Introduction to Apache Spark

After this video you will be able to..

- List the main motivations for the development of Spark
- Draw the Spark stack as a layer diagram
- Explain the functionality of the components in the Spark stack

Why Spark?

Hadoop MapReduce Shortcomings

restrict pipelines

Only for Map and Reduce based computations

Relies on reading data from HDFS

Native support for Java only

No interactive shell support

No support for streaming

spark was made to overcome this shortcomings and provide an expressive cluster computing environment

- interactive querying
- efficient iterative analytics
- streaming data processing

Basics of Data Analysis with Spark

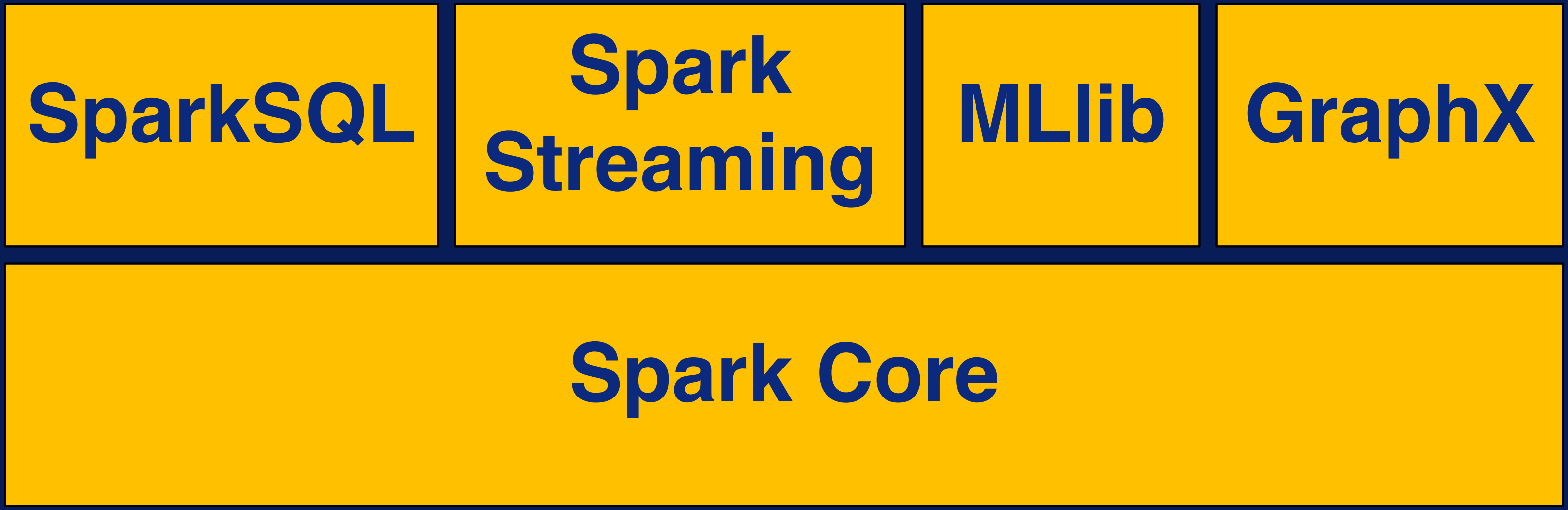
Expressive programing model

In-memory processing

Support for diverse workloads

Interactive shell

The Spark Stack



SparkSQL

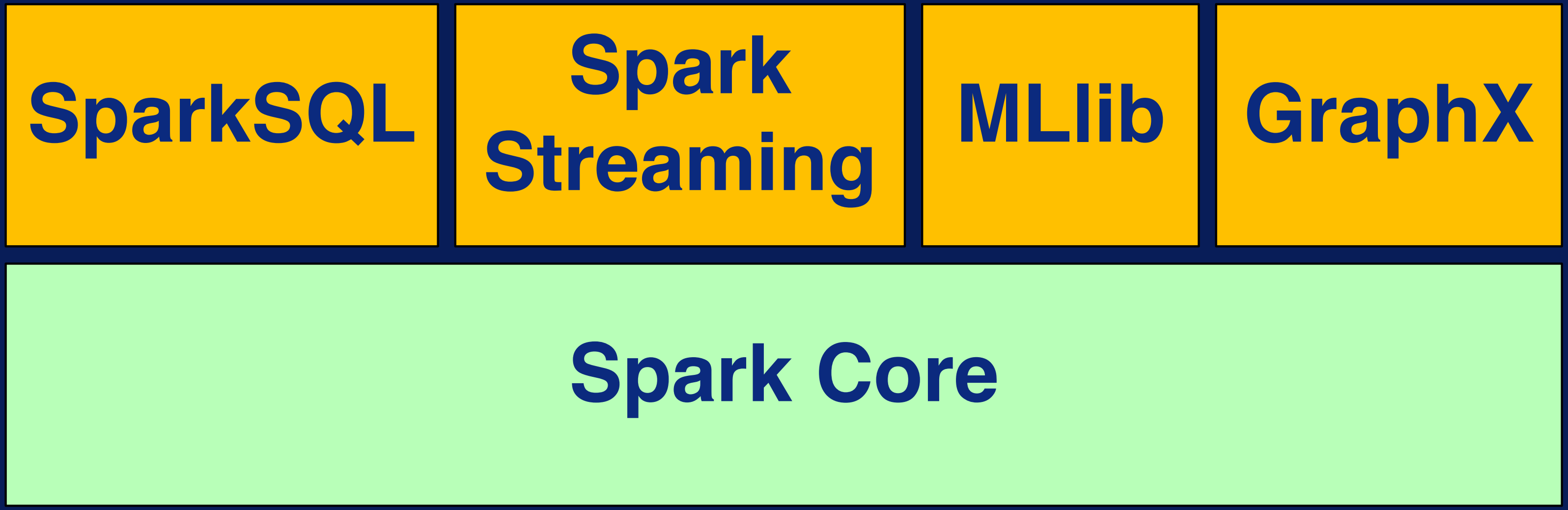
Spark Streaming

MLlib

GraphX

Spark Core

The Spark Stack


SparkSQL

Spark Streaming

MLlib

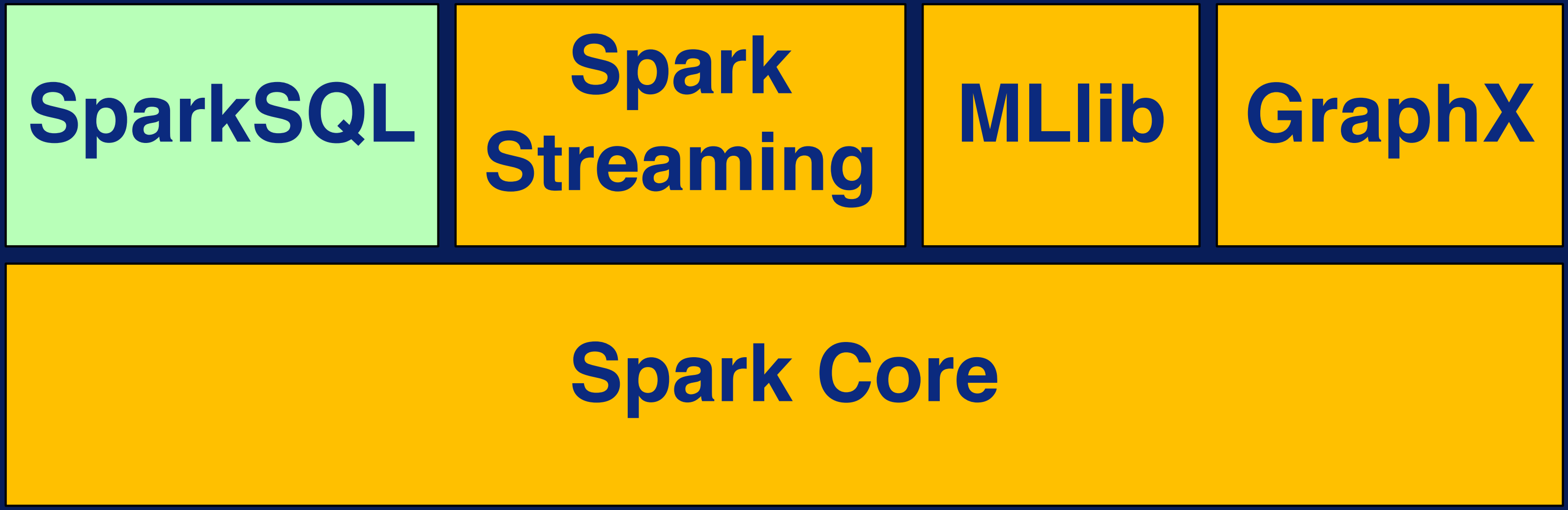
GraphX

Spark Core

- where the core capabilities of the spark framework are implemented
    - support for distributed scheduling
    - memory management
    - fault tolerance
- interaction with schedulers
- APIs for resilient distributed datasets (RDDs)

main programming abstraction

The Spark Stack



SparkSQL

Spark Streaming

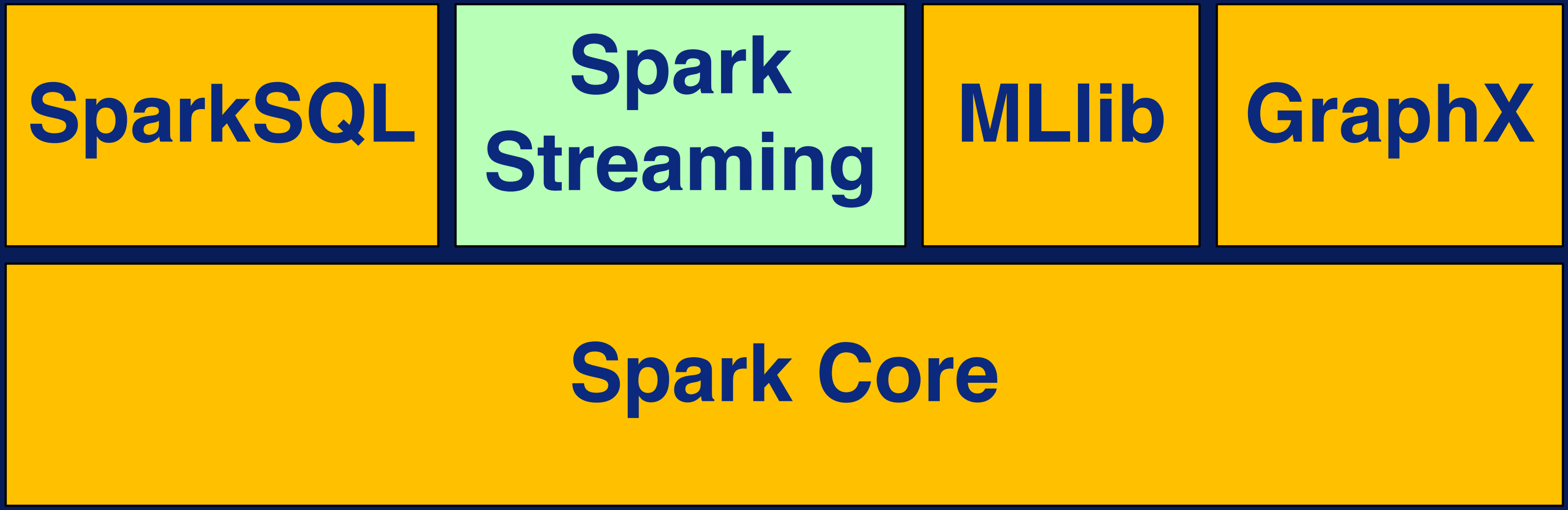
MLlib

GraphX

Spark Core

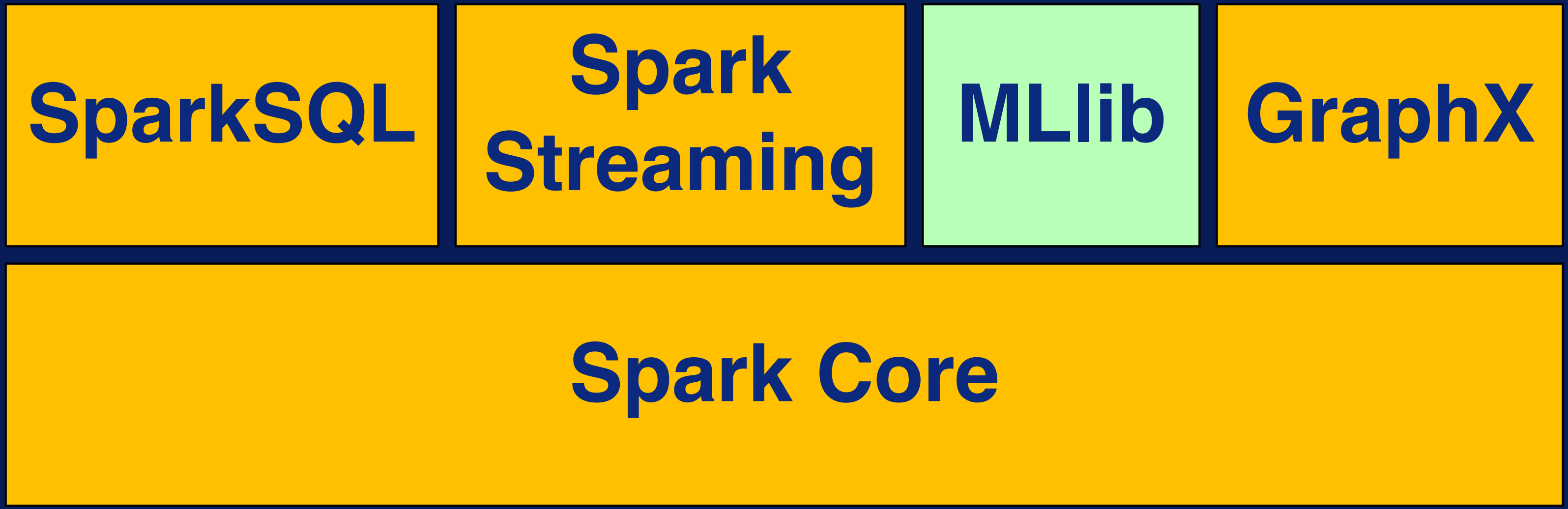
- provides querying structured and unstructured data through a common query languages
  - connects to many data sources
  - APIs to convert query results to RDDs

The Spark Stack



- for streaming data manipulations
- enables creating small aggregates of data incoming from streaming data ingestion systems (micro batches — aggregated datasets) ⇝ can be converted to RDD

The Spark Stack



SparkSQL

Spark Streaming

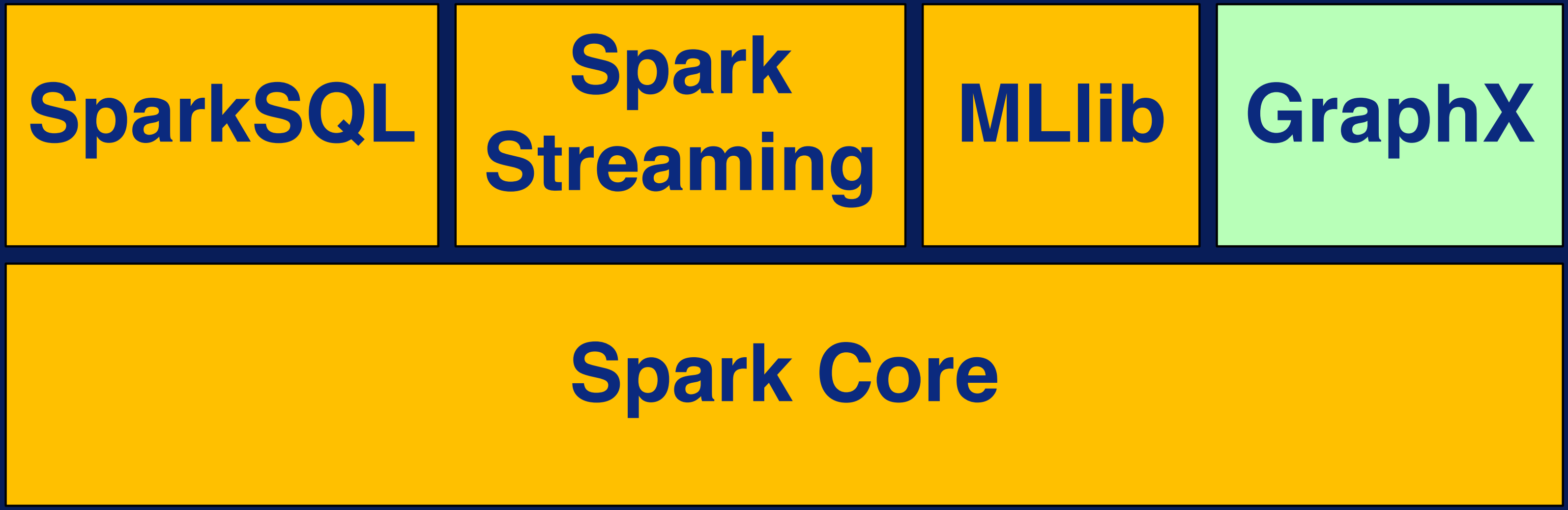
MLlib

GraphX

Spark Core

- spark's native library for machine learning algorithms as well as model evaluation
- designed to scale out using spark

The Spark Stack



SparkSQL

Spark Streaming

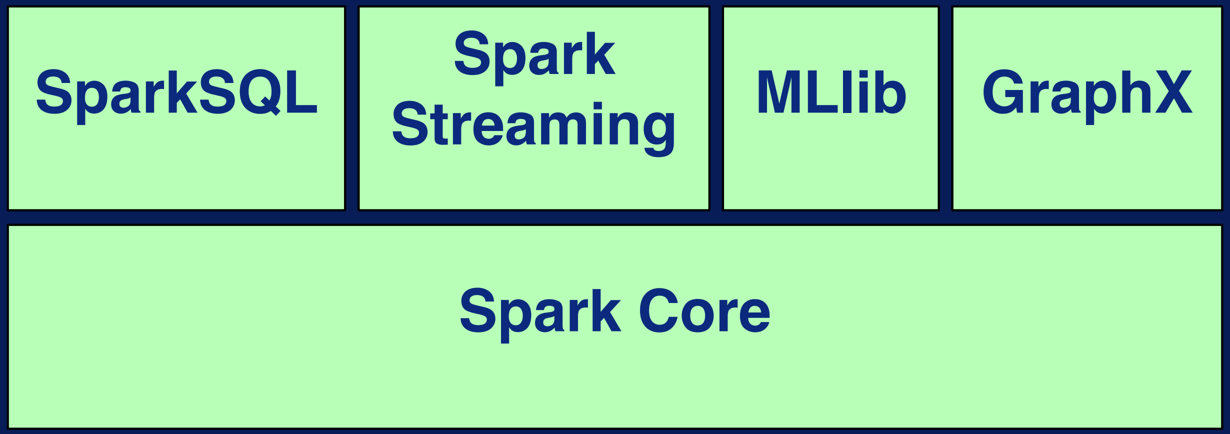
MLlib

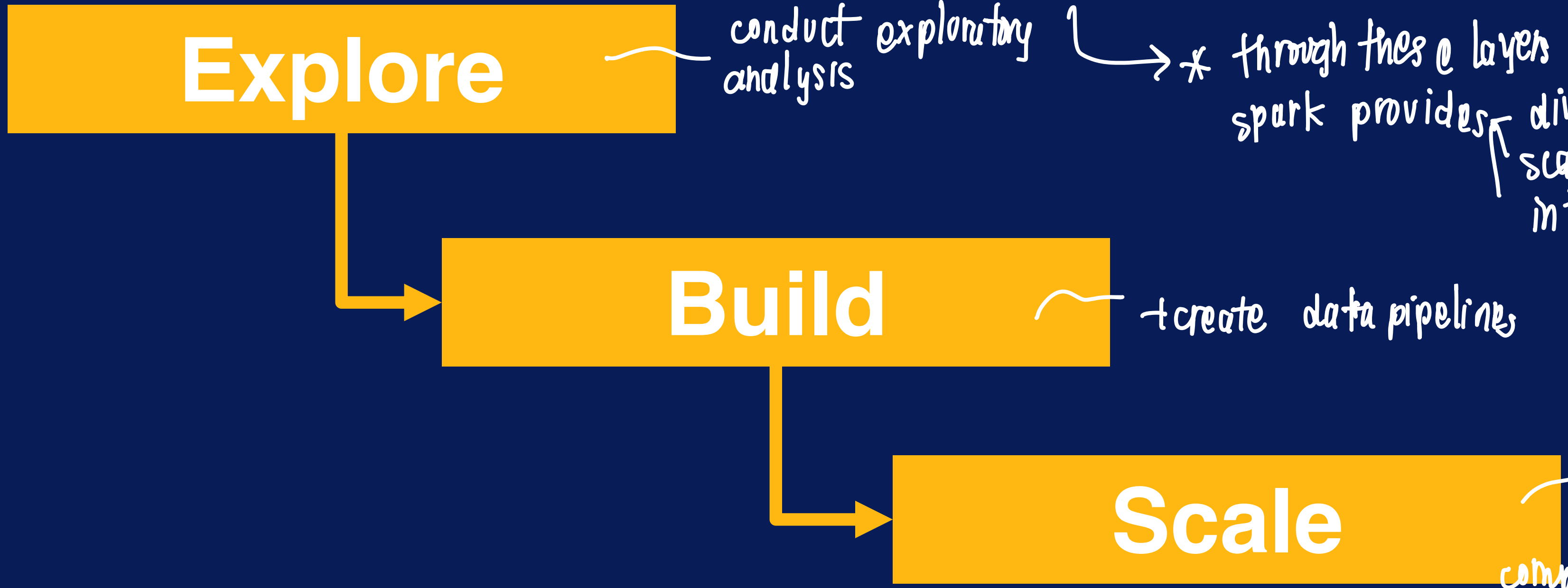
GraphX

Spark Core

- graph analytics library
- enables the vertex edge data model of graphs to be converted into RDDs

+ scalable implementations of a graph processing algorithms





conduct exploratory analysis

* through these layers spark provides diverse, scalable, interactive management and analysis (of big data

+create data pipelines

+ enabling scaling across commodity computing clusters & cloud environments