Data Engineering

DataBase Foundations

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

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

2024-I

Outline

max 2 weeks

70 Graph => 4900

edge

volume price load

1 2 3 ... 14 / 54

k8s

REDIS

RL => rewards

stable Baselines

Q-Learning

FastAPI

1. Data Engineering

Facebook Prophet

XGBoost

2. Exploratory Data Analysis

Random Forest

Gradient Boost

Outline

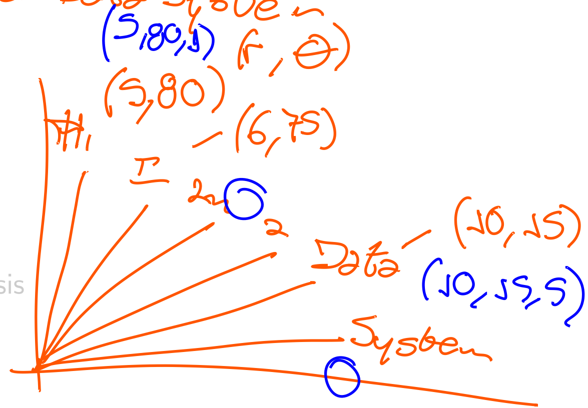
Hi I am a Data System

Vectors

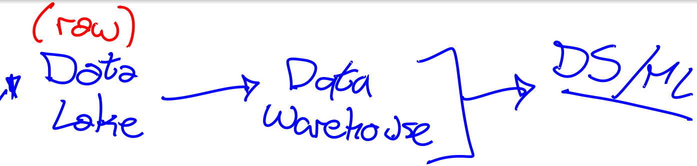
1. Data Engineering

cod	word
Hi	1
I	2
am	3
⋮	⋮
system	6

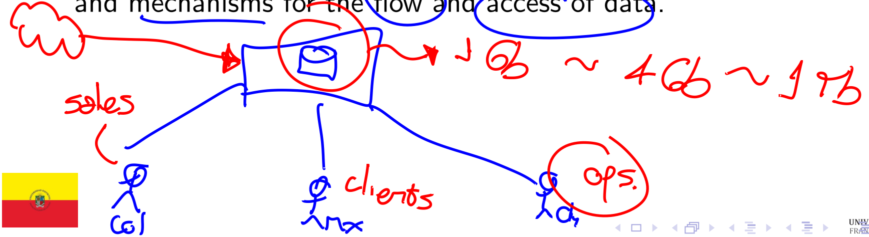
2. Exploratory Data Analysis



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.

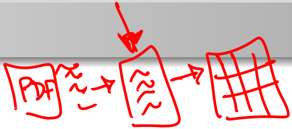


Why is important Data Engineering?

(Airflow) Pipelines

Extract Transform Load

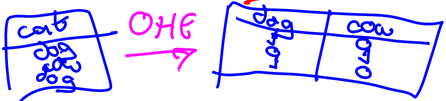
Extract Load → Transform



- 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~~ is the process of **~~building and maintaining~~ the architecture** that allows data scientists to perform their work.

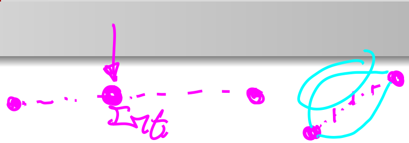
Types → Numerical
→ Categorical X → One Hot Encoder

NeRF



Data Science

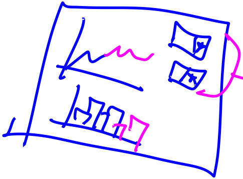
DS
MLE
SE
DE



- Data Science is the process of **extracting knowledge** from data.
- Data Science is the process of **analyzing and interpreting complex digital data**.

ML

- Data Science is the process of **creating models** that can predict future outcomes. Forecasting
- Data Science is the process of **creating visualizations** to help understand data.



Dashboards

Parametos

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**.

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

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

Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

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

Thanks!

Questions?

Repo:

github.com/EngAndres/ud-public/tree/main/courses/databases-foundations