小象科技
ChinaHadoop.cn

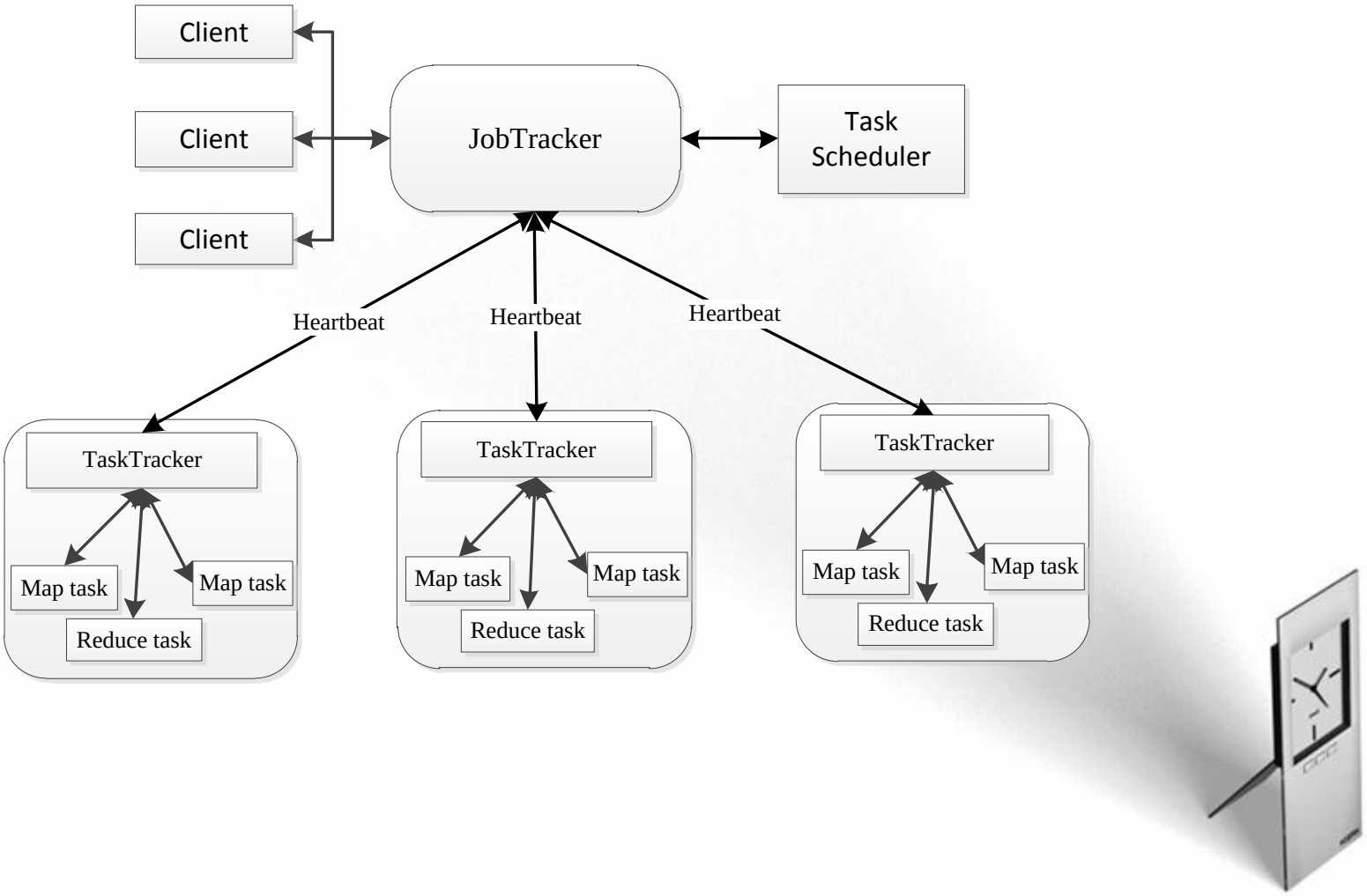
YARN应用场景、原理与资源调度

讲师：董西成

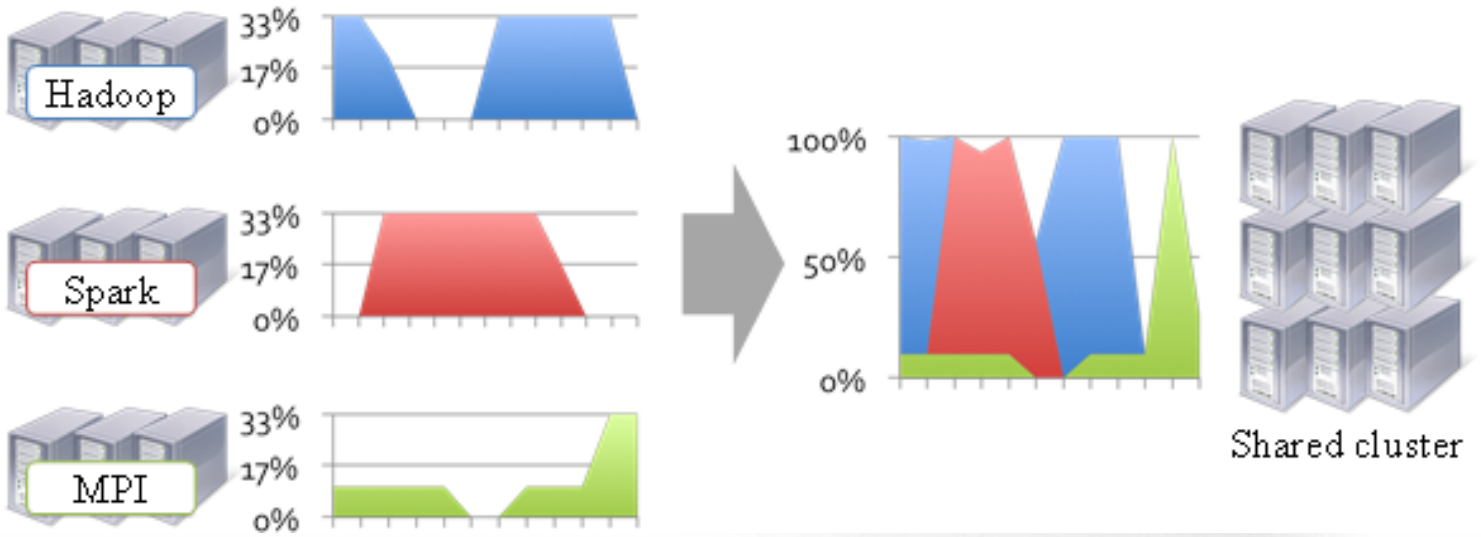
目录

小象科技
ChinaHadoop.cn

YARN产生背景—MapReduce 1.0固有问题



YARN产生背景—资源利用率



YARN产生背景—运维成本与数据共享

小象科技 ChinaHadoop.cn

- **运维成本**

  如果采用“一个框架一个集群”的模式，则可能需要多个管理员管理这些集群，进而增加运维成本，而**共享模式**通常需要少数管理员即可完成多个框架的统一管理。

- **数据共享**

  随着数据量的暴增，跨集群间的数据移动不仅需花费更长的时间，且硬件成本也会大大增加，而**共享集群模式**可让多种框架共享数据和硬件资源，将大大减小数据移动带来的成本。

YARN产生背景—总结

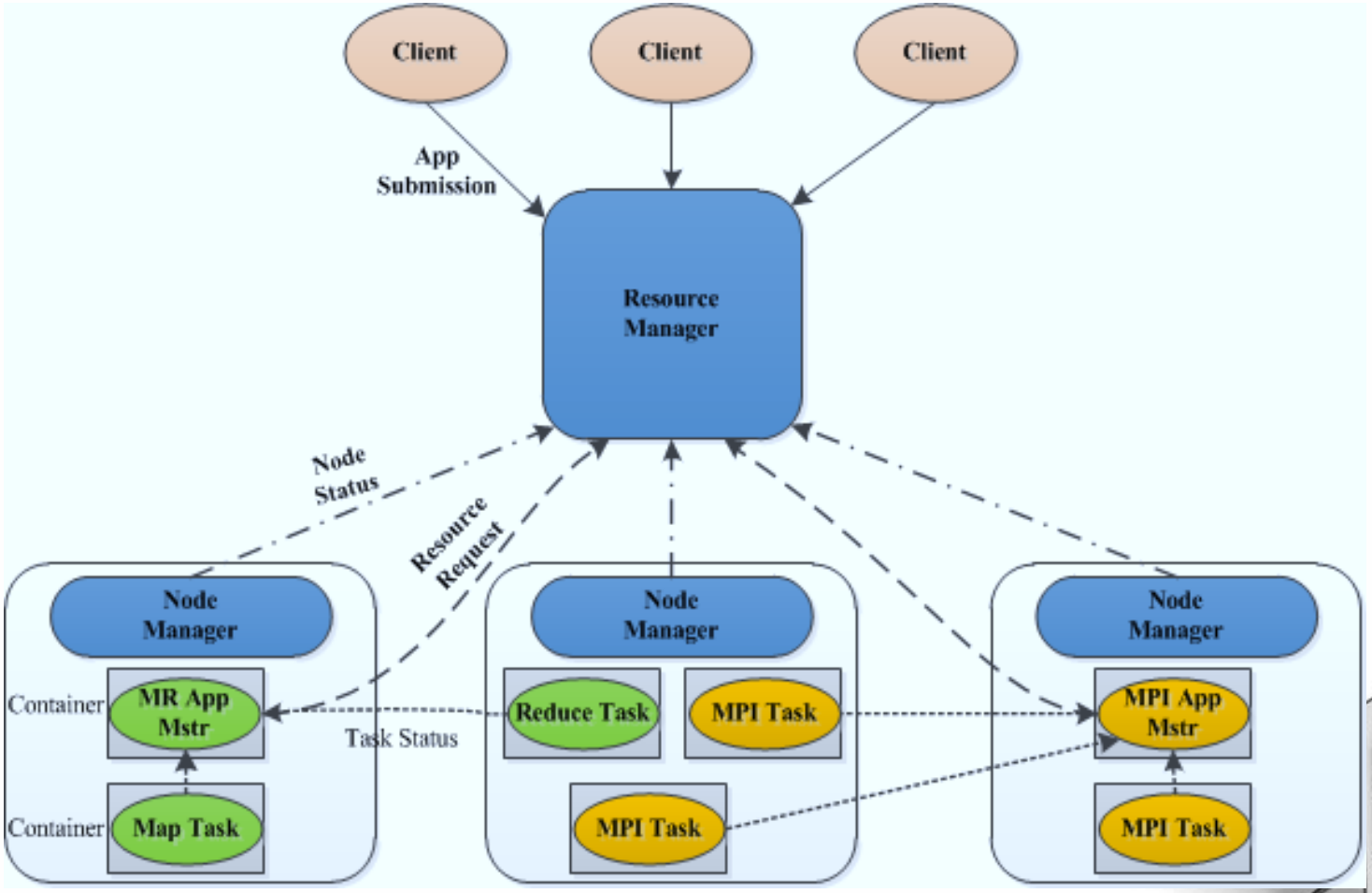
小象科技 ChinaHadoop.cn

- 直接源于MRv1在几个方面的缺陷
  - ✓扩展性受限
  - ✓单点故障
  - ✓难以支持MR之外的计算
- 多计算框架各自为战，数据共享困难
  - ✓**MR：离线计算框架**
  - ✓**Storm：实时计算框架**
  - ✓**Spark：内存计算框架**

目录

小象科技
ChinaHadoop.cn

YARN基本架构


Client
Client
Client
App Submission
Resource Manager
Node Status
Resource Request
Node Manager
Container
MR App Mstr
Container
Map Task
Task Status
Node Manager
Reduce Task
MPI Task
MPI Task
Node Manager
MPI App Mstr
MPI Task

ResourceManager

小象科技 ChinaHadoop.cn

- 整个集群只有一个，负责集群资源的统一管理和调度
- 详细功能
  - ✓处理客户端请求
  - ✓启动/监控ApplicationMaster
  - ✓监控NodeManager
  - ✓资源分配与调度

NodeManager

小象科技 ChinaHadoop.cn

- 整个集群有多个，负责单节点资源管理和使用
- 详细功能
  - ✓ 单个节点上的资源管理和任务管理
  - ✓ 处理来自ResourceManager的命令
  - ✓ 处理来自ApplicationMaster的命令

ApplicationMaster

小象科技 ChinaHadoop.cn

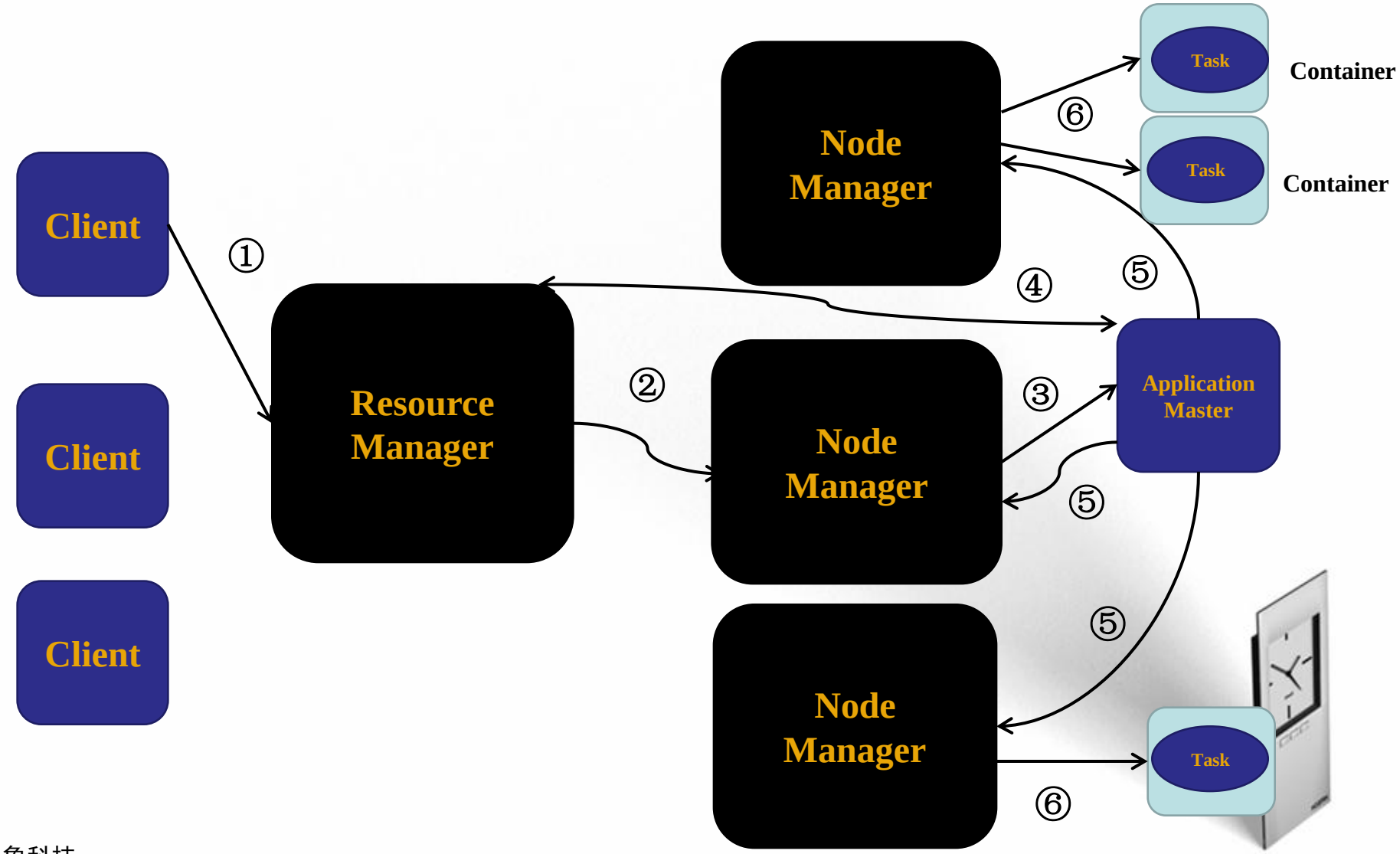
- 每个应用有一个，负责应用程序的管理
- 详细功能
  - ✓数据切分
  - ✓为应用程序申请资源，并进一步分配给内部任务
  - ✓任务监控与容错

Container

小象科技 ChinaHadoop.cn

- 对任务运行环境的抽象
- 描述一系列信息
  - ✓任务运行资源（节点、内存、CPU）
  - ✓任务启动命令
  - ✓任务运行环境

YARN运行过程剖析


Client
Client
Client
①
Resource Manager
②
Node Manager
Node Manager
Node Manager
③
④
⑤
⑤
⑤
⑥
⑥
Application Master
Task
Container
Task
Container
Task

YARN容错性

小象科技 ChinaHadoop.cn

ResourceManager

- ✓ 存在单点故障；
- ✓ 正在基于ZooKeeper实现HA。

NodeManager

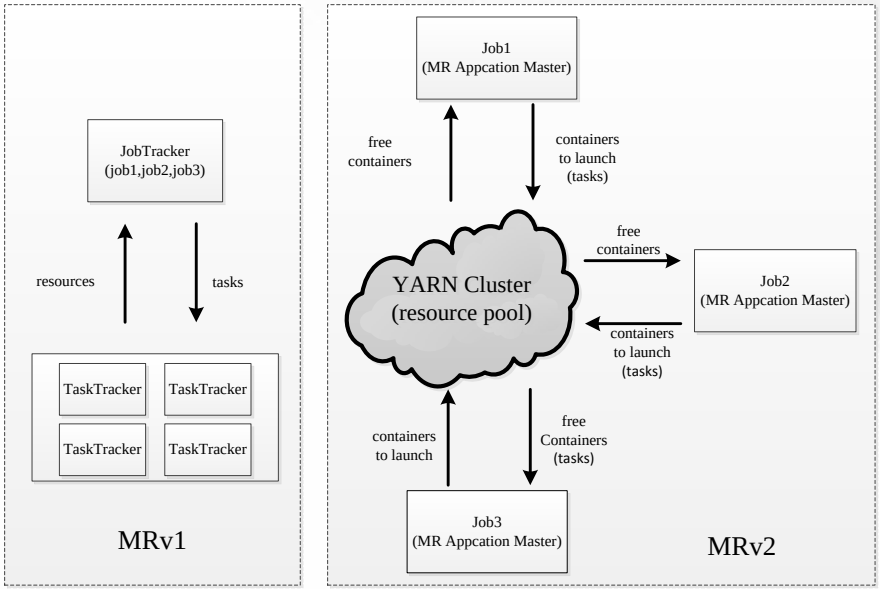
- ✓ 失败后，RM将失败任务告诉对应的AM；
- ✓ AM决定如何处理失败的任务。

ApplicationMaster

- ✓ 失败后，由RM负责重启；
- ✓ AM需处理内部任务的容错问题；
- ✓ RMAppMaster会保存已经运行完成的Task，重启后无需重新运行。

YARN调度框架

- **双层调度框架**
  - ✓ RM将资源分配给AM
  - ✓ AM将资源进一步分配给各个Task
- **基于资源预留的调度策略**
  - ✓ 资源不够时，会为Task预留，直到资源充足
  - ✓ 与“all or nothing”策略不同（Apache Mesos）



YARN资源调度器

小象科技 ChinaHadoop.cn

- **多类型资源调度**
  - ✓ 采用DRF算法（论文：“Dominant Resource Fairness: Fair Allocation of Multiple Resource Types”）
  - ✓ 目前支持CPU和内存两种资源
- **提供多种资源调度器**
  - ✓ FIFO
  - ✓ Fair Scheduler
  - ✓ Capacity Scheduler
- **多租户资源调度器**
  - ✓ **支持资源按比例分配**
  - ✓ **支持层级队列划分方式**
  - ✓ **支持资源抢占**

YARN资源隔离方案

小象科技 ChinaHadoop.cn

- **支持内存和CPU两种资源隔离**
  - ✓ 内存是一种“决定生死”的资源
  - ✓ CPU是一种“影响快慢”的资源
- **内存隔离**
  - ✓ 基于线程监控的方案
  - ✓ 基于Cgroups的方案
- **CPU隔离**
  - ✓ 默认不对CPU资源进行隔离
  - ✓ 基于Cgroups的方案

YARN支持的调度语义

小象科技 ChinaHadoop.cn

支持的语义

- ✓ 请求某个特定节点/机架上的特定资源量
- ✓ 将某些节点加入（或移除）黑名单，不再为自己分配这些节点上的资源
- ✓ 请求归还某些资源

不支持的语义

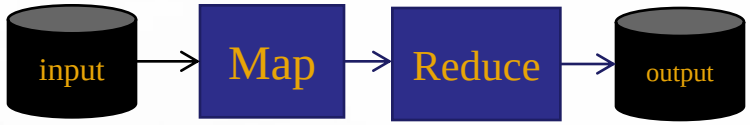
- ✓ 请求任意节点/机架上的特定资源量
- ✓ 请求一组或几组符合某种特质的资源
- ✓ 超细粒度资源
- ✓ 动态调整Container资源

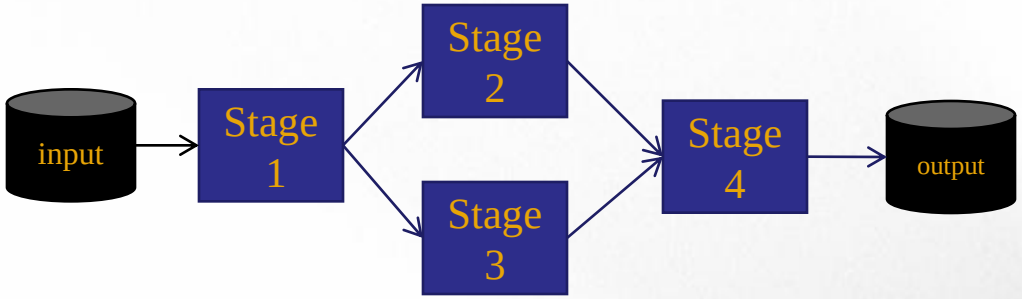
目录

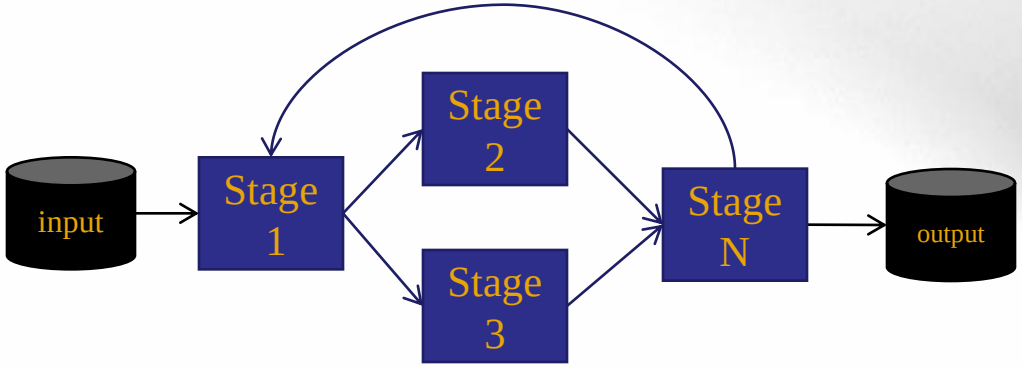
小象科技
ChinaHadoop.cn

应用程序种类繁多


input
Map
output


input
Map
Reduce
output


input
Stage 1
Stage 2
Stage 3
Stage 4
output


input
Stage 1
Stage 2
Stage 3
Stage N
output

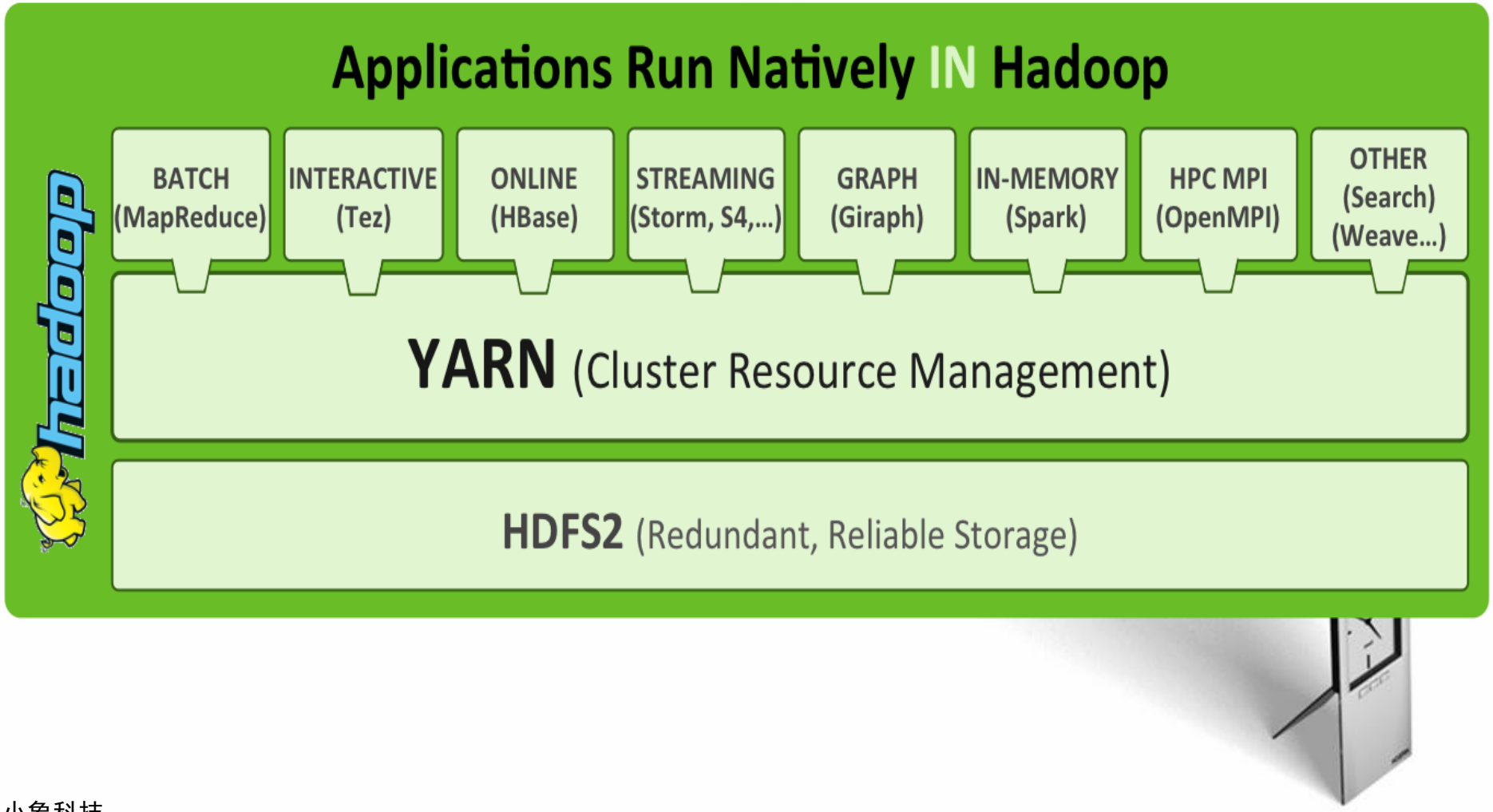
YARN设计目标

小象科技 ChinaHadoop.cn

- 通用的统一资源管理系统
  - ✓ 同时运行长应用程序和短应用程序
- 长应用程序
  - ✓ 通常情况下，永不停止运行的程序
  - ✓ Service、HTTP Server等
- 短应用程序
  - ✓ 短时间（秒级、分钟级、小时级）内会运行结束的程序
  - ✓ MR job、Spark Job等

以YARN为核心的生态系统

小象科技
ChinaHadoop.cn



运行在YARN上的计算框架

小象科技 ChinaHadoop.cn

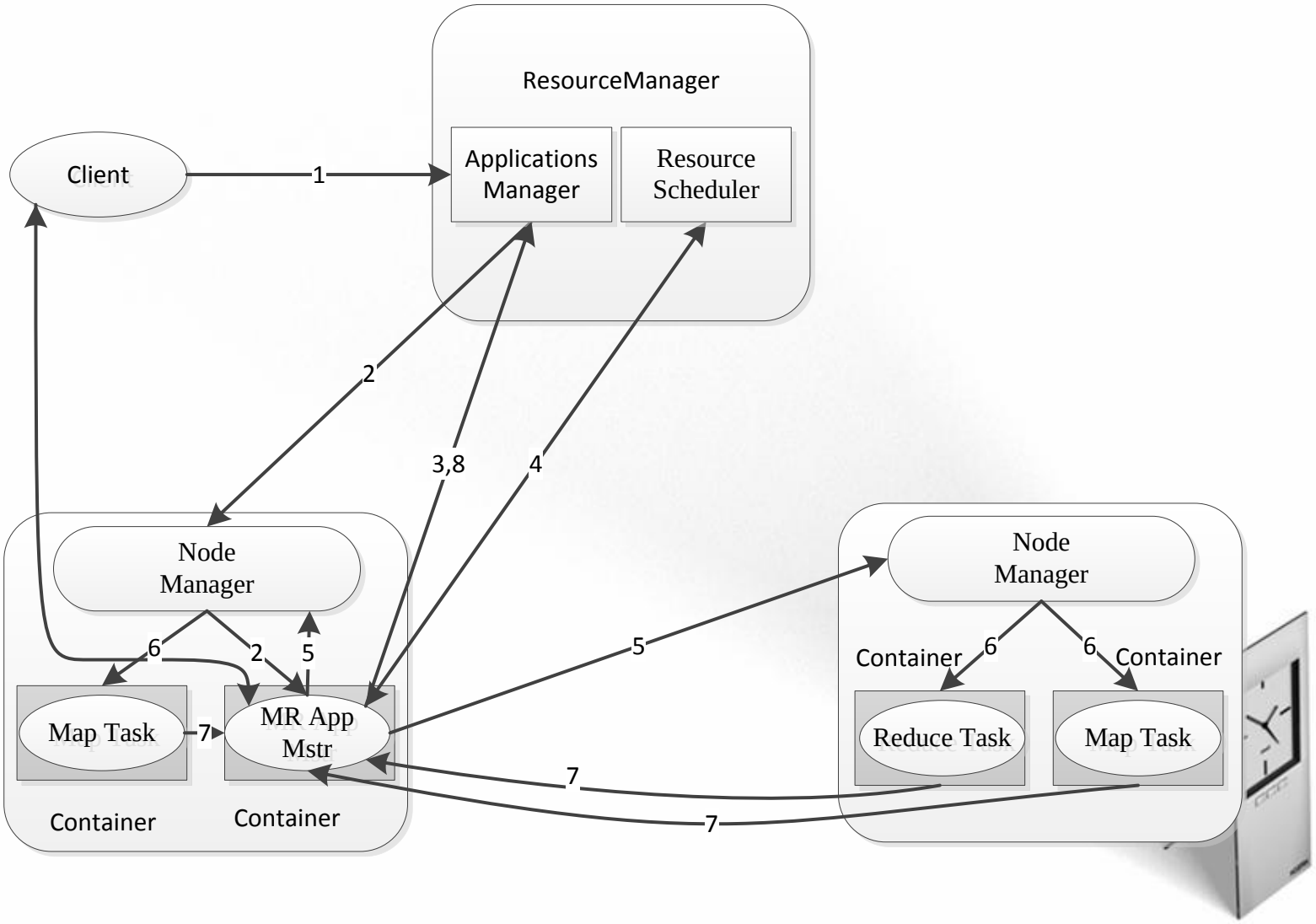
- **离线计算框架：MapReduce**
- **DAG计算框架：Tez**
- **流式计算框架：Storm**
- **内存计算框架：Spark**

离线计算框架MapReduce

小象科技 ChinaHadoop.cn

- 将计算过程分为两个阶段，Map和Reduce
  - ✓Map 阶段并行处理输入数据
  - ✓Reduce阶段对Map结果进行汇总
- Shuffle连接Map和Reduce两个阶段
  - ✓Map Task将数据写到本地磁盘
  - ✓Reduce Task从每个Map Task上读取一份数据
- 仅适合离线批处理
  - ✓具有很好的容错性和扩展性
  - ✓适合简单的批处理任务
- 缺点明显
  - ✓启动开销大、过多使用磁盘导致效率低下等

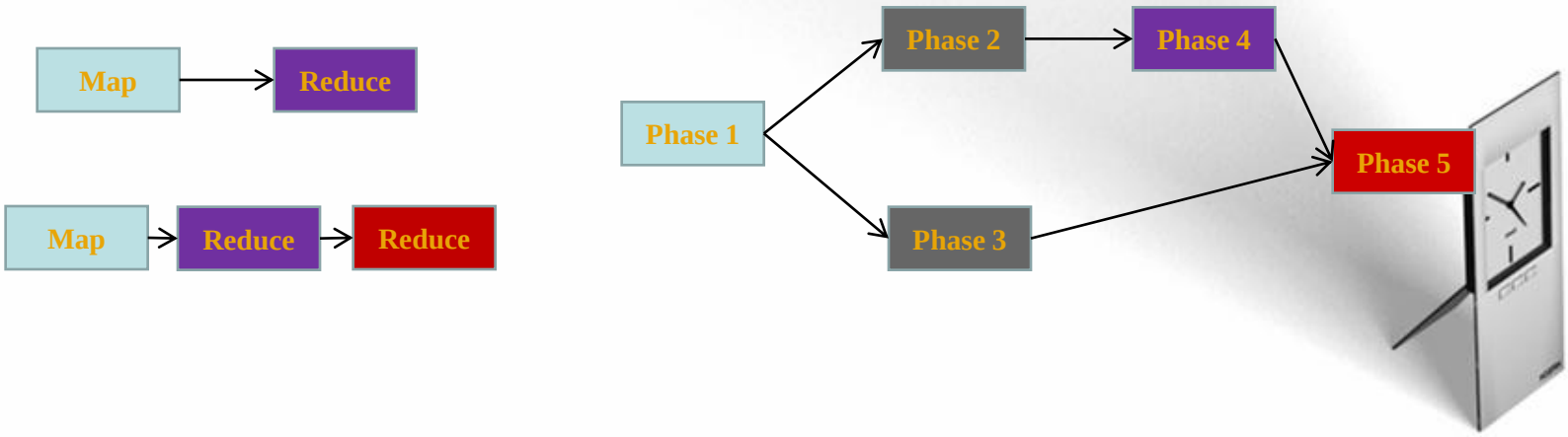
MapReduce On YARN



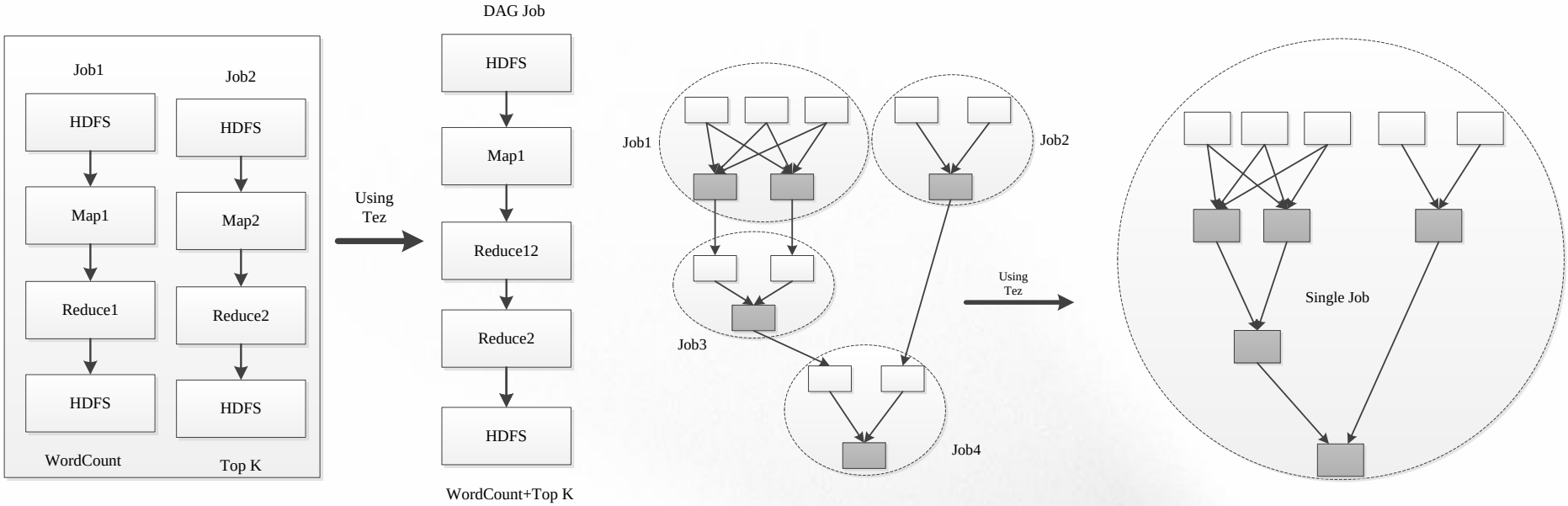
DAG计算框架Tez

小象科技 ChinaHadoop.cn

- 多个作业之间存在数据依赖关系，并形成一个依赖关系有向图（Directed Acyclic Graph），该图的计算称为“DAG计算”
- Apache Tez：基于YARN的DAG计算框架
  - ✓运行在YARN之上，充分利用YARN的资源管理和容错等功能；
  - ✓提供了丰富的数据流（dataflow）API；
  - ✓扩展性良好的“Input-Processor-Output”运行时模型；
  - ✓动态生成物理数据流关系。

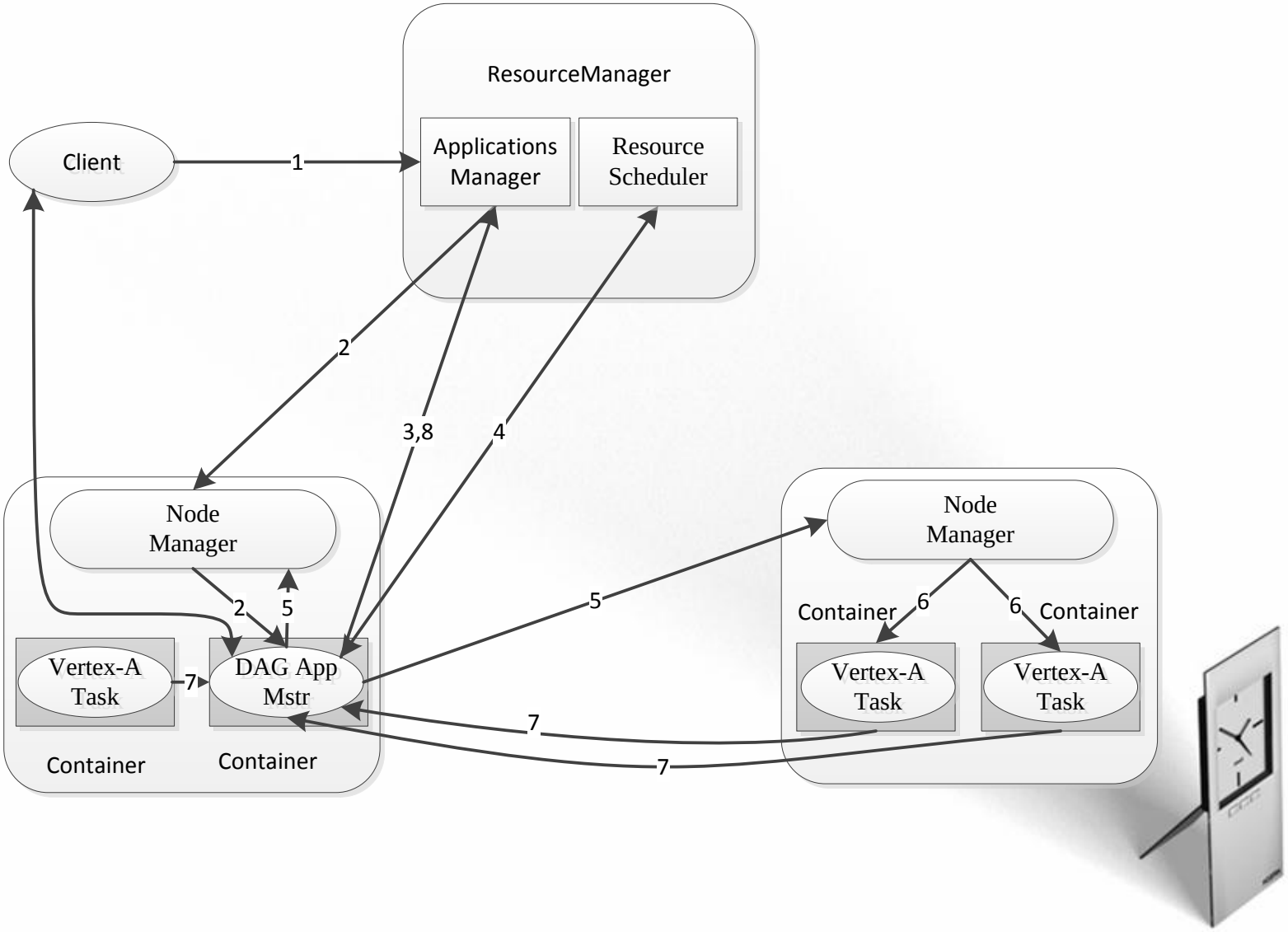


DAG计算框架Tez




```
SELECT a.state, COUNT(*), AVERAGE(c.price)
  FROM a
  JOIN b ON(a.id = b.id)
  JOIN c ON(a.itemId = c.itemId)
  GROUP BY a.state
```

Tez On YARN


ResourceManager
Applications Manager
Resource Scheduler
Client
1
2
3,8
4
Node Manager
2
5
Vertex-A Task
7
DAG App Mstr
Container
Container
5
Node Manager
Container
6
6
Container
Vertex-A Task
Vertex-A Task
7
7

Tez 优化技术

- ApplicationMaster缓冲池
  - ✓作业提交到AMPoolServer服务上
  - ✓预启动若干个ApplicationMaster，形成一个ApplicationMaster缓冲池
- 预先启动Container
  - ✓ApplicationMaster启动时可以预先启动若干个Container
- Container重用
  - ✓任务运行完成后，ApplicationMaster不会马上注销它使用的Container，而是将它重新分配给其他未运行的任务

Tez 应用场景

小象科技 ChinaHadoop.cn

- 直接编写应用程序
  - ✓Tez提供了一套通用编程接口
  - ✓适合编写有依赖关系的作业
- 优化Pig、Hive等引擎
  - ✓下一代Hive：Stinger
  - ✓好处1：避免查询语句转换成过多的MapReduce作业后产生大量不必要的网络和磁盘IO
  - ✓好处2：更加智能的任务处理引擎

流式计算框架Storm

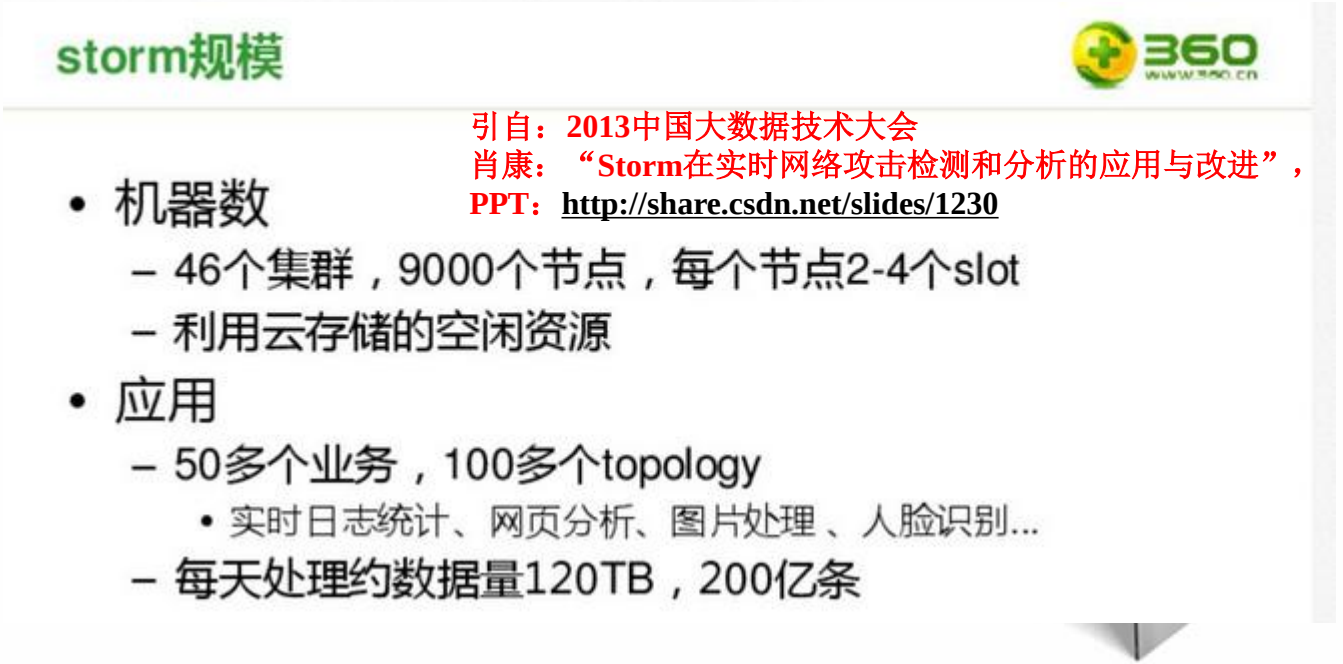
小象科技 ChinaHadoop.cn

➢流式（Streaming）计算，是指被处理的数据像流水一样不断流入系统，而系统需要针对每条数据进行实时处理和计算，并永不停止（直到用户显式杀死进程）；

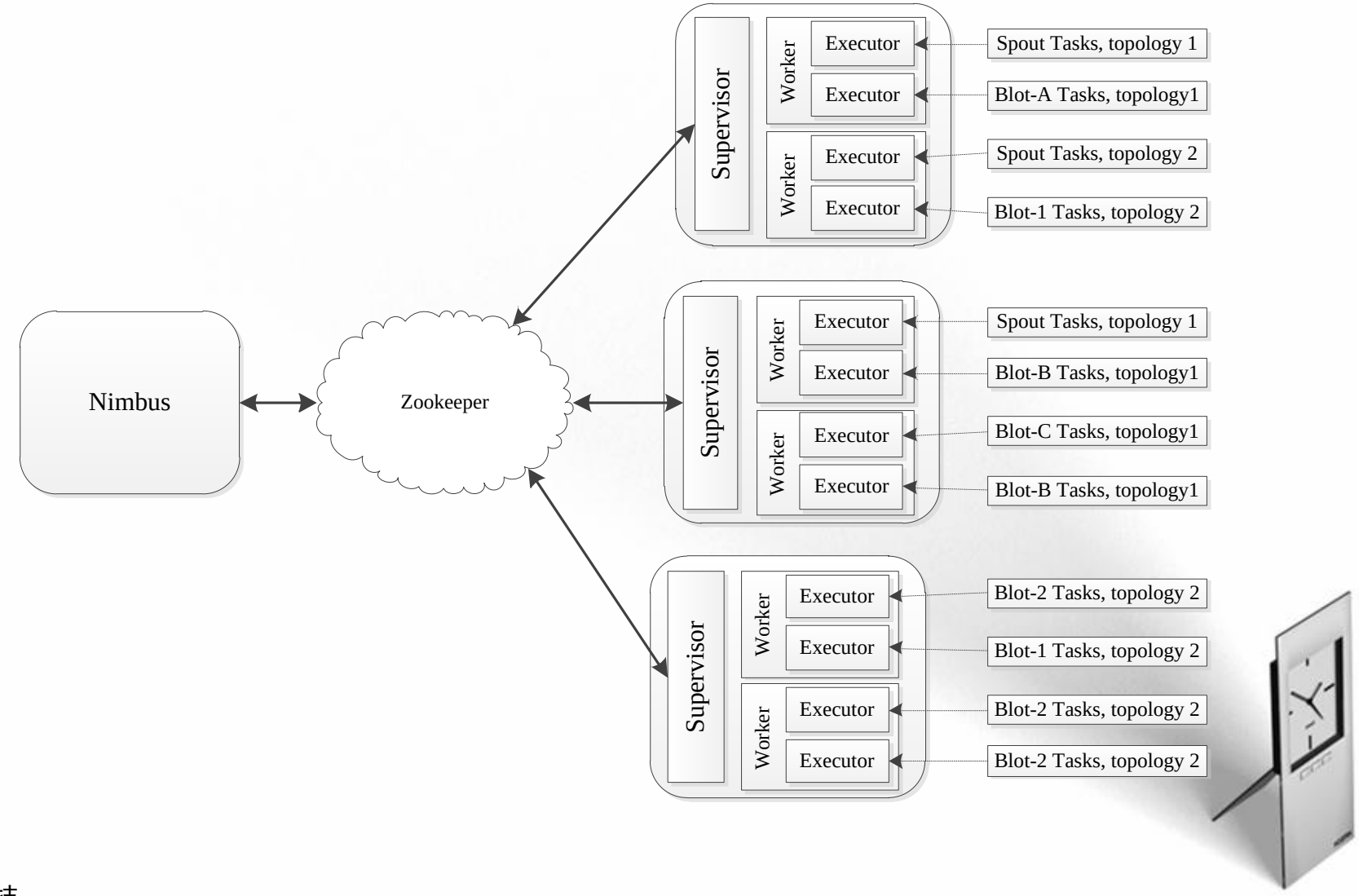
➢传统做法：由消息队列和消息处理者组成的实时处理网络进行实时计算；

- ✓缺乏自动化
- ✓缺乏健壮性
- ✓伸缩性差

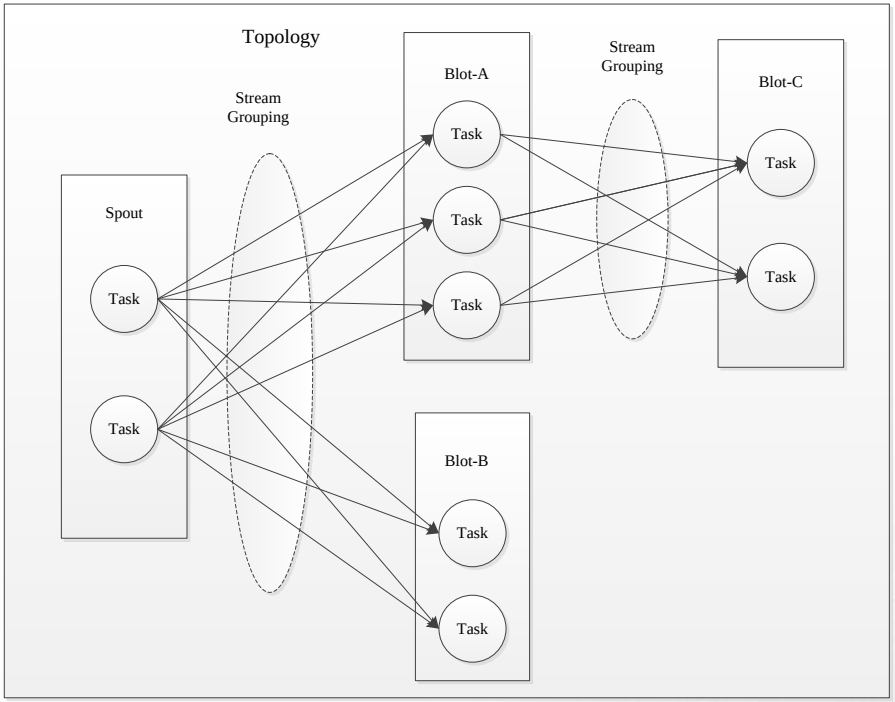
➢**Storm出现。**



流式计算框架Storm

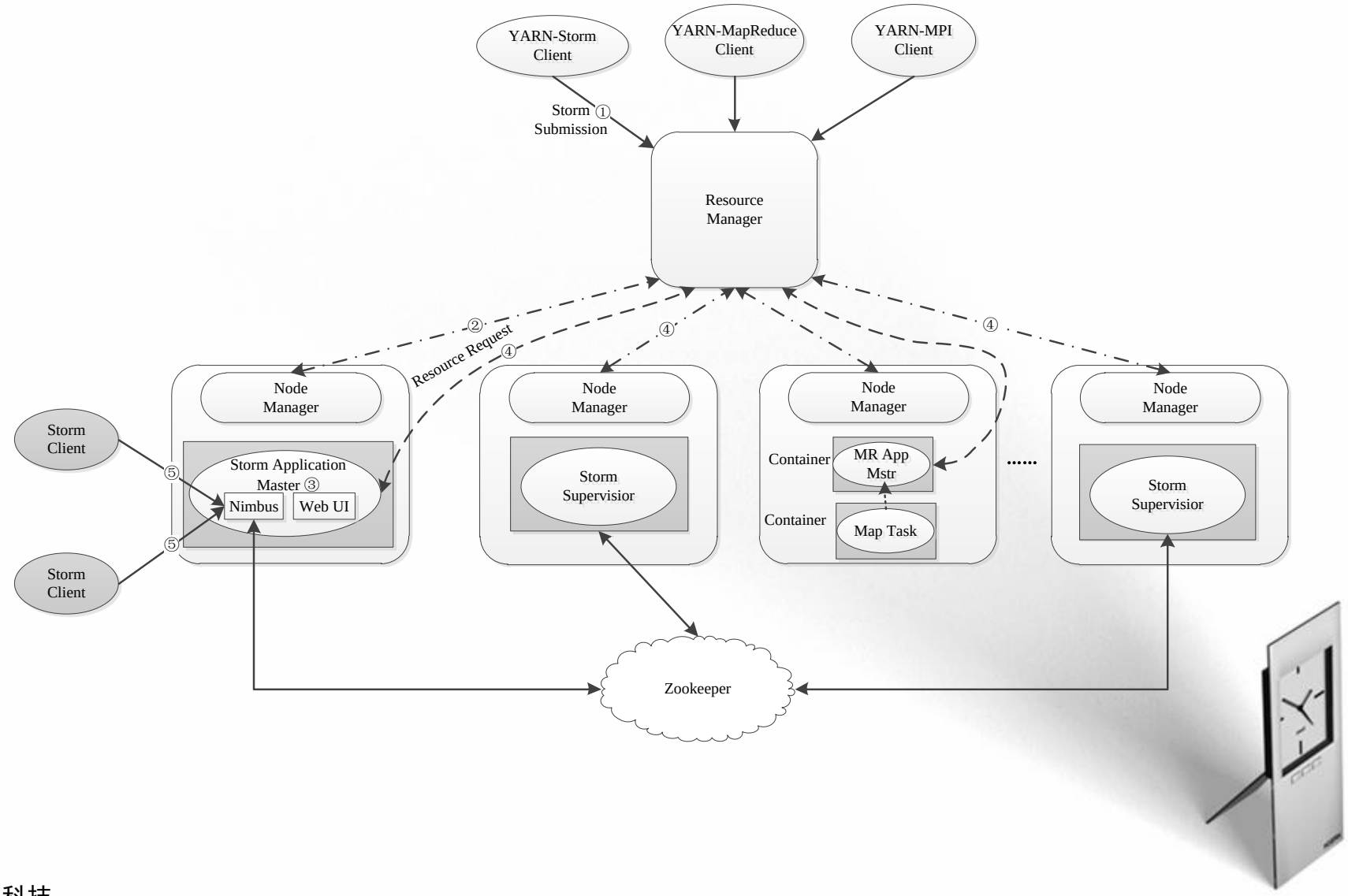


流式计算框架Storm

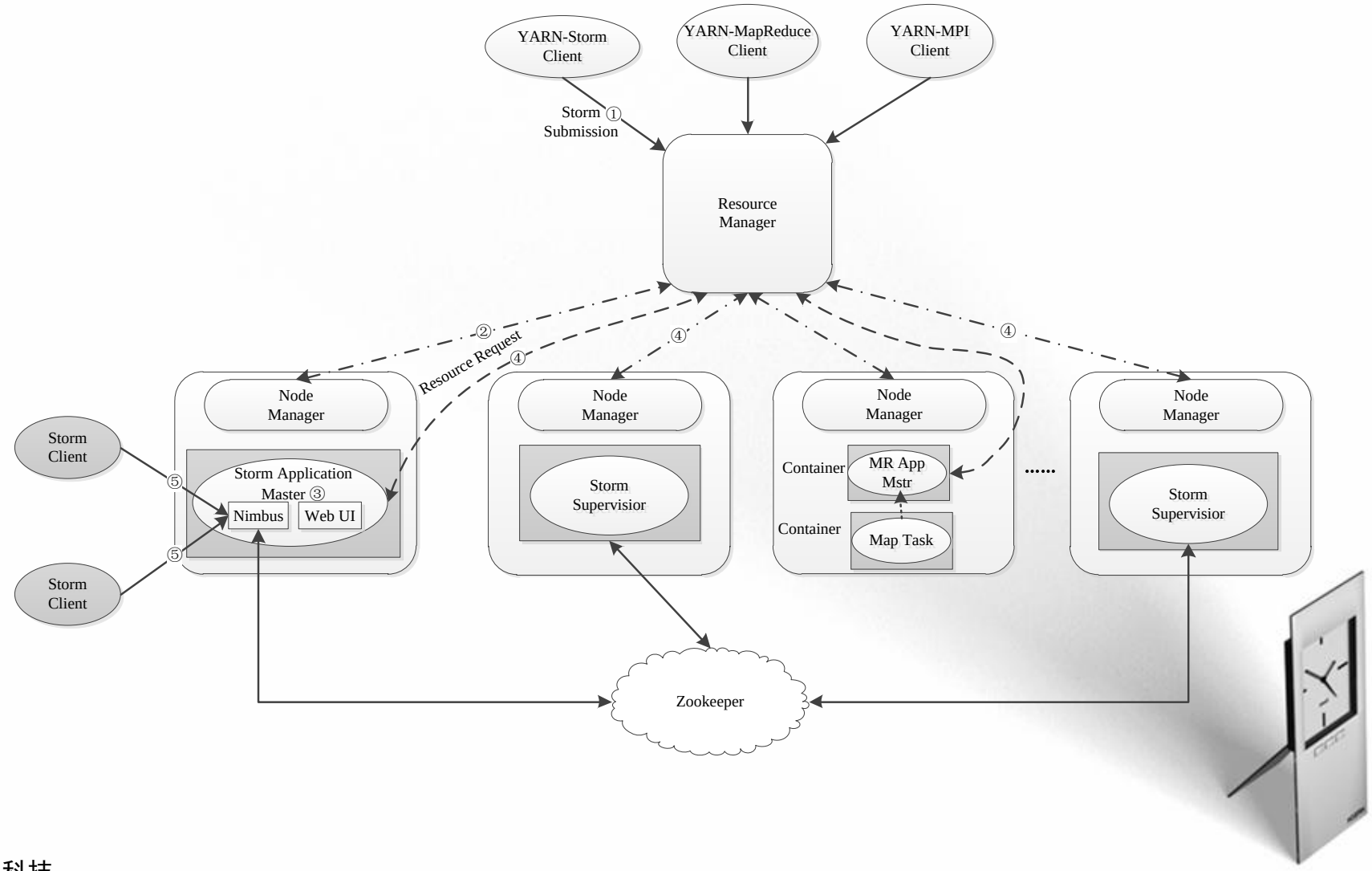


	Hadoop MapReduce（MRv1）	Storm
系统服务	JobTracker	Nimbus
	TaskTracker	Supervisor
	Child	Worker
应用程序名称	Job	Topology
编程模型	Map/Reduce	Spout/Blot
	Shuffle	Stream Grouping

Storm On YARN



Storm On YARN



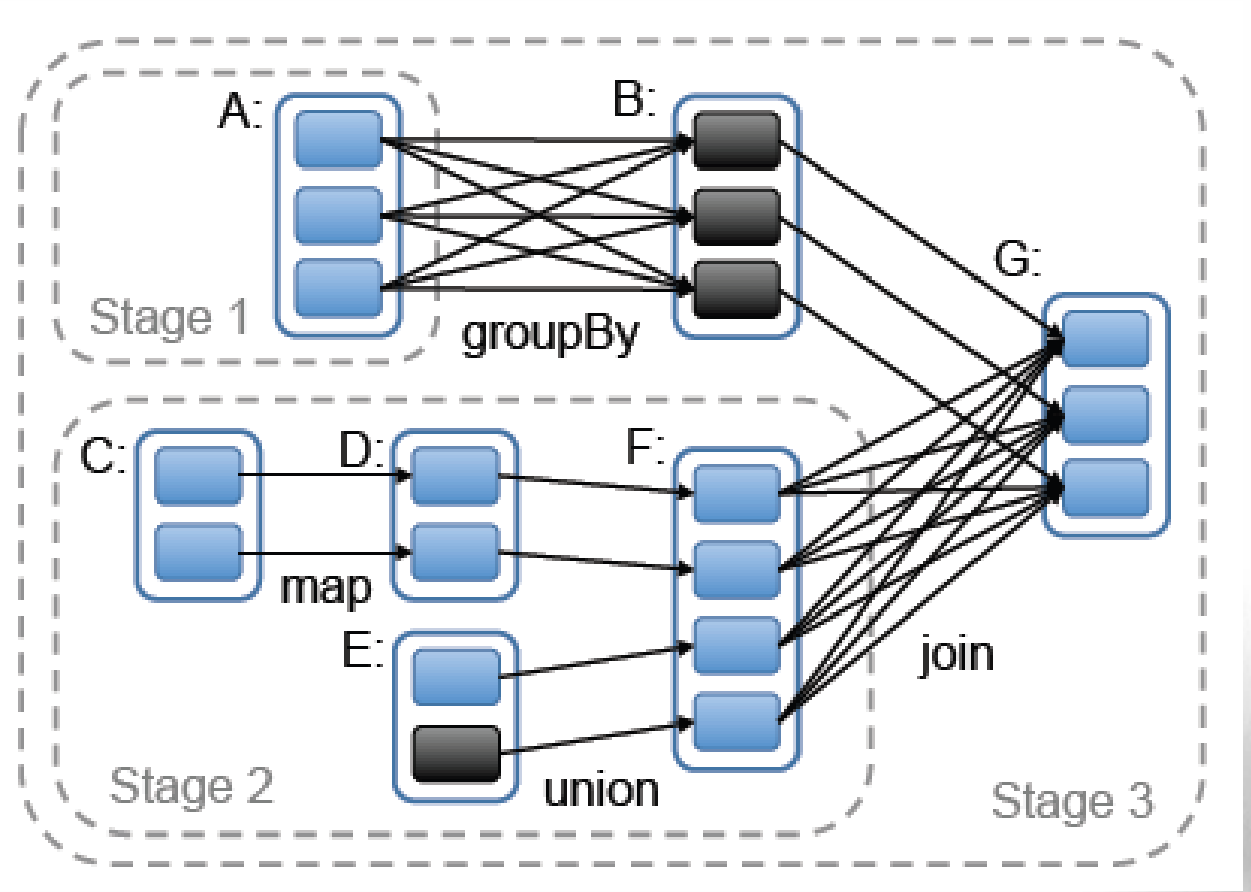
内存计算框架Spark

小象科技 ChinaHadoop.cn

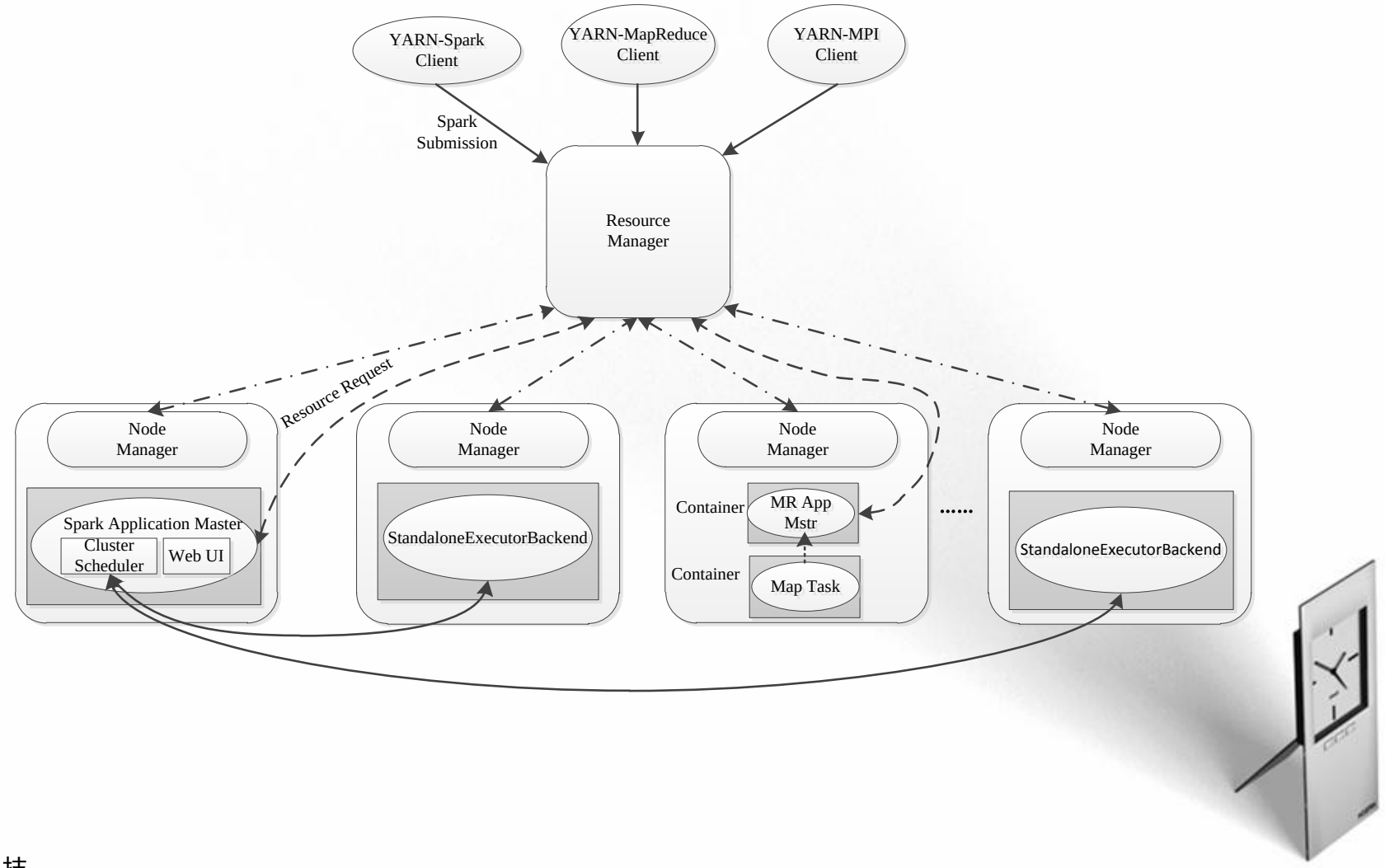
- 克服MapReduce在迭代式计算和交互式计算方面的不足；
- 引入RDD（Resilient Distributed Datasets）数据表示模型；
- RDD是一个有容错机制，可以被并行操作的数据集合，能够被缓存到内存或磁盘上。

引自：“基于Spark on Yarn的淘宝数据挖掘平台”，
PPT：http://vdisk.weibo.com/s/dn9d7A_XuVrf

内存计算框架Spark

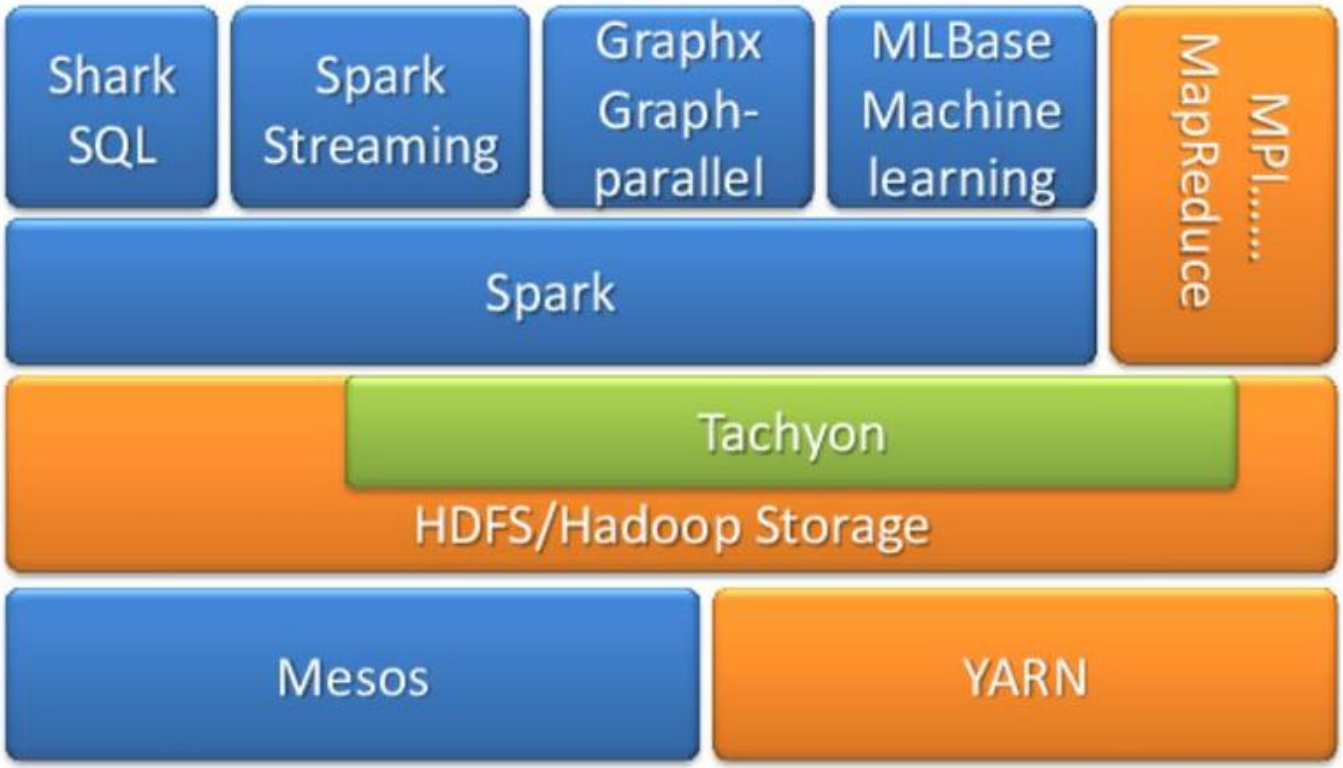


Spark On YARN



Spark 生态系统

小象科技 ChinaHadoop.cn



其他运行在YARN上的计算框架

- **Hoya：HBase on YARN；**

https://github.com/hortonworks/hoya/

- **LLAMA：Impala On YARN**

http://cloudera.github.io/llama/

- **Kafka On YARN**

https://github.com/kkasravi/kafka-yarn

目录

小象科技 ChinaHadoop.cn

MapReduce 2.0与YARN

小象科技 ChinaHadoop.cn

- 一个MR应用程序的成功运行需要若干模块：
  - ✓ 任务管理和资源调度
  - ✓ 任务驱动模块（MapTask、ReduceTask）
  - ✓ 用户代码（Mapper、Reducer…）
- MapReduce 2.0和YARN区别：
  - ✓ YARN是一个资源管理系统，负责资源管理和调度
  - ✓ MapReduce只是运行在YARN上的一个应用程序
  - ✓ 如果把YARN看做“android”，则MapReduce只是一个“app”

MapReduce 2.0组成

小象科技 ChinaHadoop.cn

- **MapReduce2.0组成：**
  - ✓ **YARN(整个集群只有一个)**
  - ✓ **MRAppMaster（一个应用程序一个）**
  - ✓ **用户代码（Mapper、Reducer…）**
- **MapReduce 1.0和MapReduce 2.0区别：**
  - ✓ **MapReduce 1.0是一个独立的系统，直接运行在Linux之上**
  - ✓ **MapReduce 2.0则是运行YARN上的框架，且可与多种框架一起运行在YARN上**

目录

小象科技
ChinaHadoop.cn

主要内容

小象科技 ChinaHadoop.cn

- Hadoop YARN产生背景
- Hadoop YARN基本构成
- Hadoop YARN应用
- MapReduce 2.0与YARN

引用链接

小象科技 ChinaHadoop.cn

- **Hadoop；**
  http://hadoop.apache.org/
- **Tez**
  http://tez.incubator.apache.org/
- **Storm**
  http://storm-project.net/
- **Storm On YARN**
  https://github.com/yahoo/storm-yarn
- **Spark**
  http://spark.incubator.apache.org/
- **Spark On YARN**
  https://github.com/apache/incubator-spark/tree/branch-0.8/yarn