ARCHITECTURE DAY05

内容

上午	09:00 ~ 09:30	作业讲解和回顾
	09:30 ~ 10:20	大数据
	10:30 ~ 11:20	Hadoop
	11:30 ~ 12:00	
下午	14:00 ~ 14:50	Hadoop安装与配置
	15:00 ~ 15:50	
	16:10 ~ 17:10	HDFS
	17:20 ~ 18:00	总结和答疑

大数据

大数据介绍

大数据的由来

知识讲解

- 大数据
  - 随着计算机技术的发展，互联网的普及，信息的积累已经到了一个非常庞大的地步，信息的增长也在不断的加快，随着互联网、物联网建设的加快，信息更是爆炸是增长，收集、检索、统计这些信息越发困难，必须使用新的技术来解决这些问题

什么是大数据（续1）

知识讲解

- 大数据能做什么
  - 企业组织利用相关数据分析帮助他们降低成本、提高效率、开发新产品、做出更明智的业务决策等
  - 把数据集合并后进行分析得出的信息和数据关系性，用来察觉商业趋势、判定研究质量、避免疾病扩散、打击犯罪或测定即时交通路况等
  - 大规模并行处理数据库，数据挖掘电网，分布式文件系统或数据库，云计算平和可扩展的存储系统等

大数据特性

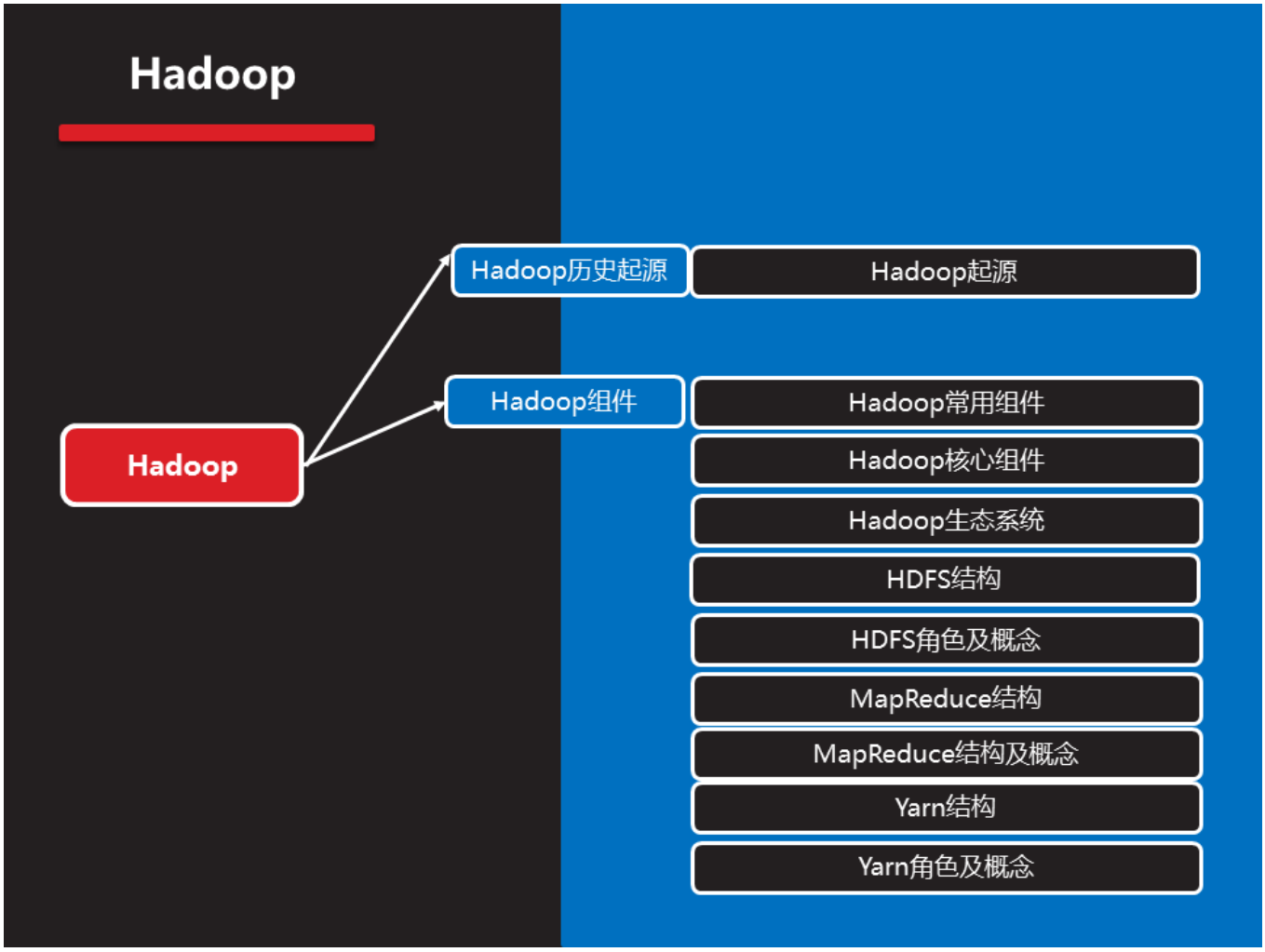
知识讲解

大数据与Hadoop

Tedu.cn 达内教育

知识讲解

- Hadoop是什么
  - Hadoop是一种分析和处理海量数据的软件平台
  - Hadoop是一款开源软件，使用JAVA开发
  - Hadoop可以提供一个分布式基础架构
- Hadoop特点
  - 高可靠性、高扩展性、高效性、高容错性、低成本


Hadoop
Hadoop
Hadoop历史起源
Hadoop起源
Hadoop组件
Hadoop常用组件
Hadoop核心组件
Hadoop生态系统
HDFS结构
HDFS角色及概念
MapReduce结构
MapReduce结构及概念
Yarn结构
Yarn角色及概念


Tedu.cn
达内教育
Hadoop历史起源

Hadoop起源（续1）

知识讲解

- BigTable
  - BigTable是存储结构化数据
  - BigTable建立在GFS，Scheduler，Lock Service和MapReduce之上
  - 每个Table都是一个多维的稀疏图

Hadoop起源（续2）

知识讲解

- GFS、MapReduce和BigTable三大技术被称为Google的三驾马车，虽然没有公布源码，但发布了这三个产品的详细设计论
- Yahoo资助的Hadoop，是按照这三篇论文的开源Java实现的，但在性能上Hadoop比Google要差很多
  - GFS - - -> HDFS
  - MapReduce - - -> MapReduce
  - BigTable - - -> Hbase

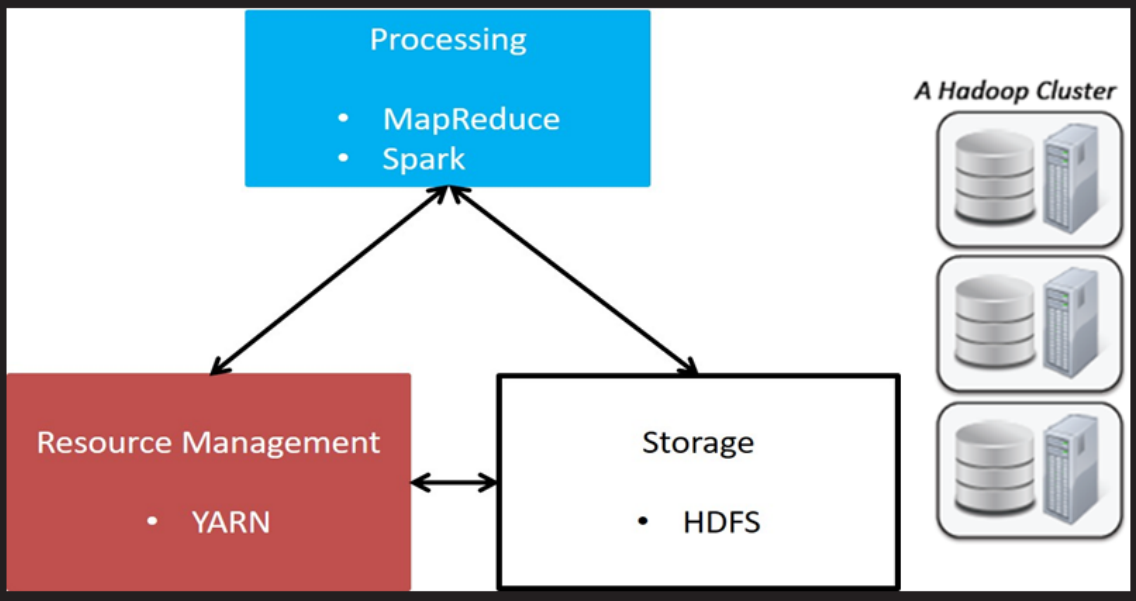
Hadoop常用组件

知识讲解

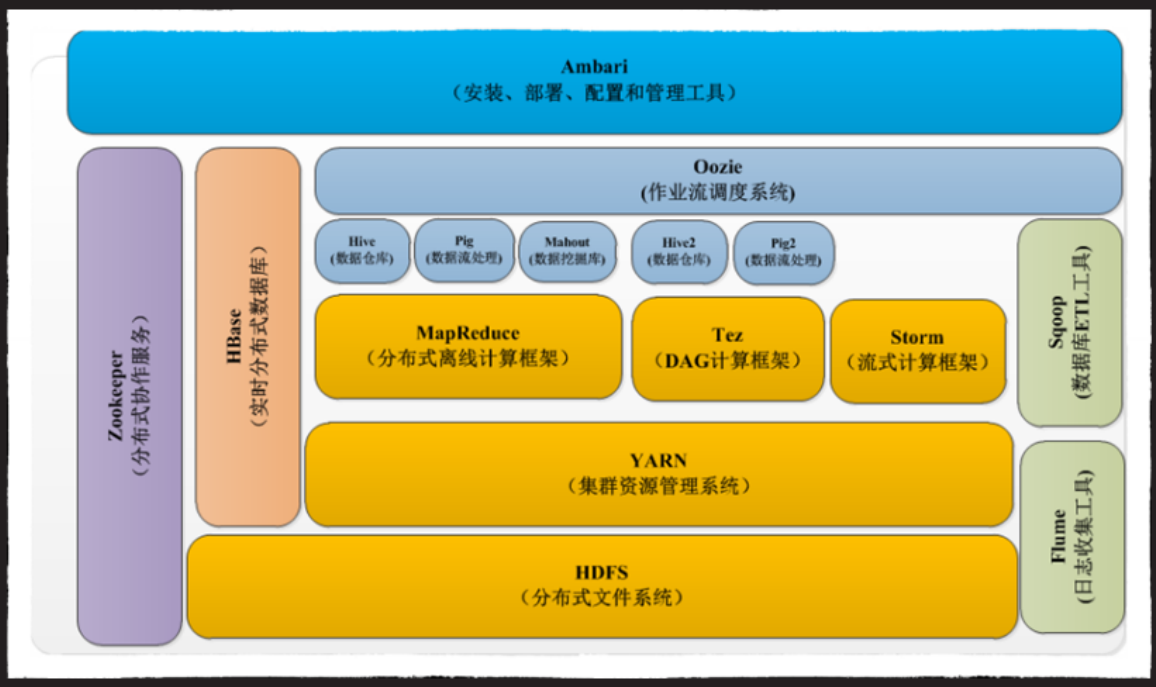
- HDFS：Hadoop分布式文件系统（核心组件）
- MapReduce：分布式计算框架（核心组件）
- Yarn：集群资源管理系统（核心组件）
- Zookeeper：分布式协作服务
- Hbase：分布式列存数据库
- Hive：基于Hadoop的数据仓库
- Sqoop：数据同步工具
- Pig：基于Hadoop的数据流系统
- Mahout：数据挖掘算法库
- Flume：日志收集工具

Hadoop核心组件

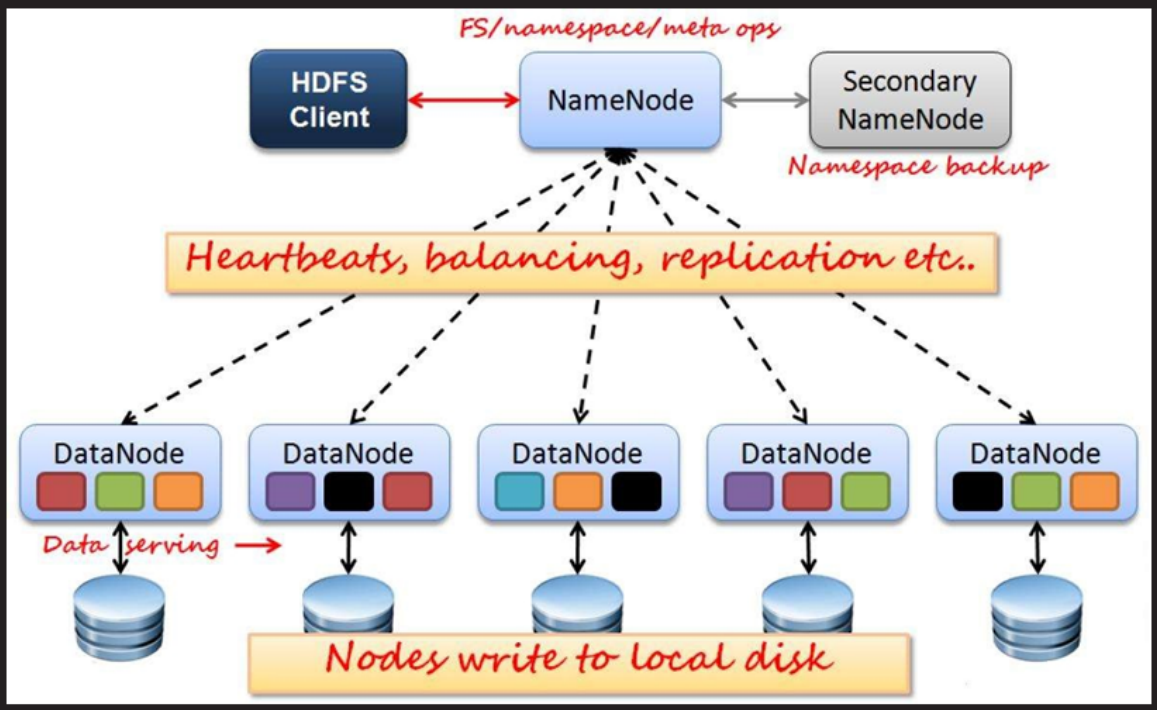
知识讲解



Hadoop生态系统



HDFS结构



HDFS角色及概念

Tedu.cn 达内教育

知识讲解

- Hadoop体系中数据存储管理的基础，是一个高度容错的系统，用于在低成本的通用硬件上运行
- 角色和概念
  - Client
  - Namenode
  - Secondarynode
  - Datanode

HDFS角色及概念（续2）

Tedu.cn 达内教育

知识讲解

- DataNode
  - 数据存储节点，存储实际的数据
  - 汇报存储信息给NameNode
- Client
  - 切分文件
  - 访问HDFS
  - 与NameNode交互，获取文件位置信息
  - 与DataNode交互，读取和写入数据

HDFS角色及概念（续3）

Tedu.cn 达内教育

- Block
  - 每块缺省128MB大小
  - 每块可以多个副本

知识讲解

MapReduce角色及概念

Tedu.cn 达内教育

知识讲解

- 源自于Google的MapReduce论文，JAVA实现的分布式计算框架

- 角色和概念
  - JobTracker
  - TaskTracker
  - Map Task
  - Reducer Task

MapReduce角色及概念（续2）

Tedu.cn 达内教育

知识讲解

- Map Task：解析每条数据记录，传递给用户编写的map()并执行，将输出结果写入本地磁盘
  - 如果为map-only作业，直接写入HDFS
- Reducer Task：从Map Task的执行结果中，远程读取输入数据，对数据进行排序，将数据按照分组传递给用户编写的reduce函数执行

Yarn角色及概念（续3）

知识讲解

- Container
  - 对任务运行行环境的抽象，封装了CPU 、内存等
  - 多维资源以及环境变量、启动命令等任务运行相关的信息资源分配与调度
- ApplicationMaster
  - 数据切分
  - 为应用程序申请资源，并分配给内部任务
  - 任务监控与容错

Yarn角色及概念（续4）

知识讲解

- Client
  - 用户与Yarn交互的客户端程序
  - 提交应用程序、监控应用程序状态，杀死应用程序等

Hadoop介绍

单机模式

知识讲解

- Hadoop的单机模式安装非常简单
  - 获取软件
    http://hadoop.apache.org
  - 安装配置Java环境，安装jps工具
    安装Openjdk和Openjdk-devel
  - 设置环境变量，启动运行
  - hadoop-env.sh
    JAVA_HOME=""

单机模式（续1）

知识讲解

- Hadoop的单机模式安装很简单，只需配置好环境变量即可运行，这个模式一般用来学习和测试Hadoop的功能
  - 测试 --- 统计词频

```
# cd  /usr/local/hadoop
# mkdir  input
# cp *.txt input/
# ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.3.jar wordcount input output
```

案例1：安装Hadoop

课堂练习

1. 单机模式安装Hadoop
2. 安装JAVA环境
3. 设置环境变量，启动运行

伪分布式

Tedu.cn 达内教育

知识讲解

- 伪分布式
  - 伪分布式的安装和完全分布式类似，区别是所有角色安装在一台机器上，使用本地磁盘，一般生产环境都会使用完全分布式，伪分布式一般是用来学习和测试Hadoop的功能
  - 伪分布式的配置和完全分布式配置类似