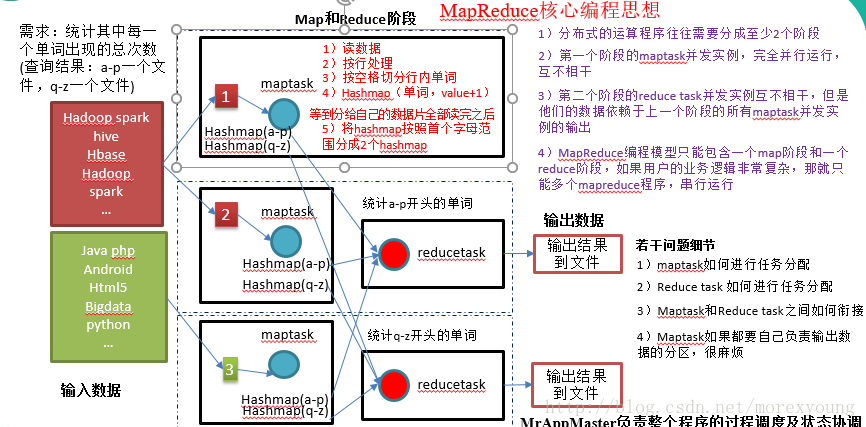
# 学习心得

1813003 李昱龙

MapReduce特性：

Mapreduce是一个分布式运算程序的编程框架，是用户开发“基于hadoop的数据分析应用”的核心框架.

Mapreduce核心功能是将用户编写的业务逻辑代码和自带默认组件整合成一个完整的分布式运算程序，并发运行在一个hadoop集群上



一个完整的mapreduce程序在分布式运行时有三类实例进程：

1）MrAppMaster：负责整个程序的过程调度及状态协调。

2）MapTask：负责map阶段的整个数据处理流程。

3）ReduceTask：负责reduce阶段的整个数据处理流程。

1.5 MapReduce编程规范（八股文）

用户编写的程序分成三个部分：Mapper，Reducer，Driver(提交运行mr程序的客户端)

1）Mapper阶段

       （1）用户自定义的Mapper要继承自己的父类

       （2）Mapper的输入数据是KV对的形式（KV的类型可自定义）

       （3）Mapper中的业务逻辑写在map()方法中

       （4）Mapper的输出数据是KV对的形式（KV的类型可自定义）

       （5）map()方法（maptask进程）对每一个<K,V>调用一次

2）Reducer阶段

       （1）用户自定义的Reducer要继承自己的父类

       （2）Reducer的输入数据类型对应Mapper的输出数据类型，也是KV

       （3）Reducer的业务逻辑写在reduce()方法中

       （4）Reducetask进程对每一组相同k的<k,v>组调用一次reduce()方法

3）Driver阶段

整个程序需要一个Drvier来进行提交，提交的是一个描述了各种必要信息的job对象