# 大数据培训笔记

讲师 刘成武

## 教学控屏IP

10.25.131.131

电话:15908114489

ppt

hadoop-PPT

链接：http://pan.baidu.com/s/1kVBsW8R 密码：o0m3

大数据jar包：

http://pan.baidu.com/s/1nvgfSil

gs0a

spark课程包

http://pan.baidu.com/s/1i4A0ch3

b7gb

hadoop-PPT

链接：http://pan.baidu.com/s/1kVBsW8R 密码：o0m3

大数据jar包：

http://pan.baidu.com/s/1nvgfSil

gs0a

spark课程包

http://pan.baidu.com/s/1i4A0ch3

b7gb

## 概述

## Hadoop

不能用常规手段处理的数据。

处理复杂数据。

分布式计算是大数据的技术基础。

分布式特点：

1. 容错性
2. 高可用性
3. 可恢复性
4. 一致性
5. 可扩展性

### 节点

节点代表程序个体，指的进程。

### 通信

节点之间相互隔离，通信唯一方式是通过不可靠的网络。

ActiveMQ

### 存储

#### 副本

默认存储3份数据。

### 哈希函数

只有一个输入不管长度如何，都返回m个输出中的一个输出。

位运算—快

1. 随机输入产生的输出服从均匀分布。
2. 相同输入总会产生相同输出。

#### 误差问题

### 数据分布方式

#### 哈希方式

保证随机性，均匀分布

#### 按数据范围方式

1-10一份 10-20一份…

#### 数据量分布

### 数据一致性

#### 1，中心化副本控制协议

##### 副本的一致性

案例:

1 1 1 1 1

update

2 1 1 1 1

再读取，容易脏读

备份5份，

（n+1）/2

#### Primary-seondary协议

##### 数据更新

##### 数据读取

##### 数据同步

#### paxos协议

**zookeeprer使用该协议。**

四个角色

proposer：提议者

acceptor:决策者

client:议题者

#### cap原理

一致性

可用性

分区容忍性

#### BASE理论

BASE理论是对CAP理论的延伸，核心思想是即使无法做到强一致性(Strong Consistency,Cap的一致性就是强一致性)，但应用可以采用合适的方式达到最终一致性。

**ACID和BASE的区别与联系**

**ACID最求强一致性模型。**

**BASE支持的是大型分布式系统，通过牺牲强一致性获得高可用性。**

**案例:**

**user\_key**

**1 100001 liu updatetime**

## 存储管理

NOSQL数据库分类—按存储类型

Key-value redis.

列存储 hbase

文档存储 monogDB 查看录像、查看武器资料(视频、报纸…数据源 json存储)

图存储

对象存储

## 大数据计算模式

### 批处理 mapreduce spark。。。

最适合于完成大数据批处理的计算，模式是MapReduce.

MapReduce是一个单输入、两阶段(Map和Reduce)

### 交互式处理 shark，hive…

### 流式计算 s4,storm

### 迭代计算 spark twister

### 图计算 pregel,giraph。。。

### 内存计算redis,dremel

## 大数据应用

### 数据挖掘

（数据库技术、统计学、可视化、高性能计算、人工智能、机器学习）

#### 数据分析

包括数据分析（数据整理、归类），画曲线、出报表、机器学习

基于相关性:没有因果关系 （1000人中，900的人抽烟，而且这900人都有小孩，认为:抽象跟有小孩有关系。）

基于推导:有因果关系,机器学习

#### 数据挖掘（带有价值取向）

案例:买菜,50元，列车买的菜品(蔬菜\鱼…)-🡪数据分析

在菜品清单中，xx曾说过吃xx,于是买xx。--》数据挖掘。

## HDFS存储框架

Hadoop实现了一个分布式文件系统(HADOOP distributed file system)，简称HDFS

### 环境搭建

#### 安装jdk

tar –zxvf xxx.tar.gz

配置环境变量:

profile ect/profile

source 刷新

#### 配置ip

ip配置文件:

etc/sysconfig/network-scripts/ifcfg-ens33

#### 配置免密登录

ssh-keygen

HDFS文件系统🈶三部分组成：与文件管理有关的软件、被管理文件以及实施文件管理…

**“块”**处理文件

### 数据块

#### NameNode

#### DataNode

访问流程

1. 首先，用户的应用程序通过HDFS的客户端程序将文件名发送至NameNode.
2. NameNode接收到文件名后，在HDFS目录中检索文件名对应的数据块，再根据数据块信息找到保存数据的DataNode地址，将这些地址送会给客户端。
3. 客户端收到DataNodel地址后，与DataNode并行地进行数据传输操作，同时将操作结果..

### HADOOP IO

1,什么事数据完整性

检查校验和

1. 掌握基本的基于文件的数据结构

SequenceFile读写小文件，key/value

1. 了解常用的压缩算法
2. 基本文件类型

#### SequenceFile

#### MapFile

带有索引

思考:SequenceFile与MapFile的区别。

#### 压缩

常见的压缩算法:

Gzip

Bizp2

CodeC

本地库

#### 序列化

Hadoop定义了两个序列号相关接口：

writable 和Comparable

WritableComparable 可序列号可比较接口

##### 重要的数据类型

Text---字符串

ByteWritable

NullWritable

ObjectWritable

GenericWritable

## MapReduce计算框架

二次排序

map-join

reduce-join

|  |  |  |
| --- | --- | --- |
| 函数 | 输入键值对 | 输出键值对 |
| Map（） | <k1,v1> | <k2,v2> |
| Reduce（） | <k2,{v2}> | <k3,v3> |

### 词频统计原理

Map输入

这一步由MapReduce框架自动完成，其中偏移量(即Key值)包括了回车所占的字符数，value为一行内容。

Map输入 Map输出

Hello a <0,’hello a’> <hello,1> <a,1>

Hello b <8,”hello b”> <hello,1> <b,1>

Hello c <16,”hello c”> <hello,1> <c,1>

Reduce输入

这一步由MapReduce框架自动完成，

Map输出 Reduce输入 reduce输出

<hello,1> <a,1> <a,1> <a,1>

<hello,1> <b,1> shuffle()-🡪 <b,1> ---------🡪reduce() <b,1>

<hello,1> <c,1> <c,1> <c,1>

<hello,{1，1，1}> <hello,3>

## HIVE应用

## 协同过滤算法

在电商行业的应用。

## 作业

1，大数据概念与hadoop关系

2，大数据数据分配方式

3，如何保证数据一致性，及一些原理

4，计算模式

1. 数据分析过程
2. FileSystem 增删操作