Spark计算框架详解

讲师：Yasaka

中国移动项目展示

- kafka(移动的程控交换机获取数据)
- storm（实时计算）
- hbase+zookeeper（数据存储）
- servlet+ehcart(数据展示)
- get 'cell_monitor_table','29448-000001_2015-07-24'

天猫双十一大屏幕


双十一成交额
2014年11月11日10:27:34
¥28,922,907,916
GMV generated on Singles Day
$4,729,829,586
无线成交占比
Mobile GMV as a % of Total GMV
46.9%
成交额（亿）
289
231
173
0
快报

天猫双十一大屏幕

2014 阿里巴巴双十一之
交易额里程碑

1 亿
55 秒
2013 年

1 分 13 秒
2014 年

50
38 分 5 秒
2013 年

100 亿
5 小时 49 分
2013 年

38 分 28 秒
2014 年

数据来源：阿里巴

截止03:00
天猫1111购物狂欢节
省份购买排名

广东省	912,088,366
浙江省	866,004,840
江苏省	785,598,039
上海	680,749,295
北京	547,641,712
四川省	504,118,389
山东省	421,682,351
湖北省	405,981,034
湖南省	369,851,972
福建省	366,121,688

讲课内容

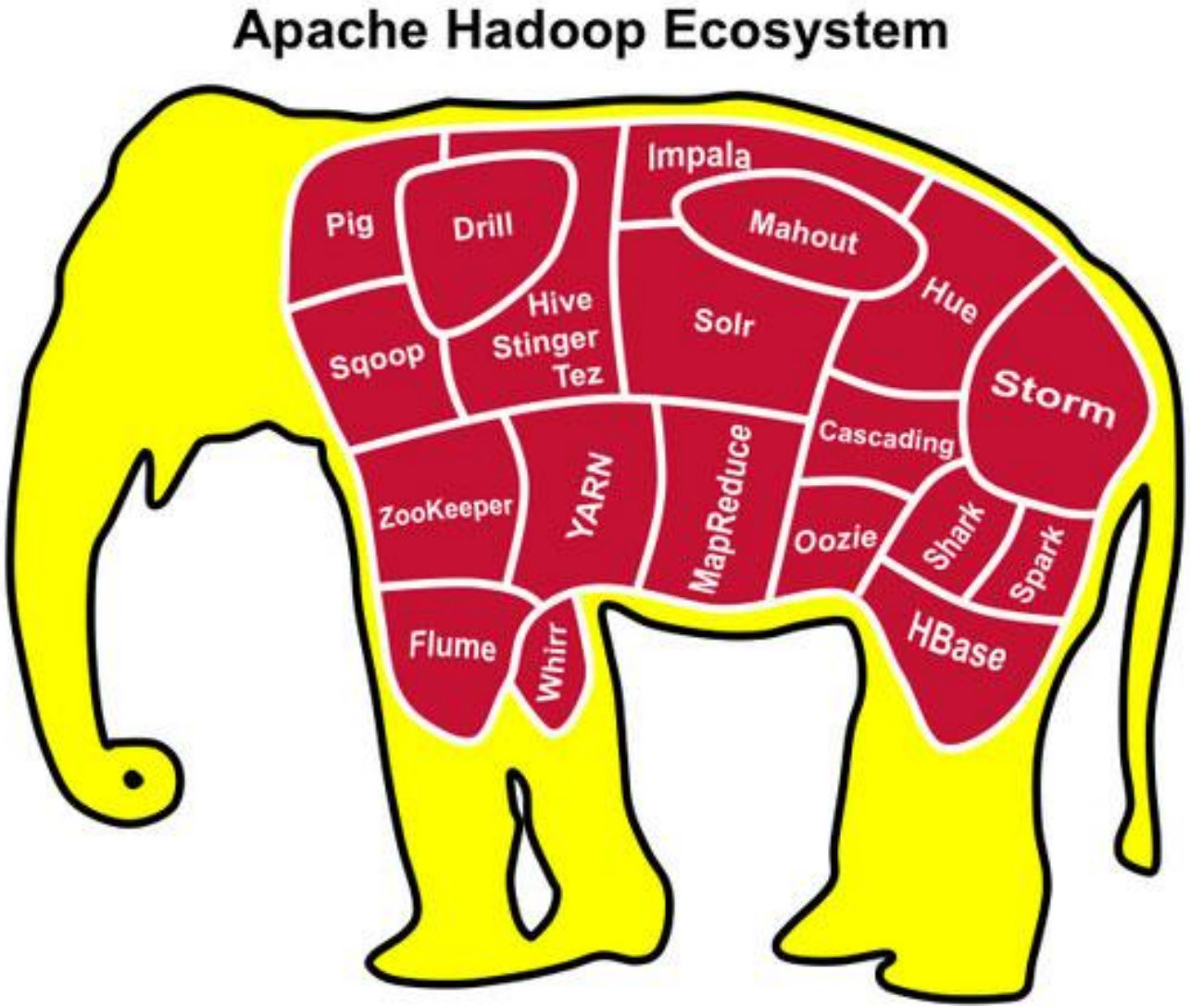
- 什么是大数据计算框架？
- 公司里为什么需要大数据计算框架？
- 哪些技术属于大数据计算框架？
- 什么是Spark？解决了哪些问题？
- Spark架构是什么？特点是什么？
- 如何搭建Spark？如何运行Spark？
- 剖析API
- 运行样例程序
- 音乐推荐系统展示

什么是大数据计算框架？

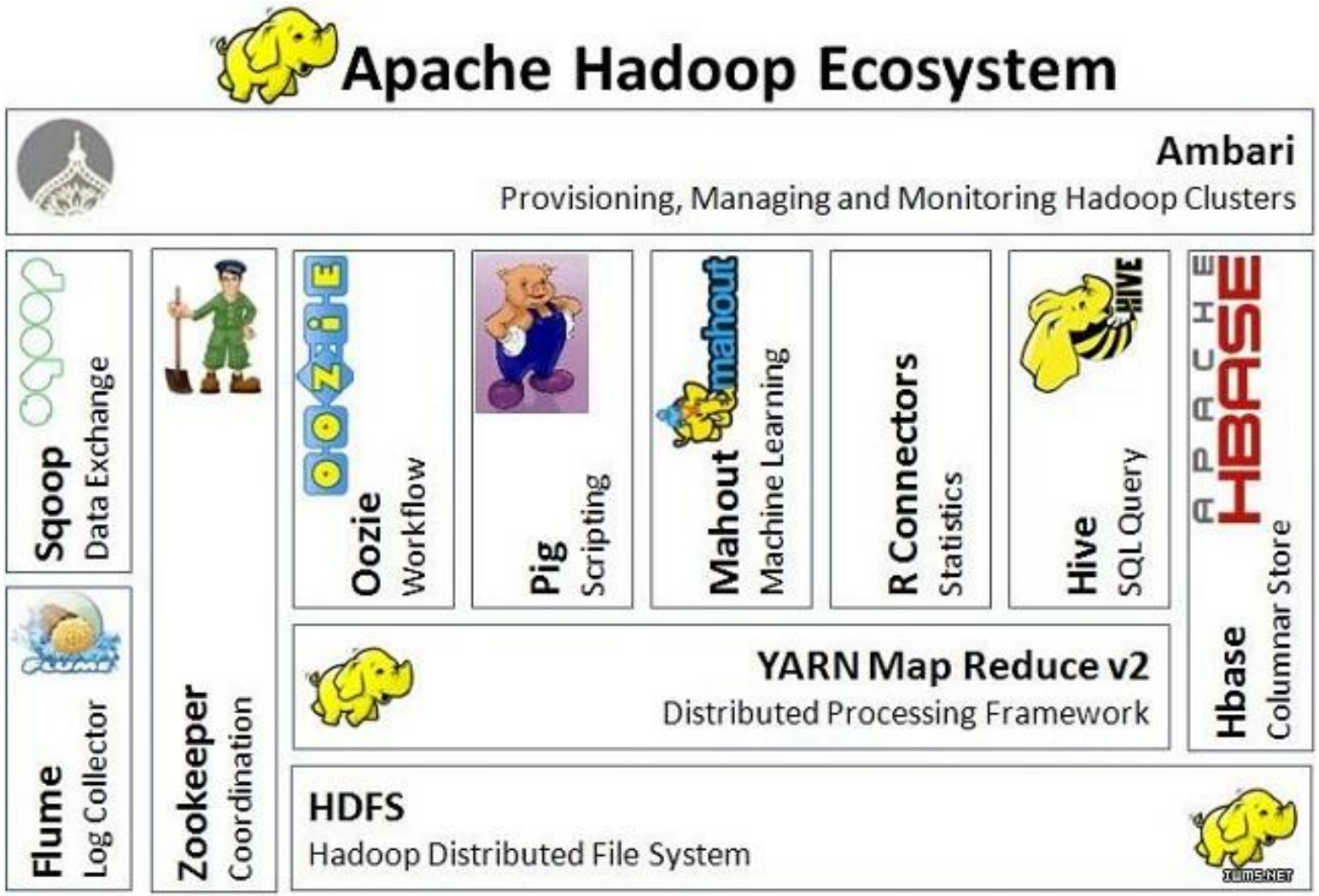
Hadoop

Storm

什么是大数据计算框架？


Apache Hadoop Ecosystem
Pig
Drill
Impala
Mahout
Hive
Stinger
Tez
Solr
Hue
Sqoop
Storm
Cascading
ZooKeeper
YARN
MapReduce
Oozie
Shark
Spark
Flume
Whirr
HBase

什么是大数据计算框架？


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
APACHE HBASE
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

公司里为什么需要大数据计算框架？

- 高并发
- 大数据
- 云计算
- 虚拟化

天猫双十一大屏幕

腾讯QQ在线人数统计

百度迁徙项目

京东推荐系统

最佳组合


鸭鸭(YAYA)男士加厚外套保暖冬装羽绒服男 A-9
¥299.00


鸭鸭(YAYA)冬装男士轻羽绒保暖男装羽绒服男
¥199.00


鸭鸭2015秋冬男短款时尚羽绒马甲背心D-5502
¥249.00

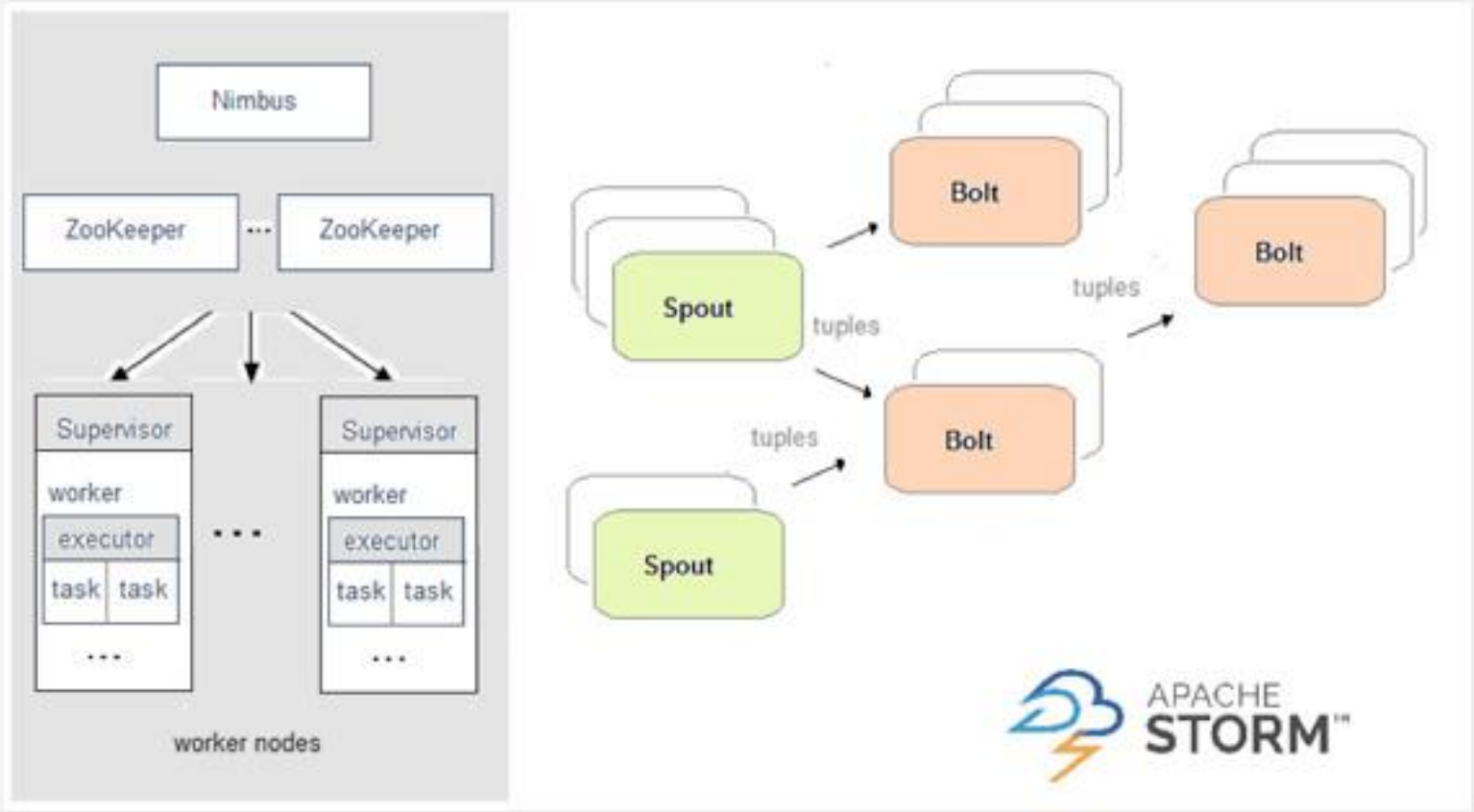

鸭鸭（yaya）男士短款加厚秋冬韩版羽绒服 A-5
¥379.00


鸭鸭(YAYA)男款加厚带帽毛领 商务休闲羽绒服
¥559.00


鸭鸭（YAYA）男款中老年款可脱卸帽90%鸭绒
¥309.00

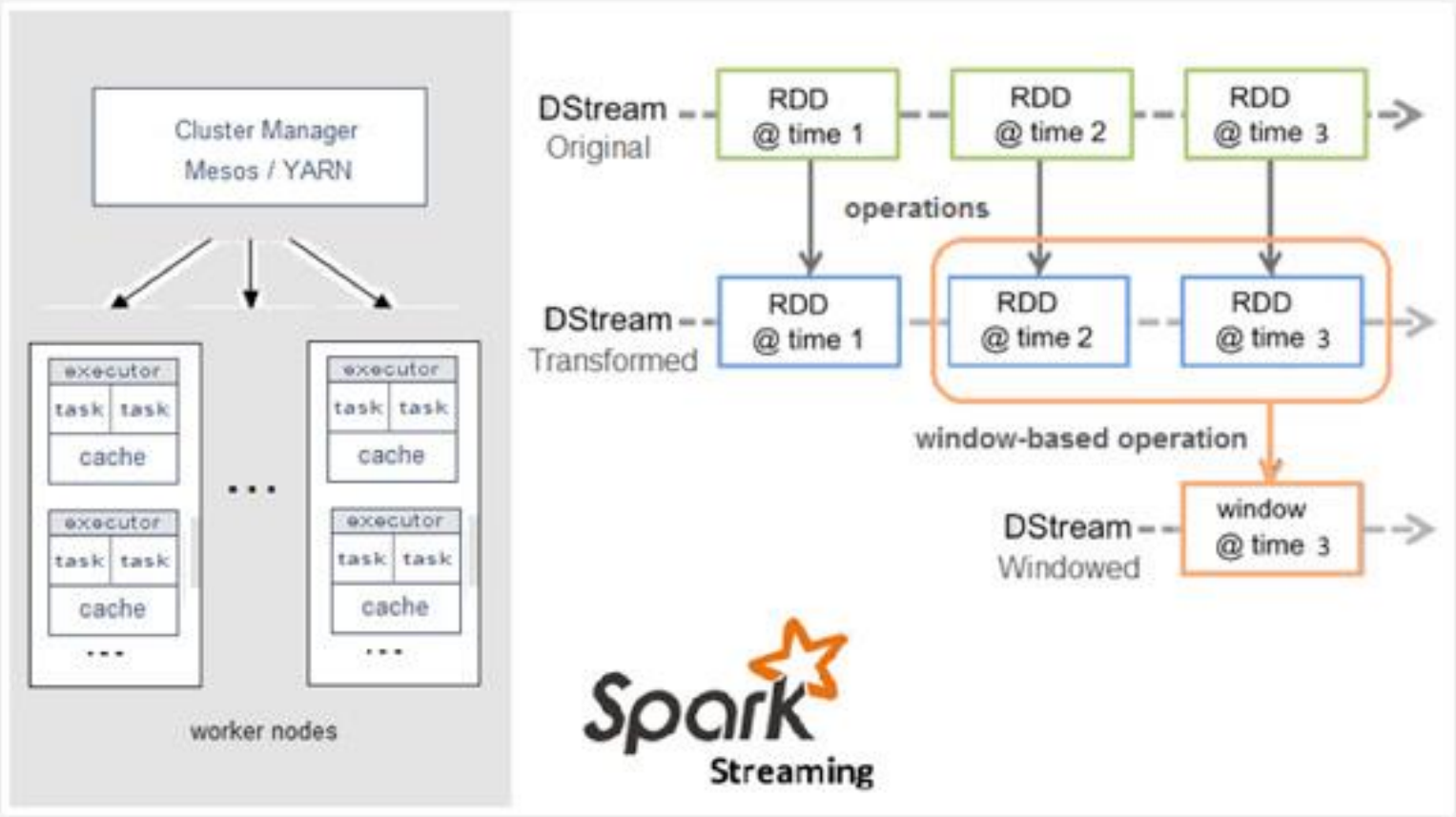
哪些技术属于大数据计算框架？

- Storm



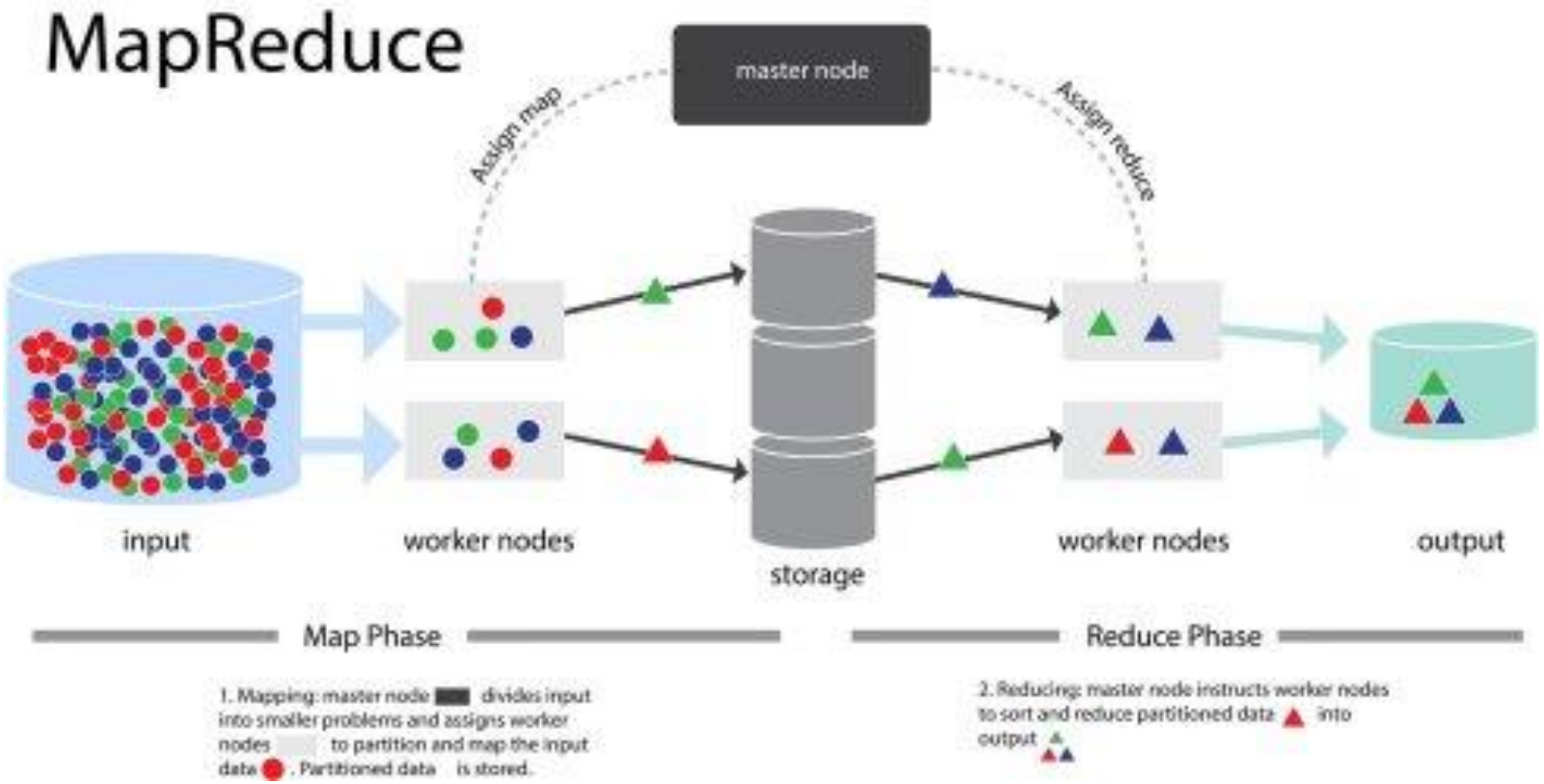
哪些技术属于大数据计算框架？

- Spark



哪些技术属于大数据计算框架？

- Hadoop→MapReduce+GFS+BigTable



什么是Spark？

- Spark，是一种通用的大数据计算框架，正如传统大数据技术Hadoop的MapReduce、Hive引擎，以及Storm流式实时计算引擎等
- 2009年，Spark诞生于伯克利大学的AMPLab实验室；2010年，伯克利大学正式开源了Spark项目；2013年，Spark成为了Apache基金会下的项目； 2014年，Spark以飞快的速度称为了Apache的顶级项目； 2015年~，Spark在国内IT行业变得愈发火爆
- 支持多语言

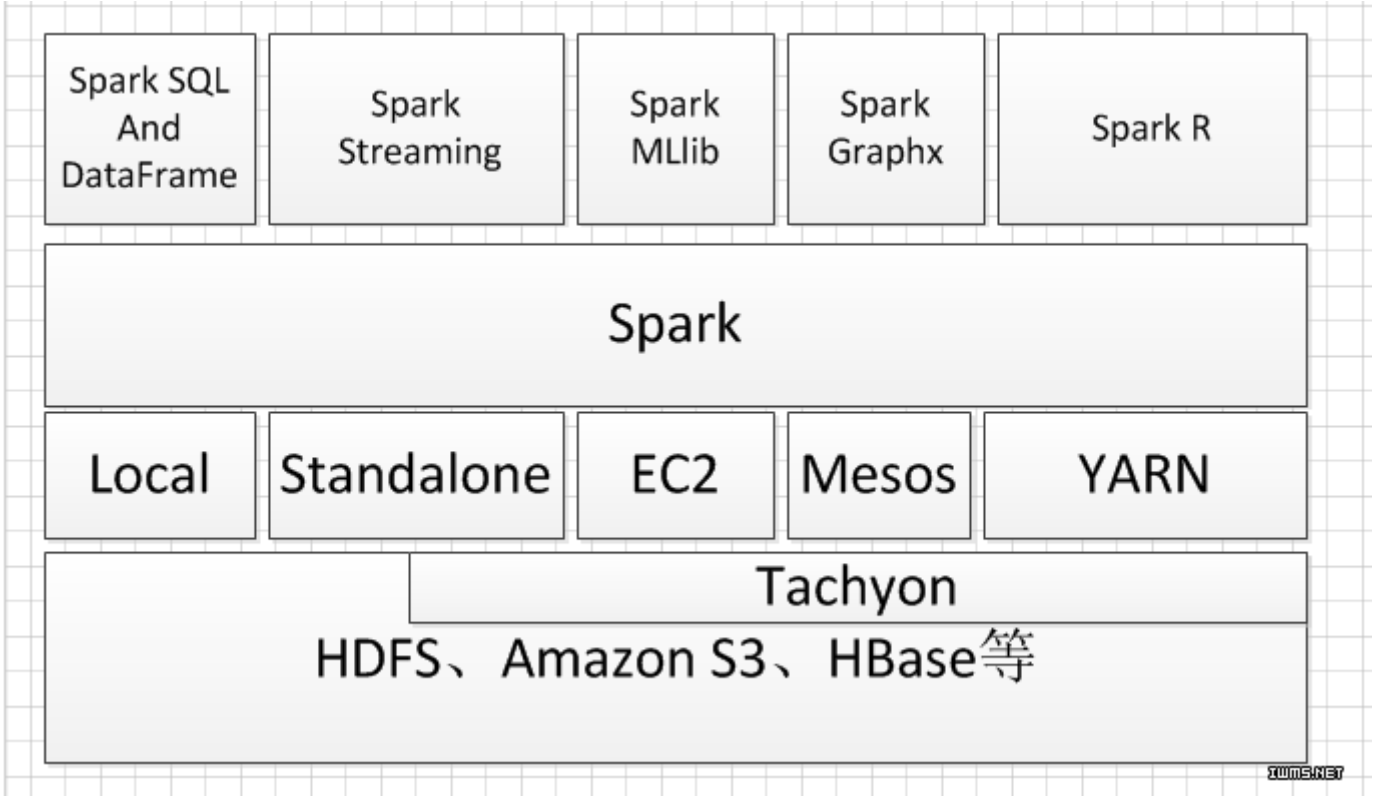
什么是Spark？

- Spark包含了大数据领域常见的各种计算框架：比如Spark Core用于离线计算，Spark SQL用于交互式查询，Spark Streaming用于实时流式计算，Spark MLlib用于机器学习，Spark GraphX用于图计算。
- Spark主要用于大数据的计算，而Hadoop以后主要用于大数据的存储（比如HDFS、Hive、HBase等），以及资源调度（Yarn）。
- Spark+Hadoop的组合，是未来大数据领域最热门的组合，也是最有前景的组合！

解决了哪些问题？

- 互联网广告实时流量统计
- 互联网数据质量实时监控
- 交通超速频发路段监控
- 交通基于GPS的实时路况分析
- 移动互联语音实时墙
- 运营商网络流量流向实时分析
- 中国移动小区基站预警
- 建设智慧城市
- 推荐系统构建

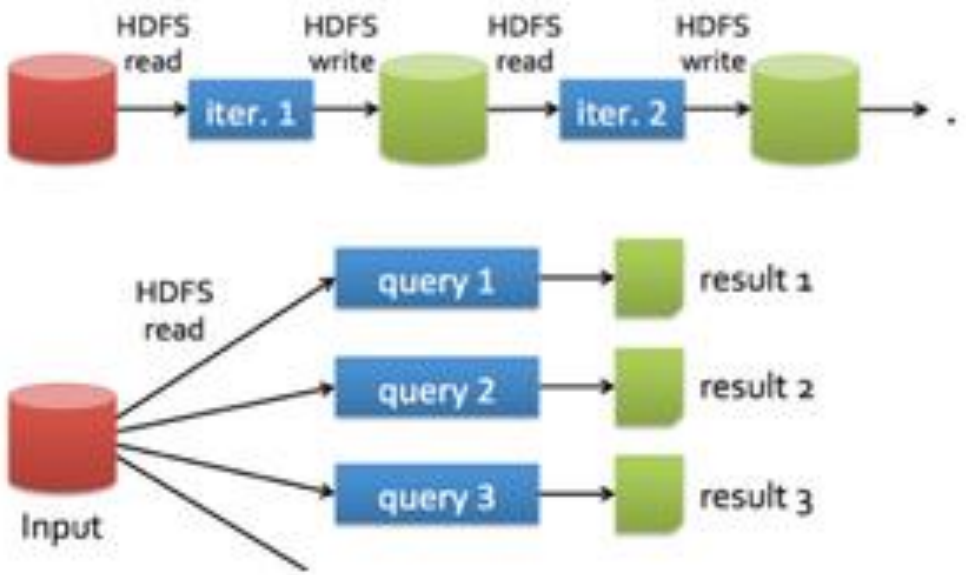
Spark架构是什么？


Spark SQL And DataFrame
Spark Streaming
Spark MLlib
Spark Graphx
Spark R
Spark
Local
Standalone
EC2
Mesos
YARN
Tachyon
HDFS、Amazon S3、HBase等

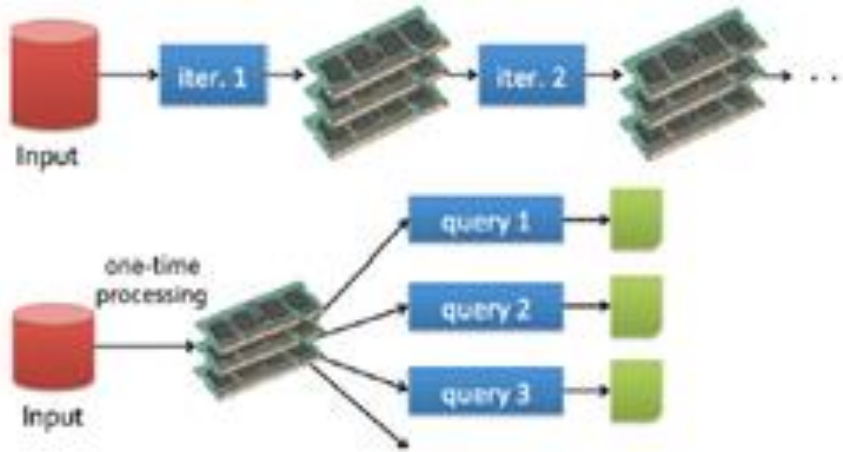
Spark特点是什么？

- 速度快
- 容易上手开发
- 超强的通用性
- 集成hadoop
- 很高的活跃度

Hadoop Data Sharing

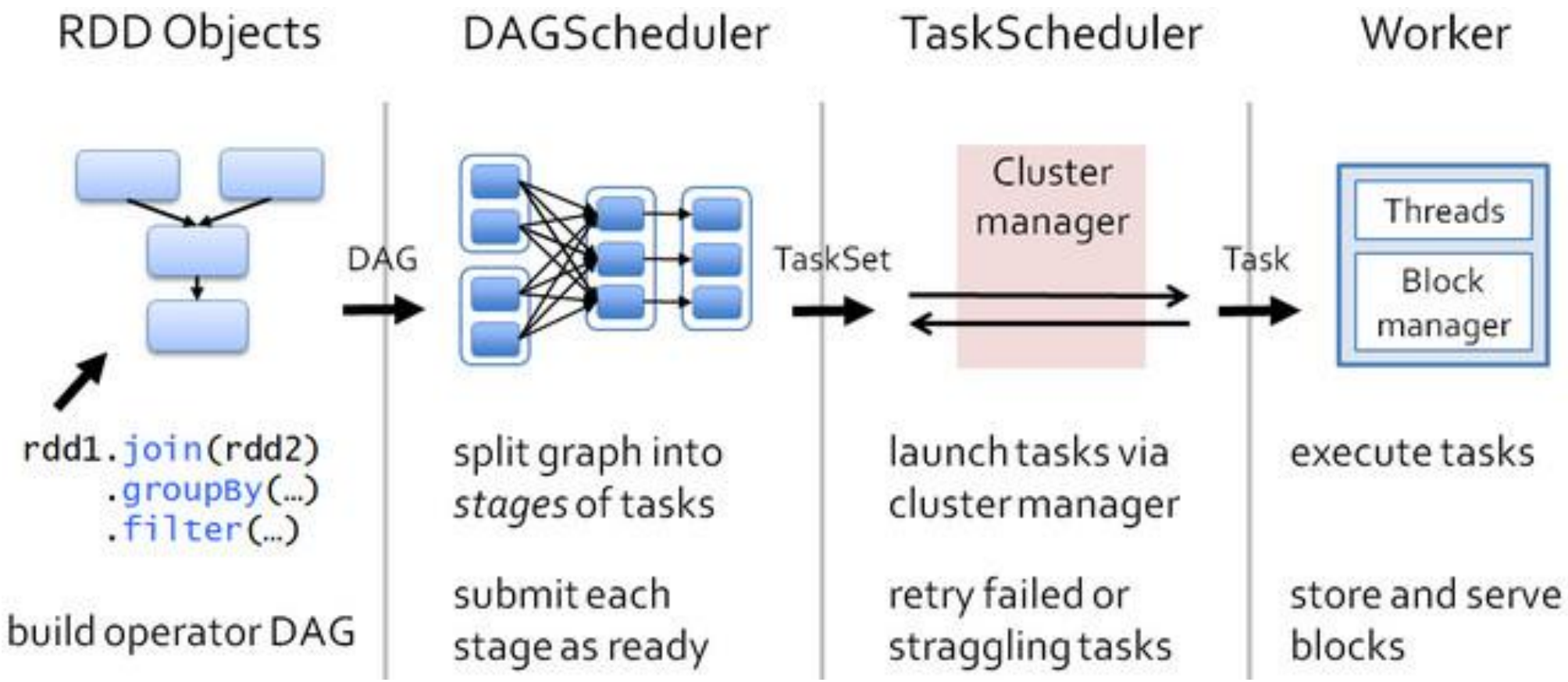


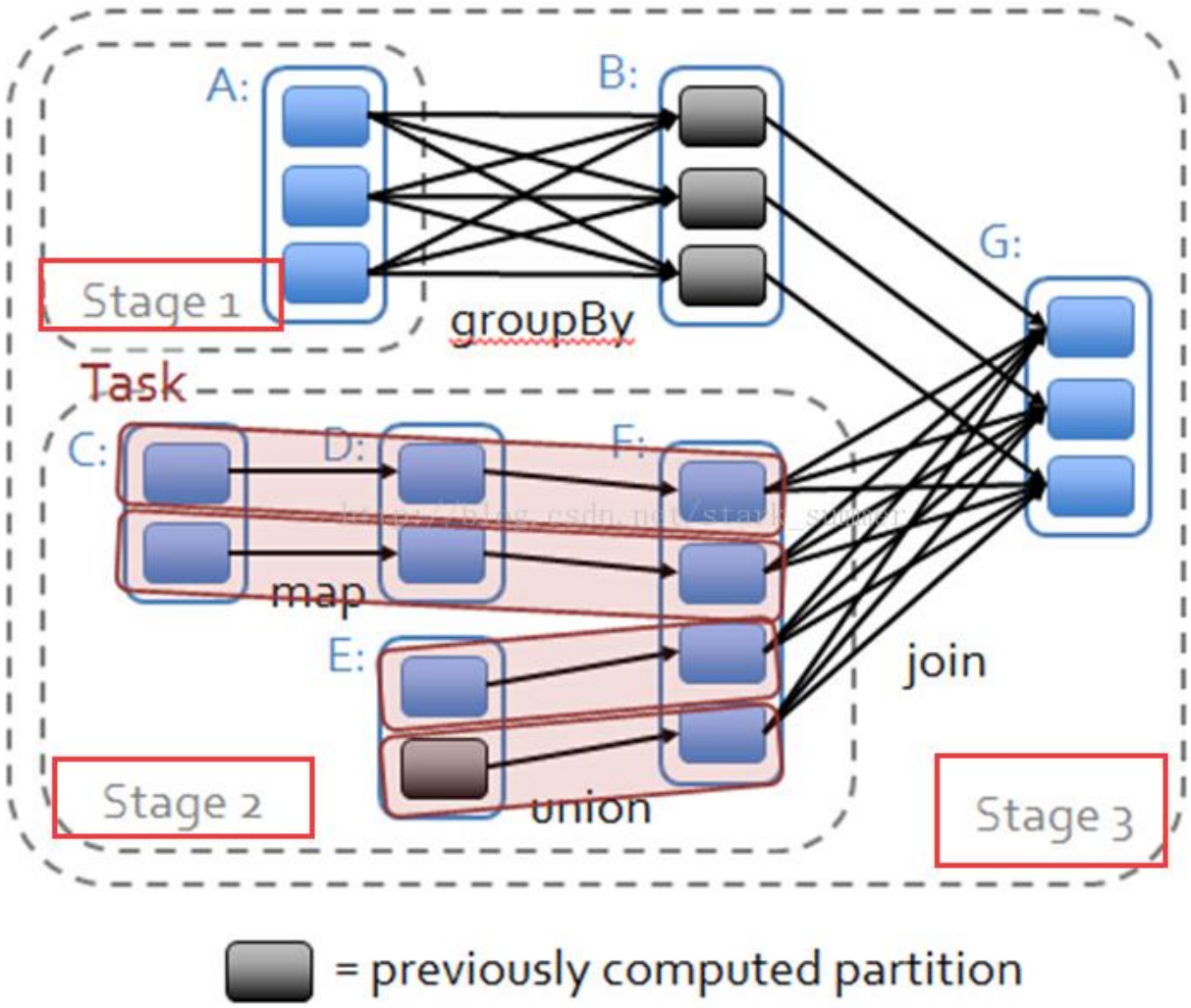
Spark Data Sharing



DAG & in-memory

任务调度图


RDD Objects
DAGScheduler
TaskScheduler
Worker
DAG
TaskSet
Task
Cluster manager
Threads
Block manager
rdd1.join(rdd2)
.groupBy(…)
.filter(…)
build operator DAG
split graph into stages of tasks
submit each stage as ready
launch tasks via cluster manager
retry failed or straggling tasks
execute tasks
store and serve blocks


A:
B:
G:
Stage 1
groupBy
Task
C:
D:
F:
map
E:
join
Stage 2
union
Stage 3
= previously computed partition

如何搭建Spark？

- 单机模式

- 独立集群→原生集群模式
- 验证：访问http://localhost:8080，运行Pi example

- YARN集群
  - Cluster
  - Client
- 验证：访问http://localhost:8088，运行Pi example

如何运行Spark？

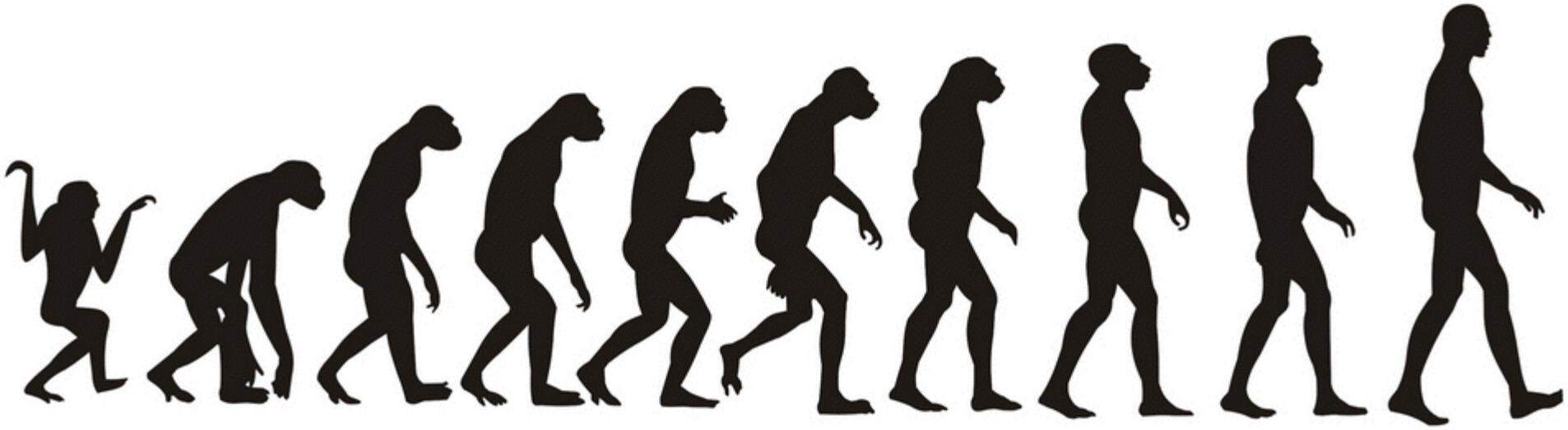
- local单机模式：
- ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master local[1] ./lib/spark-examples-1.3.1-hadoop2.4.0.jar 100
- standalone集群模式：
- ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master spark://spark001:7077 --executor-memory 1G --total-executor-cores 1 ./lib/spark-examples-1.3.1-hadoop2.4.0.jar 100
- Yarn集群模式：
- ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn-cluster --executor-memory 1G --num-executors 1 ./lib/spark-examples-1.3.1-hadoop2.4.0.jar 100

- SparkContext / RDD

Transformations	*map*(*f* : T ⇒ U)	:	RDD[T] ⇒ RDD[U]
	filter(*f* : T ⇒ Bool)	:	RDD[T] ⇒ RDD[T]
	flatMap(*f* : T ⇒ Seq[U])	:	RDD[T] ⇒ RDD[U]
	sample(*fraction* : Float)	:	RDD[T] ⇒ RDD[T] (Deterministic sampling)
	groupByKey()	:	RDD[(K, V)] ⇒ RDD[(K, Seq[V])]
	reduceByKey(*f* : (V, V) ⇒ V)	:	RDD[(K, V)] ⇒ RDD[(K, V)]
	union()	:	(RDD[T], RDD[T]) ⇒ RDD[T]
	join()	:	(RDD[(K, V)], RDD[(K, W)]) ⇒ RDD[(K, (V, W))]
	cogroup()	:	(RDD[(K, V)], RDD[(K, W)]) ⇒ RDD[(K, (Seq[V], Seq[W]))]
	crossProduct()	:	(RDD[T], RDD[U]) ⇒ RDD[(T, U)]
	mapValues(*f* : V ⇒ W)	:	RDD[(K, V)] ⇒ RDD[(K, W)] (Preserves partitioning)
	sort(*c* : Comparator[K])	:	RDD[(K, V)] ⇒ RDD[(K, V)]
	partitionBy(*p* : Partitioner[K])	:	RDD[(K, V)] ⇒ RDD[(K, V)]
Actions	*count*()	:	RDD[T] ⇒ Long
	collect()	:	RDD[T] ⇒ Seq[T]
	reduce(*f* : (T, T) ⇒ T)	:	RDD[T] ⇒ T
	lookup(*k* : K)	:	RDD[(K, V)] ⇒ Seq[V] (On hash/range partitioned RDDs)
	save(*path* : String)	:	Outputs RDD to a storage system, *e.g.*, HDFS

运行样例程序

- spark-shell 操作 wordcount



音乐推荐系统展示

```
Rating(2093760,2814,0.030518022934870874)
Rating(2093760,1001819,0.029811221324290038)
Rating(2093760,1300642,0.029380838727060564)
Rating(2093760,4605,0.02891767945591 4086)
Rating(2093760,1007614,0.028859793426105194)
```

```
Some((2814,50 Cent))
Some((4605,Snoop Dogg))
Some((1007614,Jay-Z))
Some((1001819,2Pac))
Some((1300642,The Game))
```

```
10771838 2093760 1180 1
10771839 2093760 1255340 3
10771840 2093760 378 1
10771841 2093760 813 2
10771842 2093760 942 7
```

结论

- 大家如果通过本堂课的讲解，能够较为全面地对Spark有一个感性得认识，就能意识到，Spark在大数据领域中，是未来的一个趋势和方向！

- Spark目前正在变得越来越火爆，招聘的企业正在越来越多，而且目前国内spark人才可以说是稀缺！！！在目前，以及未来，完全供不应求！因此这种趋势，以及这种现状，就决定了，对于我们个人来说，目前进行spark的学习以及研究，完全是未来一个获取快速升值的机会！！！

我们是？

- 使用最新版本
- 从零起步
- 涵盖Spark所有功能
- 一线互联网项目实战
- 结合源码对Spark内核进行深度剖析
- 全程配图详解
- 讲解Spark性能调优

你属于哪类？

- Java / J2EE开发工程师
- Hadoop开发工程师
- Spark入门级别的，或者只有一定基础的
- 在校或者刚毕业的学生

- Volume
- Velocity
- Variety
- Value