Master Executive di II Livello

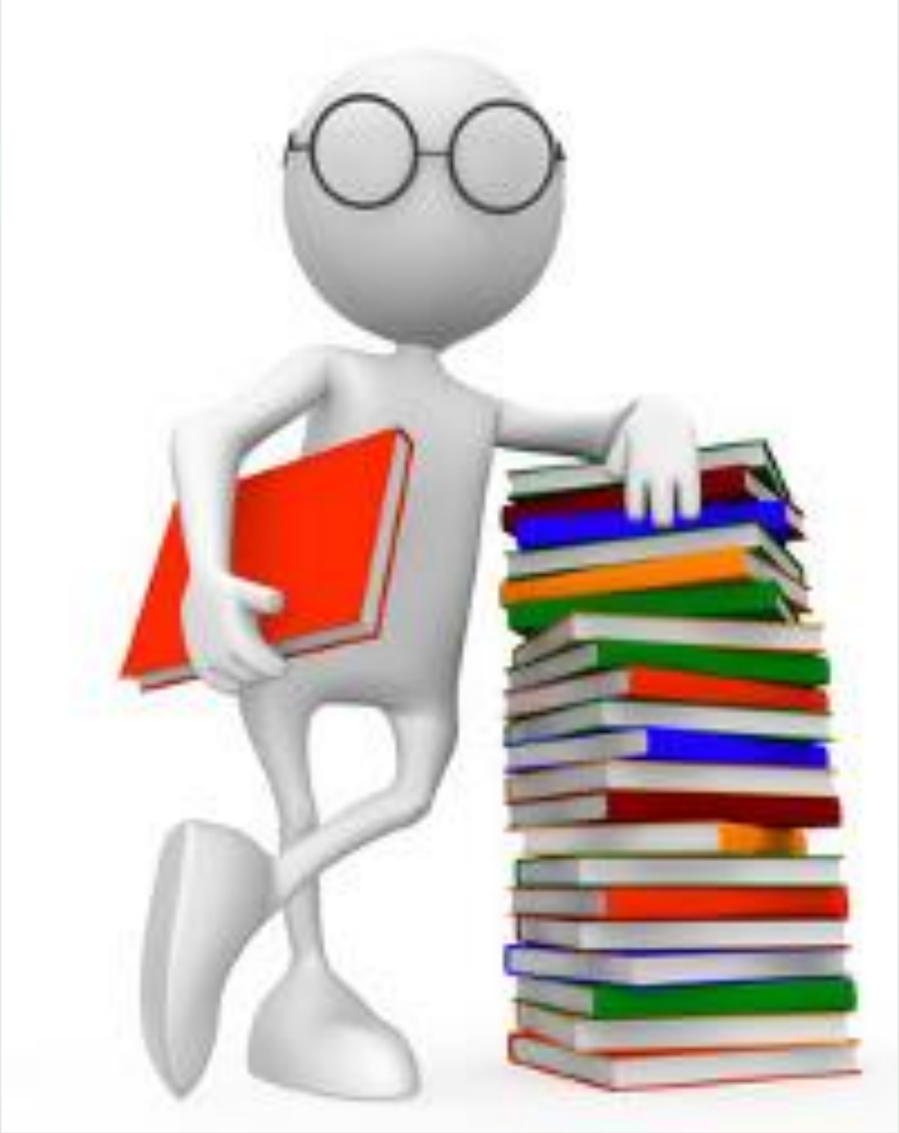
BIG DATA ANALYSIS AND BUSINESS INTELLIGENCE

Introduction to Spark

Vamsi Krishna Varma Gunturi
Data science intern at ISTAT
vamsivarmagunturi@gmail.com

fondazione INUIT TOR VERGATA

Topics



- What is Spark ?
- Key features of Spark
- Spark vs MapReduce
- Spark architecture and components
- Spark streaming
- Spark SQL
- Spark MLLib
- GraphX
- Spark with Hadoop
- Spark in different industries
- Spark applications
- Who is using Spark ?

What is Spark ?

- Apache Spark is an open-source distributed general-purpose cluster-computing framework.
- Write your SPARK scripts using either Python or Java or the Scala programming language, Scala being preferred.
- Extremely fast, 100x faster than traditional Hadoop approach as it relies on in-memory computing.
- Very versatile it can do things like handle SQL queries that can do machine learning across an entire cluster of information
- Can also handle streaming data in real time and also graph analysis.

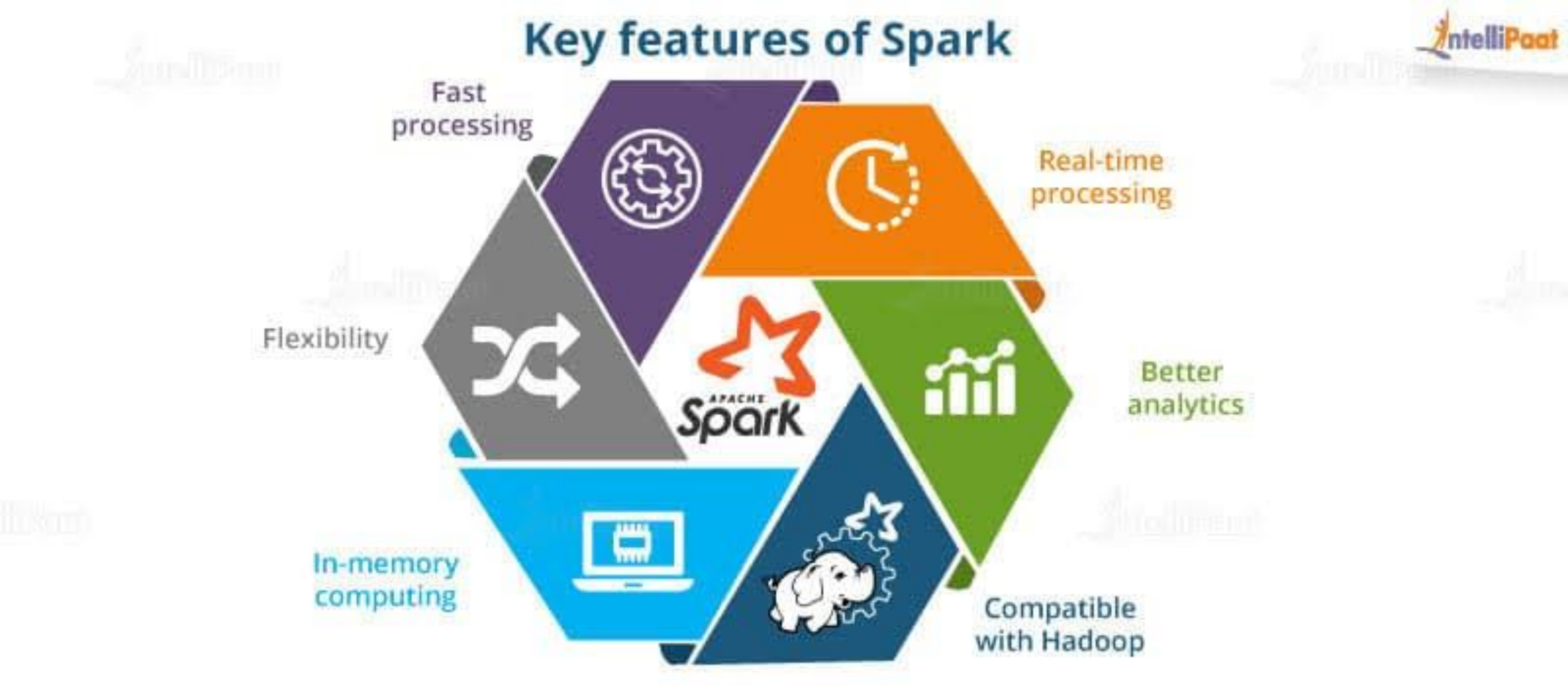
What is Spark ?

- Spark and its RDDs were developed in 2012 in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs.

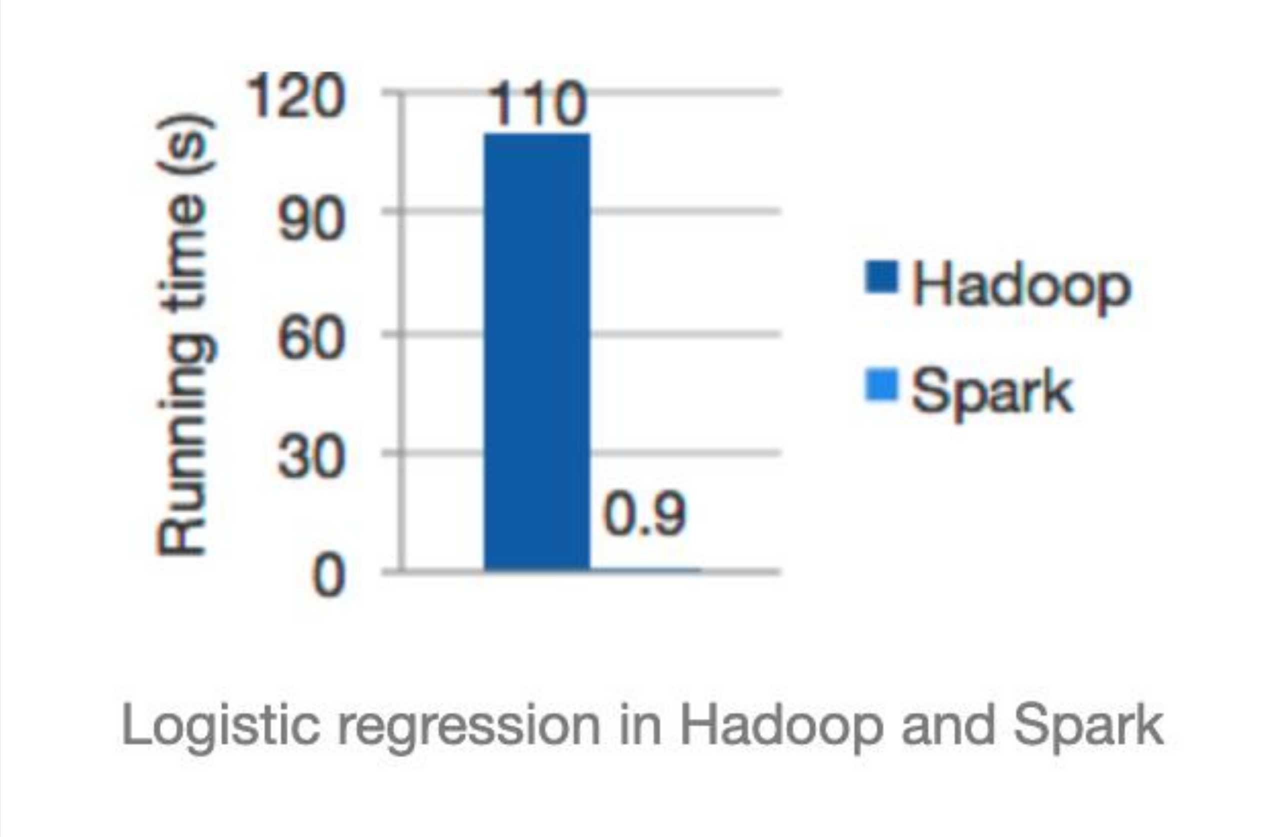
- Spark's RDDs function as a working set for distributed programs that offers a (deliberately) restricted form of distributed shared memory.

- It became one of the largest open source communities that includes over 200 contributors. The prime reason behind its success was its ability to process heavy data faster than ever before.

Introduction to Spark



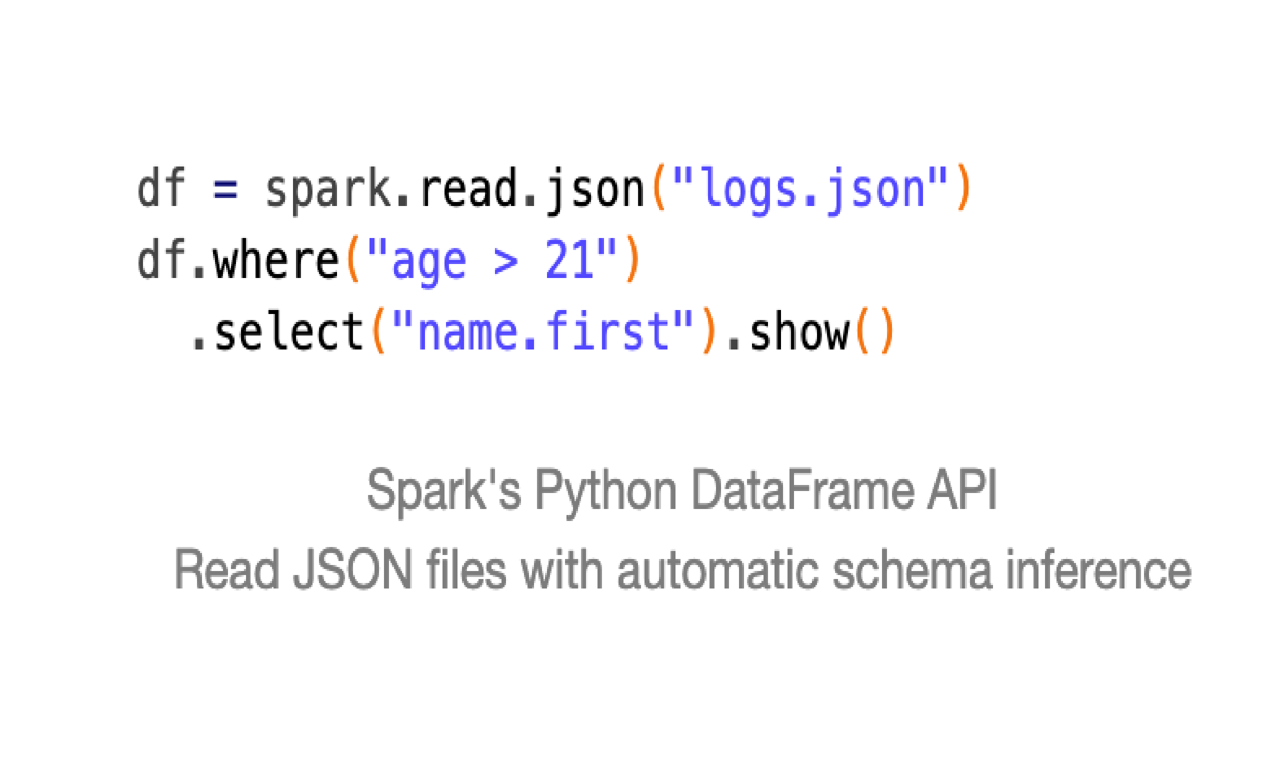
In-memory computing



Logistic regression in Hadoop and Spark

- Spark stores the data in the RAM of servers which allows him to access it quickly and in turn accelerating the speed of analytics.
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.
- Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.
- DAG engine (directed acyclic graph) optimizes workflows.
- Spark facilitates the implementation of both iterative algorithms and interactive/exploratory data analysis. Due to this latency is reduced by several orders of magnitude compared to Hadoop MapReduce which follow linear flow.

Flexibility

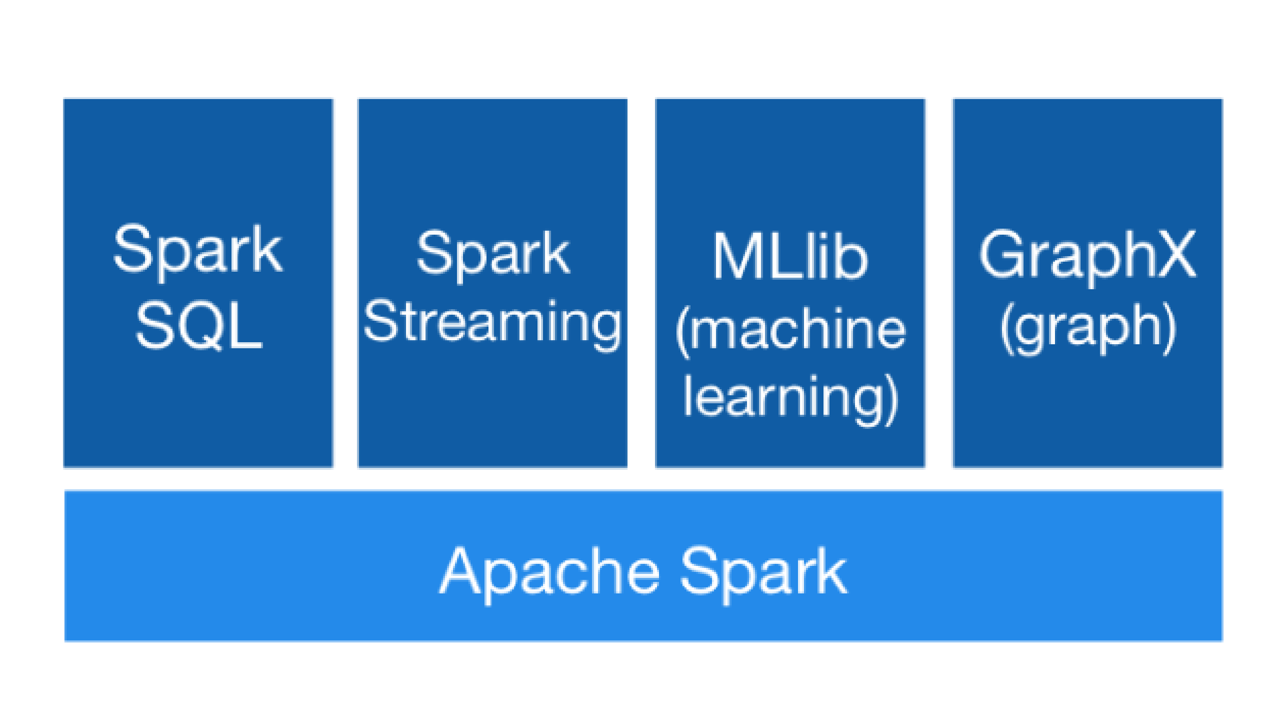


```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

- Write applications quickly in Java, Scala, Python, R, and SQL.
- Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala, Python, R, and SQL shells.

Generality



- Combine SQL, streaming, and complex analytics.
- Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

Runs Everywhere

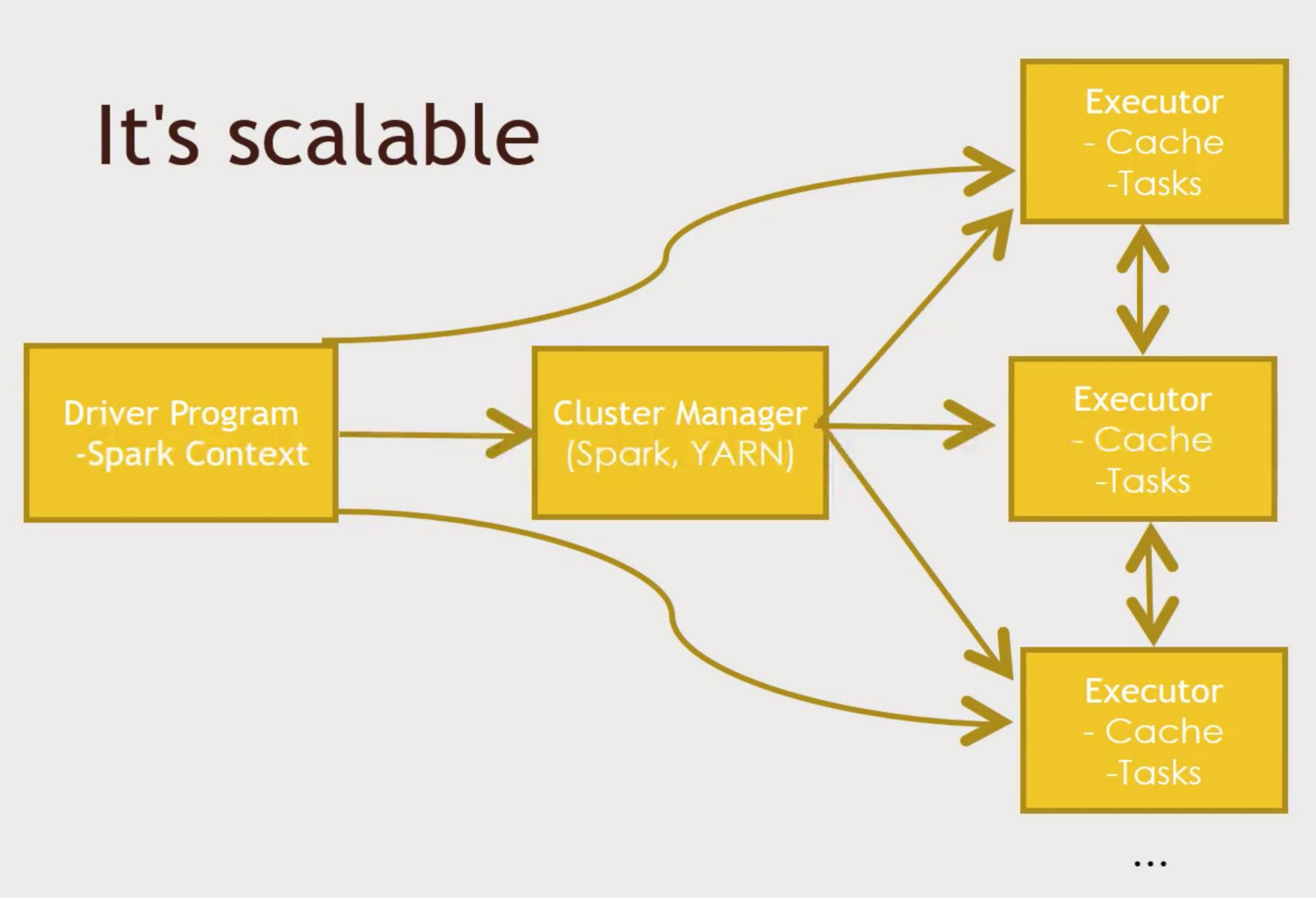
- Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.
- You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes. Access data in HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources.

Other Features

- **Real-time processing** – Spark is able to process real-time streaming data. Unlike MapReduce which processes the stored data, Spark is able to process the real-time data and hence is able to produce instant outcomes.

- **Better analytics** – Contrasting to MapReduce that includes Map and Reduce functions, Spark includes much more than that. Apache Spark consists of a rich set of SQL queries, machine learning algorithms, complex analytics, etc. With all these functionalities, analytics can be performed in a better fashion with the help of Spark.

- **Compatible with Hadoop** – Spark is not only able to work independently, it can work on top of Hadoop as well. Not just this, it is certainly compatible with both the versions of Hadoop ecosystem.



It's scalable

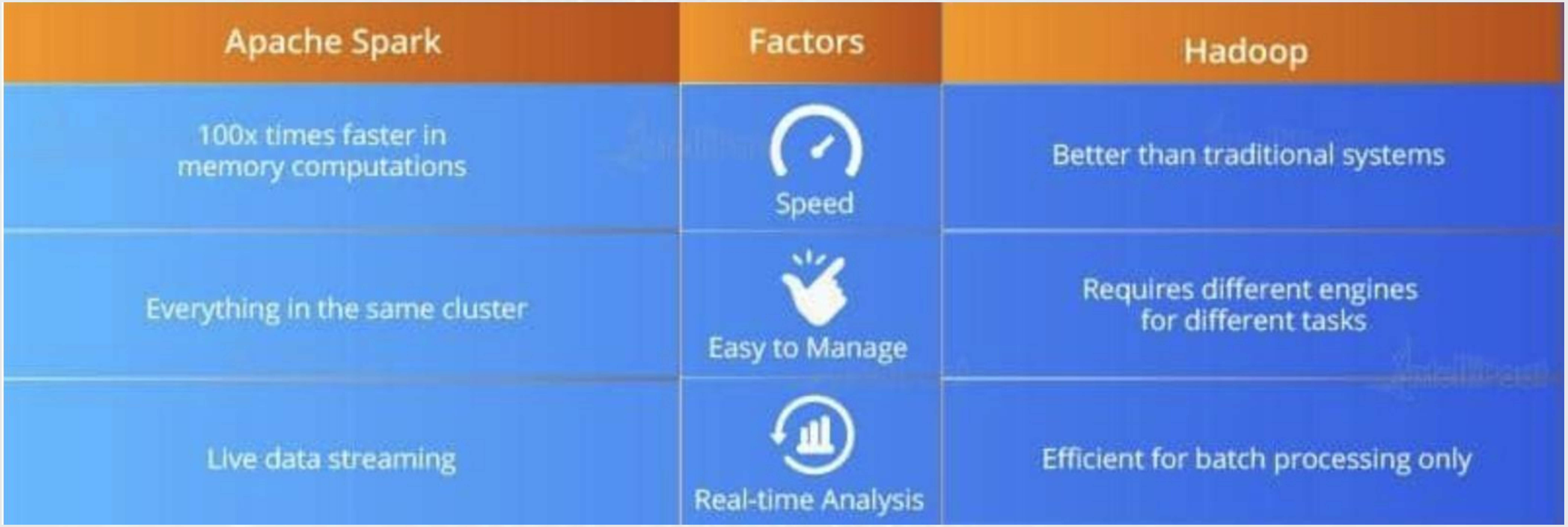
Its Hot

IntelliPaat
ebay
YAHOO!
阿里巴巴 Alibaba.com
amazon.com
NETFLIX

Its not that hard

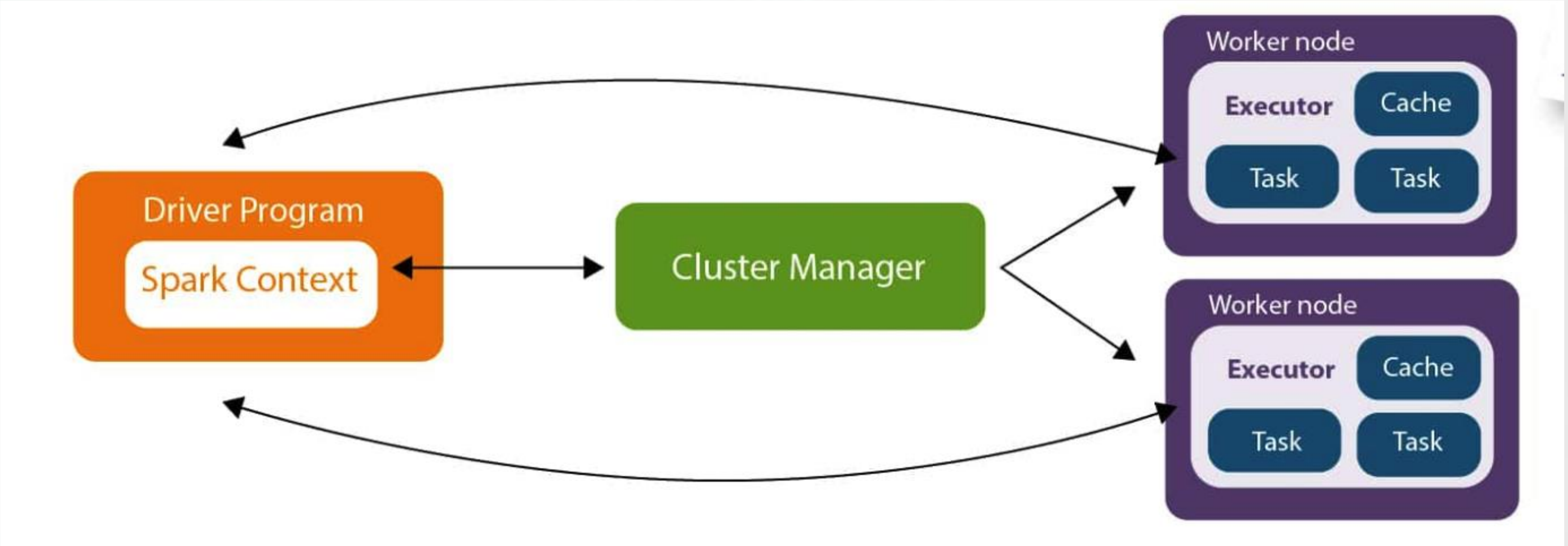
- Code in Python, Java or Scala
- Built around one main concept: the Resilient distributed dataset(RDD).

Spark vs MapReduce



Apache Spark	Factors	Hadoop
100x times faster in memory computations	Speed	Better than traditional systems
Everything in the same cluster	Easy to Manage	Requires different engines for different tasks
Live data streaming	Real-time Analysis	Efficient for batch processing only

Spark architecture



Spark architecture

- Apache Spark Framework uses a master–slave architecture which consists of a driver, which runs as a master node, and many executors which run across the worker nodes in the cluster.
- Apache Spark can be used for batch processing and real-time processing as well.

Spark architecture

- Apache Spark requires a cluster manager and a distributed storage system.
- For cluster management, Spark supports standalone, Hadoop YARN, Apache Mesos or Kubernetes.
- For distributed storage, Spark can interface with a wide variety, including Alluxio, HDFS, MapR File System, Cassandra, OpenStack Swift, Amazon S3, Kudu, Lustre file system, or a custom solution can be implemented.
- Spark also supports a pseudo-distributed local mode, usually used only for development or testing purposes, where distributed storage is not required and the local file system can be used instead; in such a scenario, Spark is run on a single machine with one executor per CPU core.

Spark Driver

- Calls the main program of an application and also creates the Spark Context which consists of all the basic functionalities.
- Contains various other components like DAG Scheduler, Task Scheduler, Backend Scheduler, and Block Manager which are responsible for translating the user written code into jobs which are actually executed on the cluster.
- The Spark Driver and the Spark Context collectively watch over the job execution within the cluster.
- The Spark Driver works with the Cluster Manager to manage various other jobs.

Cluster Manager

- Cluster Manager does all the resource allocating work. And then, the job is split into multiple smaller tasks which are further distributed onto the worker nodes.
- The SparkContext can work with various Cluster Managers, like Standalone Cluster Manager, Yet Another Resource Navigator (YARN), or Mesos, which allocate resources to containers in the worker nodes. The work is done inside the containers.

Worker nodes

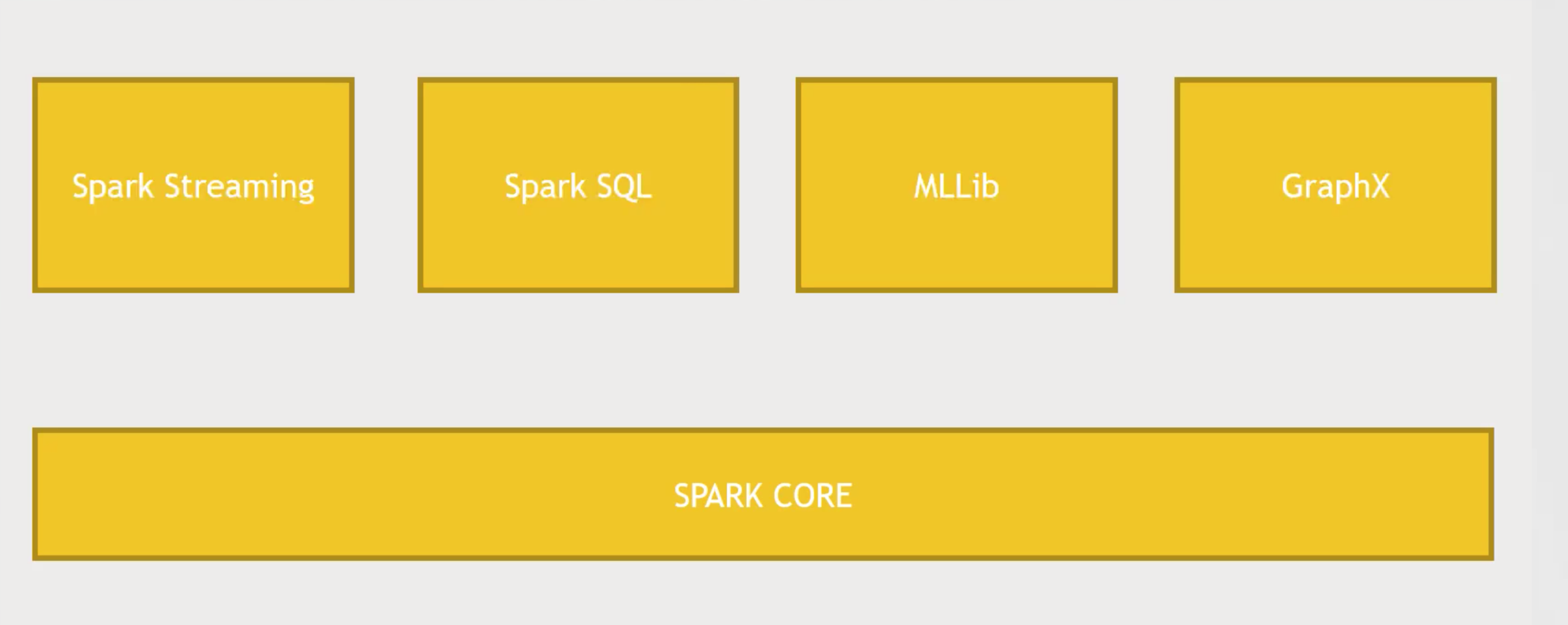
- Whenever an RDD is created in the Spark Context, it can be distributed across many worker nodes and can also be cached there.
- **Worker nodes** execute the tasks which are assigned by the Cluster Manager and returns it back to the Spark Context.
- **Executor** is responsible for the execution of these tasks. Lifetime of executors is same as that of the Spark Application. If you want to increase the performance of the system, you can increase the number of workers so that the jobs can be divided into more logical portions.

Spark abstractions

- **Resilient Distributed Dataset** (RDD): is an immutable (read-only), fundamental collection of elements or items that can be operated on many devices at the same time (parallel processing). Each dataset in an RDD can be divided into logical portions, which are then executed on different nodes of a cluster. Directed Acyclic Graph

- **Directed Acyclic Graph (DAG)**: is the scheduling layer of Apache Spark Architecture that implements stage-oriented scheduling. Compared to MapReduce, which creates a graph in two stages, Map and Reduce, Apache Spark Architecture can create DAGs that contains many stages.

Components of Spark



Spark Streaming

Spark SQL

MLLib

GraphX

SPARK CORE

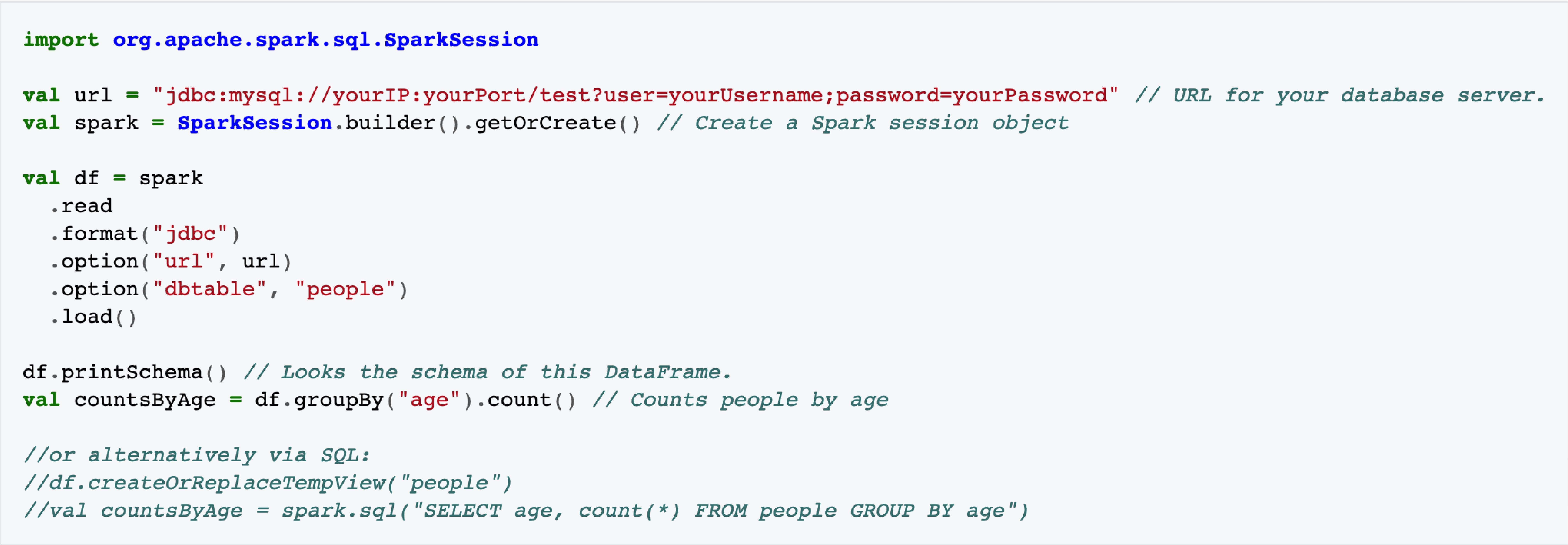
Spark streaming

- Instead of doing batch processing on data, you can actually input data in real time.
- It ingests data in mini-batches and performs RDD transformations on those mini-batches of data.
- Uses Spark Core's fast scheduling capability to perform streaming analytics.
- Imagine you have a fleet of web servers that are producing logs, that log data can be ingested as its being produced in Spark. And then analyse the data in some window of time and then output the results of the analysis to a data base or some NoSQL data store, all with a few line of code in Spark streaming.
- In Spark 2.x, a separate technology based on Datasets, called Structured Streaming, that has a higher-level interface is also provided to support streaming.

Spark SQL

- SQL interface to Spark.
- Use SQL like functions to transform your datasets.
- Introduced a data abstraction called DataFrames which provides support for structured and semi-structured data
- Very hot technology currently. A lot of optimization work is focussed on Spark SQL right now, namely datasets, which are an abstraction for RDD’s.
- Allows to do more optimizations beyond DAG’s. It also provides SQL language support, with command-line interfaces and ODBC/JDBC server.
- Spark SQL provides a domain-specific language (DSL) to manipulate DataFrames in Scala, Java, or Python.

Spark SQL



```
import org.apache.spark.sql.SparkSession

val url = "jdbc:mysql://yourIP:yourPort/test?user=yourUsername;password=yourPassword" // URL for your database server.
val spark = SparkSession.builder().getOrCreate() // Create a Spark session object

val df = spark
  .read
  .format("jdbc")
  .option("url", url)
  .option("dbtable", "people")
  .load()

df.printSchema() // Looks the schema of this DataFrame.
val countsByAge = df.groupBy("age").count() // Counts people by age

//or alternatively via SQL:
//df.createOrReplaceTempView("people")
//val countsByAge = spark.sql("SELECT age, count(*) FROM people GROUP BY age")
```

Spark MLLib

- Entire library of Machine learning and data mining tools that you can run on a dataset that is loaded on to Spark.
- It is very challenging if you want to transform your machine learning problem to Map reduce.
- Has high level classes to extract meaning from data.
- You can do every possible machine learning tasks such as clustering, linear regression, logistic regression etc..,
- 9 times as fast as the disk-based implementation used by Apache Mahout

MLLib algorithms

Many common machine learning and statistical algorithms have been implemented and are shipped with MLlib including:

- Summary statistics, correlations, stratified sampling, hypothesis testing, random data generation
- Classification and Regression: support vector machines, logistic regression, linear regression, decision trees, naive Bayes classification, Decision Tree, Random Forest
- Cluster analysis methods like k-means
- Dimensionality reduction techniques such as singular value decomposition (SVD), and principal component analysis (PCA)
- Optimization algorithms such as stochastic gradient descent(SGD), limited-memory BFGS (L-BFGS)

Spark GraphX

- Distributed graph-processing framework on top of Apache Spark.
- For example, you have a social network graph (a graph of friends, friends of friends etc..,). To analyse the properties of graph such as who is connected to who or shortest paths, GraphX provides a extensible way of analysing graphs through a lot of utility functions.
- GraphX is unsuitable for graphs that need to be updated since it is based on RDD which is immutable dataset.
- Provides two separate APIs for implementation of massively parallel algorithms: a Pregel abstraction, and a more general MapReduce-style API.
- Can be viewed as in-memory version of Apache Giraph which relies on MapReduce

Spark with Hadoop

- Spark as a binding technology for Hadoop but not its replacement. However, Spark can run separately from Hadoop, where it can run on a standalone cluster.
- Spark used on top of Hadoop can leverage its storage and cluster management. Though Spark does not provide its own storage system, we can take advantage of Hadoop for that.
- By this, we can make a powerful production environment using Hadoop capabilities. Spark can also use YARN Resource Manager for easy resource management.
- Spark can easily handle tasks scheduling across a cluster.

Spark with Hadoop

- Apache Spark can use the disaster recovery capabilities of Hadoop. We can leverage Hadoop with Spark to receive better cluster administration and data management.
- Spark together with Hadoop provides better data security.
- Spark Machine Learning provides capabilities that are not properly utilized in Hadoop MapReduce.
- Using a fast computation engine like Spark, these Machine Learning algorithms can now execute faster since they can be executed in memory.

Industries using Spark

1
Banking
2
E-commerce
3
Healthcare
4
Media
5
Travel

Industries using Spark

- **Banking**: Mainly used here for financial fraud detection, credit risk assessment, customer segmentation, and advertising. Also to analyze social media profiles, forum discussions, customer support chat, and emails. So Spark helps back to make better business decisions.

- **E-commerce**: Spark Machine Learning, along with streaming, can be used for real-time data clustering. Businesses can combine results with other data sources to provide better recommendations to their customers. Recommendation systems are mostly used in the e-commerce industry to show new trends.

- **Travel**: TripAdvisor uses Spark to compare different travel packages from different providers. It scans through hundreds of websites to find the best and reasonable hotel price, trip package, etc..,

Industries using Spark

- **Healthcare**: Apache Spark is a powerful computation engine to perform advanced analytics on patient records. It helps keep track of the patients’ health records easily. The healthcare industry uses Spark to deploy services to get insights like patient feedbacks, hospital services, and to keep track of medical data.

- **Media**: Many gaming companies use Apache Spark for finding patterns from their real-time in-game events. With this, they can derive further business opportunities like adjusting the complexity-level of the game automatically according to the players’ performance. Yahoo uses Spark for targeted marketing, customizing news pages based on readers’ interests, and so on. They use tools such as Machine Learning algorithms for identifying the ‘readers’ interests’ category. Eventually, they categorize such news stories in various sections and keep the reader updated on a timely basis.

IntelliPaat
Apache Spark Applications
APACHE Spark™
Machine Learning
Fog Computing
Event Detection
Interactive Analysis

Spark Applications

- **Machine Learning** – Apache Spark is equipped with a scalable Machine Learning Library called as MLlib that can perform advanced analytics such as clustering, classification, dimensionality reduction, etc. Some of the prominent analytics jobs like predictive analysis, customer segmentation, sentiment analysis, etc., make Spark an intelligent technology.

- **Fog computing** – With the influx of big data concepts, IoT has acquired a prominent space for the invention of more advanced technologies. Based on the theory of connecting digital devices with the help of small sensors this technology deals with a humongous amount of data emanating from numerous mediums. This requires parallel processing which is certainly not possible on cloud computing. Therefore Fog computing which decentralizes the data and storage uses Spark streaming as a solution to this problem.

Spark Applications

- **Event detection** – The feature of Spark streaming allows the organization to keep track of rare and unusual behaviours for protecting the system. Institutions like financial institutions, security organizations, and health organizations use triggers to detect the potential risk.
- **Interactive analysis** – Among the most notable features of Apache Spark is its ability to support interactive analysis. Unlike MapReduce that supports batch processing, Apache Spark processes data faster because of which it can process exploratory queries without sampling.

Who is using Spark ?

Some of the most popular companies that are using Spark are –

- **Uber** - Uses Kafka, Spark Streaming, and HDFS for building a continuous ETL pipeline.
- **Pinterest** - Uses Spark Streaming in order to gain deep insight into customer engagement details.
- **Conviva** - The pinnacle video company Conviva deploys Spark for optimizing the videos and handling live traffic.
- **Trip advisor** - Uses Spark to compare different travel packages from different providers

Master Executive di II Livello

BIG DATA ANALYSIS AND BUSINESS INTELLIGENCE

Vamsi Krishna Varma Gunturi
Data science intern at ISTAT
vamsivarmagunturi@gmail.com

Grazie

fondazione INUIT TOR VERGATA