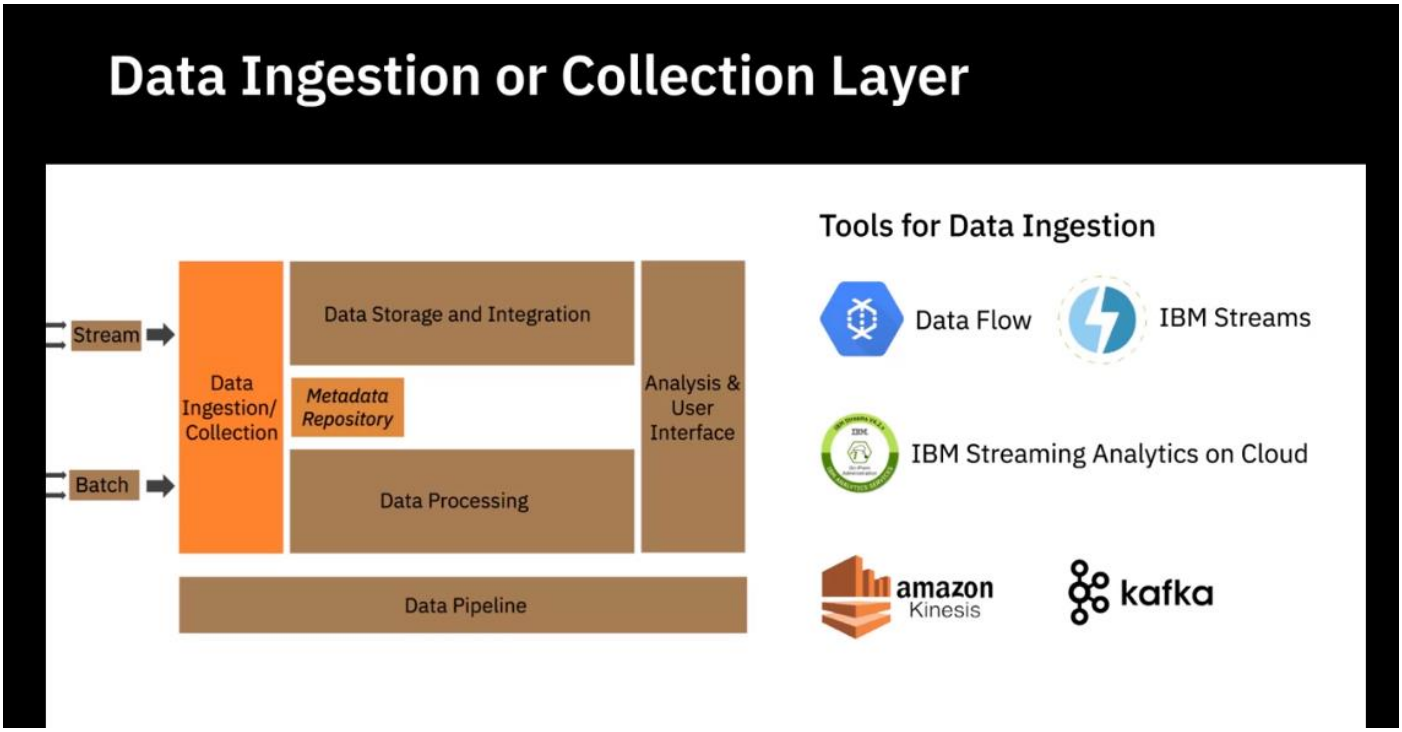
Week 3

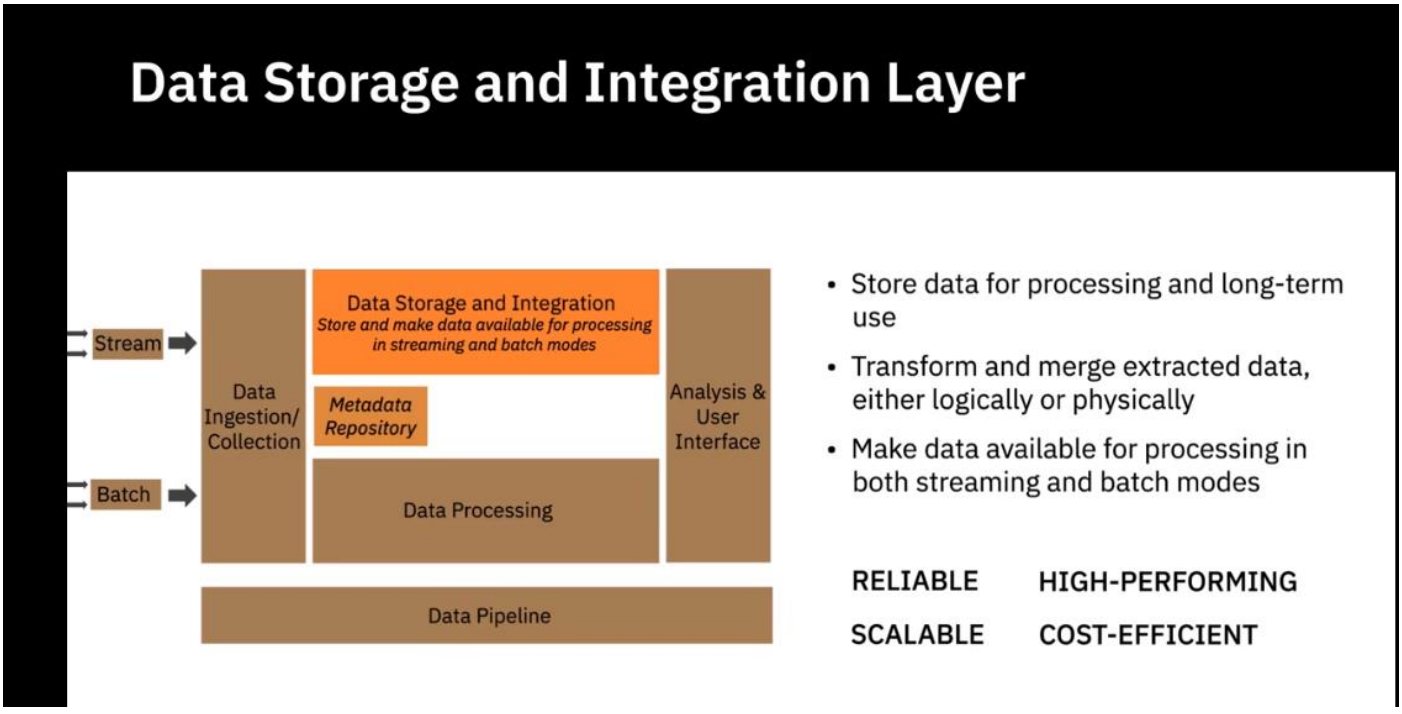
Architecting the Data Platform

Data Ingestion or Collection Layer

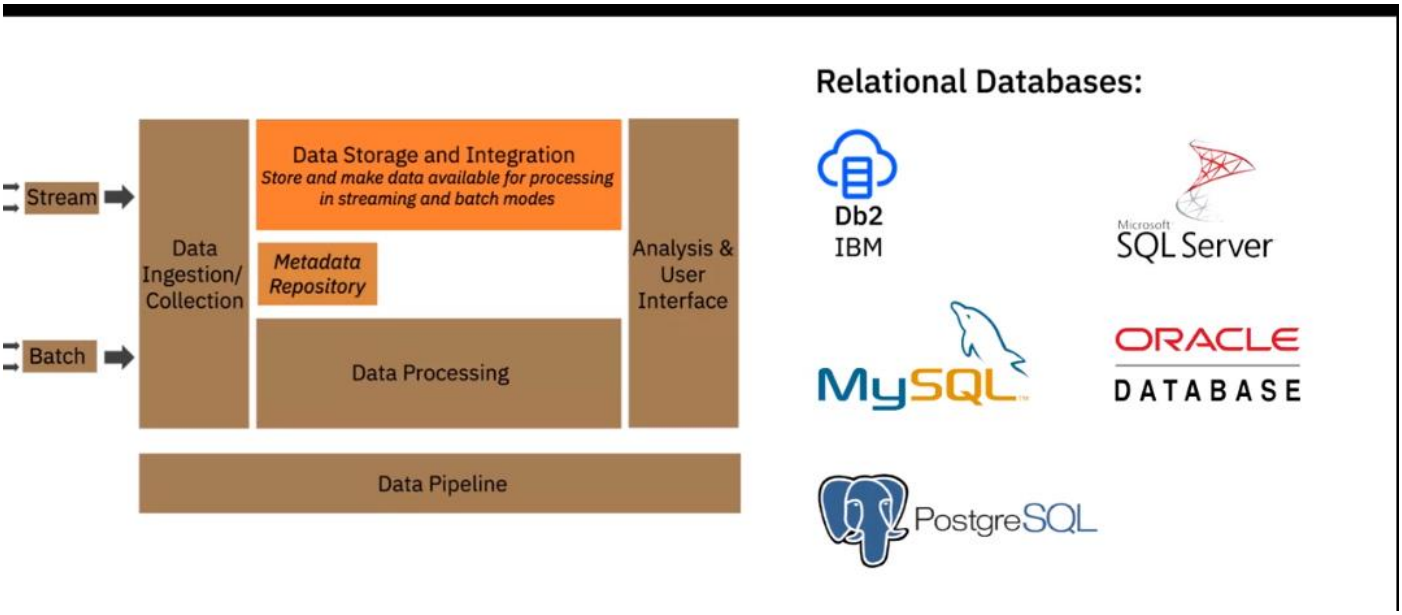
Tools for data ingestion



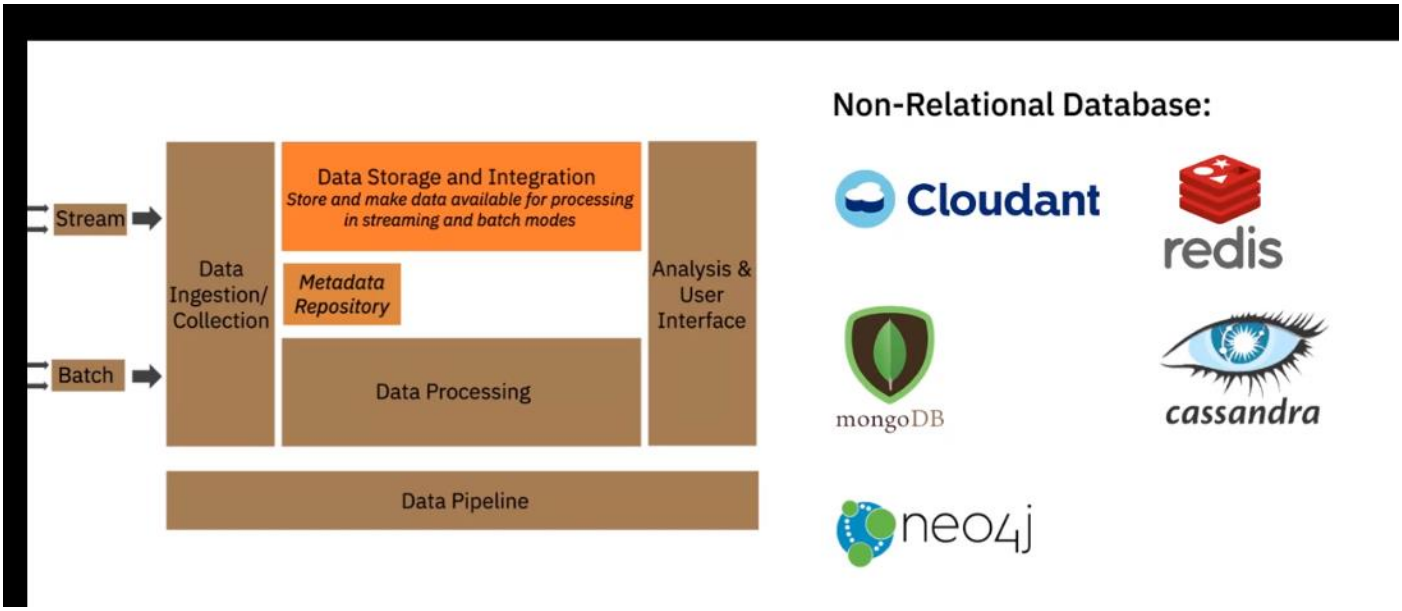
Data Storage Layer



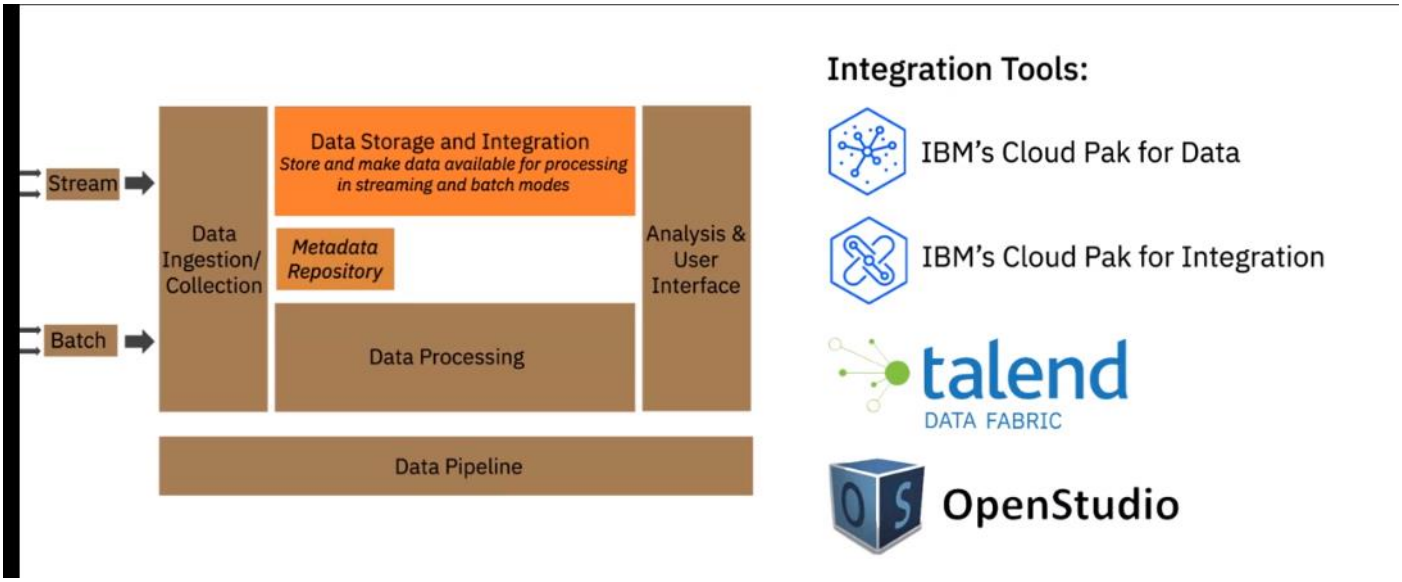
Ralational Databases



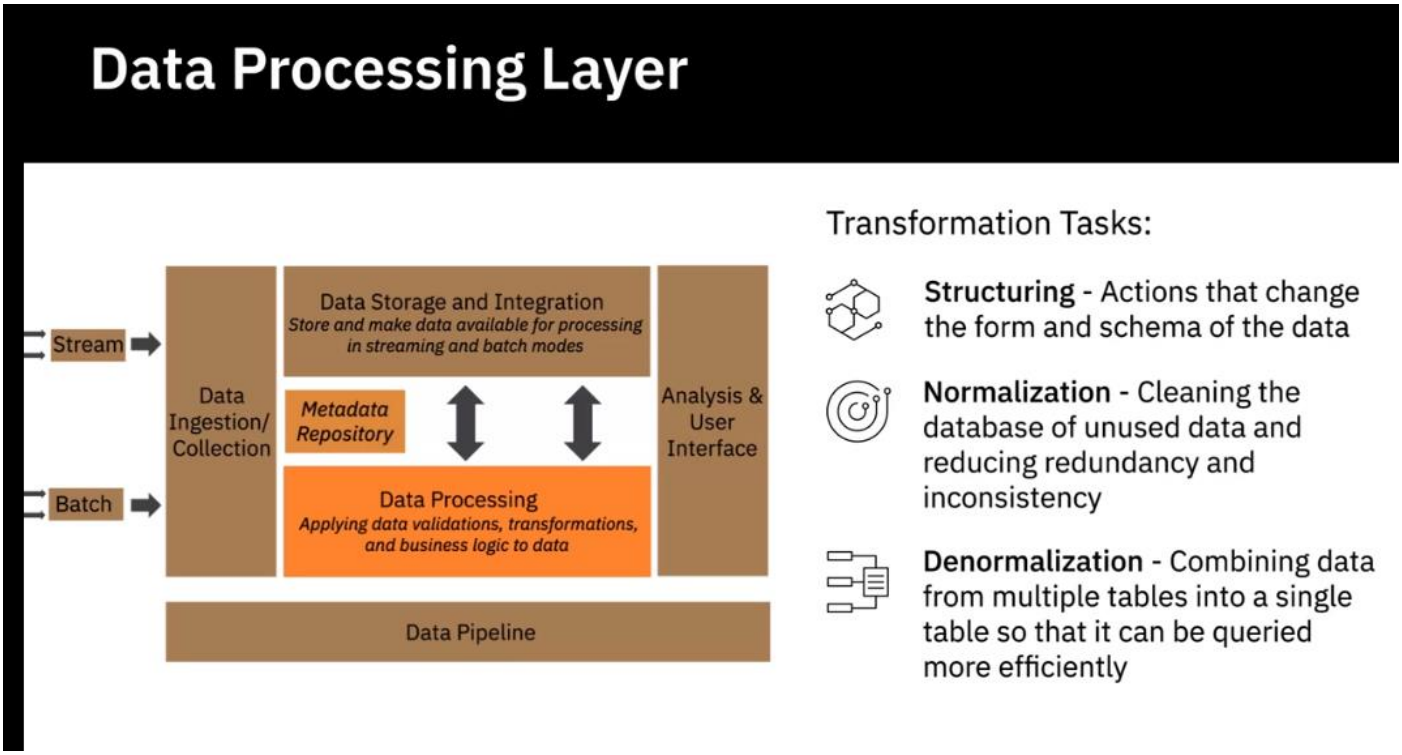
Database as a service

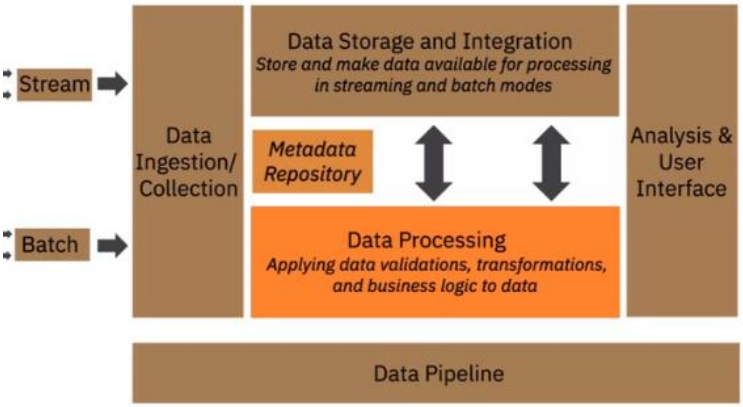


Integration Tools



Data Processing Layer

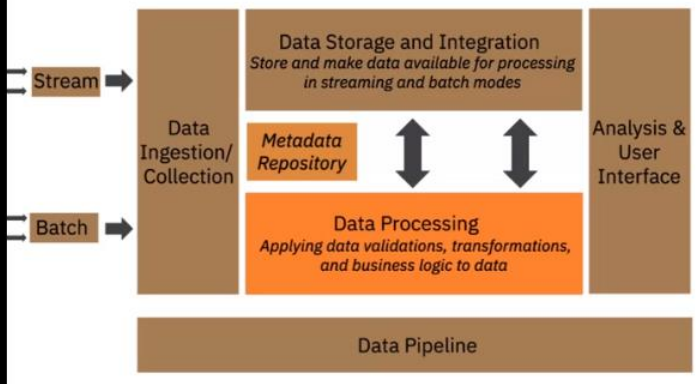




Transformation Tasks:

Data Cleaning - Fixing irregularities in data to provide credible data for downstream applications and uses

Data Processing Layer



Spreadsheets

TRIFACTA Wrangler

OpenRefine

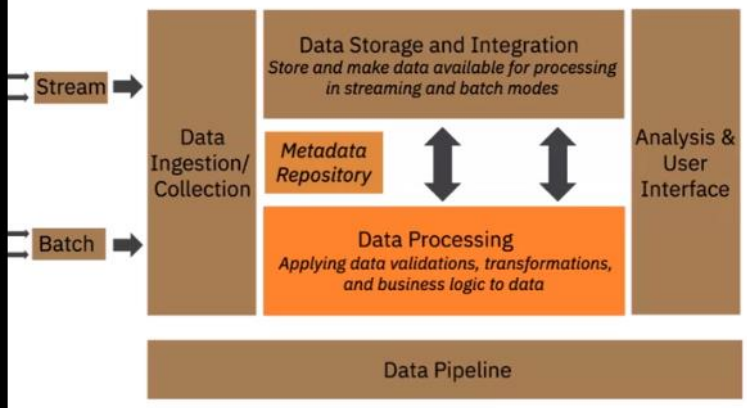
python

Google DataPrep

R

Watson Studio Refinery

Data Processing Layer



Storage and Processing may not always be performed in separate layers.

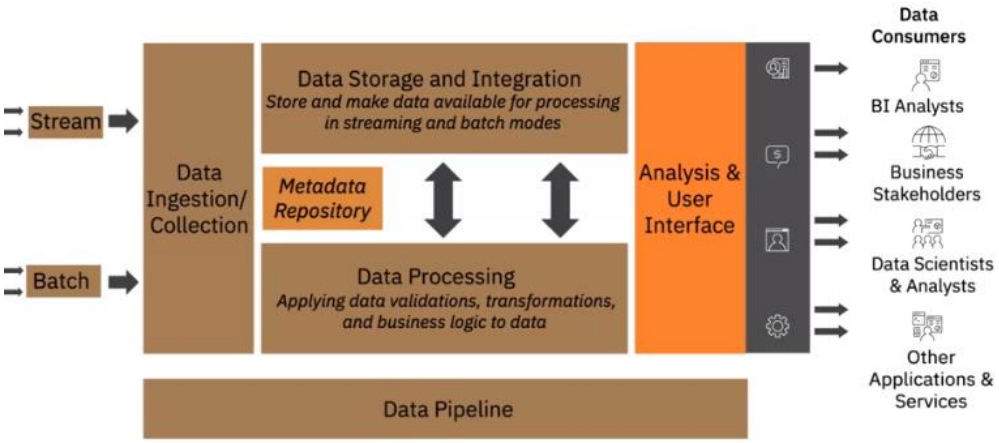
Relational Databases

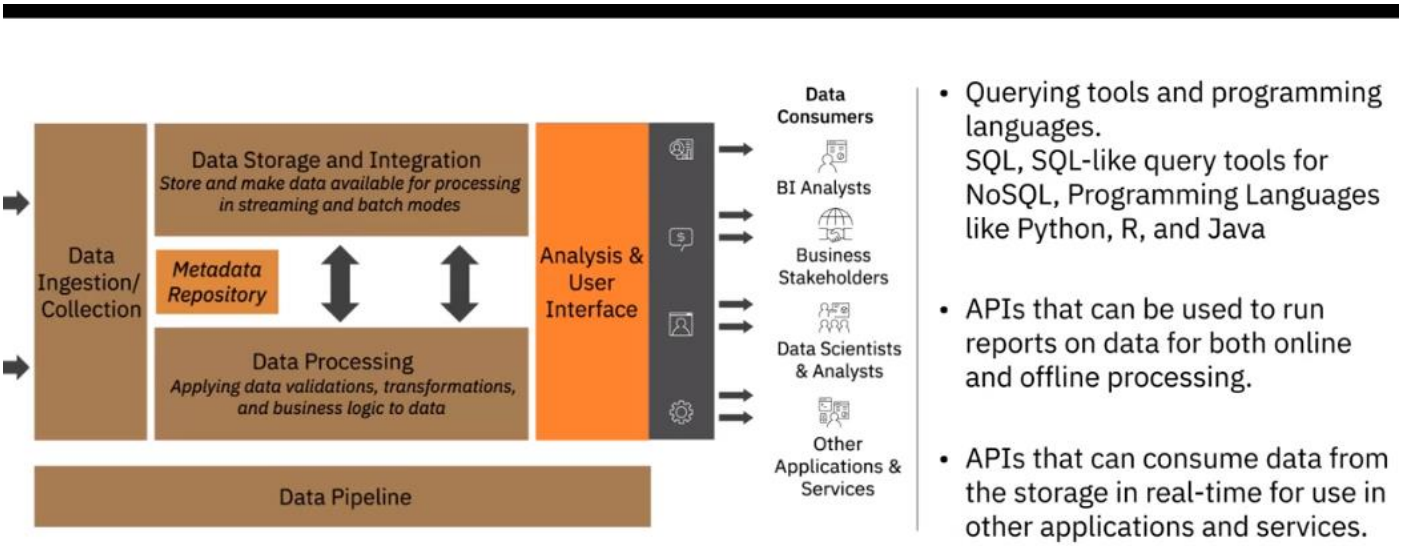
Storage and **Processing** can occur in the same layer.

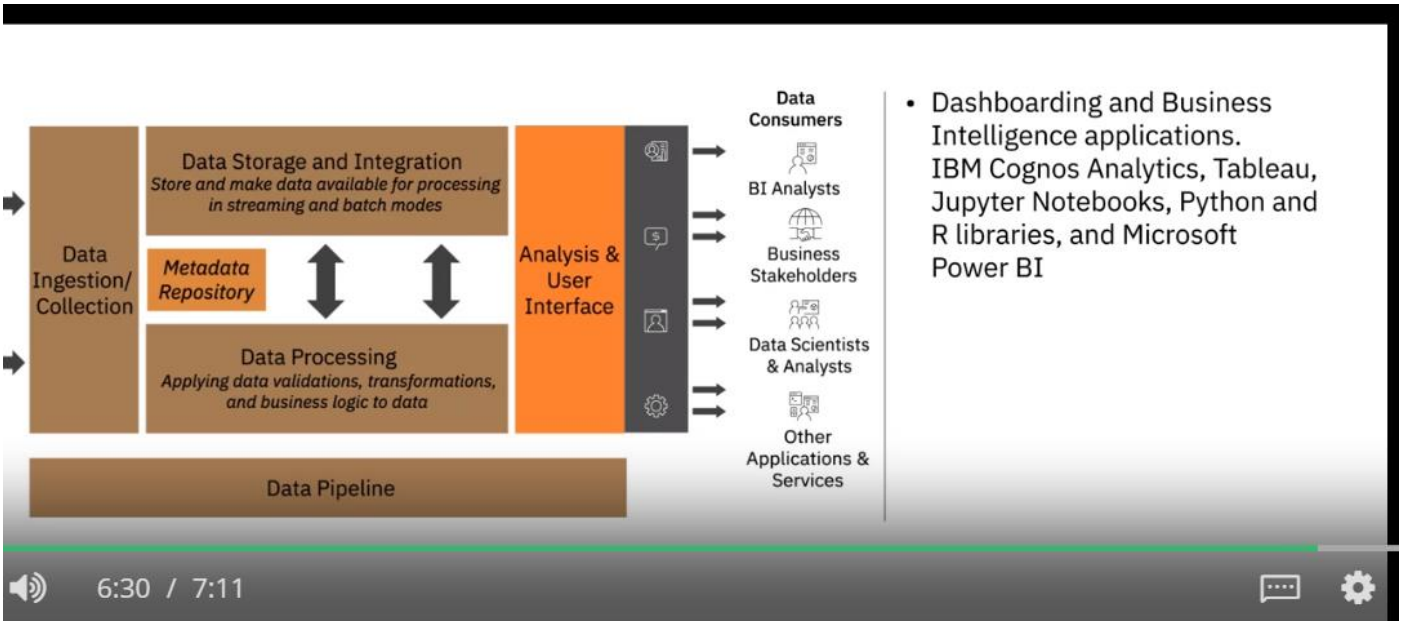
Big Data Systems

Data can first be stored in **Hadoop File Distribution System**, or **HDFS**, and then processed in a data processing engine like **Spark**.

Analysis User Interface Layer

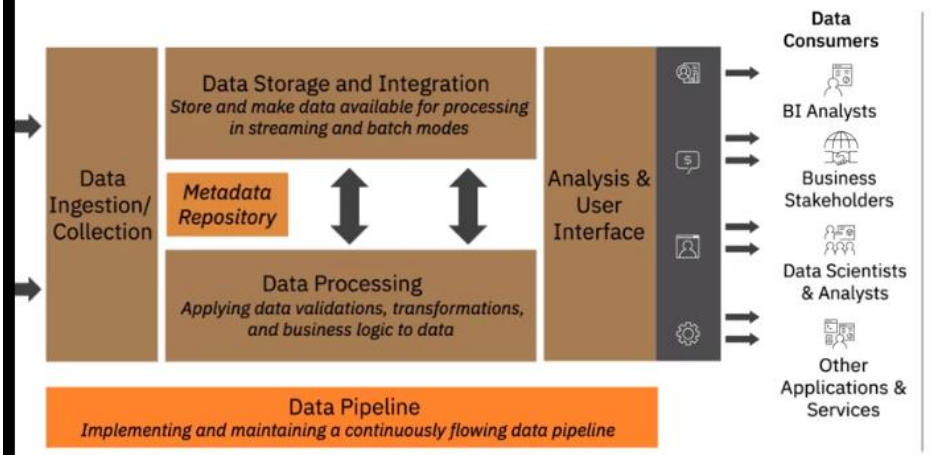






Data Pipeline Layer

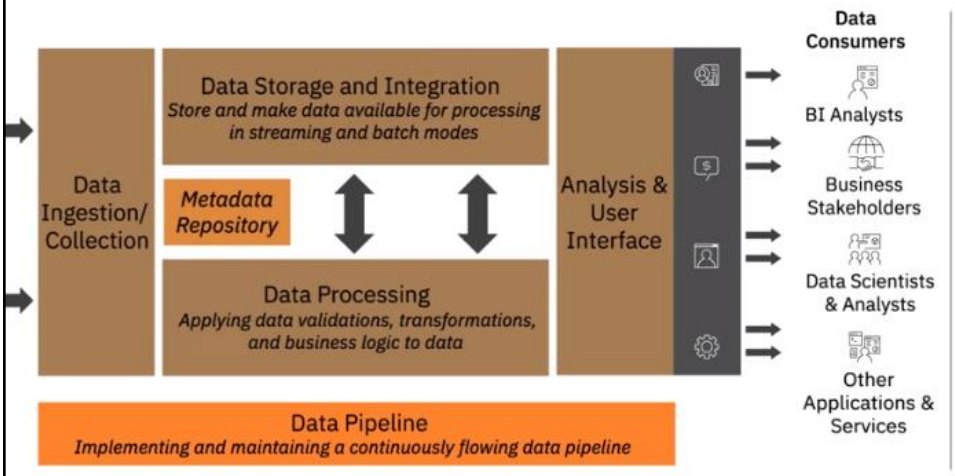
Data Pipeline Layer



Overlaying the Data Ingestion, Data Storage and Integration, and Data Processing layers is the Data Pipeline layer with the Extract, Transform, and Load tools.

This layer is responsible for implementing and maintaining a continuously flowing data pipeline.

Data Pipeline Layer



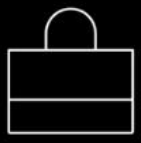
Apache Airflow

Data Flow

Factors for Selecting and Designing Data Stores

Introduction

A Data Store, or Data Repository, refers to data that has been collected, organized, and isolated so that it can be:



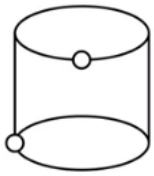
Used for business operations

Mined for reporting and data analysis.

Considerations of designing a data store

Considerations for designing a Data Store

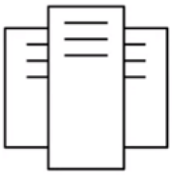
A repository can be:



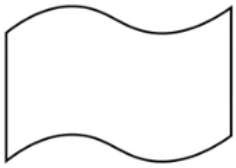
Database

Data Warehouse

Data Mart



Big Data Store



Data lake

Considerations for designing a Data Store

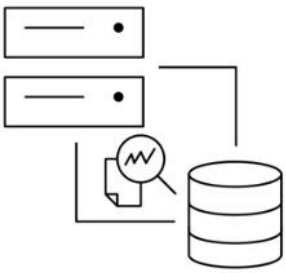
Primary considerations for designing a data store:

- Type of data
- Volume of data
- Intended use of data
- Storage considerations
- Privacy, Security, and Governance needs

Type of Data

Type of Data

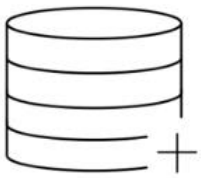
There are multiple types of databases and selecting the right one is a crucial part of designing.



- **Input**
- **Storage**
- **Search and Retrieval**
- **Modification**

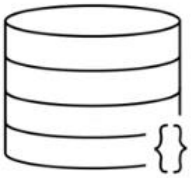
Type of Database

Type of Data



RDBMS

Best used for structured data, which has a well-defined schema and can be organized into a tabular format.



NoSQL

Best used for semi-structured and unstructured data, data that is schema-less and free-form.

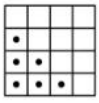
Type of nosql database

Type of Data

Four types of NoSQL Databases:

Key-value

Document-based



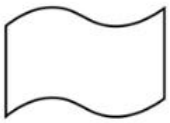
Column-based

Graph-based

Volume of Data

Volume of Data

Volume / Scale of data:



Data Lake

- Store large volumes of raw data in its native format, straight from its source
- Store both relational and non-relational data at scale without defining the data's structure and schema

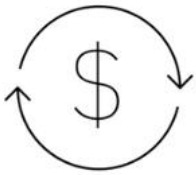
Big Data Store

- Store data that is high-volume, high-velocity, of diverse types, needs distributed processing for fast analytics
- Big Data Stores split large files across multiple computers allowing parallel access to them
- Computations run in parallel on each node where data is stored

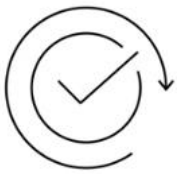
Intended use of Data

Intended use of Data

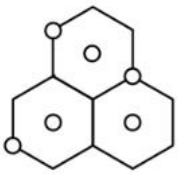
How you intend to use the data you are collecting:



Number of Transactions

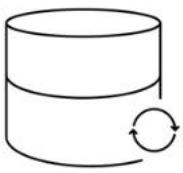


Frequency of Updates



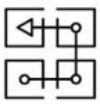
Type of Operations

Response Time



Backup and Recovery

Intended use of Data



Transactional Systems used for capturing high-volume transactions, need to be designed for high-speed read, write, and update operations.

Analytical Systems need complex queries to be applied to large amounts of historical data aggregated from transactional systems. They need faster response times to complex queries.

Schema design, indexing, and **partitioning** strategies have a big role to play in performance of systems based on how data is getting used.

Intended use of Data

The intended use of data also drives **scalability** as a design consideration.

Normalization is another important consideration at the design stage.

Optimal use of storage space

Makes database maintenance easier

Provides faster access to data

Normalisation: Transactional Data

Storage Considerations

Performance, availability Integrity and recoverability

Storage Considerations

Design considerations from the perspective of storage:

Performance, Availability, Integrity, and Recoverability of Data

- **Performance** - Throughput and Latency

  Throughput: Rate at which information can be read from and written to the storage

  Latency: Time it takes to access a specific location in storage.

Storage Considerations

Design considerations from the perspective of storage:

Performance, Availability, Integrity, and Recoverability of Data

- **Availability** - Storage solution must enable you to access your data when you need it, without exception. There should be no downtime.

Storage Considerations

Design considerations from the perspective of storage:

Performance, Availability, Integrity, and Recoverability of Data

- **Recoverability -** Storage solution should ensure you can recover your data in the event of failures and natural disasters.

Privacy Security and Governance

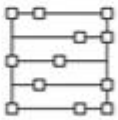
Privacy, Security, and Governance

A secure data strategy is a layered approach.

It includes:

Access Control

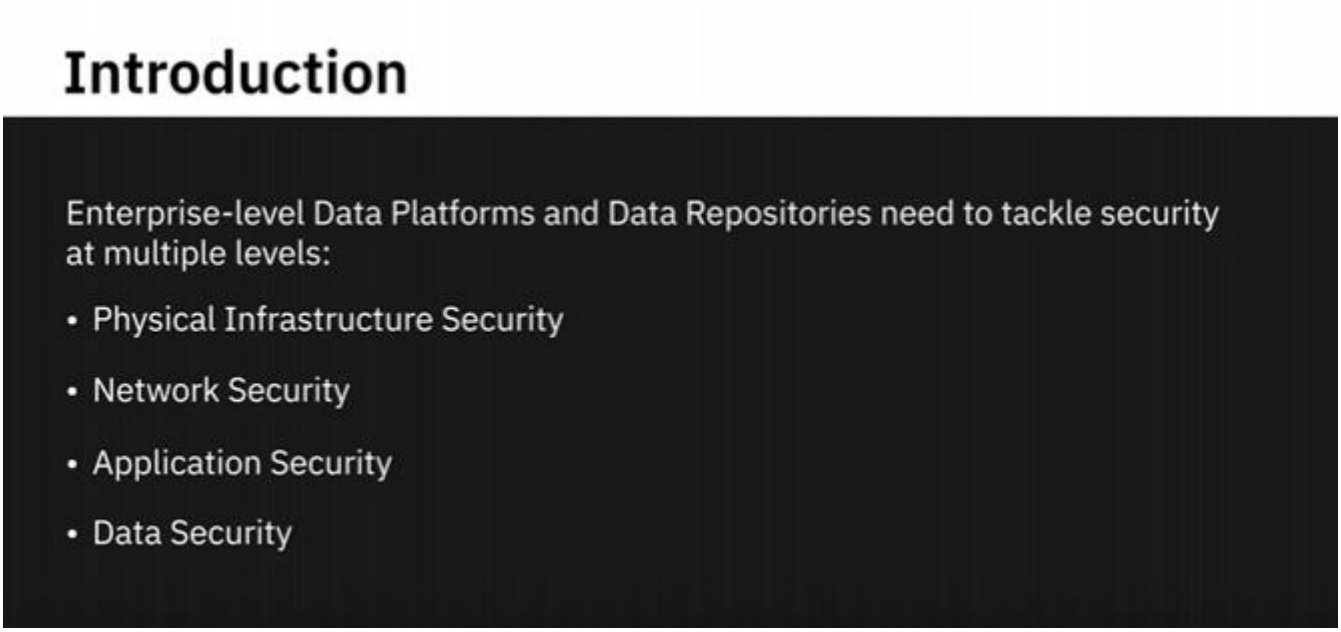
Multizone Encryption



Data Management

Monitoring Systems

Security



The CIA Triad

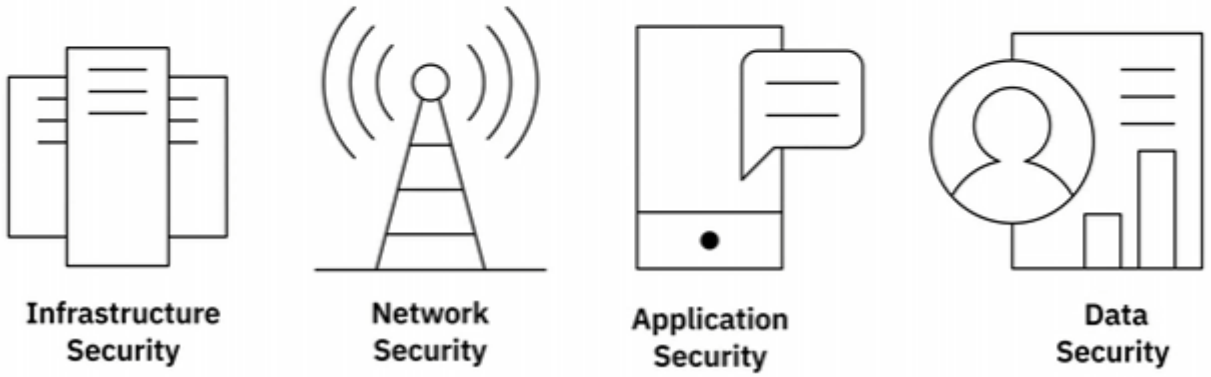
The CIA Triad

Key components to creating an effective strategy for information security include:

- **Confidentiality** through controlling unauthorized access
- **Integrity** through validating that your resources are trustworthy and have not been tampered with
- **Availability** by ensuring authorized users have access to resources when they need it

Application

The CIA Triad



Physical Infrastructure Security

Physical Infrastructure Security

Measures to ensure physical infrastructure security:

- **Access to the perimeter** of the facility based on authentication
- Round the clock **surveillance for entry and exit points** of the facility
- **Multiple power feeds** from independent utility providers with **dedicated generators and UPS battery backup**
- **Heating and cooling mechanisms** for managing the temperature and humidity levels in the facility
- Factoring in **environmental threats** before considering the location of the facility

Network Security

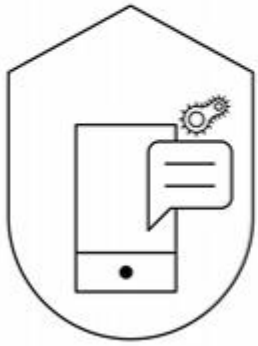
Network Security

Network security is vital to keep interconnected systems and data safe.

- **Firewalls** to prevent unauthorized access to private networks
- **Network Access Control** to ensure endpoint security by allowing only authorized devices to connect to the network
- **Network Segmentation** to create silos, or virtual local area networks, within a network
- **Security Protocols** to ensure attackers cannot tap into data while it is in transit
- **Intrusion Detection** and **Intrusion Prevention** systems to inspect incoming traffic for intrusion attempts and vulnerabilities

Application Security

Application Security



Application Security is critical for keeping customer data private and ensuring applications are fast and responsive.

- **Threat modeling** to identify relative weaknesses and attack patterns related to the application
- **Secure design** that mitigates risks
- **Secure coding** guides and practices that prevent vulnerabilities
- **Security testing** to fix problems before the application is deployed and to validate that it is free of known security issues

Data Security

Data Security

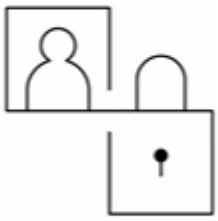
Data is either at rest in storage, or in transit, between systems, applications, services, and workloads.

Authentication

Data Security

Authentication systems verify you are who you say you are.

Passwords



Tokens

Biometrics

Authorization ensures users access information based on their role and privileges.

Data at Rest and Data in Transit

Data Security

Data at rest:

- Includes files, objects, and storage
- Stored physically in a database, data warehouse, tapes, offsite backups, and mobile devices
- Can be protected by encryption

Data in transit:

- Moving from one location to another over the internet
- Can be protected using encryption methods such as HTTPS, SSL, and TLS

Monitoring and Intelligence

Viewpoints: Importance of Data Security

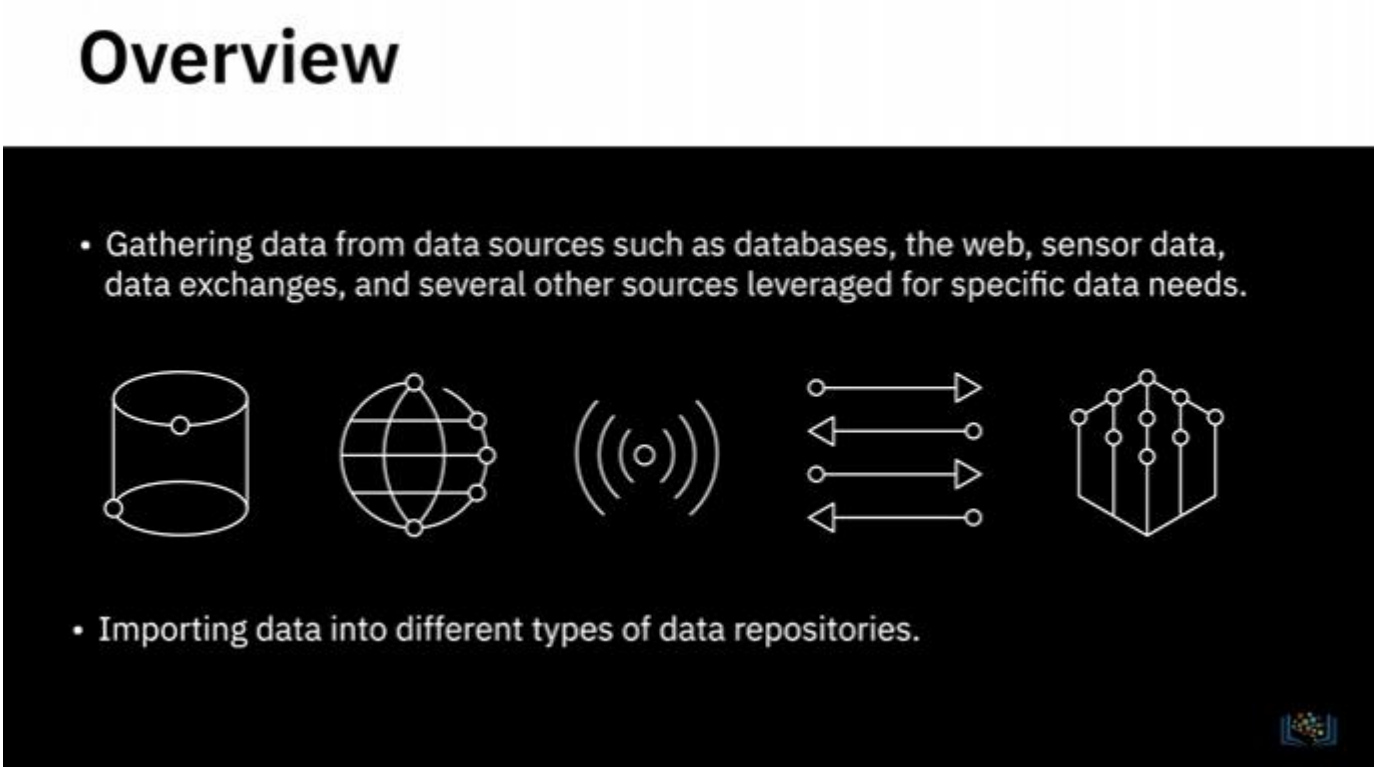
In this video, we will listen
to several data professionals talk about the importance of
data security as it relates to data engineering. So if your data is not secure,
nothing else about it really is going to matter in the long run.
There are many companies who are examples of serious data
breaches, and while that won't necessarily put you out of
business these days, that depends on the size of your
organization somewhat to begin with. If you're a smaller
organization and you lose access to your own data, that's
certainly a potentially company-ending problem. You really
can't understate the importance of data security. It's often
something we think of, oh we'll, harden it after go live or we'll
harden it just before go live. But really it's something that
you have to be thinking about data security along every step
of the way. The importance of security of data can't be stressed enough. A few years
ago, The Economist magazine published an article saying the
world's most valuable resource is no longer oil, but data. So if data is the most valuable
resource an organization has, it needs to be secured and protected. And failure to do so
can have catastrophic consequences. So data security, governance, compliance, these

are not just things that a data engineer needs to worry about. This needs to be an important part of not just the data architecture, but the entire organizational processes and every part of the organization needs to keep on top of these things. We have to make sure that when we have separate production and non-production environments that our production data doesn't get unmasked into a lower environment. There are absolutely data breaches that can happen that way. We have to think about making sure that people who have access to production data really need that access to production data. The important thing is to understand the security levels and roles for the tool you're using, and make sure that each user is getting the least privilege they need in order to do their jobs. It's when we give people more access that we really open things up to data breaches and data security problems. Most of the threats to your data are actually more likely to come from within your organization than without. It may be exciting, or whatever, to think of somebody coming from outside of your organization to get to your data, but most often that threat is going to come from inside your organization. A lot of the things that we hear about in the data profession are things like people at prescription drug places accessing the prescriptions of celebrities. There are things along those lines that we have to be very careful in protecting the data. Data is a natural resource these days, right? The world of Big Data, data is oil, that's our new natural resource. So it's our job to make sure that that financial aspect of valuable data is protected and that our organization is not at risk for being in the news as one of those causing a data breach or even worse, being put out of business because of a data breach. I consider part of data security, the ability to restore my data and the ability to recover if a particular set of hardware fails. To me, that's part of security, because if we lose the data because of a hardware failure, that means we don't have access to the data,

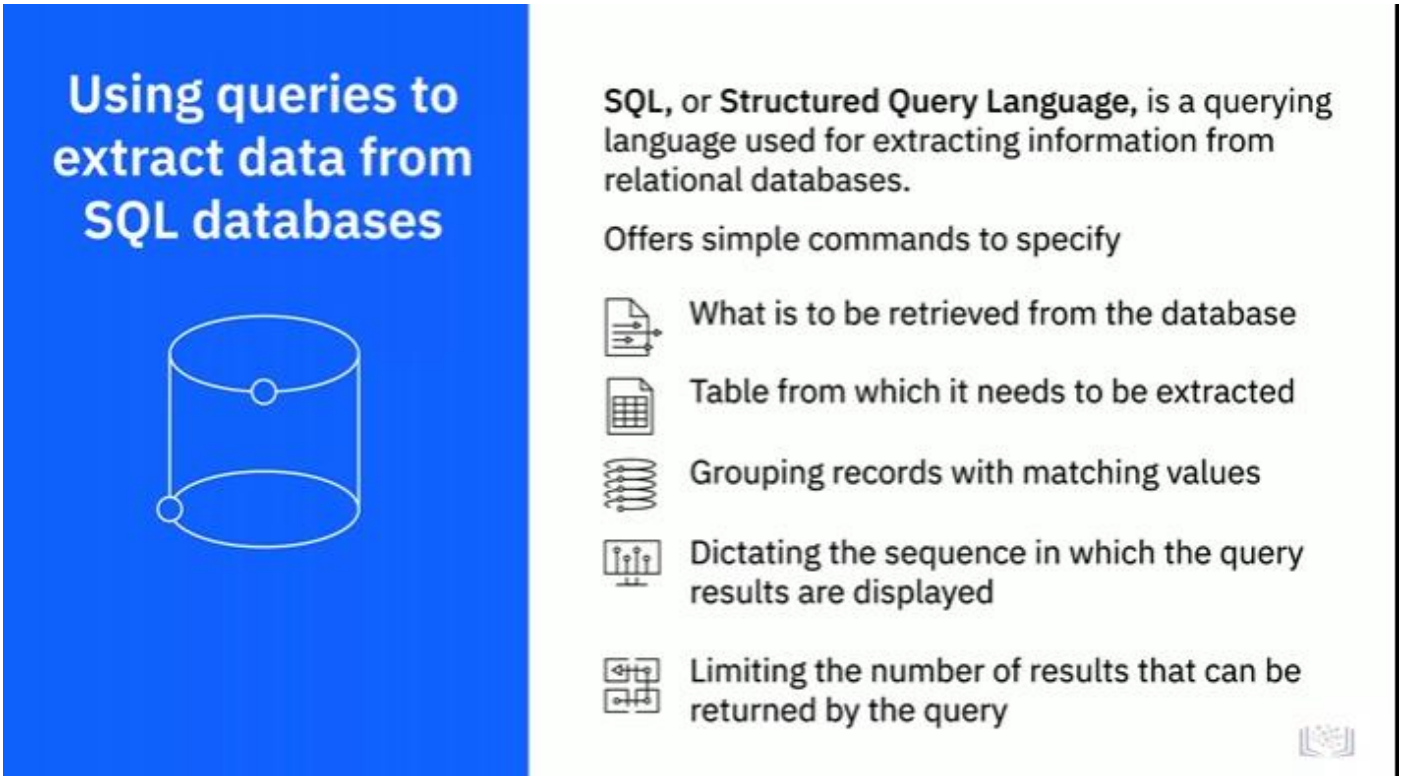
which is part of data security in my own estimation.

Data Collection and Data Wrangling

How to Gather and Import Data

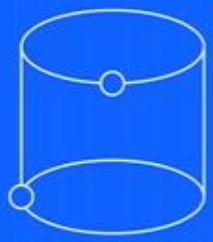


Using queries to extract data from SQL databases



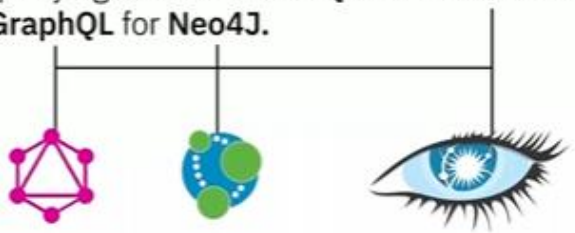
Nosql

Using queries to extract data from SQL databases



Non-relational databases can be queried using **SQL** or **SQL**-like query tools.

Some non-relational databases come with their own querying tools such as **CQL** for **Cassandra** and **GraphQL** for **Neo4J.**



APIs

APIs

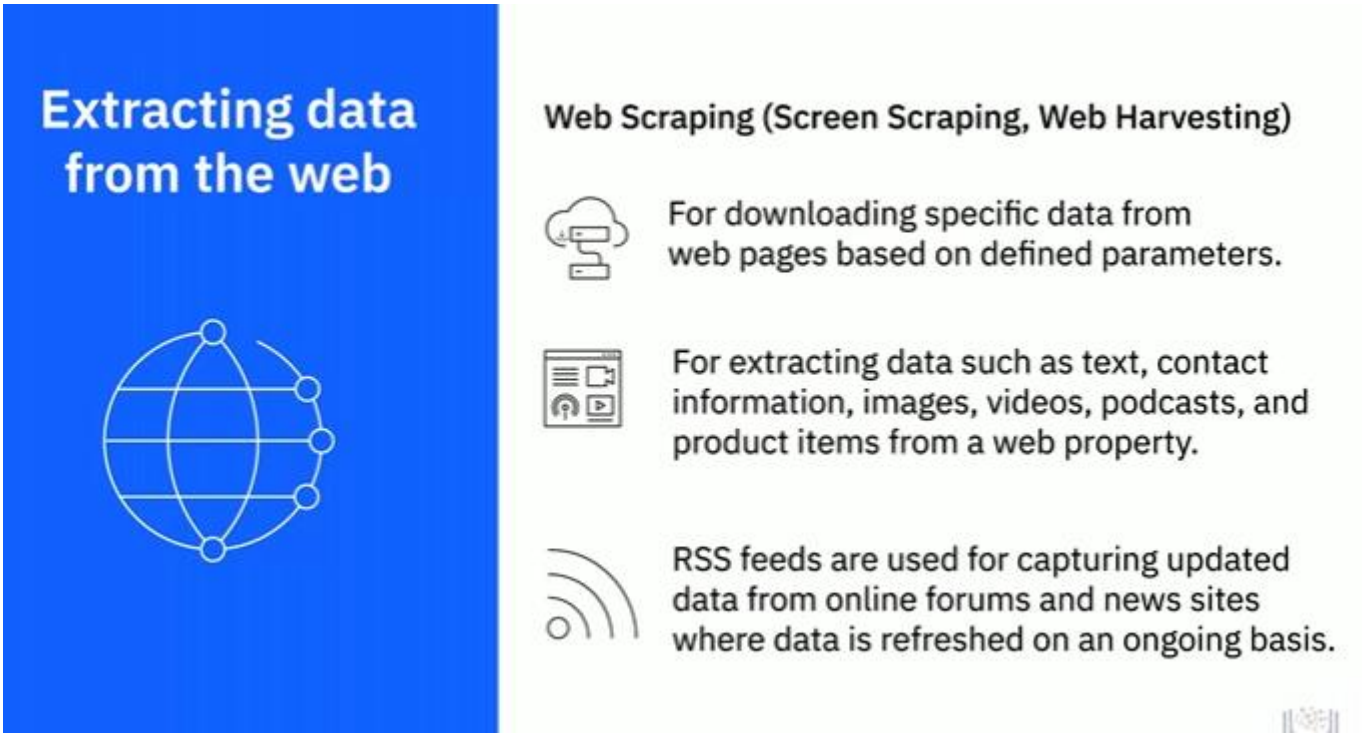
Application Programming Interfaces (or APIs)

Popularly used for extracting data from a variety of data sources.

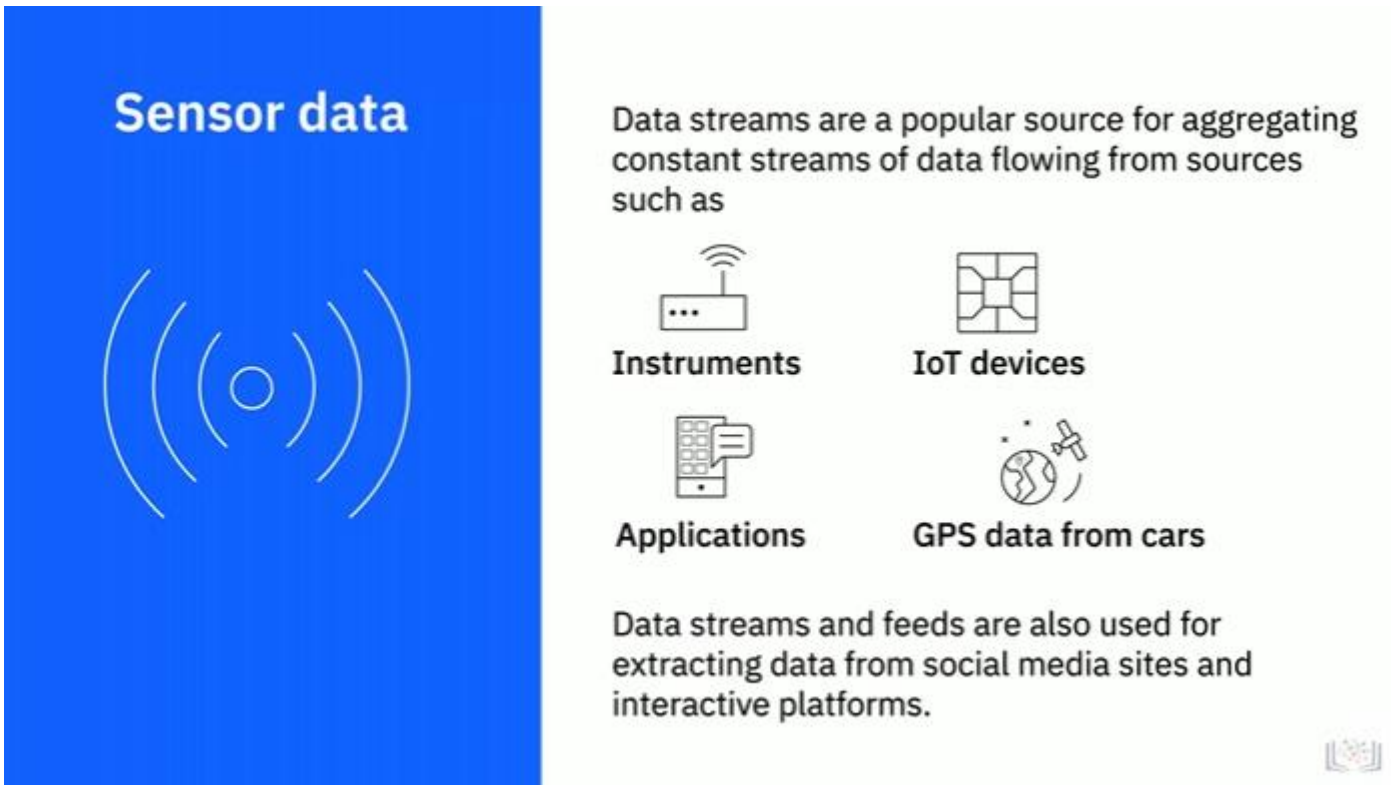
Are invoked from applications that require the data and access an endpoint containing the data. Endpoints can include databases, web services, and data marketplaces.

Also used for data validation.

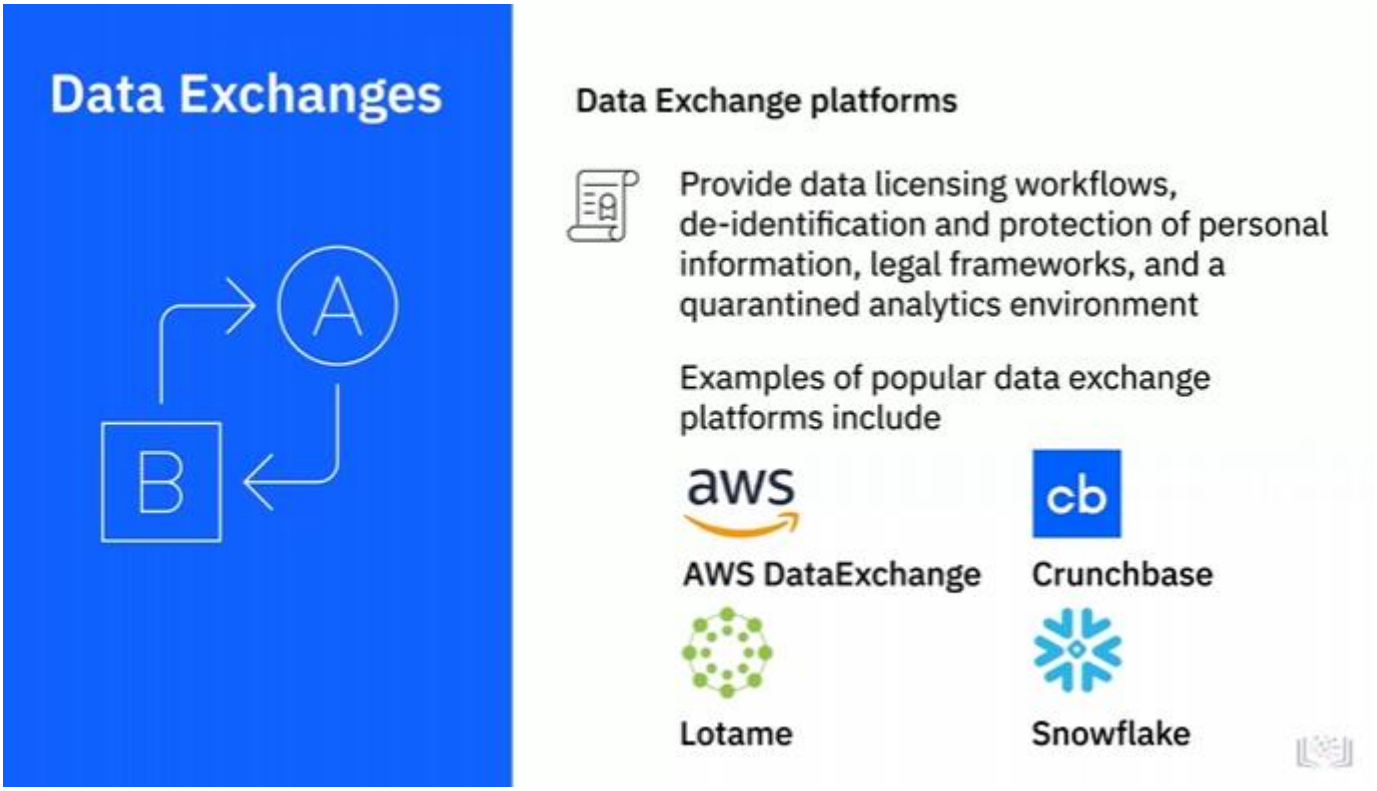
Extracting data from the Web



Sensor Data



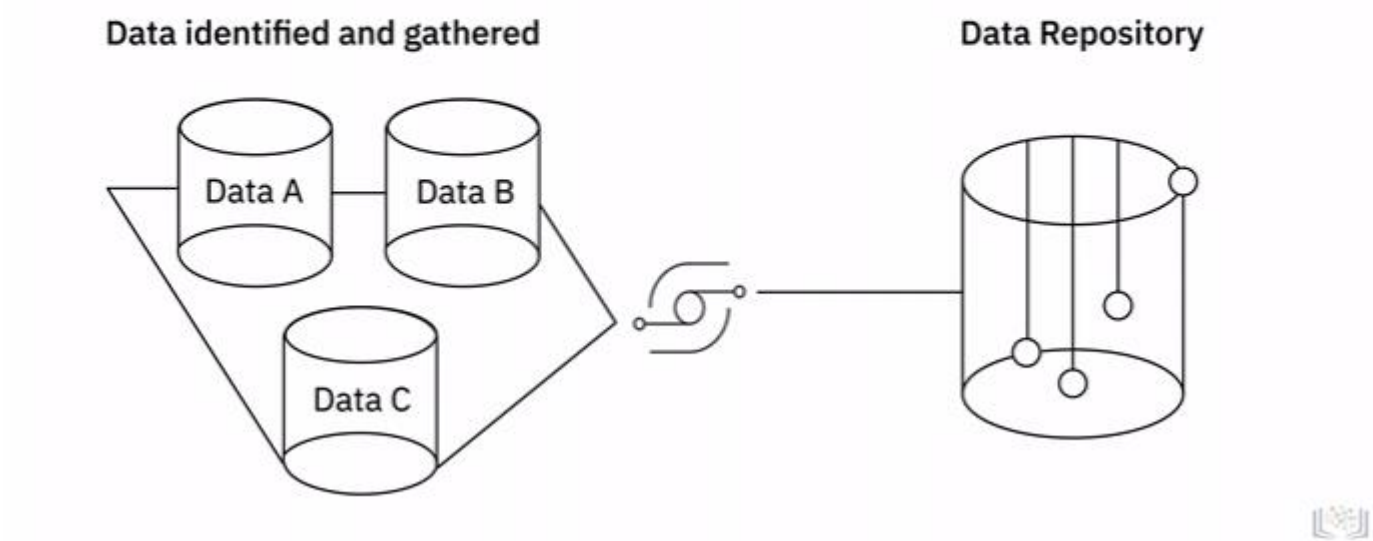
Data Exchanges



Other Sources

Importing Data

Importing data



Combined view or single interface

Data type and destination repositories

Dstructured

Data types and destination repositories

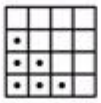
Specific data repositories are optimized for certain types of data.

Structured data

- Relational databases store structured data with a well-defined schema
- Sources include data from OLTP systems, spreadsheets, online forms, sensors, network and web logs
- Can also be stored in NoSQL databases

Data types and destination repositories

Specific data repositories are optimized for certain types of data.



Semi-structured data

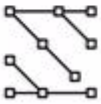
Sources include emails, XML, zipped files, binary executables, and TCP/IP protocols

Can be stored in NoSQL clusters

XML and JSON are commonly used for storing and exchanging semi-structured data

Data types and destination repositories

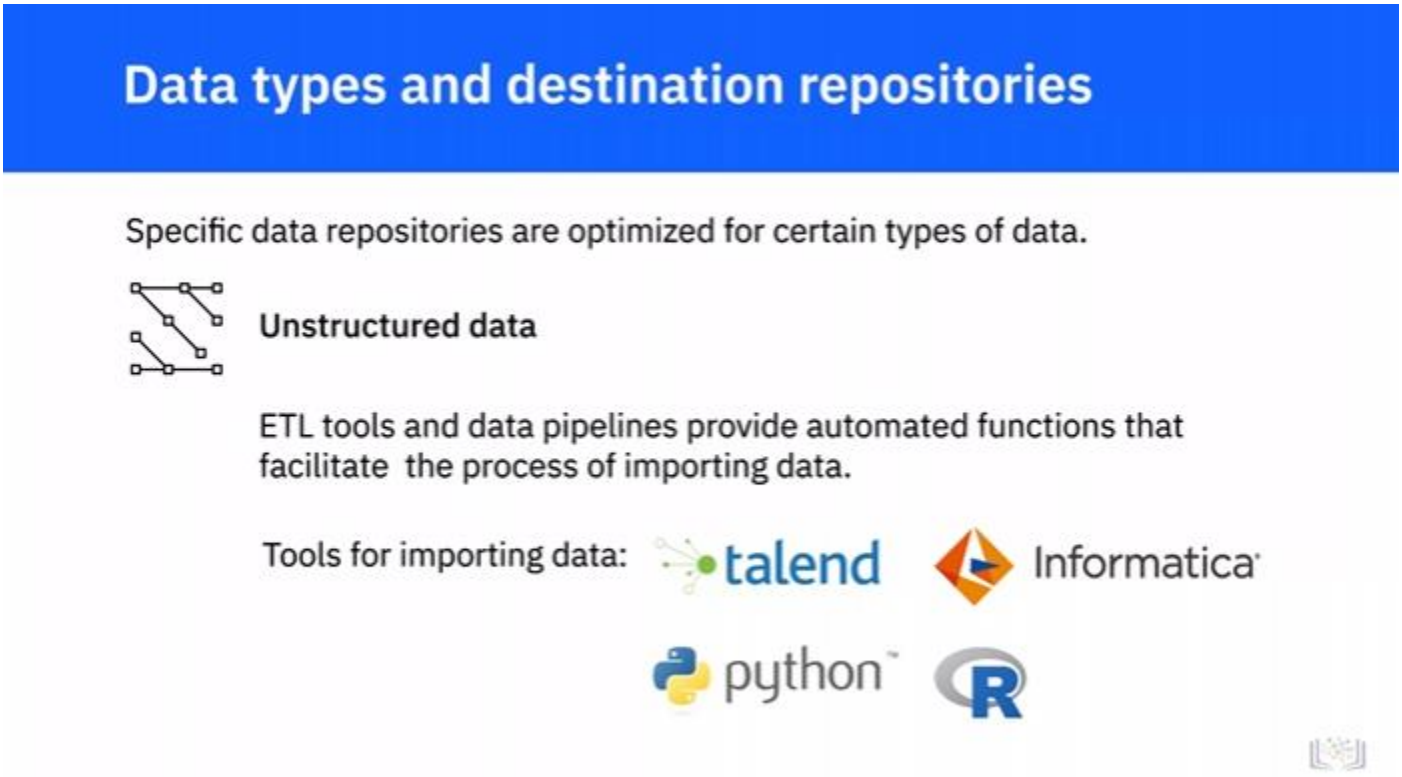
Specific data repositories are optimized for certain types of data.



Unstructured data

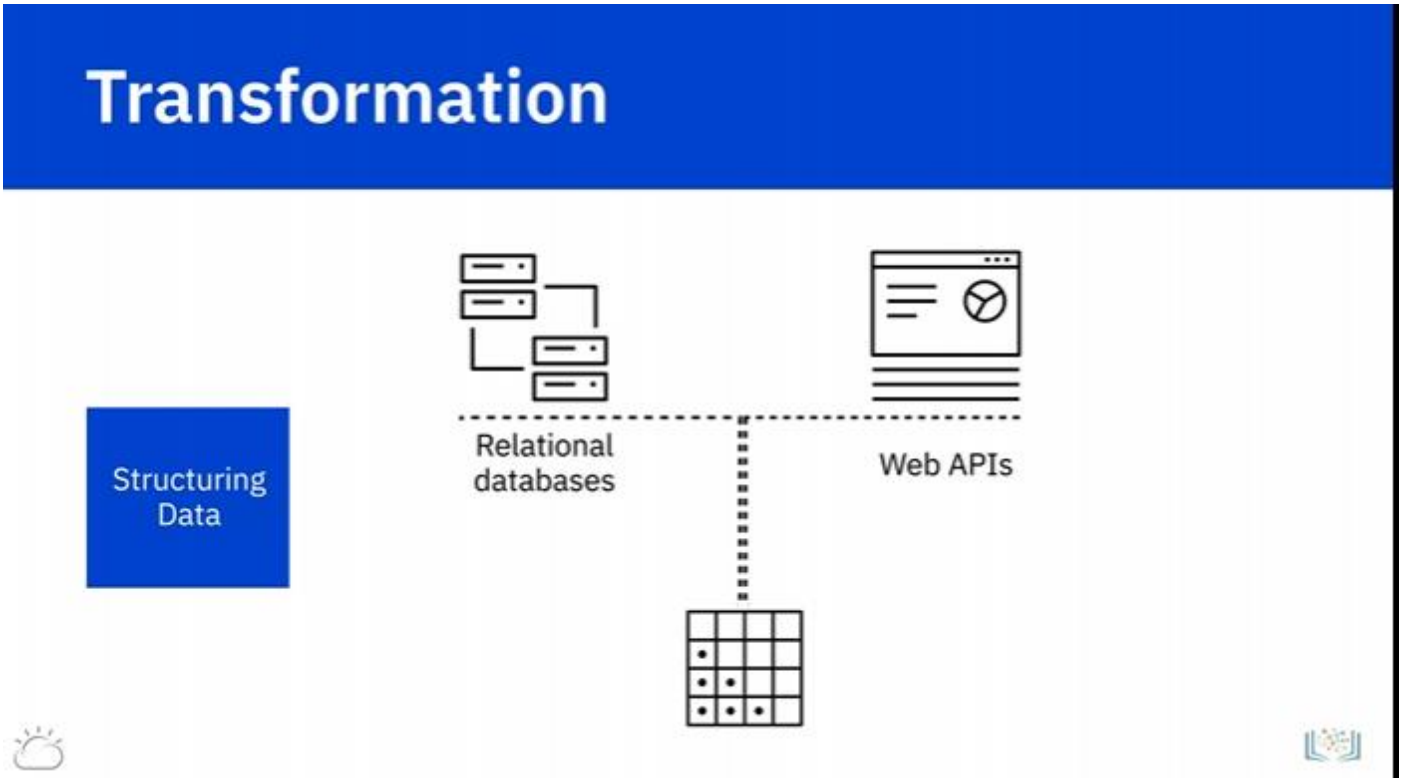
Sources include web pages, social media feeds, images, videos, documents, media logs, and surveys

Can be stored in NoSQL databases and data lakes

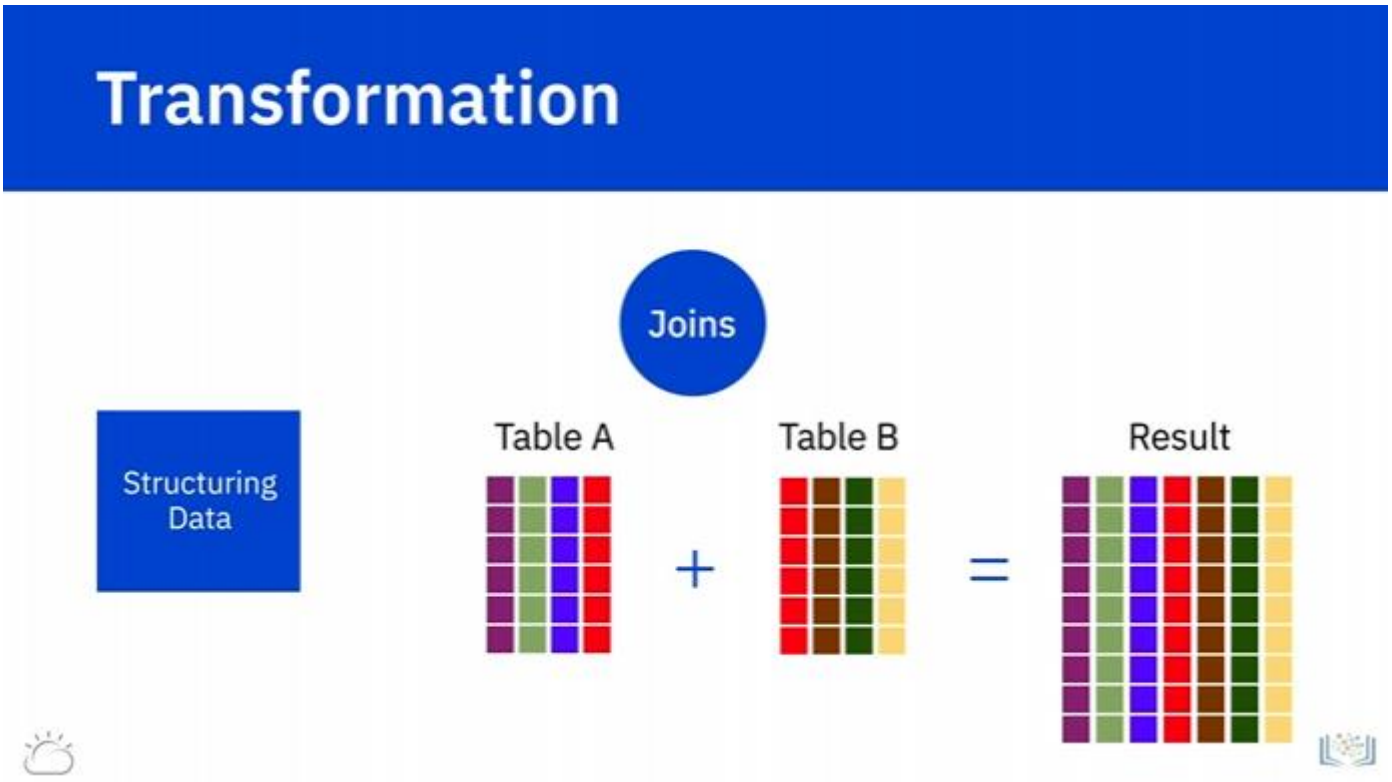


Data Wrangling

Transformation

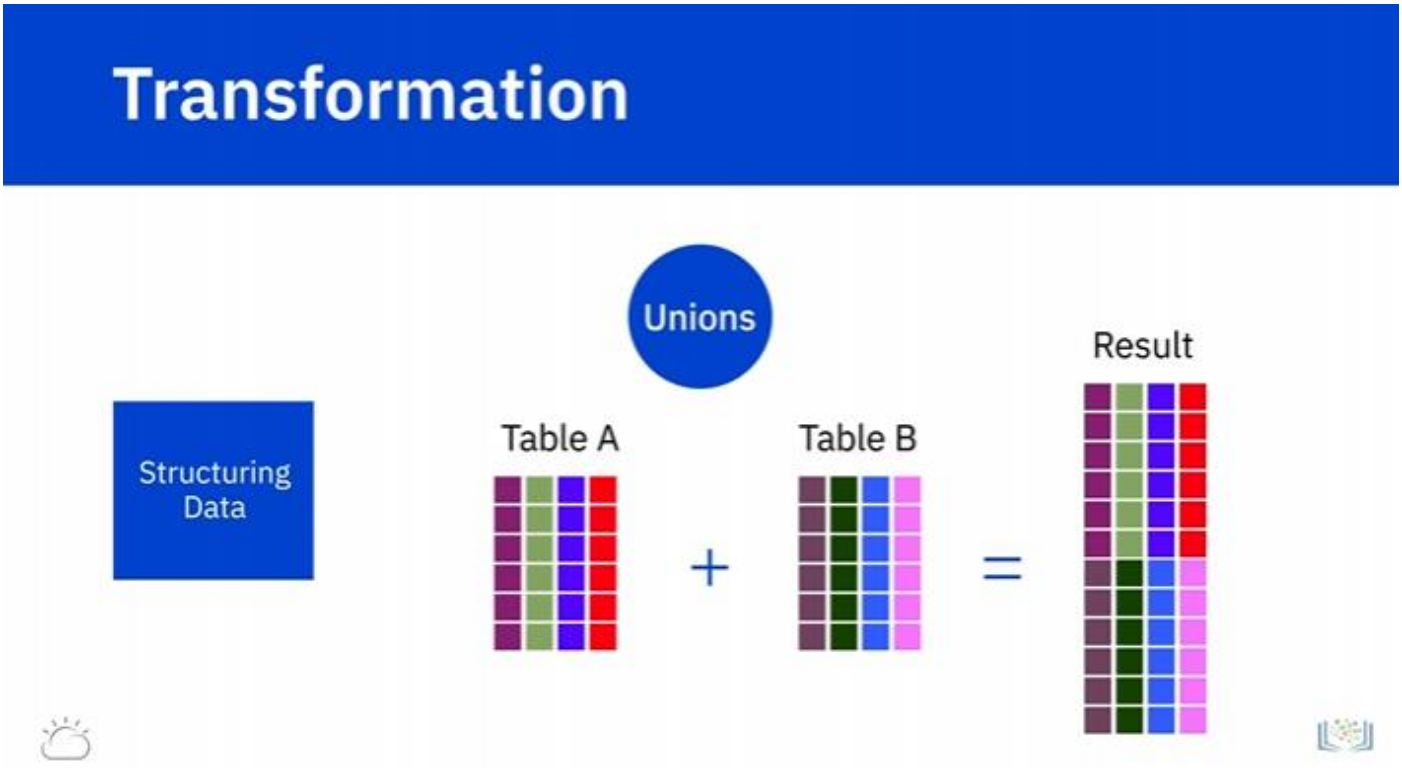


Joins and Unions



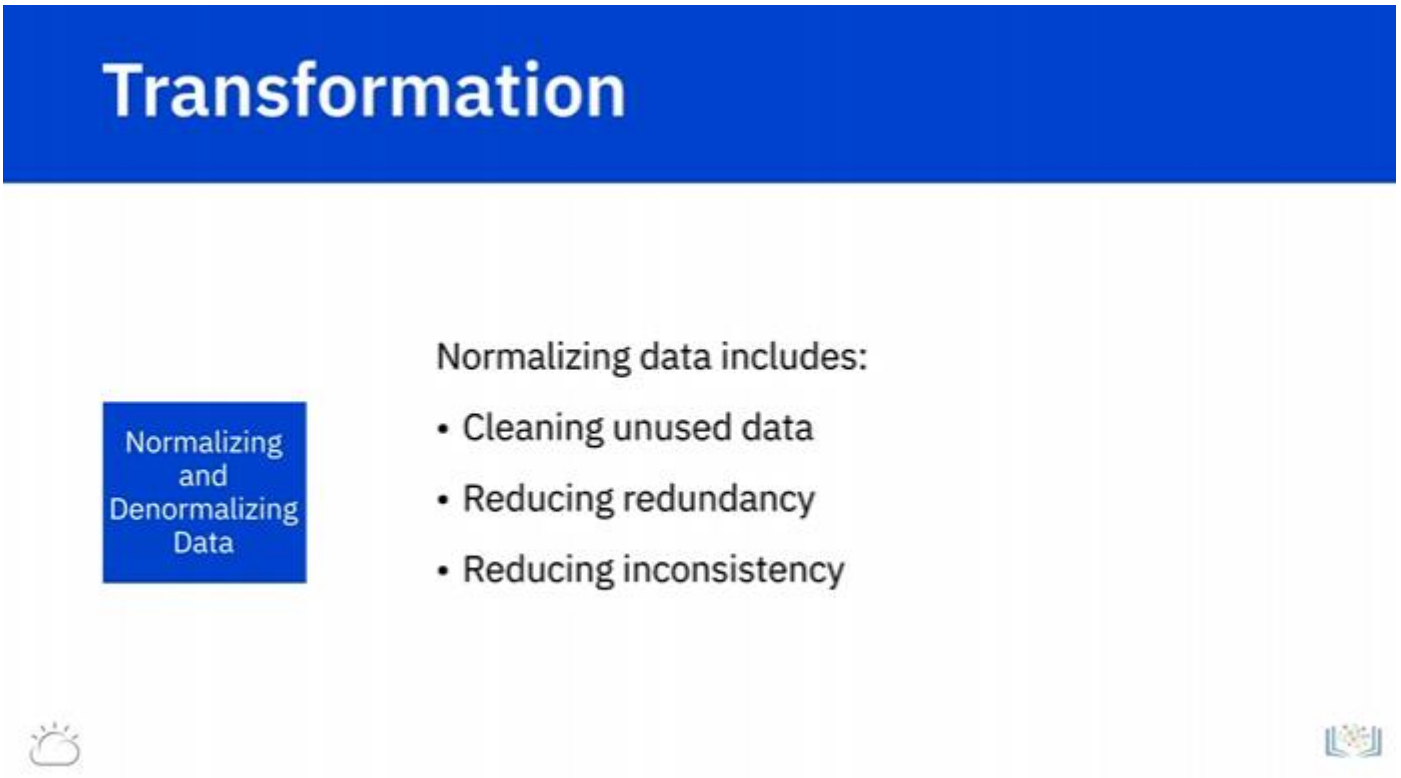
When two tables are joined together, columns from the first source table are combined with columns from the second source table—in the same row.
So, each row in the resultant table contains columns from both tables.

Unions



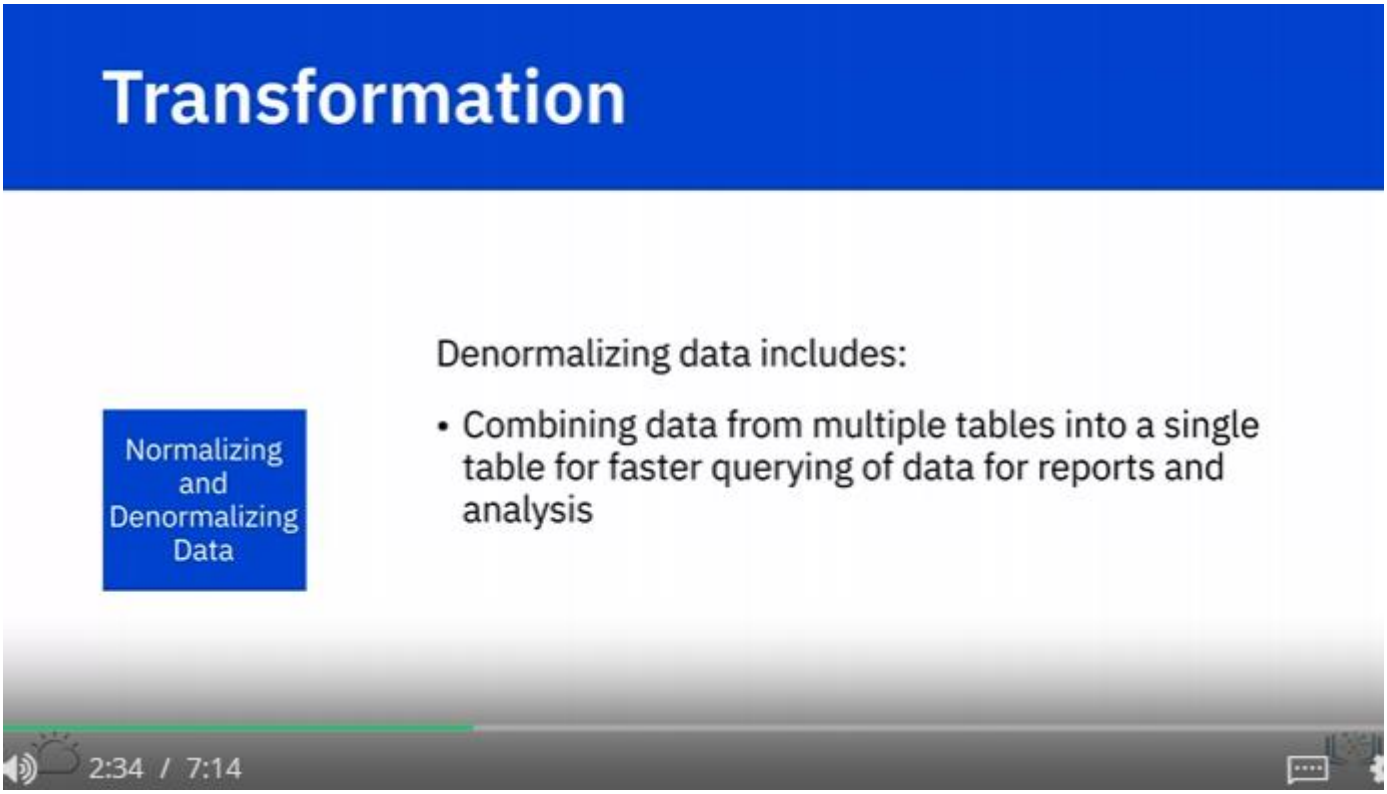
Unions combine rows.
Rows of data from the first source table are combined with rows of data from the second source table into a single table.
Each row in the resultant table is from one source table or another.

Normalisation and Denormalization



Transformation can also include normalization and denormalization of data.

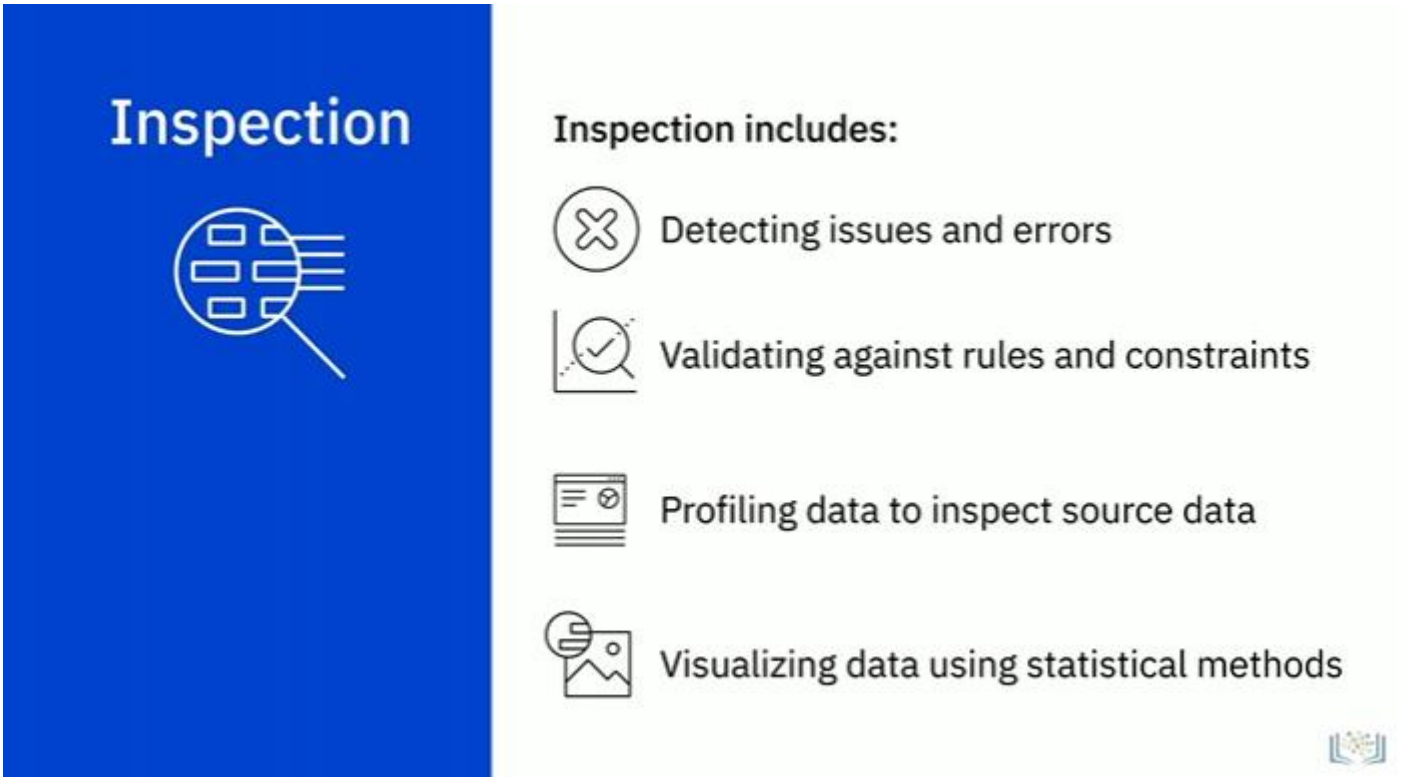
Normalization focuses on cleaning the database of unused data and reducing redundancy and inconsistency.
Data coming from transactional systems, for example, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized.



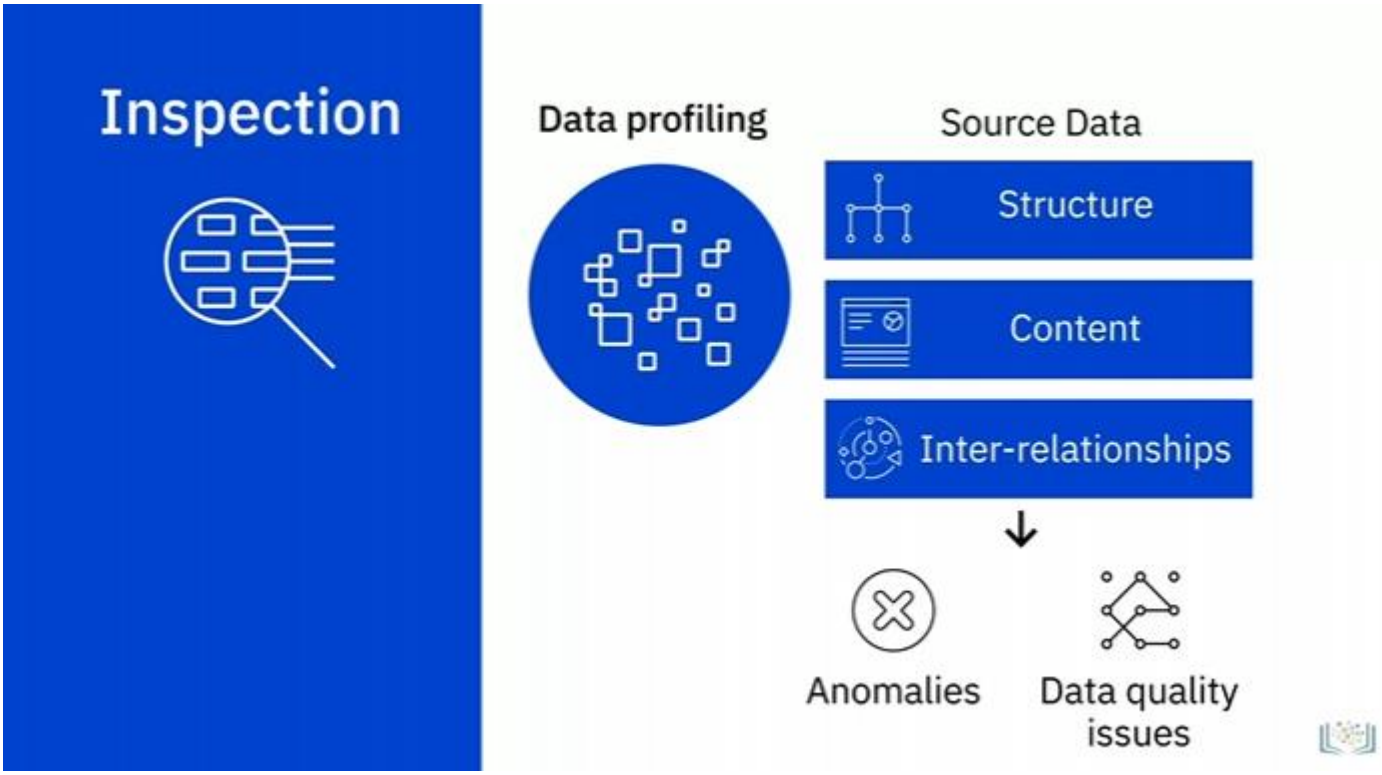
Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster.
For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis.

Cleaning

Data cleaning workflow step 1: Inspection

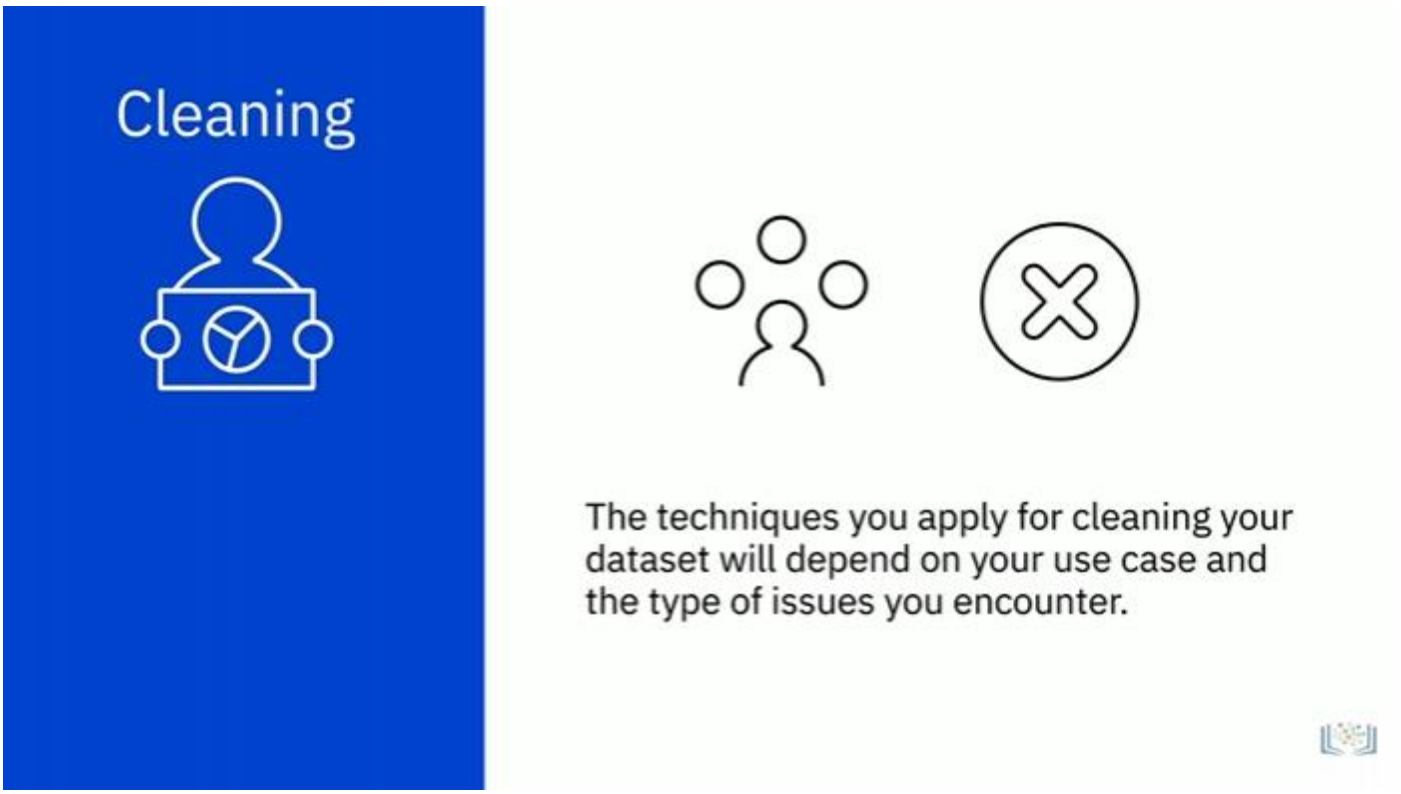


Data profiling



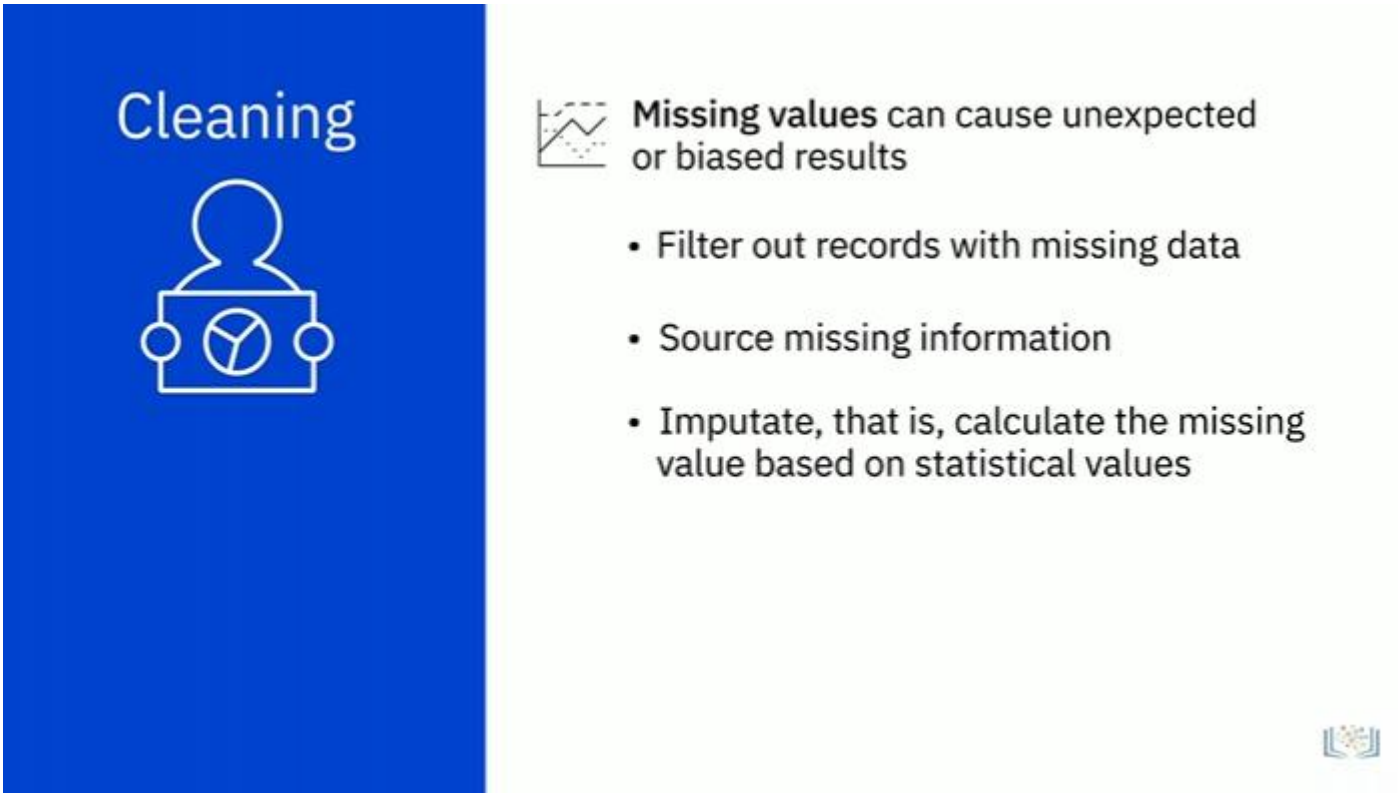
Visualizing the data

Cleaning of Data

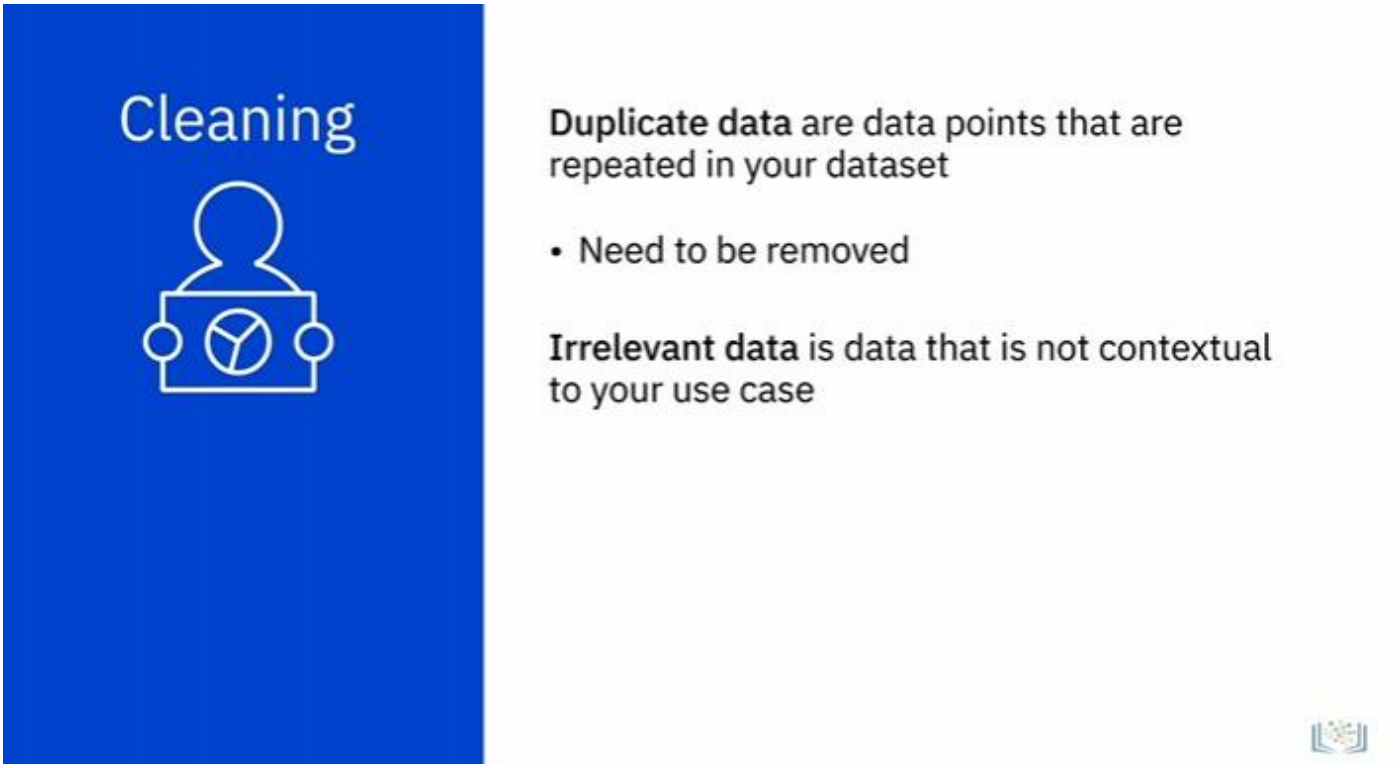


Common Issues

Dealing with Missing Values



Duplicate Data



Data Type Conversion

Cleaning

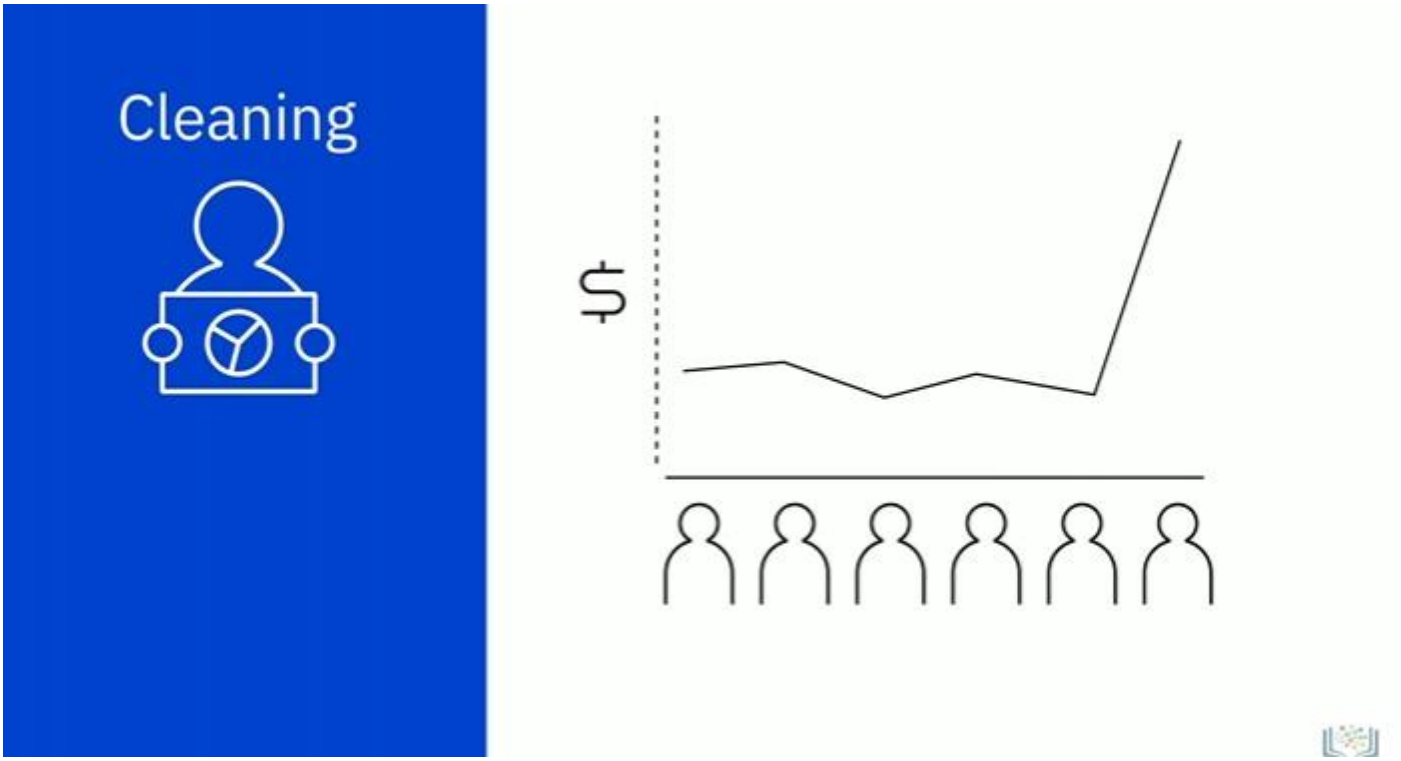
Data type conversion is needed to ensure that values in a field are stored as the data type of that field

Standardizing data is needed to ensure date-time formats and units of measurement are standard across the dataset

Cleaning

Syntax errors, such as white spaces, extra spaces, typos, and formats need to be fixed

Outliner



Tools for Data Wrangling

Excel Power Query/ Spreadsheet

Excel Power Query / Spreadsheets

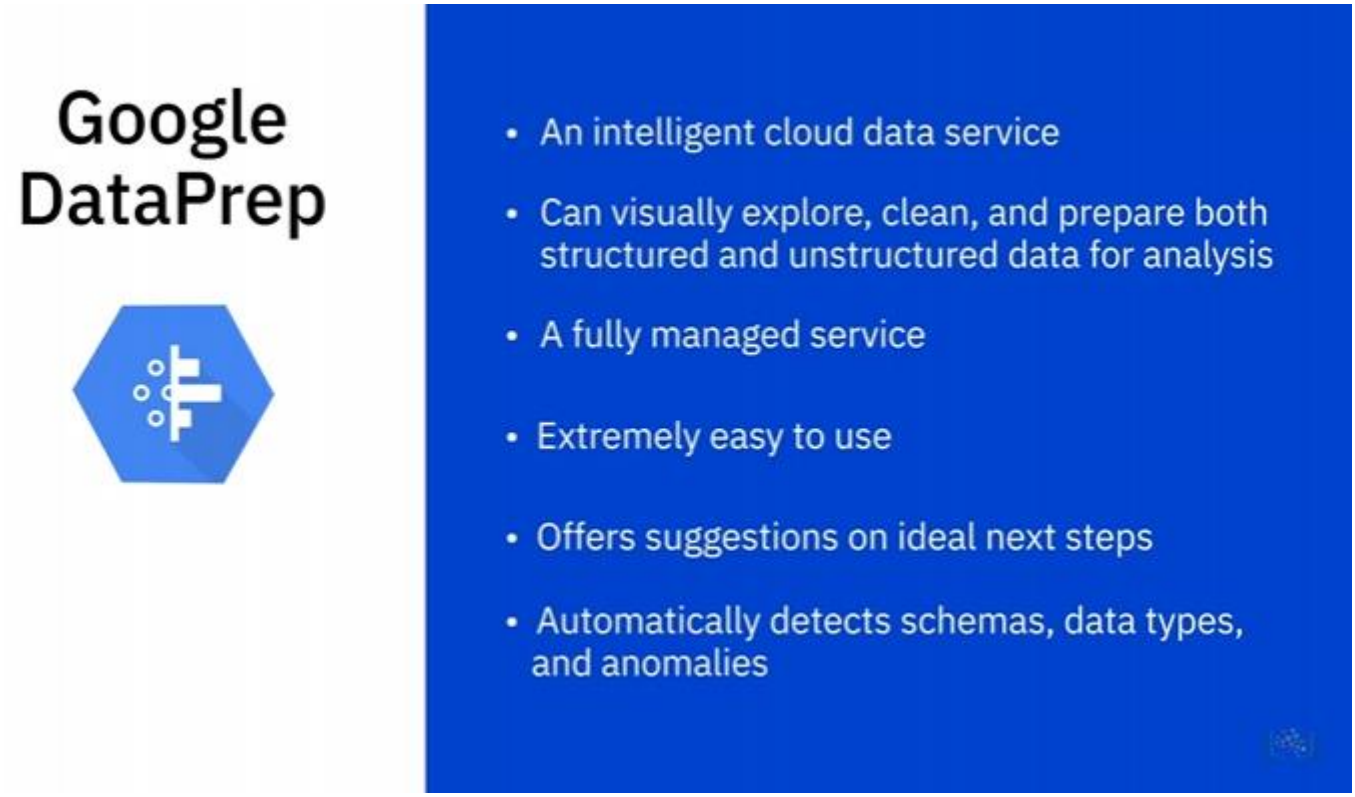
- Spreadsheets such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data.
- Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets are add-ins that allow importing data from different sources and cleaning and transforming data as needed.

OpenRefine

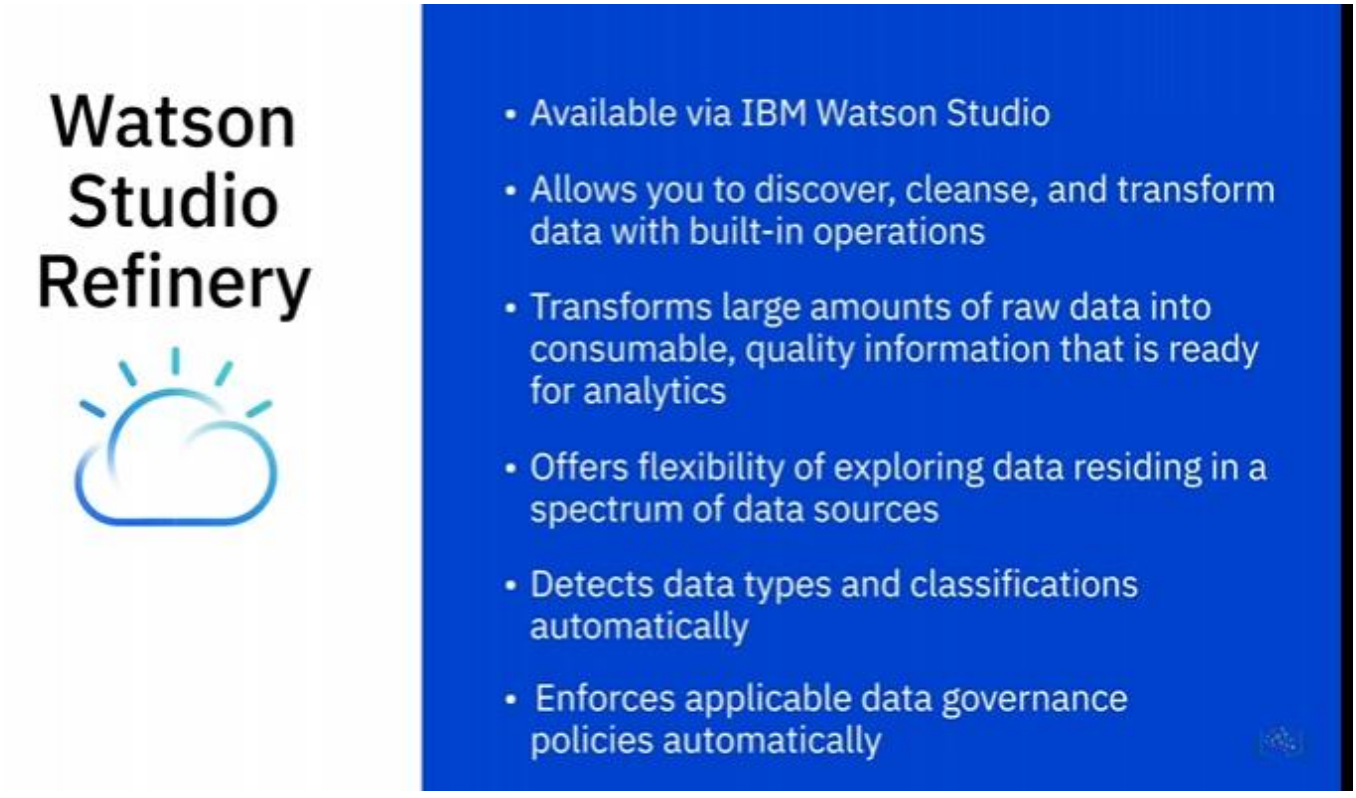
OpenRefine

- Open-source tool
- Can import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON
- Can clean data, transform it from one format to another, and extend data with web services and external data
- Easy to learn
- Easy to use
- Offers menu-based operations

Google DataPrep



Watson Studio Refinery

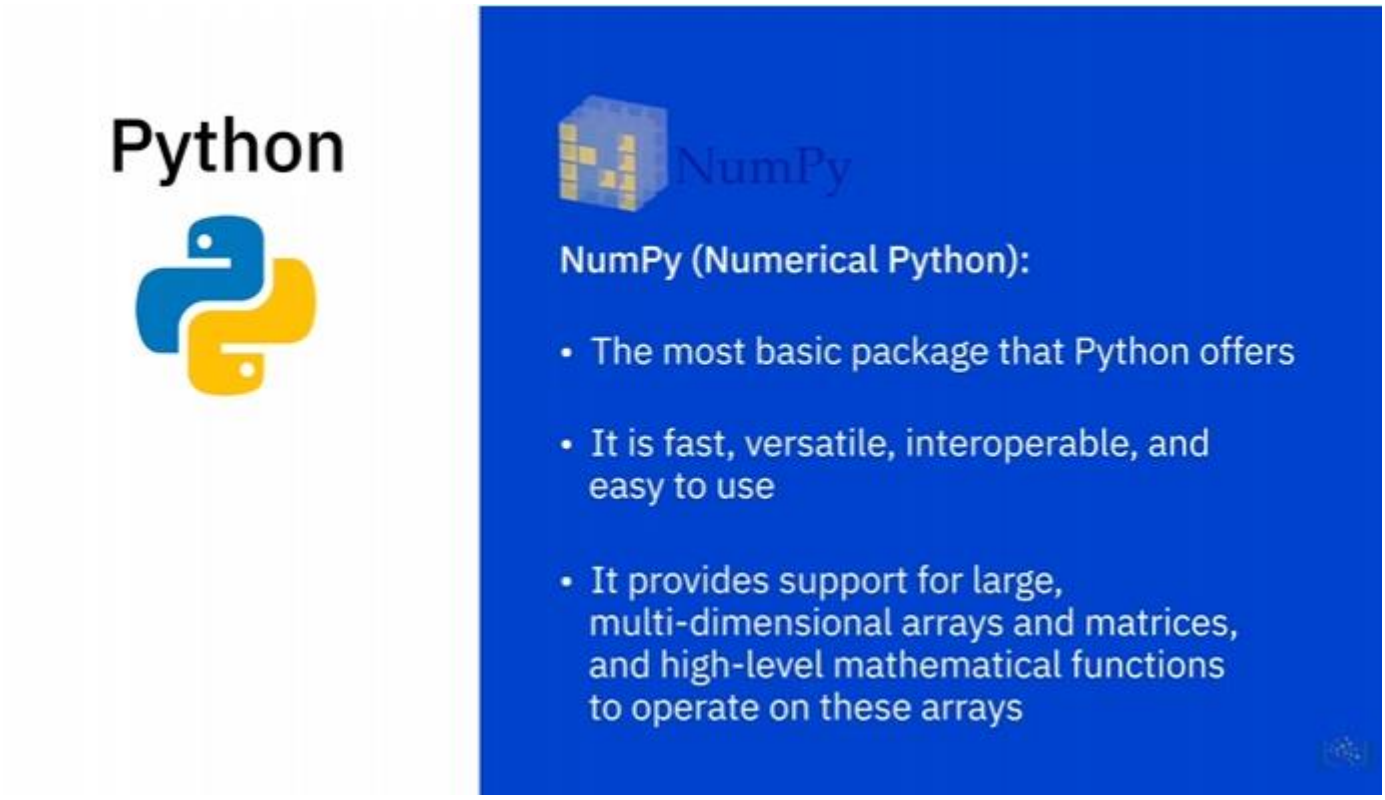


Trifacta Wrangler

Python

Jupyter Notebook

Numpy


Python
NumPy
NumPy (Numerical Python):
• The most basic package that Python offers
• It is fast, versatile, interoperable, and easy to use
• It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays

Pandas

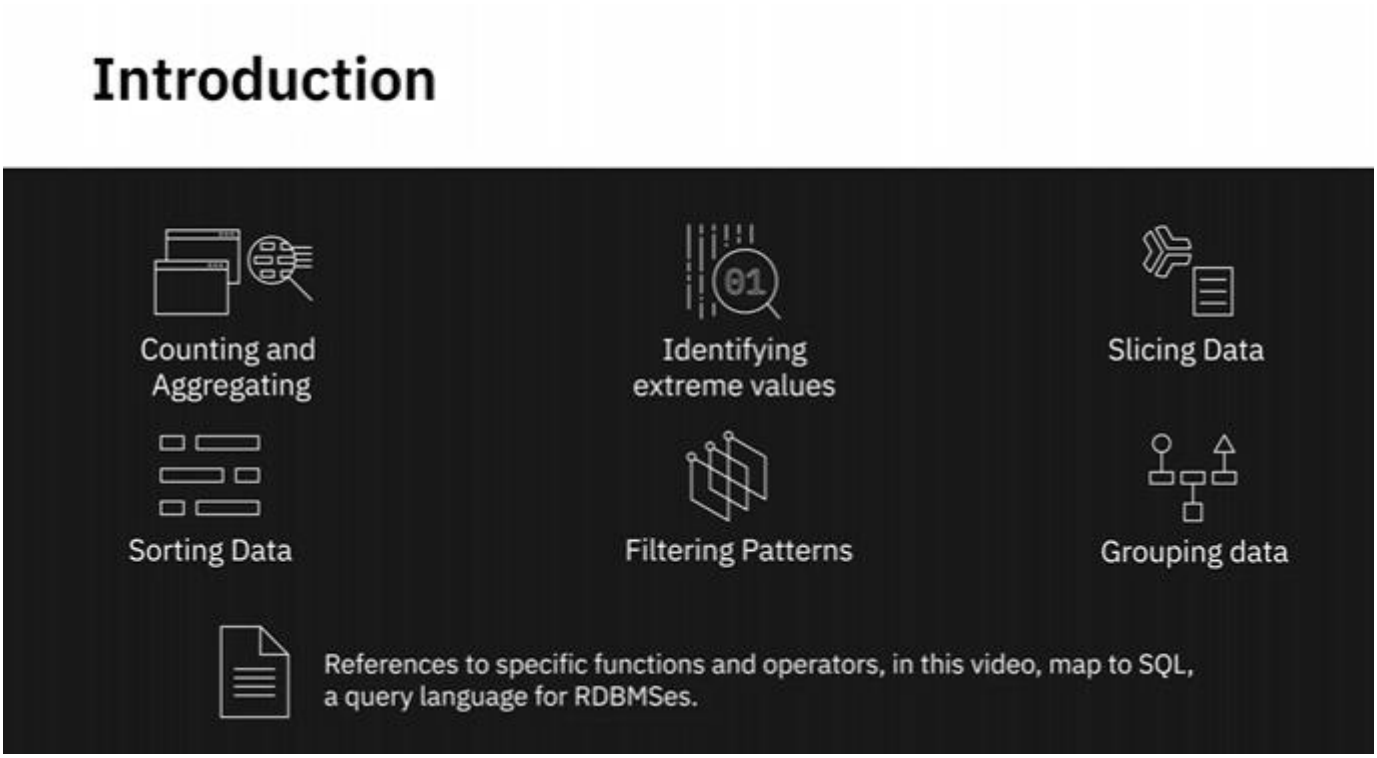

Python
pandas
Pandas:
• Designed for fast and easy data analysis operations
• Allows complex operations such as merging, joining, and transforming huge chunks of data using simple, single-line commands
• Helps prevent common errors that result from misaligned data coming in from different sources

R

Tools for Data Wrangling

Querying Data, performing tunning and troubleshooting

Querying and Analyzing Data



Counting

Example

Counting

Counting the number of rows of data, or records, in the data set.

Customer Name	State	City	Pin	Date of Purchase	Car Model	Sale Amount	Dealer
Allen Perl	NM	Albuquerque	87109	15-Jan-2020	Mitsubishi	1,000.00	Car Nation
Anthony Whitney	NM	Albuquerque	87109	3-Jan-2020	Dodge	1,200.00	Sonic Auto
Thomas Owens	TX	Tulia	79088	25-Dec-2019	Oldsmobile	1,500.00	Thilia Motors
Melvin Schmitz	NM	Albuquerque	87109	20-Dec-2019	Porsche	1,150.00	Car Chambers
Muriel Exley	CA	North Valley	87107	7-Jan-2020	323i	1,850.00	Car Nation
Alfonso Frazier	NM	Santa Fe	87501	15-Jun-2020	328i	1,650.00	Car Nation
Cecil Games	TX	Springlake	79082	12-Dec-2019	528i	1,525.00	Larry Car Dealers
Edward Turner	TX	Amarillo	79109	25-Jul-2020	Century	1,500.00	Quality Kings
Amy Randle	TX	Amarillo	79109	30-May-2020	Regal	1,800.00	Fair Car Dealers
Rafael Middleton	TX	Claude	79019	9-Sep-2020	Park Avenue	1,900.00	Fair Car Dealers
Linda Garcia	IN	Fountain	80817	19-Dec-2019	LeSabre	1,550.00	The Dealers
Quinn Perry	IN	Fountain	80817	27-Feb-2019	DeVille	1,750.00	The Dealers
Phyllis White	OC	Anaheim	92805	25-Dec-2019	Seville	25,500.00	Phyllis Auto
Lisa Guest	CA	Antioch	60002	3-Jan-2019	Eldorado	37,000.00	Melvin Motors

Sample data set includes data for customer who have purchased used cars.

Counting

```
mysql> select distinct Dealer from CarSaleDetails;
+-------------------+
| Dealer            |
+-------------------+
| Car Nation        |
| Sonic Auto        |
| Thilia Motors     |
| Car Chambers      |
| Larry Car Dealers |
| Quality Kings     |
| Fair Car Dealers  |
| The Dealers       |
| Phyllis Auto      |
| Melvin Motors     |
+-------------------+
```

Displaying the unique car dealers in the data set.

Counting

```
mysql> select count(distinct Dealer) from CarSaleDetails;
+-------------------------+
| count(distinct Dealer) |
+-------------------------+
|                       8 |
+-------------------------+
```

Counting the total number of unique, or distinct, car dealers.

SELECT COUNT(DISTINCT Dealer) from CarSaleDetails;

Aggregation

Aggregation

Aggregation functions help to provide an overview of the data set from different perspectives.

```
mysql> select sum(Sale_Amount) as sum_all from CarSaleDetails;
+--------------+
|   sum_all    |
+--------------+
|        80875 |
+--------------+
```

Calculating the sum of a numeric column.

```
mysql> select avg(Sale_Amount) as avg_all from CarSaleDetails;
+--------------+
|   avg_all    |
+--------------+
|      5776.79 |
+--------------+
```

Calculating the average value of a numeric column.

Aggregation

```
mysql> select stddev(Sale_Amount) as std_all from CarSaleDetails;
+----------------+
|   std_all      |
+----------------+
|   11028.28734  |
+----------------+
```

Calculating the standard deviation to see how spread out the cost of a used car is.

Aggregation

Customer Name	State	City	Pin	Date of Purchase	Car Model	Sale Amount	Dealer
Allen Perl	NM	Albuquerque	87109	15-Jan-2020	Mitsubishi	1,000.00	Car Nation
Anthony Whitney	NM	Albuquerque	87109	3-Jan-2020	Dodge	1,200.00	Sonic Auto
Thomas Owens	TX	Tulia	79088	25-Dec-2019	Oldsmobile	1,500.00	Thilia Motors
Melvin Schmitz	NM	Albuquerque	87109	20-Dec-2019	Porsche	1,150.00	Car Chambers
Muriel Exley	CA	North Valley	87107	7-Jan-2020	323i	1,850.00	Car Nation
Alfonso Frazier	NM	Santa Fe	87501	15-Jun-2020	328i	1,650.00	Car Nation
Cecil Games	TX	Springlake	79082	12-Dec-2019	528i	1,525.00	Larry Car Dealers
Edward Turner	TX	Amarillo	79109	25-Jul-2020	Century	1,500.00	Quality Kings
Amy Randle	TX	Amarillo	79109	30-May-2020	Regal	1,800.00	Fair Car Dealers
Rafael Middleton	TX	Claude	79019	9-Sep-2020	Park Avenue	1,900.00	Fair Car Dealers
Linda Garcia	IN	Fountain	80817	19-Dec-2019	LeSabre	1,550.00	The Dealers
Quinn Perry	IN	Fountain	80817	27-Feb-2019	DeVille	1,750.00	The Dealers
Phyllis White	OC	Anaheim	92805	25-Dec-2019	Seville	25,500.00	Phyllis Auto
Lisa Guest	CA	Antioch	60002	3-Jan-2019	Eldorado	37,000.00	Melvin Motors

Calculating the standard deviation to see how spread outthe cost of a used car is.

The average cost for a used car is a little less than USD 6,000.

```
mysql> select stddev(Sale_Amount) as std_all from CarSaleDetails;
+----------------+
|   std_all      |
+----------------+
|   11028.28734  |
+----------------+
```

The standard deviation is over USD 11,000.

Extreme Value Identification

Extreme Value Identification

Identifying extreme values in a data column.

```
mysql> select max(Sale_Amount) as max_price from CarSaleDetails;
+--------------+
|  max_price   |
+--------------+
|        37000 |
+--------------+
```

Calculating the maximum value in a column.

```
mysql> select min(Sale_Amount) as min_price from CarSaleDetails;
+---------------+
|   min_price   |
+---------------+
|          1000 |
+---------------+
```

Calculating the minimum value in a column.

Slicing Data

Slicing Data

Finding customers based on a specific condition or set of conditions.

```
mysql> select Dealer from CarSaleDetails where Dealer City in ("Albuquerque", "Texline");
+---------------+
|    Dealer     |
+---------------+
|   Car Nation  |
|   Sonic Auto  |
| Thilia Motors |
+---------------+
```

Slicing the data set to retrieve data for customers who:

- Live in a certain area
- Have purchased their car from dealers in a specific area
- Have spent between USD 1,000–2,000 for their car
- Have spent between USD 1,000–2,000 for their car and live in a specific area

Sorting Data

Sorting Data

Sorting data helps to arrange data in a meaningful order, making it easier to understand and analyze.

```
mysql> select * from CarSaleDetails order by Date_of_Purchase;
+--------------------+---------------+-------------+--------------------+
|      Dealer        |   Car Model   |  Sale Amount |  Date of Purchase |
+--------------------+---------------+-------------+--------------------+
|   Melvin Motors    |   Eldorado    |  37,000     |    27-Feb-2019     |
+--------------------+---------------+-------------+--------------------+
|    The Dealers     |   DeVille     |  1,750      |    12-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    The Dealers     |   LeSabre     |  1,550      |    19-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    Phyllis Auto    |   Seville     |  25,000     |    20-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
| Larry Car Dealers  |   528i        |  1,525      |    25-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    Car Chambers    |   Porsche     |  1,150      |    25-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    Thilia Motors   |   Oldsmobile  |  1500       |    03-Jan-2020     |
+--------------------+---------------+-------------+--------------------+
|    Sonic Auto      |   Dodge       |  1200       |    07-Jan-2020     |
+--------------------+---------------+-------------+--------------------+
```

Sorting the data set on date of purchase to see if more cars are purchased on festival days.

Filtering Patterns

Filtering Patterns

Filtering patterns to perform partial matches of data values.

Customer Name	State	City	Pin	Date of Purchase	Car Model	Sale Amount	Dealer
Allen Perl	NM	Albuquerque	87109	15-Jan-2020	Mitsubishi	1,000.00	Car Nation
Anthony Whitney	NM	Albuquerque	87109	3-Jan-2020	Dodge	1,200.00	Sonic Auto
Thomas Owens	TX	Tulia	79088	25-Dec-2019	Oldsmobile	1,500.00	Thilia Motors
Melvin Schmitz	NM	Albuquerque	87109	20-Dec-2019	Porsche	1,150.00	Car Chambers
Muriel Exley	CA	North Valley	87107	7-Jan-2020	323i	1,850.00	Car Nation
Alfonso Frazier	NM	Santa Fe	87501	15-Jun-2020	328i	1,650.00	Car Nation
Cecil Games	TX	Springlake	79082	12-Dec-2019	528i	1,525.00	Larry Car Dealers
Edward Turner	TX	Amarillo	79109	25-Jul-2020	Century	1,500.00	Quality Kings
Amy Randle	TX	Amarillo	79109	30-May-2020	Regal	1,800.00	Fair Car Dealers
Rafael Middleton	TX	Claude	79019	9-Sep-2020	Park Avenue	1,900.00	Fair Car Dealers
Linda Garcia	IN	Fountain	80817	19-Dec-2019	LeSabre	1,550.00	The Dealers
Quinn Perry	IN	Fountain	80817	27-Feb-2019	DeVille	1,750.00	The Dealers
Phyllis White	OC	Anaheim	92805	25-Dec-2019	Seville	25,500.00	Phyllis Auto
Lisa Guest	CA	Antioch	60002	3-Jan-2019	Eldorado	37,000.00	Melvin Motors

Equal To Operator returns records in which a data value matches a certain value.

Like Operator helps specify a pattern to return records that match a data value partially.

```
mysql> select * from CarSaleDetails where Pin like "871%";
+--------------------+---------------+-------------+--------------------+
|      Dealer        |   Car Model   |     Pin     |  Date of Purchase  |
+--------------------+---------------+-------------+--------------------+
|   Melvin Motors    |   Eldorado    |    87109    |    27-Feb-2019     |
+--------------------+---------------+-------------+--------------------+
|    The Dealers     |   DeVille     |    87109    |    12-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    The Dealers     |   LeSabre     |    87107    |    19-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
|    Phyllis Auto    |   Seville     |    87132    |    20-Dec-2019     |
+--------------------+---------------+-------------+--------------------+
```

Select * from CarSaleDetails where

Grouping Data

Grouping Data

Grouping data based on a commonality.

Customer Name	State	City	Pin	Date of Purchase	Car Model	Sale Amount	Dealer
Allen Perl	NM	Albuquerque	87109	15-Jan-2020	Mitsubishi	1,000.00	Car Nation
Anthony Whitney	NM	Albuquerque	87109	3-Jan-2020	Dodge	1,200.00	Sonic Auto
Thomas Owens	TX	Tulia	79088	25-Dec-2019	Oldsmobile	1,500.00	Thilia Motors
Melvin Schmitz	NM	Albuquerque	87109	20-Dec-2019	Porsche	1,150.00	Car Chambers
Muriel Exley	CA	North Valley	87107	7-Jan-2020	323i	1,850.00	Car Nation
Alfonso Frazier	NM	Santa Fe	87501	15-Jun-2020	328i	1,650.00	Car Nation
Cecil Games	TX	Springlake	79082	12-Dec-2019	528i	1,525.00	Larry Car Dealers
Edward Turner	TX	Amarillo	79109	25-Jul-2020	Century	1,500.00	Quality Kings
Amy Randle	TX	Amarillo	79109	30-May-2020	Regal	1,800.00	Fair Car Dealers
Rafael Middleton	TX	Claude	79019	9-Sep-2020	Park Avenue	1,900.00	Fair Car Dealers
Linda Garcia	IN	Fountain	80817	19-Dec-2019	LeSabre	1,550.00	The Dealers
Quinn Perry	IN	Fountain	80817	27-Feb-2019	DeVille	1,750.00	The Dealers
Phyllis White	OC	Anaheim	92805	25-Dec-2019	Seville	25,500.00	Phyllis Auto
Lisa Guest	CA	Antioch	60002	3-Jan-2019	Eldorado	37,000.00	Melvin Motors

```
mysql> select sum(Sale_Amount) as area_sum, Pin from CarSaleDetails group by Pin;
+--------------+--------------+
|      Pin     |   area_sum   |
+--------------+--------------+
|     87109    |   3,350.00   |
+--------------+--------------+
|     80817    |   3,300.00   |
+--------------+--------------+
|     79109    |   5,200.00   |
+------------------------------+
```

Total amount spent by customers, pincode-wise.

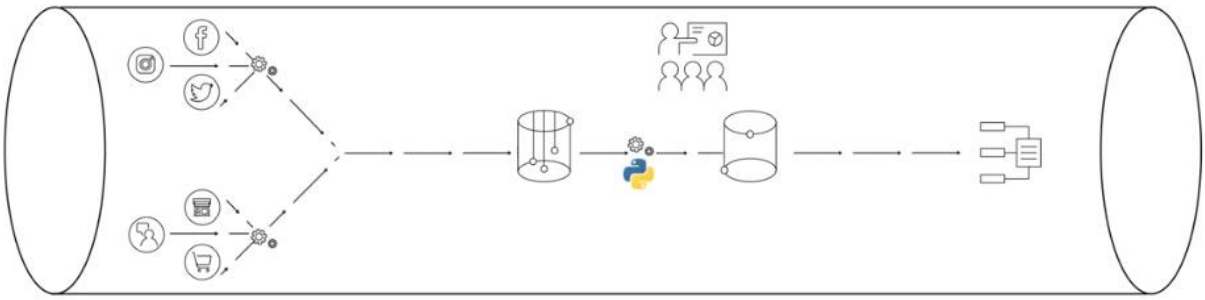
Performance Tuning and Troubleshooting

Introduction

One of the key responsibilities of a Data Engineer is to monitor and optimize systems and data flows for **performance** and **availability**.

Data Pipelines

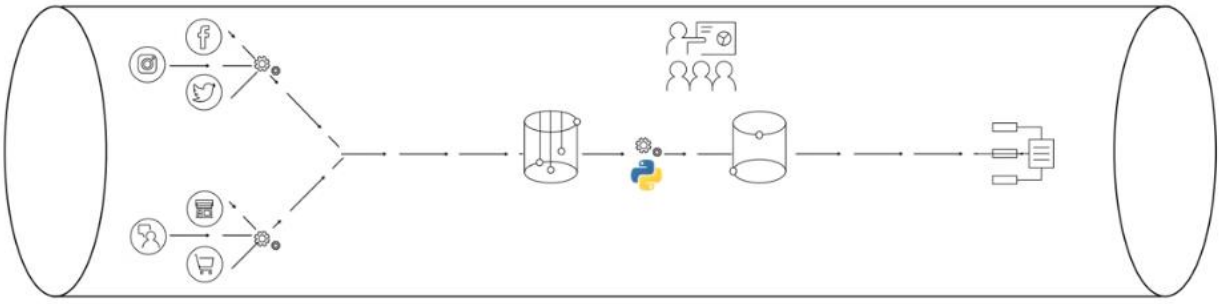
Data Pipelines



A data pipeline typically runs with a combination of complex tools and can face several different types of performance threats.

Data Pipelines Performance Threats

Data Pipelines - Performance Threats



Performance threats include:

- Scalability in the face of increasing data sets and workloads
- Application failures
- Scheduled jobs not functioning accurately
- Tool incompatibilities

Data Pipeline Performance Metrics

Data Pipelines - Performance Metrics

Latency - The time it takes for a service to fulfill a request

Failures - The rate at which a service fails

Data Pipelines - Performance Metrics

Resource **utilization** and utilization patterns

Traffic - Number of user requests received in a given period

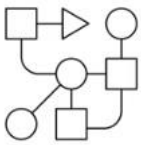
Troubleshooting

Data Pipelines - Troubleshooting

Collect information about the incident to ascertain if observed behavior is an issue.

Check if you're working with all the **right versions** of software and source codes.

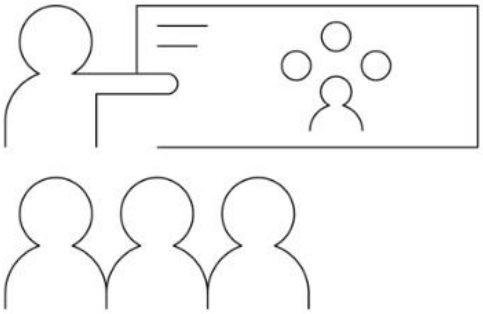
Data Pipelines - Troubleshooting



Check the **logs and metrics** early on in your troubleshooting process to isolate whether an issue is related to infrastructure, data, software, or a combination of these.

Reproduce the issue in a test environment. This can be an iterative and time-consuming process.

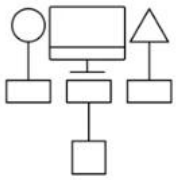
Database Optimization for Performance



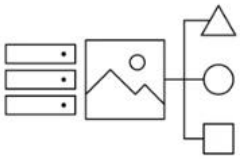
Performance Metrics for Databases:

- System outages
- Capacity utilization
- Application slowdown
- Performance of queries
- Conflicting activities and queries being executed simultaneously
- Batch activities causing resource constraints

Database Optimization for Performance

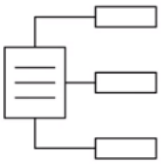


Capacity Planning - Determining the optimal hardware and software resources required for performance.

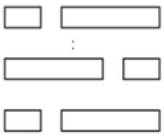


Database Indexing - Locating data without searching each row in a database resulting in faster querying.

Database Optimization for Performance



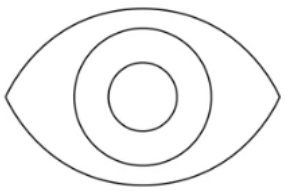
Database Partitioning - Dividing large tables into smaller, individual tables, improving performance and data manageability.



Database Normalization - Reducing inconsistencies arising out of data redundancy and anomalies arising out of update, delete, and insert operations on databases.

Monitoring Systems

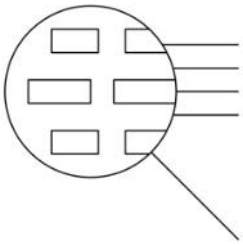
Monitoring Systems



Monitoring and alerting systems help us collect quantitative data about our systems and applications in real time.

These systems give visibility into the performance of data pipelines, data platforms, databases, applications, tools, queries, scheduled jobs, and more.

Monitoring Systems



Database Monitoring Tools take frequent snapshots of the performance indicators of a database.

This helps to:

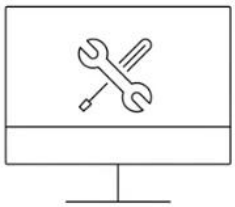
> Track when and how a problem started to occur.

> Isolate and get to the root of the issue.

Monitoring Systems

Application Performance Management Tools measure and monitor the performance of applications and amount of resources utilized by each process. This helps in proactive allocation of resources to improve application performance.

Monitoring Systems



Tools for Monitoring Query Performance gather statistics about query throughput, execution, performance, resource utilization and utilization patterns for better planning and allocation of resources.

Monitoring Systems

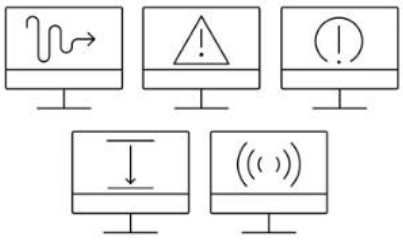
Job-level Runtime Monitoring breaks up a job into a series of logical steps which are monitored for completion and time to completion.

Monitoring Systems

Monitoring Amount of Data being Processed through a data pipeline helps to assess if size of workload is slowing down the system.

Maintenance Schedules

Maintenance Schedules



Preventive maintenance routines generate data that we can use to identify systems and procedures responsible for faults and low availability.

These routines can be:

- Time-based – Planned as scheduled activities at pre-fixed time intervals.
- Condition-based – Performed when there is a specific issue or a decrease in performance.

Lab: Explore your dataset using SQL queries using Datasette

Dataset:

```
select * from exercise03_car_sales_data
```

price	mileage	engType	year	model
3300	350	Diesel	1995	Discovery
3500	200	Petrol	2003	Freelander
3550	255	Diesel	2001	Discovery
3700	124	Petrol	2005	Freelander
3900	290	Diesel	1998	Range Rover
6600	298	Diesel	1997	Discovery
6900	167	Gas	2003	Freelander
7000	190	Diesel	2004	Discovery
7800	355	Diesel	2003	Range Rover
8900	295	Gas	2000	Range Rover
11000	270	Diesel	1996	Range Rover
12999	164	Diesel	2008	Range Rover Sport
14000	230	Diesel	2007	Range Rover
14900	111	Diesel	2007	Freelander
15900	189	Petrol	2007	Freelander
16500	213	Diesel	2006	Discovery
16700	240	Petrol	2005	Discovery
17000	188	Diesel	2007	Freelander
19500	169	Gas	2007	Range Rover Sport
19999	106	Gas	2005	Range Rover Sport
21000	136	Diesel	2010	Freelander
22000	240	Petrol	2003	Range Rover
22500	168	Diesel	2007	Range Rover Sport
22999	167	Gas	2006	Range Rover Evoque
23000	140	Diesel	2007	Range Rover Sport
24777	198	Gas	2008	Range Rover
25000	175	Gas	2007	Range Rover
25300	127	Diesel	2007	Range Rover
26500	260	Gas	2008	Range Rover

Commands Ran

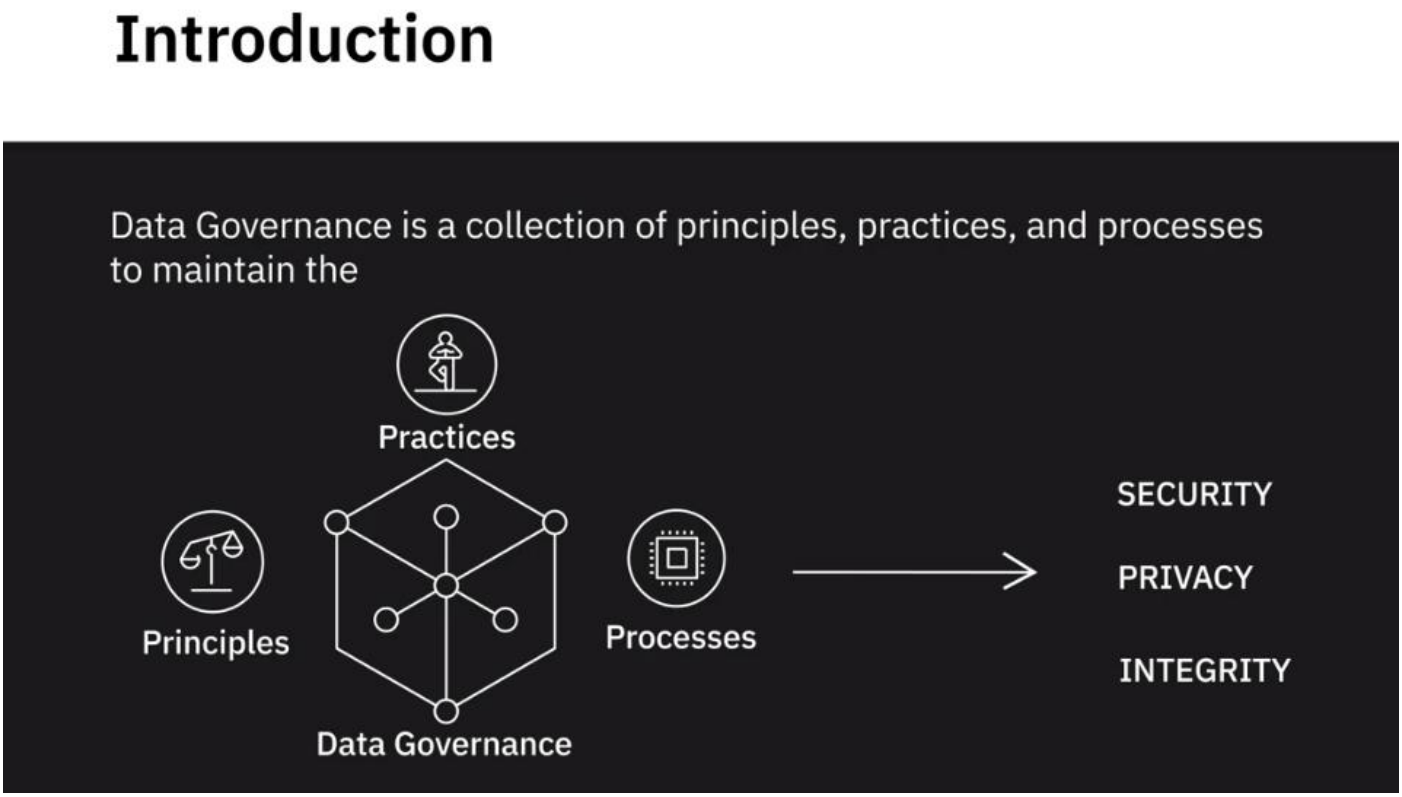
Execute the following query to check the maximum price.

select max(price) as max_price from exercise03_car_sales_data

Execute the following query to display the distinct models.

select distinct(model) from exercise03_car_sales_data;

Governance and Compliance



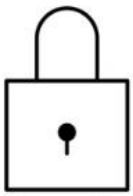
Data that needs governance

Personal Data

Data that needs Governance

Personal: Personal Information (PI) and Sensitive Personal Information (SPI)

- Can be traced back to an individual
- Can be used to identify an individual
- Can be used to cause harm to an individual



General Data Protection Regulation (GDPR)

California created the California Consumer Privacy Act

Data that needs Governance

Industry-specific regulations:

Health Insurance Portability and Accountability Act (**HIPAA**) for Healthcare

Payment Card Industry Data Security Standard (**PCI DSS**) for Retail

Sarbanes Oxley (**SOX**) for Finance

Compliance

Compliance

Compliance covers the processes and procedures through which an organization adheres to regulations and conducts its operations in a **legal and ethical manner.**

Establish controls and checks in order to comply with regulations

Maintain a verifiable audit trail to establish adherence to regulations

Compliance

Compliance is an ongoing process requiring a blend of

PEOPLE

PROCESS

TECHNOLOGY

Data Lifecycle

Governance regulations require enterprises to **know their purpose** and **maintain transparency in their actions** at each step of the data lifecycle.

Acquisition

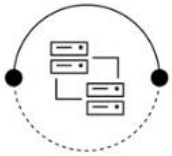
In the **Data Acquisition Stage**, you need to:

- Identify data that needs to be collected and the legal basis for procuring the data.
- Establish the intended use of data, published as a privacy policy.
- Identify the amount of data you need to meet your defined purposes.

Data Lifecycle: Processing

Data Lifecycle

Governance regulations require enterprises to **know their purpose** and **maintain transparency in their actions** at each step of the data lifecycle.



Processing

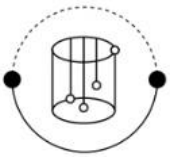
In the **Data Processing Stage**, you need to:

- Flesh out details of how exactly you are going to process personal data.
- Establish your legal basis for the processing of personal data.

Data Lifecycle: Storage

Data Lifecycle

Governance regulations require enterprises to **know their purpose** and **maintain transparency in their actions** at each step of the data lifecycle.



Storage

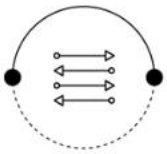
In the **Data Storage Stage**, you need to:

- Define where you will store the data.
- Establish specific measures you will take to prevent internal and external security breaches.

Sharing

Data Lifecycle

Governance regulations require enterprises to **know their purpose** and **maintain transparency in their actions** at each step of the data lifecycle.



Sharing

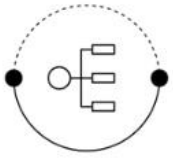
In the **Data Sharing Stage**, you need to:

- Identify third-party vendors in your supply chain that will have access to the collected data.
- Establish how you will hold third-party vendors contractually accountable to regulations.

Retention and Disposal

Data Lifecycle

Governance regulations require enterprises to **know their purpose** and **maintain transparency in their actions** at each step of the data lifecycle.



Retention and Disposal

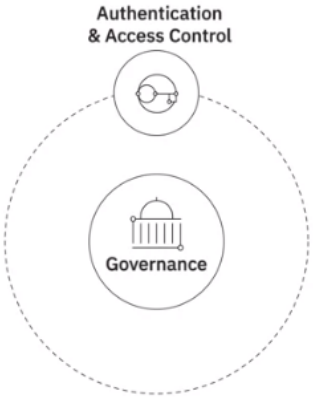
In the **Data Retention and Disposal Stages**, you need to:

- Define policies and processes you will follow for the retention and deletion of personal data after a designated time.
- Define how you will ensure deleted data is removed from all locations, including third-party systems.

Technology as an enabler

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



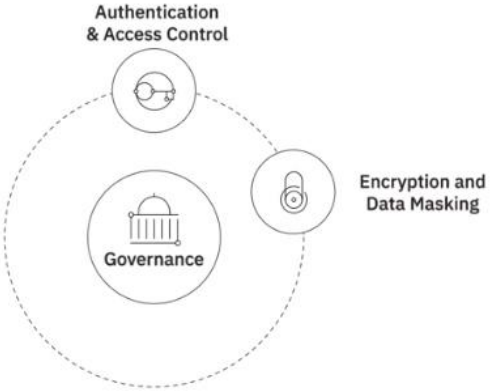
Authentication and Access Control

- Layered authentication processes
- Combination of passwords, tokens, and biometrics, to prevent unauthorized access
- Authentication systems verify that you are who you say you are

Data encryption

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



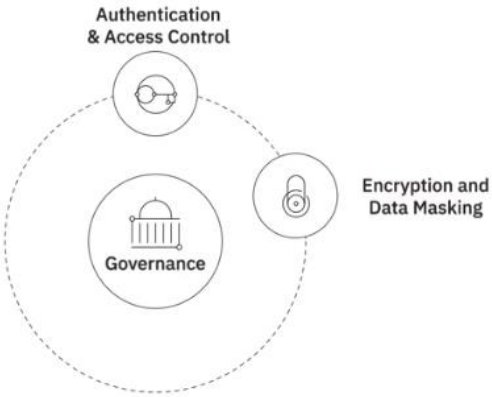
Encryption and Data Masking

- Encryption converts to an encoded format that can only be legible once it is decrypted via a secure key
- Encryption of data is available for:
  - Data at rest
  - Data in transit

Data Masking

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.

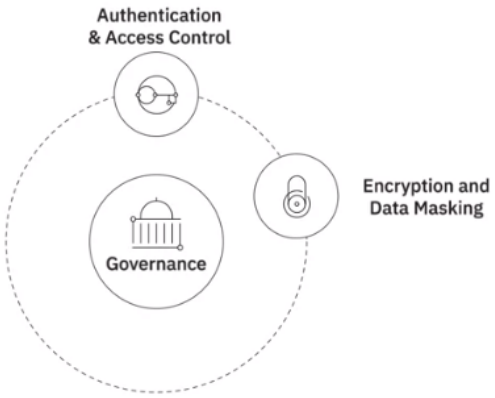


Encryption and Data Masking

- Data Masking provides anonymization of data for downstream processing and pseudonymization of data
- Anonymization abstracts the presentation layer without changing the data in the database itself

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



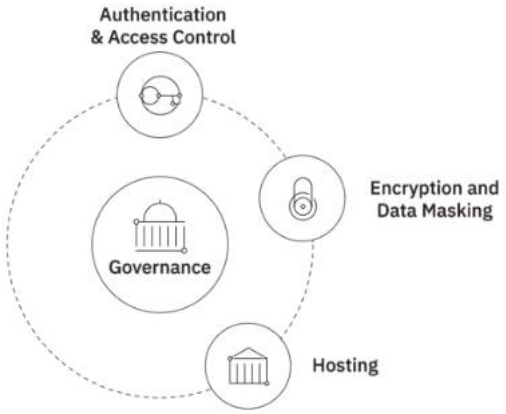
Encryption and Data Masking

- Pseudonymization of data replaces personally identifiable information with artificial identifiers so that it cannot be traced back to an individual's identity

Hosting

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



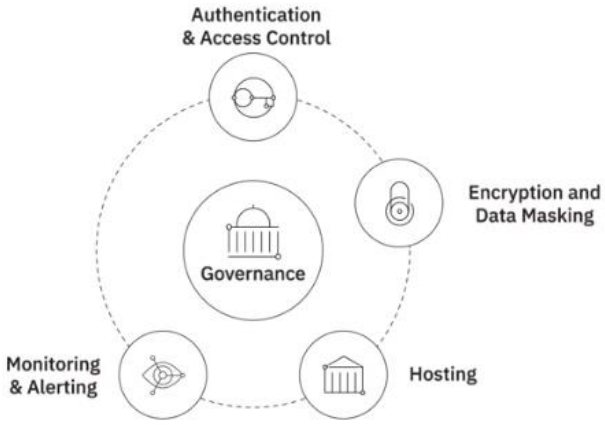
Hosting

On-premise and cloud-based systems provide hosting options that comply with the requirements and restrictions for international data transfers

Monitoring and alerting

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.

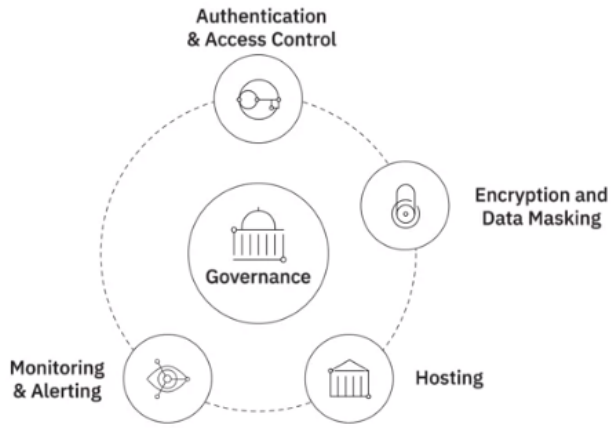


Monitoring and Alerting

- Security monitoring proactively monitors, tracks, and reacts to security violations across infrastructure, applications, and platforms
- Monitoring systems provide detailed audit reports that track access and other operations on the data

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



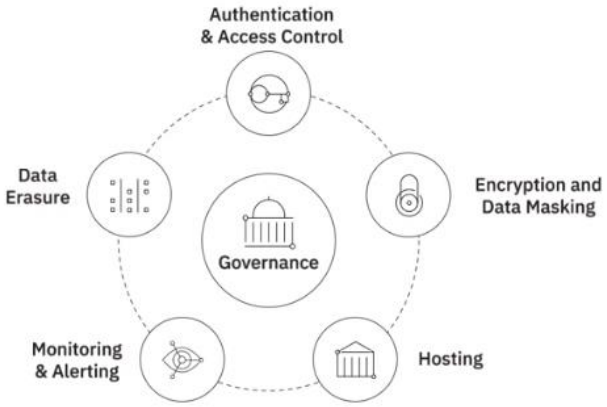
Monitoring and Alerting

- Alerting functionalities flag security breaches so immediate remedial actions can be triggered
- Alerts are based on the severity and urgency level of a breach

Data Erasure

Technology as an Enabler

Today's tools and technologies provide several controls for ensuring organizations comply to governance regulations.



Data Erasure

A software-based method of permanently clearing data from a system by overwriting.

Data erasure prevents deleted data from being retrieved.

Optional: Overview of the DataOps Methodology

Gartner defines DataOps as a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and consumers across an organization. DataOps aims to create predictable delivery and change management of data, data models, and related artifacts. DataOps uses technology to automate data delivery with the appropriate levels of security, quality, and metadata to improve the use and value of data in a dynamic environment." (Source: https://blogs.gartner.com/nick-heudecker/hyping-dataops/) A small team working on a simpler or limited number of use cases can meet business requirements efficiently. As data pipelines and data infrastructures get more complex, and data teams and consumers grow in size, you need development processes and efficient collaboration between teams to govern the data and analytics lifecycle. From data ingestion and data processing to analytics and reporting, you need to reduce data defects, ensure shorter cycle times, and ensure 360-degree access to quality data for all stakeholders. DataOps helps you achieve this through metadata management, workflow and test automation, code repositories, collaboration tools, and orchestration to help manage complex tasks and workflows. Using the DataOps methodology ensures all activities occur in the right order the right security permissions. It helps set in a continual process that allows you to cut wastages, streamline steps, automate processes, increase throughput, and improve continually. Several DataOps Platforms are available in the market, some of the popular ones being IBM DataOps, Nexla, Switchboard, Streamsets, and Infoworks.

DataOps Methodology:

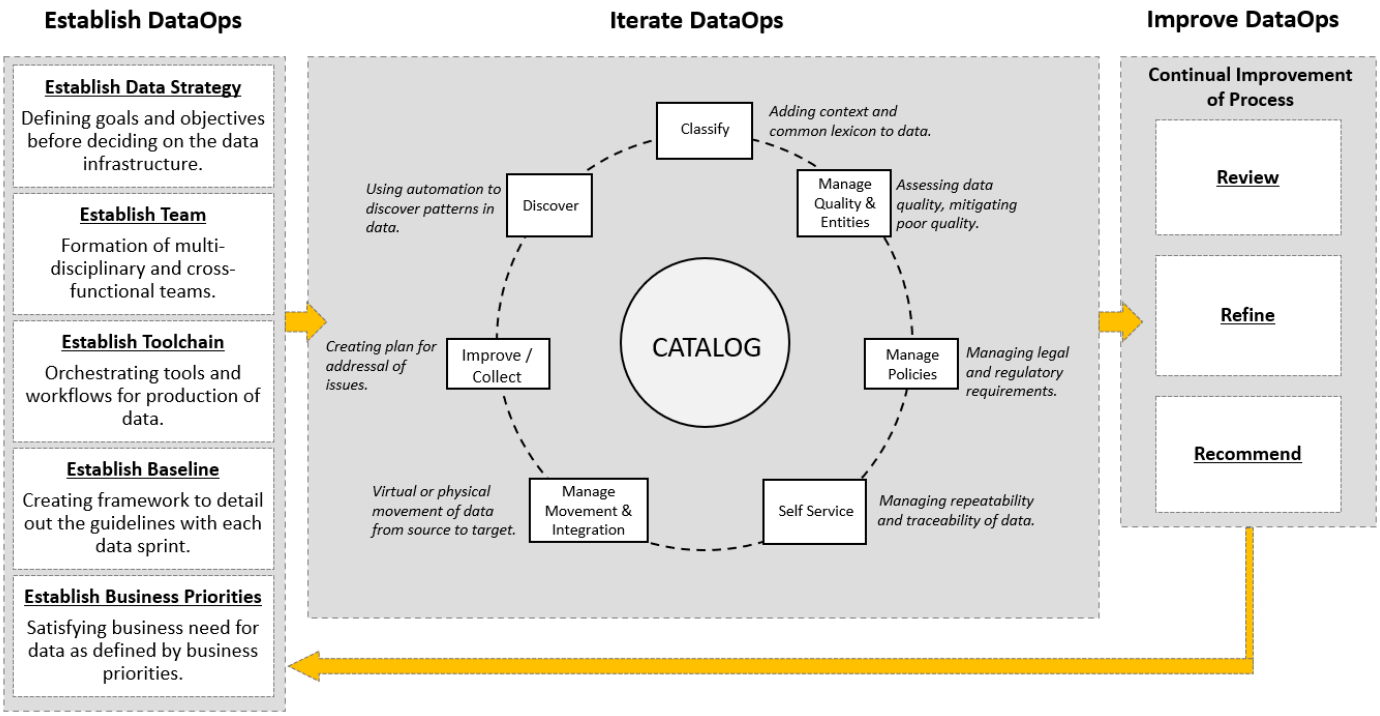
The purpose of the DataOps Methodology is to enable an organization to utilize a repeatable process to build and deploy analytics and data pipelines. Successful implementation of this methodology allows an organization to know, trust, and use data to drive value.

It ensures that the data used in problem-solving and decision making is relevant, reliable, and traceable and improves the probability of achieving desired business outcomes. And it does so by tackling the challenges associated with inefficiencies in accessing, preparing, integrating, and making data available.

In a nutshell, the DataOps Methodology consists of three main phases:

- The **Establish DataOps Phase** provides guidance on how to set up the organization for success in managing data.
- The **Iterate DataOps Phase** delivers the data for one defined sprint.
- The **Improve DataOps Phase** ensures learnings from each sprint is channeled back to continually improve the DataOps process.

The figure below presents a high-level overview of these phases and the key activities within each of these phases.



Benefits of using the DataOps methodology:

Adopting the DataOps methodology helps organizations to organize their data and make it more trusted and secure. Using the DataOps methodology, organizations can:

- Automate metadata management and catalog data assets, making them easy to access.
- Trace data lineage to establish its credibility and for compliance and audit purposes.
- Automate workflows and jobs in the data lifecycle to ensure data integrity, relevancy, and security.
- Streamline the workflow and processes to ensure data access and delivery needs can be met at optimal speed.
- Ensure a business-ready data pipeline that is always available for all data consumers and business stakeholders.
- Build a data-driven culture in the organization through automation, data quality, and governance.

As a data practitioner, using the methodology can help you reduce development time, cut wastages and duplication of effort, increase your productivity and throughput, and ensure that your actions produce the best possible quality of data.

With DataOps, data professionals, consumers, and stakeholders can collaborate more effectively towards the shared goal of creating valuable insights for business. While implementing the methodology will require systemic change, time, and resources, but in the end, it makes data and analytics more efficient and reliable.

Interestingly, it also opens up additional career opportunities for you as a data engineer. **DataOps Engineers** are technical professionals that focus on the development and deployment lifecycle rather than the product itself. And as you grow in experience, you can move into more specialist roles within DataOps, contributing to defining the data strategy, developing and deploying business processes, establishing performance metrics, and measuring performance.