

The Pilot Of Spark

2017.4

XenRon

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

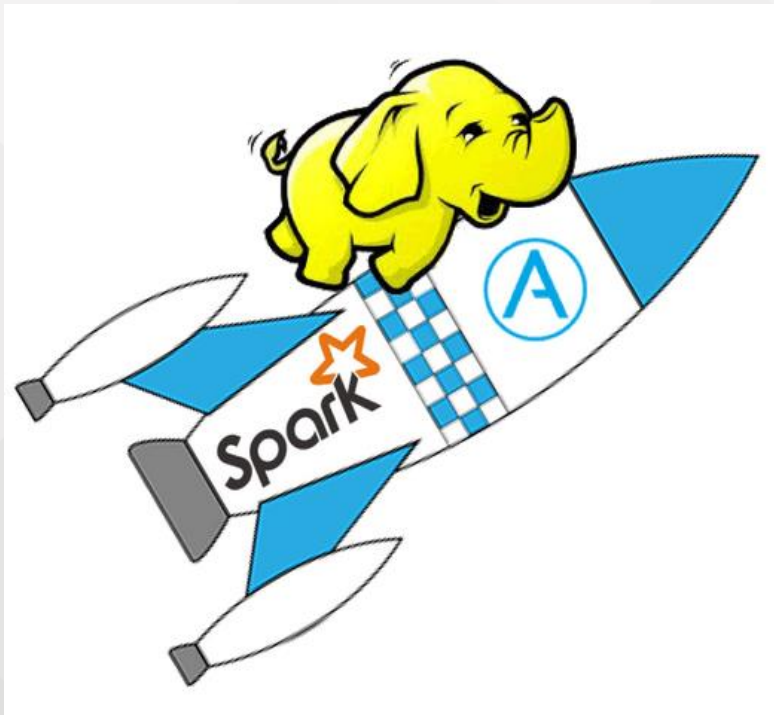
CONTENTS

Preliminary Topics 01

Spark Environment 02

Spark Architecture 03

Spark RDD 04





chew the fat

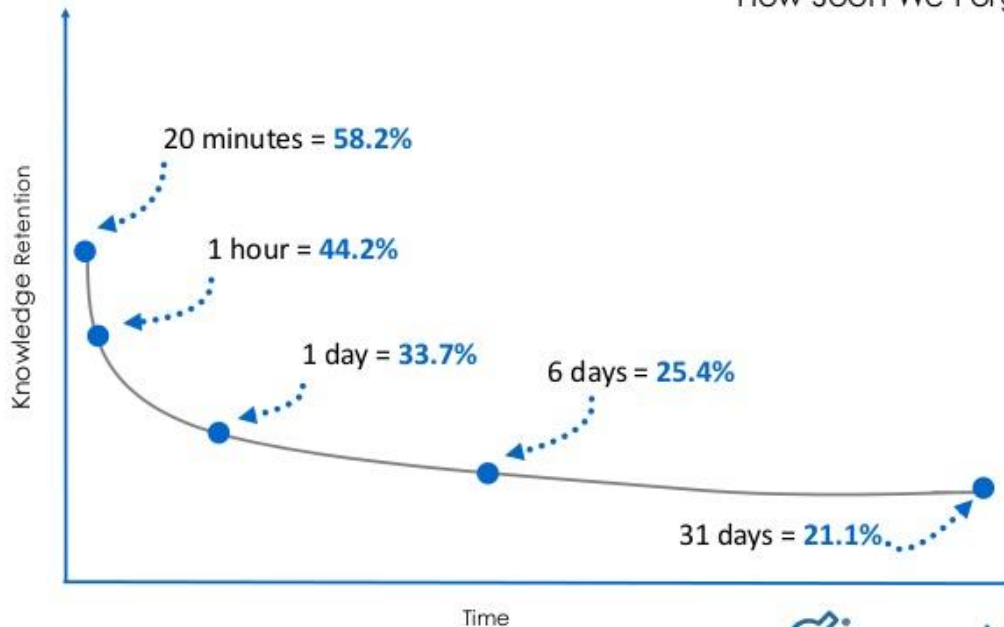
Forgetting Curve

Hermann Ebbinghaus

4

Ebbinghaus Curve

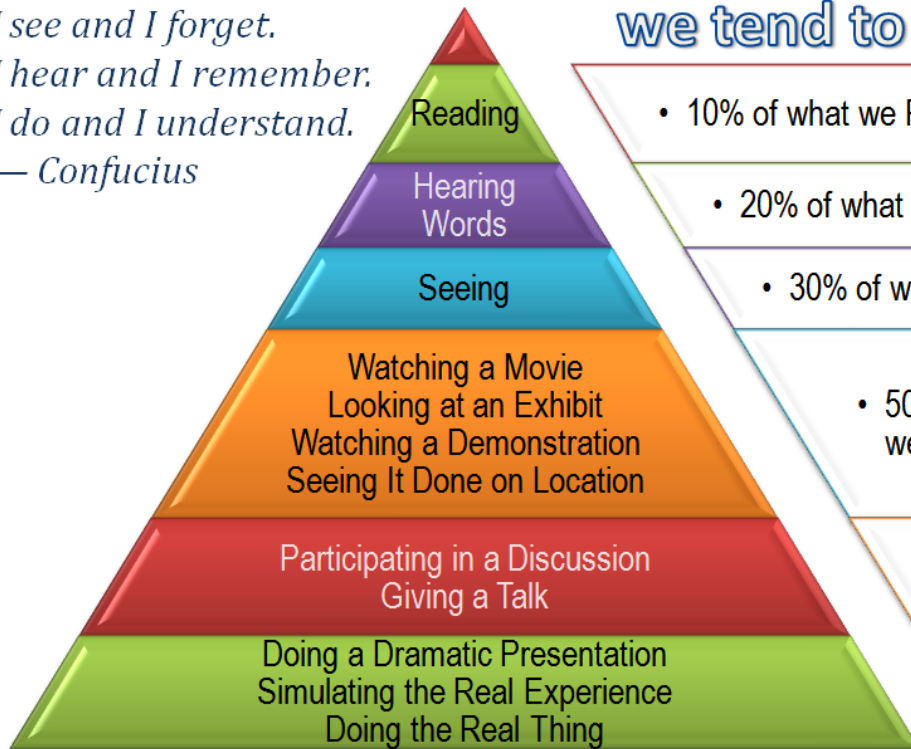
How Soon We Forget



Source: Purdue University Graph

The Cone Of Learning

*I see and I forget.
I hear and I remember.
I do and I understand.
— Confucius*



After 2 weeks,
we tend to remember ...

- 10% of what we READ
- 20% of what we HEAR
- 30% of what we SEE
- 50% of what we SEE & HEAR
- 70% of what we SAY
- 90% of what we SAY & DO

P
a
s
s
i
v
e

A
c
t
i
v
e

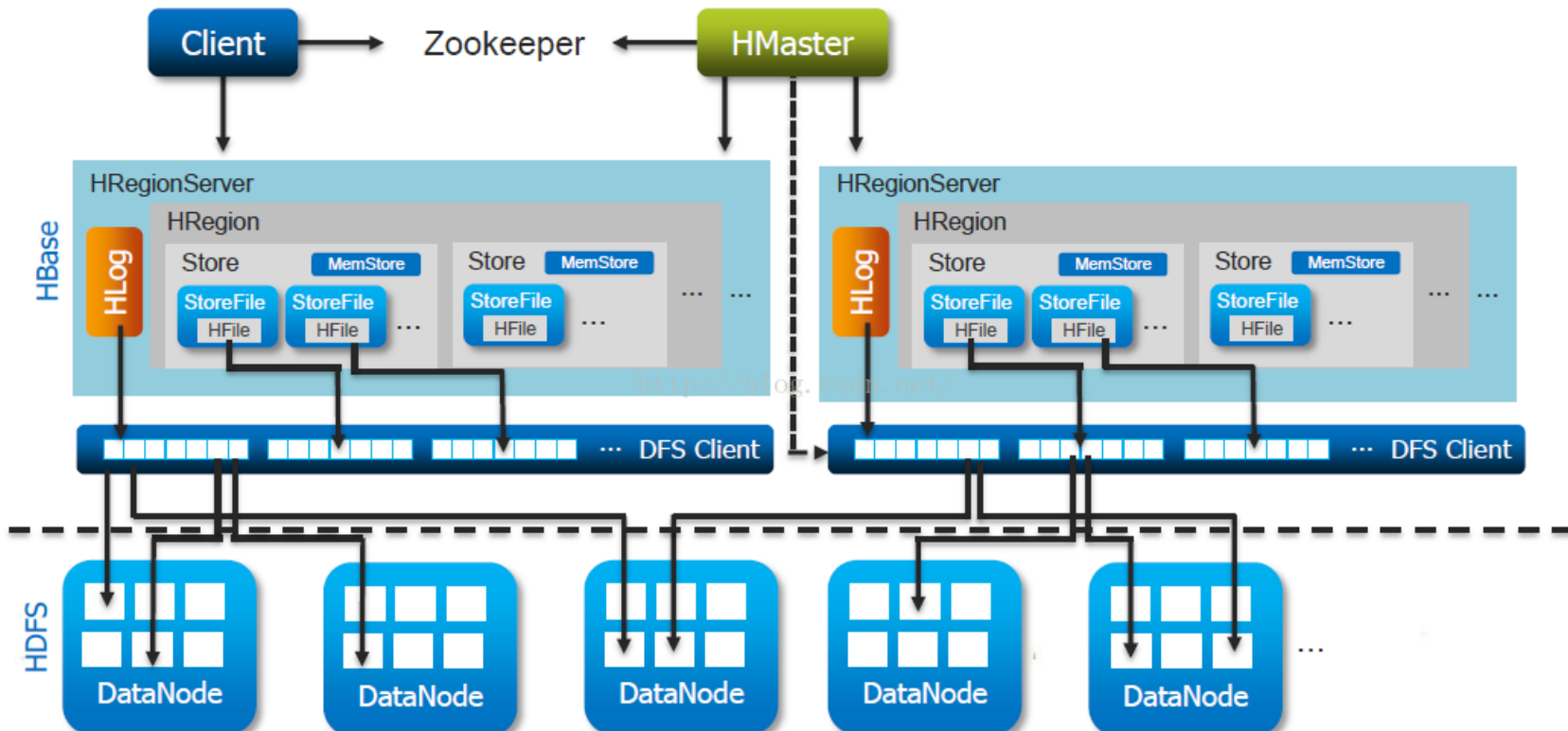
Source: Edgar Dale (1969)

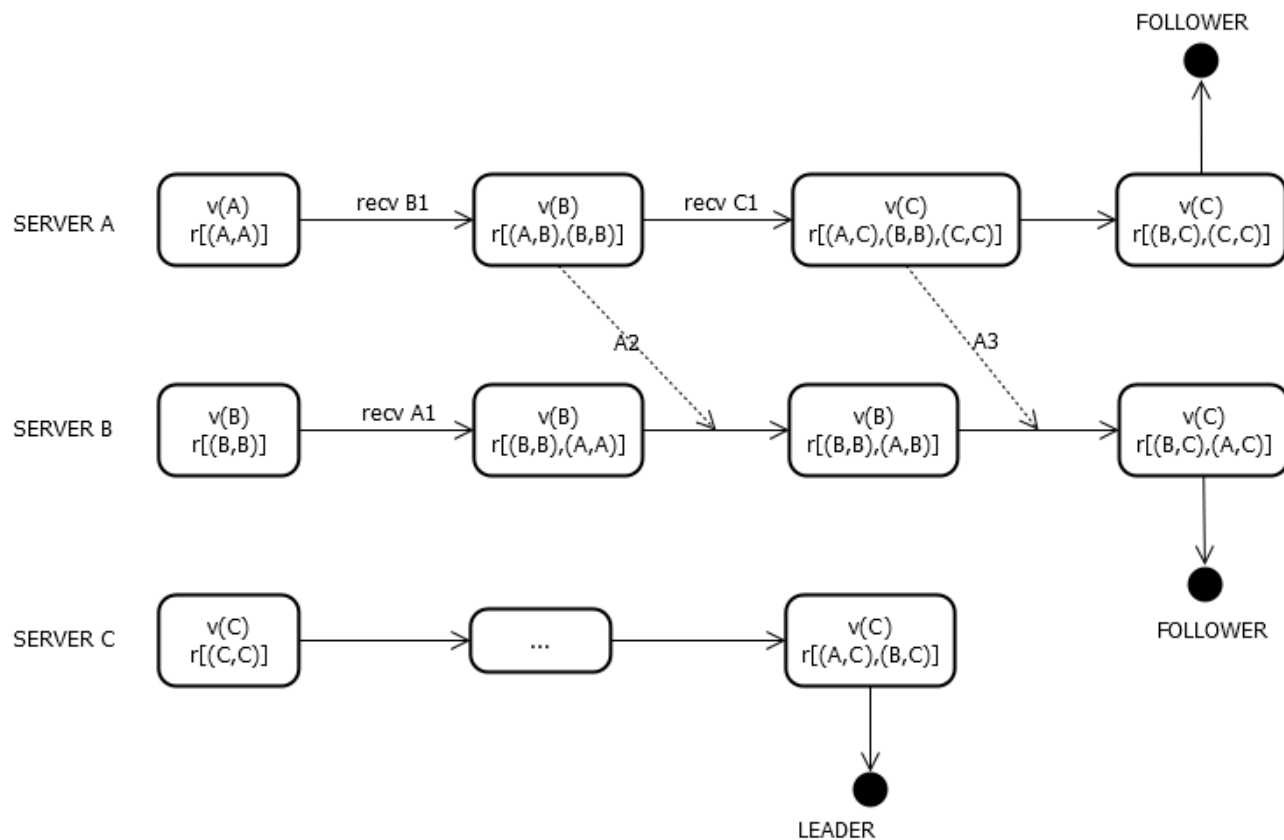


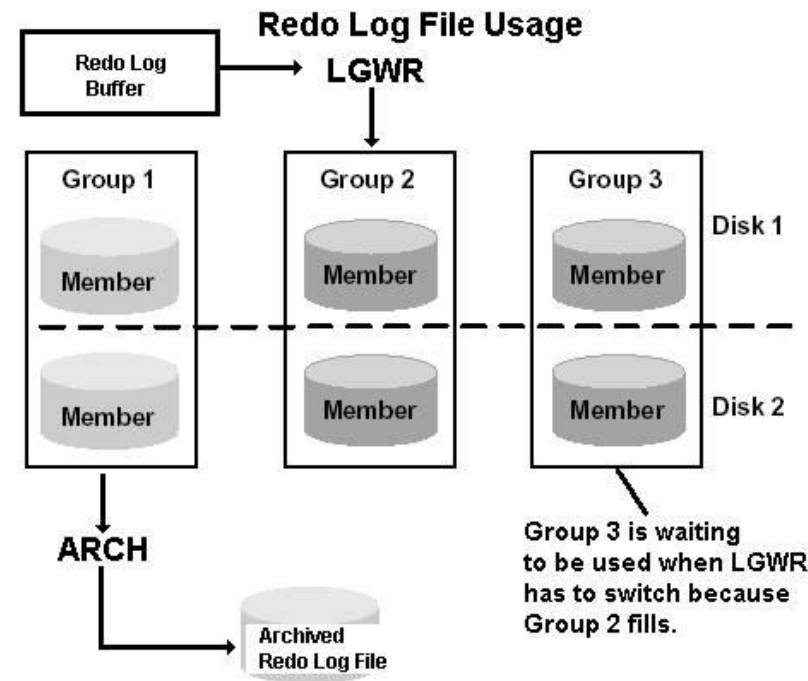
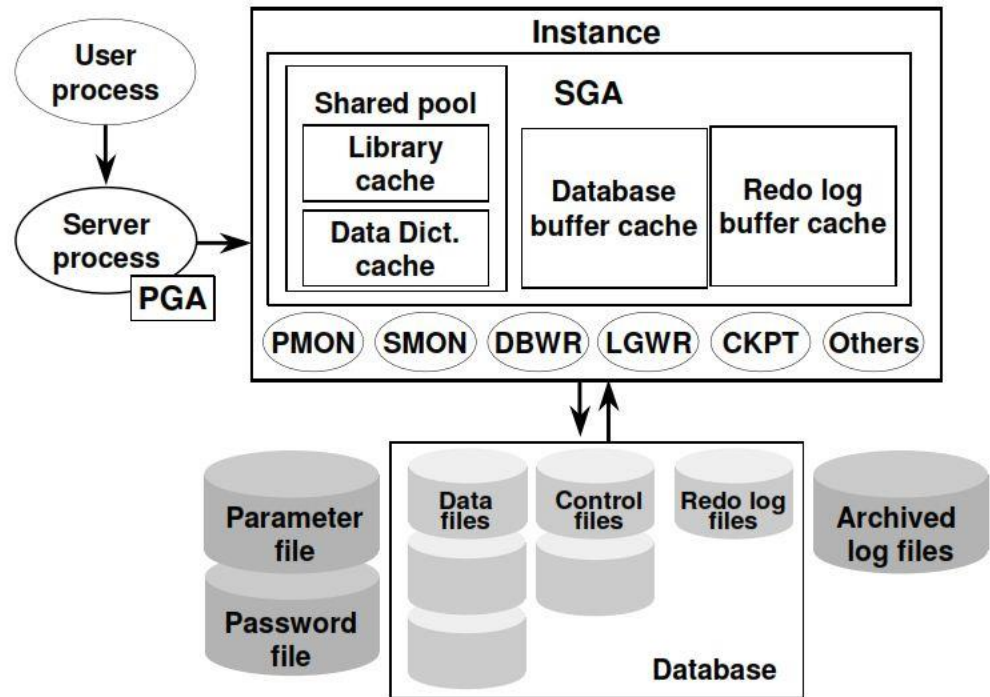
Review



Review

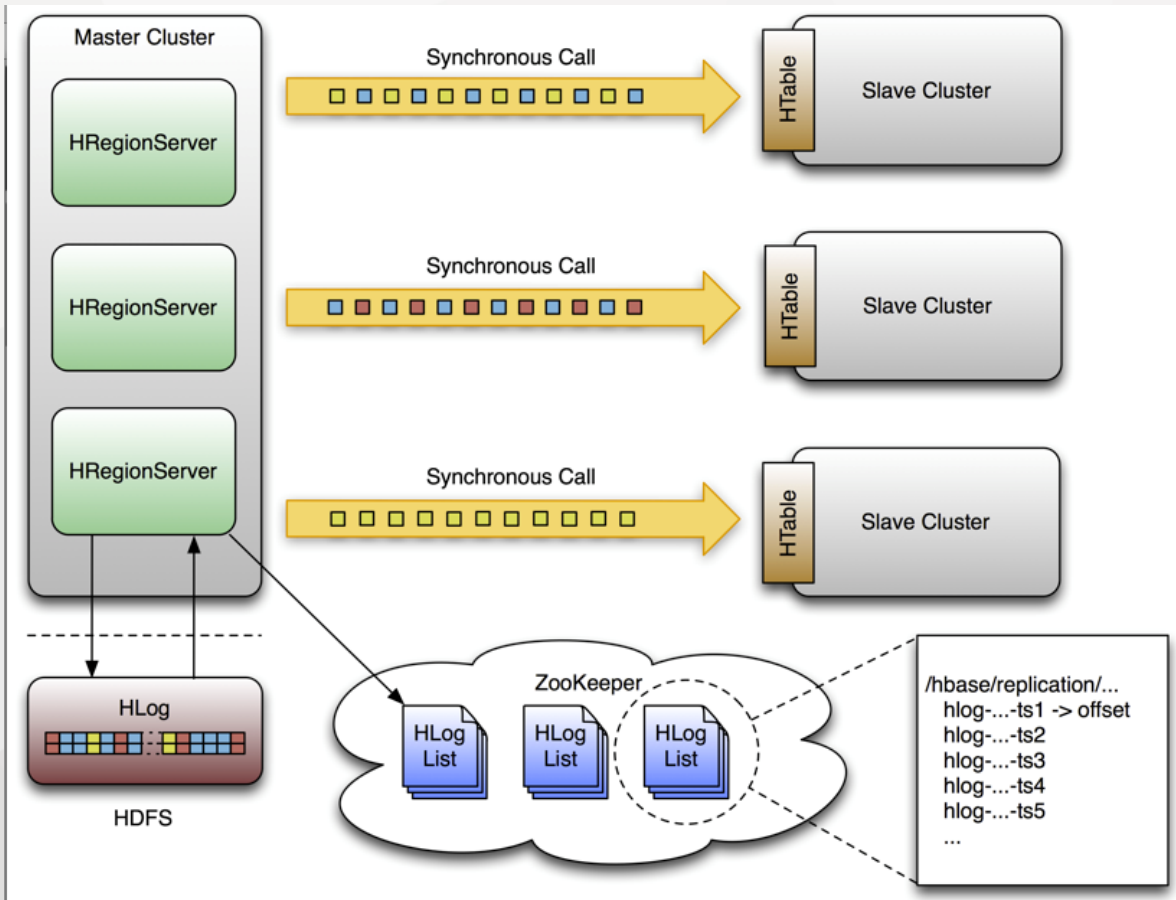


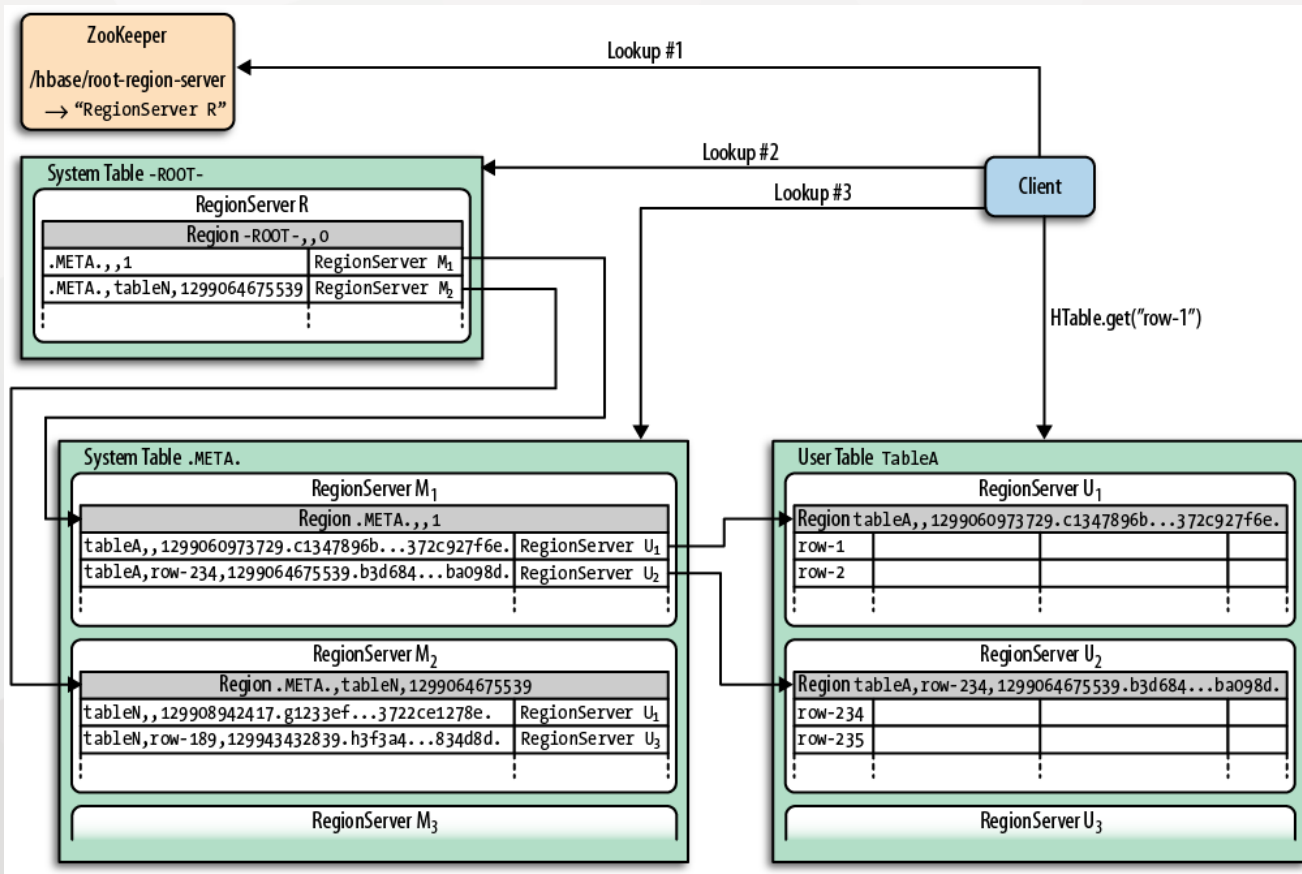




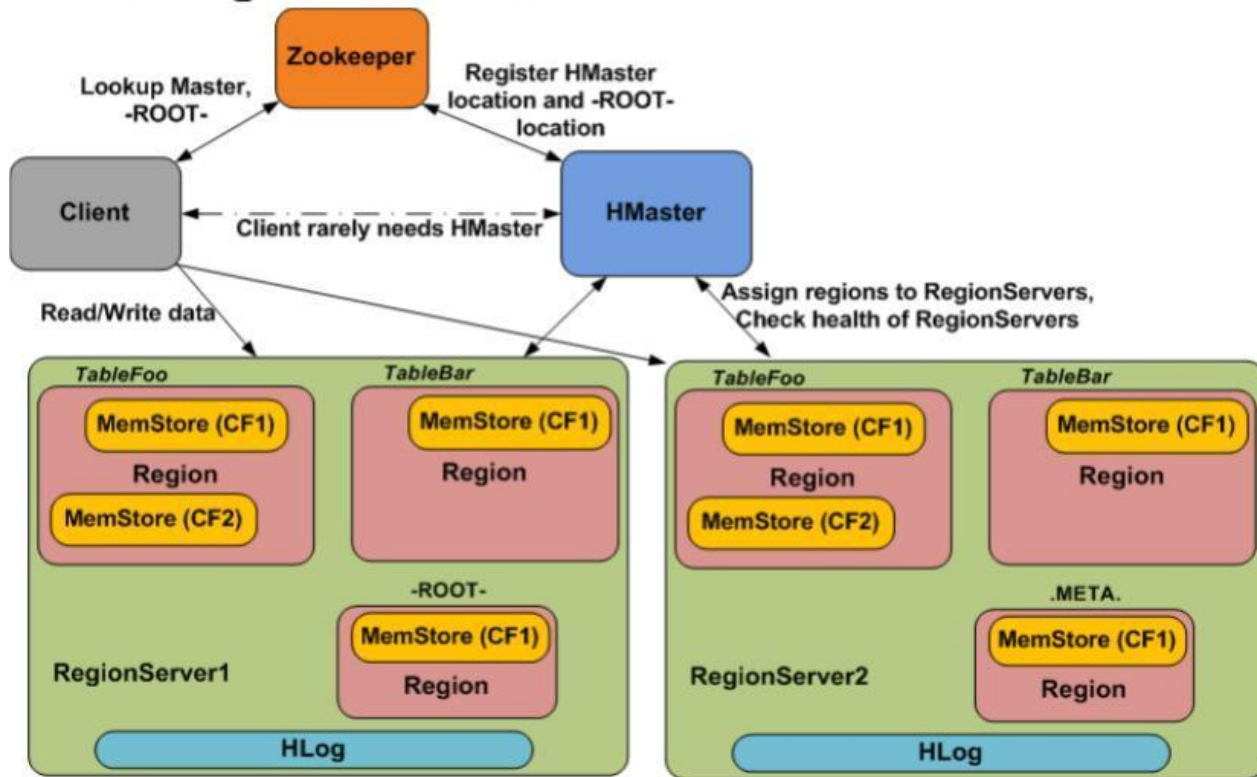
Review

10





HBase high-level architecture



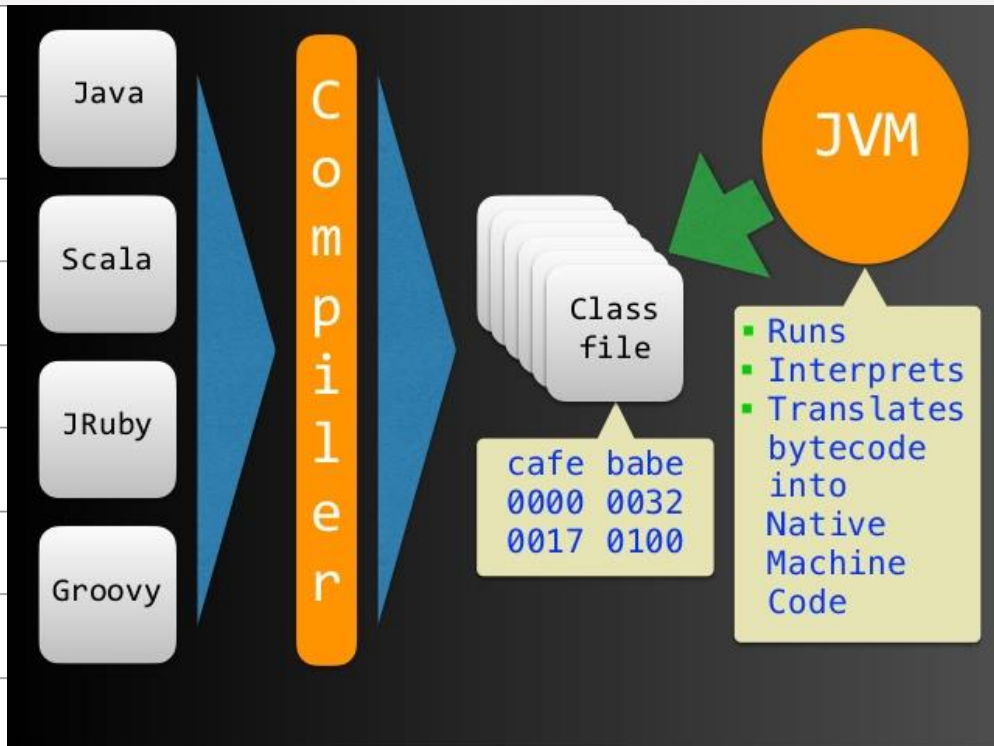
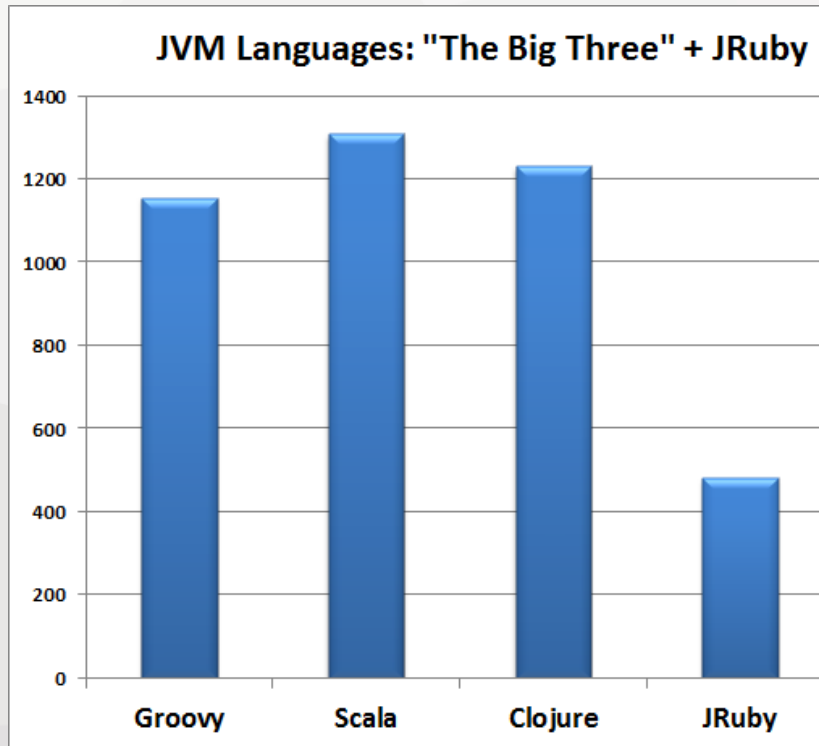
PART1



Preliminary Topics
事前準備



Big Three





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;  
}
```



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")
```

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
3       xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">
4   <modelVersion>4.0.0</modelVersion>
5   <groupId>info.solidsoft.rnd</groupId>
6   <artifactId>spock-10-groovy-24-gradle-maven</artifactId>
7   <version>0.0.1-SNAPSHOT</version>
8   <properties>
9     <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
10    <surefire.version>2.18.1</surefire.version>
11  </properties>
12  <build>
13    <plugins>
14      <plugin>
15        <groupId>org.codehaus.gmavenplus</groupId>
16        <artifactId>gmavenplus-plugin</artifactId>
17        <version>1.4</version>
18        <executions>
19          <execution>
20            <goals>
21              <goal>compile</goal>
22              <goal>testCompile</goal>
23            </goals>
24          </execution>
25        </executions>
26      </plugin>
27      <plugin>
28        <artifactId>maven-surefire-plugin</artifactId>
29        <version>${surefire.version}</version>
30        <configuration>
31          <includes>
32            <include>/**/*.java</include> <!-- Yes, .java extension -->
33            <include>/**/*.Test.java</include> <!-- Just in case having "normal" JUnit tests -->
34          </includes>
35        </configuration>
36      </plugin>
37    </plugins>
38  </build>
39  <dependencies>
40    <dependency>
41      <groupId>org.codehaus.groovy</groupId>
42      <artifactId>groovy-all</artifactId>
43      <version>2.4.1</version>
44    </dependency>
45    <dependency>
46      <groupId>org.spockframework</groupId>
47      <artifactId>spock-core</artifactId>
48      <version>1.0-groovy-2.4</version>
49      <scope>test</scope>
50    </dependency>
51  </dependencies>
52 </project>

```

pom.xml

```

1 apply plugin: 'groovy'
2
3 group = "info.solidsoft.rnd"
4 version = "0.0.1-SNAPSHOT"
5
6 repositories {
7   mavenCentral()
8 }
9
10 dependencies {
11   compile 'org.codehaus.groovy:groovy-all:2.4.1'
12   testCompile 'org.spockframework:spock-core:1.0-groovy-2.4'
13 }
14
15

```

build.gradle

```

1 rootProject.name = 'spock-10-groovy-24-gradle-maven'
2

```

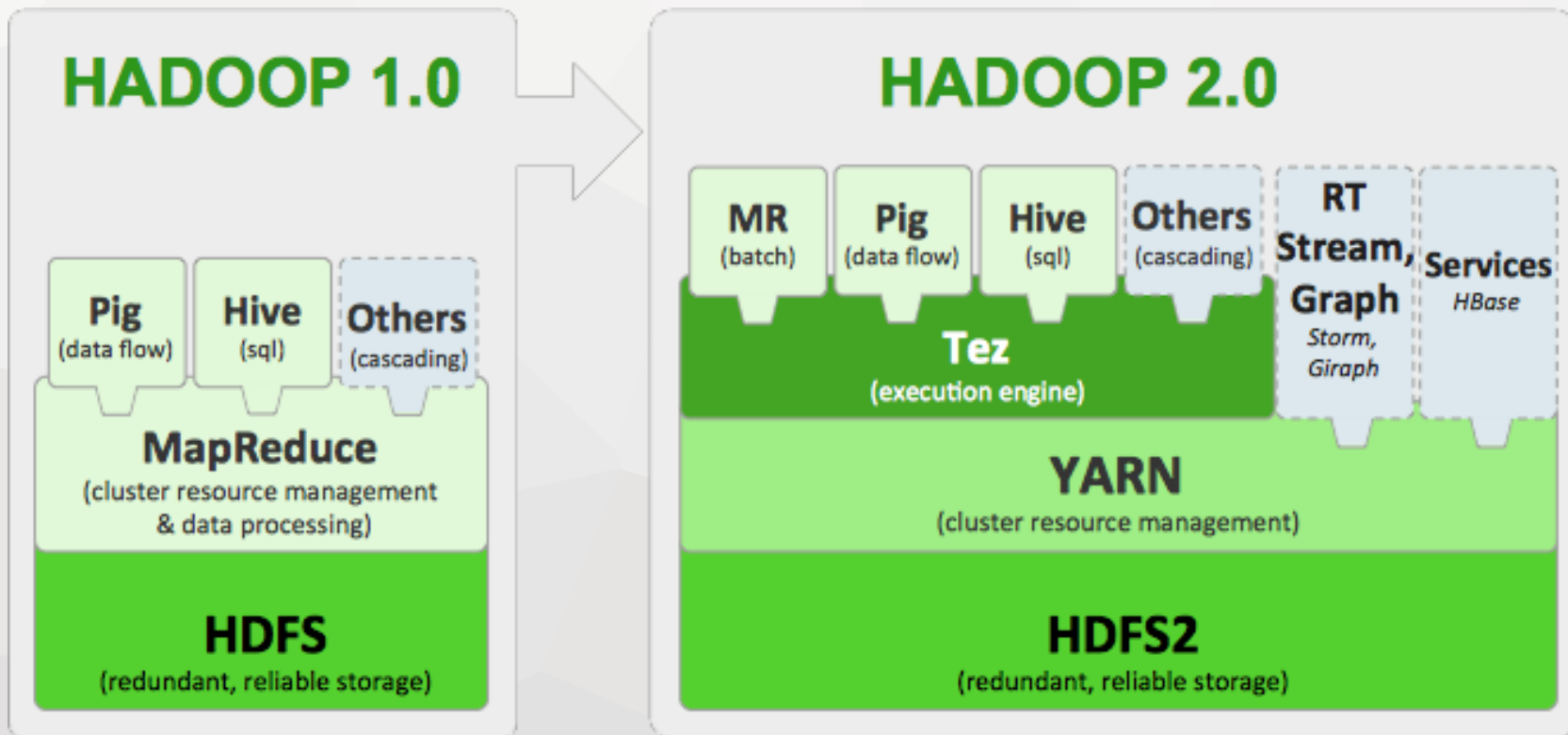
settings.xml



maven



gradle



Applications Run Natively IN Hadoop

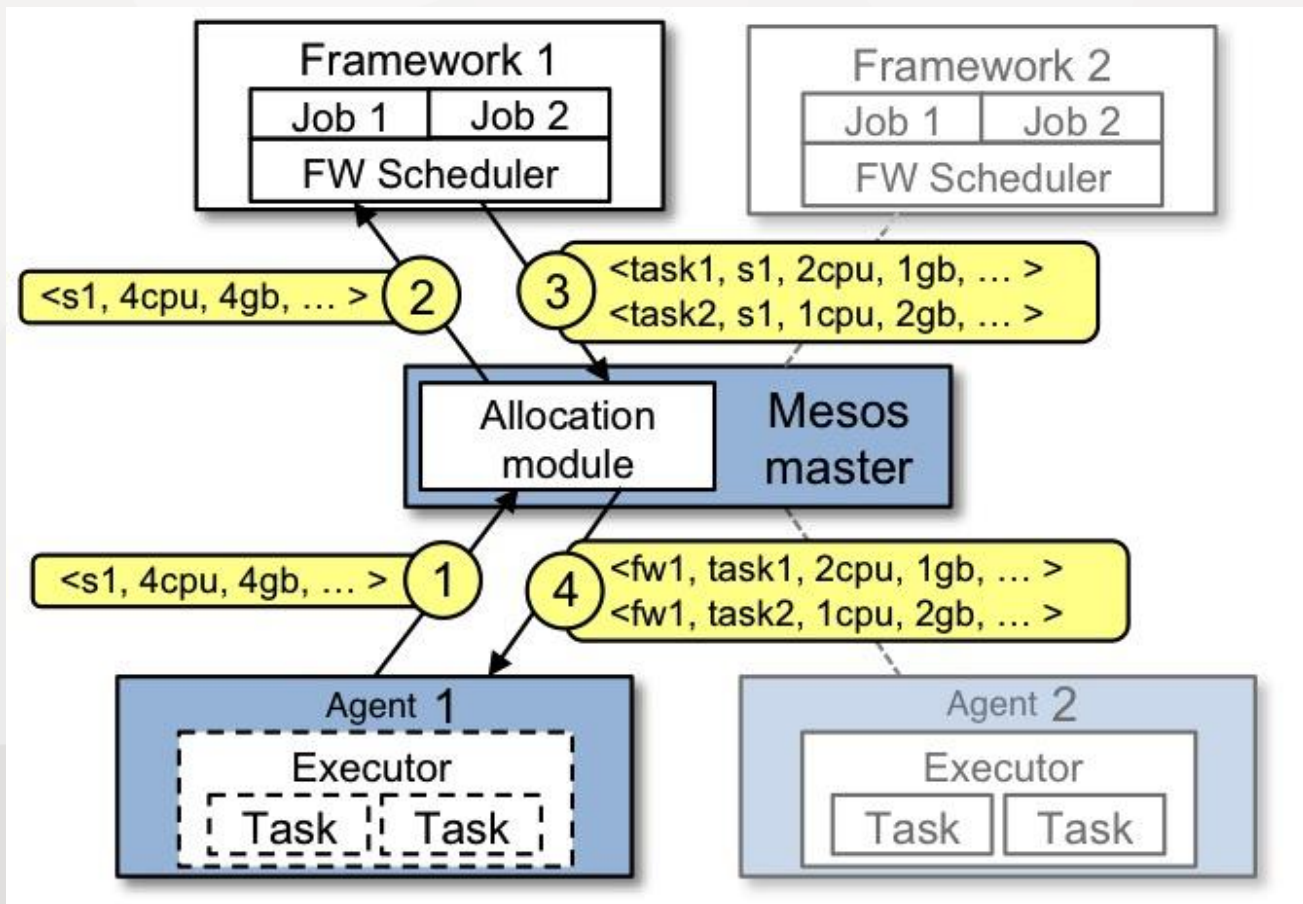
Pig Script	Hive SQL	HBase NoSQL	Accumulo NoSQL	Storm Stream	Solr Search	Spark In-Memory	Cascading Java	Others ISV Engines
----------------------	--------------------	-----------------------	--------------------------	------------------------	-----------------------	---------------------------	--------------------------	---------------------------------

YARN: Data Operating System

HDFS

(Hadoop Distributed File System)



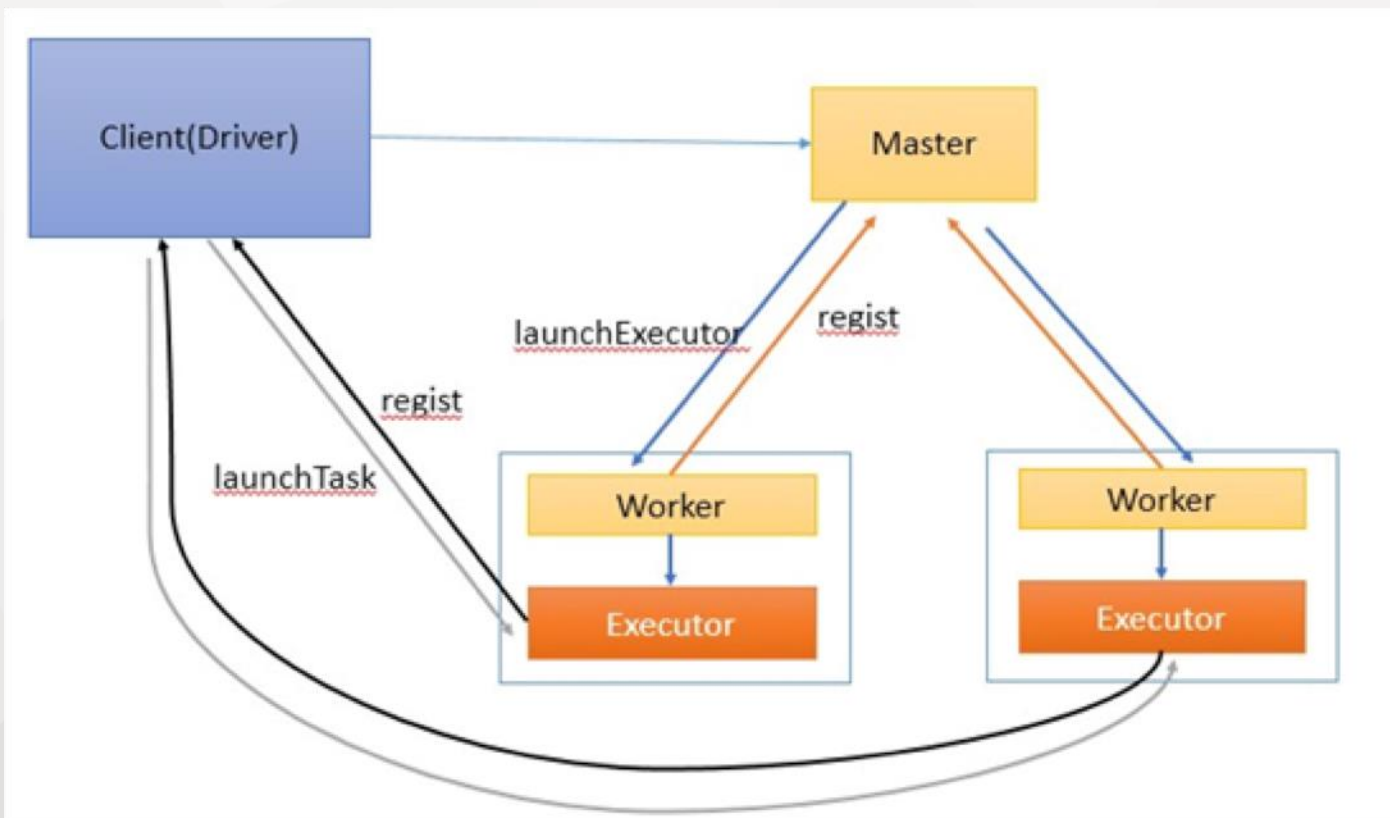




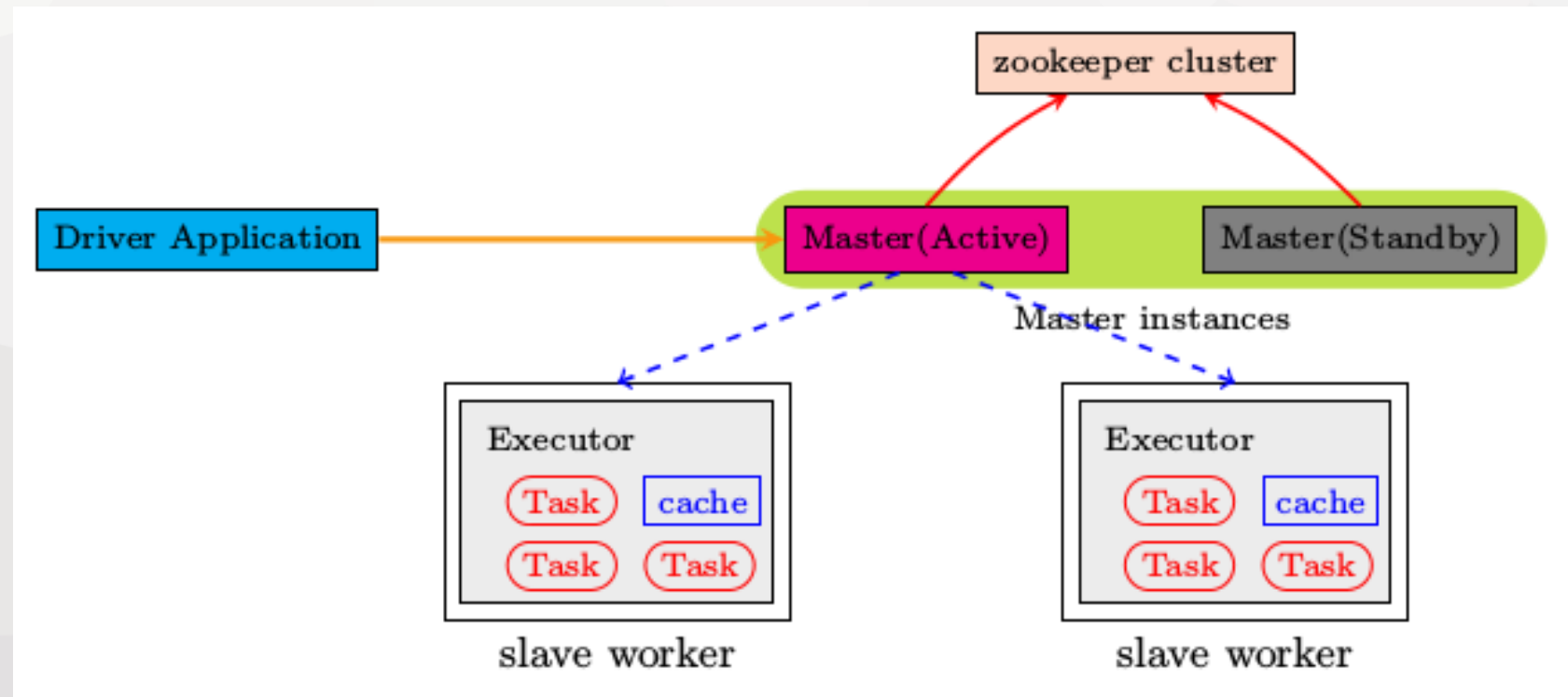
PART2

Spark Environment

Spark Standalone



Spark Standalone HA



[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](#)

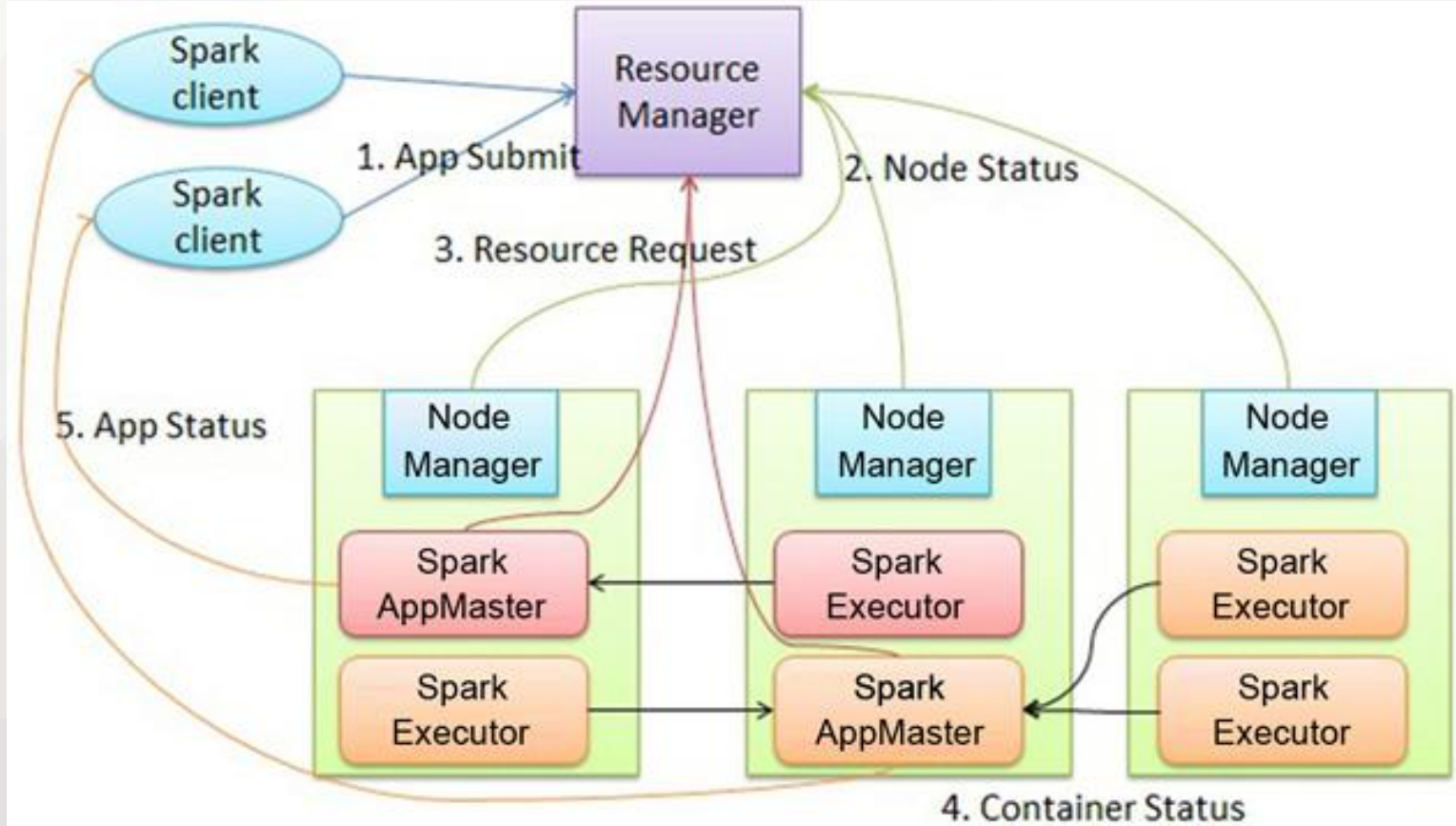
Built-in Libraries:

[SQL and DataFrames](#)
[Spark Streaming](#)
[MLlib \(machine learning\)](#)
[GraphX \(graph\)](#)

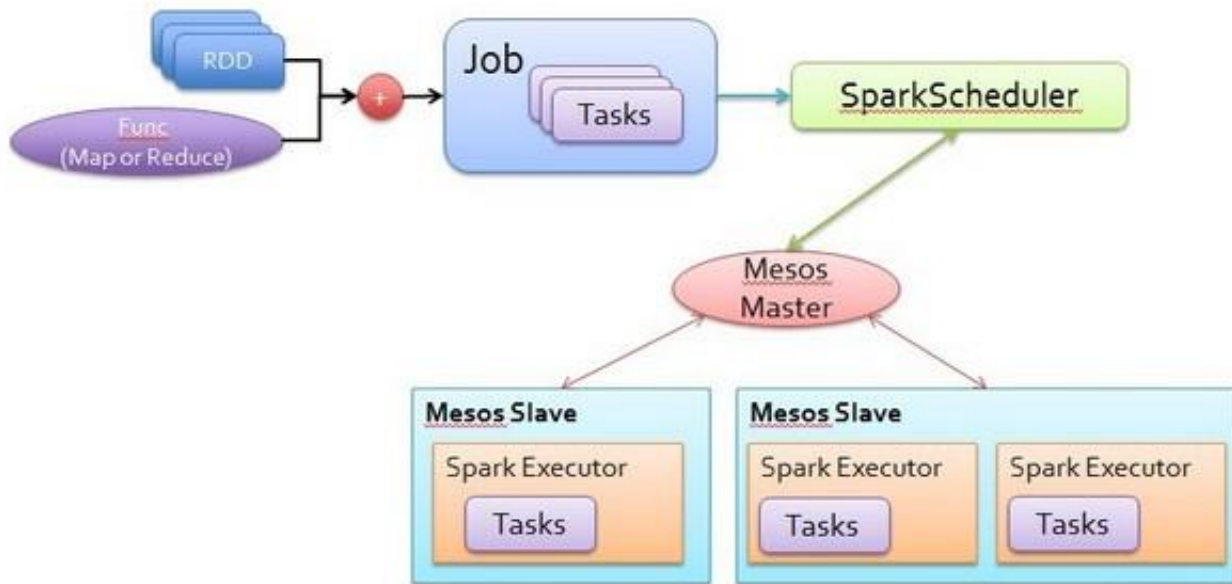
[Third-Party Projects](#)

Spark On Yarn

27



Spark On Mesos





PART3



Spark Architecture

Resource Management

Standalone

YARN

Mesos

Spark Ecosystems

Spark SQL

Spark Streaming

BlinkDB

Spark Machine Learning

GraphX

Tachyon

Spark Core

BACK TO BASICS

Spark DataFrame API



Java



Scala

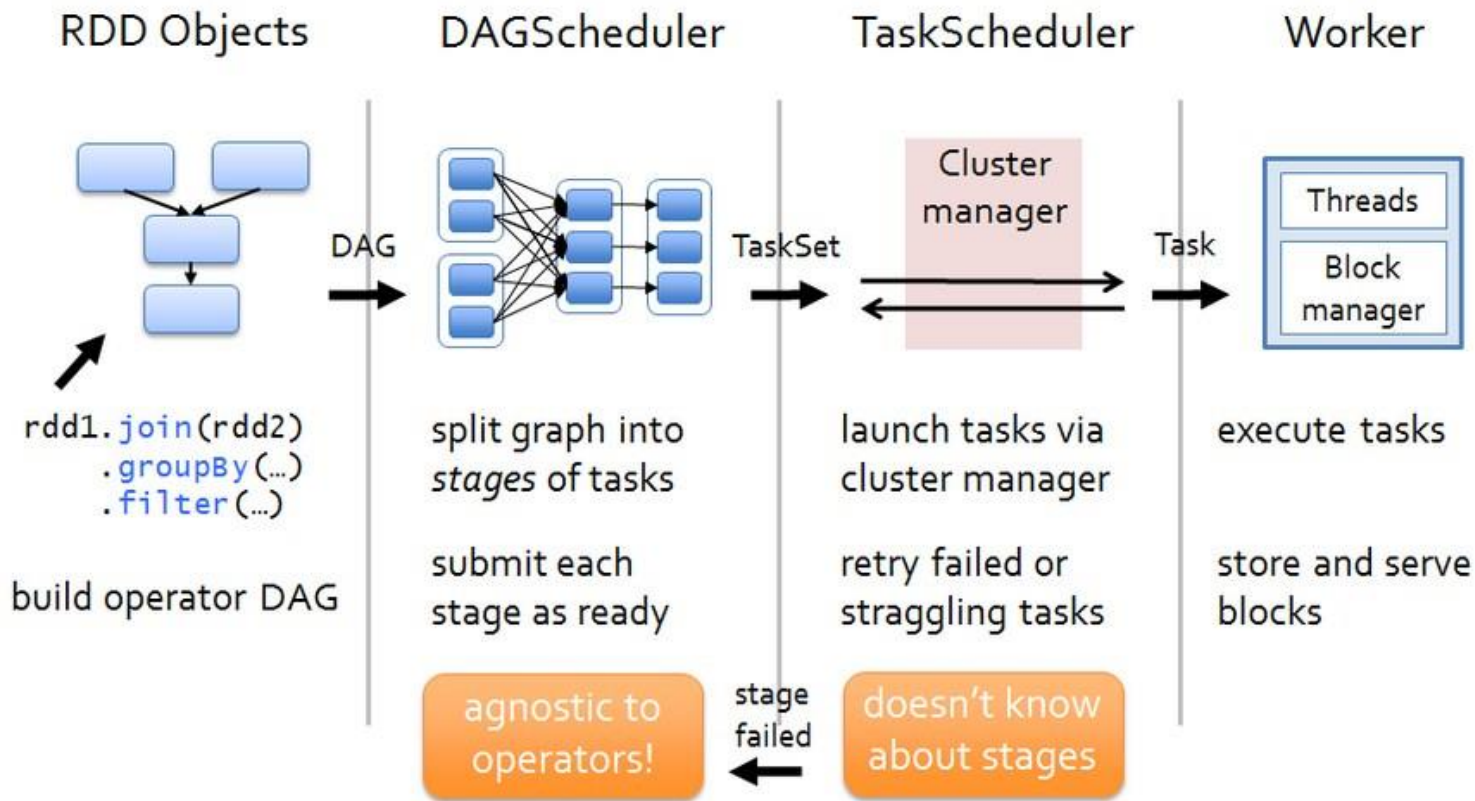


Python



R

Spark Core



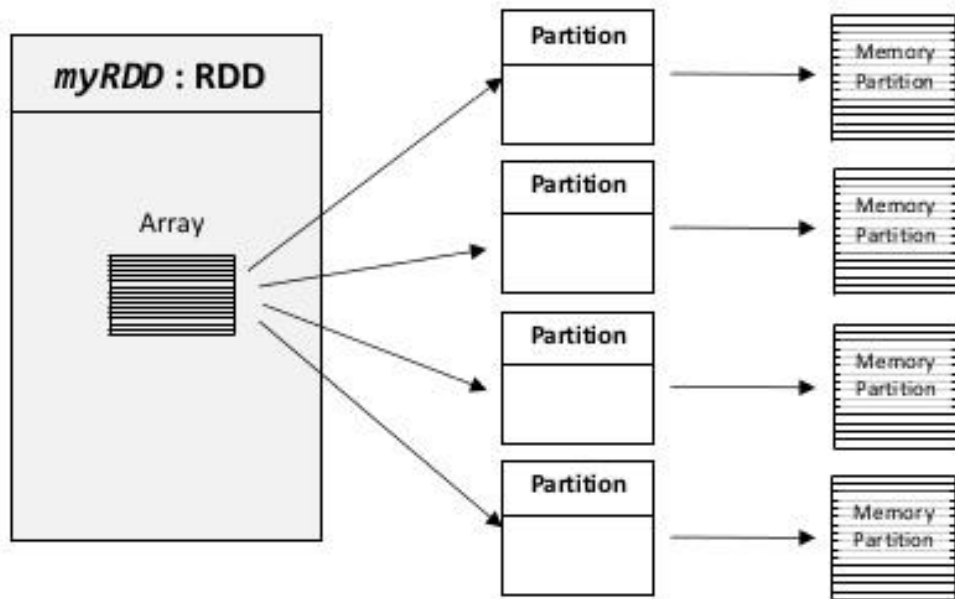


PART4



Spark RDD

What is an RDD?

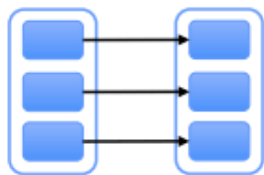


Some RDD Characteristics

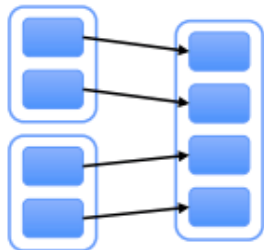
- Hold references to Partition objects
- Each Partition object references a subset of your data
- Partitions are assigned to nodes on your cluster
- Each partition/split will be in RAM (by default)

Dependency Types

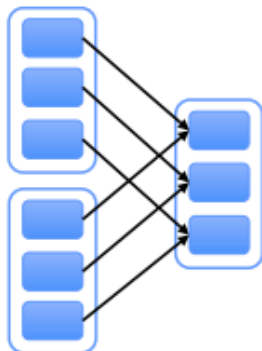
“Narrow” (pipeline-able)



map, filter

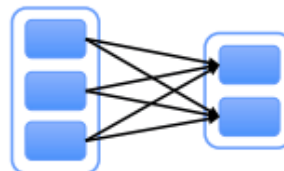


union

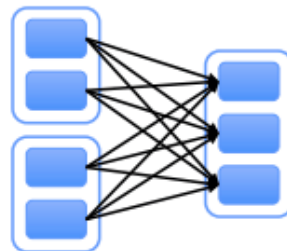


join with inputs
co-partitioned

“Wide” (shuffle)

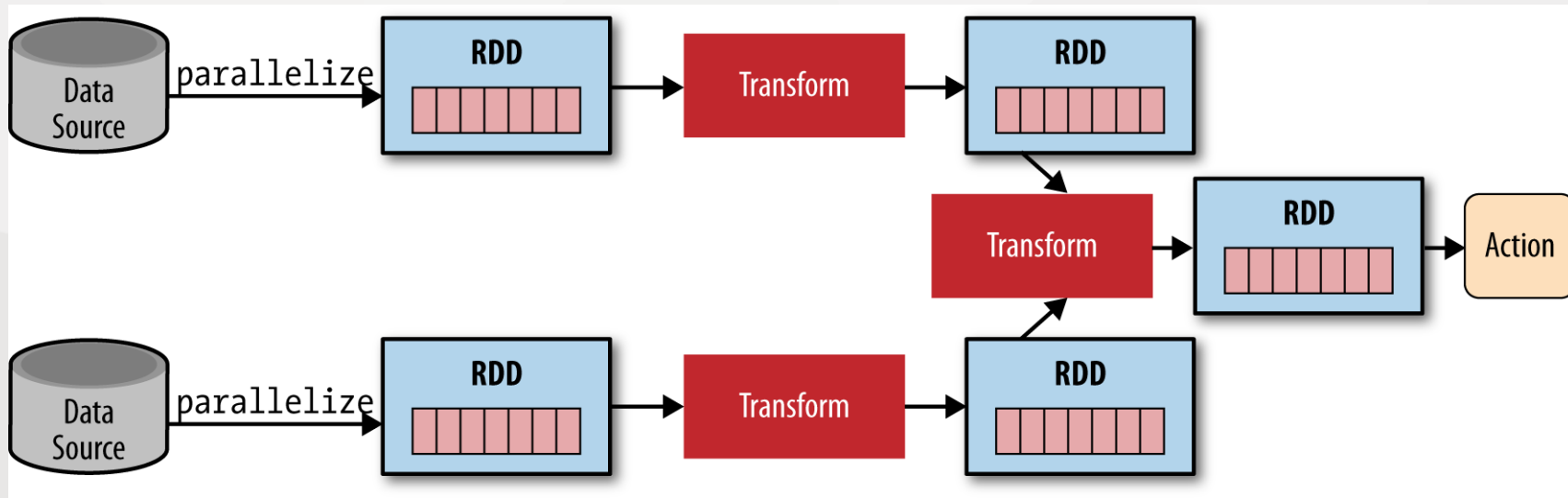


groupByKey on
non-partitioned data



join with inputs not
co-partitioned

Dependency Type



A stylized illustration of a computer monitor. The monitor has a dark blue frame and a white screen. On the screen, the words "The End" are written in a bold, dark blue, sans-serif font. The monitor is supported by a simple, dark blue stand. The background consists of a light gray, low-poly geometric pattern.

The End