Data Engineering

DataBase Foundations

Author: Eng. Carlos Andrés Sierra, M.Sc.
cavirguezs@udistrital.edu.co

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-III

Outline

1. Data Engineering
2. Exploratory Data Analysis

EDA

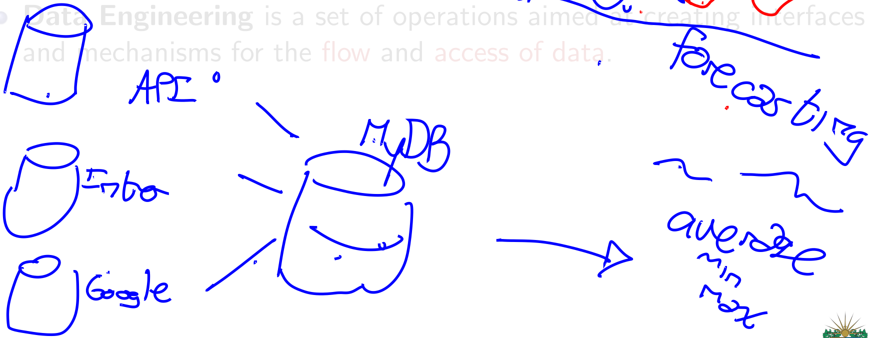
Outline

1. Data Engineering

2. Exploratory Data Analysis

What is Data Engineering?

- **Data Engineering** is the aspect of data science that focuses on practical applications of data collection and analysis.
- **Data Engineers** are responsible for building and maintaining the architecture that allows data scientists to perform their work.
- **Data Engineering** is a set of operations aimed at creating interfaces and mechanisms for the flow and access of data.



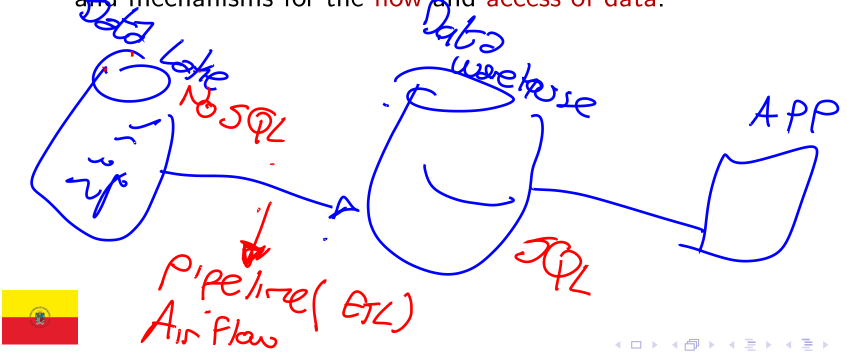
What is Data Engineering?

- **Data Engineering** is the aspect of data science that focuses on practical applications of data collection and analysis.
- **Data Engineers** are responsible for building and maintaining the architecture that allows data scientists to perform their work.
- **Data Engineering** is a set of operations aimed at creating interfaces and mechanisms for the flow and access of data.

Data Science

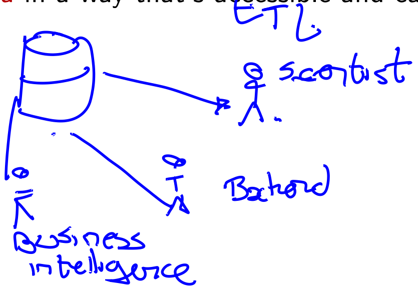
What is Data Engineering?

- **Data Engineering** is the aspect of data science that focuses on practical applications of data collection and analysis.
- **Data Engineers** are responsible for building and maintaining the architecture that allows data scientists to perform their work.
- **Data Engineering** is a set of operations aimed at creating interfaces and mechanisms for the flow and access of data.

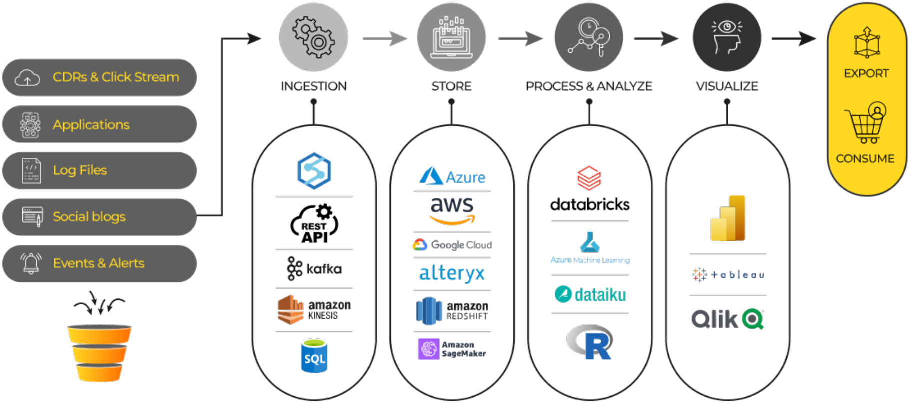


Why is important Data Engineering?

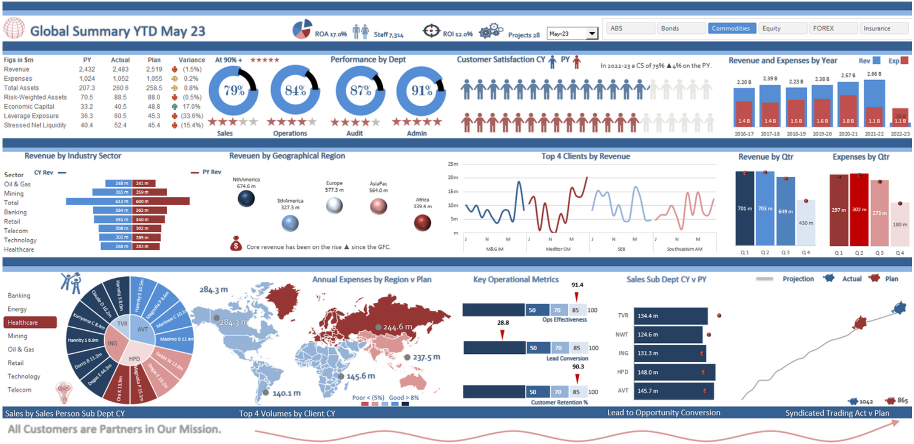
- **Data Engineering** is the foundation of the high-quality data that is necessary for effective data science.
- **Data Engineering** is the process of collecting, transforming, and storing data in a way that's accessible and easy to analyze.



Data Engineering Architecture



Case of Study: Dashboards



Figs in $m	PY	Actual	Plan	Variance
Revenue	2,432	2,483	2,519	(1.5%)
Expenses	1,024	1,052	1,055	0.2%
Total Assets	207.3	260.6	258.5	0.8%
Risk-Weighted Assets	70.5	88.5	88.0	(0.5%)
Economic Capital	33.2	40.5	48.8	17.0%
Leverage Exposure	36.3	60.5	45.3	(33.6%)
Stressed Net Liquidity	40.4	52.4	45.4	(15.4%)

Data Science

- **Data Science** is the process of extracting knowledge from data.
- **Data Science** is the process of analyzing and interpreting complex digital data.
- **Data Science** is the process of creating models that can predict future outcomes.
- **Data Science** is the process of creating visualizations to help understand data.

Data Science

- **Data Science** is the process of extracting knowledge from data.
- **Data Science** is the process of analyzing and interpreting complex digital data.
- **Data Science** is the process of creating models that can predict future outcomes.
- **Data Science** is the process of creating visualizations to help understand data.

Data Science

- **Data Science** is the process of extracting knowledge from data.
- **Data Science** is the process of analyzing and interpreting complex digital data.
- **Data Science** is the process of creating models that can predict future outcomes.
- **Data Science** is the process of creating visualizations to help understand data.

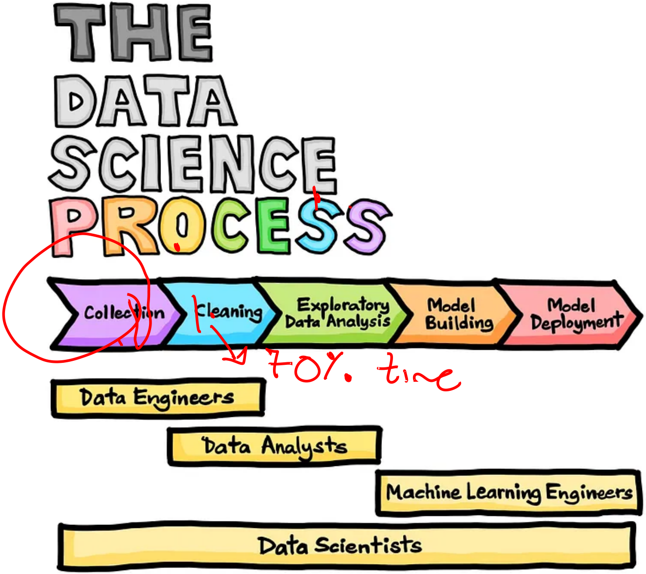
AI

Forecasting

Data Science

- **Data Science** is the process of extracting knowledge from data.
- **Data Science** is the process of analyzing and interpreting complex digital data.
- **Data Science** is the process of creating models that can predict future outcomes.
- **Data Science** is the process of creating visualizations to help understand data.

Data Science Workflow



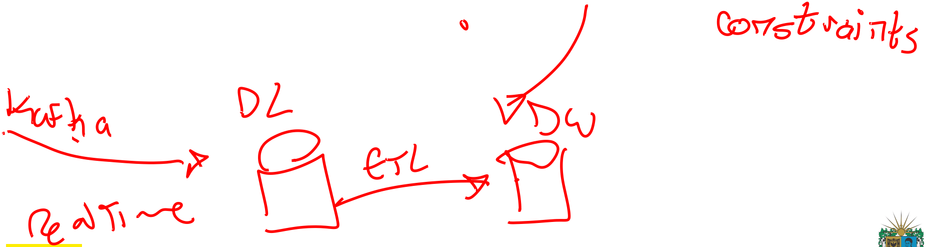
DBOps vs Data Engineer

- **DBOps** is responsible for the operation of the database.
- **DBOps** is responsible for the performance of the database.
- **DBOps** is responsible for the security of the database.
- **Data Engineer** is responsible for the data architecture.
- **Data Engineer** is responsible for the data quality.
- **Data Engineer** is responsible for the data flow.

Data Architect

DBOps vs Data Engineer

- **DBOps** is responsible for the operation of the database.
- **DBOps** is responsible for the performance of the database.
- **DBOps** is responsible for the security of the database.
- **Data Engineer** is responsible for the data architecture.
- **Data Engineer** is responsible for the data quality.
- **Data Engineer** is responsible for the data flow.



Outline

1 Data Engineering

2 Exploratory Data Analysis

EDA

What is Exploratory Data Analysis?

- Exploratory Data Analysis (EDA) is an **approach** to analyzing data sets to **summarize** their **main characteristics**.
- Exploratory Data Analysis (EDA) is the process of **visualizing and analyzing data** to extract insights.
- Exploratory Data Analysis (EDA) is the process of **understanding the data** before building a model.

ANN
RF
XGBoost

- Exploratory Data Analysis (EDA) is the process of **cleaning and preparing data** for analysis.

format

- Exploratory Data Analysis (EDA) is the process of **identifying patterns** in the data.

Techniques using for EDA

- **Descriptive Statistics:** is the process of summarizing data using statistical measures.

mean

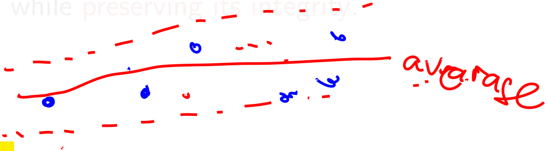
variance

standard deviation

min

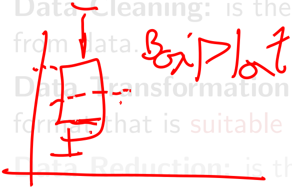
max

- **Data Visualization:** is the process of creating visual representations of data.
- **Data Cleaning:** is the process of removing errors and inconsistencies from data.
- **Data Transformation:** is the process of transforming data into a format that is suitable for analysis.
- **Data Reduction:** is the process of reducing the size of the data while preserving its integrity.



Techniques using for EDA

- **Descriptive Statistics:** is the process of summarizing data using statistical measures.
- **Data Visualization:** is the process of creating visual representations of data.
- **Data Cleaning:** is the process of removing errors and inconsistencies from data.
- **Data Transformation:** is the process of transforming data into a format that is suitable for analysis.
- **Data Reduction:** is the process of reducing the size of the data while preserving its integrity.



Techniques using for EDA

- **Descriptive Statistics:** is the process of summarizing data using statistical measures.
- **Data Visualization:** is the process of creating visual representations of data.
- **Data Cleaning:** is the process of removing errors and inconsistencies from data.
- **Data Transformation:** is the process of transforming data into a format that is suitable for analysis.
- **Data Reduction:** is the process of reducing the size of the data while preserving its integrity.

Techniques using for EDA

- **Descriptive Statistics:** is the process of summarizing data using statistical measures.
- **Data Visualization:** is the process of creating visual representations of data.
- **Data Cleaning:** is the process of removing errors and inconsistencies from data.
- **Data Transformation:** is the process of transforming data into a format that is suitable for analysis.
- **Data Reduction:** is the process of reducing the size of the data while preserving its integrity.

format

new column

Techniques using for EDA

- **Descriptive Statistics:** is the process of summarizing data using statistical measures.
- **Data Visualization:** is the process of creating visual representations of data.
- **Data Cleaning:** is the process of removing errors and inconsistencies from data.
- **Data Transformation:** is the process of transforming data into a format that is suitable for analysis.
- **Data Reduction:** is the process of reducing the size of the data while preserving its integrity.

↓ colomns

How to improve data quality?

- **Data Quality** is the process of ensuring that data is accurate, complete, and reliable.
- **Data Quality** is the process of ensuring that data is consistent and up-to-date.
- **Data Quality** is the process of ensuring that data is free from errors and inconsistencies.
- **Data Quality** is the process of ensuring that data is of high quality and can be trusted.
- **Data Quality** is the process of ensuring that data is fit for purpose and can be used effectively.

Outline

1 Data Engineering

2 Exploratory Data Analysis

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/databases-foundations*