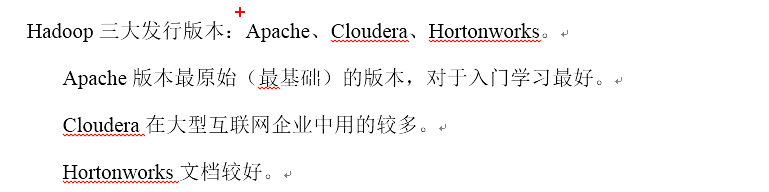


Hadoop发行版本



海量数据存储和计算

大量，多样性，高速，低价值密度

应用场景：

物流仓储

旅游：分析对应地区人的口味

超市

商品广告推广

保险

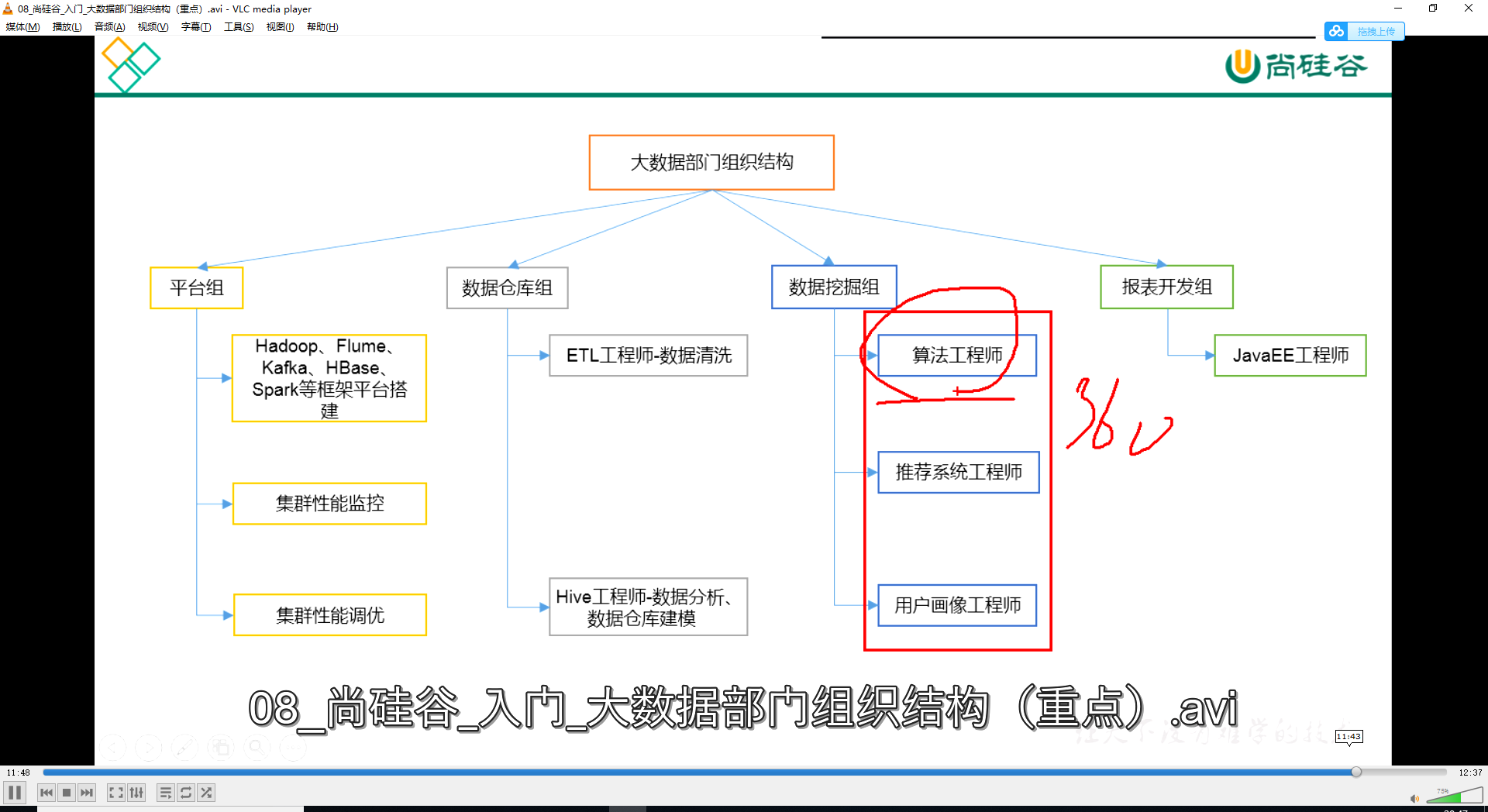
金融

房产

征信

人工智能

大数据部门组织结构



Hadoop 是一个由Apache 基金会所开发的分布式系统基础架构

解决海量的数据的存储和海量数据的分析计算



生态圈

Nutch



Apache 版本最原始的基础版本，对于入门学习

Cloudera 在大型互联网企业中用的较多 收费

Hortonworks 文档较好

高可用

高扩展：在集群间分配任务数据，可方便的扩展数以前计的节点

高可靠：Hadoop 底层维护多个数据副本，即使Hadoop 某个计算元素或存储出现故障，也不会导致数据丢失

高效性：在MapReduce的思想下，Hadoop是并行工作，以加快任务处理速度

高容错性：能够自动将失败的任务重新分配

Hadoop 1.x

组成：

Common辅助工具

HDFS 数据存储

MapReduce：计算+资源调度 （耦合性比较大）

Hadoop 2.x

Common辅助工具

HDFS 数据存储

nameNode:文件的元数据，如文件名，文件目录接口，文件属性（生成时间，副本数，文件权限）(目录)

dataNode:在本地文件系统存储文件块数据，以及块数据的校验和4

Secondary NameNode(2nn):用来监控HDFS 状态的辅助后台程序，每隔一段时间获取FDFS元数据的快照。（详情）

Yarn 资源调度

ResourceManger 服务器节点 ==服务器 处理客户端请求

Node manger运行的任务job

NodeManage（NM）主要作用如下：

1. 管理单节点上的资源
2. 处理来自ResourceManger的命令
3. 处理来自ApplicationMaster的命令

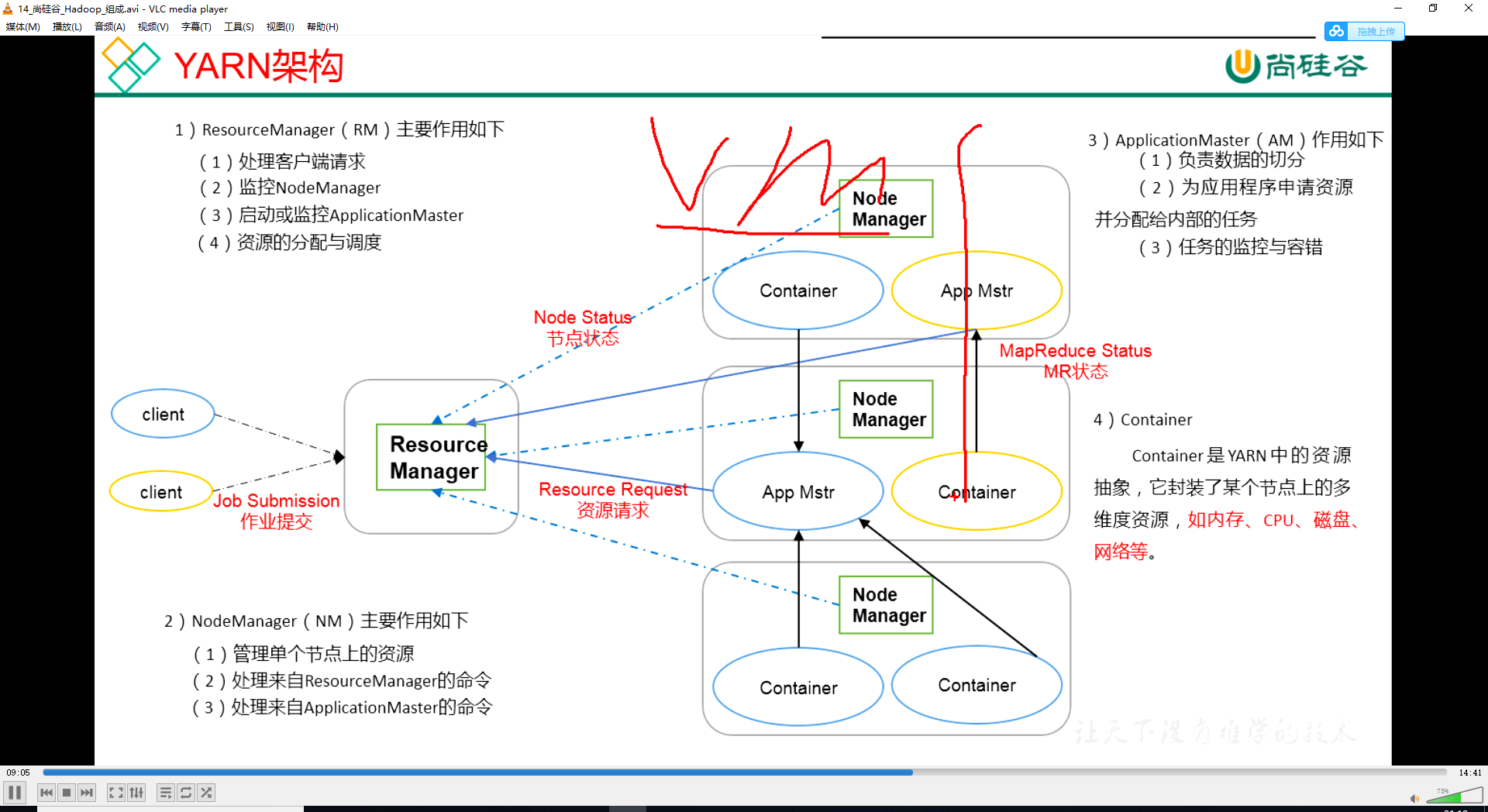
资源调度

ApplicationMaster作用

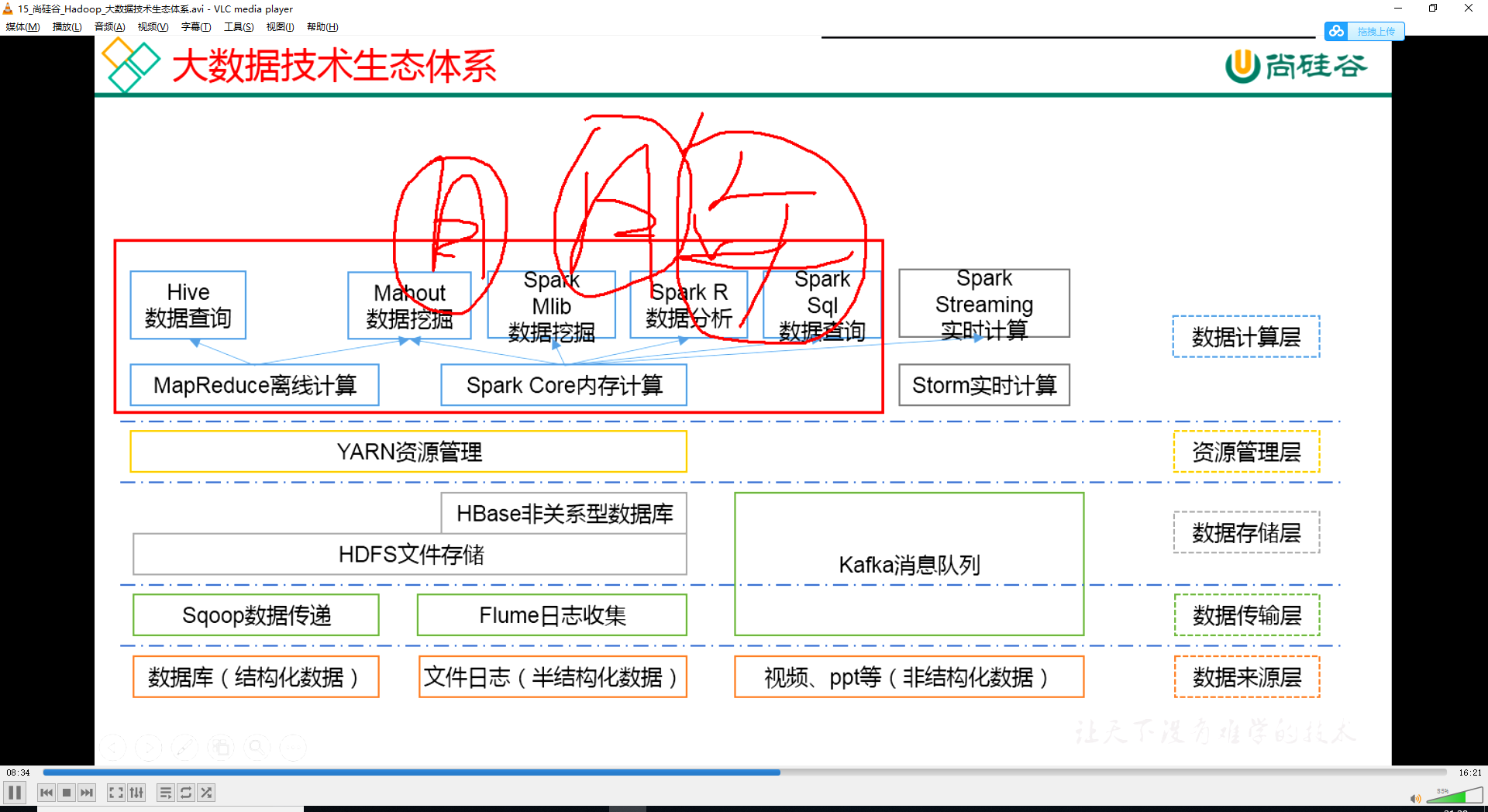
1. 负责数据的切分
2. 为应用程序申请资源并分配给内部的任务
3. 任务的监控与容错

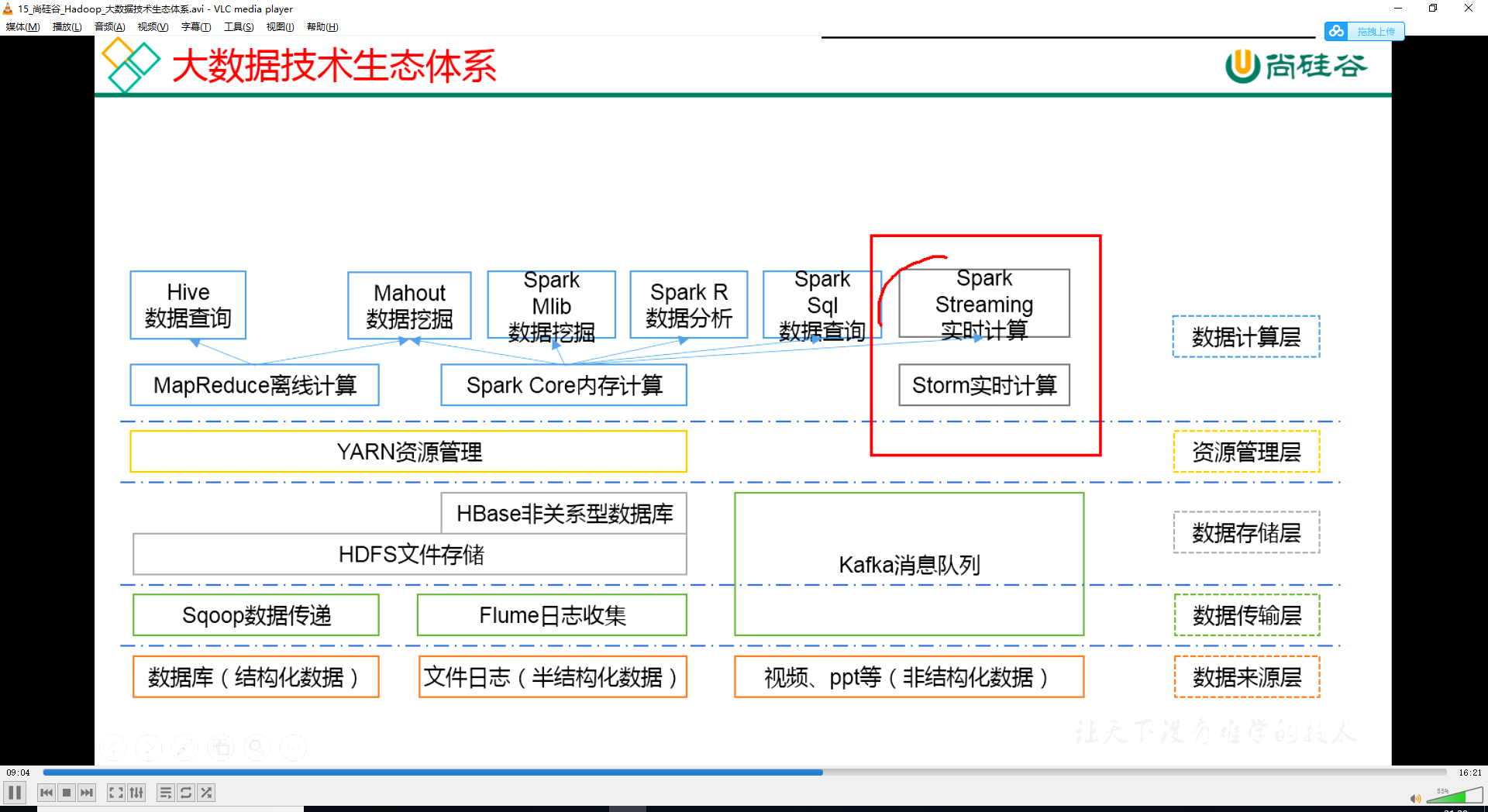
Container

1.Container 是YARN中的资源抽象，它分装了某个节点上的多维度资源

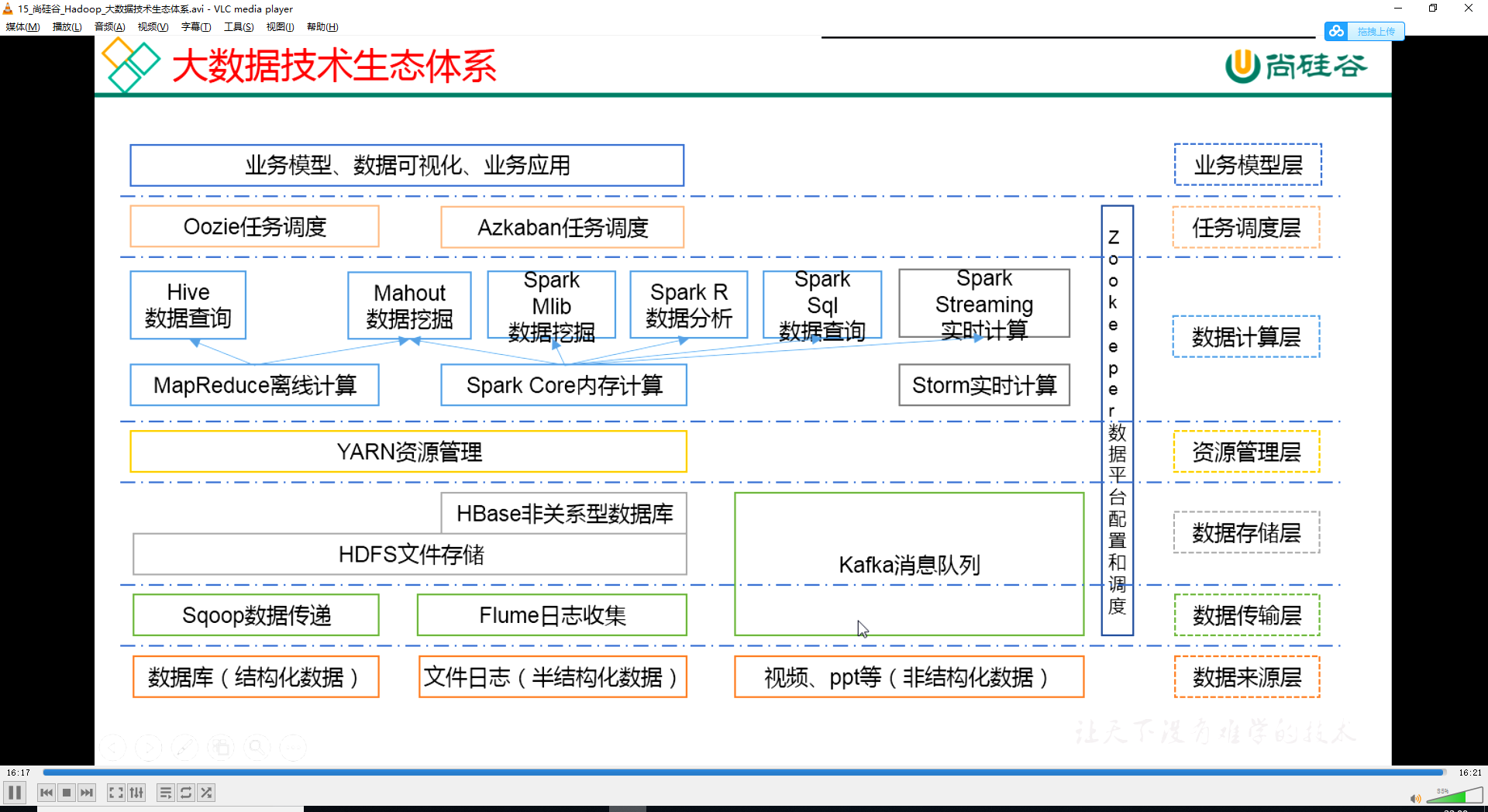


MapReduce：计算





Batch 处理



Hdfs-site.xml

3台服务器上存储3分数据 其他集群上保存，挂掉也执行在其他机器上保存

启动集群

a:格式化NameNode

格式化NameNode （第一次启动时格式化，以后就不要总格式化）

Bin/hdfs namenode -format

hdfs namenode -format

启动namenode

hadoop-daemon.sh start namenode -----启动名称节点

hadoop-daemon.sh start datanode ------启动数据节点

jps 不生效

原因：全局变量hadoop java 没有生效。解决办法：需要source/etc/profile 文件 java生效

8088 端口连不上

Hfds 端口：50070

bin/hdfs dfs -ls / 文件体统命令

bin/hdfs dfs -put LICENSE.txt /user 文件上传命令

Hadoop-env.sh 配置文件java jdk



格式化namenode

Nn: clusterID=CID-76e7156f-483b-4abf-bf57-f1d78263e481

Data :clusterID=CID-76e7156f-483b-4abf-bf57-f1d78263e481

格式化后namenode 会变，就报错

先删除dataNode数据

启动dfs

sbin/start-dfs.sh

启动yarn 资源分配

sbin/start-yarn.sh

启动yarn

resourcemanager

yarn-daemon.sh start resourcemanager

nodemanager

yarn-daemon.sh start nodemanager

zookeeper

bin/hbase-daemon.sh start zookeeper

master

bin/hbase-daemon.sh start master

启动hbase

bin/hbase-daemon.sh start zookeeper

bin/hbase-daemon.sh start master

bin/hbase-daemon.sh start regionserver

Habse

WAL 预写日志文件

向Hbase 写入数据

-存HLog ，在存储在HDFS上

-存Region

1）内存

· 2）spill到磁盘、文件、hdfs 文件

公式：内存\*0.4/60

Region RegionServer 管理

memStor storefile

./hbase/.tmp :临时目录，当对表做创建和删除的时候，会将表move到该目录下，然后进行操作

/hbase/data：核心目录，存储hbase表的数据

默认情况下，目录下有两个文件

/hbase/data/default

在用户创建表的时候，meta存储信息 ，没有指定namespace时命名空间，表就创建在此目录下

/hbase/data/hbase

系统内部创建的表

