杭州第4次SPARK MEETUP

SPARK STREAMING使用和概要图

@时金魁　挖财

TEXT

AGENDA

- 1. 如何搭配
- 2. 持久化到HBase
- 3. phoenix
- 4. mesos
- 5. Streaming图

TEXT

- 吞吐：Kafka, Spark Streaming, HBase

- source: kafka, flume, mqtt, zeroMQ...
- target: HBase, hdfs, tachyon, phoenix...

HBASE的问题

- 1. 压力大（开spark和hbase背压）
- 2. 超时
- 3. 连接数过多
- 4. 过长的定长rowkey
  - "16,xxoold#16,xy_id#16,ooxxld"

TEXT

SCHEMA过多怎么选择?

HBase、 Hive、 Phoenix、 Parquet...

- 一切为了索引
- 直接写HFile
- 可选的还有elasticsearch + kibana
  - 收集和可视化metric信息
  - mesos: /metrics/snapshot, /state, /health, /master/health ...
  - marathon: /metrics,
  - hadoop

MESOS选择的理由

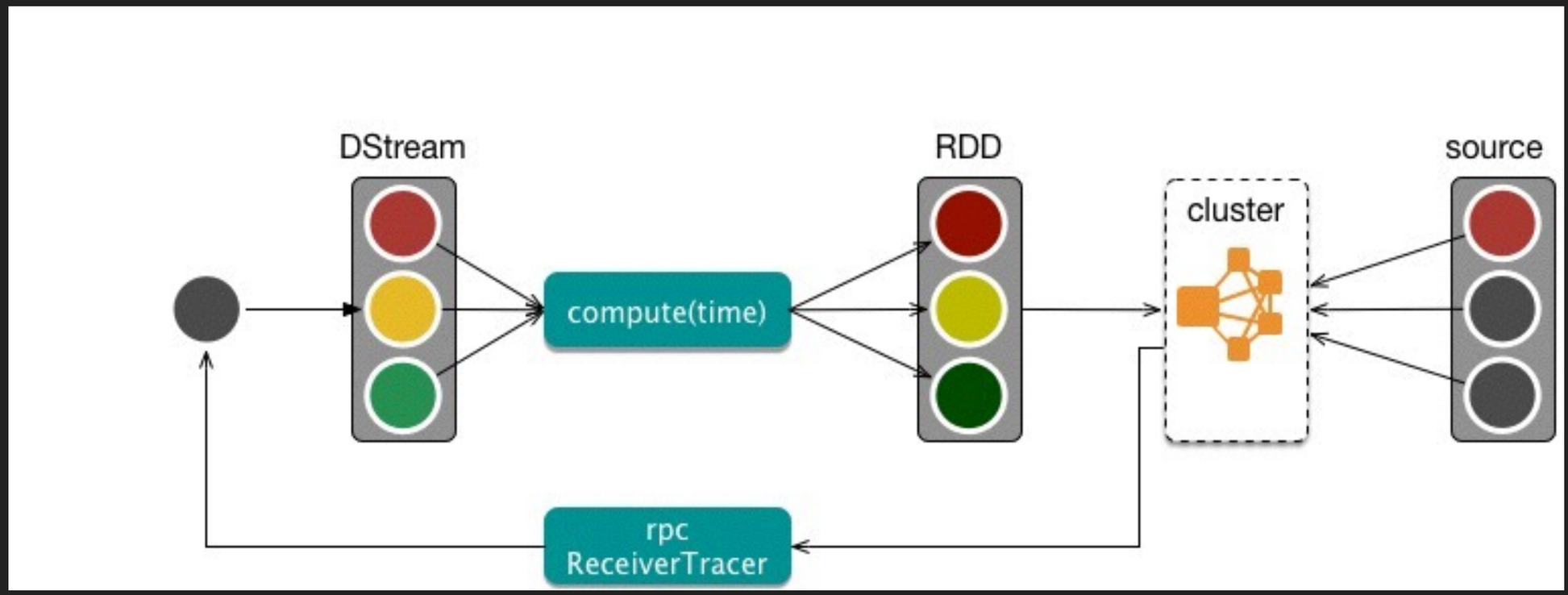
- 1.0.0-RC1
- 持久化, job状态不丢失
- 多容器支持:mesos和docker
- Restful

- marathon
- fine-grained
- chronos, dcos
- 网络隔离/服务发现
- ...

TEXT

- Spark Streaming, Phoenix, HBase
- running on: Mesos, Marathon, ES

STREAMING图


DStream
RDD
source
cluster
compute(time)
rpc
ReceiverTracer

Thanks

https://github.com/pusuo/spark_meetup