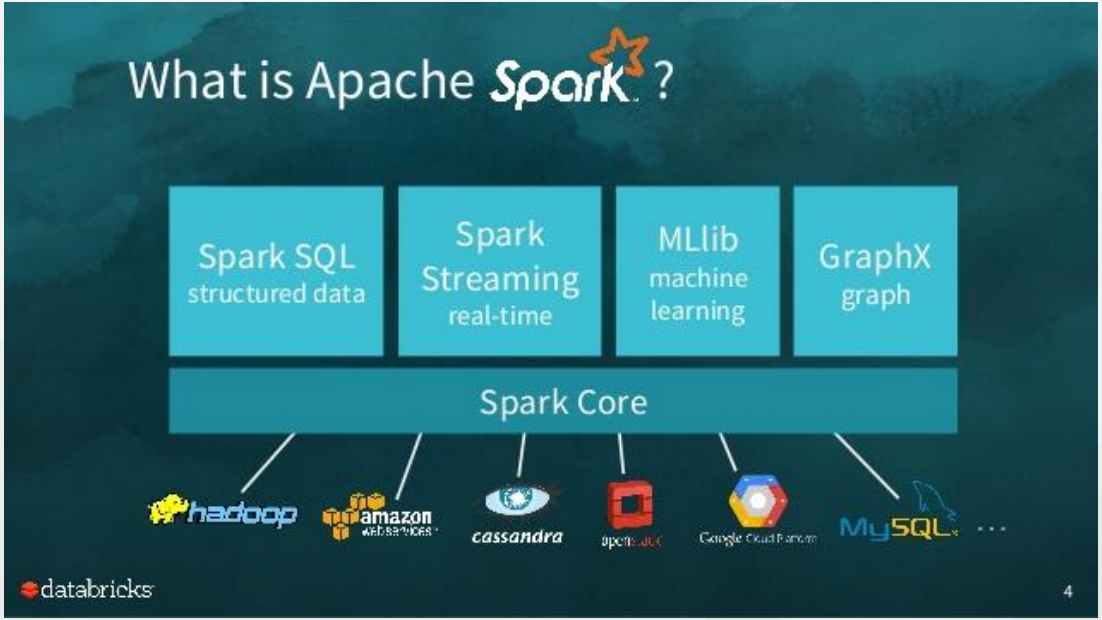
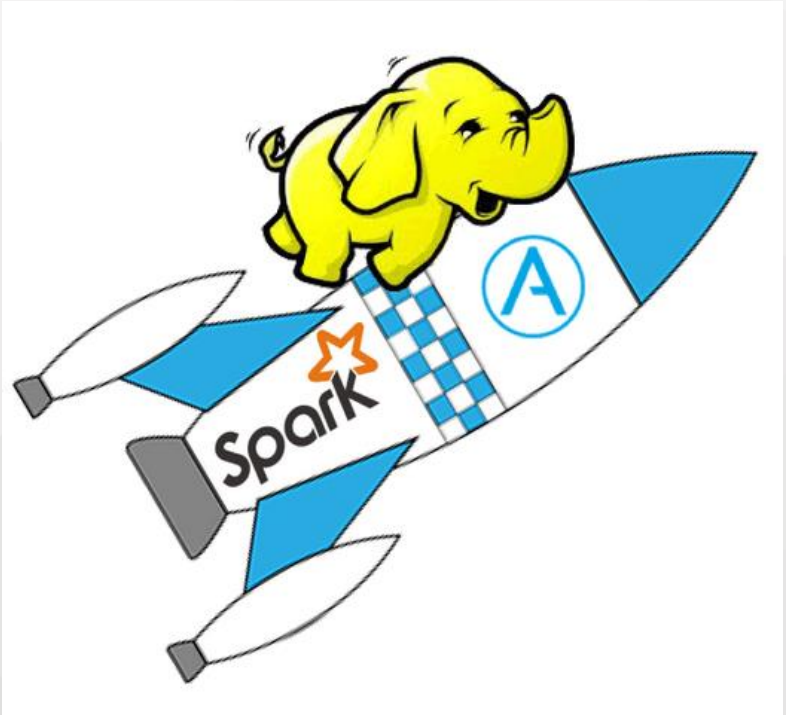


The Pilot Of Spark

2017.5 XenRon

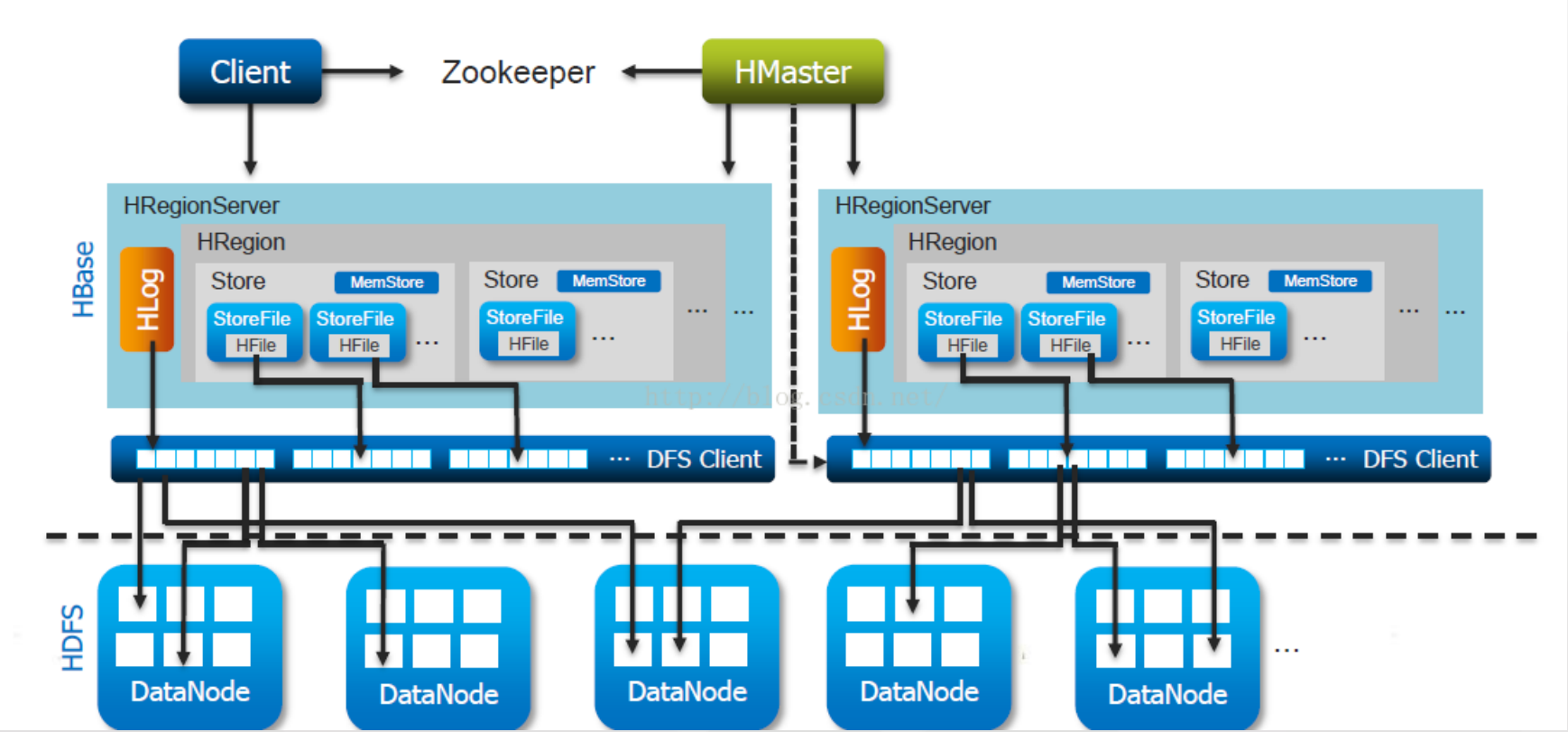
http://spark.apache.org/docs/latest/

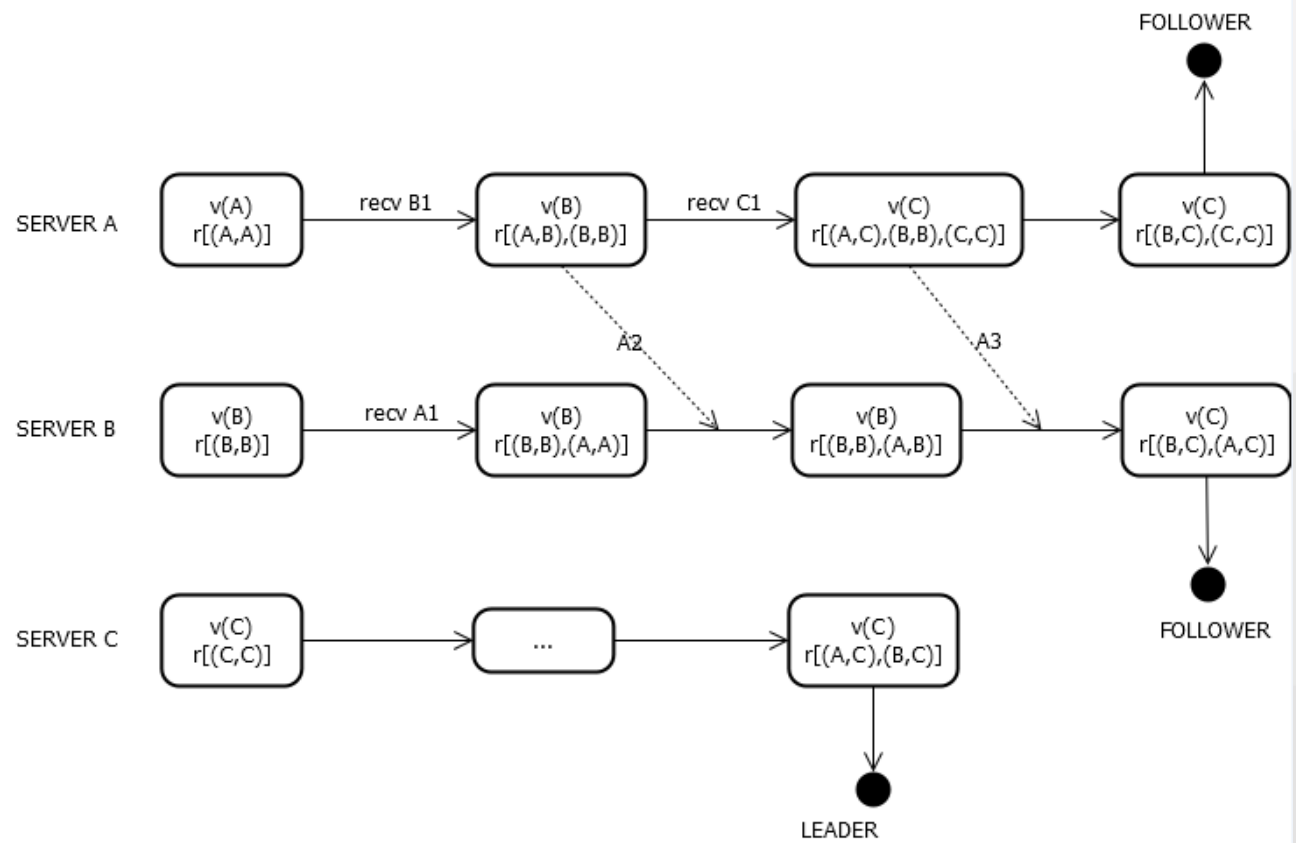
CONTENTS



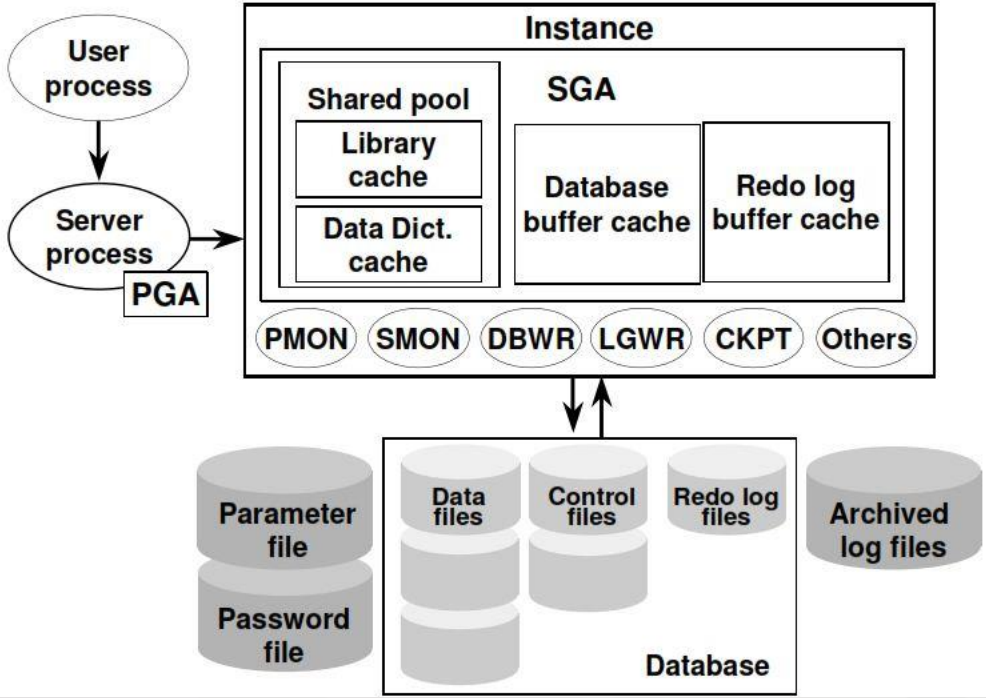
Review

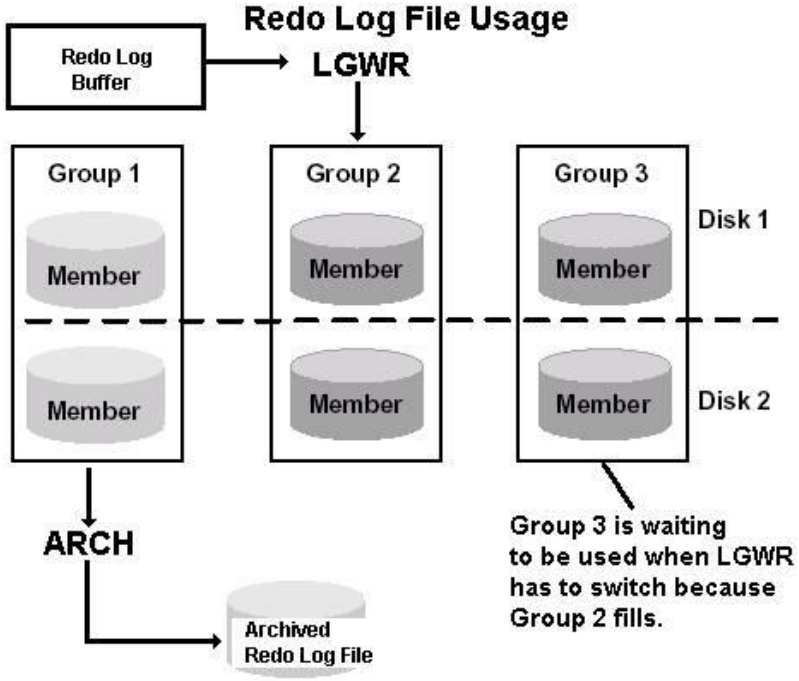
Zookeeper


Client
Zookeeper
HMaster
HBase
HRegionServer
HRegion
HLog
Store
MemStore
StoreFile
HFile
DFS Client
HDFS
DataNode

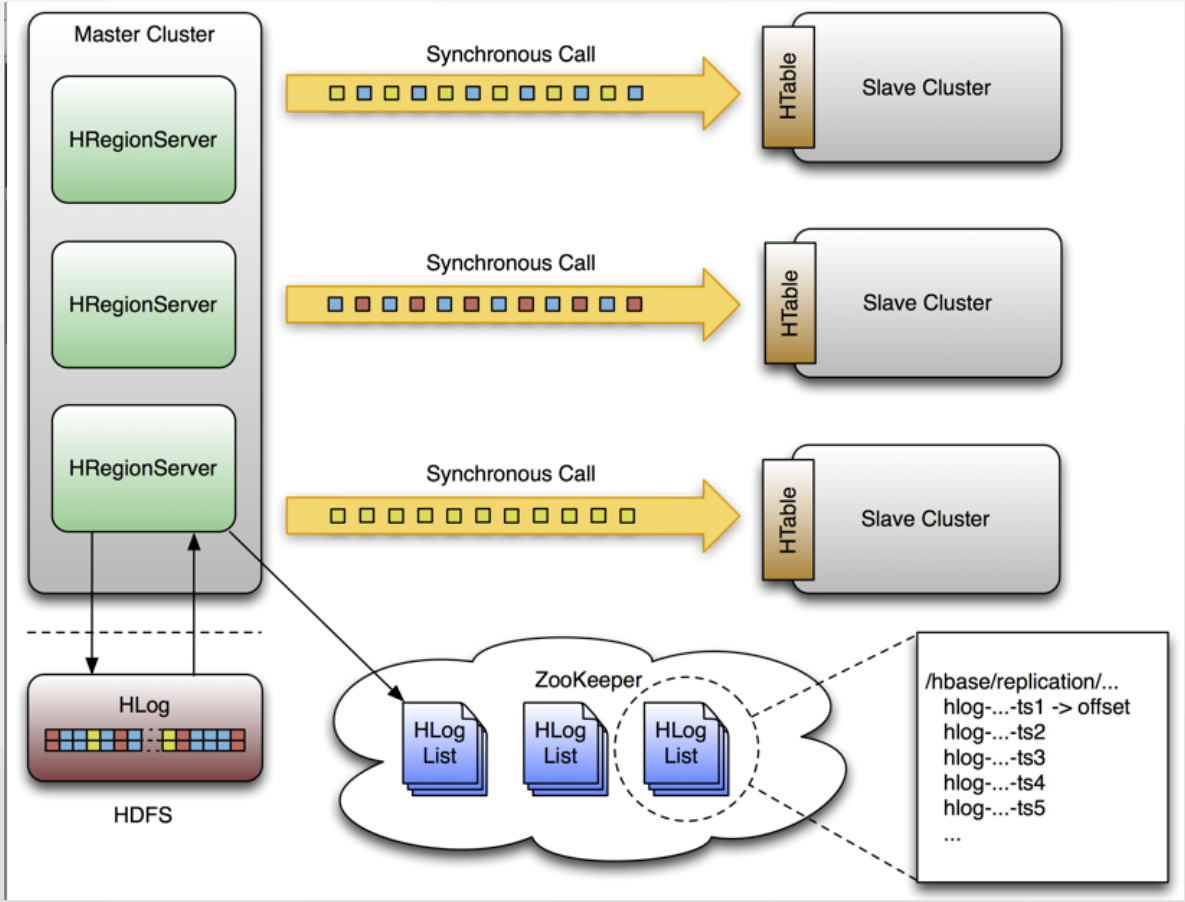

FOLLOWER
SERVER A
v(A)
r[(A,A)]
recv B1
v(B)
r[(A,B),(B,B)]
recv C1
v(C)
r[(A,C),(B,B),(C,C)]
v(C)
r[(B,C),(C,C)]
A2
A3
SERVER B
v(B)
r[(B,B)]
recv A1
v(B)
r[(B,B),(A,A)]
v(B)
r[(B,B),(A,B)]
v(C)
r[(B,C),(A,C)]
FOLLOWER
SERVER C
v(C)
r[(C,C)]
...
v(C)
r[(A,C),(B,C)]
LEADER

Oracle's log

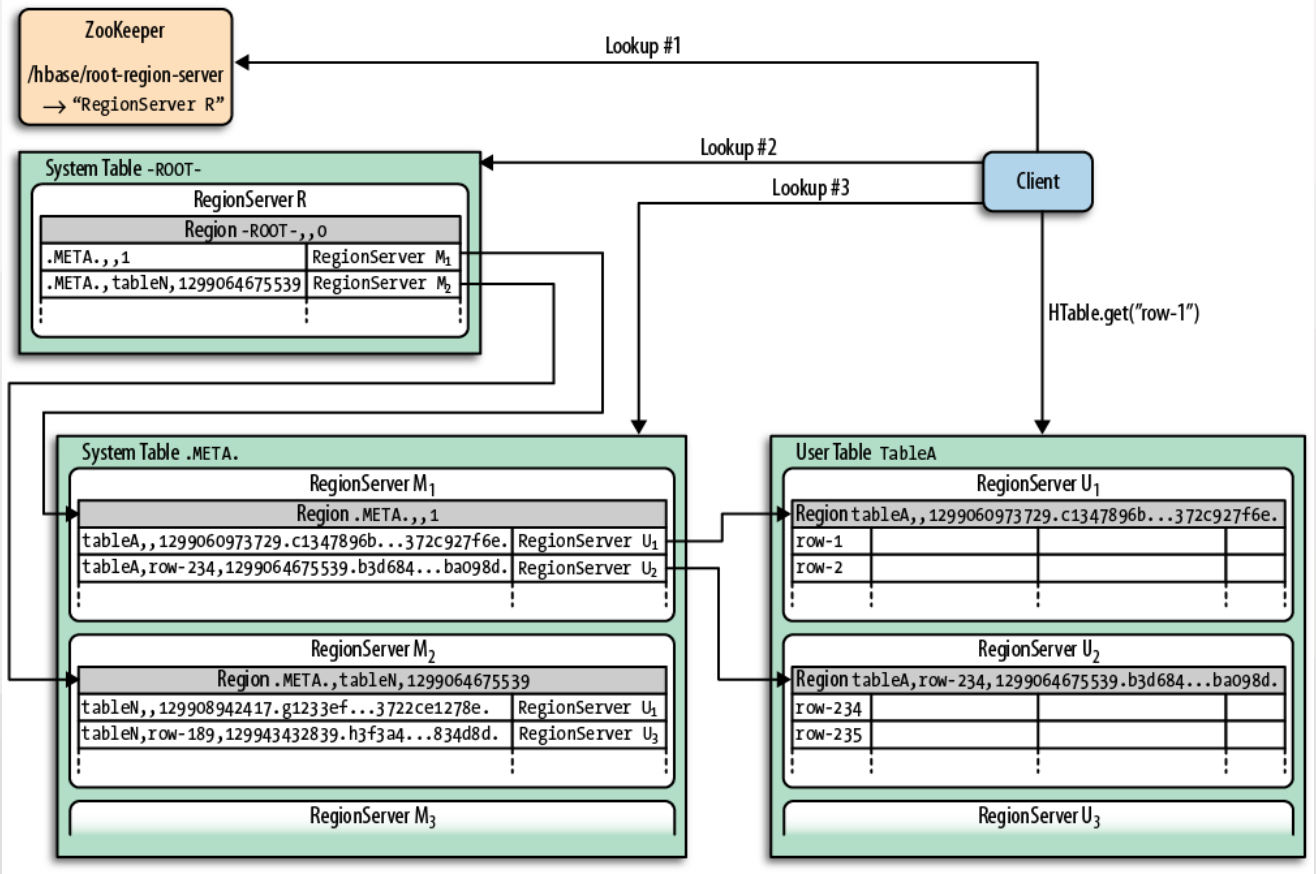

User process
Server process
PGA
Instance
Shared pool
Library cache
Data Dict. cache
SGA
Database buffer cache
Redo log buffer cache
PMON
SMON
DBWR
LGWR
CKPT
Others
Parameter file
Password file
Data files
Control files
Redo log files
Archived log files
Database


Redo Log File Usage
Redo Log Buffer
LGWR
Group 1
Group 2
Group 3
Disk 1
Disk 2
Member
ARCH
Archived Redo Log File
Group 3 is waiting to be used when LGWR has to switch because Group 2 fills.

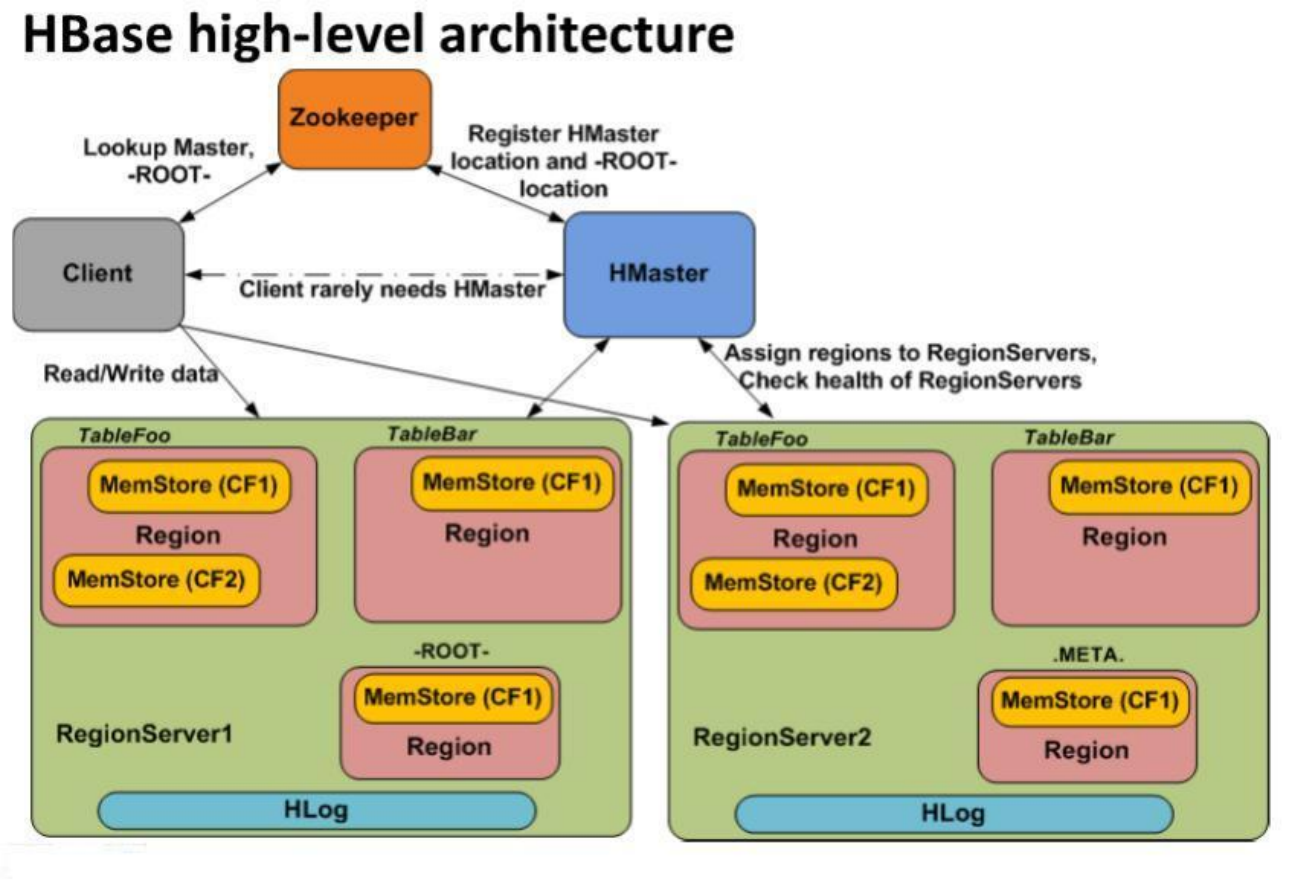
Review


Master Cluster
HRegionServer
HRegionServer
HRegionServer
Synchronous Call
Synchronous Call
Synchronous Call
HTable
Slave Cluster
HTable
Slave Cluster
HTable
Slave Cluster
HLog
HDFS
ZooKeeper
HLog List
HLog List
HLog List
/hbase/replication/...
hlog-...-ts1 -> offset
hlog-...-ts2
hlog-...-ts3
hlog-...-ts4
hlog-...-ts5
...

Review


ZooKeeper
/hbase/root-region-server
→ "RegionServer R"
Lookup #1
Lookup #2
Lookup #3
Client
HTable.get("row-1")
System Table -ROOT-
RegionServer R
Region -ROOT-,,0
.META.,,1
RegionServer M1
.META.,tableN,1299064675539
RegionServer M2
System Table .META.
RegionServer M1
Region .META.,,1
tableA,,1299060973729.c1347896b...372c927f6e.
RegionServer U1
tableA,row-234,1299064675539.b3d684...ba098d.
RegionServer U2
RegionServer M2
Region .META.,tableN,1299064675539
tableN,,129908942417.g1233ef...3722ce1278e.
RegionServer U1
tableN,row-189,129943432839.h3f3a4...834d8d.
RegionServer U3
RegionServer M3
User Table TableA
RegionServer U1
Region tableA,,1299060973729.c1347896b...372c927f6e.
row-1
row-2
RegionServer U2
Region tableA,row-234,1299064675539.b3d684...ba098d.
row-234
row-235
RegionServer U3

Review

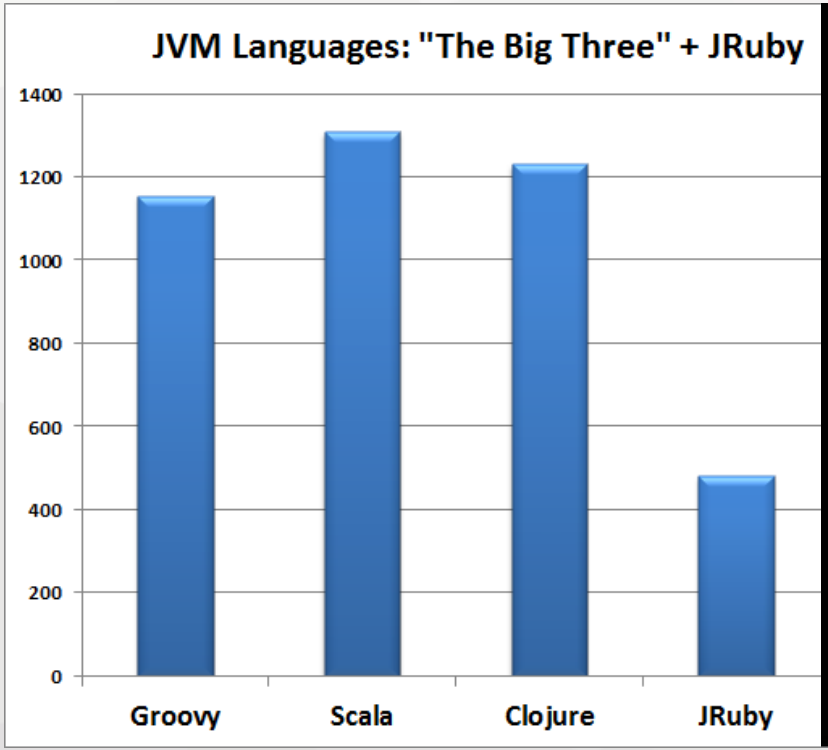

HBase high-level architecture
Zookeeper
Lookup Master, -ROOT-
Register HMaster location and -ROOT- location
Client
Client rarely needs HMaster
HMaster
Read/Write data
Assign regions to RegionServers, Check health of RegionServers
TableFoo
TableBar
MemStore (CF1)
Region
MemStore (CF2)
-ROOT-
.META.
RegionServer1
RegionServer2
HLog

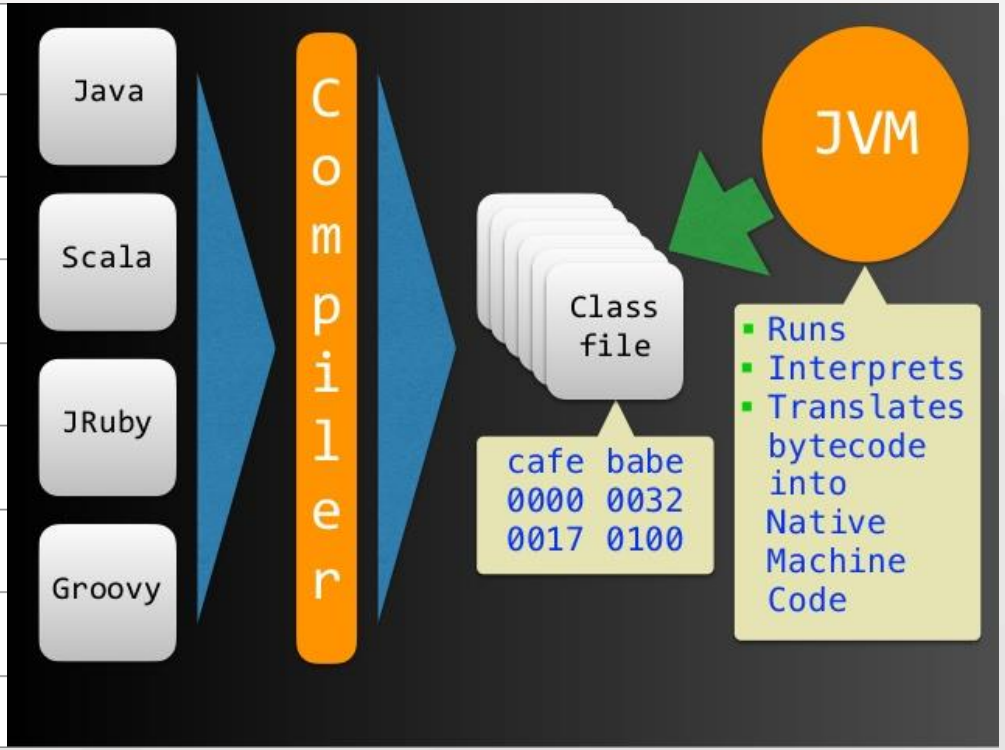
Preliminary Topics
事前準備

JVM


Java
1.0 1996
1.2 1998
1.3 2000
1.4 2002
5.0 2004
6.0 2006
7.0 2011
8.0 2013
Scala
2001
1.0 2003
2.0 2006
2.10 2013
2003
1.0 2007
2.0 2012
FANTOM
2005
THOMPSON 2008
1.0.64 2012
Clojure
2007
1.0 2009
1.4 2012
ceylon
2010
M1 2011
M1 2012
Kotlin
2011
M1 2012
M5 2013
Xtend
1.0 2011
2.3 2012

Big Three


JVM Languages: "The Big Three" + JRuby
1400
1200
1000
800
600
400
200
0
Groovy
Scala
Clojure
JRuby


Java
Scala
JRuby
Groovy
Compiler
Class file
cafe babe
0000 0032
0017 0100
JVM
Runs
Interprets
Translates bytecode into Native Machine Code


Scala
def products = orders.flatMap(o => o.products)
public List<Product>getProducts(){
List<Product> products = new ArrayList<Product>();
for (Order order : order) {
products.addAll(order.getProducts());
}
return products;
}
Java

BUILD TOOL – SBT

http://www.scala-sbt.org/

```
name := "hello world"

version := "0.0.1"

scalaVersion := "2.11.1"

resolvers ++= Seq (
Resolver.mavenLocal,
Resolver.sonatypeRepo ("releases"),
Resolver.typesafeRepo ("releases")
)

libraryDependencies ++=
Seq ("org.scala-lang" % "scala-compiler" % "2.11.1")

addSbtPlugin ("com.typesafe.play" % "sbt-plugin" % "2.3.1")
```

Sbt


```xml
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">
    <modelVersion>4.0.0</modelVersion>
    <groupId>info.solidsoft.rnd</groupId>
    <artifactId>spock-10-groovy-24-gradle-maven</artifactId>
    <version>0.0.1-SNAPSHOT</version>
    <properties>
        <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
        <surefire.version>2.18.1</surefire.version>
    </properties>
    <build>
        <plugins>
            <plugin>
                <groupId>org.codehaus.gmavenplus</groupId>
                <artifactId>gmavenplus-plugin</artifactId>
                <version>1.4</version>
                <executions>
                    <execution>
                        <goals>
                            <goal>compile</goal>
                            <goal>testCompile</goal>
                        </goals>
                    </execution>
                </executions>
            </plugin>
            <plugin>
                <artifactId>maven-surefire-plugin</artifactId>
                <version>${surefire.version}</version>
                <configuration>
                    <includes>
                        <include>**/*Spec.java</include> <!-- Yes, .java extension -->
                        <include>**/*Test.java</include> <!-- Just in case having "normal" JUnit tests -->
                    </includes>
                </configuration>
            </plugin>
        </plugins>
    </build>
    <dependencies>
        <dependency>
            <groupId>org.codehaus.groovy</groupId>
            <artifactId>groovy-all</artifactId>
            <version>2.4.1</version>
        </dependency>
        <dependency>
            <groupId>org.spockframework</groupId>
            <artifactId>spock-core</artifactId>
            <version>1.0-groovy-2.4</version>
            <scope>test</scope>
        </dependency>
    </dependencies>
</project>
```

pom.xml

maven

```
apply plugin: 'groovy'

group = "info.solidsoft.rnd"
version = "0.0.1-SNAPSHOT"

repositories {
    mavenCentral()
}

dependencies {
    compile 'org.codehaus.groovy:groovy-all:2.4.1'

    testCompile 'org.spockframework:spock-core:1.0-groovy-2.4'
}
```

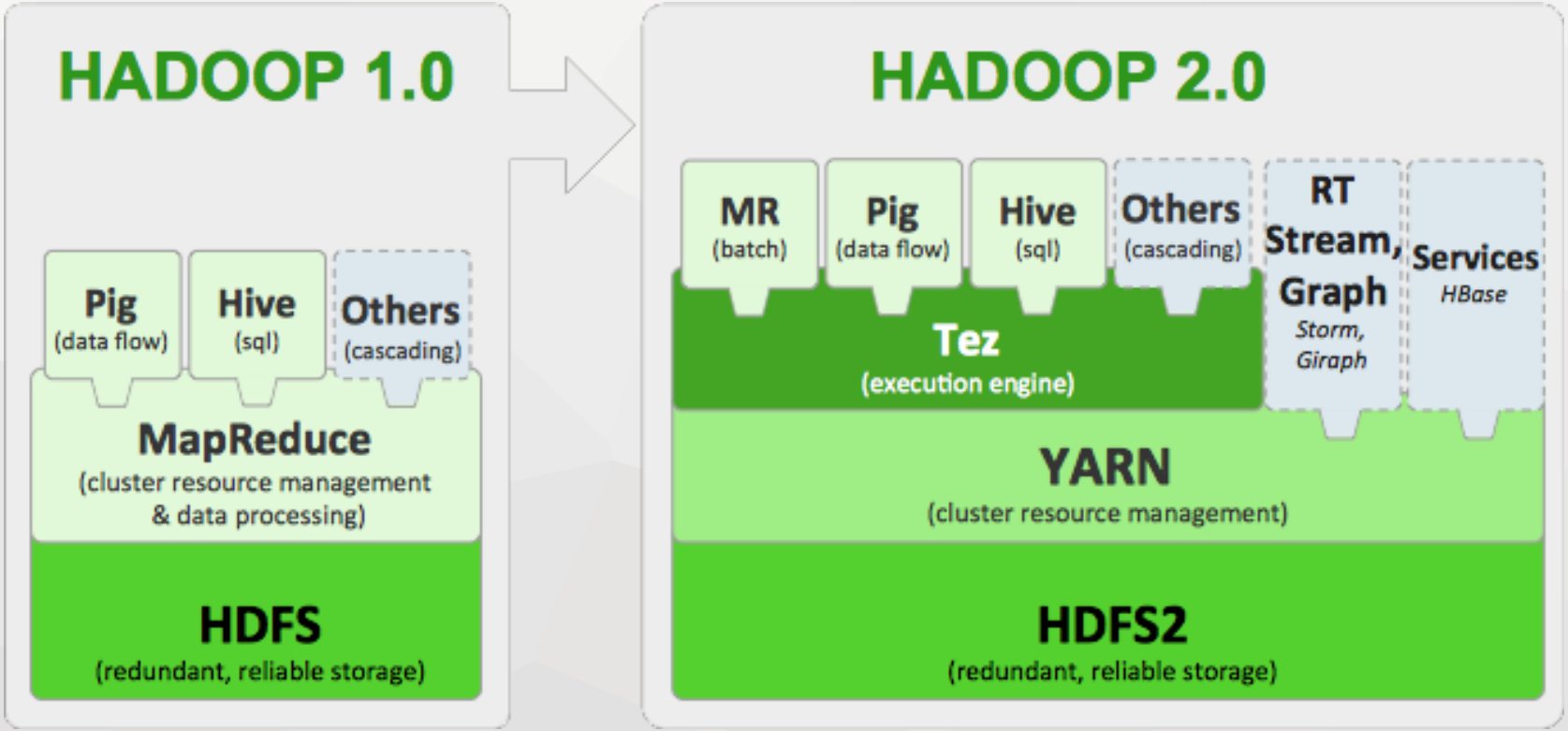
build.gradle

```
rootProject.name = 'spock-10-groovy-24-gradle-maven'
```

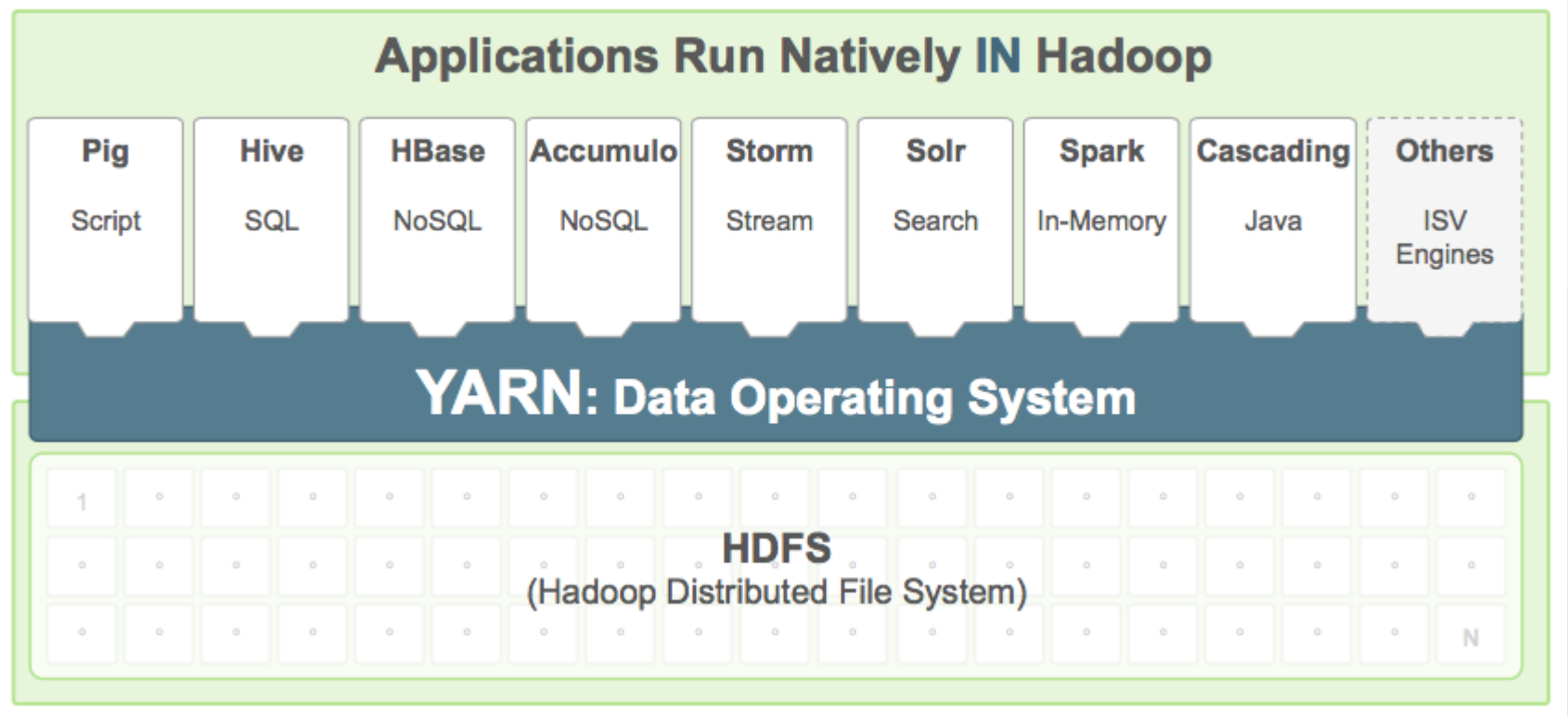
settings.xml

gradle

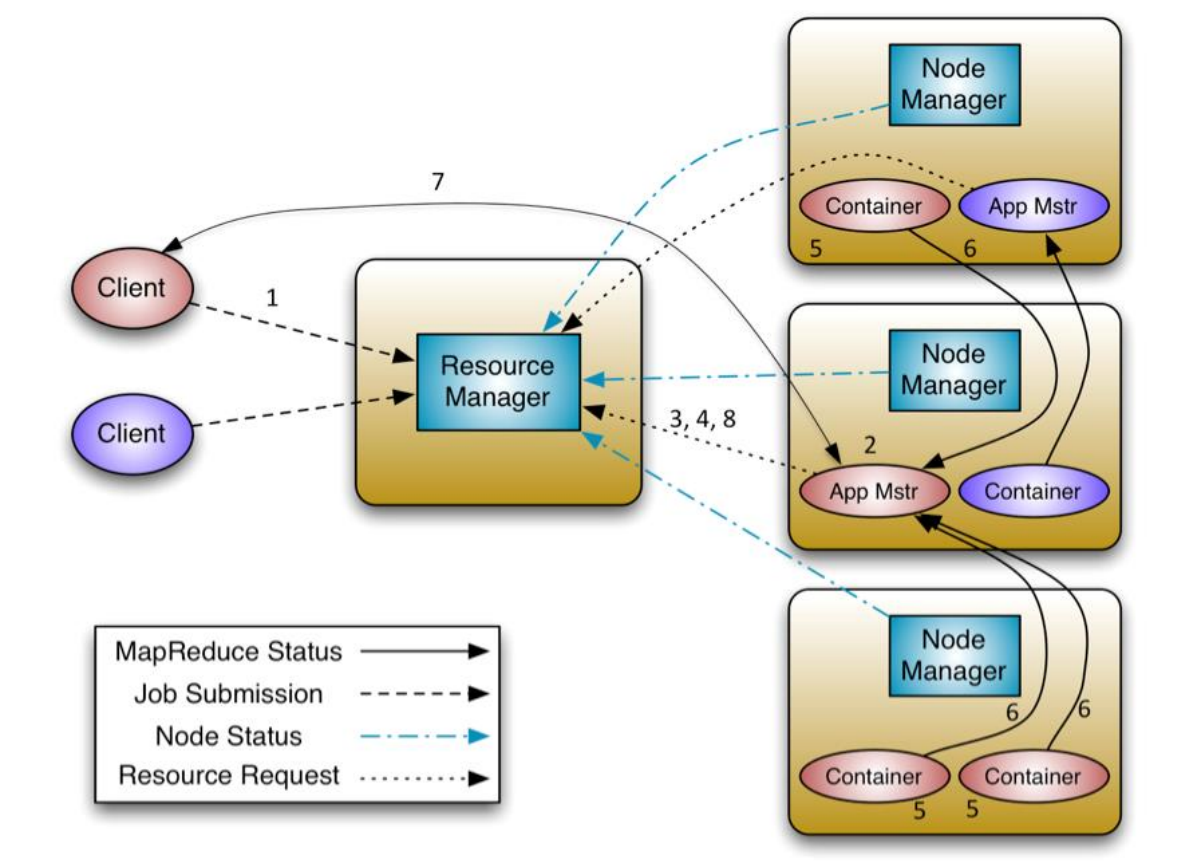
Yarn


HADOOP 1.0
Pig
(data flow)
Hive
(sql)
Others
(cascading)
MapReduce
(cluster resource management
& data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MR
(batch)
Pig
(data flow)
Hive
(sql)
Others
(cascading)
RT
Stream,
Graph
Storm,
Giraph
Services
HBase
Tez
(execution engine)
YARN
(cluster resource management)
HDFS2
(redundant, reliable storage)

Yarn

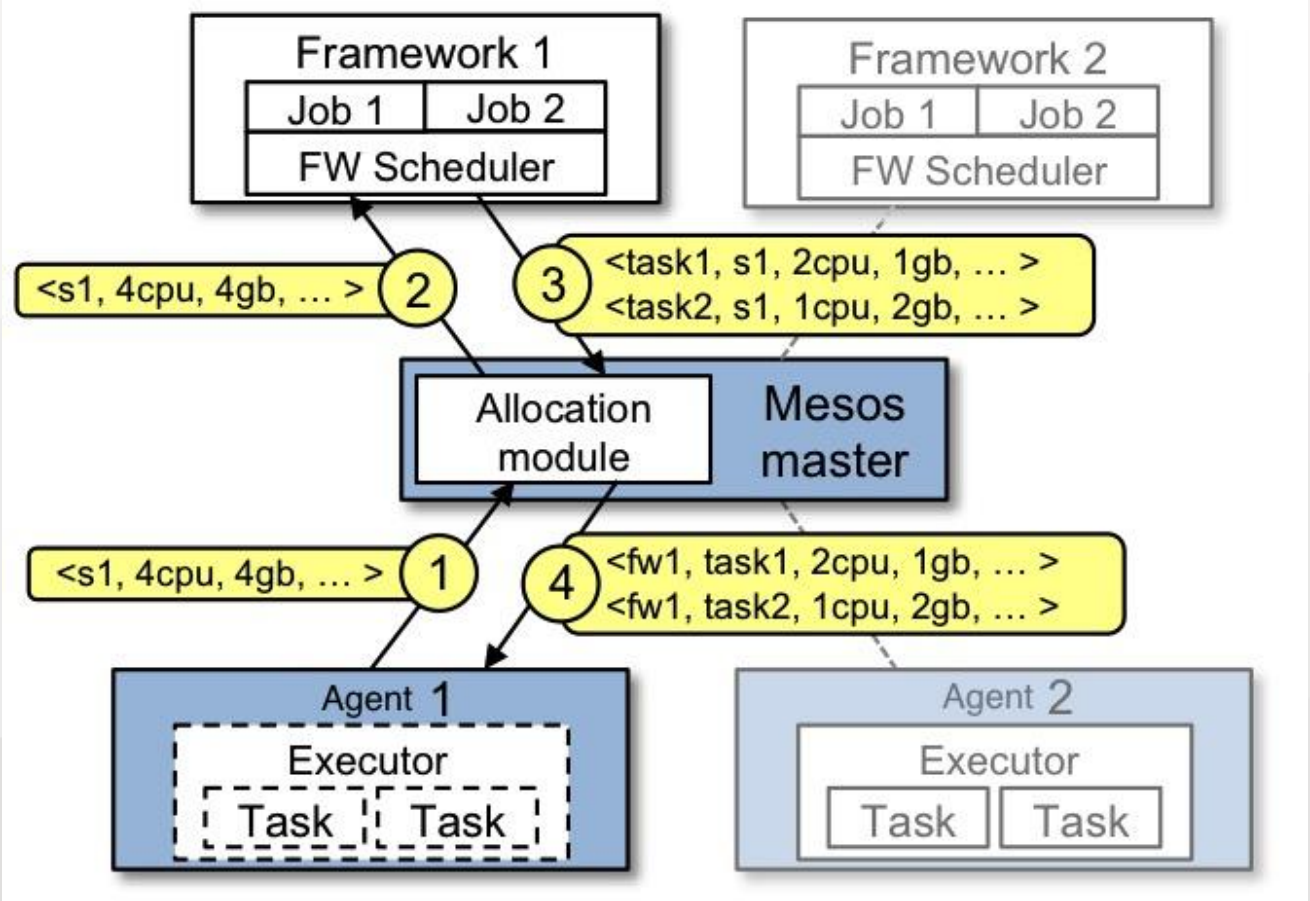

Applications Run Natively IN Hadoop
Pig
Script
Hive
SQL
HBase
NoSQL
Accumulo
NoSQL
Storm
Stream
Solr
Search
Spark
In-Memory
Cascading
Java
Others
ISV Engines
YARN: Data Operating System
1
N
HDFS
(Hadoop Distributed File System)

Yarn



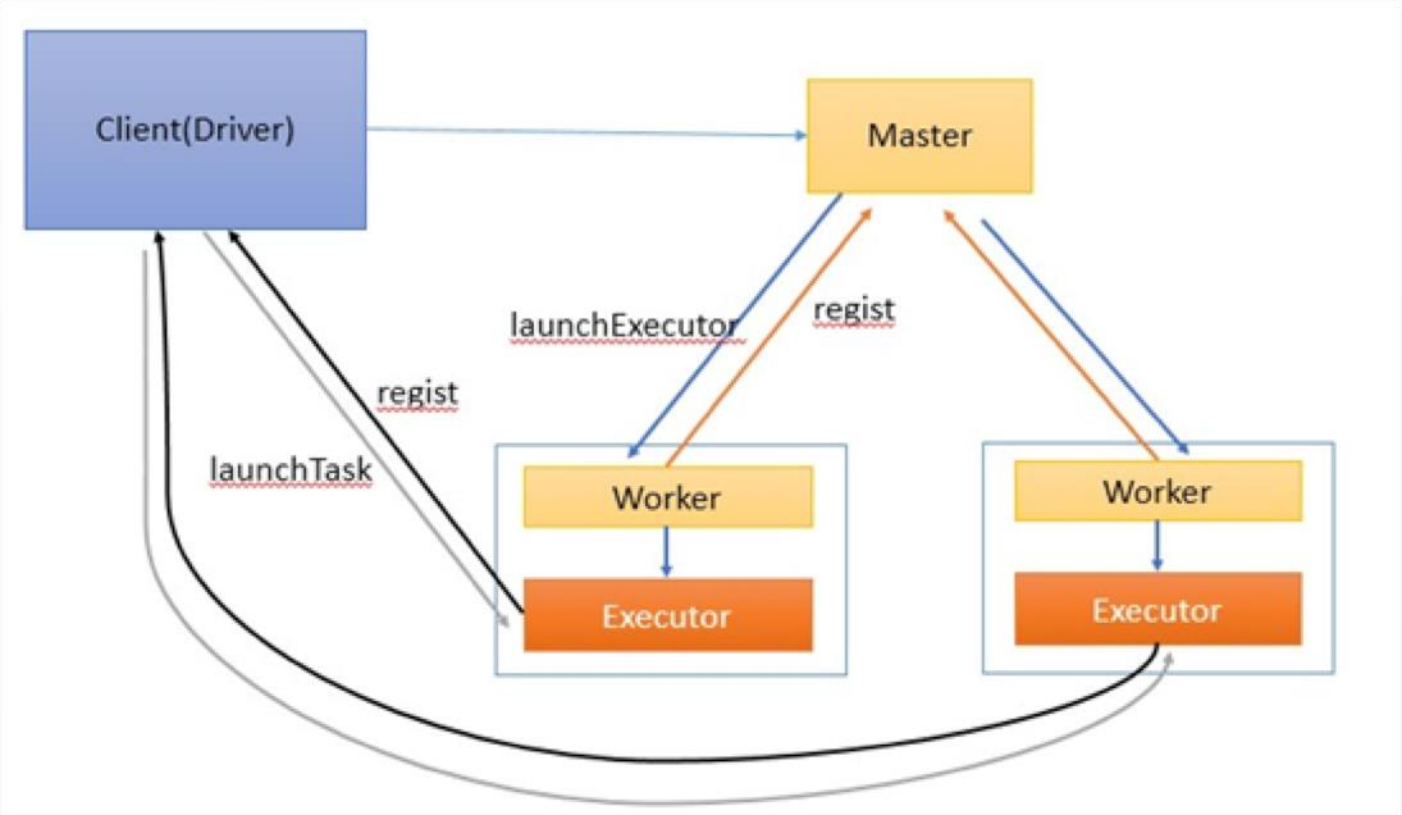
https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/WritingYarnApplications.html

Mesos

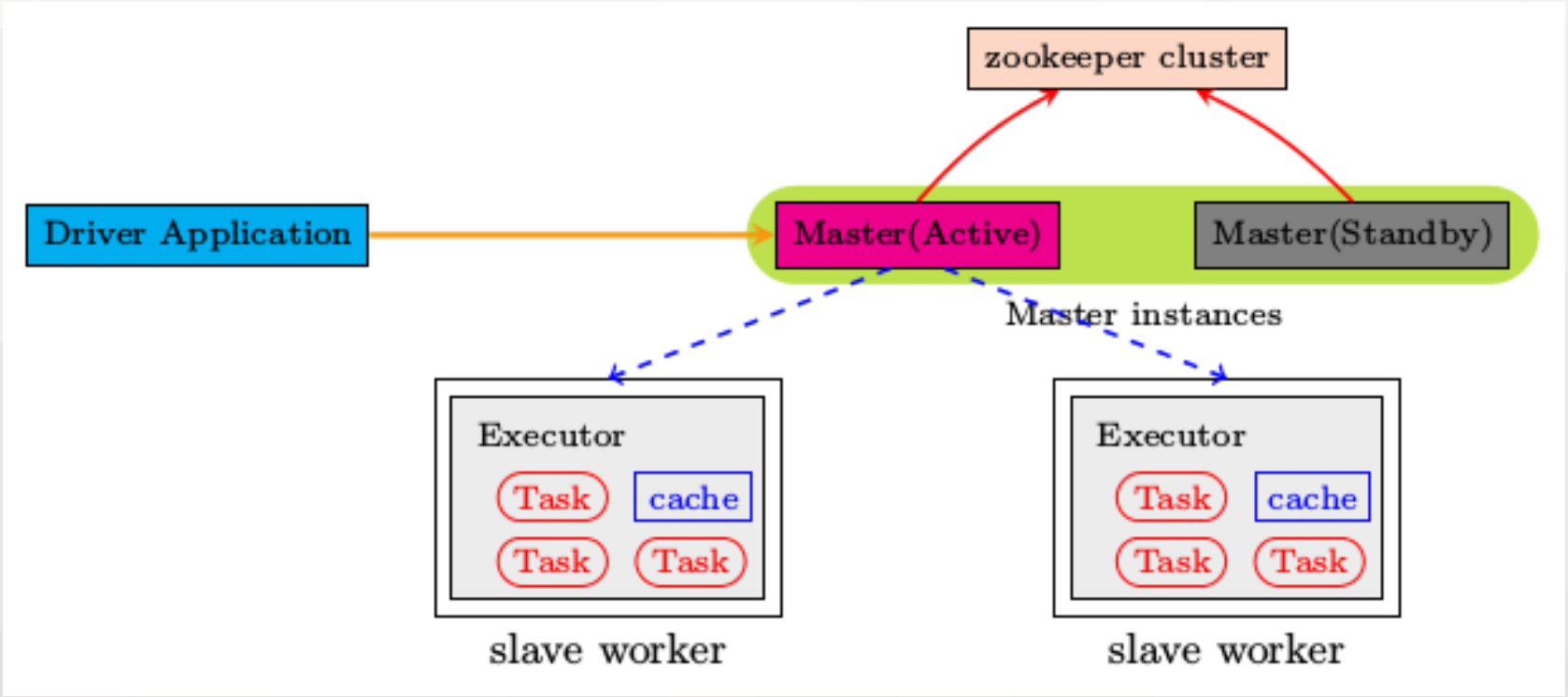

Framework 1
Job 1
Job 2
FW Scheduler
Framework 2
Job 1
Job 2
FW Scheduler
<s1, 4cpu, 4gb, … >
2
3
<task1, s1, 2cpu, 1gb, … >
<task2, s1, 1cpu, 2gb, … >
Allocation module
Mesos master
<s1, 4cpu, 4gb, … >
1
4
<fw1, task1, 2cpu, 1gb, … >
<fw1, task2, 1cpu, 2gb, … >
Agent 1
Executor
Task
Task
Agent 2
Executor
Task
Task

Spark Environment

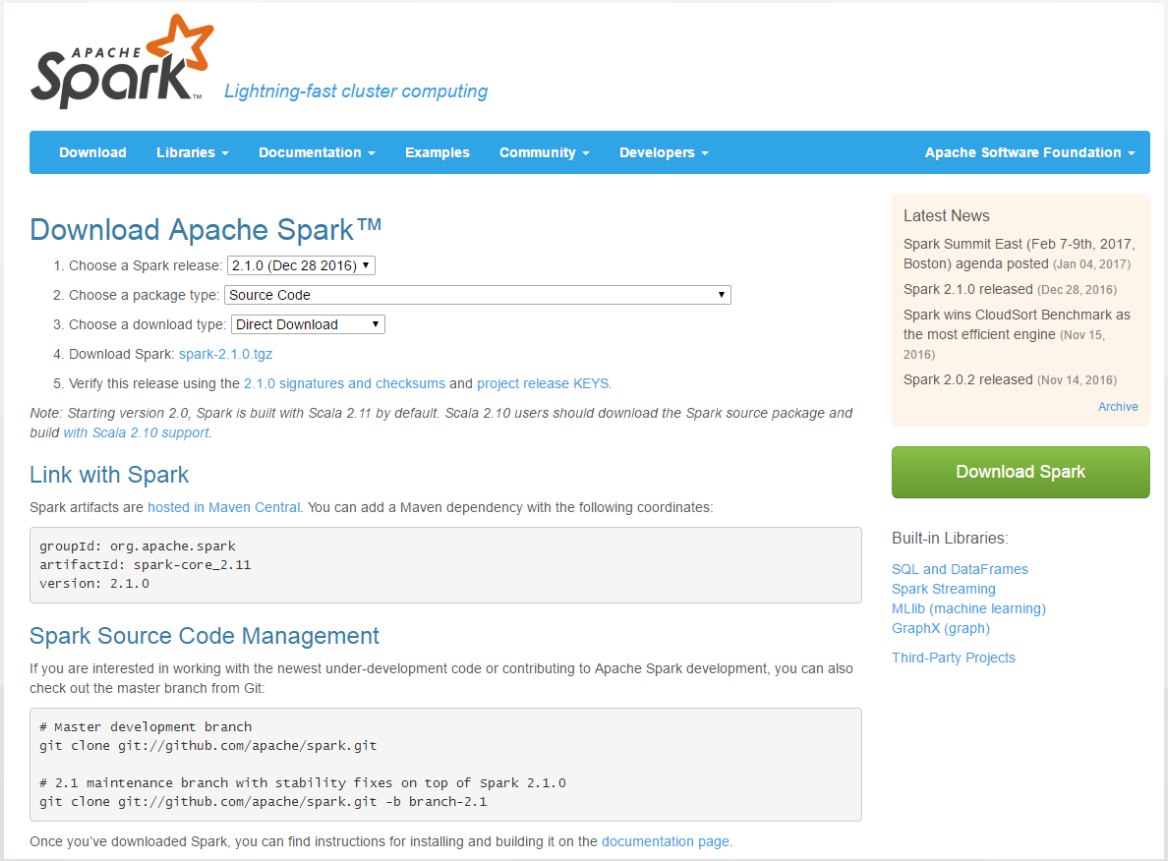
Spark Standalone



Spark Standalone HA



Spark Source Compile



APACHE Spark™ *Lightning-fast cluster computing*

Download | Libraries | Documentation | Examples | Community | Developers | Apache Software Foundation

Download Apache Spark™

1. Choose a Spark release: 2.1.0 (Dec 28 2016)
2. Choose a package type: Source Code
3. Choose a download type: Direct Download
4. Download Spark: spark-2.1.0.tgz
5. Verify this release using the 2.1.0 signatures and checksums and project release KEYS.

Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

Link with Spark

Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.11
version: 2.1.0
```

Spark Source Code Management

If you are interested in working with the newest under-development code or contributing to Apache Spark development, you can also check out the master branch from Git:

```
# Master development branch
git clone git://github.com/apache/spark.git

# 2.1 maintenance branch with stability fixes on top of Spark 2.1.0
git clone git://github.com/apache/spark.git -b branch-2.1
```

Once you've downloaded Spark, you can find instructions for installing and building it on the documentation page.

Latest News

Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)

Spark 2.1.0 released (Dec 28, 2016)

Spark wins CloudSort Benchmark as the most efficient engine (Nov 15, 2016)

Spark 2.0.2 released (Nov 14, 2016)

Archive

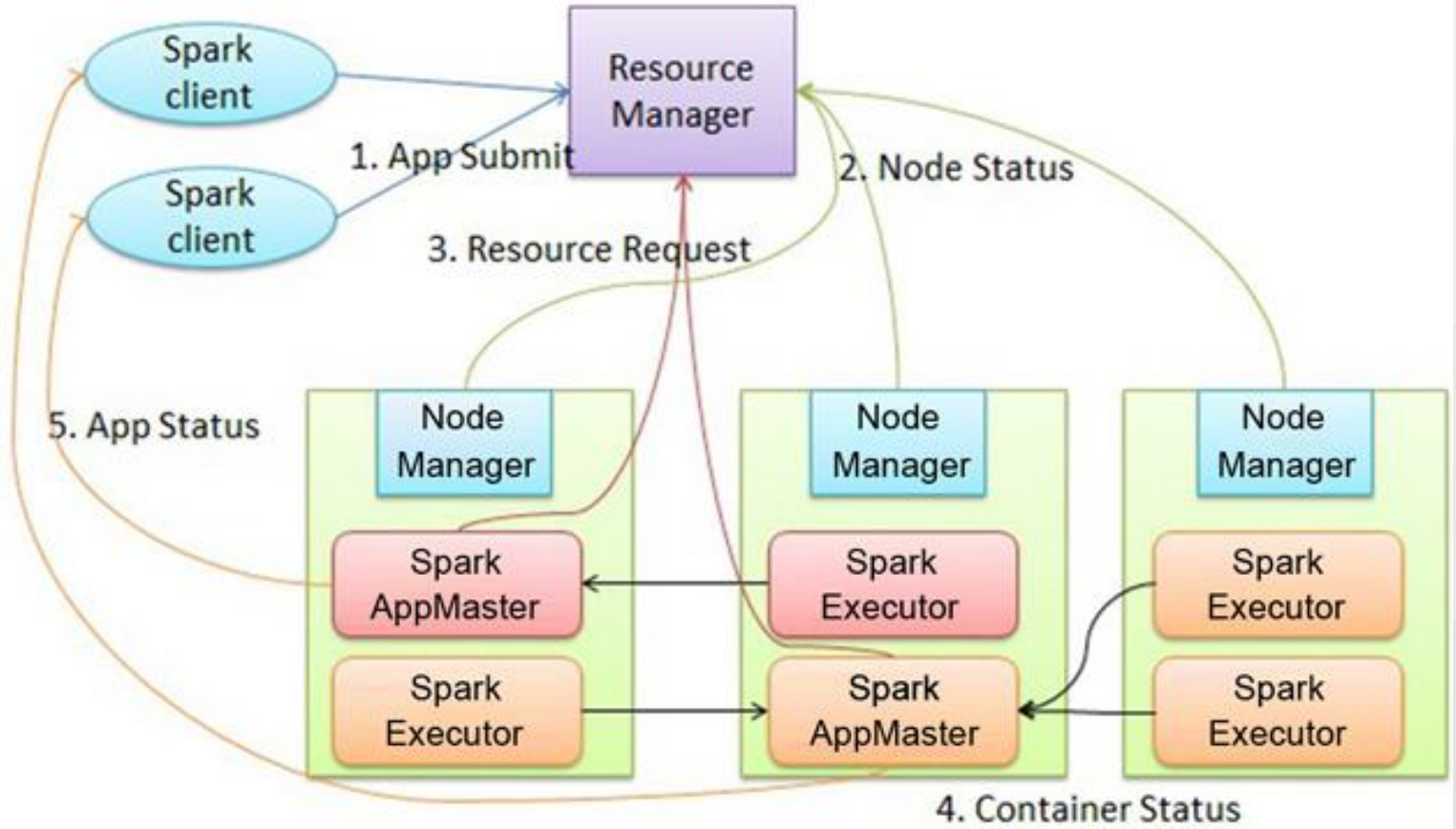
Download Spark

Built-in Libraries:

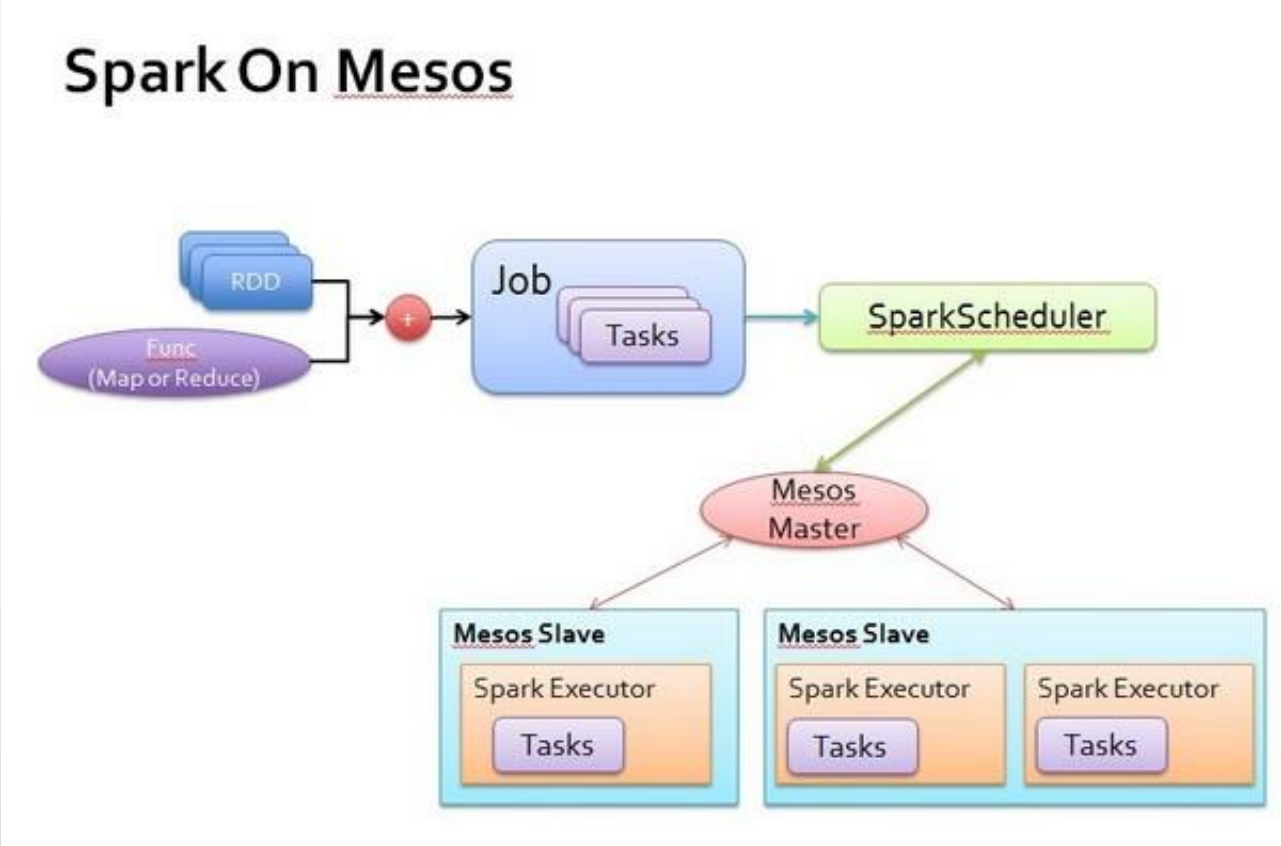
SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Third-Party Projects

Spark On Yarn

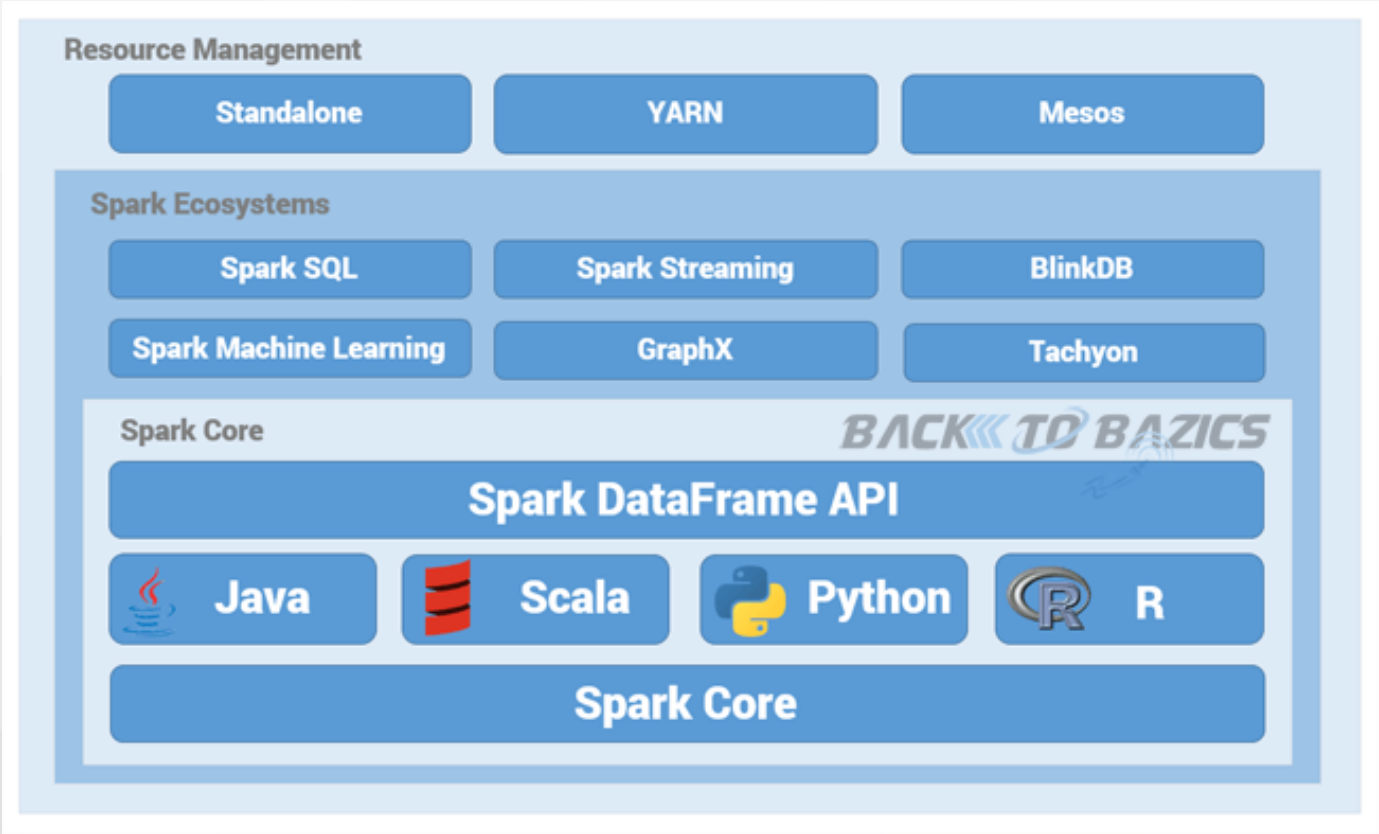


Spark On Mesos

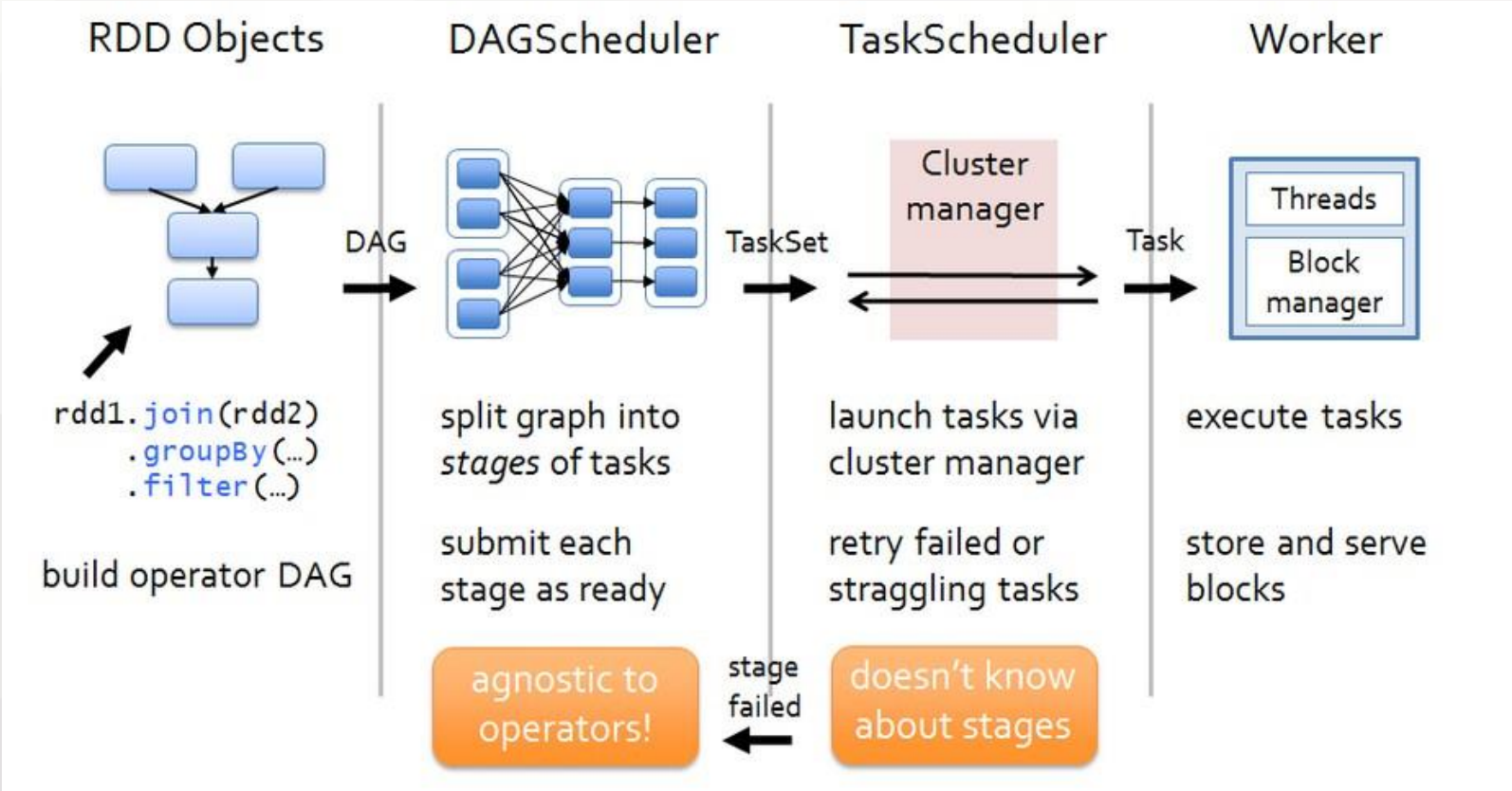


Spark Architecture

Ecosystems

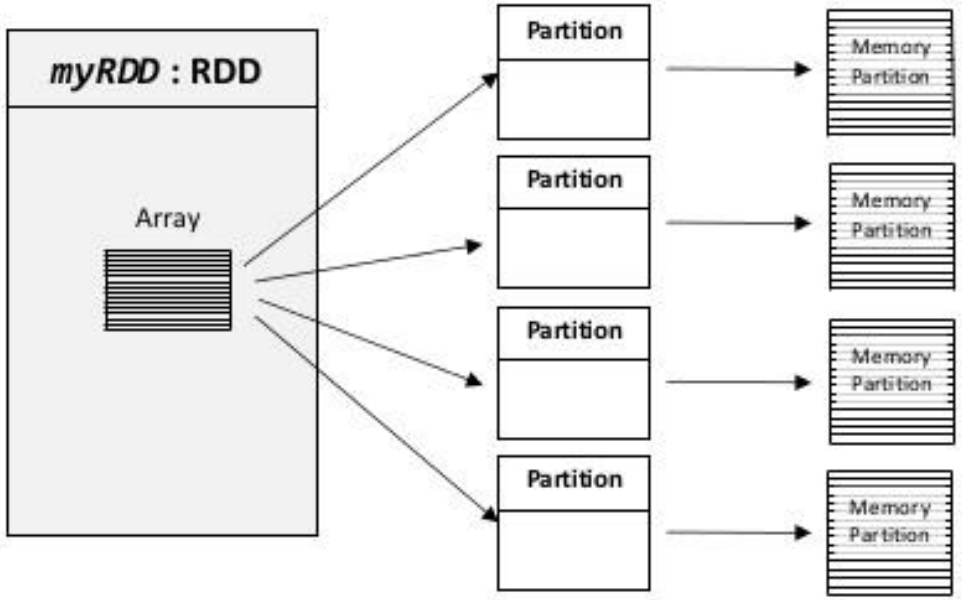

Resource Management
Standalone
YARN
Mesos
Spark Ecosystems
Spark SQL
Spark Streaming
BlinkDB
Spark Machine Learning
GraphX
Tachyon
Spark Core
BACK TO BAZICS
Spark DataFrame API
Java
Scala
Python
R
Spark Core

Architecture



Spark RDD

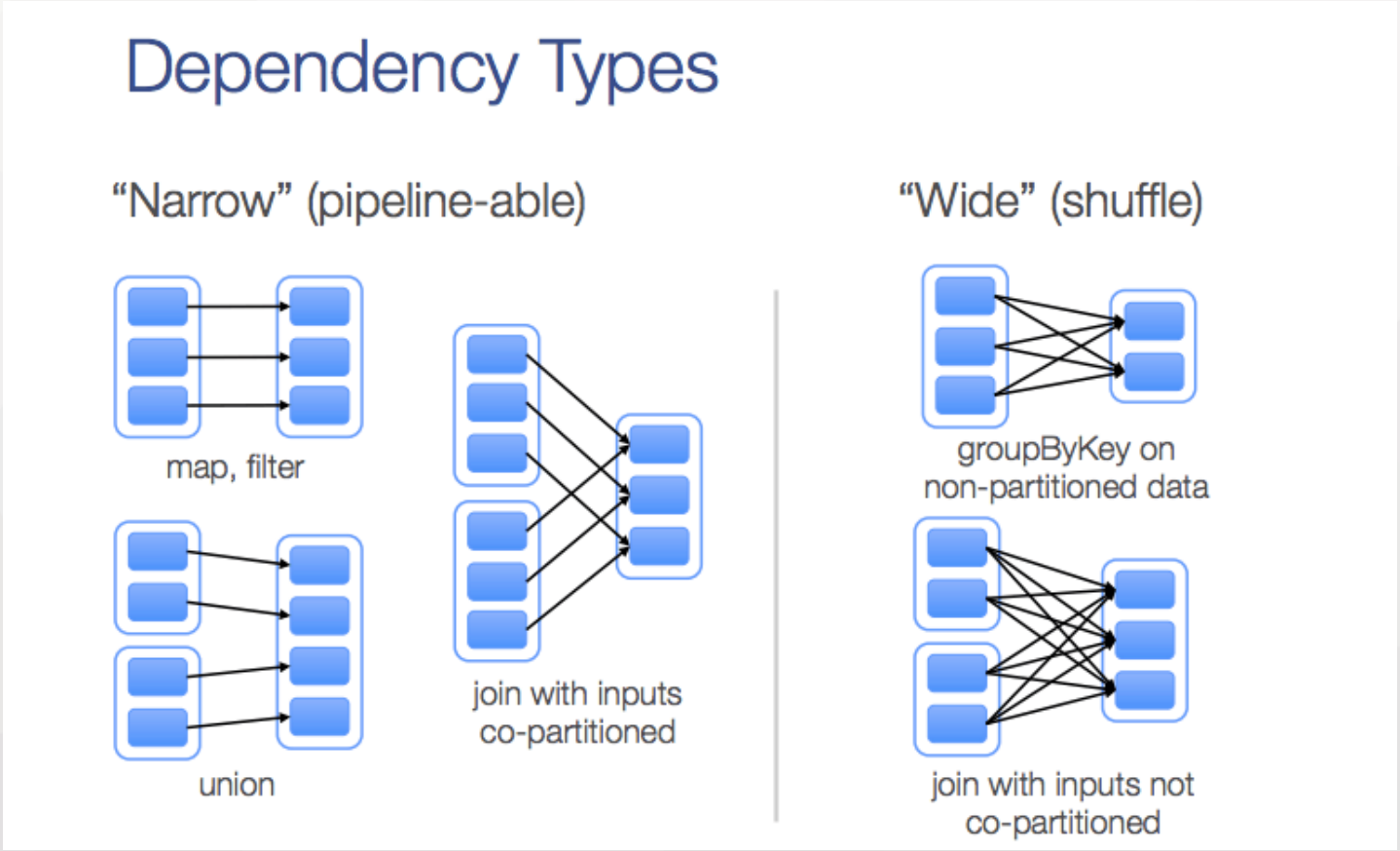
What is an RDD?



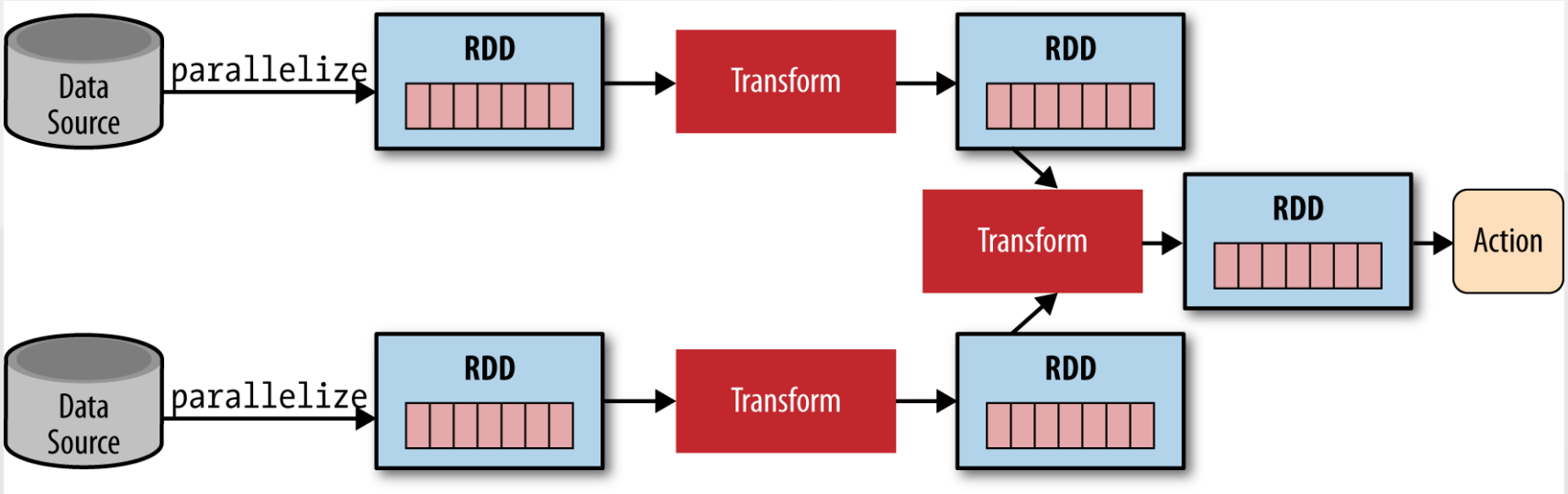
Some RDD Characteristics

- Hold references to Partition objects
- Each Partition object references a subset of your data
- Partitions are assigned to nodes on your cluster
- Each partition/split will be in RAM (by default)

Dependency Type



Dependency Type




The End