit Bank Corporation's analytics dashboard

Applying data stewardship

In the context of our case study

The HR team's analytics dashboard vision faces data quality issues, jeopardising key metric integrity and utility.

To combat these issues the data engineering team:

- Explores data quality best practices with data stewardship as the main focus
- Considers a ‘hive-mind’ approach for refining data quality
- Implements strategies based on standards to ensure more accurate datasets

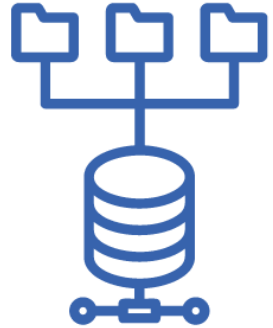
Credit Bank Corporation’s data engineering team

Data stewardship

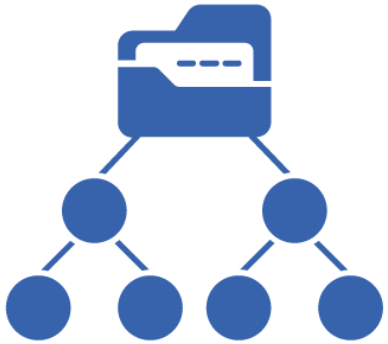
How to implement?

Assignment of Data Ownership

Data Stewardship



Ensuring Data Provenance



Automated Indexing of Data

Quality Metrics Tracking

Regular Data Audits

BPP

Overcoming data challenges with data stewardship

In the context of our case study

Data stewardship techniques that help to overcome challenges in data quality for our HR analytics dashboard, include:

Data cleansing initiatives

Data validation processes

Improvement of data entry procedures

Enhanced data reporting

Feedback loops for continuous improvement

Credit Bank Corporation's HR analytics dashboard

Knowledge check poll

Which of the following is NOT a best practice for data stewardship?

A) Assignment of data ownership

B) Ensuring data provenance

C) Automated indexing of data

D) Ignoring feedback loops for continuous improvement

Feedback: D – Ignoring feedback loops for continuous improvement is not a best practice for data stewardship.

Submit your responses to the chat!

BPP

Focus on Sustainability

Focus on sustainability

Data compression techniques

Here’s what you need to know:

- Data compression techniques reduce the environmental impact of data storage and transfer
- Advanced compression algorithms decrease the size of data at rest and in transit
- This reduction leads to lower storage and bandwidth requirements
- Consequently, it results in lower energy consumption and carbon footprint

Brotli Compression: *A lossless data compression algorithm developed by Google*

Focus on sustainability

Green data transfer

Organisations can also adopt green data transfer practices to minimise their environmental impact.

These practices include:

- Smart Data Transfer Scheduling
- Green Protocols for Data Transfer
- Regular Audits of Data Efficiency

Green Data Transfer

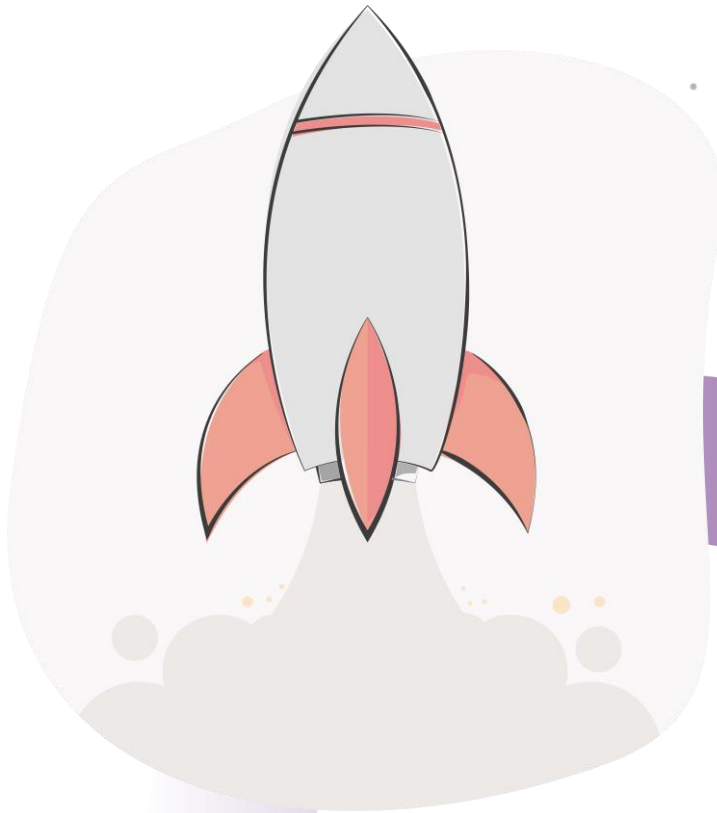
Image source:
InformationSecurityManagement.co.uk

Focus on sustainability

Strategies to support net-zero goals

If you remember to implement the following data quality strategies, you are well on your way to successful data stewardship:

- Avoiding data duplication
- Lean data management
- Rationalising data storage
- Data lifecycle management



Knowledge check poll

Building Careers Through Education

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

Which of the following is NOT a recommended practice for green data transfer?

A) Utilising data compression techniques

B) Scheduling data transfers during peak hours

C) Choosing energy-efficient transfer protocols

D) Conducting regular audits of data efficiency

Feedback: B - Scheduling data transfers during peak hours is not a recommended practice for green data transfer.

Submit your responses to the chat!

BPP

Data Quality Management Beyond Data Stewardship

Beyond data quality management

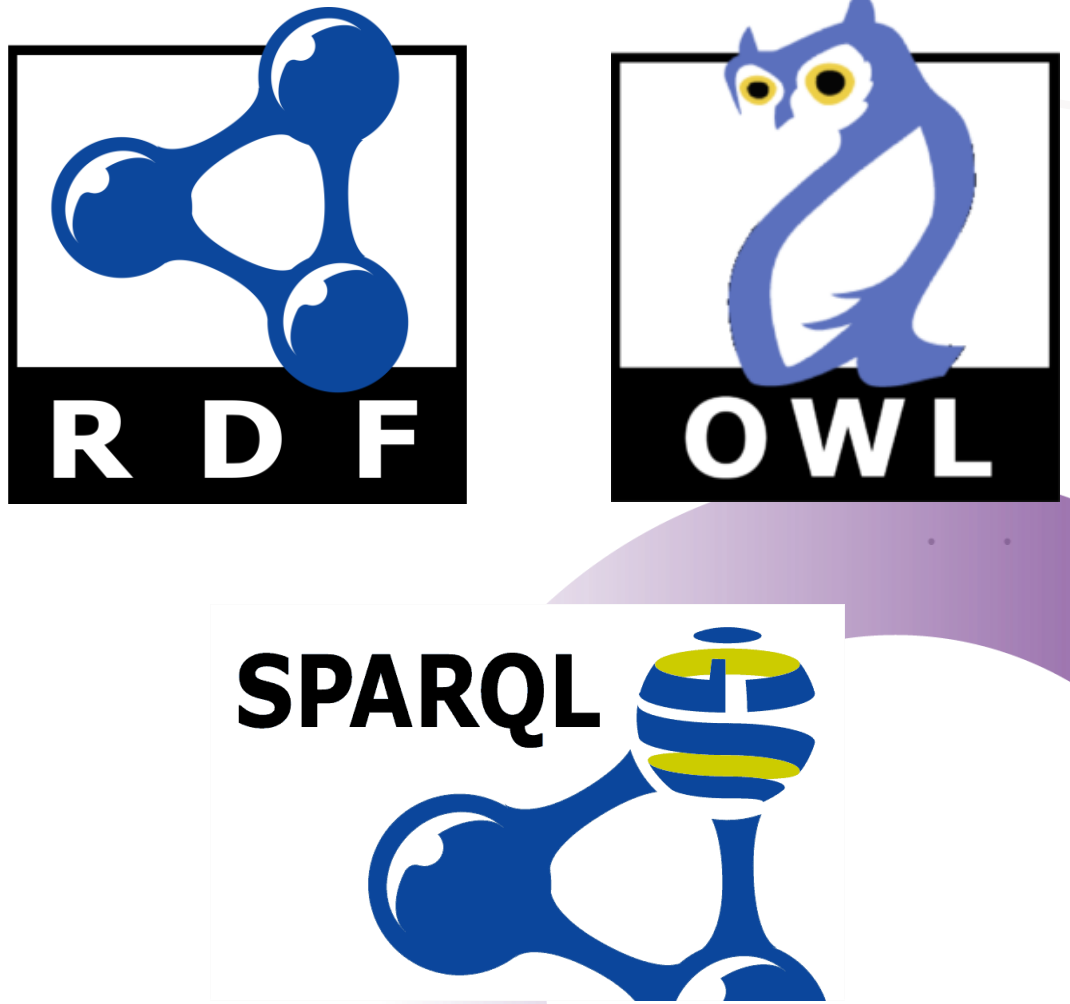
Semantic validity in linked data

In the context of linked data, semantic validity refers to ensuring that data accurately represents the real-world constructs it is intended to model.

This is particularly important for facilitating interoperability and meaningful data integration across diverse datasets.

Semantic validity models include:

- RDF (Resource Description Framework)
- SPARQL
- OWL (Web Ontology Language)



Semantic validity models

BPP

Beyond data quality management

Best practices for semantic validation

Here's what you need to know:

- Consistent use of vocabularies and ontologies ensures adherence to predefined terms and definitions
- Enhances interoperability and semantic clarity of data
- Enables effective communication and interpretation across different systems and stakeholders

Semantic validity models

BPP

Implementing a data quality framework

A step-by-step guide (DMAIC)

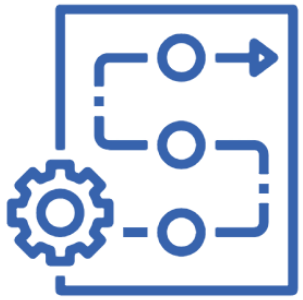
Developing a robust data quality framework is essential for organisations to ensure the reliability, accuracy, and integrity of their data assets.

Here's how:

Define	Measure	Analyse	Improve	Control
• Clearly define the objectives	• Establish metrics and KPIs	• Conduct in-depth analysis	• Implement corrective actions	• Establish monitoring mechanisms

The role of data governance

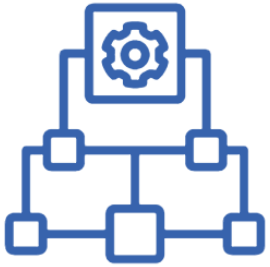
Supporting data quality



Establishing policies, procedures, and standards

The Role of Data Governance in Data Quality

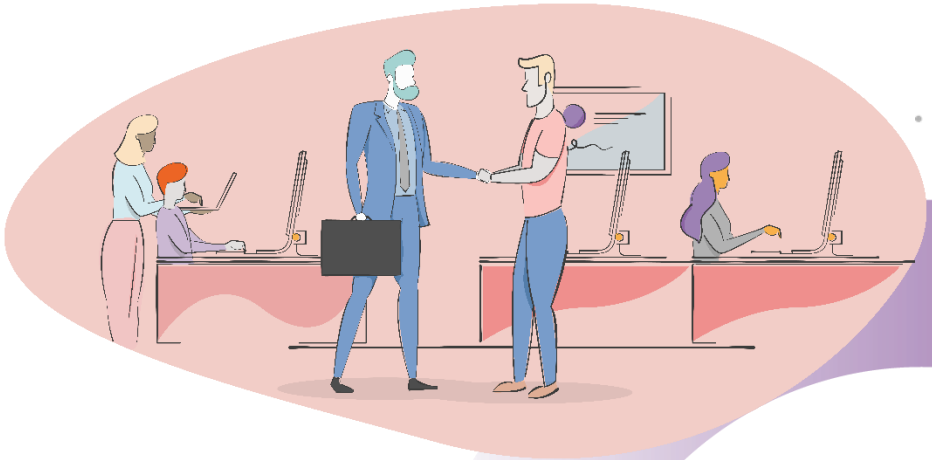
Ensures accountability and responsibility



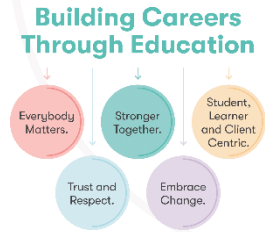
Provides a framework for decision-making

Facilitates collaboration and communication

Practical: Data Quality Testing



Apply hands-on exercise

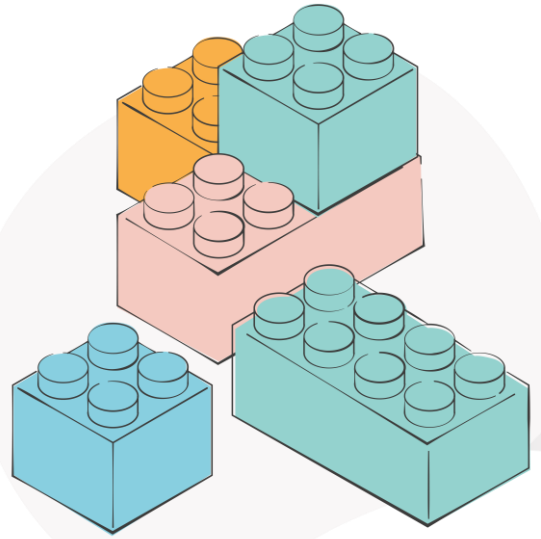


Your homework (part 1)

Your brief is as follows:

1. Find & download IBM HR Analytics dataset from Kaggle
2. Analyse metadata & identify potential data quality issues
3. Prepare data: retain specific columns, create metadata sheets
4. Document findings, upload file to GitHub
5. Review KSBs covered, reflect in learning journal

Step-by-step guidance is provided here: M1T2 Apply: Intro to data



Apply exercise

BPP

Apply hands-on exercise

Your homework (part 2)

In this exercise, you create a high-quality dataset by consolidating data from three separate low-quality datasets.

These are currently provided as three separate sheets:

- Mwanza_r, Pwani_r, DSM(ilala)_r

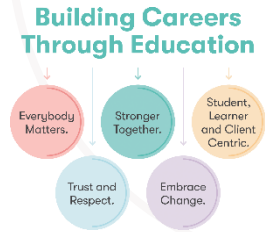
You have received the brief below in an email:

```
From: <datascientist@company.com>
To: <dataengineer@company.com>
Re: Dataset usability - healthcare facility data

Good morning, as part of our new healthcare facility project for Tanzania, we would like to
perform clustering on the data for the three localities Mwanza, Pwani and Ilala. The data has
been sourced from local authorities, they all use slightly different formats. Can you please
pool this into a single dataset and make sure any quality issues are addressed. Please also
create a schema for the data.

Many thanks,
```

Step-by-step guidance is provided here: M1T2 Apply: Intro to data

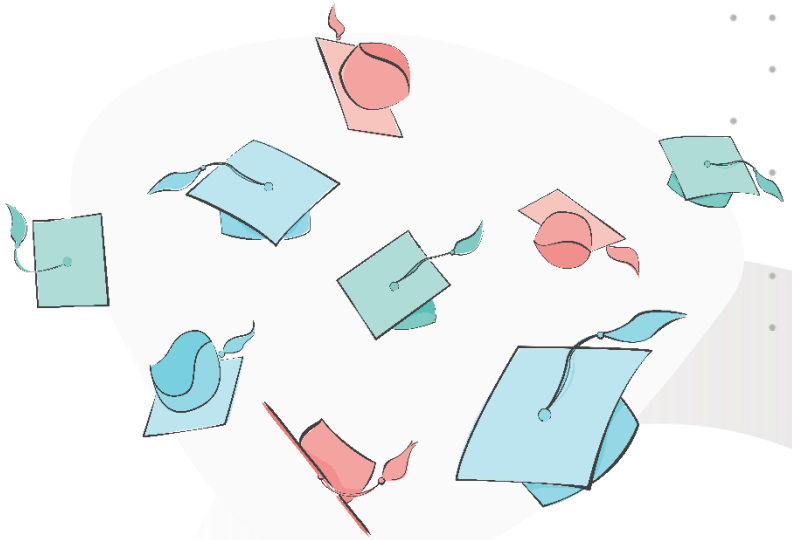
BPP

Key Learning Summary



The key takeaways from this session are as follows:

- Data quality encompasses dimensions like accuracy, integrity, consistency, timeliness, and reliability
- Following data and metadata standards like FAIR principles, Dublin Core, etc. helps improve data findability, accessibility, interoperability, and reusability
- Data lineage tracking and use of unique identifiers enable transparent data management and enhance data integrity/compliance
- Data wrangling, cleansing, and stewardship techniques help improve data quality
- Data compression and green data transfer practices aid sustainability and net-zero goals by reducing storage/bandwidth requirements
- Implementing a robust data quality framework (e.g. DMAIC) supported by data governance is key for maintaining high data quality standards



BPP

Thank you

Do you have any questions, comments, or feedback?