HADOOP基础知识

- 一、什么是hadoop： 是技术框架，针对海量数据，是分布式系统+分布式计算框架，不是数据库，HBASE才是数据库，更新速度快，只能在linix系统安装
- 二、Hadoop来源于谷歌的三大论文
  - （1）Google File System ： 这是一个可扩展的分布式文件系统，从根本上说：文件被分割成很多块，使用冗余的方式储存在商用机器集群上，应用点为存储海量网页信息
  - （2）MAP-REDUCE：描述了大数据的分布式计算方式，主要思想是将任务分解然后在多台处理能力较弱的计算节点中同时 处理，然后将结果合并从而完成大数据处理。应用点为计算网页排名；
  - （3）BIGTABLE：一个列式存储数据库，构建在GFS和MAP-REDUCE之上，不支持联结和SQL类查询。应用点为存储海量网页信息
- 三、hadoop的起源
  - 1、hadoop之父Doug Cutting 辞职回家研发，雅虎支持,到12年后开始大推，根据谷歌的三大论文做出了HADOOP
  - 2、GFS 对应 HDFS 、 MAPREDUCE 对应 MAP-REDUCE、BIGTABLE 对应 HBASE
  - 3、Hadoop思想：集群处理，一般服务器放在偏远地区
- 四、hadoop框架：由很多组件组成，数据采集--数据储存、--资源管理、协调服务等
- 五、HADOOP组件介绍
  - （1）数据采集部分
    - 1) Flume :实时数据收集工具，用于非结构化数据收集，常用于收集日志信息
    - 2) kafka：实时数据收集工具，高吞吐量的发布式订阅消息系统，消息队列---偏发数据开发、java语言
    - 3) Sqoop： SQL -to - Hadoop ，连接传统关系型数据库和Hadoop 的桥梁，把关系型数据库的数据导入到Hadoop系统（如HDFS）中，或者把数据从Hadoop系统里抽取并导出到关系型数据库里
  - （2）数据存储部分
    - （1）HDFS---分布式文件系
      - 1) 源于Google 的GFS论文，是一个分布式文件系统
      - 2) 特点
        - 良好的扩展性，因为是分布式系统，多了就减硬盘、少了就加硬盘
        - 高容错型，分布式系统，一部分硬盘或机器挂了，也不会影响其他硬盘
        - 适合PB级以上海量数据的存储
      - 3) 基本原理
        - 将文件切分成等大的数据块，存储到多台机器上

- 将数据切分、容错、负载均衡等功能透明化
- 可将HDFS看成一个容量大、具有高容错性的磁盘

4）应用场景

- 海量数据的可靠性存储
- 数据归档

- （2）HBASE---是一种Hadoop中的列式数据库，不支持表连接、不支持SQL查询，列式

（3）计算引擎部分：引擎工具

- 1） MAP-REDUCE --源于Goggle 的MAPREDUCE论文，是一个分布式计算框架

2）特点

- 良好的扩展性
- 高容错性
- 适合PB级以上海量数据的离线处理

3） MapReduce（分布式计算框架）运行原理

- ① 输入拆分：输入到mapreduce工作被划分成固定大小的块叫做input splits，输入拆分是由单个映射消费输入块；
- ② 映射-mapping ： 这是在map-reduce程序执行的第一个阶段，在这个阶段中的每个分隔数据被传递给映射函数来产生输出值
- ③ 重排-shuffling：类似于分组，这个阶段消耗映射阶段的输出，它的任务是合并映射阶段输出的相关记录
- ④ reducing ： 在这一阶段，从重排阶段输出值汇总。这个阶段结合来自重排阶段值，并返回一个输出值，总之，这一阶段汇总了完整的数据集

（4）数据处理部分：HIVE（数据仓库工具）

- 1） HIVE （基于MR的数据仓库）编译工具：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，并提供类SQL查询功能

2） HIVE相关

- ① 由facebook 开源，最初用于解决海量结构化的日志数据统计问题：ETL工具
- ② 构建在hadoop之上的数据仓库：数据计算使用MR，数据存储使用HDFS;
- ③ HIVE定义了一种类SQL查询语言---HQL，类似SQL，但不完全相同
- ④ 通常用于进行离线数据处理（采用mapreduce）：可认为是一个HQL---MR的语言翻译器

（5）资源管理和协调部分

- 1） YARN --是一个资源调度平台，负责为运算程序提供服务器运算资源，相当于一个分布式的操作系统平台，而mapreduce等运算程序则相当于运行于操作系统之上的应用程序
- 2） ZOOKEEPER---是一个分布式的，开放源码的分布式应用程序协调服务

- 六、版本介绍
  - 1、目前企业用的多的是2015年的2.7版本，最新的是2016年的3.0版本
  - 2、市面上的Hadoop版本
    - Apache Hadoop
      -社区开源版本，集群不是很大时可以使用
    - CDH（Cloudera Distributed Hadoop）
      -是Cloudera 发行的一个收费版本
    - HDP（Hortonworks Data Platform）
      是Hortonworks 发行的一个收费版本