Apache Spark

Session 2 - Getting Started

WELCOME - KNOWBIGDATA

- ❑ Expert Instructors
- ❑ CloudLabs
- ❑ Lifetime access to LMS
  - ❑ Presentations
  - ❑ Class Recording
  - ❑ Assignments + Quizzes
  - ❑ Project Work

- ❑ Real Life Project
- ❑ Course Completion Certificate
- ❑ 24x7 support
- ❑ KnowBigData - Alumni
  - ❑ Jobs
  - ❑ Stay Abreast (Updated Content, Complimentary Sessions)
  - ❑ Stay Connected

COURSE CONTENT

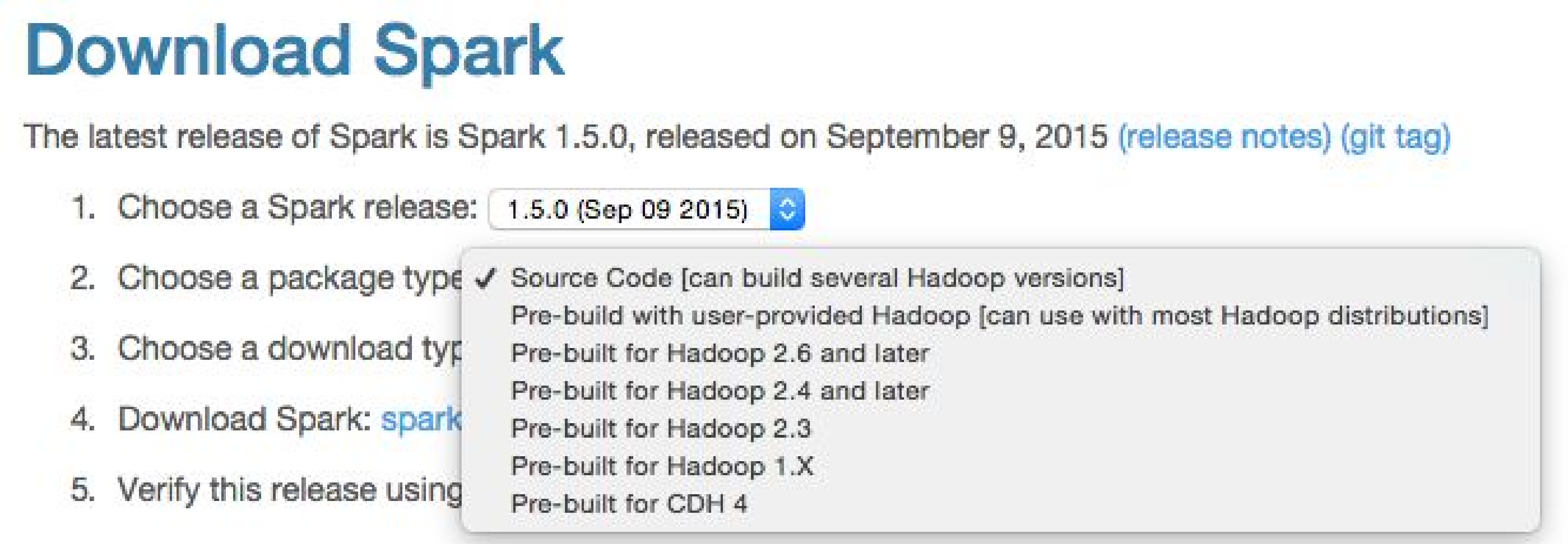
	I	Introduction to Big Data with Apache Spark
➡	II	Downloading Spark and Getting Started
	III	Programming with RDDs
	IV	Working with Key/Value Pairs
	V	Loading and Saving Your Data
	VI	Advanced Spark Programming
	VII	Running on a Cluster
	VIII	Tuning and Debugging Spark
	IX	Spark SQL, SparkR
	X	Spark Streaming
	XI	Machine Learning with MLlib, GraphX

About Instructor?

2014	**KnowBigData**	Founded
2014 2012	**Amazon**	Built High Throughput Systems for Amazon.com site using in-house NoSql.
2012	**InMobi**	Built Recommender that churns 200 TB
2011 2006	**tBits Global**	Founded tBits Global Built an enterprise grade Document Management System
2006 2002	**D.E.Shaw**	Built the big data systems before the term was coined
2002	**IIT Roorkee**	Finished B.Tech.

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 -xzvf spark*.tgz
7. Copy it to /usr/lib and make a clean link /usr/lib/spark



Download Spark

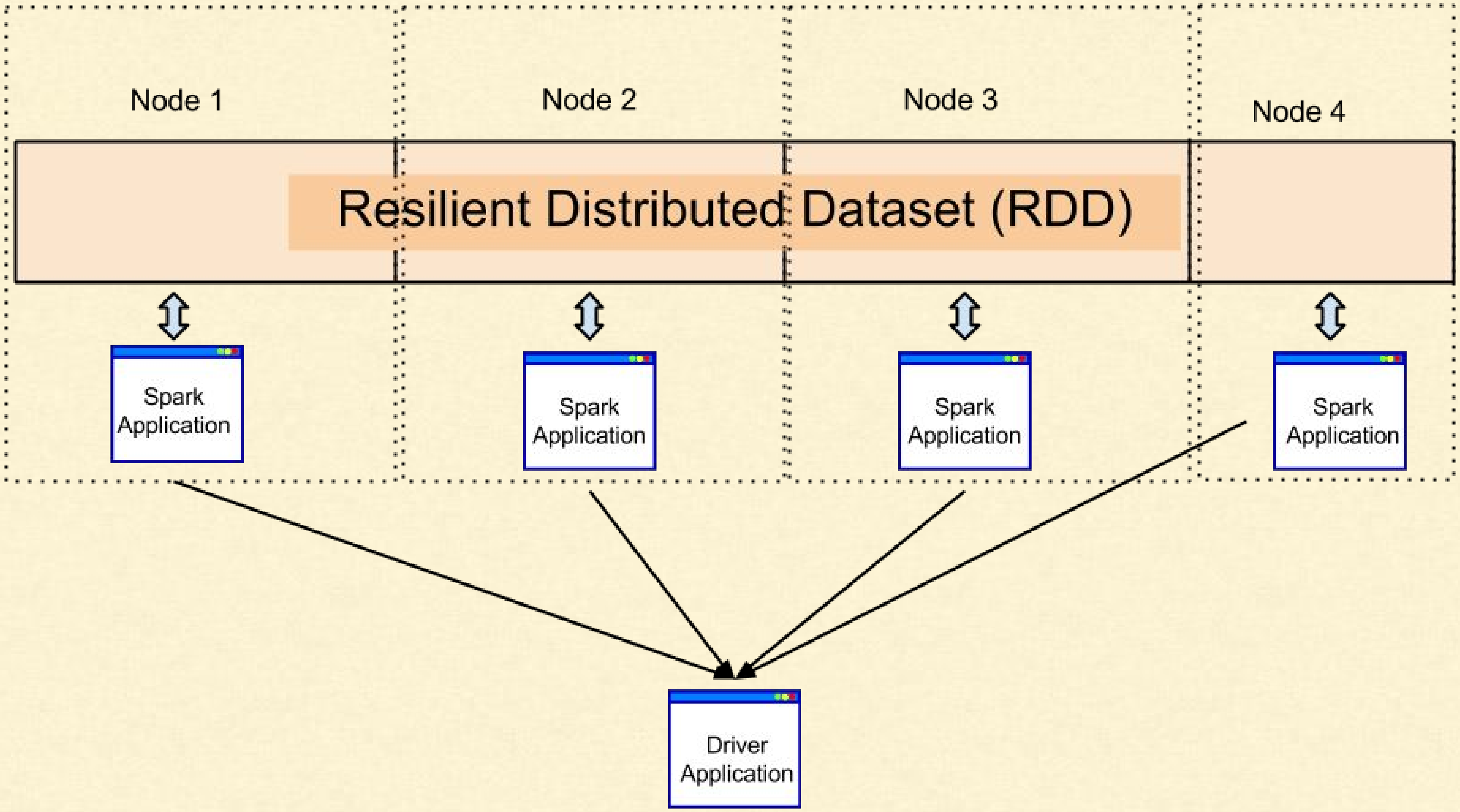
The latest release of Spark is Spark 1.5.0, released on September 9, 2015 (release notes) (git tag)

1. Choose a Spark release: 1.5.0 (Sep 09 2015)
2. Choose a package type
3. Choose a download typ
4. Download Spark: spark
5. Verify this release using

✓ Source Code [can build several Hadoop versions]
Pre-build with user-provided Hadoop [can use with most Hadoop distributions]
Pre-built for Hadoop 2.6 and later
Pre-built for Hadoop 2.4 and later
Pre-built for Hadoop 2.3
Pre-built for Hadoop 1.X
Pre-built for CDH 4

Architecture Overview

Spark Driver Launches work.

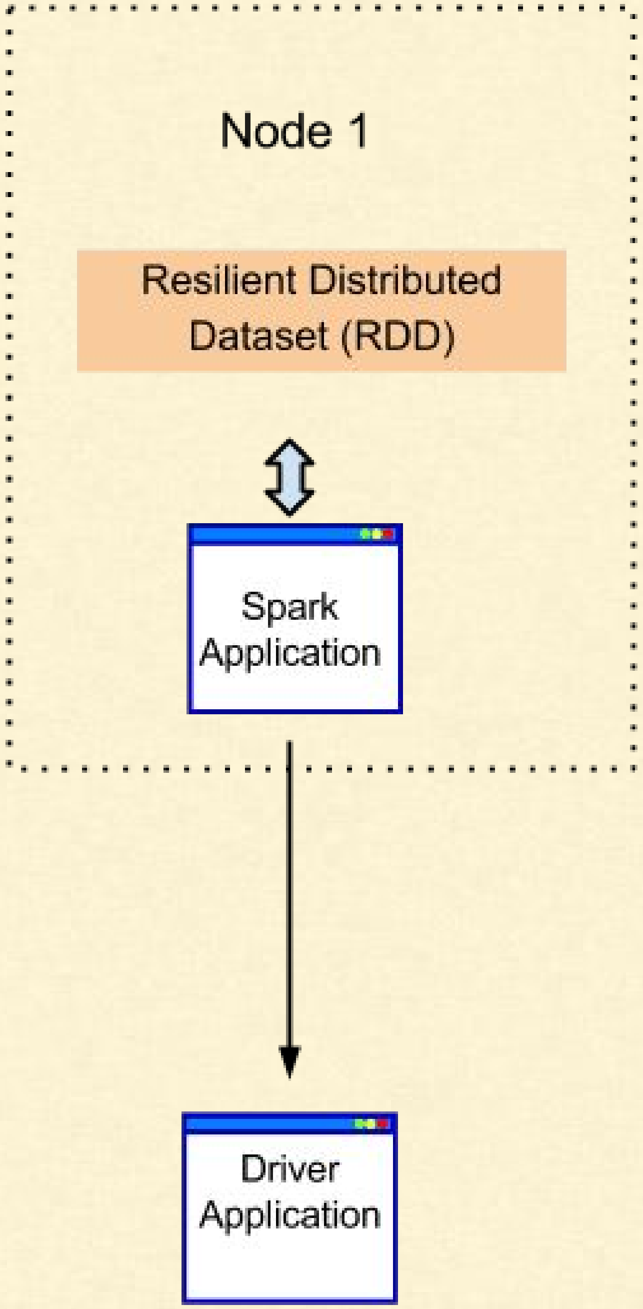


Getting Started - Two Modes

Standalone	**Over Cluster**
1. Doesn't need resource manager 2. Multiple core - parallel computing 3. Install spark on all nodes. a. Inform all nodes about each other b. Launch spark on all nodes. c. The spark nodes will discover each other	1. For production environment 2. On resource managers e.g. a. YARN b. Mesos

Getting Started - Standalone

- Spark without any resource manager on a machine
- Used for testing
- Or utilizing the mutli core abilities of machine
  - For parallel processing
- Any command out of all the binaries launched
  - without --master
  - or with --master local

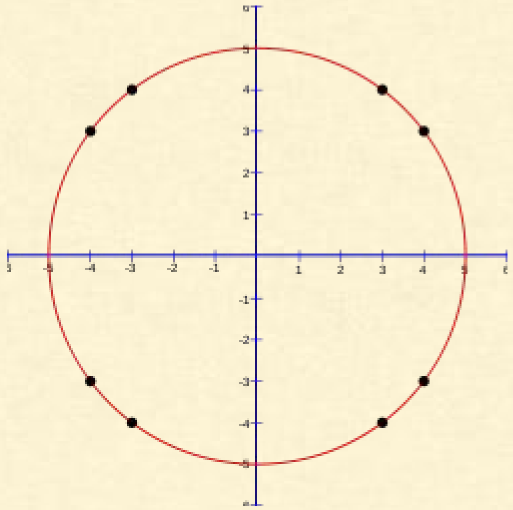


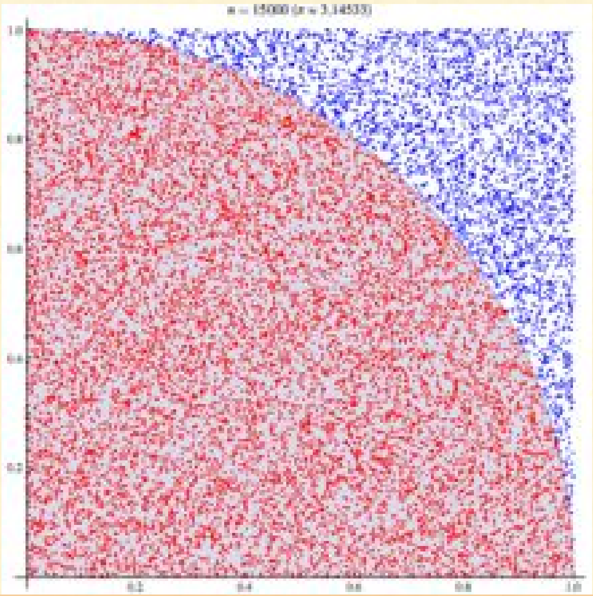
Getting Started - Standalone

- To run example:
  - *./bin/spark-submit --class org.apache.spark.examples.SparkPi --master local lib/spark-examples*.jar 10*
  - *./bin/spark-submit --class org.apache.spark.examples.SparkPi lib/spark-examples*.jar 10*
- To check the status, use:
  - http://hadoop1.knowbigdata.com:4040/

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





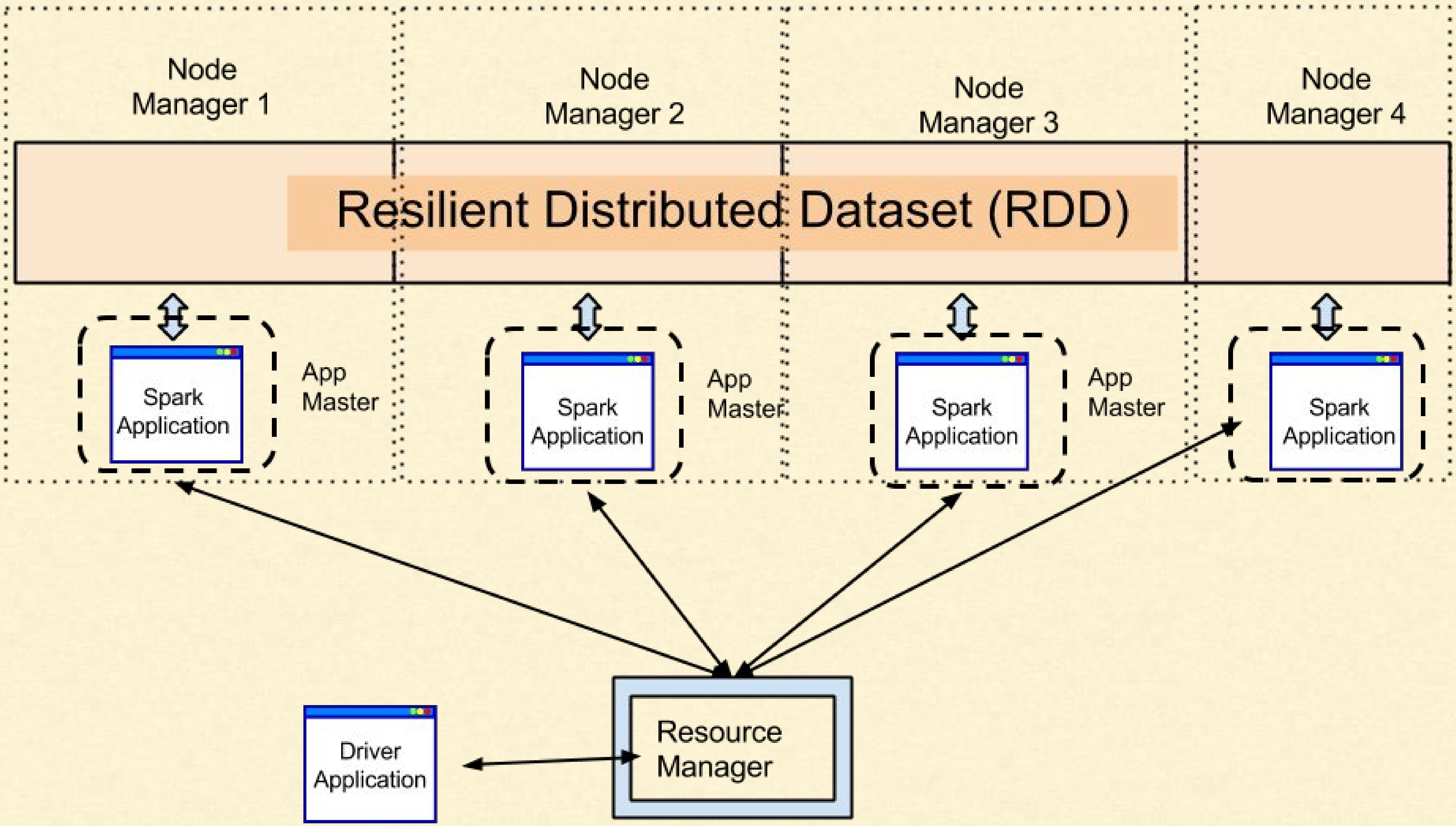
Getting Started - Binaries Overview

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

Running on Hadoop / YARN : Two Modes

As yarn-client	**yarn cluster**
1. Driver runs on the client 2. Client can’t disconnect 3. --master yarn-client	1. Driver runs on Application master insite yarn 2. Client can disconnect after starting 3. --master yarn-cluster

Architecture Yarn Client Mode



1. Driver Application is runs outside yarn
   a. On machine where it is launched
2. If Driver Application shuts down the process is killed
3. Does not have resilience but is quicker to run.

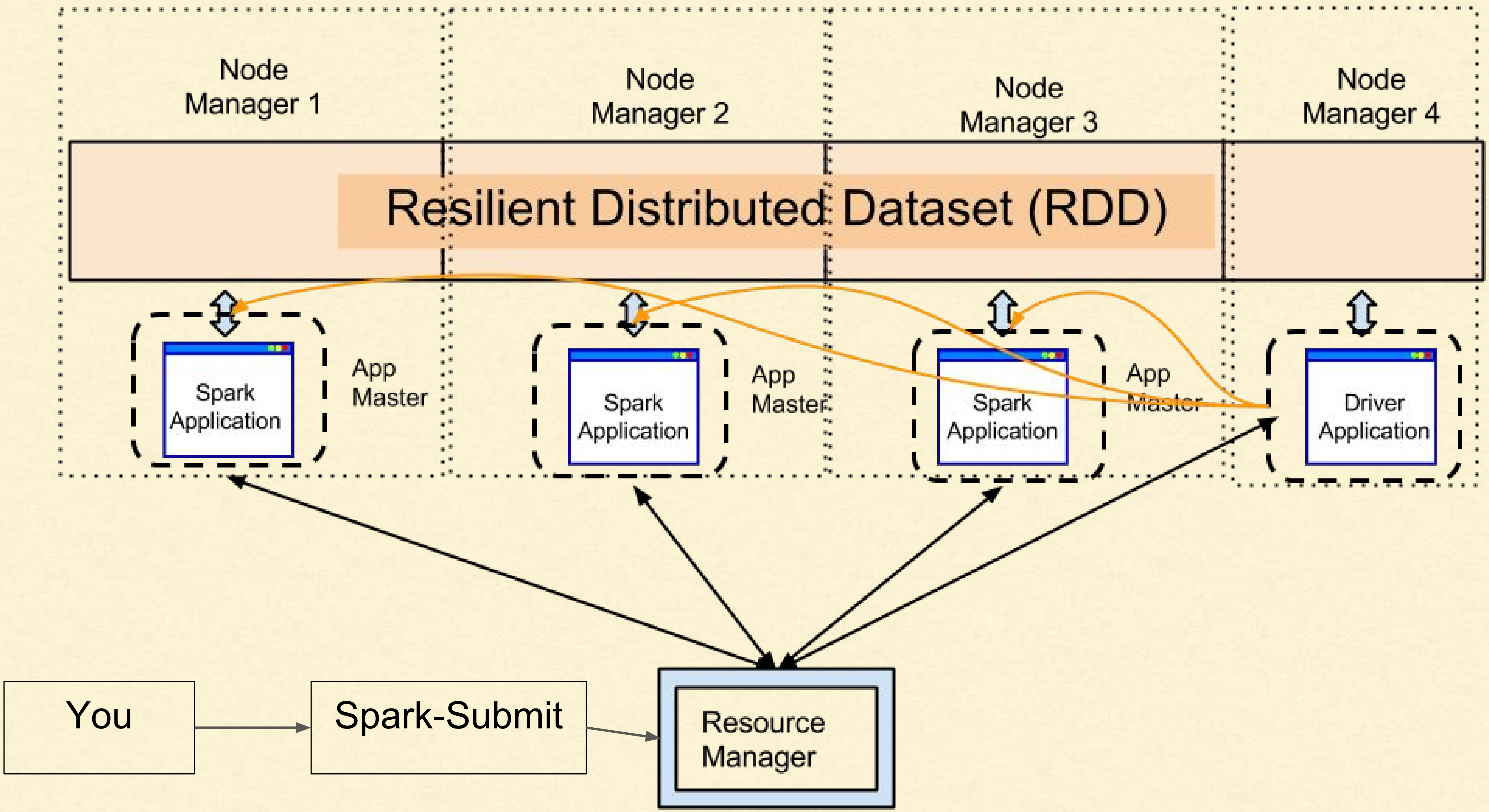
Architecture Yarn Client Mode - Example

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

cd /usr/spark2.6
spark-submit --class org.apache.spark.examples.SparkPi \
    --master yarn-client \
    --num-executors 2 \
    --driver-memory 100m \
    --executor-memory 100m \
    --executor-cores 1 \
    lib/spark-examples*.jar \
    100
```

- To check the status, use:
  - http://hadoop1.knowbigdata.com:4040/
  - http://hadoop1.knowbigdata.com:8088/cluster

Architecture Yarn Cluster Mode



1. Driver Application runs inside yarn in application master
2. If launcher shuts down the process continues like a batch process
   a. in background
3. Preferred way to run the long running processes

Architecture Yarn Client Mode - Example

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

./bin/spark-submit --class org.apache.spark.examples.SparkPi \
    --master yarn-cluster \
    --num-executors 2 \
    --driver-memory 100m \
    --executor-memory 100m \
    --executor-cores 1 \
    lib/spark-examples*.jar \
    10
```

To check the status, use:

- http://hadoop1.knowbigdata.com:4040/
- http://hadoop1.knowbigdata.com:8088/cluster

Summary

1. How to download and get started?
2. What are various Binaries?
3. Understood various modes
    1. Standalone
    2. On Cluster
        1. On Yarn
            1. Yarn-client
            2. Yarn-cluster

Apache Spark

Thank you.

+1 419 665 3276 (US)
+91 803 959 1464 (IN)

reachus@knowbigdata.com

Subscribe to our Youtube channel for latest videos - https://www.youtube.com/channel/UCxugRFe5wETYA7nMH6VGyEA