

AWS Summit

AWS 技术峰会 · 北京 2014

EMR在畅游的实践

搜狐畅游系统运维部 DBA 张文

2014-11-26



业务及团队背景

- 内部创业团队
- 产品面向海外
- 业务快速增长
- 团队大数据分析方面薄弱
- 海外业务构建在AWS

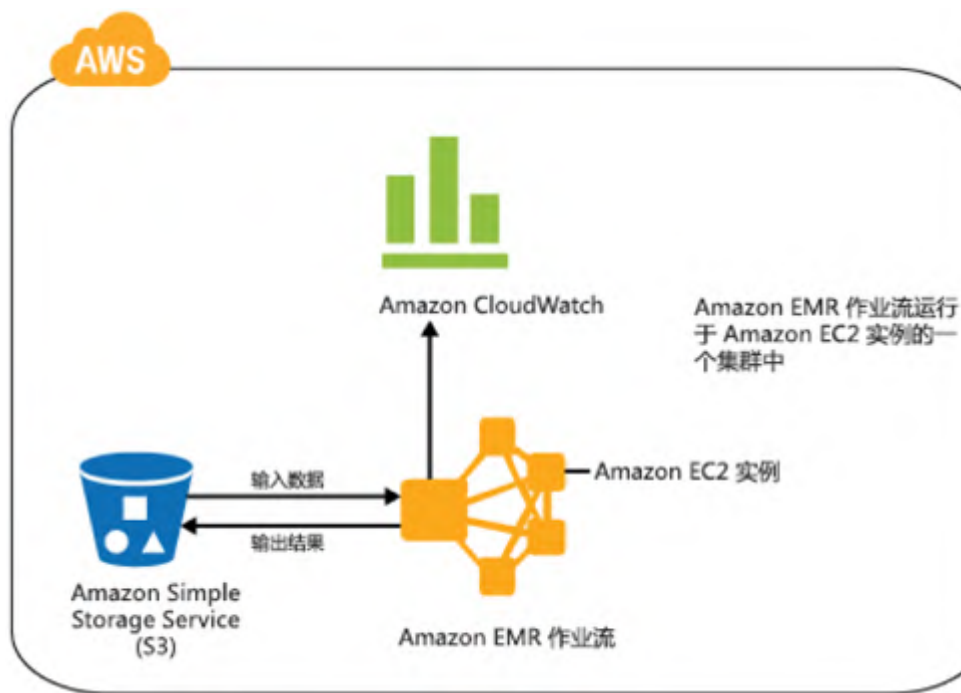
数据分析经历的三个阶段

- Mysql 阶段
- 自建hadoop集群阶段
- 逐渐转向EMR （redshift和EMR对比）

EMR简述

- Amazon Elastic MapReduce (Amazon EMR)
 - 运行在EC2实例上面的Hadoop集群
 - 可以处理或者存储（PT 级）数据，集群规模根据需要定制。
 - 简化Hadoop 及相关大数据应用程序的运行。
 - 启动便捷快速

EMR架构



业务架构



EMR的特性

- 可调整集群大小 {可动态扩展的集群}
- 仅按实际用量付费
- 易于使用
- 使用Amazon S3 或者HDFS
- 平行的集群
- Hadoop 应用程序支持 {Cascading hbase hive hue impala spark }
- 商业智能工具 {tableau talend microstrategy}

Redshift的特点

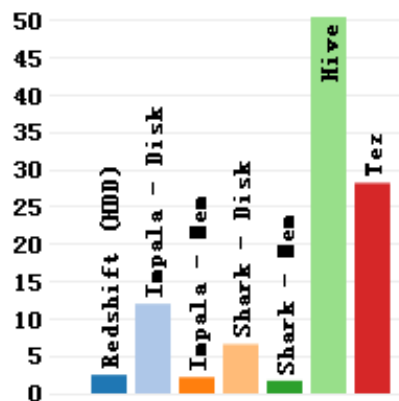
- 使用简单 （服务级别）
- 数据加载效率高
- 查询效率高
- 海量数据分析的趋势，在线实时分析

不足之处

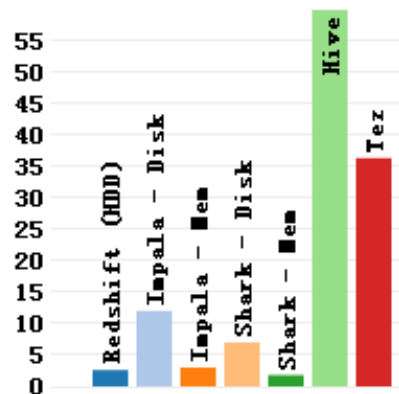
- 灵活性不够，不能跑一些推荐算法
- 无法灵活扩展，无法应对峰值需求

Redshift 和EMR性能测试对比

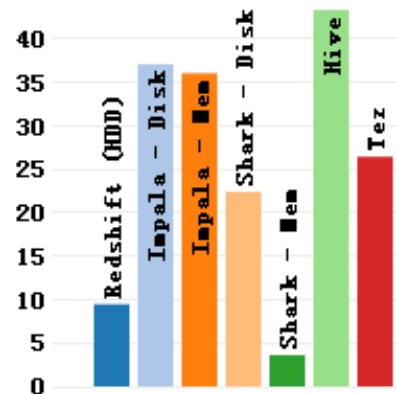
Query 1A
32,888 results



Query 1B
3,331,851 results



Query 1C
89,974,976 results



EMR应用的业务场景

- 已经在用的业务场景

- 推荐系统
- 统计分析系统

- 后期

- 搜索系统
- 爬虫的数据处理
- 广告推荐系统

EMR的集群类型

- Hive 集群 目前使用比较多
- 自定义JAR (map&reduce) 集群 推荐系统正在使用
- 流式集群
- Pig 集群
- HBase 集群

EMR的数据安全性

- 启用异常保护
- 性能和数据安全
- HDFS与S3

EMR的高可用

- 元数据用单独的数据库存储
- MapReduce程序，日常统计程序，定期同步备份到S3
- 关键数据备份到S3，原始数据存储在S3
- 如果不在乎MONEY，MR也不错

- 故障恢复
 - 万一集群出现问题，该怎么办？
 - 重建整个集群，大概30min——1h左右

EMR自动扩展

- 根据负载自动扩展
- 根据任务量自动扩展
 - 自动扩展时，使用spot实例也很省钱
 - 自动扩展可以有效应对临时性的大的查询需求和周期性的大查询需求
- 手动扩展
 - 一条命令而已
 - elastic-mapreduce --jobflow jobid \
--modify-instance-group groupid --instance-count 20

EMR最佳实践建议

- 1、数据传输 s3cmd并行传输
- 2、数据压缩
- 3、数据分区
- 4、S3 和 HDFS 并存 分别放不同的数据
- 5、架构的可恢复性
- 6、根据不同的业务特点选择不同的实例类型

成本节约

➤ 合理规划

- m1.xlarge(4c*15G 4*420)
- 通过**竞价型实例**降低成本

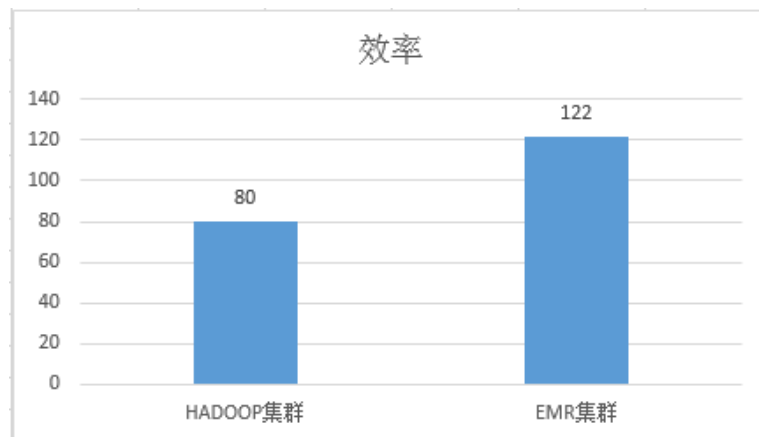
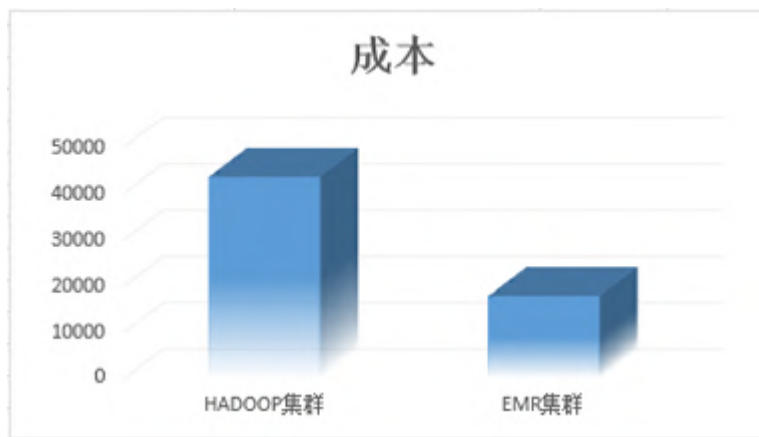
➤ 购买预留实例

- 通过预留实例节省（50%—70%）

➤ 灵活使用集群扩展

- 根据负载水平，任务数量灵活调整集群的规模，如：每天只有8个小时的高峰期，其他时间使用很少，那就在这8个小时内提高集群的规模，8小时之外缩减集群规模。

效率于成本对比



DEAMO 演示

➤ 经典案例 字数统计

- `elastic-mapreduce --create - --instance-type m1.xlarge --num-instances 3 \`
`--bootstrap-action s3://elasticmapreduce/bootstrap-actions/install-ganglia --stream \`
`--input s3://elasticmapreduce/samples/wordcount/input \`
`--output s3://dba.xxxx.com/emr/output/2014-05-09 \`
`--mapper s3://elasticmapreduce/samples/wordcount/wordSplitter.py \`
`--reducer aggregate \`
`--log-uri s3://dba.xxxx.com/emr/log/`

THANKS!

2014.12.12

AWS Summit

AWS 技术峰会 · 北京 2014

