



Introduction

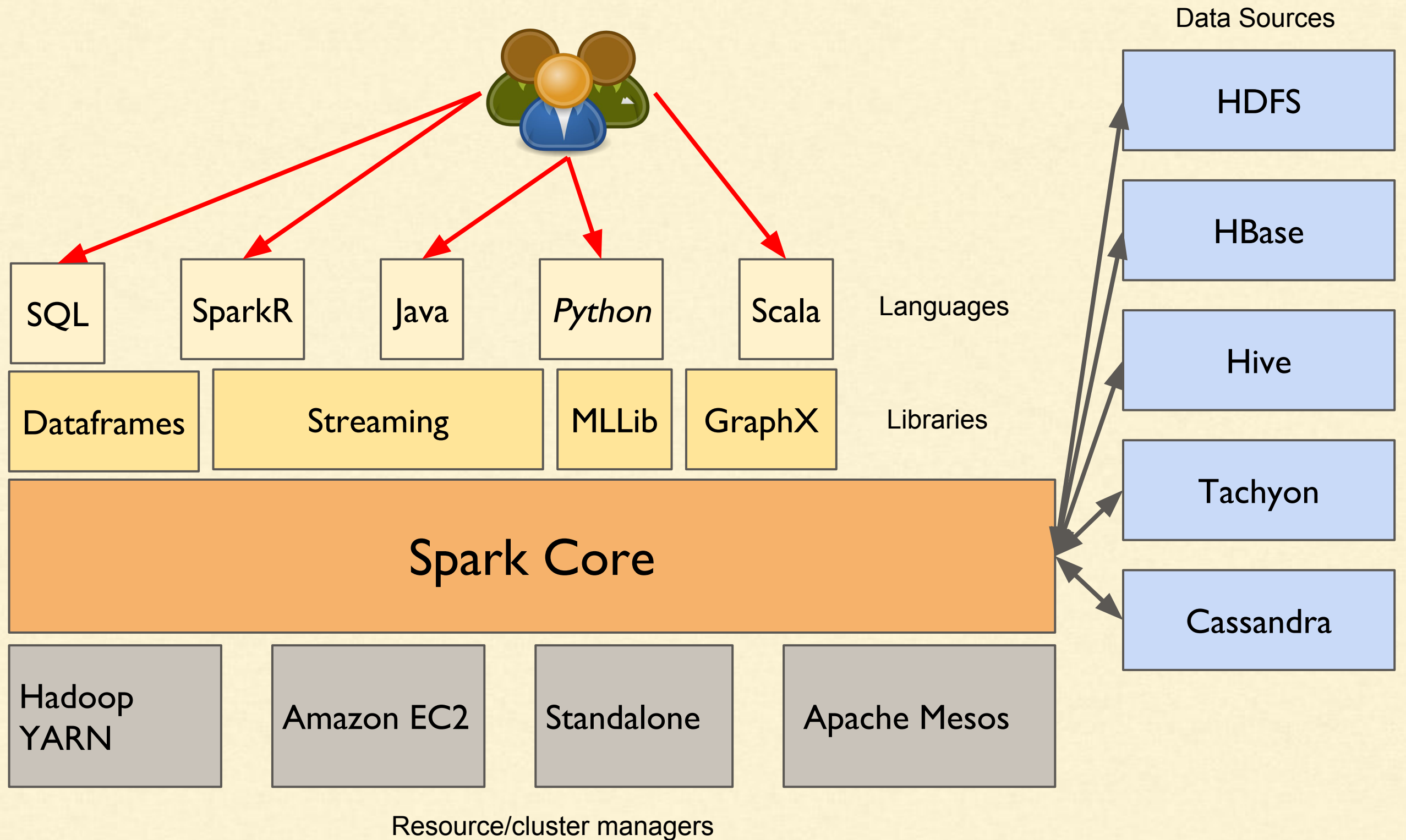




- Really fast MapReduce
 - 100x faster than Hadoop MapReduce in memory,
 - 10x faster on disk.
- Builds on similar paradigms as MapReduce
- Integrated with Hadoop

Spark Core - A fast and general engine for large-scale data processing.

Spark Architecture



Why Apache Spark?

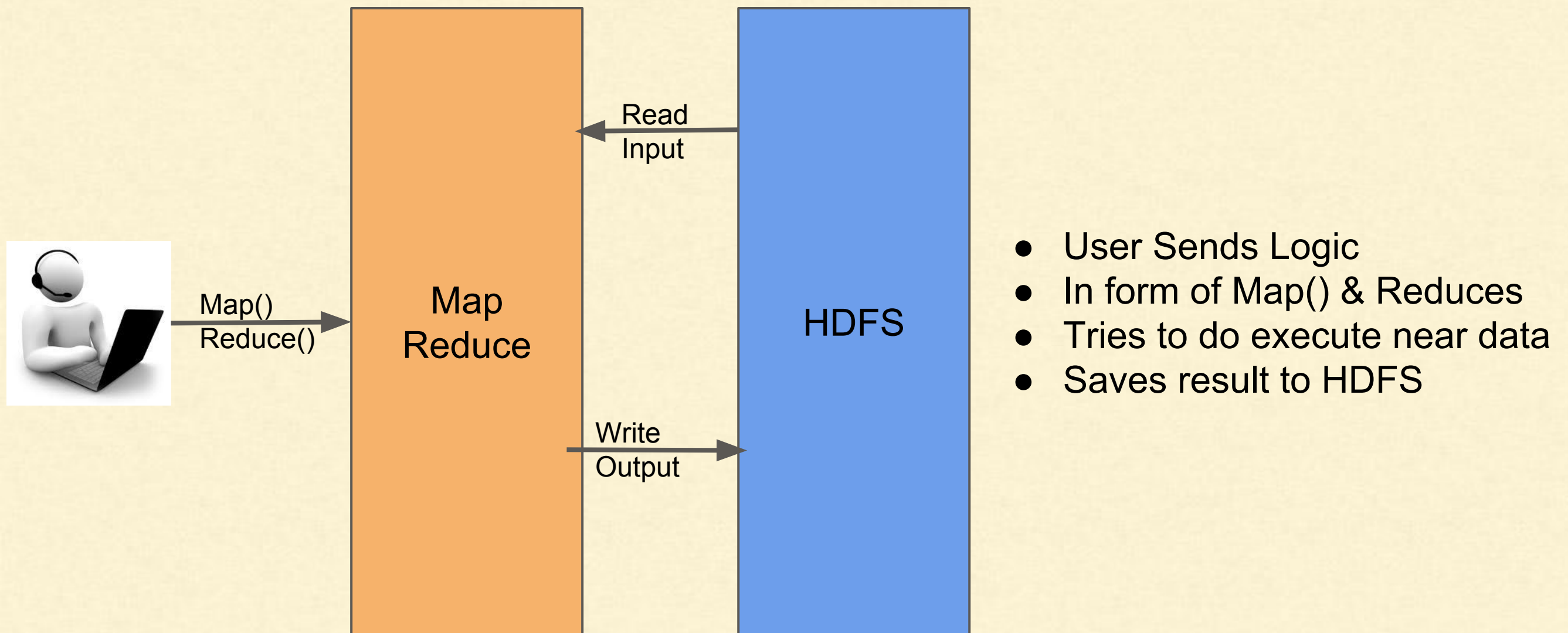
Or

Why is Apache Spark faster than MapReduce?



Why Apache Spark?

Hadoop Map Reduce



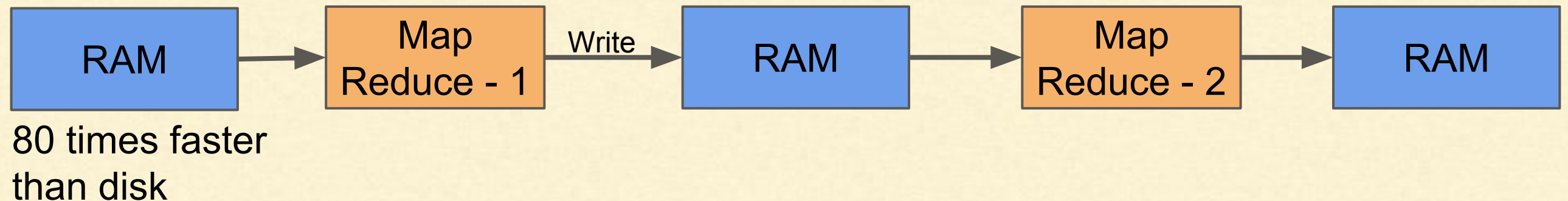
Hadoop Map Reduce - Multiple Phases



Shortcoming of Map Reduce

1. Batchwise Design
 - a. Every map-reduce cycle reads from and writes to HDFS
 - b. Heavy Latency
2. Converting logic to Map-Reduce paradigm is difficult
3. In-memory computing was not possible

Shortcoming of Map Reduce



Latency Numbers Every Programmer Should Know

Read 1 MB sequentially from SSD*	1 ms	~1GB/sec SSD, 4X memory
Disk seek	10 ms	20x datacenter roundtrip
Read 1 MB sequentially from disk	20 ms	80x memory, 20X SSD
Send packet CA→Netherlands→CA	150 ms	

See: <https://gist.github.com/jboner/2841832>

Getting Started - CloudeXLab

We have already installed the Apache Spark on CloudeXLab.
So, you don't have install anything.


You simply have to login into Web Console
and
Get started with commands.

Getting Started - Downloading

1. Find out hadoop version:
 - `[student@hadoop1 ~]$ hadoop version`
 - `Hadoop 2.4.0.2.1.4.0-632`
2. Go to <https://spark.apache.org/downloads.html>
3. Select the release for your version of hadoop & Download
4. On servers you could use `wget`
5. Every download can be run in standalone mode
6. Unzip - `tar -xzf spark*.tgz`
7. In this folder, the *bin* folder contains the spark commands

Download Spark

The latest release of Spark is Spark 1.5.0, released on September 9, 2015 ([release notes](#)) ([git tag](#))

1. Choose a Spark release: 1.5.0 (Sep 09 2015) 
2. Choose a package type ☒ Source Code [can build several Hadoop versions]
☐ Pre-built with user-provided Hadoop [can use with most Hadoop distributions]
3. Choose a download type ☐ Pre-built for Hadoop 2.6 and later
☐ Pre-built for Hadoop 2.4 and later
4. Download Spark: [spark-1.5.0-bin-hadoop2.tgz](#)
☐ Pre-built for Hadoop 2.3
☐ Pre-built for Hadoop 1.X
☐ Pre-built for CDH 4
5. Verify this release using `md5sum`

Getting Started - Binaries Overview

Binary	Description
<i>spark-shell</i>	Runs spark scala interactive commandline
<i>pyspark</i>	Runs python spark interactive commandline
<i>sparkR</i>	Runs R on spark (/usr/spark2.6/bin/sparkR)
<i>spark-submit</i>	Submit a jar or python application for execution on cluster
<i>spark-sql</i>	Runs the spark sql interactive shell

Starting Spark With Scala Interactive Shell

```
$ spark-shell
```

```
16/10/14 10:04:50 INFO SessionState: Created HDFS directory: /tmp/hive/sandeeppg26d35
16/10/14 10:04:50 INFO SessionState: Created local directory: /tmp/sandeeppgiri95
16/10/14 10:04:50 INFO SessionState: Created HDFS directory: /tmp/hive/sandeeppg26d35/_tmp_space.db
16/10/14 10:04:50 INFO SparkILoop: Created sql context (with Hive support)..
SQL context available as sqlContext.

scala> █
```

It is basically the scala REPL or interactive shell with one extra variable “sc”.
Check `dir(sc)` or `help(sc)`

Starting Spark With Python Interactive Shell

```
$ pyspark
```

```
██████████ version 1.5.0
```

```
Using Python version 2.6.6 (r266:84292, Jan 22 2014 09:42:36)  
SparkContext available as sc, HiveContext available as sqlContext.  
>>>
```

It is basically the python interactive shell
with one extra variable “sc”.

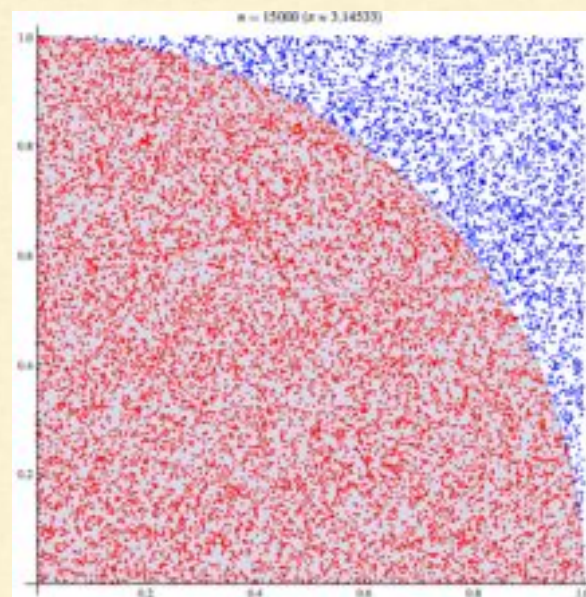
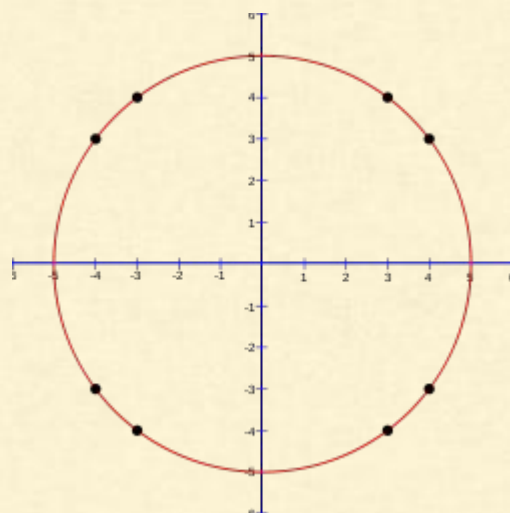
Check `dir(sc)` or `help(sc)`

Getting Started - spark-submit

- To run example:
 - `spark-submit --class org.apache.spark.examples.SparkPi /usr/hdp/current/spark-client/lib/spark-examples-*.jar 10`

The example computes the area of circle of a radius 1 by counting total number of squares.

- See https://en.wikipedia.org/wiki/Approximations_of_%CF%80#Summing_a_circle.27s_area
- Code: <https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/spark/examples/SparkPi.scala>



Getting Started - spark-submit

```
sandeep — sandeepgiri9034@ip-172-31-60-179:~ — ssh sandeepgiri9034@e.cloudxlab.com — 121x44
17/03/15 20:53:03 INFO TaskSetManager: Finished task 5.0 in stage 0.0 (TID 5) in 115 ms on localhost (10/10)
17/03/15 20:53:03 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
17/03/15 20:53:03 INFO DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:36) finished in 0.560 s
17/03/15 20:53:03 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:36, took 0.825483 s
Pi is roughly 3.143212
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/api,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
17/03/15 20:53:03 INFO ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
17/03/15 20:53:03 INFO SparkUI: Stopped Spark web UI at http://172.31.60.179:4043
17/03/15 20:53:03 INFO DAGScheduler: Stopping DAGScheduler
17/03/15 20:53:03 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/03/15 20:53:03 INFO MemoryStore: MemoryStore cleared
17/03/15 20:53:03 INFO BlockManager: BlockManager stopped
17/03/15 20:53:03 INFO BlockManagerMaster: BlockManagerMaster stopped
17/03/15 20:53:03 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
17/03/15 20:53:03 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
17/03/15 20:53:03 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
17/03/15 20:53:03 INFO SparkContext: Successfully stopped SparkContext
17/03/15 20:53:03 INFO ShutdownHookManager: Shutdown hook called
17/03/15 20:53:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-c0c1ed73-bad6-45ff-ae88-deefbf967bd8
[sandeepgiri9034@ip-172-31-60-179 ~]$
```

Getting Started - Binaries Overview

Binary	Description
<i>spark-shell</i>	Runs spark scala interactive commandline
<i>pyspark</i>	Runs python spark interactive commandline
<i>sparkR</i>	Runs R on spark (/usr/spark2.6/bin/sparkR)
<i>spark-submit</i>	Submit a jar or python application for execution on cluster
<i>spark-sql</i>	Runs the spark sql interactive shell

Getting Started - CloudxLab

To launch Spark on Hadoop,
Set the Environment Variables pointing to Hadoop.

```
export YARN_CONF_DIR=/etc/hadoop/conf/  
export HADOOP_CONF_DIR=/etc/hadoop/conf/
```

Getting Started - CloudxLab

We have installed other versions too:

1. `/usr/spark2.0.1/bin/spark-shell`
2. `/usr/spark1.6/bin/spark-shell`
3. `/usr/spark1.2.1/bin/spark-shell`



Introduction

Thank you!

