Hadoop Training Sheet

1. 操作系统

- Linux操作系统（检查操作系统版本号）
- Linux操作系统概述
- 安装Linux操作系统

  > CentOS、Ubuntu

- Linux Shell相关工具的使用

  > Xshell、Xftp

- Linux图形界面、字符界面
- Linux基本命令
  - 查看主机名：hostname
  - 硬件配置：df、free、ethtool
  - 文件/文件夹操作：cd、mkdir、rm、cp、mv、touch、du、lsof
  - 查看文本：cat、less、tail、vim、vi
  - 查找类：grep、find
  - 压缩解压缩：tar、gzip、zip、unzip
  - 软件安装类：rpm、yum、apt-get
  - 帮助类：man
  - 时间类：date
  - IO类：lostat
  - 权限类：sudo、chown、chmod、chgrp
  - 端口连接类：netstat、ping、telnet
  - 启停服务类：etc/init.d/mysqld [start|stop|restart]

    - 网页类：elinks http://192.168.1.210:60010
    - 挂载类：mount、umount
- 用户、组群和权限管理
- 文件系统管理
- 软件包管理与系统备份
- Linux网络配置
- Linux基本服务配置

  > DNS服务、DHCP服务、HTTP服务、FTP服务、SMTP服务、POP3服务

- Linux Shell命令
    - 文件及文本常用命令：tar、find、cut、wc、split、grep、head、tail、sed、awk
    - 系统运行状况命令：top、watch、free、mpstat、vmstat、lostat、pidstat、df、du
    - 系统运行进程命令：ps、nice、renice、lsof、pgrep、pkill
    - 追踪命令：strace
    - 排序命令：sort
    - 删除重复行：uniq
    - 正则表达式
- Linux Shell脚本编写
    - 定时备份系统日志
    - 自动监控其它主机硬盘及内存使用状况
    - 自动化安装JDK、Tomcat

2. 数据库

- 关系型数据库原理
- 在Linux上安装Mysql、SQL-Server、DB2、Oracle数据库
- DDL、DML、DCL语法

  > Mysql、SQL-Server、Oracle、DB2

- SQL基础
    - 基本语句：insert、select、where、update、delete、join、group by、having、desc、asc、limit、isnull、等
    - 函数：日期函数、嵌套函数、字符串函数、数字函数、聚集函数
- SQL高级
    - PL/SQL、if、case、loop、while、for、游标、视图、索引、存储过程、事务、SQL编程
- 数据库管理
    - 容量规划
    - 安全
    - 性能
    - 对象
    - 存储管理
    - 变化管理
    - 任务调度
    - 网络管理
    - 故障排查
    - 管理工具
        - Mysql：Workbench、Navicat
        - SQL-Server：SSMSE
        - Oracle：OEM、PL/SQL developer、Toad
        - DB2：DB2top、Toad for db2
- 备份与恢复
    - 文件：参数文件、控制文件、数据文件、转储文件、重做日志
    - 备份：冷备份、热备份
    - 还原和恢复：备份控制文件、归档日志文件
- 数据库优化
    - 表：建立分区表、重建索引

    - I/O：将数据、日志、索引放在不同I/O设备
    - 切分：横/纵向分割表
    - 硬件：升级CPU、内存、网络带宽等
    - 业务分离：OLTP与OLAP分离
    - 字段选取：where后的查询字段避免冗余

3. 大数据

一、原生**Hadoop**

- Hadoop框架
    - 大数据概念及应用场景
    - Hadoop介绍
    - Hadoop组件
- HDFS组件
    - HDFS读取过程
    - HDFS基本命令：cat、chgrp、chmod、chown、cp、df、du、find、get、ls、lsr、mkdir、mv、put、rm、rmdir、rmr、tail、stat等
- Hive组件
    - Hive表结构与数据存储
    - Hive与RDBMS区别
    - Hive数据库与表
    - 基本HiveQL语法
    - 向Hive装载数据
- Sqoop组件
    - Sqoop工作原理
    - Sqoop数据流转
- Flume组件

- Flume工作原理
- Flume参数配置
- 实时将系统日志文件导入HDFS

- HBase组件
  - HBase概念及应用场景
  - HBase与RDBMS联系与区别
  - HBase表结构与数据存储

二、TDH发行版本

安装前准备

- 操作系统版本 CentOS 6.3-6.5/REHL 6.3-6.5/ Suse SP2-SP3/操作系统是否干净?
- 是否需要配置sudo用户安装TDH?
- 机器硬件配置 CPU/MEM是否满足要求? / 系统根分区大于300G?/千兆以上网络?
- 是否配置了SSD?
- 是否操作系统双盘RAID1，数据盘RAID0?
- 配置是否对称同构

  （1）磁盘同构： 数据盘对应的每块磁盘是否一样大? （严禁大小磁盘混合做数据盘，例如300G /mnt/disk1, 2.7T /mnt/disk2）

  （2）网络同构： 每台机器网卡配置是否相同?

  （3）CPU/内存大小是否同构:
- 系统时间是否正确。 > date -s '2015-11-11 09:45:00'
- 确认网络解析是用/etc/hosts文件还是DNS server。

  （1）推荐使用hosts文件解析。

  （2）若用hosts文件解析，确保/etc/resolv.conf 为空或隐掉。并保证/etc/nsswitch.conf 中 files 解析在DNS解析之前

  （3）各节点尽可能的在一个网段
- hostname只能是以字母和数字的组合(中间允许'-')，不能有"," / "." / "_"等特殊字符。

TDH安装与运维

- 安装
  - root安装、非root安装
  - 配置RACK（机柜命名一定要以'/' 开头，如 /default）
  - 添加节点、添加硬盘、升级Licence
- 配置检查
  - Zookeeper的重要配置 Zookeeper 配置个数是否检查？（奇数个，10个节点以下3个，10-50个节点5个）
  - HDFS的重要配置 HDFS 的1 个目录配置是否只包含 /mnt/disk*的数据盘，SSD是否排除在外?
  - YARN的重要配置

    （1）YARN 的2个目录配置是否只包含 /mnt/disk*的数据盘，SSD是否排除在外?

    （2）YARN 的 vcore/Mem配置是否配置成了1个core对应2G内存?
  - Inceptor的重要配置 （1）Inceptor 是否配置了HiveServer2 （推荐 Kerberos+LDAP HA模式）

    （2）Inceptor 的 fastdisk 是否配置了SSD?

    （3）Inceptor 的localdir 配置里是否只包含 /mnt/disk* ，SSD是否排除在外?

    （4）Inceptor 的资源配置是否合理？ 每个core是否都分配了1.5-2G内存?
  - Hyperbase的重要配置
  - Hmaster个数是否为奇数？（3个或者5个）
  - Fair Schedule配置
- 日志相关
  - Zookeeper的日志位置（/var/log/zookeeper1）
  - HDFS的日志位置（/var/log/hdfs1）
  - YARN的日志位置（/var/log/yarn1）
  - Hyperbase的日志位置（/var/log/hyperbase1）
  - Inceptor的日志位置（/var/log/inceptor1）

- 服务启停
  - 查看机器已启动的服务
  - 各服务启停的顺序
  - Zookeeper的启停
  - HDFS的启停
  - Hyperbase的启停
  - YARN的启停
  - Inceptor的启停
- 管理页面
  - HDFS/YARN/Hyperbase/Inceptor重要的管理界面
  - HDFS健康状态的检查
  - YARN状态的检查
  - Hyperbase状态的检查
  - Inceptor运行状态的检查
- 安全相关
  - 开启Kerberos
  - 添加/删除用户
- HDFS状态检查
  - 查看HDFS状态
  - 查看损坏文件
  - fsimage和editlog存放的位置
- Inceptor操作
  - Hive、Inceptor默认的分隔符
  - 创建外表以及数据存放位置
  - 创建ORC格式表及数据存放位置

    - 创建Transaction ORC表及数据存放位置
    - 创建Hyperbase外表及数据存放位置
- 数据迁移
    - Sqoop应用场景
    - Sqoop工具的使用
- 常用BI工具对接
    - Tableau对接
    - JDBC程序对接
    - SQuirreL对接
- Hyperbase操作
    - 全局索引、local索引、全文索引的概念与区别
    - localmode、clustermode的区别
    - Hyperbase.reader=true的含义
    - 4040页面上的表征
    - 怎样查看Hyperbase相关状态
    - 一个RegionServer最多host多少个Region?
    - 哪些情况会导致数据写入热点?
    - ObjectStore 程序怎么写？Json支持程度?
    - Batchinsert是什么？语法怎么写
- Bulkload相关
    - Bulkload的意义和本质
    - 各个阶段的目的
    - Bulkload有几个要点?
    - 什么是SQLbulkload?
    - SQLBulkload操作（TPCDS一张表的数据）

- TPC-DS相关
  - 什么是TPC-DS?
  - TPC-DS中有多少个SQL?
  - 怎样运行TPC-DS?
  - TPC-DS截图