Spark SQL最佳实践

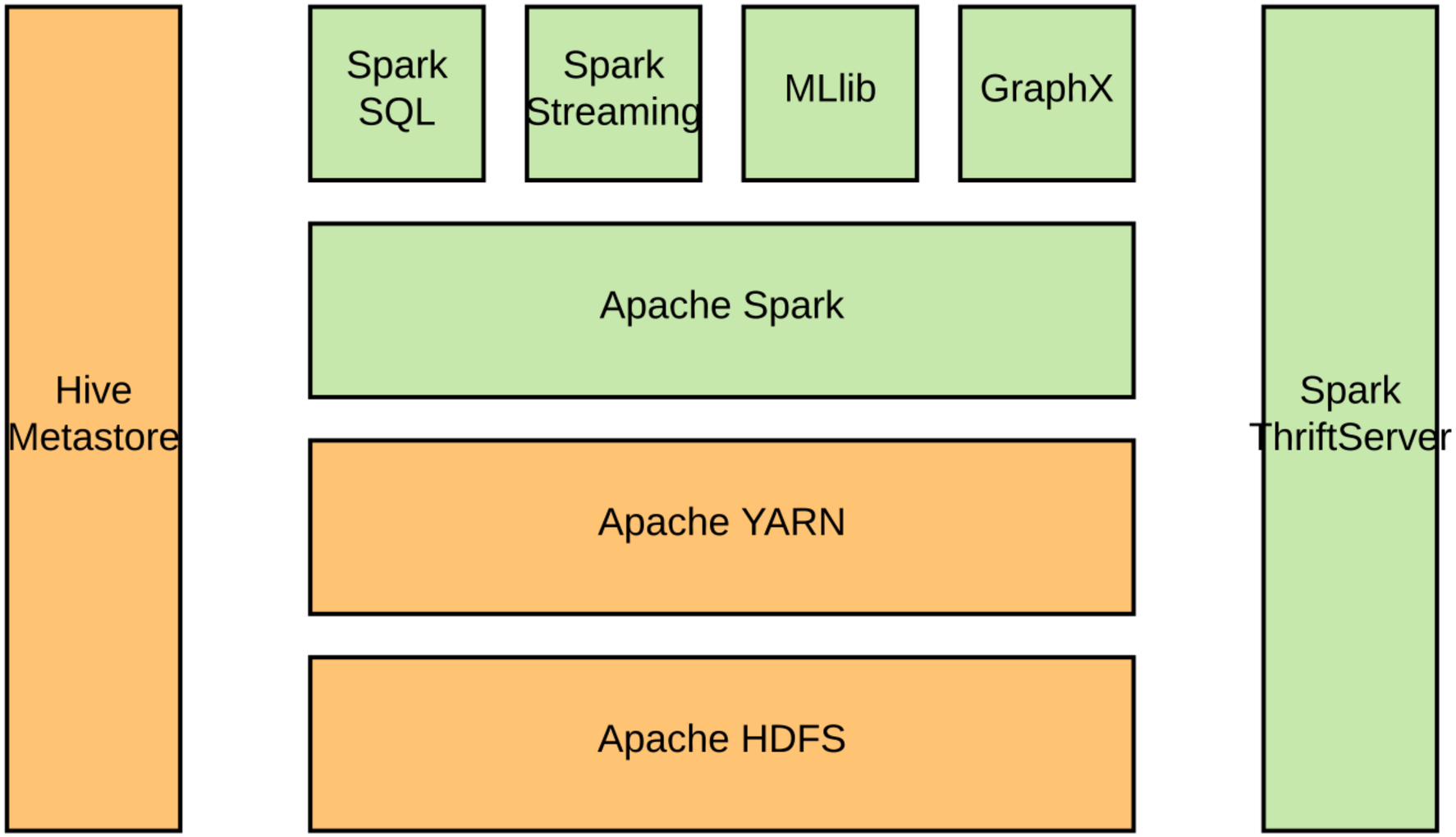
阿里巴巴计算平台事业部EMR
高级技术专家 李呈祥

目录

- Spark相关组件介绍
- 建表与ETL
- SQL Query优化

Part1: Spark相关组件介绍

Spark相关组件


Hive
Metastore
Spark
SQL
Spark
Streaming
MLlib
GraphX
Apache Spark
Apache YARN
Apache HDFS
Spark
ThriftServer

Hive Metastore

Hive元数据管理

- 库，表的基本信息，包括表名，存储类型及地址，分区信息，列等等。
- 已注册UDF相关信息。
- 用户，权限相关信息。

关注点：Metastore内存使用和RPC响应时间。

相关：分区数量

Spark ThriftServer

- Spark SQL处理
  - SQL词法解析
  - SQL语法解析
  - 逻辑执行计划生成及优化
  - 物理执行计划生成及优化

关注点：内存使用，意外崩溃。

相关：Query返回结果。

Part2: 表与ETL

表，分区与桶

- 内外表
- 数据格式
- 分区
- 桶

Tips1: 关注表的文件数量，这会影响Namenode的性能和稳定。

小文件

NameNode压力
读文件效率
数据压缩效率

Tips2: 关注表的Partition数量，可能会搞垮Hive Metastore和Spark Thrift Server。

Tips3: 数据表的列类型应该基于业务含义，CAST is bad。

- 一次写入，多次访问
- 数据存储效率
- 数据访问效率

Tips4:Join很昂贵, denormalized tables（反范式化表）可能更便宜。

范式化节省了存储空间，但存储空间却很便宜

Case Study

```
INSERT overwrite TABLE default.client_collect_cmd_p_byhour
partition(day, hour, cmd)
SELECT log_timestamp,
       t.field['header_DC'].string_type device,
       ip,
       field,
       t.thedate as day
       t.thehour as hour,
       t.field['cmd'].string_type as cmd
FROM default.client_collect t
WHERE t.thedate='2018-07-04'
AND regexp_extract(field['cmd'].string_type, '([0-9,a-z,A-Z]*)')=field['cmd'].string_type
  AND length(field['cmd'].string_type)<100
```

文件数量 =Partition基数*job并发度

Case Study

```
INSERT overwrite TABLE default.client_collect_cmd_p_byhour
partition(day, hour, cmd)
SELECT log_timestamp,
       t.field['header_DC'].string_type device,
       ip,
       field,
       t.thedate as day
       t.thehour as hour,
       t.field['cmd'].string_type as cmd
FROM default.client_collect t
WHERE t.thedate='2018-07-04'
AND regexp_extract(field['cmd'].string_type, '([0-9,a-z,A-Z]*)')=field['cmd'].string_type
  AND length(field['cmd'].string_type)<100
DISTRIBUTE BY hour, cmd;
```

文件数量 =Partition基数*job并发度

Case Study

```
INSERT overwrite TABLE default.client_collect_cmd_p_byhour
partition(day, hour, cmd)
SELECT log_timestamp,
       t.field['header_DC'].string_type device,
       ip,
       field,
       t.thedate as day
       t.thehour as hour,
       t.field['cmd'].string_type as cmd
FROM default.client_collect t
WHERE t.thedate='2018-07-04'
AND regexp_extract(field['cmd'].string_type, '([0-9,a-z,A-Z]*)')=field['cmd'].string_type
  AND length(field['cmd'].string_type)<100
DISTRIBUTE BY hour, cmd , rand(10);
```

文件数量 =Partition基数*job并发度

Case Study

```
INSERT overwrite TABLE default.client_collect_cmd_p_byhour
partition(day, hour, cmd)
SELECT log_timestamp,
       t.field['header_DC'].string_type device,
       ip,
       field,
       t.thedate as day
       t.thehour as hour,
       t.field['cmd'].string_type as cmd
FROM default.client_collect t
WHERE t.thedate='2018-07-04'
AND regexp_extract(field['cmd'].string_type, '([0-9,a-z,A-Z]*)')=field['cmd'].string_type
  AND length(field['cmd'].string_type)<100
DISTRIBUTE BY hour, cmd , rand(10);
```

表的数据目录和文件数量对于Namenode的影响，以及分区的数量对于Hive Metastore Spark Thrift Server的影响是用户在设计表和ETL语句是需要重点考虑的因素。

Part3: SQL查询的优化

Tips1: 不要在SQL使用星号

Tips2:使用Limit，除非你非常确定返回数量。

Tips3:Cross Join可能导致task数量过多，Try AE。

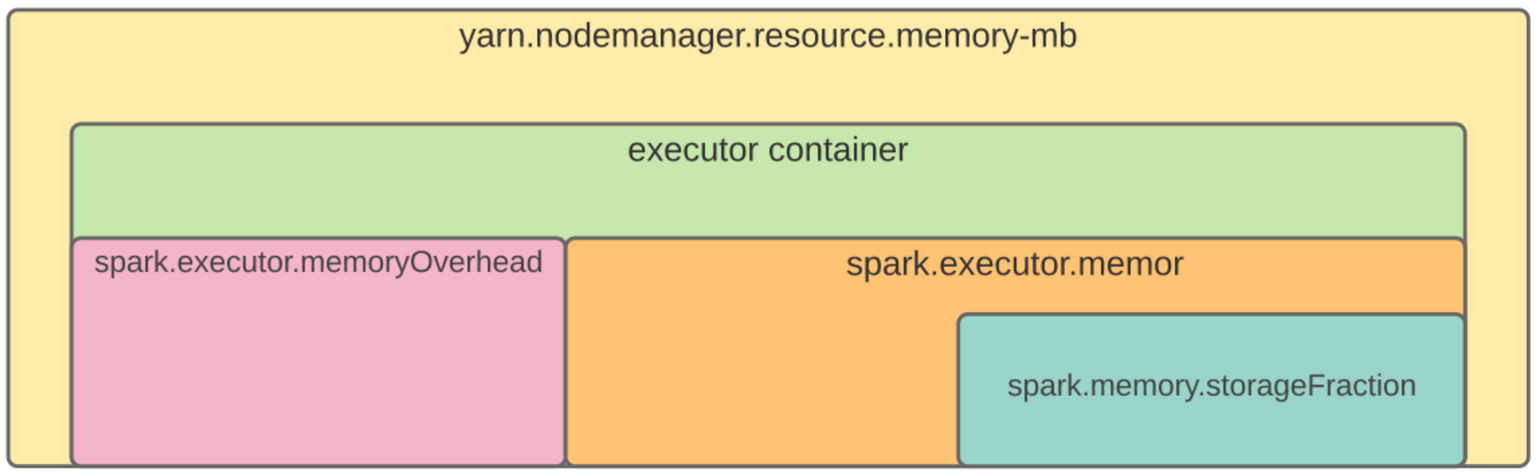
Tips4:为你的表收集Statistic信息。

CBO依赖Statistics判断最优的执行计划，确保你的表都有Statistic信息。

ANALYZE TABLE [db_name.]**table_name** COMPUTE **STATISTICS** [analyze_option]

ANALYZE TABLE [db_name.]**table_name** COMPUTE **STATISTICS FOR** COLUMNS col1 [, col2, ...]

Tips5:Container killed by YARN for exceeding memory limits



Join

- Broadcast Join，Shuffle Join
- SortMerge Join，Hash Join（spark.sql.join.preferSortMergeJoin）

- 尽量使用Broadcast Join

  spark.sql.autoBroadcastJoinThreshold, default 10M

  Adaptive Execution

- 如果无法使用Broadcast Join，尽量减少Shuffle 数据。

  Bucket Join

  Runtime Filter

Order By,Sort By,Cluster By,Distribute By

- Order By
  - 全局排序，只有1个Reduce Task用于排序。
- Sort By
  - 只保证一个Reduce内部有序。常和Distribute By一起，按Key1分组后按key2排序。
- Distribute By
  - 保证相同key的数据分发到相同的reduce task中。
- Cluster By = Distribute By + Sort By on same key

数据倾斜

- 转换成Broadcast Join
- Adaptive Execution
  - Auto Setting The Shuffle Partition Number
  - Optimizing Join Strategy at Runtime
  - Handling Skewed Join


计算，让想象力无限
Computing Gives You Indefinite Imagination
2016 杭州·云栖大会

谢谢！

Apache Spark中国技术...

1761人

扫一扫群二维码，立刻加入该群。