大数据管理系统与大规模数据分析 - June 10'rd, 2018

Computation Concepts

https://github.com/rh01

MapReduce/Hadoop

MapReduce 编程模型

MapReduce 的数据模型

- <key, value>
  - 数据由一条一条的记录组成
  - 记录之间是无序的
  - 每一条记录有一个 key，和一个 value
  - key: 可以不唯一
  - key 与 value 的具体类型和内部结构由程序员决定，系统基本上把它们看作黑匣
- MapReduce
  - Map(ik, iv) $\longrightarrow${<mk, mv>}
  - Reduce(mk, {mv})$\longrightarrow${<ok, ov>}
- Map 函数
  - 输入是一个 key-value 记录：<ik, iv>
    - 我们用‘i’代表 input
  - 输出是 0 ～多个 key-value 记录：<mk, mv>
    - 我们用’m’代表 intermediate
  - 注意：mk 与 ik 很可能完全不同
- Shuffle (由系统完成)
  - Shuffle = group by mk
  - 对于所有的 map 函数的输出，进行 group by
  - 将相同 mk 的所有 mv 都一起提供给 Reduce
- Reduce 函数
  - 输入是一个 mk 和与之对应的所有 mv
  - 输出是 0 多个 key-value 记录：<ok, ov>
    - 我们用’o’代表 output
  - 注意：ok 与 mk 可能不同

MapReduce vs. SQL

MapReduce	SQL Select
Map	Selection/projection
Shuffle	Group by
Reduce	Aggregation, Having
选择的功能更加丰富 程序实现的, 类似最简单的 SQL select, 但不支持 join	功能由数据类型和 SQL 语言标准定义 有 UDF, 但支持得不好

表 1: MapReduce vs. SQL Select

MapReduce/Hadoop 系统架构

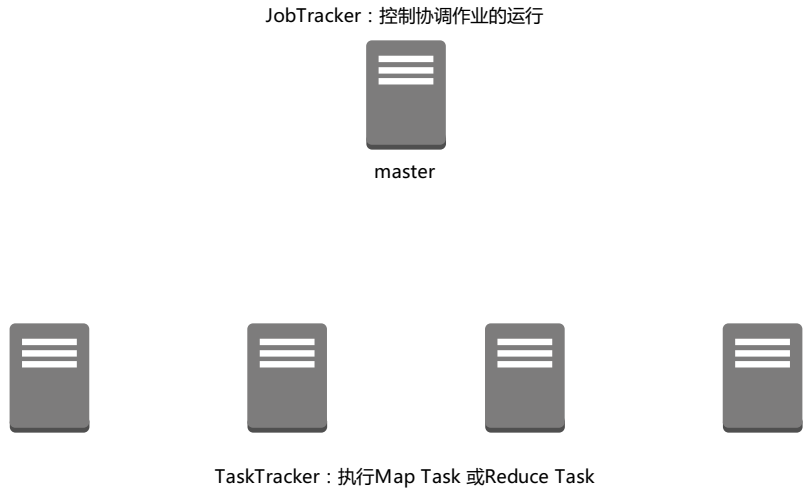

图 1：MapReduce/Hadoop 系统架构.

MR 运行流程图

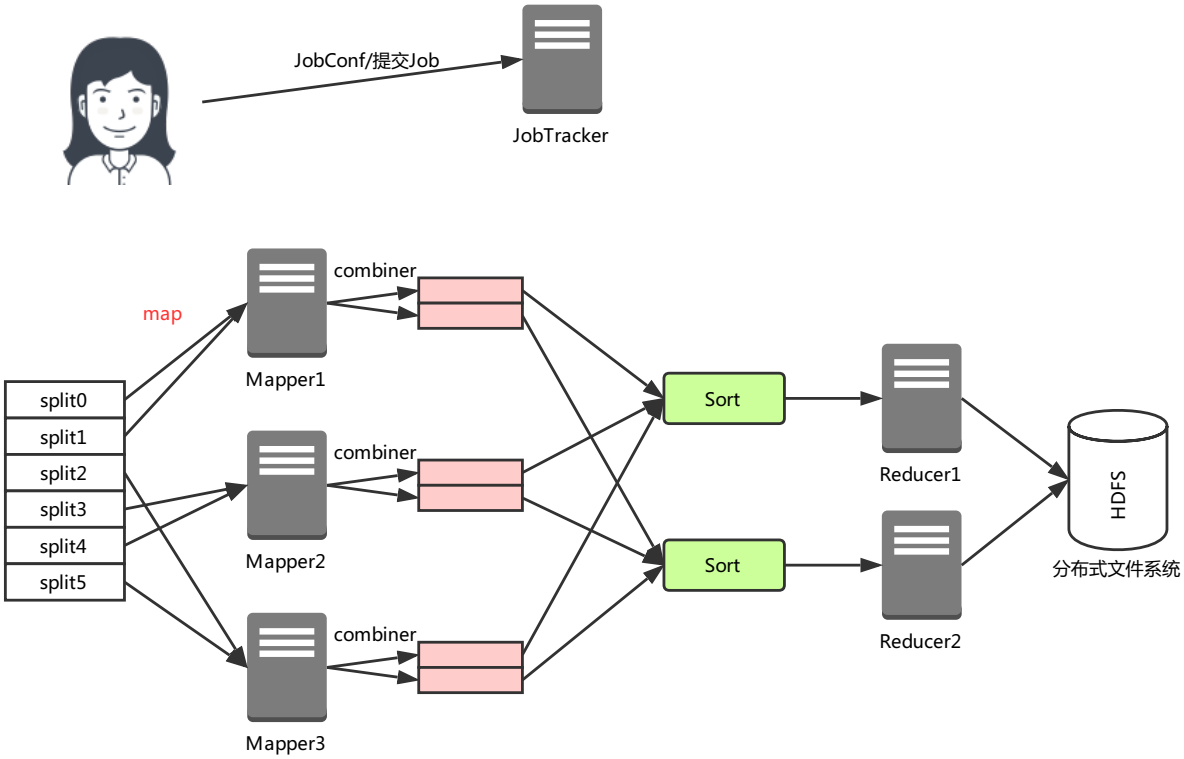

图 2：MR 运行.

MR: Fault Tolerance (容错)

- HeartBeat(心跳) 消息
  - 定期发送，向 JobTracker 汇报进度
- JobTracker 可以及时发现不响应的机器或速度非常慢的机器
  - 这些异常机器被称作 Stragglers
- 一旦发现 Straggler
  - JobTracker 就将它需要做的工作分配给另一个 worker
- Straggler 是 Mapper，将所对应的 splits 分配给其它的 Mapper
  - 输入数据是分布式文件，所以不需要特殊处理
  - 通知所有的 Reducer 这些 splits 的新对应 Mapper
  - Shuffle 时从新对应的 Mapper 传输数据
- Stragger 是 Reducer，在另一个 TaskTracker 执行这个 Reducer
  - 这个 Reducer 需要重新从 Mappers 传输数据
  - 注意：因为 Mapper 的输出是在本地文件中的，所以可以多次传输

Microsoft Dryad

- Dryad 是对 MapReduce 模型的一种扩展
  - 组成单元不仅是 Map 和 Reduce，可以是多种节点
  - 节点之间形成一个有向无环图 DAG(Directed Acyclic Graph)，以表达所需要的计算
  - 节点之间的数据传输模式更加多样
    * 可以是类似 Map/Reduce 中的 shuffle
    * 也可以是直接 1:1、1: 多、多:1 传输
  - 比 MapReduce 更加灵活，但也更复杂
    * 需要程序员规定计算的 DAG

同步图计算系统