10.04 Data Frames in Spark

What is Apache Spark?

- It is an open-source popular framework for Big Data processing and analysis

- It is written in Scala making it 10X faster than other frameworks

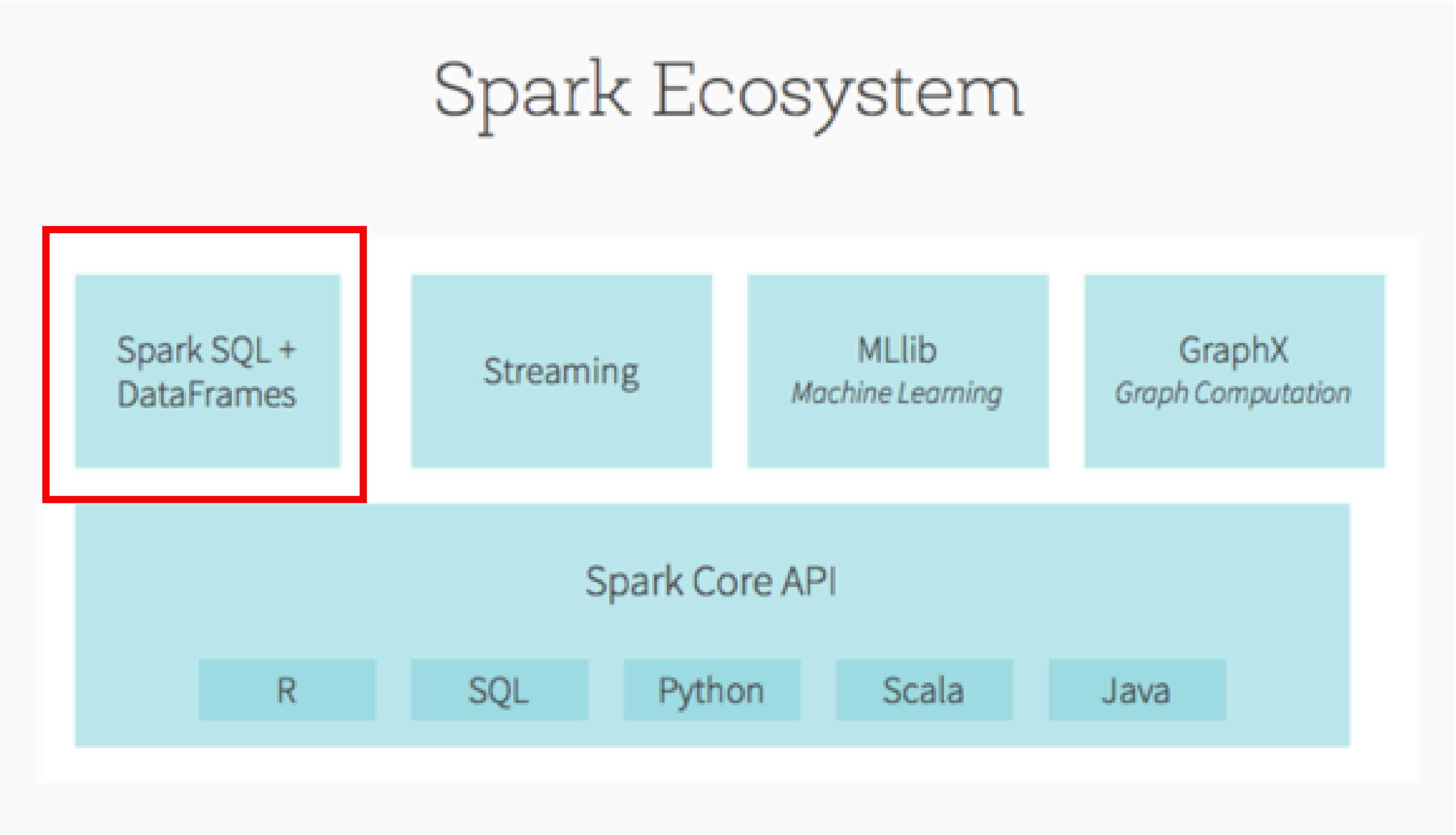
- It is a recommended platform for:
    - Stream Processing (Live Data)
    - Batch Processing
    - Large-scale SQL
    - Machine Learning

What is Apache Spark?

- Spark is a general-purpose, cluster computing framework that rapidly performs processing tasks with extensive datasets

- The framework can also distribute data processing tasks across many nodes, by itself or simultaneously with other distributed computing tools

What is Apache Spark?



Should I use Scala or Python with Spark?

1. Python is slower but very easy to use, while Scala is faster and moderately easy to use
2. Scala provides access to latest features of Spark, as Spark is written in Scala
3. Language choice for programming in Spark depends on features that best fit the project needs, as each one has its own pros and cons
4. Python is more analytical oriented while Scala is more engineering oriented but both are great languages for building Data Science applications
5. Overall, Scala would be more beneficial in order to utilise the full potential of Spark
    1. The complex syntax is worth learning if you really want to do out-of-the-box machine learning over Spark

Data Frames in Spark

- A DataFrame is a two-dimensional labeled data structure with columns of potentially different types

- You can think of a DataFrame like a spreadsheet, a SQL table, or a dictionary of series objects

- For more information and examples, see the Quickstart on the Apache Spark documentation website

Databricks Overview

- Founded by members of the original Spark development team
- Databricks offers an optimised version of Spark via a platform that unifies data science and data engineering
- It offers interactive notebooks, and it provides full enterprise security that any large organization would require
- You can also connect to other business intelligence tools using Databricks and because they're one of the main contributors you can be confident that anything that is created by the Databricks developers will later end up in Apache Spark's main code base

Databricks Clusters

- A Databricks cluster is a set of computation resources and configurations on which you run data engineering, data science, and data analytics workloads, such as production ETL pipelines, streaming analytics, ad-hoc analytics, and machine learning.

- You run these workloads as a set of commands in a notebook or as an automated job. Databricks makes a distinction between all-purpose clusters and job clusters. You use all-purpose clusters to analyze data collaboratively using interactive notebooks. You use job clusters to run fast and robust automated jobs.
  - You can create an all-purpose cluster using the UI, CLI, or REST API. You can manually terminate and restart an all-purpose cluster. Multiple users can share such clusters to do collaborative interactive analysis.
  - The Databricks job scheduler creates a job cluster when you run a job on a new job cluster and terminates the cluster when the job is complete. You cannot restart a job cluster.