hw01

1. 简述Hadoop的基本原理和组成及优势

1.1 基本原理

Hadoop框架中最核心的设计就是：HDFS 和MapReduce
* HDFS是Hadoop分布式文件系统，具有高容错性、高伸缩性，允许用户基于廉价硬件部署，构建分布式存储系统，为分布式计算存储提供了底层支持
* MapReduce提供简单的API，允许用户在不了解底层细节的情况下，开发分布式开行程序，利用大规模集群资源，解决传统单机无法解决的大数据处理问题
* 设计思想起源于Google GFS、MapReduce Paper

1.2 组成

- Hadoop是Apache的一个开源的分布式计算平台，以HDFS分布式文件系统和MapReduce分布式计算框架为核心，为用户提供了一套底层透明的分布式基础设施
- Hadoop框架中最核心设计就是：HDFS和MapReduce。HDFS提供了海量数据的存储,MapReduce提供了对数据的计算。

1.3 优势

- 弹性可扩展

通过简单增加集群节点，线性扩展集群存储和计算资源
* 健壮高容错

故障检测和自动恢复，允许通用硬件失效而丌影响整个集群可用性
* 成本低廉

采用廉价通用硬件部署，无需高端设备
* 简单易用

API简单，允许用户不了解底层情况下，写出高效的分布式计算应用程序

2. 简述Hadoop 1.0和2.0的区别

- Hadoop1.0即第一代Hadoop，由分布式存储系统HDFS和分布式计算框架MapReduce组成，其中HDFS由一个NameNode和多个DateNode组成，MapReduce由一个JobTracker和多个TaskTracker组成。

- Hadoop2.0即第二代Hadoop为克服Hadoop1.0中的不足：针对Hadoop1.0单NameNode制约HDFS的扩展性问题，提出HDFS Federation，它让多个NameNode分管不同的目录进而实现访问隔离和横向扩展，同时彻底解决了NameNode单点故障问题；针对Hadoop1.0中的MapReduce在扩展性和多框架支持等方面的不足，它将JobTracker中的资源管理和作业控制分开，分别由ResourceManager（负责所有应用程序的资源分配）和ApplicationMaster（负责管理一个应用程序）实现，即引入了资源管理框架Yarn。同时Yarn作为Hadoop2.0中的资源管理系统，它是一个通用的资源管理模块，可为各类应用程序进行资源管理和调度，不仅限于MapReduce一种框架，也可以为其他框架使用，如Tez、Spark、Storm等

3. 简述Hive的基本原理和特点

3.1 基本原理

- Hive是基于Hadoop的一个数据仓库工具
- 可以将结构化的数据文件映射为一张数据库表，并提供简单的类SQL(HQL)查询功能，可以将HQL语句转换为MapReduce任务进行运行
- 学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计
- 适合数据仓库的ETL和统计分析

3.2 特点

- 简单易用

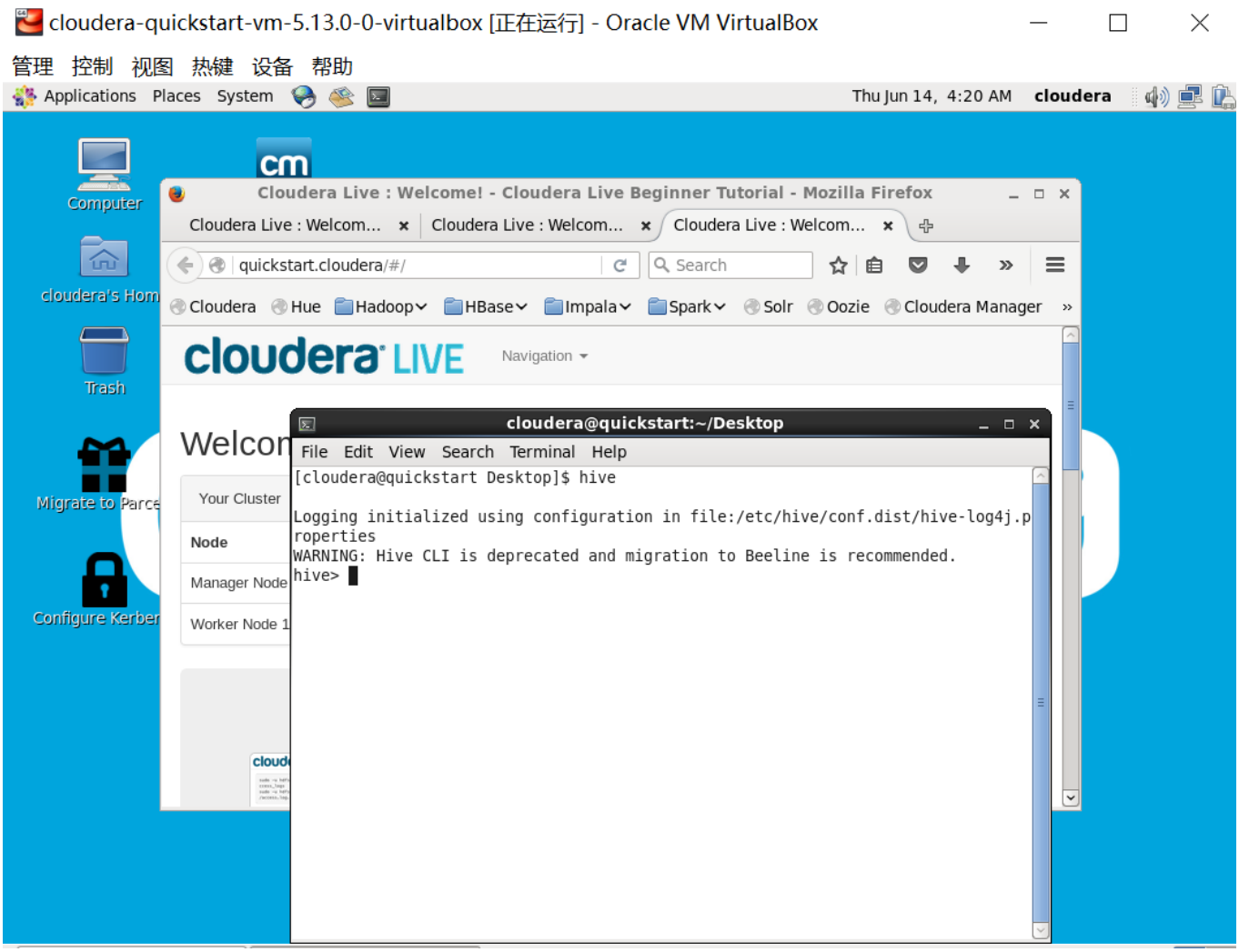
基于SQL表达式语法，兼容大部分SQL-92语义和部分SQL-2003扩展语义
* 可扩展

Hive基于Hadoop实现，可以自由的扩展集群的规模，一般情况下不需要重启服务
* 延展性

Hive支持用户自定义函数，用户可以根据自己的需求来实现自己的函数
* 容错性

Hadoop良好的容错性，节点出现问题SQL仍可完成执行

4. 能够独立部署好CDH的单机实验环境（开机后截图即可）


cloudera-quickstart-vm-5.13.0-0-virtualbox [正在运行] - Oracle VM VirtualBox
管理 控制 视图 热键 设备 帮助
Applications Places System
Thu Jun 14, 4:20 AM cloudera
Computer
cloudera's Hom
Trash
Migrate to Parce
Configure Kerber
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Live : Welcom...
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcom
Your Cluster
Node
Manager Node
Worker Node 1
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>