大数据学习之路（转）

转载

simao310 于 2020-04-29 09:58:15 发布 59 收藏
原文链接：https://www.cnblogs.com/xing901022/p/6195422.html
版权
转自

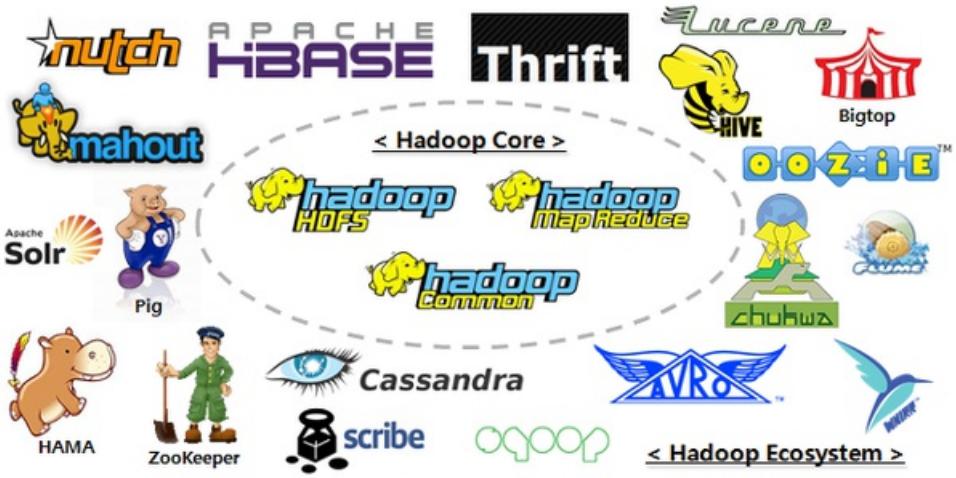
作者：xingoo

出处：http://www.cnblogs.com/xing901022

目前正在结合机器学习理论学习MLlib源码

前言

在学习大数据之前，先要了解他解决了什么问题，能给我们带来什么价值。一方面，以前IT行业发展没有那么快，系统的应用也不完善，数据库足够支撑业务系统。但是随着行业的发展，系统运行的时间越来越长，搜集到的数据也越来越多，传统的数据库已经不能支撑全量数据的存储工作；另一方面，数据越来越多，单机的计算已经成为瓶颈。因此，基于分布式的大数据系统崭露头角。那么大数据系统里面都有什么东西呢？可以参考下面的图



在存储上，hdfs的分布式存储可以任意水平扩展，可以解决数据存储的难题。在计算上，从最初的MapReduce，把任务水平拆分，多台机器并行计算，再汇总结果；到基于Spark的内存计算，改造Mapreduce每次数据落盘以及编程方式的痛点。

有了存储和计算框架，周边就衍生出了很多管理、缓存相关的技术，比如：

- yarn解决多租户资源调度的难题，
- flume解决数据传输的难题，
- sqoop解决分布式存储数据与传统DB数据之间的转换，
- oozie解决了大数据计算任务的调度，
- kafka提供了发布订阅机制的消息队列，
- zookeeper可以帮助用户完成主备的选举，
- hive在hdfs的基础上提供了数仓的功能，
- hbase则基于hdfs实现列式数据库....

上面都是hadoop生态的，由于hadoop中计算模型普遍是mapreduce，但是它的编程风格和计算机制让很多人使用不便。因此后来spark逐渐代替了mapr成为主流的计算框架。Spark也有它自己的生态，但是由于hadoop更多更早的被应用到企业，所以spark也可以无缝的集成hadoop生态中的产品。spark更多只是扮演一个计算的框架，在这个框架上，提供了基本的计算模块core，基于sql的计算引擎spark sql，对接实时数据的流式计算spark streaming，算法相关的mlib以及图计算相关的graphx。

这些框架都在这个大数据生态中扮演了自己重要的角色，他们协同工作就可以帮助我们解决很多难题。由于我也是接触不久，所以就按照自己学习和工作涉及的内容，在下面按照各个章节进行介绍，后续也会持续的更新。希望对所有对大数据感兴趣的

学习必备

在学习大数据的过程中，需要具备的能力或者知识，在这里简单的罗列一下：

- 语言基础：需要会使用shell脚本、java和scala(这俩语言主要是用于日常代码和阅读源代码)
- 工具：IDE如eclipse或者idea，虚拟机和secureCRT连接工具
- 书籍：《Hadoop权威指南》《Hadoop YARN权威指南》《Spark快速大数据分析》《从Paxos到zookeeper分布式一致性原理与实践》《Hive编程指南》其他的书籍阅读后再推荐吧
- 博客：董的博客
- 进阶：阅读官方文档（帮你了解它都能做什么）、源代码（帮你了解它是怎么做的）

hdfs

hdfs是大数据系统的基础，它提供了基本的存储功能，由于底层数据的分布式存储，上层任务也可以利用数据的本地性进行分布式计算。hdfs思想上很简单，就是namenode负责数据存储位置的记录，datanode负责数据的存储。使用者client会先访问namenode询问数据存在哪，然后去datanode存储；写流程也基本类似，会先在namenode上询问写到哪，然后把数据存储到对应的datanode上。所以namenode作为整个系统的灵魂，一旦它挂掉了，整个系统也就无法使用了。在运维中，针对namenode的高可用变得十分关键。

- 2016-07-28 单节点部署Hadoop教程
- 2016-07-28 Hadoop HDFS 用户指南

mapreduce

hive

hive基于hdfs构建了数据仓库系统，它以hdfs作为存储，依赖于数据库(嵌入式的数据库derby或者独立的数据mysql或oracle)存储表schema信息，并完成基于sql自动解析创建mapreduce任务(由于mapreduce计算效率比较差，目前官方推荐的是底层计算模型采用tez或者spark)。所以hive可以理解为：hdfs原始存储+DB Schema信息存储+SQL解析引擎+底层计算框架组成的数据仓库。

官方文档

- 2016-08-13 Hive初识
- 2016-08-16 Hive部署入门教程
- 2016-08-23 《Hive编程指南》—— 读后总结
- 2016-08-23 Hive数据的导入导出
- 2016-08-24 Hive连接JOIN用例详解
- 2016-08-30 循序渐进，了解Hive是什么！
- 2016-08-31 手把手教你搭建Hive Web环境

spark

spark是现在大数据中应用最多的计算模型，它与java8的stream编程有相同的风格。封装了很多的计算方法和模型，以延迟执行的方式，在真正需要执行的时候才进行运算。既可以有效的做计算过程的容错，也可以改善我们的编程模型。

官方文档

- 2016-08-05 《Spark大数据处理》—— 读后总结
- 2016-09-03 《Spark快速大数据分析》—— 第三章 RDD编程
- 2016-09-05 《Spark快速大数据分析》—— 第五章 数据读取和保存
- 2016-09-06 《Spark快速大数据分析》—— 第六章 Spark编程进阶
- 2016-09-13 《Spark快速大数据分析》—— 第七章 在集群上运行Spark
- 2016-09-21 [大数据之Spark]——快速入门
- 2016-10-09 [大数据之Spark]——Transformations转换入门经典实例
- 2016-10-10 [大数据之Spark]——Actions算子操作入门实例
- 2017-02-18 Spark源码分析之Spark Shell（上）
- 2017-02-19 Spark源码分析之Spark Shell（下）
- 2017-02-21 Spark源码分析之Spark-submit和Spark-class
- 2017-02-23 Spark SQL 用户自定义函数UDF、用户自定义聚合函数UDAF 教程
- 2017-02-26 基于Spark UI性能优化与调试——初级篇
- 2017-04-06 Spark Stage切分 源码剖析——DAGScheduler
- 2017-04-16 Spark源码分析之分区器的作用
- 2018-01-10 Spark源码分析 之 Driver和Excutor是怎么跑起来的?(2.2.0版本)
- 2018-01-19 Spark Client启动原理探索
- 2018-06-02 Structured Streaming教程(1) —— 基本概念与使用
- 2018-06-04 Structured Streaming教程(2) —— 常用输入与输出
- 2018-06-05 Structured Streaming教程(3) —— 与Kafka的集成
- 2018-07-05 Spark MLlib 之 特征处理StringIndexer、IndexToString使用说明以及源码剖析
- 2018-07-07 Spark MLlib 之 Vector向量深入浅出
- 2018-07-09 Spark MLlib 之 aggregate和treeAggregate从原理到应用
- 2018-07-11 Spark MLlib 之 大规模数据集的相似度计算原理探索

oozie

oozie提供了大数据场景下各种任务的调度，比如shell脚本、spark任务、mapreduce任务、sqoop任务、hive查询以及普通的java程序等等。它的编译是生态圈里面最复杂的，由于以来的各个版本不同，需要指定特定的版本，因此没有成型的一键部署包。

官方文档

- 2016-09-22 oozie快速入门
- 2016-11-17 Oozie分布式任务的工作流——邮件篇
- 2016-11-19 Oozie分布式任务的工作流——脚本篇
- 2016-11-21 Oozie调度报错——ORA-00918：未明确定义列
- 2016-11-22 Oozie分布式任务的工作流——Sqoop篇
- 2016-12-11 大数据之Oozie——源码分析（一）程序入口
- 2016-12-23 Oozie分布式任务的工作流——Spark篇
- 2017-02-28 图文并茂 —— 基于Oozie调度Sqoop
- 2017-03-01 Oozie分布式工作流——流控制
- 2017-03-02 Oozie分布式工作流——Action节点
- 2017-03-04 Oozie分布式工作流——从理论和实践分析使用节点间的参数传递
- 2017-03-07 Oozie分布式工作流——EL表达式

sqoop

sqoop支持基于sql或者表名把数据库中的数据存储到分布式环境中，数据库支持oracle\mysql等等，分布式环境可以是hdfs,hive,hbase等等，数据的导入时双向的，比如你可以把oracle中的数据读取存储到hdfs，也可以把hdfs的数据导入到oracle.

官方文档

- 2016-09-12 sqoop初探?
- 2016-09-29 什么是sqoop?
- 2016-11-23 sqoop切分任务原理

hbase

HBase是基于Hdfs之上的列式数据库，基于文件分割以及rowkey的顺序存储，能快速索引查询数据。我这边是在推荐系统中，作为推荐结果存储引擎，不过由于内容比较碎片化，Hbase写入时间比较随意，因此总会出现大量超时现象，还在持续优化中。

推荐学习资料：

1. 封神总结的HBase全网最佳学习资料汇总
2. 分布式(hadoop)内核研发面试指南
3. 封神微博
4. 阿里封神谈hadoop生态学习之路

个人总结：

- 2017-06-09 Hbase常用命令
- 2017-07-03 Hbase多版本的读写（Shell&Java API版）
- 2017-11-24 HBase跨地区机房的压测小程序——从开发到打包部署
- 2018-02-28 Spark DataFrame写入HBase的常用方式
- 2018-07-12 HBase官方文档 之 Region的相关知识

yarn

在企业中，大数据的基础平台往往是多个用户共用的，那么如何管理资源的分配，就需要yarn来处理了。Yarn默认提供了三种资源分配的策略：

- FIFO：先进先出，即按照用户提交任务的时间分配资源
- Capacity：按照队列设置队列的大小
- Fair Share：也是基于队列，只不过资源的粒度更小。

常见可以用于分配的资源可以是节点的数量，内存的大小，也可以是CPU核数。

官方文档

- 2016-12-06 yarn资源调度浅学
- 2016-12-13 大数据之Yarn——Capacity调度器概念以及配置

zookeeper

从名字来说他是动物园的管理员，实际上他是各个组件的协调者。可以实现类似主从选举、分布式事务、负载均衡等多种功能，比如HDFS HA方案、HBase的Metastore、Kafka里面的offset维护等等，由此可以见，zookeeper的重要性。

不过激发我学习zookeeper的主要原因还是因为它里面涉及了很多分布式协议的东西，从而能更好的理解分布式中的一些概念。所以，就跟着我一起深入浅出的学习吧！

主要参考：官方文档 《从Paxos到zookeeper分布式一致性原理与实践》

- 2018-03-27 Zookeeper学习笔记——1 单机版本环境搭建
- 2018-04-08 分布式理论——从ACID到CAP再到BASE
- 2018-04-09 Zookeeper学习笔记——2 Shell和Java API的使用
- 2018-04-10 跟着ZooKeeper学Java——CountDownLatch和Join的使用

最后

上面是我学习hadoop和spark的分享，更重要的是学习历程的记录，希望有兴趣学习大数据的朋友可以通过我之前的学习路线获得一些思考和借鉴。后续也会逐步的完善，等到对整体有了比较全面的了解后，会专门针对安装部署、使用实践、原理解析进行介绍。