Big Data

Big Data Course

Eng.Mostafa Nabieh

Mostafa Nabieh

وزارة الاتصالات وتكنولوجيا المعلومات

MANNING

وزارة الاتصالات وتقنية المعلومات
MINISTRY OF COMMUNICATIONS AND INFORMATION TECHNOLOGY

AI TALENTS HUB

UDACITY

PLURALSIGHT

Learning Space
YOUR SPACE TO LEARN FUTURE SKILLS

BIG DATA

Mostafa Nabieh

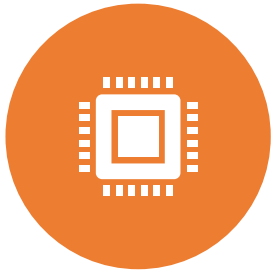
HUAWEI

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

IBM

Mostafa Nabieh

CONTENTS



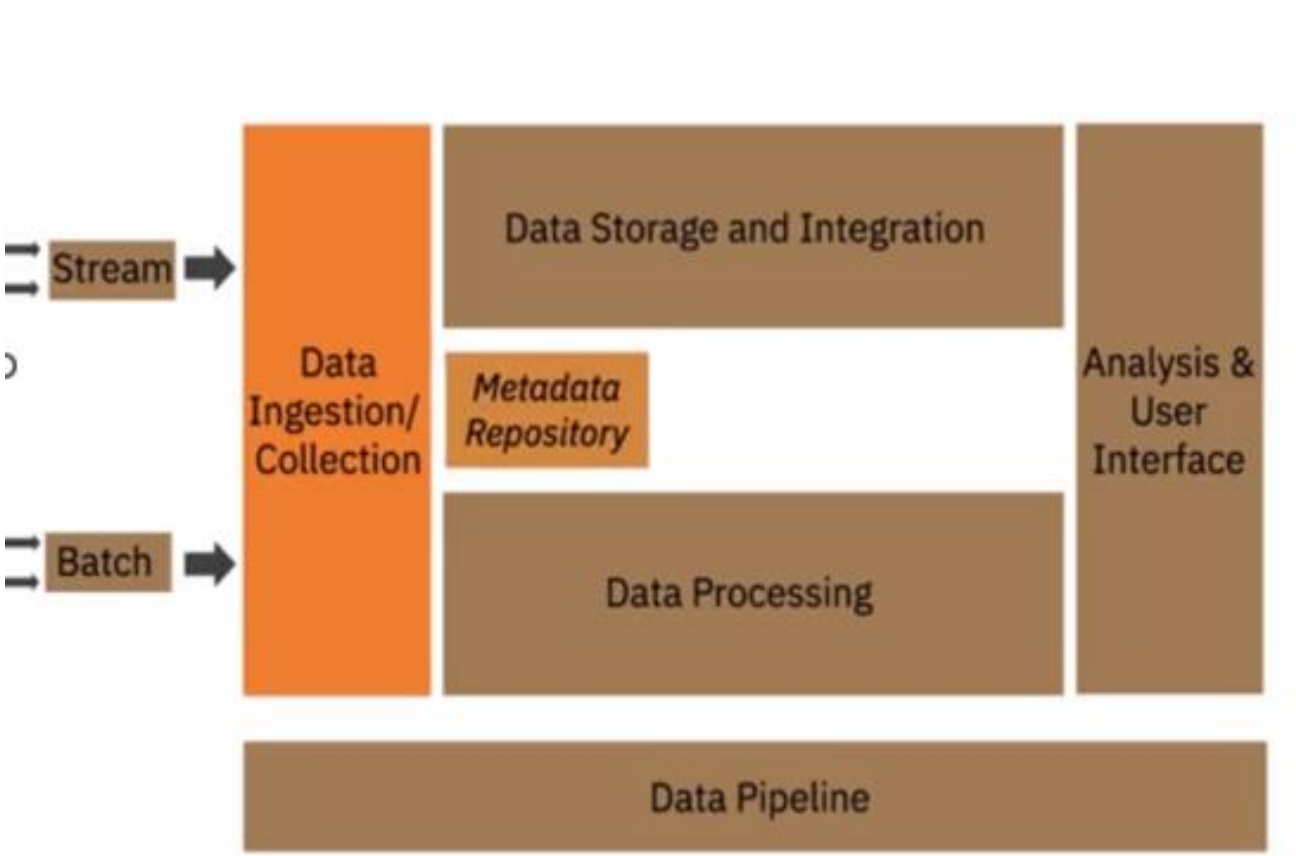
WHAT IS DATA ENGINEERING?

BIG DATA ECOSYSTEM

BIG DATA LIFECYCLE

CAREER OPPORTUNITIES

Data Ingestion or Collection Layer



- Connect to data sources
- Transfer data from data sources to the data platform in streaming and batch modes
- Maintain information about the data collected in the metadata repository

Tools for Data Ingestion

Data Flow

IBM Streams

IBM Streaming Analytics on Cloud

amazon Kinesis

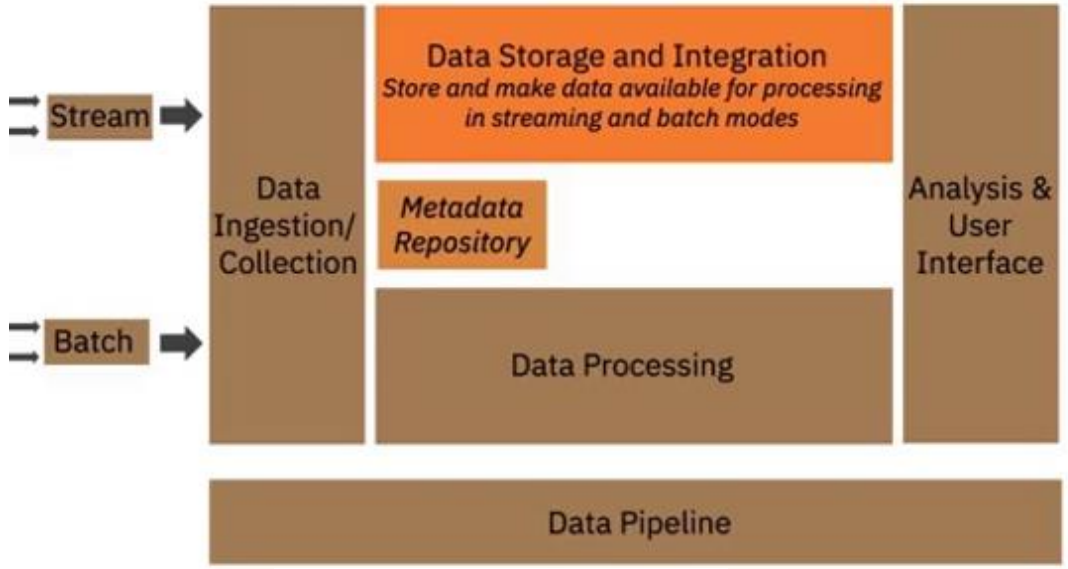
kafka

Data Storage and Integration Layer

- Store data for processing and long-term use
- Transform and merge extracted data, either logically or physically
- Make data available for processing in both streaming and batch modes

RELIABLE HIGH-PERFORMING

SCALABLE COST-EFFICIENT



Relational Databases:

Db2 IBM

Microsoft SQL Server

MySQL

ORACLE DATABASE

PostgreSQL

Integration Tools:

IBM's Cloud Pak for Data

IBM's Cloud Pak for Integration

talend DATA FABRIC

OpenStudio

Database-as-a-Service:

Db2 IBM

amazon RDS

Google Cloud SQL

Microsoft SQL Azure

Open-source Integration Tools:

boomi A Dell Technologies Business

snaplogic

Non-Relational Database:

Cloudant

redis

mongoDB

cassandra

neo4j

Platform as a Service (iPaaS):

Adeptia Integration Suite

Google Cloud's Cooperation 534

IBM's Application Integration Suite on Cloud

Informatica's Integration Cloud

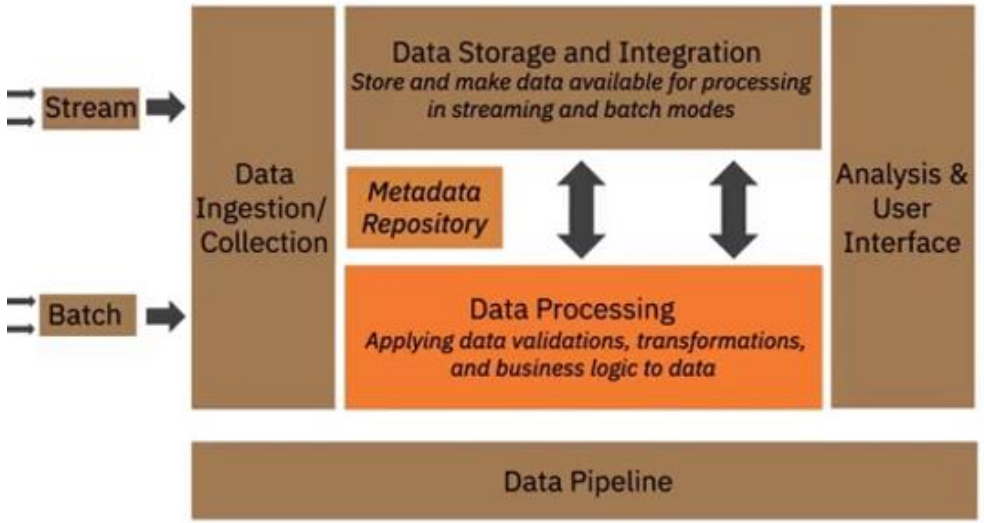
Data Processing Layer

- Read data in batch or streaming modes from storage and apply transformations
- Support popular querying tools and programming languages
- Scale to meet the processing demands of a growing dataset
- Provide a way for analysts and data scientists to work with data in the data platform

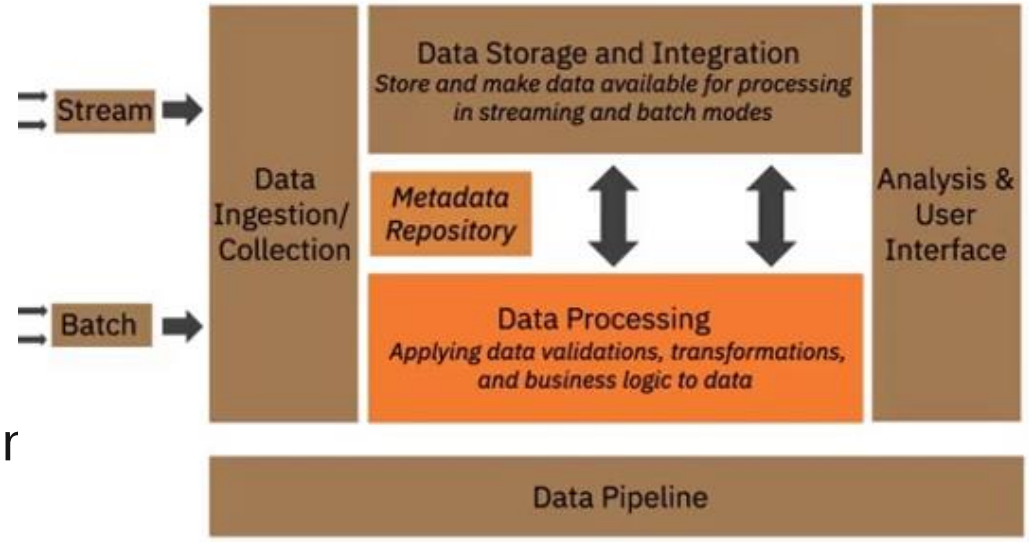
Transformation Tasks:

- **Structuring -** Actions that change the form and schema of the data
- **Normalization -** Cleaning the database of unused data and reducing redundancy and inconsistency
- **Denormalization -** Combining data from multiple tables into a single table so that it can be queried more efficiently
- **Data Cleaning -** Fixing irregularities in data to provide credible data for downstream applications and uses

There are a host of tools available for performing these transformations on data, selected based on the data size, structure, and specific capabilities of the tool.



Data Processing Layer



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

- **Storage and Processing** can occur in the same layer
- **Data can first be stored in Hadoop File Distribution System, or HDFS**, and then processed in a data processing engine like **Spark**.

Data Processing layer can also precede the Data Storage layer, where transformations are applied before the data is loaded, or stored, in the database.

Spreadsheets

TRIFACTA Wrangler

OpenRefine

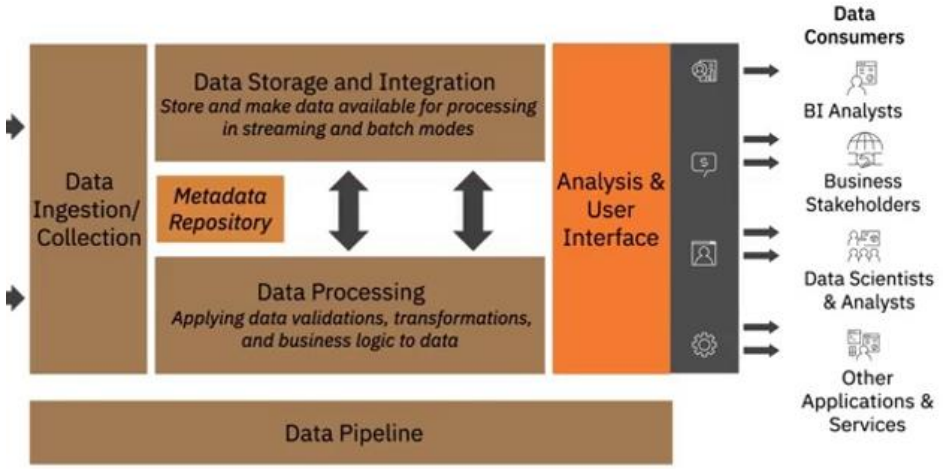
python

Google DataPrep

Watson Studio Refinery

Analysis and User Interface Layer

- Querying tools and programming languages. SQL, SQL-like query tools for NoSQL, Programming Languages like Python, R, and Java
- APIs that can be used to run reports on data for both online and offline processing.
- APIs that can consume data from the storage in real-time for use in other applications and services.
- Dashboarding and Business Intelligence applications. IBM Cognos Analytics, Tableau, Jupyter Notebooks, Python and R libraries, and Microsoft Power BI

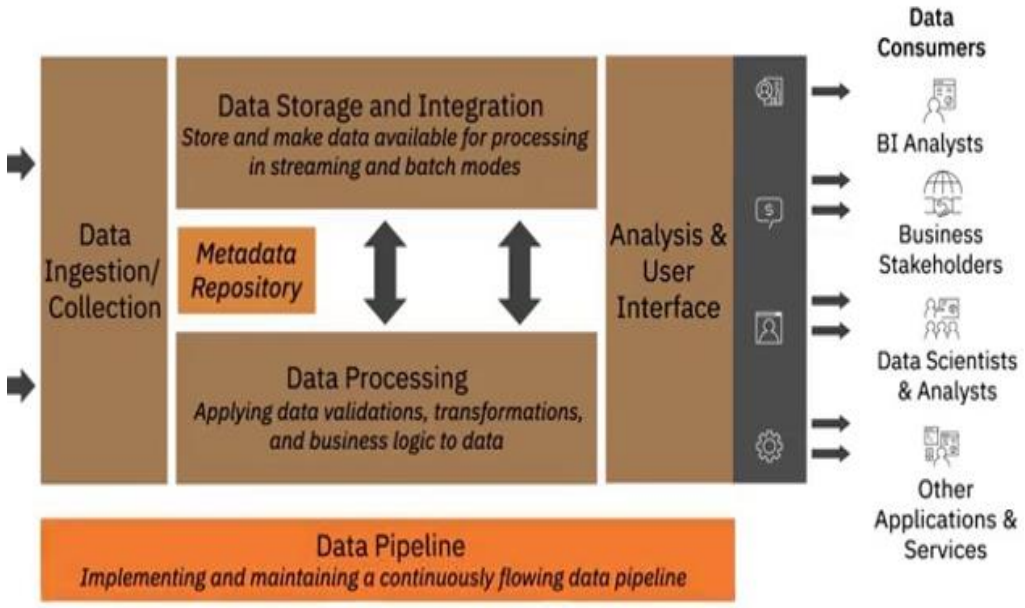


Data Pipeline Layer

- 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.

Apache Airflow

Data Flow

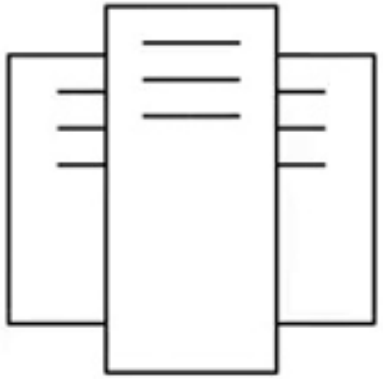



Security

The CIA Triad

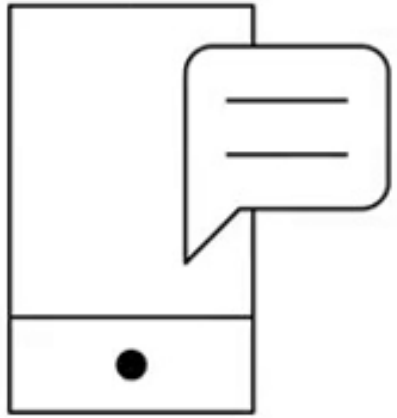
- Key components to creating an effective strategy for information security include:
  - **Confidentiality** through controlling unauthorized access
  - **Integrity** through validating that your resources are trustworthy and have not been tampered with
  - **Availability** by ensuring authorized users have access to resources when they need it

The CIA Triad



Infrastructure Security

Network Security



Application Security

Data 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

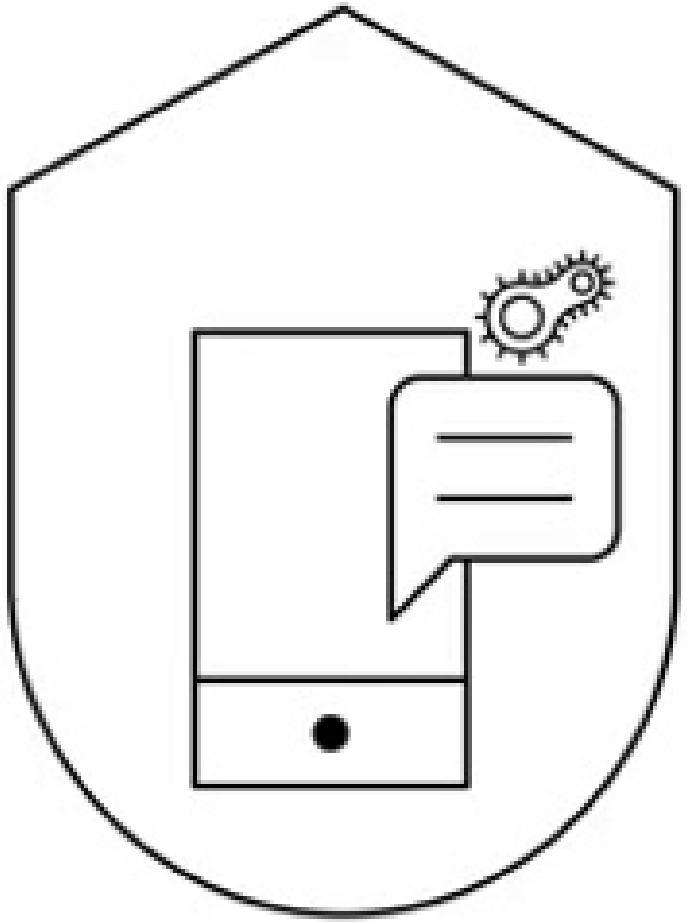
Network security is vital to keep interconnected systems and data safe.

- **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 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

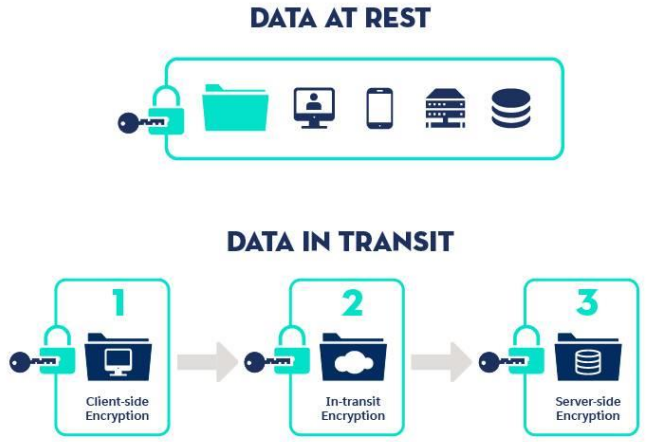
- Data is either at rest in storage, or in transit, between systems, applications, services, and workloads.
- Authentication systems verify you are who you say you are.
- Authorization ensures users access information based on their role and privileges.

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:

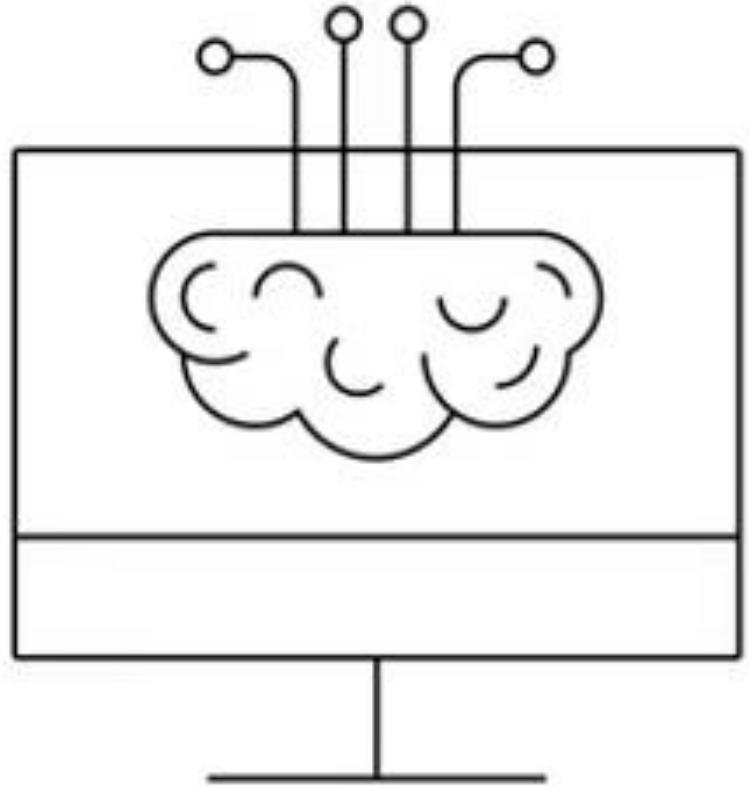
- Moving from one location to another over the internet
- Can be protected using encryption methods such as HTTPS, SSL, and TLS



Monitoring and Intelligence

Security Monitoring and Intelligence Systems:

- Create an audit history for triage and compliance purposes
- Provide reports and alerts that help enterprises react to security violations in time



Monitor >> Track >> React


Thank
you