

**Project Management**

| **Term** | **Type** | **What It Is** | **Key Features** | **Example Use** |
| --- | --- | --- | --- | --- |
| **Agile** | Methodology | A mindset for building products in **small, iterative cycles** | Flexibility, collaboration, customer feedback | “Let’s build the product step by step.” |
| **Scrum** | Framework (Agile) | A **structured Agile framework** with roles & time-boxed sprints | Roles: Product Owner, Scrum Master, Team Events: Daily Standup, Sprint Planning | “Let’s plan 2-week sprints and review progress.” |
| **Kanban** | Method / Framework | A **visual workflow method** to track work continuously | Boards with columns (To Do → Doing → Done), WIP limits | “Let’s just move tasks as they progress.” |
| **Jira** | Tool (by Atlassian) | A project tracking tool that **supports Agile, Scrum, Kanban** | Backlogs, sprint boards, burndown charts, issue tracking | “Track our dev tasks and bugs in Jira.” |
| **Notion** | Tool (All-in-one) | A flexible tool for **notes, docs, tasks, databases** | Customizable pages, Kanban boards, wikis | “Plan tasks, take meeting notes, and share docs.” |



Webscraping – python-beautiful soup

**Data Preparation**

**Raw data - Cleaning, outliers, join, transformation,**

**Comply with GDPR for AI companies for PII information**

**EDA**

**Visualization**

Heat map to indicate which time frauds are happening

Histogram

Apache spark is used for distributing computing

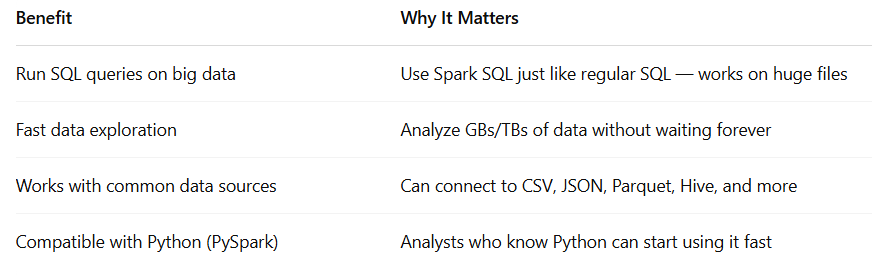
The process of dividing/distributing the work in different computers/nodes (map) and aggregate results (reduce)

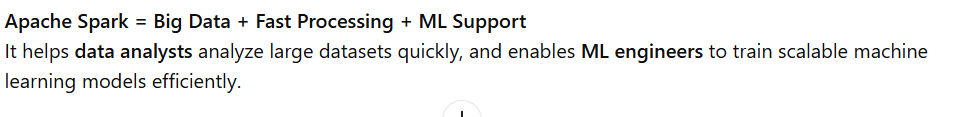
Pip install pyspark

**Apache Spark** is an **open-source, distributed computing system** that processes big data **super fast**. It can handle huge datasets across multiple machines

**Key Features:**

* **Fast** in-memory processing (100x faster than Hadoop in some cases)
* Supports **multiple languages**: Python (PySpark), Scala, Java, R
* Built-in libraries for:
  + **SQL** (Spark SQL)
  + **Machine Learning** (MLlib)
  + **Streaming Data** (Spark Streaming)
  + **Graph Processing** (GraphX)





**Feature Engineering**

Derived column, Scaling, encoding

**Model Selection and Training**

**Model Evaluation and Fine Tuning**

K-fold, Grid SearchCV, Randomized searchCV

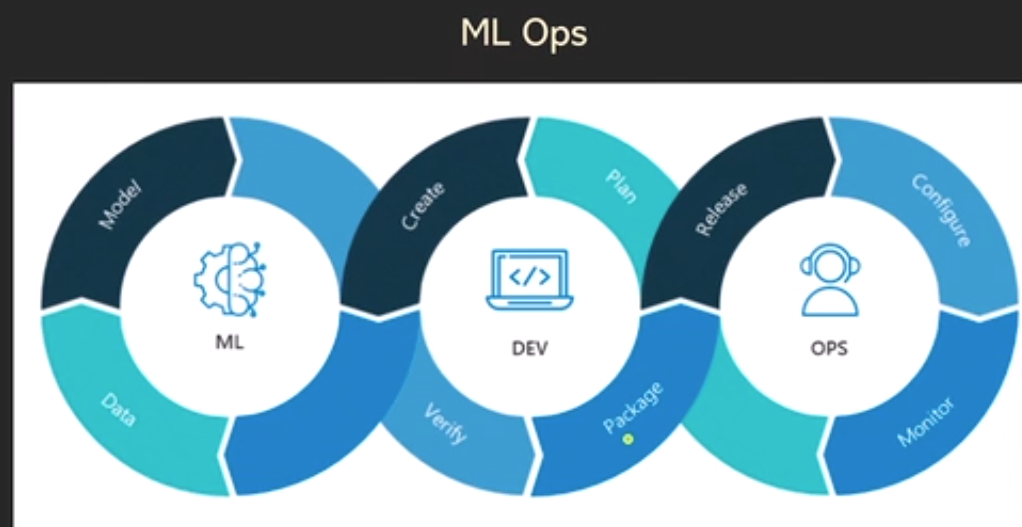
**Model Deployment**

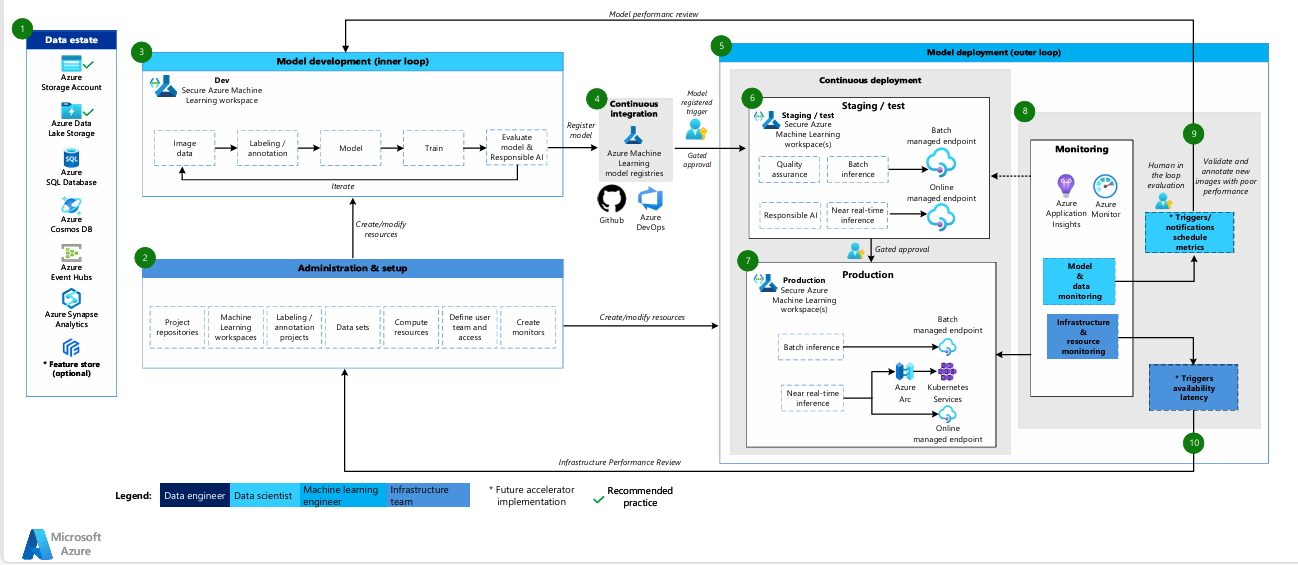
(server) fastAPI, flask , nodeS

Cloud providers : aws, azure, google cloud,

**Monitoring and Feedback using ML Ops**

Docker

****

****