TalkingData

Mobile · Data · Value

Spark & Spark Streaming

2015年01月

简介

- 数据量3-4TB
- 以前
  - mapReduce
  - hive
- 现在
  - Spark & Spark Streaming & Spark SQL

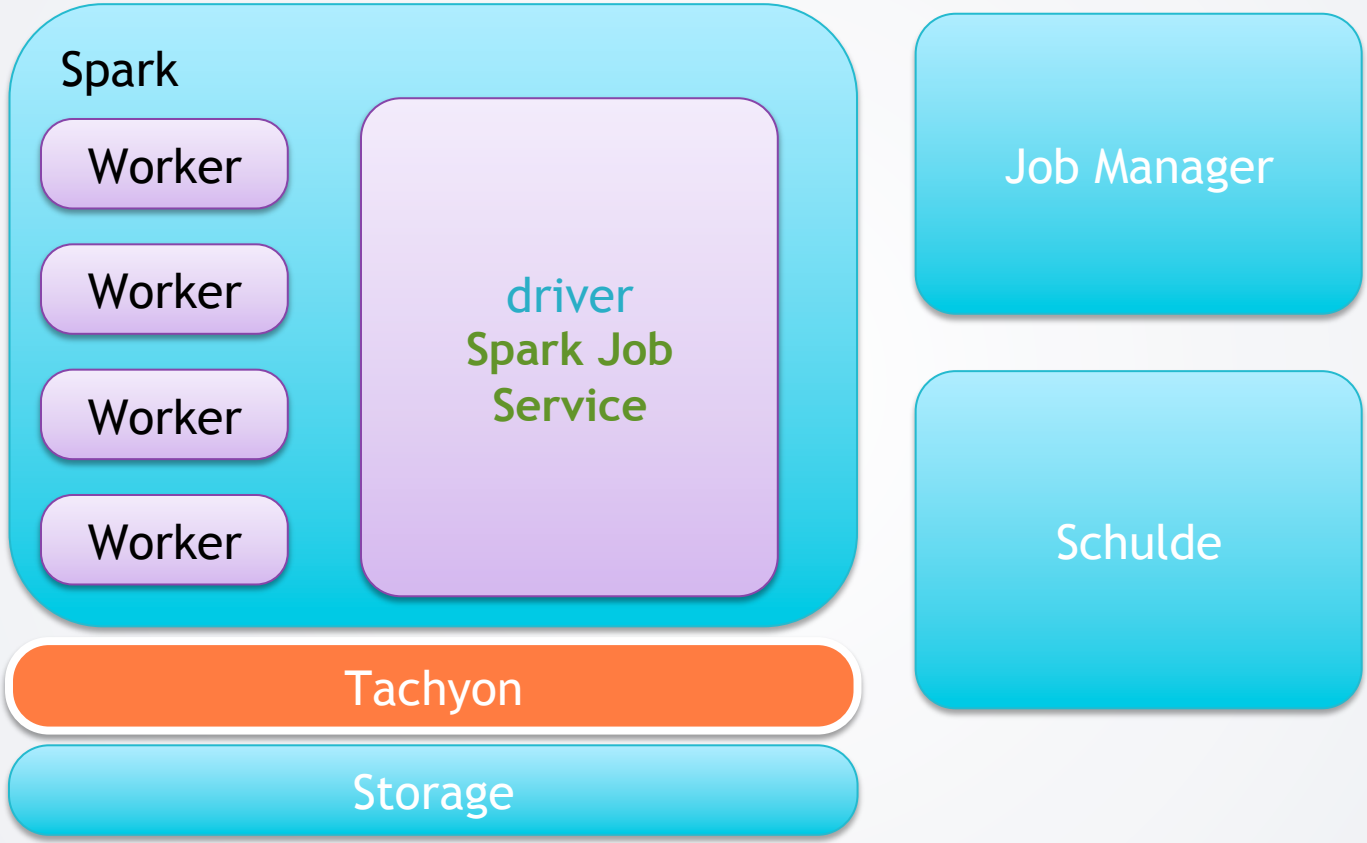
对比

	Hadoop	*Spark*
每天数据量	*1-2TB*	*3-4TB*
跑一天需要时间	*4*、*5*个小时	*1*、*2*个小时
一个月	无法忍受	*1*天
机器学习	慢	适合迭代

优化

- 每天生成汇总数据生成parquet
  - 耗时3个小时
- ad-hoc从parquet计算
  - 1天耗时 2min 左右
  - 1周耗时10min左右

优化


Spark
Worker
Worker
Worker
Worker
driver
Spark Job
Service
Job Manager
Schulde
Tachyon
Storage

Spark 问题与技巧

序列化

- 使用 kryo
- Serializable
  - 如何避免报没有序列化的错
    - 定义成 object 成员变量
    - 用函数
- 闭包
  - akka 发送变量

Timeout

- 内存溢出
- spark.akka.frameSize
- partition 太多，再shuffle

groupByKey

- reduceByKey 可以替代好多
- groupByKey 的使用

比如我们对一个key的值的序列做一个非常复杂的计算，比如求这个序列的方差、标准差，再计算3倍标准差以内的平均。

AUC

- 算法
  - roc曲线下的面积
  - 算正负样本对中，正样本score大于负样本对数
  - 对样本score排名，计算正样本rank和，再计算
- mapPartitionsWithIndex

Scala

- sc.textFile(in).collect.toSet() = ?

Scala

- sc.textFile(in).collect.toSet() = ?
- Set[String] ?

Scala

- sc.textFile(in).collect.toSet() = ?
- Set[String] ?
- Boolean

Scala

- sc.textFile(in).collect.toSet() = ?
- Set[String] ?
- Boolean
- toSet() = toSet.apply() = toSet.apply(():Unit)

Spark Streaming

祁技超