# 大数据

# Hadoop

《hadoop权威指南》

Java,linux的知识

## hafs

hadoop逻辑架构

学习大数据之前需要了解下linux

Namenode：存储文件系统元数据（文件目录结构，分块情况，每块位置，权限）存储块128M

Datanode

冗余

物理架构逻辑架构

怎样搭建一个分布式的集成环境

Xshell

HDFS集群

通过HDFS扔一个文件

通过Java程序访问HDFS，就把HDFS集群当成一个大的硬盘

Java访问HDFS

HDFS：存储

YARN：

Mapreduce：

Java访问Hadoop

namenode管理负载均衡

存储文件系统元数据（文件目录结构，分块情况，每块位置，权限）

一台namenode宕了，另一台直接补上

hadoop分为几个逻辑组件

Hdfs,存储

Namenode

Datanode

secondNameNode

journalNode

Yarn,

资源管理

Mapreduced

计算

Hdfs典型应用设计

360网盘，百度网盘

建立连接的引用

## YARN

资源调度引擎

yet another resource negation

跑各种各样的计算模型

为计算模型做资源调度：CPU的资源，内存的资源

hadoop默认hdfs,yarn ,mapreduce

hadoop自带这三个模块，耦合性不强，hadoop安装默认三个组件都有

HDFS <--> YARN <--> mapreduce

把计算程序仍到yarn去跑

World count

Hadoop jar ......wordcount /wcinput/test.txt /wcoutput/

为什么这么慢

秒杀；高并发的架构

wordCount工作流程

Input

Input--》拆分--》Map--》shuffer--》reduce--》output

Output：是一个目录，所有的结果都会汇总到目录中，所以是一堆结果

[www.sunstack.org](http://www.sunstack.org)

网络版的坦克

## MapReduce(映射、缩减)

计算引擎

HDFS：就是一个分鱼展的大硬盘

Yarn：资源调配系统，CPU和内存

MapReduce：可以用Java开发Mapreduce程序

mapReduce不是一定要依赖于yarn的

mapReduce也不用仍到集群中运行，可以在本地进行运行

Map 程序和 reduce程序需要进行自己开发

split和shuffle默认进行

泛型和序列化

跳跃式发展的思维和胆量

Mapp的输入和输出是什么

Reducer的输入和输出是什么

Java开发MapReduce

MapReduce在本地执行，数据在本地，MapReduce在本地执行，job local，调试

MapReduce在本地执行，数据在远程集群，MapReduce把数据拉到本地，mapRedece在本地执行 job local

MapReduce程序在本地windows编辑，设定好在远程运行，设定好远程的输入和输出，需要把自己打成jar文件，远程给Yarn进行分发计算

多数情况下，算法是用来干嘛的，调参数就可以了

使用maven来建立项目

手动开发mapReduce程序少，mapReduce之后出了一堆框架

真正开发的时候是Spark内存计算和Storm流式计算

mapReduce比较死，而且比较慢

Javaee是永远的需求，大数据是最新的前景

Spring springMVC Mybaitsf

Spring Boot

# HIVE

用类似SQL语言的方式做计算

将SQL语句查询，转换为MapRedece操作，省去Java编程

应用场合：静态数据分析，数据不会频繁变化，不需要实时结果响应

生产环境之中用的非常多的一种工具

从数据里面挖掘一些东西

不需要写编程的代码就可以进行大数据的运算（HiveSQL,类似sql的语言）

MR:MapReduce

Hive on MR is deprecated in Hive2 and may not be avalibale in the future version .Consider using a different execution engine (i.e. Spark,tez) or using Hive1x releases.

复杂的东西简单化

大数据工资高在于数据挖掘的地方和机器学习的地方

Hive解决95%的内容还有5%进行Java拓展

数据清洗

首先创建一个表，映射文件

Create table sgr(qtime string,qid string,qword string,url string) row format delimited fields terminated by ‘,’;

Load data inpath /sougou.dic into table sgr; //只是建立了一个映射，没有拷贝文件

在集群中的某一个集群里面装了Hive，整个结果导出到文件里面也可以导出到数据库里面，然后通过网页进行展现

大数据分析的结果，通过网页进行展现，那应该用到storm,spark,Hive,Hadoop(HDFS,Yarn,MapReduce);

学完javase直接学习大数据

javeee跑出来很多数据来源，利用大数据进行分析

大数据学习路线：

Hadoop hive hbase storm spark

# Spark

大数据开发工程师，spark是核心

不需要有hadoop基础的

内存计算引擎

使用scala开发，

支持Java，scala，python开发接口，企业中使用scala,python进行操作

scala运行在jvm机器上

运行模式

local模式

本地启动一个web视图，计算结束关闭

standalone模式，常用

yarn模式，常用

mesos模式

核心概念RDD

Resilient Distributed Datasets，是一个弹性的，分布式的数据集

数据集合：

提交命令spark submit

Spark shell

扁平化

大数据最近比较热的话题

# 大数据技术框架

# 机器学习

应用领域

语音识别

自然语言处理

自动驾驶

计算机视觉

推荐系统

无人机

垃圾邮件识别

机器人

BI商业智能：从一堆以前的数据里面挖掘出来，对商业有价值的东西，数据挖掘

经验的积累与特征值的灌输

## 基本概念

训练集（training Samples）

测试集(TestSamples)

特征值(FetureVector)

目标概念(Label)

机器学习三要素

数据，算法，模型

机器学习研究的是从数据中通过选取合适的算法，自动的归纳和逻辑或者规则，并根据这个归纳的结果（模型）与新数据来进行预测

机器学习最难的部分

把现实中的问题提炼成一个机器学习的问题，需要的是对问题本身的深刻洞察，有一天也许整个数据清洗到模型选择和交叉验证 都自动化了，但始终有一个东西不能被机器所替代，那就是如何从一个全新的领域，去提取机器学习有助于解决的最重要的问题。

分类

回归

监督学习

有监督学习

## 机器学习的基本步骤

获取数据：

获取(sqoop,)，清洗，存储(hdfs)

数据拆分训练集合合测试集合

用特征向量训练算法

在测试集上评估算法

算法迭代与改进

实际应用

获取更多的数据

在别人算法的基础上可以拿来用，避免重复造轮子

## 常用算法

监督学习

分类

KNN(最近邻)

DT(Decision Tree)

SVM(supprt vector machine)

NN(neruel network)

NB(naive bayes)

回归

LR(Linear Rearession)

NLR(Noe -)

无监督学习

K-Means