**大数据平台说明文档**

**主要应用技术组件：**

Hive Sqoop HBase HDFS Spark Kafka Zookeeper Flume Kettle

**开源框架搭建大数据平台：** cloudera-manager

(Cloudera’s Distributed Including Apache Hadoop 简称 CDH)

(Hortonworks Data Platform 简称 HDP)

准备工作：

CDH管理工具 CDH大数据平台软件包

Windows工具：Xshell、Xftp

注意事项：

datanode：存储业务数据，存储空间足够大，且每个datanode的存储空间尽量保持一致

namenode：对存储空间要求不高，存储计算节点的元数据信息

spark：内存要求较高，单节点2T数据64G内存

service iptables stop 所有节点关闭防火墙 chkconfig ip6tables off

**Flume分布式海量日志采集系统：**

ApacheFlume是一个高可靠性、高可用性的分布式海量日志收集聚合传输系统

传输的基本单位是event，文本文件的一行记录

flume运行的核心是agent，是一个完整的数据收集工具

含有三个核心组件source、channel、sink

**Zookeeper分布式协调服务：**

一个分布式开源应用程序协调服务，基于zookeeper实现同步服务、配置维护和命名服务。

最终一致性、可靠性、实时性、等待无关、原子性、顺序性

核心：原子广播 Zab协议（恢复模型、广播模式）机制保证各个server之间同步

**Kafka分布式消息发布订阅系统：**

由LinkedIn开发的一个分布式消息系统，使用Scala语言编写

解耦、冗余、扩展性、灵活性、可恢复性、顺序保证、缓冲、异步通信

Producer：消息生产者 向broker发消息的客户端

Consumer：消息消费者 向broker取消息的客户端

Topic：一个消息队列主题

Pritition：每个topic会被划分成多个partition，每条消息在文件中的位置称为offset(偏移量)

Messsge：消息是kafka处理的对象 消息是按topic存储的

Broker：一台kafka服务器就被认为是一个broker

每个partition保证消息顺序，但不保证总体消息顺序

防止出现阻塞

**Spark流计算集群平台：**

SparkCore RDD：Resident可恢复、Distributed分布式、Datasets数据集

Spark的RDD操作：transformation转化、action行动

核心组件：SparkSQL操作结构化数据、SparkStreaming实时数据流计算、MLlib机器学习库

集群管理器：standalone独立模式、Hadoop Yarn、Apache Mesos

driver驱动器：提交作业发起集群操作 executor执行器：执行任务运行task

spark-submit提交应用：

./bin/spark-submit - -master spark://localhost:7077 - -deploy-mode cluster - -class com.weifang.SparkExample - -name “Example” - -total-exexutor-cores 300 - -executor-memory 10g SparkExample.jar (独立集群模式)

**Redis内存数据库**

**Key-value**

**HDFS分布式文件存储系统：**

高容错性、数据批处理、Block块作为基本存储单元

安装部署：

若干台虚拟机集群，一个master节点，剩下的作为slave节点

scp命令ftp协议发送文件到集群中各个节点 或者克隆虚拟机

配置所有节点hosts文件 vim /etc/hosts

安装jdk 配置jvm环境 配置所有节点无密钥访问，

修改~/etc/Hadoop下的若干.xml配置文件

~/sbin下start-dfs或者stop-dfs jps查看所有节点状态是否成功启动

netstat -ano | grep 8088 查看端口占用情况

telnet 192.168.1.1 8088 ping指定ip的端口

Master主节点守护进程：

ResourceManager、SecondaryNameNode、NameNode

Slava从节点守护进程：

NodeManager、DataNode

hdfs主要命令：

hadoop dfs -put 本地 远程 上传本地文件

hadoop dfs -get 远程 本地 下载远程文件

hadoop dfs -ls 显示hdfs所有文件

hadoop dfs -ls / 显示指定目录下文件

hadoop dfs -cat / 获取指定文件里内容

hadoop dfs -mkdir /user/test 创建文件夹

**HBase分布式数据库组件：**

分布式面向列的开源数据库，高可靠性、高性能、面向列、可伸缩

**Hive数据仓库工具：**

基于Hadoop的数据仓库工具，可以将结构化的数据文件映射成数据库表，并提供简单的sql查询功能，将sql转换成简单MapReduce任务进行运行。可以用来进行数据提取转化加载（ETL），并且定义了简单的类SQL语言HQL。Hive并不适用于需要低延迟的应用，最佳使用场景是大数据集的批处理操作，比如网络日志的分析。

Hive元数据的集中存放处metastore，Hive内嵌derby数据库作为引擎

MySql作为Hive的外部存储引擎

内部表 外部表 分区表 同表

**Sqoop数据转换工具:**

主要在Hadoop（HIve）与传统数据库（mysql、oracle、postgresql）间进行数据的传递。Sqoop专为大数据批量传输设计，能够分割数据集并创建分布式任务来处理。

Sqoop支持增量更新，将新纪录添加到最近一次导出的数据源上，或者指定修改的时间戳。

Sqoop默认的并行机制在数据不均匀时容易出现节点负载不均衡现象。

主要命令：

导入数据：sqoop import -connect jdbc:mysql://hostname:port/database -username root -password 123456 -table example -m 1 hdfs默认保存路径：/user/grid/example/part-m-0

生成hive表：sqoop create-hive-table -connect jdbc:mysql://localhoust/yidong -table example -fields-terminated-by “,” 为已导入特定数据集创建hive表指定分隔符

Sqoop会将导入的数据保存为逗号分隔的文本文件

导入的文本格式数据可直接供SparkStreaming运行

Sqoop -query查询中加入where子句来限定需要导入的数据表记录

**Kettle组件：**

采用纯Java实现的开源ETL工具，属于开源商务智能软件Pentaho的一个重要组成部分。

transformation转换：完成数据ETL（Extract、Transform、Load）工作

job作业：定义一个完成整个工作流的控制

kettle组件：spoon (IDE工具)、kitchen、pan、carte菜单(http服务器)

**Apache Airflow任务调度工作流管理平台 :**

Airflow是一个Airbnb的Workflow开源项目，是一个工作流分配管理系统，通过有向非循环图的方式管理任务流程，设置任务依赖关系和时间调度。Airflow独立于要运行的任务，只需要把任务的名字和运行方式提供给Airflow作为一个task即可。

Airflow直接使用Python写DAG Definition，突破了文本文件表达能力的限制。

在分布式环境中，Airflow通过自动重启任务的方式来重置任务状态。

**大数据语言：**

Java Scala

Python：科学计算库 机器学习库 开源组件动态库

R：C

CDH不支持SparkSQL