Data Engineering and MLOps in Business

Intro to API & DataBases

Primoz Konda

AAUBS

March 11, 2024

pk@business.aau.dk

Outline

1 Introduction

2 API

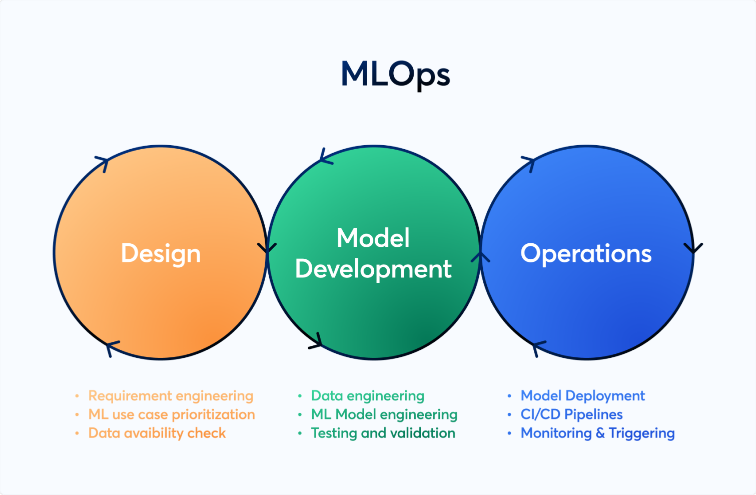
3 Data Bases

4 Exercise

Why do we need to know MLOps?

- Bridging the Gap Between Experimentation and Production
- Ensuring Model Reliability and Scalability
- Facilitating Collaboration Among Teams
- Automating the Machine Learning Lifecycle
- Improving Model Monitoring and Management

MLOps



- Requirement engineering
- ML use case prioritization
- Data avaibility check

- Data engineering
- ML Model engineering
- Testing and validation

- Model Deployment
- CI/CD Pipelines
- Monitoring & Triggering

Serverless Machine Learning

The ML Platform to Build, Maintain, and Monitor ML Systems

- Hopsworks.ai
- From ML Models to MLOps to ML Systems
- Create your account ('Start for free')
- Works with GitHub

Application Programming Interface?

Definition

1. It is a set of defined rules that enable different applications to communicate with each other.
2. It acts as an intermediary layer that processes data transfers between systems.

Usability

APIs simplify software development and innovation by enabling applications to exchange data and functionality easily and securely

Example: 'ex_jokes.py'

Types regarding availability:

- **Public API** is open and available for use by any outside developer or business. These are also called open or external APIs
- **Partner API** is only available to specifically selected and authorized outside developers or API consumers. It facilitates business-to-business activities
- **Private API** is intended only for use within the enterprise to connect systems and data within the business
- **Composite API** is a sequence of tasks that run synchronously as a result of the execution and not at the request of a task.

Types regarding protocol:

1. **REST** (Representational State Transfer) is a web services API and crucial for modern web applications
2. **SOAP** (Simple object access protocol) is a well-established protocol but comes with strict rules, rigid standards
3. **RPC** (Remote Procedure Call protocol) is the oldest and simplest type of API with a goal for the client to execute code on a server
4. **Event-driven** or asynchronous APIs transmit information in quasi-real-time. The advantage is that it allows the source to send a response only when the information is new or has changed, useful for stock exchanges

Requests:

- GET - is a read-only operation and doesn't change the state of the resource but only retrieve data
- POST - sends data to a server to create or update a resource and is often used when submitting web forms
- PUT - is used to update an existing resource with new data
- PATCH - is used to apply partial modifications to a resource
- DELETE - used to delete a specified resource from the server

Example: 'ex_finance.py', 'ex_news.py'

FastAPI

FastAPI is a modern, fast (high-performance) web framework for building APIs with Python 3.7+.

- It is fast
- Supports asynchronous code (async and await commands). It can perform multiple tasks concurrently. In that way, it doesn't need to wait until one called to be answered and can continue with a new request
- Short development line 7-8 lines of code

What is an API Wrapper?

Definition

An API Wrapper is a set of programming instructions that acts as an intermediary layer between an application and the web API it intends to communicate with. It simplifies the API calls by providing more intuitive functions that are easier to work with.

- Simplifies the process of making API calls.
- Handles data formatting, error handling, and connection logic.
- Can provide additional functionality like caching responses.

Example: 'ex_datasets.py'

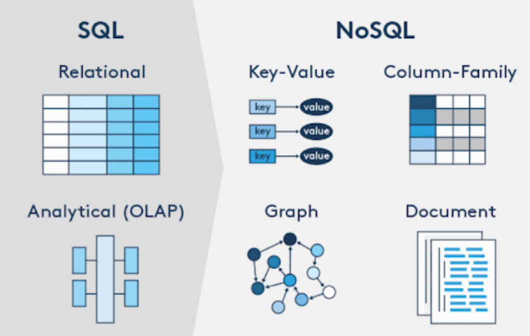
Why We Need Databases in MLOps

- Feature Store: Centralize feature storage and management for consistency across projects.
- Data Versioning: Track changes in datasets to ensure reproducibility.
- Experiment Tracking: Store experiment metadata and results for comparison.
- Model Management: Keep versions of models for deployment and rollback.

Types of Databases Used in MLOps

- Relational Databases (SQL): PostgreSQL, MySQL for structured data.
- NoSQL Databases: MongoDB, Cassandra for unstructured or semi-structured data.
- Time Series Databases: InfluxDB for time-dependent data.
- Graph Databases: Neo4j for complex, interconnected data relationships.
- Feature Stores: Specialized databases like Feast for managing and serving features to ML models.

Types of Databases

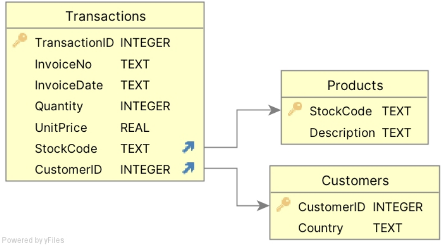


SQL Databases

SQL (Structured Query Language) Databases, also known as relational databases, are characterized by their structured format using tables.

- **ACID Properties:** Ensures database transactions are processed reliably (Atomicity, Consistency, Isolation, Durability).
- **Schema-based:** Requires a predefined schema for data storage, enhancing data integrity.
- **Relational Data Management:** Efficiently manages data across multiple tables through relationships.

SQL Databases



Transactions

🔑	TransactionID	INTEGER	
	InvoiceNo	TEXT	
	InvoiceDate	TEXT	
	Quantity	INTEGER	
	UnitPrice	REAL	
	StockCode	TEXT	↗
	CustomerID	INTEGER	↗

Products

🔑	StockCode	TEXT
	Description	TEXT

Customers

🔑	CustomerID	INTEGER
	Country	TEXT

Powered by yFiles

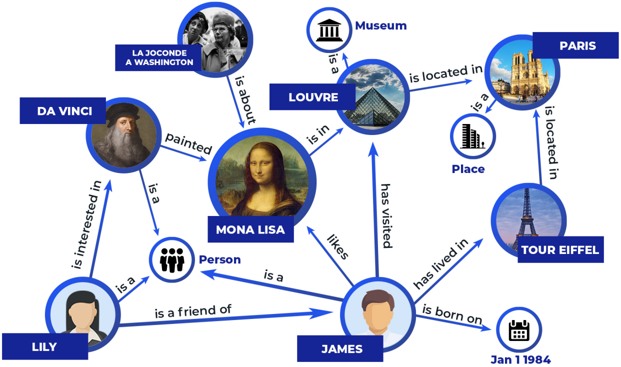
Example: 'db_create.py'

Knowledge Graph Databases

Knowledge Graph Databases are designed to store interlinked descriptions of entities — objects, events, or concepts — providing a framework that integrates data using a graph.

- **Graph Structure:** Data is represented as nodes (entities) and edges (relationships), enabling direct representation of relationships.
- **Semantic Querying:** Supports queries that understand the meaning of data, allowing for more nuanced data retrieval.
- **Schema-less or Schema-flexible:** Can evolve over time without requiring predefined schema modifications.

Knowledge Graph Database



How to get weather data from online API?

TASK: We want to create a simple web app showing recent and historical weather and air pollution data.

There are many websites but not all offer APIs, e.g., DMI.

We can get data for free on https://open-meteo.com/.

Exercise competition

You have 15 minutes to get data on Air Quality (PM 2.5) for Roskilde and Beijing

Hint: AirQuality API is in OtherAPIs :)