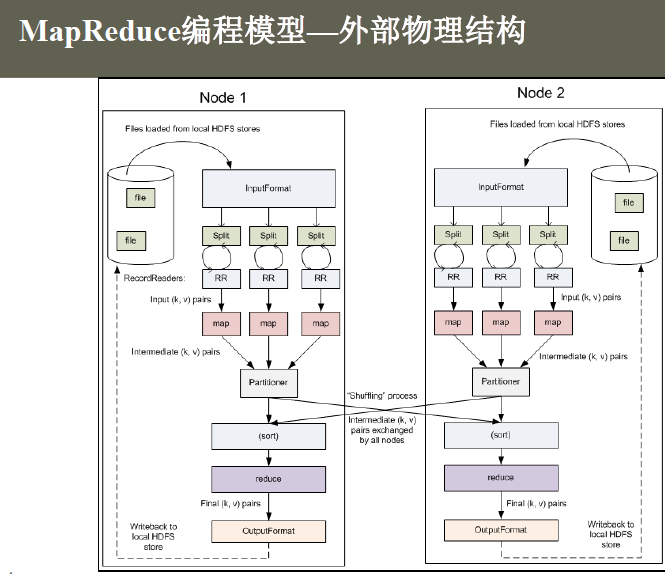
编程模型：

1. MapReduce将整个运行过程分为两个阶段：Map和Reduce阶段
2. Map阶段有一定数量的Map Task组成
   1. 输入数量数据格式解析：InputFormat
   2. 输入数据处理：Mapper
   3. 数据分组：Partitioner
3. Reduce阶段是由一定数量的Reduce Task组成
   1. 数据远程拷贝
   2. 数据按照key排列
   3. 数据处理：Reducer
   4. 数据输出格式：OutPutFormat



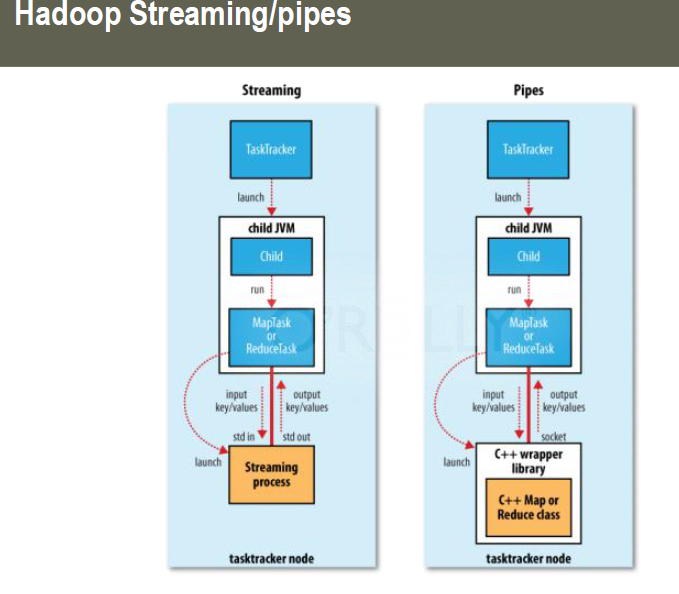
1. Map阶段：
   1. InputFormat（默认TextInputFormat）
   2. Mapper
   3. Combiner（local Reducer）
   4. Partitioner
2. Reducer
   1. Reducer
   2. OutPutFormat（TextOutPutFormat）

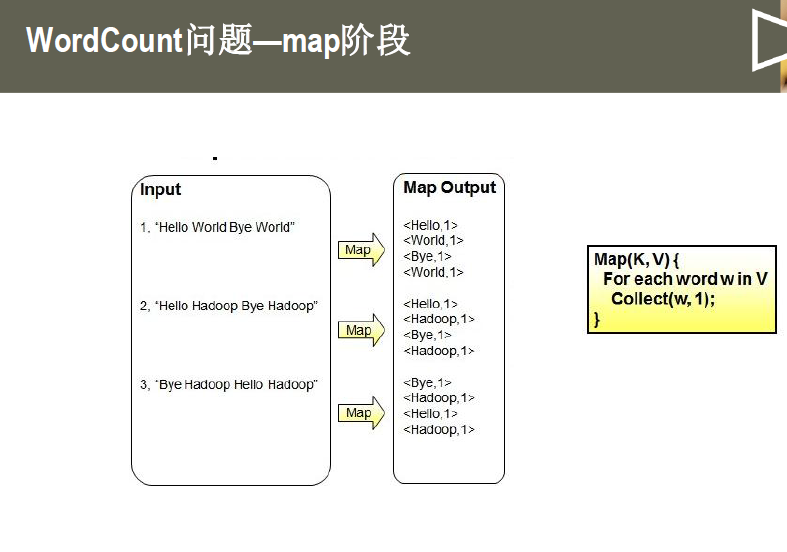
MapReduce编程接口：

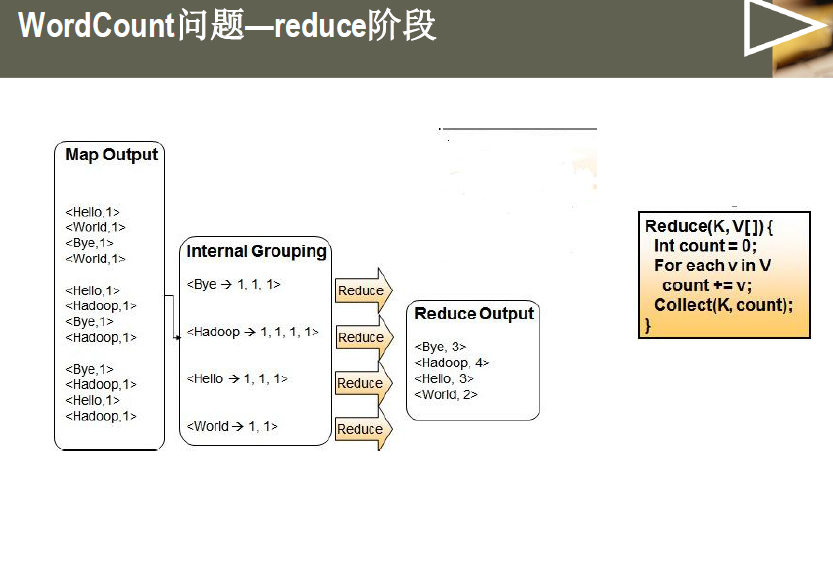
1. Hadoop提供了三种编程方式
   1. Java （原始方式）
   2. Hadoop Streaming（支持多语言）
   3. Hadoop Pipes（支持C++）
2. Java编程接口是所有方式的基础
3. 不同的编程接口知识暴露给用户的方式不同而已，内部引擎都是一样的。（最后都转化为一种）
4. 不同编程效率。（数据密集型的都会小于原生Java的效率,多一次数据转化，进程间通信）
5. Java编程接口组成：
   1. 就API：org.apache.hadoop.mapred
   2. 新API：org.apache.hadoop.mapreduce
6. 新API具有良好的扩展性；
   1. 两种编程接口只是暴露给用户的形式不同而已，内部引擎是一样的；
7. 旧API可以完全兼容Hadoop2.0，但是新API不行。

Hadoop Streaming：

1. 与Linux管道机制一致。
2. 通过标准输入输出实现进程间通信。
3. 标准输入输出是任何语言都有的。
4. 几个例子：
   1. Cat 1.txt | grep “dog” | sort
   2. Cat 1.txt | python grep.py | java sort.jar

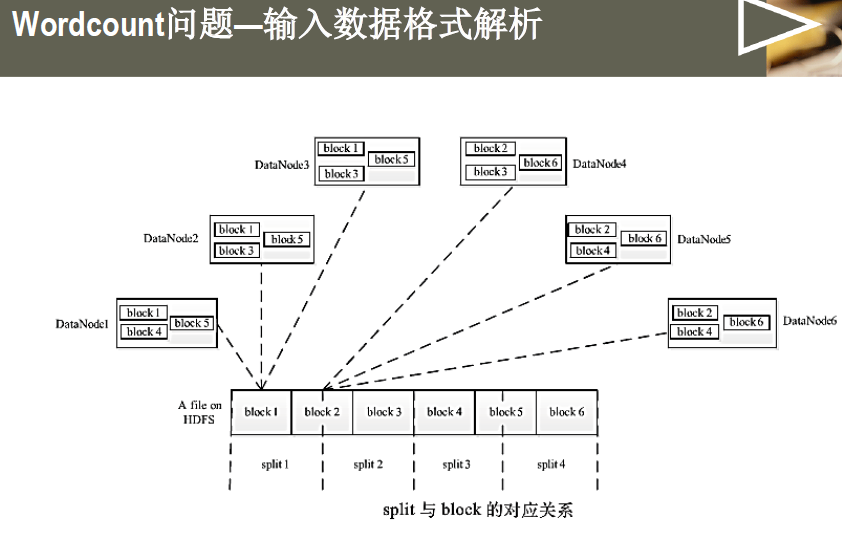






TextInputFormat：

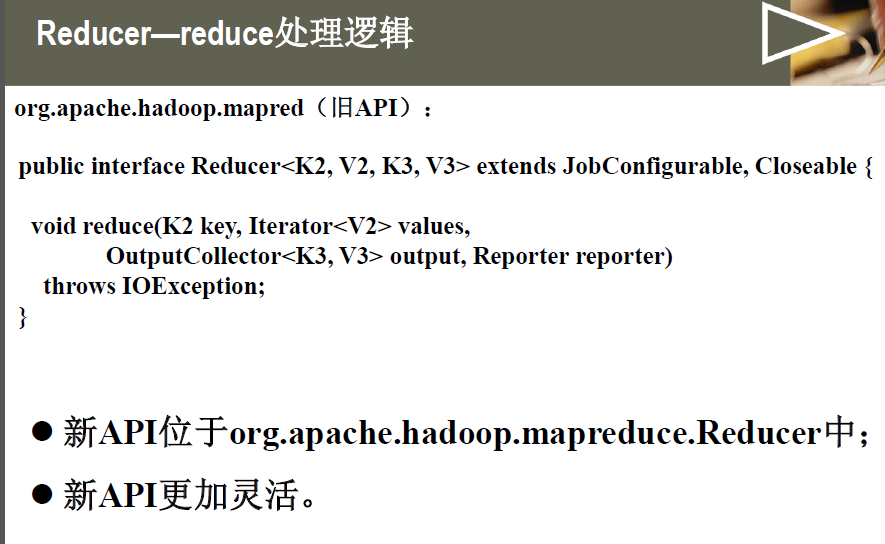
1. 每个Map Task处理一个split；
2. 每个split大小等于block；
3. 如果最后一行数据被截断，则读取后一个block前半段部分；
4. 转换成key/value对，key是偏移量，value是行内容

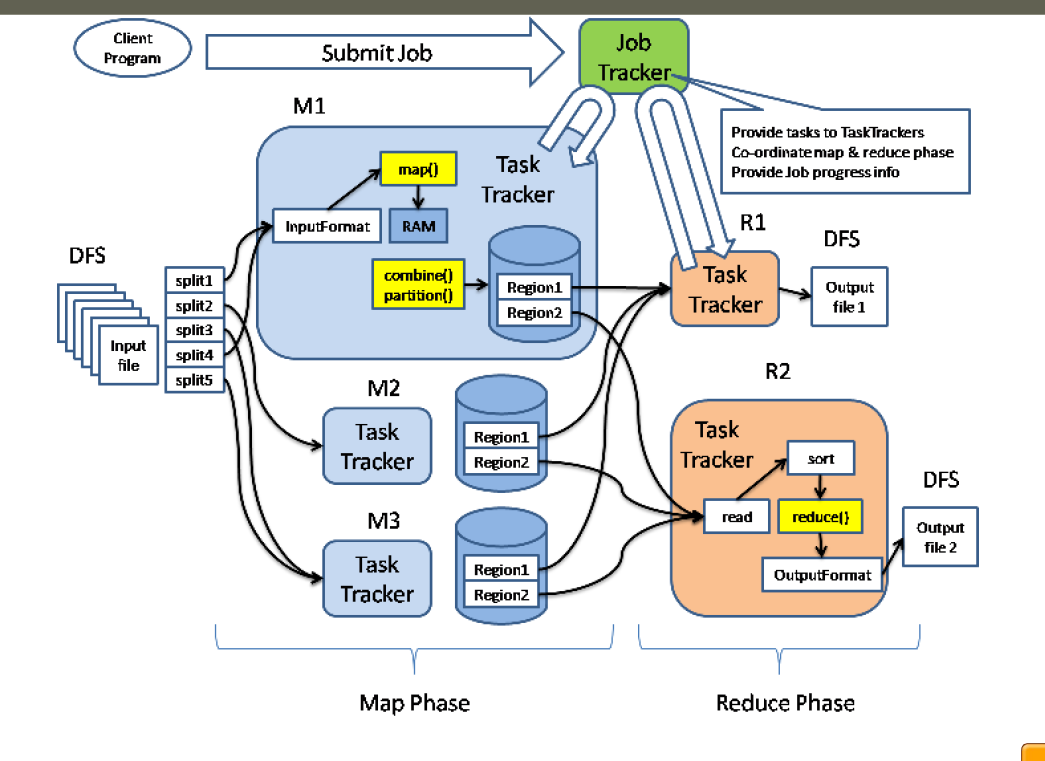








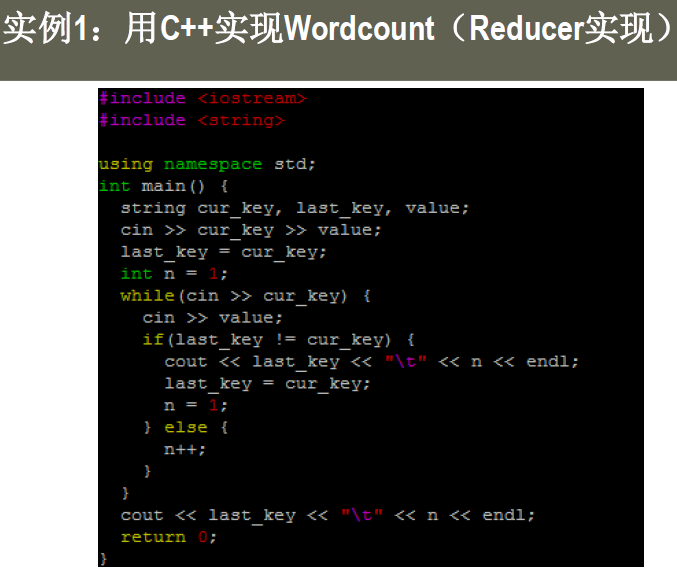




多语言编程：

1. 以标准输入作为输入：
   1. C++：cin
   2. Scanf
2. 以标准输出作为输出
   1. C++：count
   2. C：printf
3. 可实现Mapper和Reducer，其他组件（InputFormat、Partitioner等需要用Java语言实现）



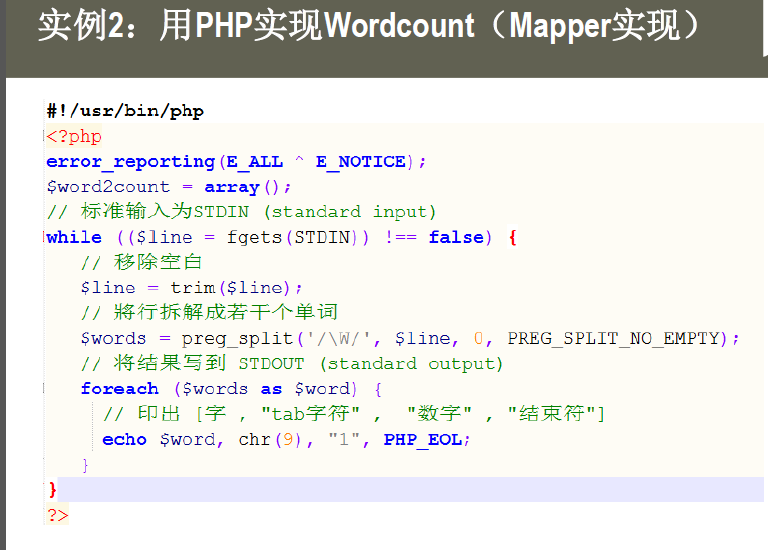


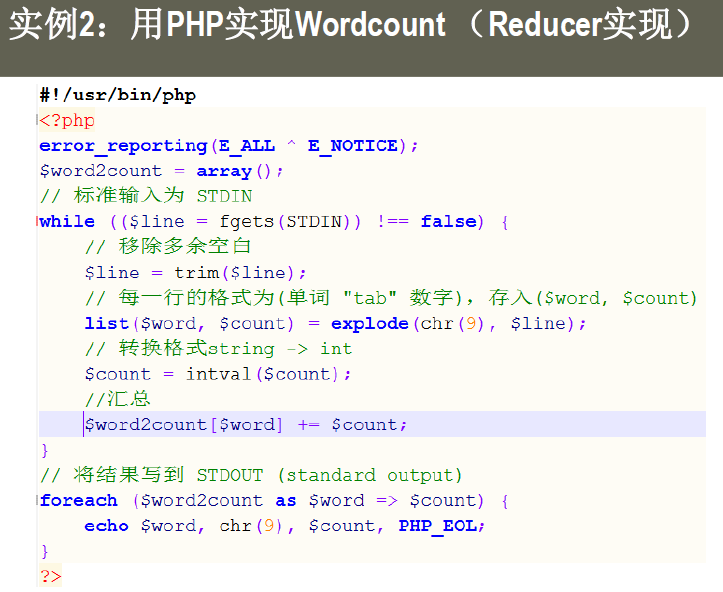




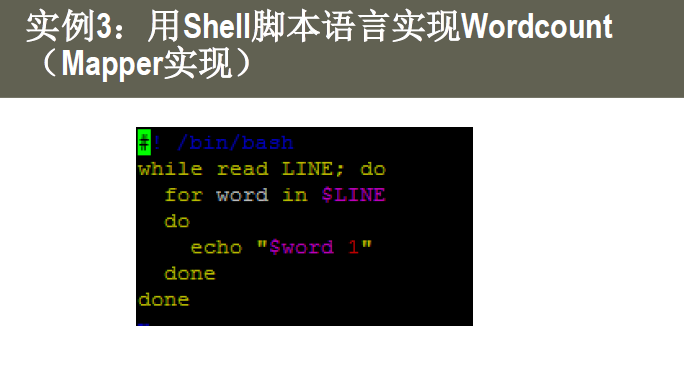
Streaming程序运行方式说明：

1. 区分通用参数和命令行参数，通用参数应放在命令参数前面，否则不起作用。通用参数有7个：
   1. Conf –D –fs –jt –files 0libjars –archives
2. “-file”或者”-files”参数，设置要奋发到各个节点上的文件，对于Mapper和Reducer文件，必须要用或者”-files”或者”-file”指定。
   1. -files mapper,reducer
   2. –file mapper –file reducer
3. 每次运行程序前，清空输出目录。





Shell：







Hadoop Streaming 高级编程：

1. 定义Hadoop Counter
   1. 使用标准错误输出，格式为：
      1. Reporter:counter:<group>,<counter>,<amount>
2. 在运行过程中战士状态信息
   1. 使用标准错误输出，格式为
      1. Reporter:status:<message>
3. 打印调试信息
   1. 使用标准错误输出，调试信息将被保存stderr文件中
4. 获取conf配置信息，比如reduce task个数等
   1. 在环境变量中，所有的”.”被替换成了”\_”

Int main(int args,char \*argv[],char \*env[]){…}

比如”mapreduce.job.reducers”被换成了”mapreducer\_job\_reducers”



Java与Streaming编程比较：

1. Java编程：
   1. Hadoop最原始开发语言；
   2. 支持所有功能，其他编程方式的基础
2. Streaming编程：
   1. 仅适用于开发Mapper和Reducer，其他组件需要采用Java实现
   2. 天生支持文本格式，但二进制格式支持较弱（新版的Hadoop支持）；
   3. 通常用于简单的文本数据处理，加快开发效率。