Enterprise Hadoop Solutions

新浪微博 @Mr-Robot1992

作者: whoami

现有几大Hadoop平台比较

Hortonworks HDP 2

- 特点

纯开源
原装Hadoop(做了rpm包)
版本跟进快
Hive+tez(主推)
企业级安全(不开源)
监控集成方案：
ganglia+nagios
Spark合作(原装)
贡献Hadoop源代码为主？
权限控制目前最好 ？ 高度可视化？

- 缺点

升级(坑,各种版本升级方式还不同?)
配置管理?
添加删除节点?
维护升级成本？

Cloudera CDH 5

- 特点

CM小白式安装
监控,使用可视化程度高
CM修改配置方便
小白式升级
资源分配设置可视化
文档非常详细
Impala(主推),维护成本低
Spark合作(功能阉割)
修复Hadoop缺陷为主(bug)
权限控制不够完善,可视化程度一般？

- 缺点

CM不开源,企业级(Navigator付费)
版本跟进相对保守
社区版本不透明,
功能有些许小限制？

目录

> HDP Cluster Install

- http://www.itweet.cn/2015/08/31/hdp-install/
- http://docs.hortonworks.com/HDPDocuments/Ambari-2.2.0.0/bk_Installing_HDP_AMB/content/_ambari_repositories.html

› HDFS / MapReduce / Yarn / Hive / Impala / Oozie/sentry/zookeeper/Hue/sqoop2

- 进阶应用
- 集群资源分配和调优

> Hive Tpcds

- TPC-Hive bulid package
- 生成测试数据

> Hive/Impala SQL的强大功能

- Hive2.0/impala-kudu,逐渐向传统RDBMS挑战。

> Hive vs Impala

-几个Hints

- Hive的Map Join
- Impala的SHUFFLE Join（ partitioned join）

> Presto vs sparksql vs Drill vs impala vs tez vs other

- 各种SQL-on-Hadoop的痛
- 各种限制,各种坑,稳定性成谜？各种的营销噱头,没有解决实际问题？

Hadoop平台介绍

Hadoop虽然在技术上已经得到验证、认可甚至已经到了成熟期。其中最能代表Hadoop发展轨迹的莫过于商业公司推出的Hadoop发行版了。自从2008年Cloudera成为第一个Hadoop商业化公司，并在2009年推出第一个Hadoop发行版后，很多大公司也加入了做Hadoop产品化的行列。

“发行版”这个词是开源文化特有的符号，看起来任何一个公司只要将开源代码打个包，再多多少少加个佐料就能有一个“发行版”，然而背后是对海量生态系统组件的价值筛选、兼容和集成保证以及支撑服务。

Cloudera提出了Hybrid Open Source的架构：核心组件名称叫CDH（Cloudera's Distribution including Apache Hadoop），开源免费并与Apache社区同步，用户无限制使用，保证Hadoop基本功能持续可用，不会被厂家绑定；数据治理和系统管理组件闭源且需要商业许可，支持客户可以更好更方便的使用Hadoop技术，如部署安全策略等。Cloudera也在商业组件部分提供在企业生产环境中运行Hadoop所必需的运维功能，而这些功能并不被开源社区所覆盖，如无宕机滚动升级、异步灾备等。

Hortonworks采用了100%完全开源策略，产品名称为HDP（Hortonworks Data Platform）。所有软件产品开源，用户免费使用，Hortonworks提供商业的技术支持服务。与CDH相比，管理软件使用开源Ambari，数据治理使用Atlas，安全组件使用Ranger而非Sentry，SQL继续紧抱Hive大腿。

MapR采用了传统软件厂商的模式，使用私有化的实现。用户购买软件许可后才能使用。其OLAP产品主推Drill，又不排斥Impala。

现在主流的公有云如AWS、Azure等都已经在原有提供虚拟机的IaaS服务之外，提供基于Hadoop的PaaS云计算服务。未来这块市场的发展将超过私有Hadoop部署。

思考

-- 大数据数据仓库选型中，集群性能指标的主要基准测试标准是？

-- 看具体的应用场景
-做报表，没有太多的adhoc查询，ok ？
-有多用户同时访问，多用户同时提交查询的话，hadoop表现就不那么好了，延迟严重，并发性没看到很好的解决方案出现？
-一个是计算，一个是数据转换，hadoop计算处理后放入rdbms or NOSQL(newsql) ？

我的2问题：
-- 是否需要讲解更多？有关错误以及解决过程？目前我把最佳实践告诉各位？
你在使用集群1-2年后才能慢慢遇到问题，才知道集群规划，参数该怎么设置
该踩的坑一个不会少?

-- 基础，百度谷歌就能解决的，就别直接来问我，在群里讨论即可？一些概念性的东西,下来多去看看资料？

泪流满面，不是说了0基础学习大数据技术吗？我们这个课程太高级？NO,NO,NO！

xxx: 请教下, linux磁盘空间不够了？怎么办？
xxx: 请教下，hive怎么实现update,delete,通过程序能实现吗？
xxx: 请教下，我这里xxx.rpm包安装不了？Xxx： 你好，我按照你的视频安装一直无法成功？怎么办？
xxx: ……

今天说多了点：问题的时候请专业一些？

Thank you

提问时间？

Blog： http://www.itweet.cn

PPT： https://github.com/itweet/course

百度云 链接：http://pan.baidu.com/s/1dDnpT6p 密码：rg03