University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Program: MSc of Data Science
Module: Big Data Tools and Techniques

Week 4 – Part 1

Apache, Spark, Databricks?

2025

Ground Rules

1. Choose a quiet place to attend the class and please concentrate during the lecture
2. Put your questions in Padlet (not Teams' chat box) and I will review them in the due time (Padlet link is in the Bb, week 4, Lecture)
3. Turn off your mic during the lecture
4. We will have 5 mins break after the first hour of the lecture (please remind me)
5. Jisc code will be shared during the break time

What is Apache?

Apache is a name for a multidimensional project. The dimensions like:

Apache Spark, **Apache** Hadoop, **Apache** Hive, **Apache** Avro, **Apache** Ambari, **Apache** Kafka, …

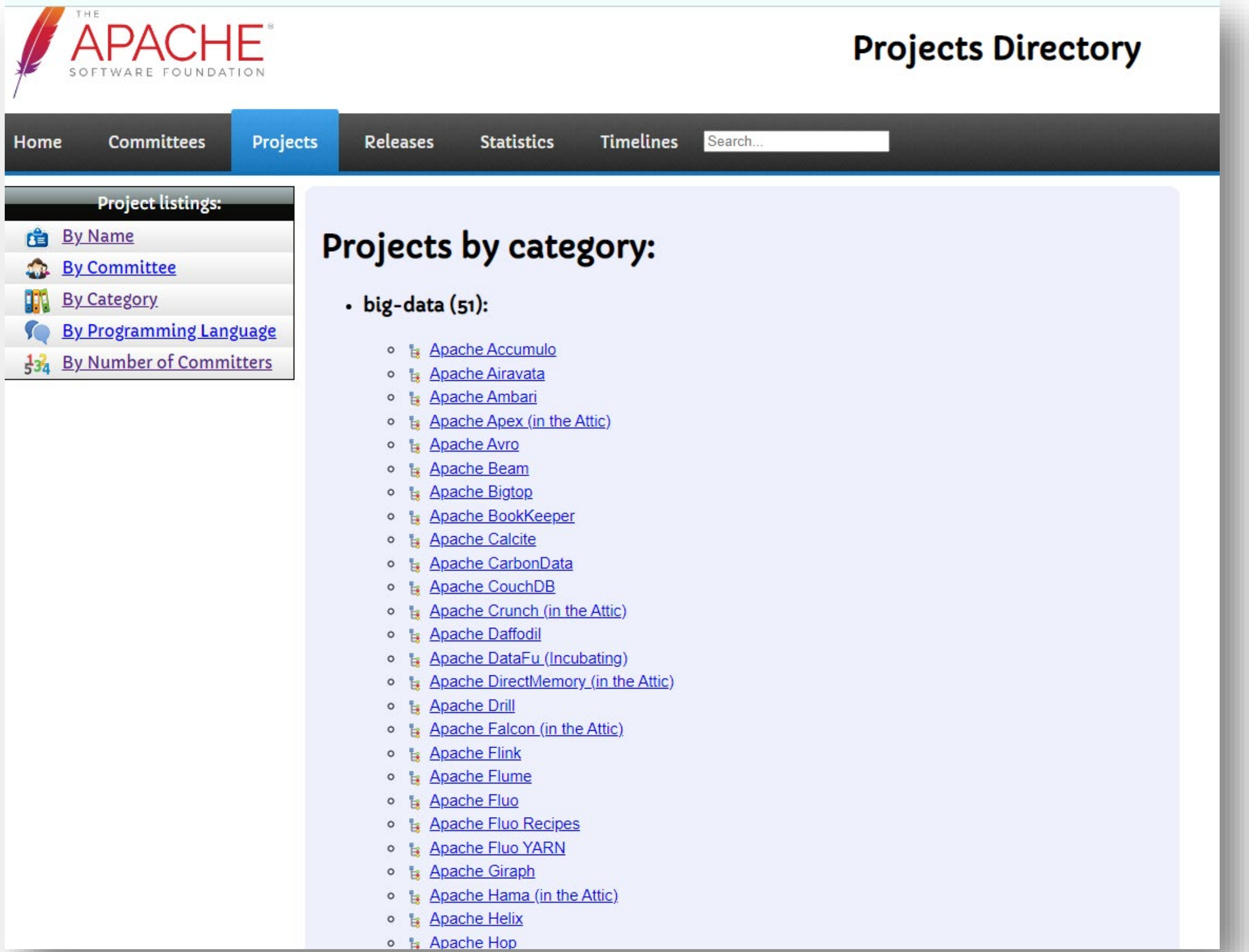
So, what is Apache exactly?

Apache project

Founded in 1999, the ASF functions as a charitable organisation in the United States, relying on contributions from individuals and corporate sponsors for funding. An all-volunteer board manages over 300 prominent Open-Source projects, among them the Apache HTTP Server, recognized as the foremost web server software globally.

https://www.apache.org/

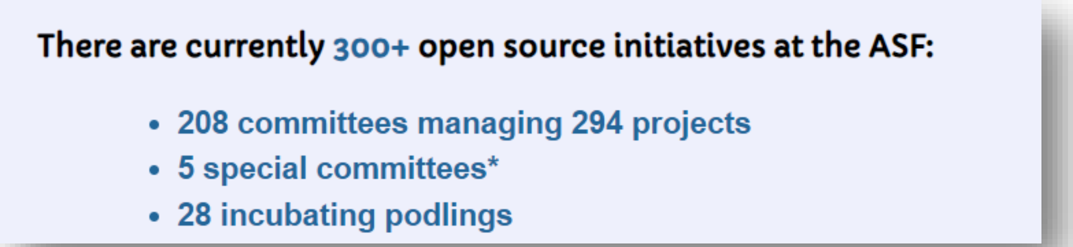
Apache Projects



Projects Directory

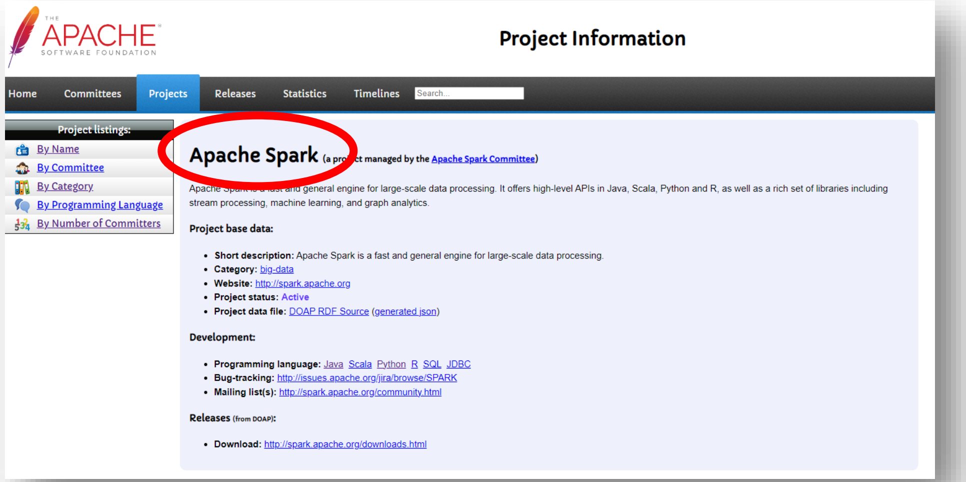
Projects by category:

- big-data (51):
  - Apache Accumulo
  - Apache Airavata
  - Apache Ambari
  - Apache Apex (in the Attic)
  - Apache Avro
  - Apache Beam
  - Apache Bigtop
  - Apache BookKeeper
  - Apache Calcite
  - Apache CarbonData
  - Apache CouchDB
  - Apache Crunch (in the Attic)
  - Apache Daffodil
  - Apache DataFu (Incubating)
  - Apache DirectMemory (in the Attic)
  - Apache Drill
  - Apache Falcon (in the Attic)
  - Apache Flink
  - Apache Flume
  - Apache Fluo
  - Apache Fluo Recipes
  - Apache Fluo YARN
  - Apache Giraph
  - Apache Hama (in the Attic)
  - Apache Helix
  - Apache Hop



There are currently 300+ open source initiatives at the ASF:

- 208 committees managing 294 projects
- 5 special committees*
- 28 incubating podlings



Project Information

Apache Spark (a project managed by the Apache Spark Committee)

Apache Spark is a fast and general engine for large-scale data processing. It offers high-level APIs in Java, Scala, Python and R, as well as a rich set of libraries including stream processing, machine learning, and graph analytics.

Project base data:

- **Short description:** Apache Spark is a fast and general engine for large-scale data processing.
- **Category:** big-data
- **Website:** http://spark.apache.org
- **Project status:** Active
- **Project data file:** DOAP RDF Source (generated json)

Development:

- **Programming language:** Java Scala Python R SQL JDBC
- **Bug-tracking:** http://issues.apache.org/jira/browse/SPARK
- **Mailing list(s):** http://spark.apache.org/community.html

Releases (from DOAP)**:**

- **Download:** http://spark.apache.org/downloads.html

https://projects.apache.org/projects.html?category

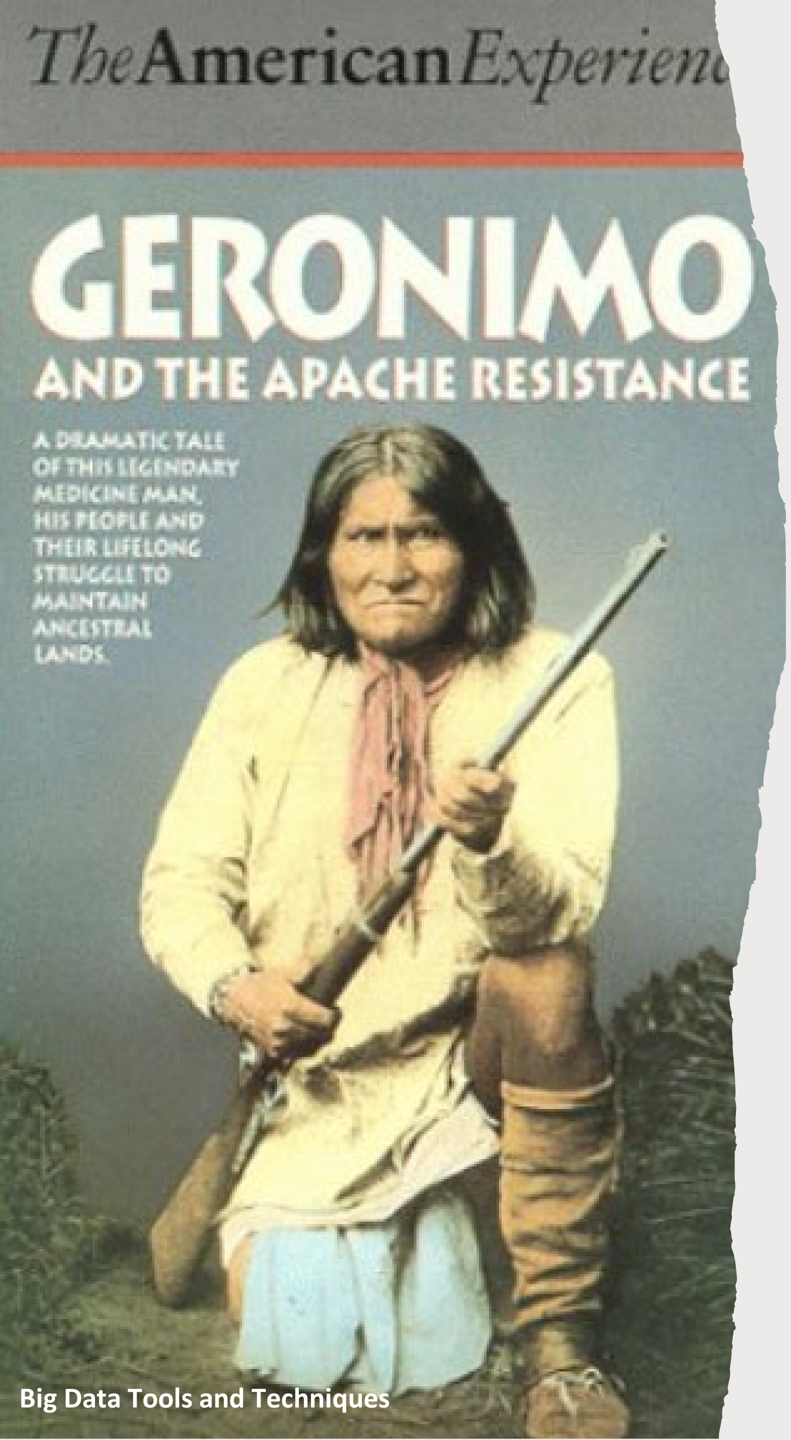
Why has been named "Apache"

A group of people calling themselves the "Apache Group" created the foundation in 1999. They had come together several years earlier, **to continue to support and maintain the HTTPD web server** written by the National Center for Supercomputing Applications (NCSA) at the University of Illinois.

That server was freely available, came with its source code, its license allowed very open modification and redistribution, but the original developers lost interest in that project and moved onto something else, leaving users with no support for the application.

Some of those users started to exchange fixes (called "patches") and information on how to prevent problems and improve the existing software. Brian Behlendorf created a mailing list on which those users could collaborate to fix, maintain and improve that software.

https://www.apache.org/foundation/how-it-works.html#what


The American Experienc
GERONIMO
AND THE APACHE RESISTANCE
A DRAMATIC TALE OF THIS LEGENDARY MEDICINE MAN, HIS PEOPLE AND THEIR LIFELONG STRUGGLE TO MAINTAIN ANCESTRAL LANDS.

Why has been named “Apache”

Co-founder Brian Behlendorf states how the name Apache was chosen:

"I had just seen a documentary about Geronimo and the last days of a Native American tribe called Apaches, right, who succumbed to the invasion from the West, from the United States, and **they were the last tribe to give up their territory and for me that almost romantically represented what I felt we were doing with this web-server project** ..."

What is Apache Spark?

Apache Spark?

Imagine you have a giant **puzzle with millions of pieces**, each representing a piece of information. Analysing all that data by hand would be impossible! That's where Spark comes in.

Think of Spark as a superpowered puzzle solver. It's a tool that can:

- Work on a single computer or many computers working together (like a team, depending on the size of the puzzle.)
- Hold some of the data in its memory (like remembering where you saw a specific colour or shape, making it much faster than searching through the whole puzzle each time.)
- Break the data down into smaller parts, allowing it to analyse different sections simultaneously (like having multiple people working on different areas of the puzzle at once.)
- Do different things with the data pieces (like sorting them by colour, shape, or size,
- or even using them to build new things, like a picture or a story.)

Apache Spark?

Is Spark a programming language? or a platform? or an architecture? what is it exactly?

Apache Spark?

Spark isn't exactly one of those categories on its own. It's more like a combination of them, serving as a powerful tool in the big data world:

It's not a programming language: While you can interact with Spark using various languages like Python, Scala, and Java, Spark itself isn't a language you directly write code in.

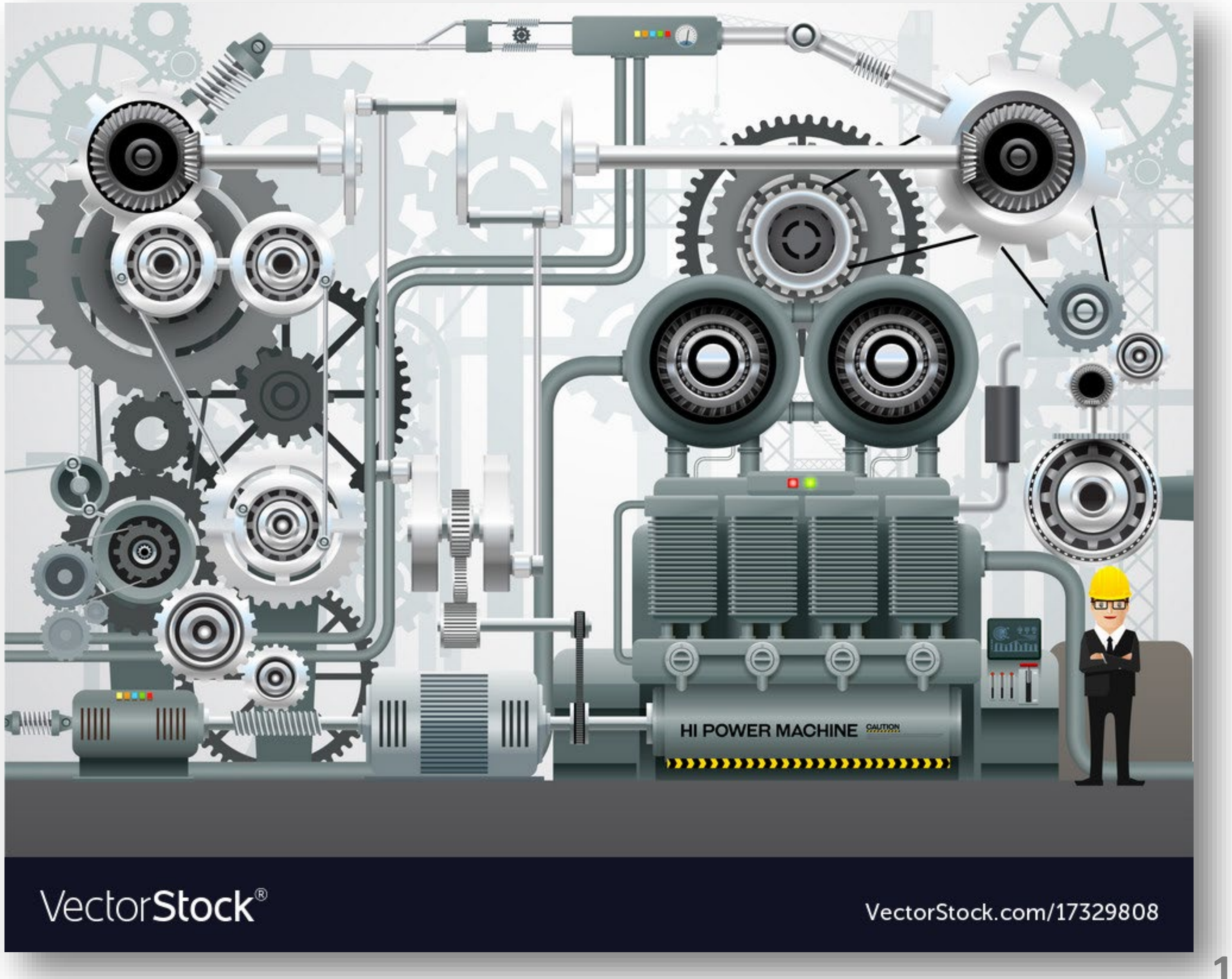
It's not just a platform: Although platforms like Databricks offer ways to use and manage Spark, Spark exists independently and can be implemented in various environments.

It's not just an architecture: While Spark utilizes a distributed architecture across clusters or single machines, the core is more than just its structure.

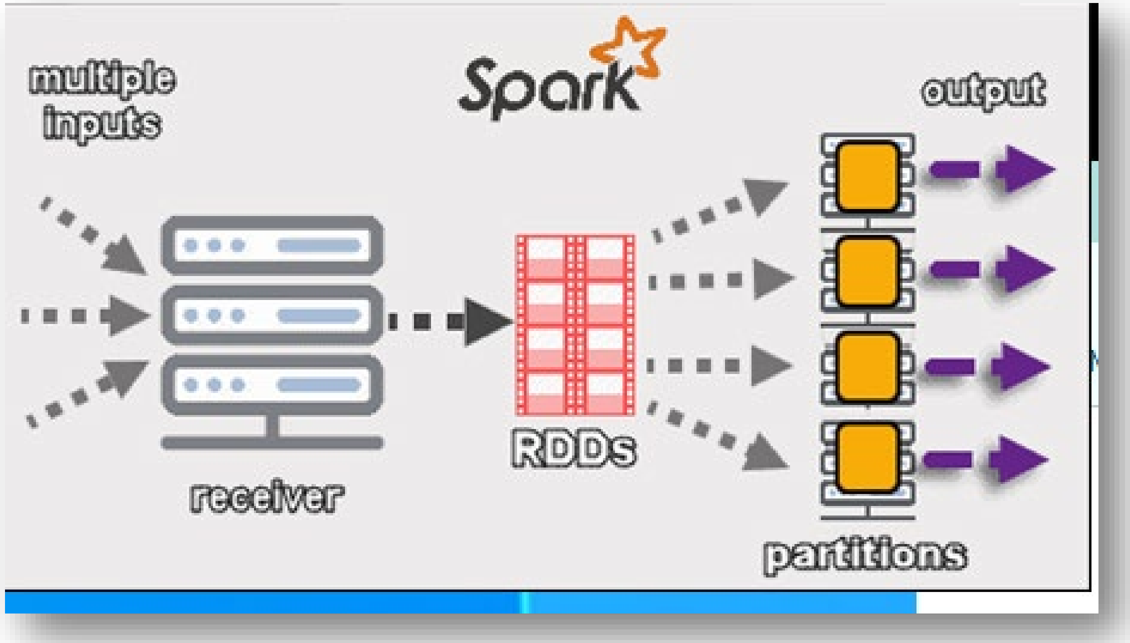
Apache Spark?

Spark is a unified analytics engine

It's a software framework specifically designed to perform large-scale data processing efficiently across various domains like data engineering, data science, and machine learning.



Apache Spark



What is Apache Spark?
Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning even on single-node machines or clusters.

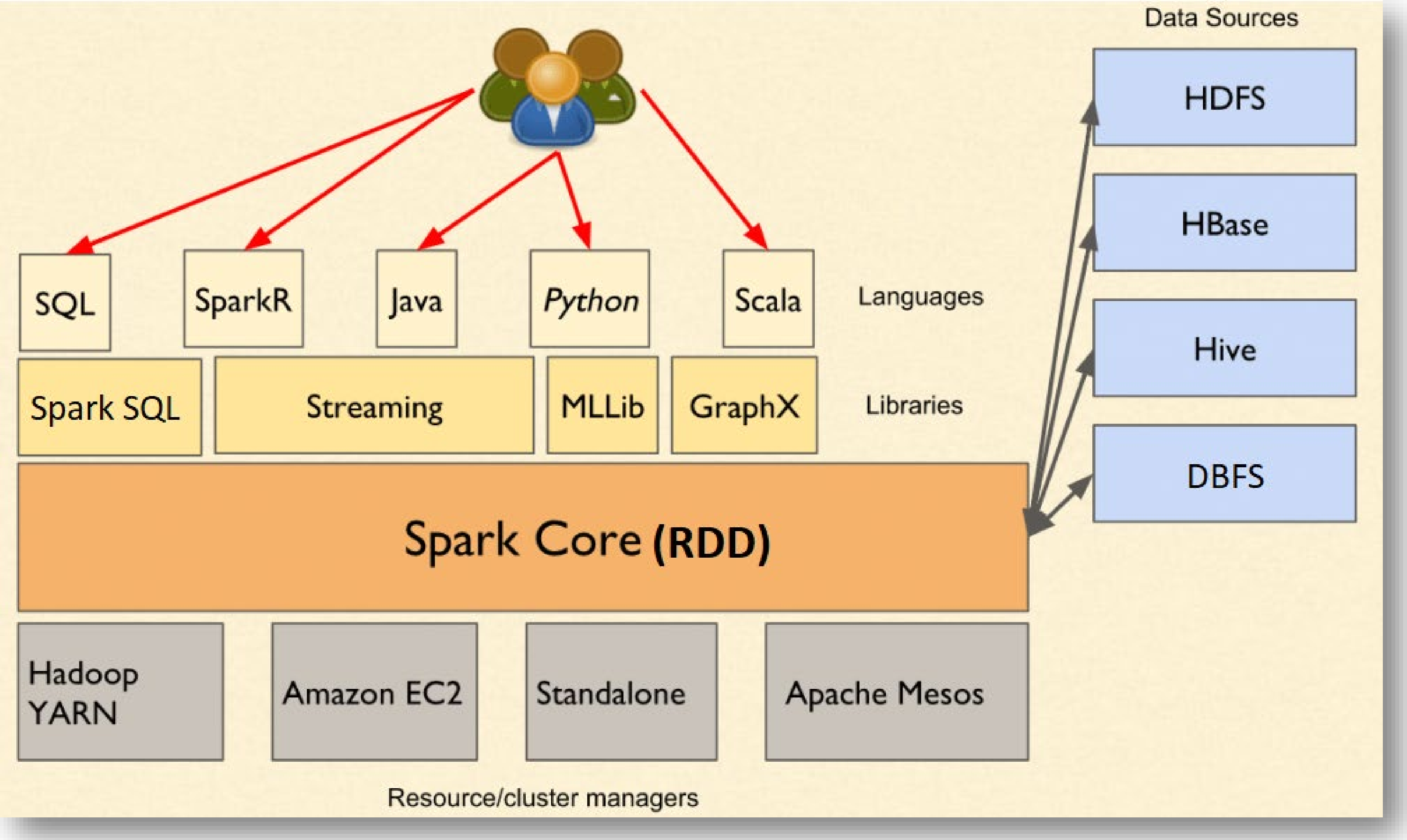
RDD is the main concept of the Apache Spark

The primary abstraction of **the Spark is the concept of RDD, which Spark uses to achieve faster and efficient MapReduce operations**.

Resilient Distributed Dataset (**RDD**) is the fundamental data structure of Spark.

https://spark.apache.org/

Apache Spark Structure



What is Databricks?

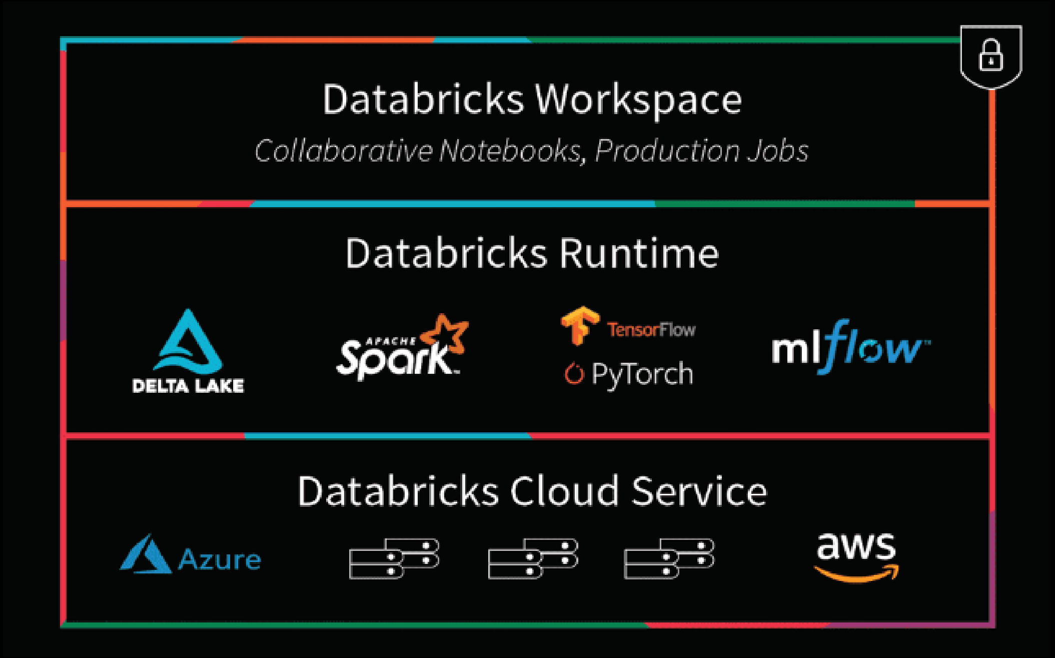
Comparing Apache Spark™ and Databricks

Apache Spark capabilities provide speed, ease of use and breadth of use benefits and include APIs supporting a range of use cases:

- Data integration and ETL
- Interactive analytics
- Machine learning and advanced analytics
- Real-time data processing

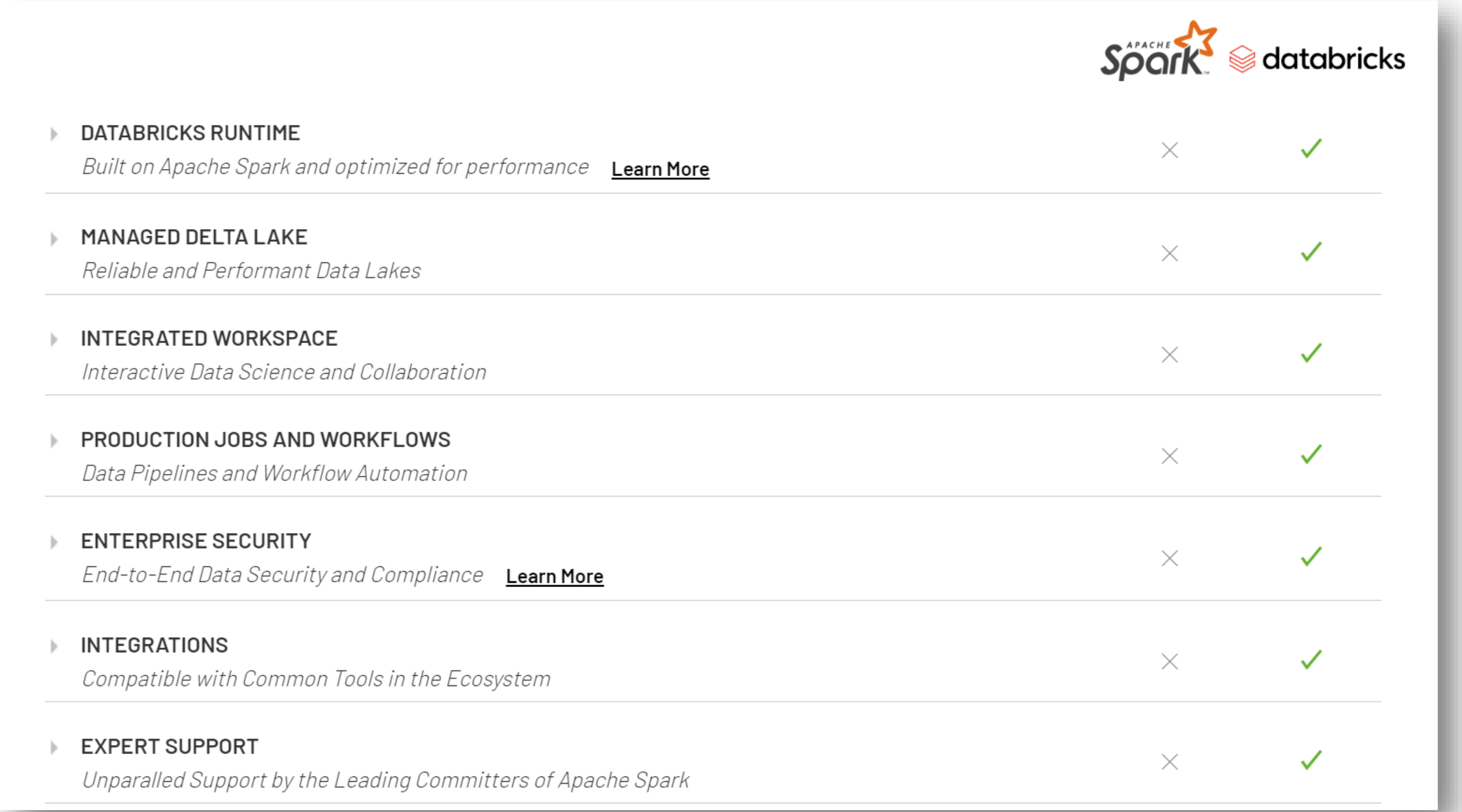
Databricks builds on top of Spark and adds:

- Highly reliable and performant data pipelines
- Productive data science at scale



Want to learn more? Visit our **platform page**.

Why use Databricks not plain Apache Spark!!



	Apache Spark	databricks
DATABRICKS RUNTIME *Built on Apache Spark and optimized for performance* Learn More	×	✓
MANAGED DELTA LAKE *Reliable and Performant Data Lakes*	×	✓
INTEGRATED WORKSPACE *Interactive Data Science and Collaboration*	×	✓
PRODUCTION JOBS AND WORKFLOWS *Data Pipelines and Workflow Automation*	×	✓
ENTERPRISE SECURITY *End-to-End Data Security and Compliance* Learn More	×	✓
INTEGRATIONS *Compatible with Common Tools in the Ecosystem*	×	✓
EXPERT SUPPORT *Unparalled Support by the Leading Committers of Apache Spark*	×	✓

https://www.databricks.com/spark/comparing-databricks-to-apache-spark

What is Databricks?

Databricks is an enterprise software company that provides Data Engineering tools for **Processing** and **Transforming** huge volumes of data to build machine learning models. Traditional Big Data processes are not only sluggish to accomplish tasks but also consume more time to set up clusters using Hadoop. However, Databricks is built on top of distributed Cloud computing environments like **Azure, AWS,** or **Google Cloud** that facilitate running applications on CPUs or GPUs based on analysis requirements. Databricks platform is said to be **100 times faster** than **Apache Spark.**

https://hevodata.com/learn/what-is-databricks/

What is Databricks?

Let us start by answering this main question of What is Databricks. Databricks, developed by the creators of **Apache Spark**, is a Web-based platform, which is also a **one-stop product** for all Data requirements, like Storage and Analysis. It can derive insights using SparkSQL, provide active connections to visualization tools such as **Power BI, Qlikview, and Tableau,** and build Predictive Models using **SparkML**. Databricks also can create interactive **displays, text,** and **code** tangibly. Databricks is an alternative to the **MapReduce** system.

What have you learnt so far?

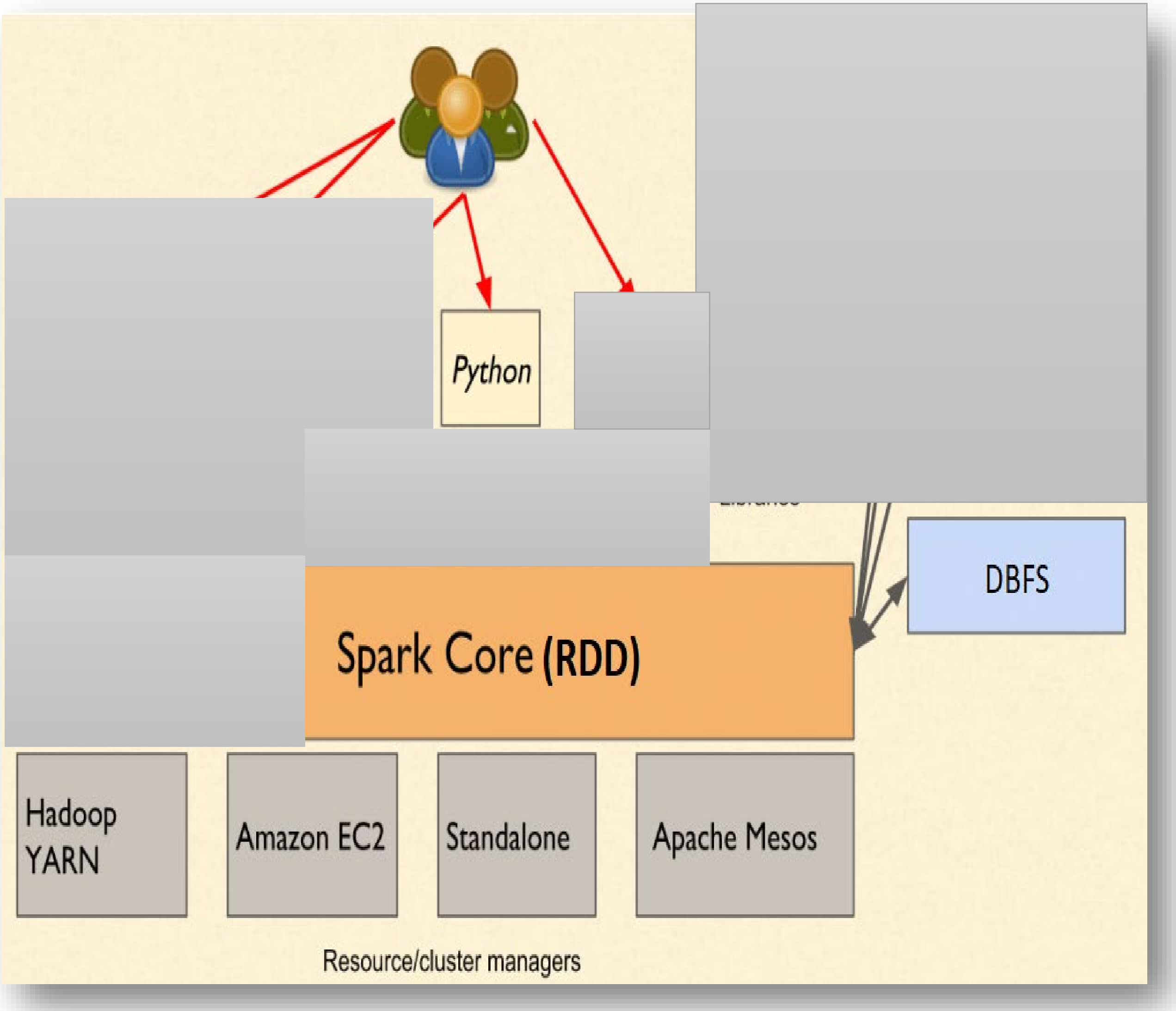
What you have learnt so far?

What you have learnt so far?

You have used **Python** to communicate with **Spark Core** inside the **Databricks** platform.

You have uploaded Data sets into **DBFS** and then you created many **RDDs**.

You have done some **Big Data analysis in its early stages** like extracting RDDs containing parts of the data sets that matter to you.



What will you learn after this?

What will you learn?

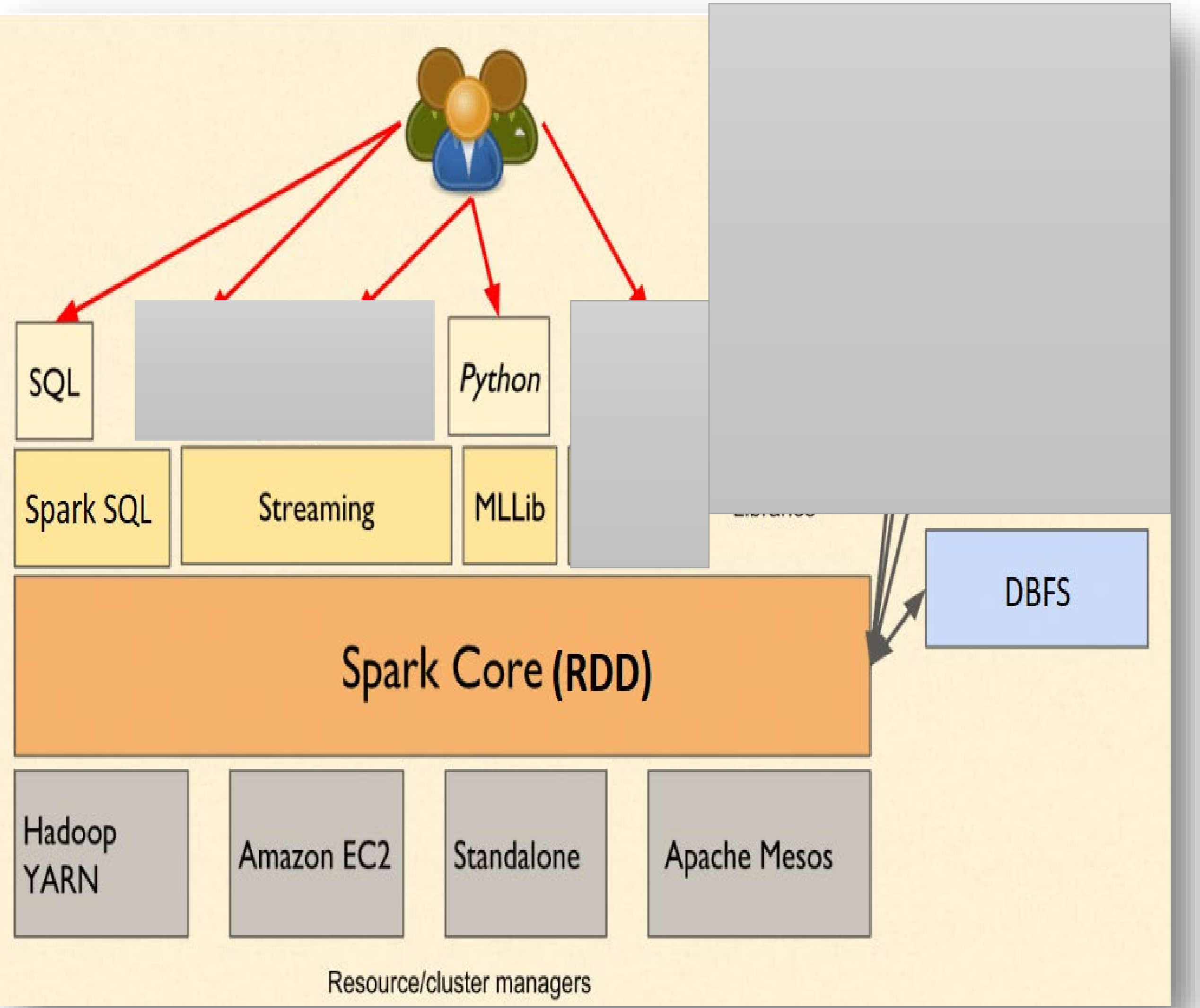
What you will learn in BDTT module.

You will use **Python** and **SQL** to work with Spark SQL, Spark Streaming and MLib.

You will create and analyse data frames (instead of RDDs)

You will execute SQL statements against data sets.

You will upload batch data into **DBFS** or bring **stream data** into DBFS and then doing advance **Big Data analysis.**


SQL
Python
Spark SQL
Streaming
MLLib
DBFS
Spark Core (RDD)
Hadoop YARN
Amazon EC2
Standalone
Apache Mesos
Resource/cluster managers