SPARK

By
Niranjan Hegde 1BM19IS103
Prashanth Jaganathan 1BM19IS115
Prateek Gummaraju 1BM19IS117
Samartha S 1BM19IS219

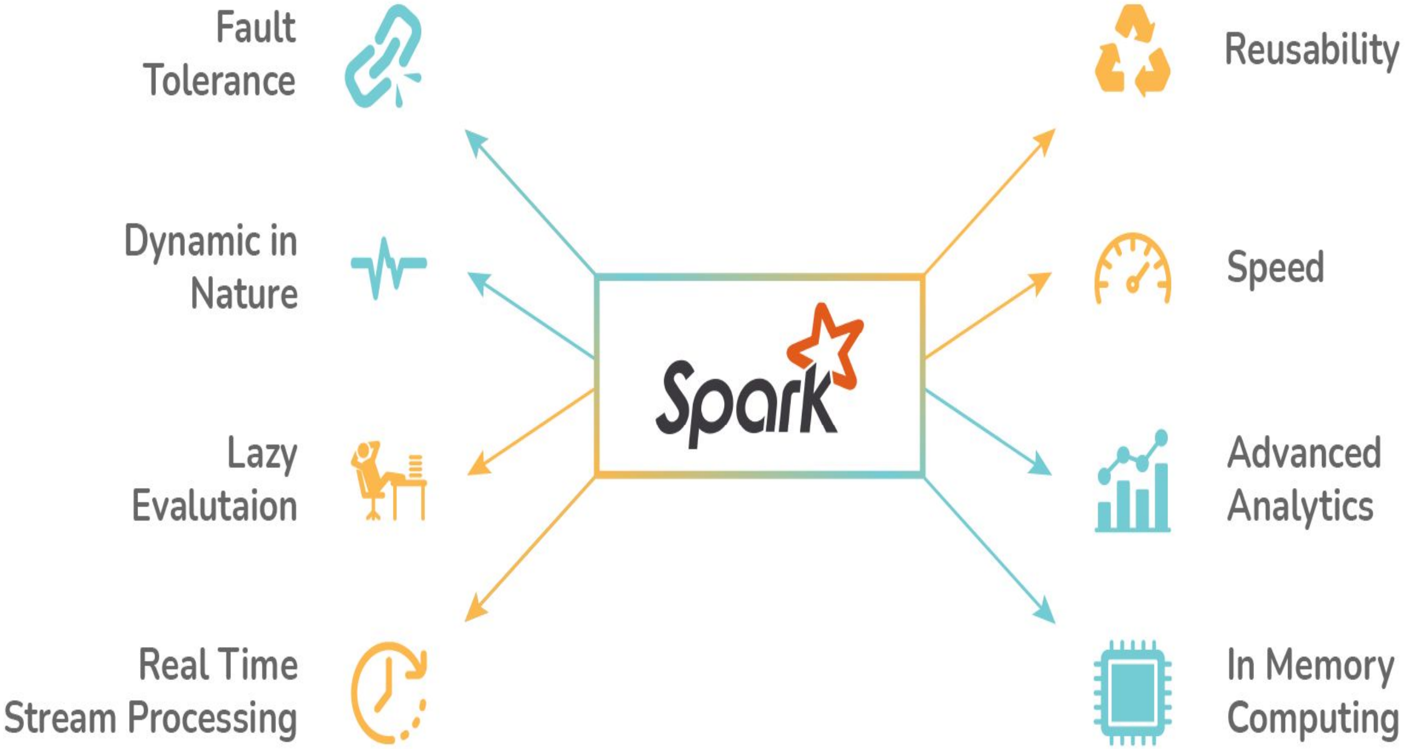
What is Apache Spark?

Apache Spark is an open source, distributed processing system used for Big Data workloads.

Developed in 2009 in UC Berkeley's AMPLab

The main feature of Spark is its **in-memory cluster computing technology** that increases the processing speed of an application

Features of Spark



Fault Tolerance

Dynamic in Nature

Lazy Evalutaion

Real Time Stream Processing

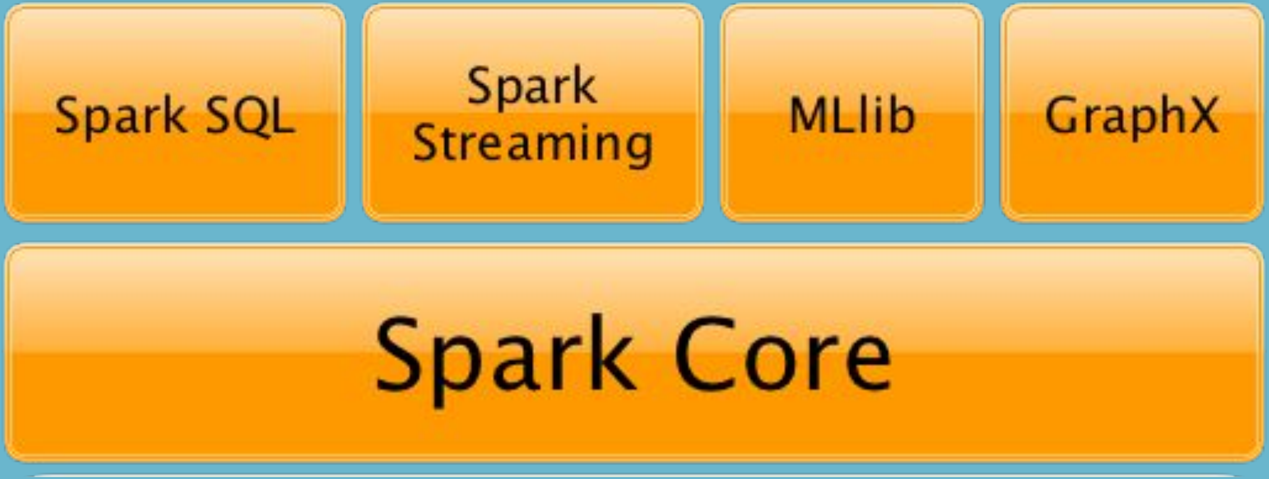
Reusability

Speed

Advanced Analytics

In Memory Computing

COMPONENTS OF SPARK



Spark SQL

Spark Streaming

MLlib

GraphX

Spark Core

SPARK
vs
HADOOP

Performance

hadoop

Hadoop is generally slow as it performs operations on the disk and cannot deliver near real-time analytics from the data

No real-time analytics

APACHE Spark

Spark runs 100 times faster in-memory, and 10 times faster on disk. If Spark runs on YARN with other resources demanding services, there could be major degradation

Faster in-memory processing

Cost

hadoop

Hadoop is less expensive as it is an **open-source software**. It requires more memory on disk which is relatively an **inexpensive commodity**

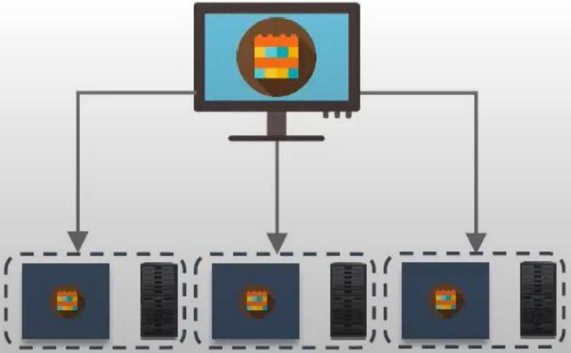
APACHE Spark

Spark is **open-source** but requires a lot of RAM to run in-memory. This **increases** the **cluster size** and its cost

Fault Tolerance

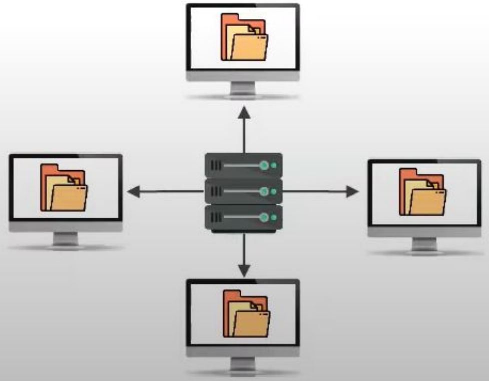
hadoop

Hadoop is **highly fault-tolerant** because it was designed to **replicate data** across many nodes. Each file is split into blocks and replicated numerous times across many machines



APACHE Spark

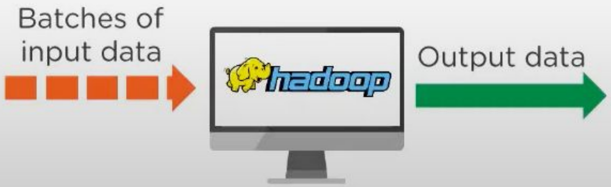
Spark uses **Resilient Distributed Datasets** (RDDs), which are fault-tolerant collections of elements that can be operated on in parallel



Data Processing

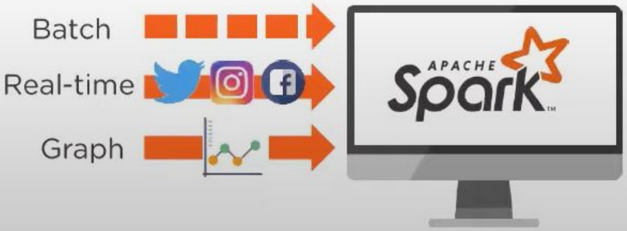
hadoop

Hadoop processes data in batches. MapReduce operates in sequential steps by reading data from the cluster, performing its operations on the data, writing the results back to the cluster



APACHE Spark

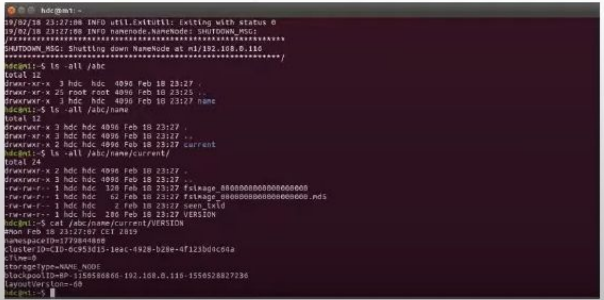
Sparks performs batch, real-time, and graph processing of data. It reads data from the cluster, performs its operation on the data, and then writes it back to the cluster



Ease of Use

hadoop

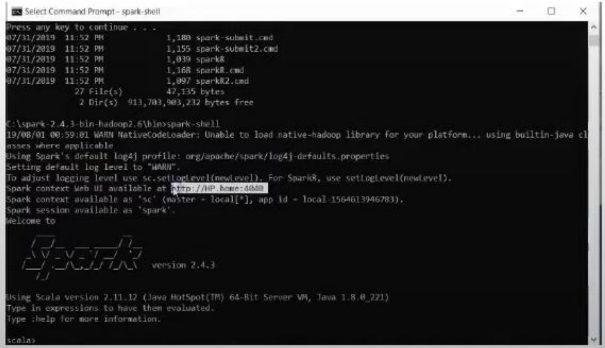
Hadoop's MapReduce has **no interactive mode** and is complex. It needs to handle low-level APIs to process the data, which requires lots of coding



```
19/02/18 23:27:08 INFO util.ExitUtil: Exiting with status 0
19/02/18 23:27:08 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at m1/192.168.0.116
************************************************************/
hdc@m1:~$ ls -all /abc
total 12
drwxr-xr-x  3 hdc  hdc  4096 Feb 18 23:27 .
drwxr-xr-x 25 root root 4096 Feb 18 23:25 ..
drwxrwxr-x  3 hdc  hdc  4096 Feb 18 23:27 name
hdc@m1:~$ ls -all /abc/name
total 12
drwxrwxr-x 3 hdc hdc 4096 Feb 18 23:27 .
drwxr-xr-x 3 hdc hdc 4096 Feb 18 23:27 ..
drwxrwxr-x 2 hdc hdc 4096 Feb 18 23:27 current
hdc@m1:~$ ls -all /abc/name/current/
total 24
drwxrwxr-x 2 hdc hdc 4096 Feb 18 23:27 .
drwxrwxr-x 3 hdc hdc 4096 Feb 18 23:27 ..
-rw-rw-r-- 1 hdc hdc  320 Feb 18 23:27 fsimage_0000000000000000000
-rw-rw-r-- 1 hdc hdc   62 Feb 18 23:27 fsimage_0000000000000000000.md5
-rw-rw-r-- 1 hdc hdc    2 Feb 18 23:27 seen_txid
-rw-rw-r-- 1 hdc hdc  206 Feb 18 23:27 VERSION
hdc@m1:~$ cat /abc/name/current/VERSION
#Mon Feb 18 23:27:07 CET 2019
namespaceID=1779844860
clusterID=CID-6c953d15-1eac-4928-b28e-4f123bd4c64a
cTime=0
storageType=NAME_NODE
blockpoolID=BP-1150586866-192.168.0.116-1550528827236
layoutVersion=-60
hdc@m1:~$ 
```

APACHE Spark™

Spark supports **user-friendly APIs** for different languages. It has an **interactive mode** and provides intermediate feedback for queries and actions



```
Press any key to continue . . .
07/31/2019  11:52 PM             1,180 spark-submit.cmd
07/31/2019  11:52 PM             1,155 spark-submit2.cmd
07/31/2019  11:52 PM             1,039 sparkR
07/31/2019  11:52 PM             1,168 sparkR.cmd
07/31/2019  11:52 PM             1,097 sparkR2.cmd
              27 File(s)         47,135 bytes
               2 Dir(s)  913,703,903,232 bytes free

C:\spark-2.4.3-bin-hadoop2.6\bin>spark-shell
19/08/01 00:59:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://HP.home:4040
Spark context available as 'sc' (master = local[*], app id = local-1564613946783).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.3
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_221)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Language Support

hadoop

Hadoop framework is developed in Java programming language. While, MapReduce applications can be written in Python, R and C++

Java

MapReduce supports programming languages

python R C++

APACHE Spark

Apache Spark is developed in Scala language and supports other programming languages like Python, R, and Java

Scala

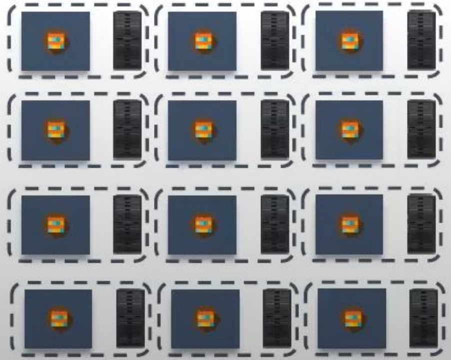
Spark supports other programming languages

python R Java

Scalability

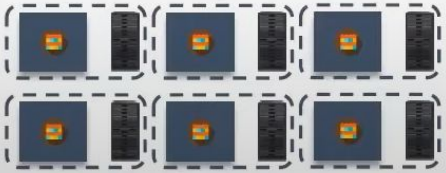
hadoop

Hadoop is **highly scalable** as we can add n number of nodes in the cluster. Yahoo reportedly used a **42,000** node Hadoop cluster



APACHE Spark

The largest known Spark cluster has **8,000** nodes. But as big data grows, it's expected that cluster sizes will increase to maintain throughput expectations.



Pyspark Demo

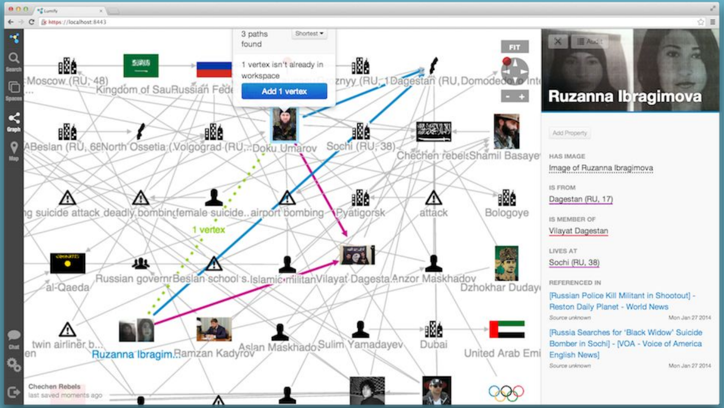
- https://colab.research.google.com/drive/1dOV2TuRV5EIjfxII2jHnNgml8l4_Dn8O
- https://colab.research.google.com/drive/1hoX7JLNGtZxUJSn3gT6msQerKWUE2Juq?ts=62beb310

Big Data Analytics and Visualization using LUMIFY

- Lumify is a big data fusion, analysis, and visualization platform. Like all big data analytics tools, it too enables you to understand connections and explore the relationship between your data.

- Lumify is considered as a good big data analytics tool because it facilitates its users to get a set of analytics options that include graph visualizations, full-text faceted search, dynamic histograms, interactive geospatial views, and collaborative workspaces that can be shared in real-time.

- Lumify offers both 2D and 3D graph visualizations with automatic layouts. It also provides a plethora of options to analyze the links between different entities in a graph.



Ruzanna Ibragimova

Add Property

HAS IMAGE
Image of Ruzanna Ibragimova

IS FROM
Dagestan (RU, 17)

IS MEMBER OF
Vilayat Dagestan

LIVES AT
Sochi (RU, 38)

REFERENCED IN
[Russian Police Kill Militant in Shootout] - Reston Daily Planet - World News
Source unknown Mon Jan 27 2014
[Russia Searches for 'Black Widow' Suicide Bomber in Sochi] - [VOA - Voice of America English News]
Source unknown Mon Jan 27 2014

- Lumify comes with specific ingest processing and interface elements for textual content, images, and videos. The platform allows you to organize your work in different workspaces.

- The platform is built on proven, scalable big data technologies. It is secure, scalable, and backed by a motivated full-time development team.

- Lumify enables users to discover complex connections and explore diverse relationships in their data through a suite of analytic options, including graph visualizations, full-text faceted search, dynamic histograms, interactive geospatial views, and collaborative workspaces shared in real-time.

- It works well in cloud environments, especially AWS.

Datawrapper

Big Data Analytics and Visualization using DataWrapper

- Datawrapper is a free, intuitive and interactive tool that does not require any coding or design knowledge in order to visualize data.

- It lets you plot data as insightful maps, charts and tables. The map, chart or table can be downloaded as PNG, PDFs or they can be embedded directly onto your website.

- Let me take you through a quick demo.

Thank you