数据平台的组成与扩展

[http://mp.weixin.qq.com/s?\_\_biz=MjM5ODIzNDQ3Mw==&mid=208452262&idx=1&sn=44a3ba983ac85d489d3dd0134d6bfda9&scene=27#wechat\_redirect](http://mp.weixin.qq.com/s?__biz=MjM5ODIzNDQ3Mw==&mid=208452262&idx=1&sn=44a3ba983ac85d489d3dd0134d6bfda9&scene=27%23wechat_redirect)

大数据存取的选择：行存储还是列存储？

<http://www.infoq.com/cn/articles/bigdata-store-choose>

spark在企业中的一些应用，文章中间有介绍。

[https://mp.weixin.qq.com/s?\_\_biz=MzAwNzIzMDY5OA==&mid=2651424952&idx=1&sn=3c529ff32f5c91cb55b4a136337594ba&chksm=80fc6d3db78be42b7b7c502e45563e4180778c91f062158f13343e49adb2d92ce74e50bd4d15&scene=0#rd](https://mp.weixin.qq.com/s?__biz=MzAwNzIzMDY5OA==&mid=2651424952&idx=1&sn=3c529ff32f5c91cb55b4a136337594ba&chksm=80fc6d3db78be42b7b7c502e45563e4180778c91f062158f13343e49adb2d92ce74e50bd4d15&scene=0%23rd)

spark sql在饿了么的应用实践

[https://mp.weixin.qq.com/s?\_\_biz=MzAxNDUwMzU3Mw%3D%3D&mid=2649026034&idx=1&sn=0baf5529981f77088edd9fa79eb039eb&scene=45#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzAxNDUwMzU3Mw%3D%3D&mid=2649026034&idx=1&sn=0baf5529981f77088edd9fa79eb039eb&scene=45%23wechat_redirect)

文章中提及到了spark2.1.1有较多坑爹的地方，建议使用spark2.2.0及以后的版本，可以发现spark2.2.0已经是spark划时代的一个版本，其在工业中的使用度已经很广泛，尤其是以前标记为实验性质的structStreaming已经在该版本取消了实验标记变为了生产级别。

Spark性能测试：

**spark standalone client**

3台机器 每台分配给executor的内存为1GM 每台机器8核，分配的cpu cores为48使用spark standalone client(spark 1.5.0)读取hbase表7538285条51s

**Spark yarn cluster**

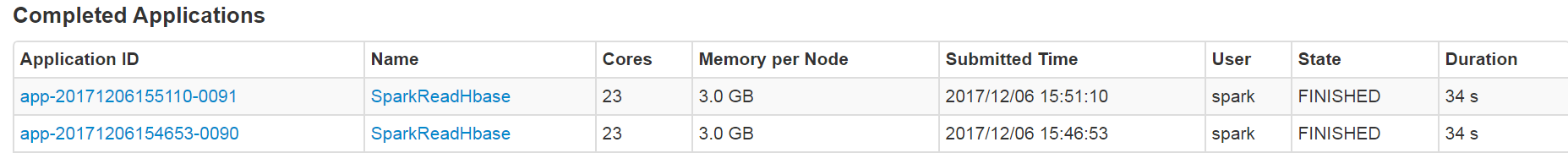
1resourcceManeger,2 nodeManager cpucoures共16个 运行时只有11个，内存共10.55G 运行时使用了9G 用时 1mins, 33sec(包含了连接mysql与写简单sql的更行语句) 共7538285条数据，hbase 5个online regions

可以知道在数据量不大时使用yarn cluster反而速度不是太快因为资源的初始化都会耗费一些时间

**Spark standalone cluster**

**1master 3 worker**

资源配置如下



读取hbase表7538285条 34s,读取百万级别的表时相较于yarn优秀

也许还可以优化，从cpu cores 内存方面可以优化到几秒甚至ms级别

约1G的数据量的表，cores共24，每个executor的内存为3G重分区的数量在500-600时较好

**日志收集最佳实践**

[https://mp.weixin.qq.com/s?\_\_biz=MzU1NDA4NjU2MA==&mid=2247486104&idx=1&sn=9b97fd92841bf09be417839893a2be94&chksm=fbe9b357cc9e3a4163c4818f83fcad29e2d83a4625fe63071d780aad38766fb215f0871e685b&scene=27#wechat\_redirect](https://mp.weixin.qq.com/s?__biz=MzU1NDA4NjU2MA==&mid=2247486104&idx=1&sn=9b97fd92841bf09be417839893a2be94&chksm=fbe9b357cc9e3a4163c4818f83fcad29e2d83a4625fe63071d780aad38766fb215f0871e685b&scene=27%23wechat_redirect)

**基于ELK stack和spark streaming的日志处理平台设计和实现**

<https://www.ibm.com/developerworks/cn/analytics/library/ba-1512-elkstack-logprocessing/index.html>

ELK Stack是一套开源日志处理平台解决方案