IS883: Deploying Generative AI

Mohannad Elhamod

BOSTON UNIVERSITY

Boston University Questrom School of Business

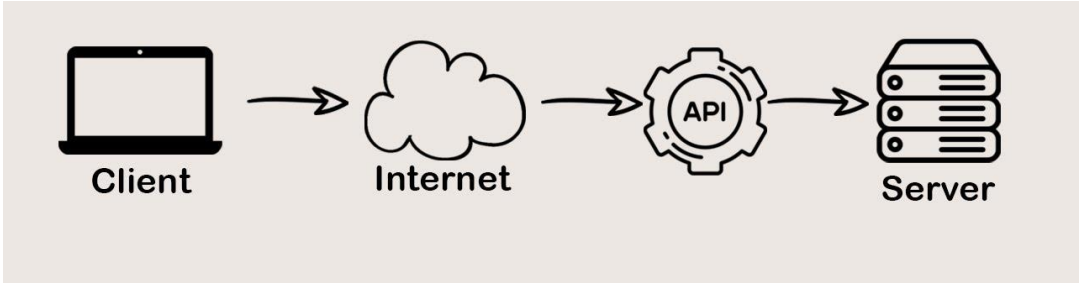
Using APIs

BOSTON UNIVERSITY

What is an API?

An Application Programming Interface (API) establishes a *"contract"* or *"glue"* that allows multiple pieces of software to communicate.

- Standardizes software development.
- Allows integration with online services.
- Is generally well-documented.



CodeWithMazn

LLM APIs

APIs that provide access to LLMs:

- You send a query/request (e.g., a prompt or a sentence)
- You get a response (e.g., a summary or a translation)

Why LLM APIs (vs. in-house)?

- Pros?
  - Less memory and compute resources needed.
  - Abstraction in terms of maintenance and updates.
  - Security?

- Cons?
  - Less control.
  - Maybe slower.
  - Cost.
  - Privacy.

OpenAI API

- Documentation
- Playground
- Guide (notebook)
- Pricing

Cloud Computing

BOSTON UNIVERSITY

Boston University Questrom School of Business

What is Cloud Computing?

The technology that allows us to access computing resources over the internet (as opposed to purchasing our own resources).

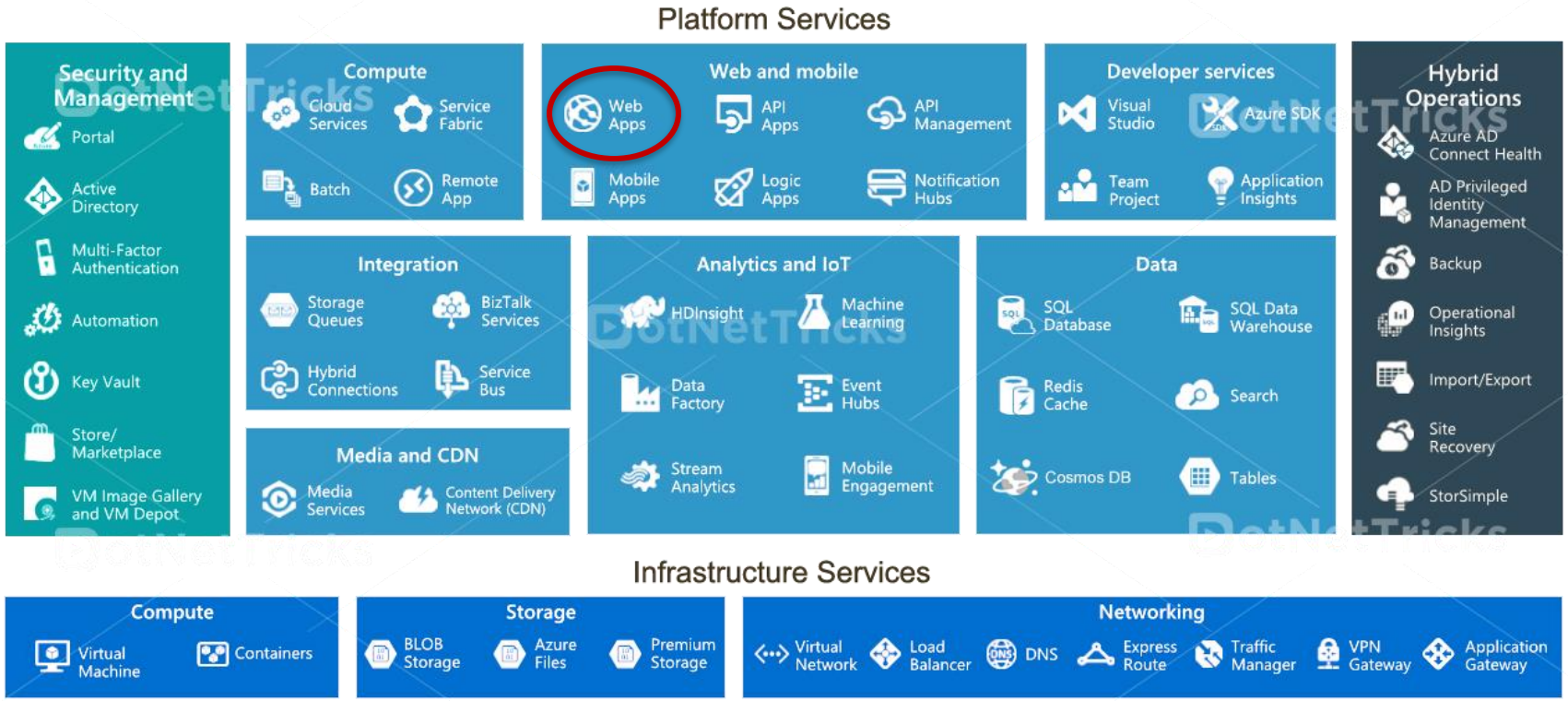
Google Cloud

Subscription vs. Buying

wearekemb

Google Cloud

Example: Microsoft Azure



Medium

Code Development Tools

BOSTON UNIVERSITY

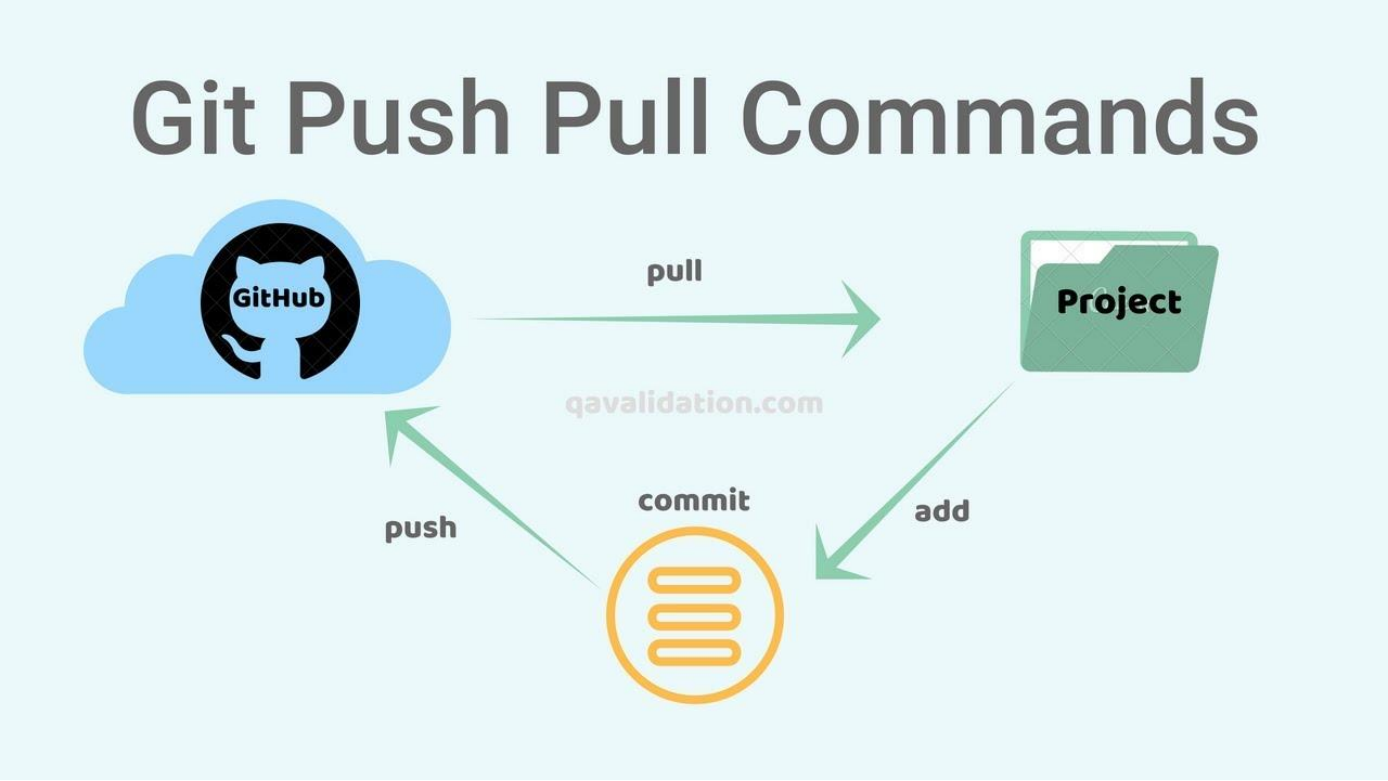
Boston University Questrom School of Business

GitHub

Widely used.

- Helps with team collaborations.
- Tracks code change history.
- Can automate tasks such as deployment.
- Allows multiple versions of the code (e.g., for different features).

The Git Workflow



In-Class Work

BOSTON UNIVERSITY