

# Shortcomings of MapReduce

# Shortcomings of MapReduce

Force your pipeline into Map and Reduce steps

Other workflows? i.e. join, filter, map-reduce-map

# Shortcomings of MapReduce

Read from disk for each  
MapReduce job

Iterative algorithms? i.e.  
machine learning

# Shortcomings of MapReduce

Only native JAVA  
programming interface

Other languages?  
Interactivity?

# Solution?

- New framework: same features of MapReduce and more
- Capable of reusing Hadoop ecosystem, e.g. HDFS, YARN...
- Born at UC Berkeley

# Solutions by Spark

Other workflows? i.e. join,  
filter, map-reduce-map

~20 highly efficient  
distributed operations, any  
combination of them

# Solutions by Spark

Iterative algorithms? i.e.  
machine learning

in-memory caching of data,  
specified by the user

# Solutions by Spark

Interactivity? Other  
languages?

Native Python, Scala (, R)  
interface. Interactive shells.

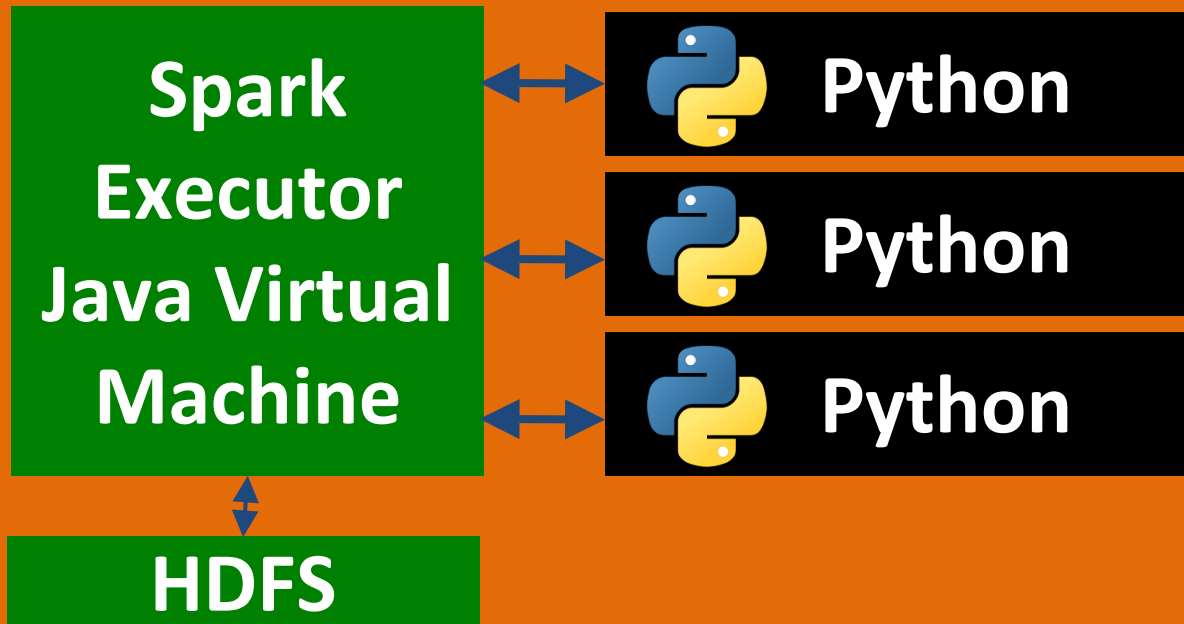


# 100TB Sorting competition

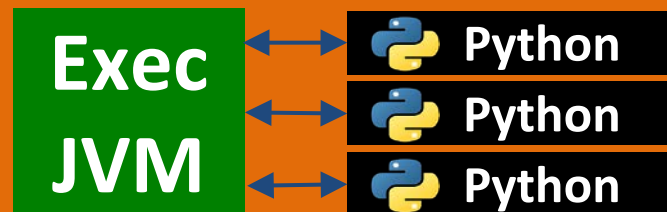
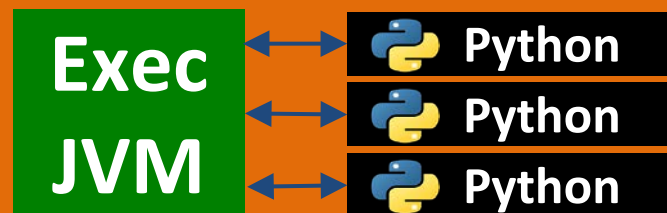
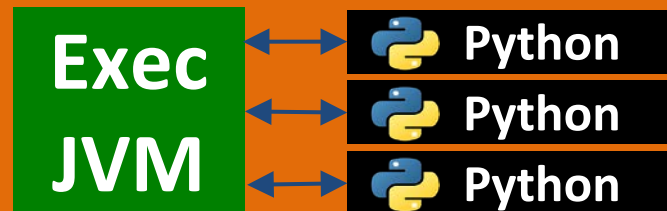
	<b>Hadoop MR Record</b>	<b>Spark Record</b>	<b>Spark 1 PB</b>
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Cluster disk throughput	3150 GB/s (est.)	618 GB/s	570 GB/s
Sort Benchmark Daytona Rules	Yes	Yes	No
Network	dedicated data center, 10Gbps	virtualized (EC2) 10Gbps network	virtualized (EC2) 10Gbps network
<b>Sort rate</b>	<b>1.42 TB/min</b>	<b>4.27 TB/min</b>	<b>4.27 TB/min</b>
<b>Sort rate/node</b>	<b>0.67 GB/min</b>	<b>20.7 GB/min</b>	<b>22.5 GB/min</b>

# Architecture of Spark

## Worker Node



# Worker Nodes



# Worker Nodes

**Cluster Manager**  
YARN/Standalone  
Provision/Restart Workers

**Exec  
JVM**

 Python  
 Python  
 Python

**Exec  
JVM**

 Python  
 Python  
 Python

**Exec  
JVM**

 Python  
 Python  
 Python

# Worker Nodes

## Driver Program

Spark  
Context

Spark  
Context

Cluster  
Manager

Exec  
JVM

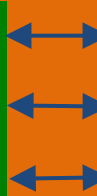
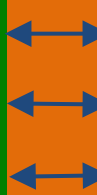
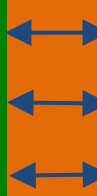
Python  
Python  
Python

Exec  
JVM

Python  
Python  
Python

Exec  
JVM

Python  
Python  
Python



# on Amazon EMR

## EC2 nodes

### Master node

