## 分布式文件系统

### 分布式文件系统需提供的功能

1. 文件系统目录树
2. 文件的读取

### 以分布式文件系统的地址为建立基础的方式

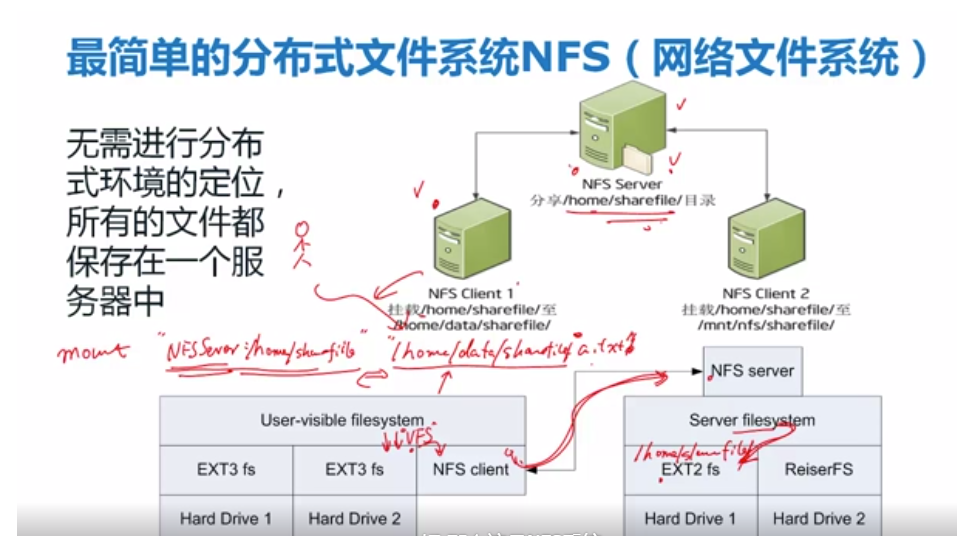
服务器->磁盘

服务器上有操作系统管理好磁盘，服务器本地有一个Local FS来定位到磁盘上。

### 分布式文件系统的本质功能

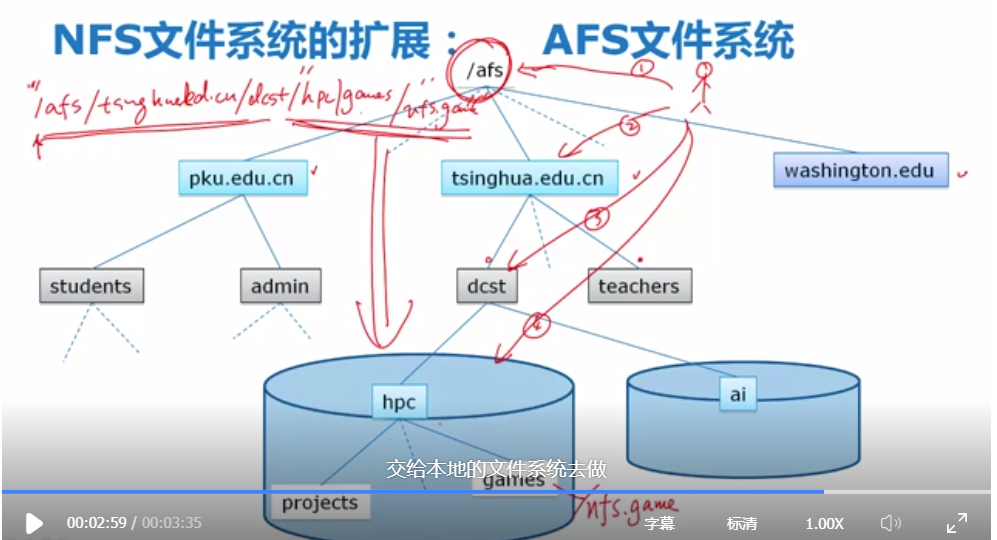
将一个以目录树表达的文件翻译为一个具体的节点，而到磁盘的定位则可以交给本地文件系统去完成。

### 分布式文件系统简单示例NFS



### NFS的扩展：AFS文件系统

每个服务器都会记录其下一级服务器。



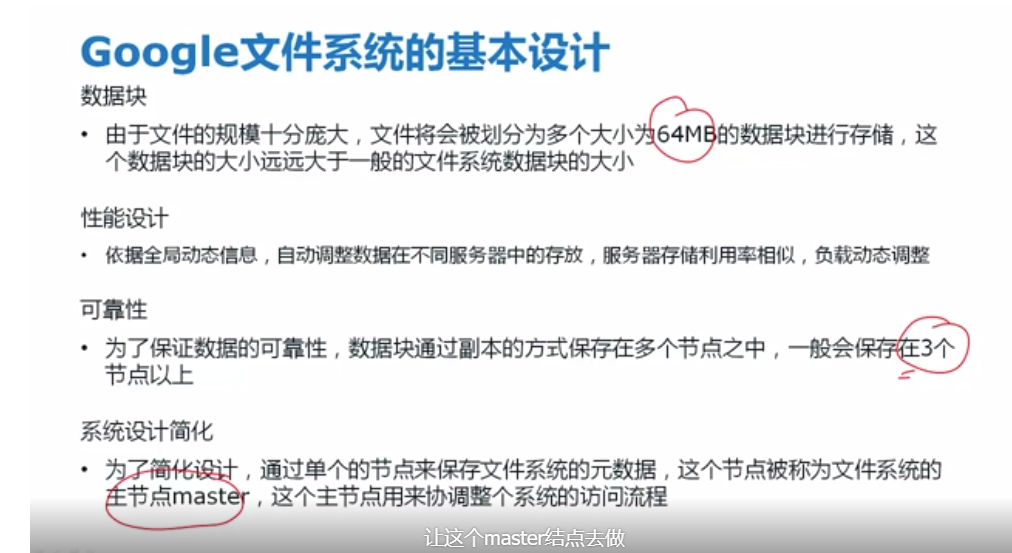
### Google文件系统

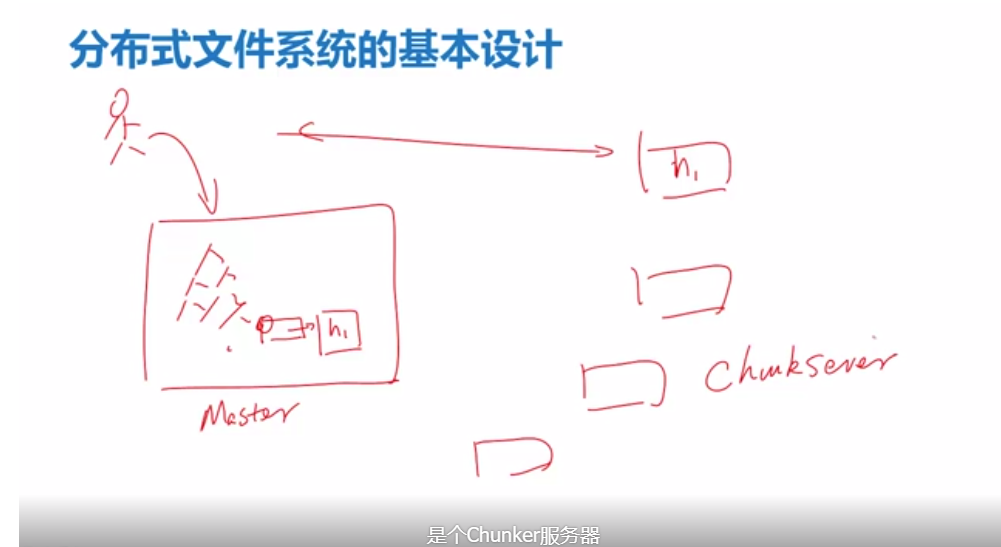
#### 倒排索引

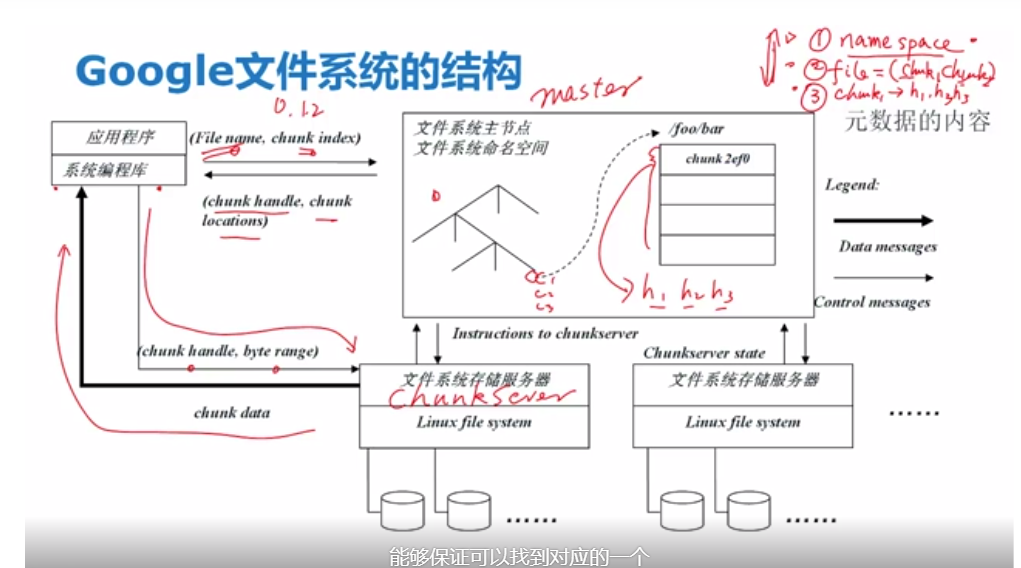
倒排索引(Inverted Index)：倒排索引是实现“单词-文档矩阵”的一种具体存储形式，通过倒排索引，可以根据单词快速获取包含这个单词的文档列表。倒排索引主要由两个部分组成：“单词词典”和“倒排文件”。

参考：<https://blog.csdn.net/starzhou/article/details/87519973>

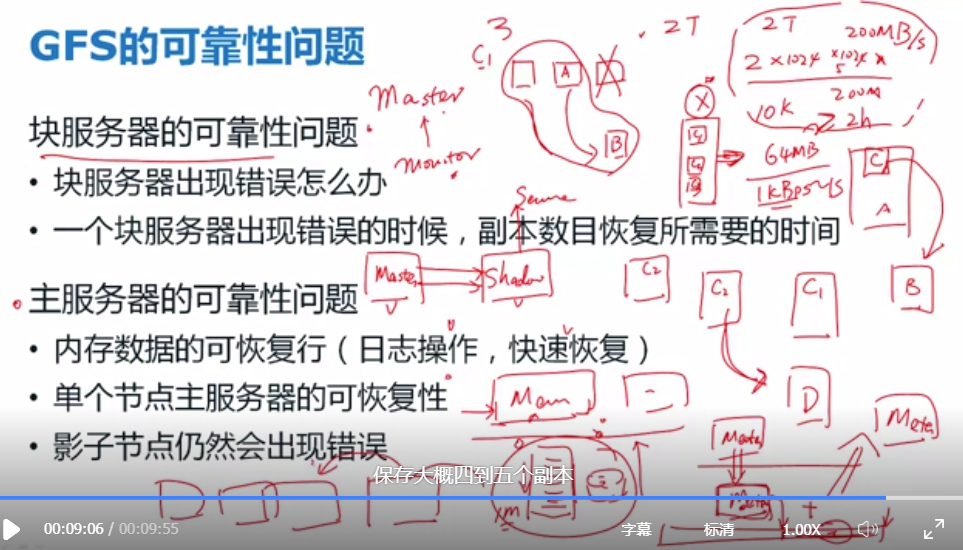
#### 谷歌文件系统的基本设计



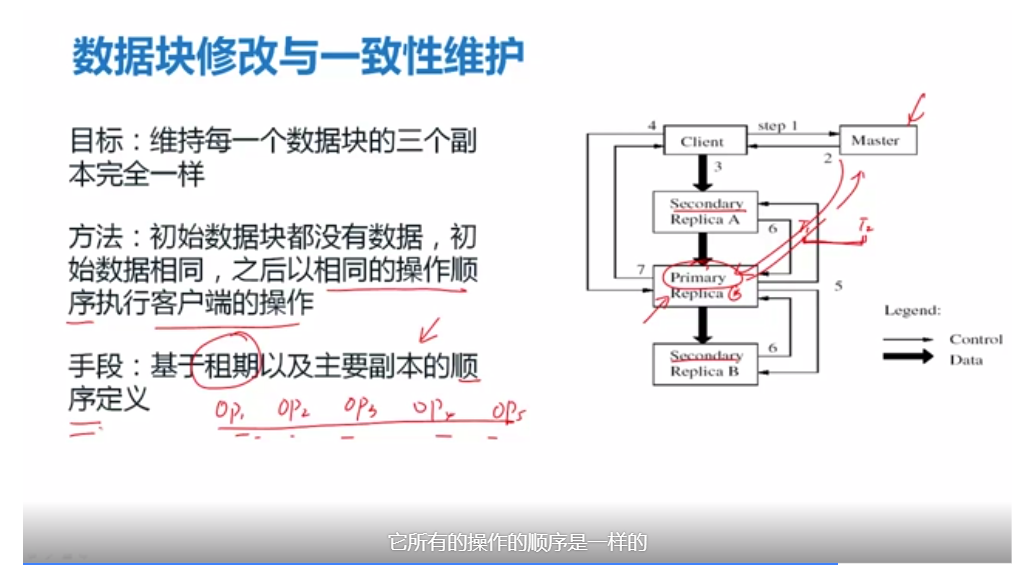


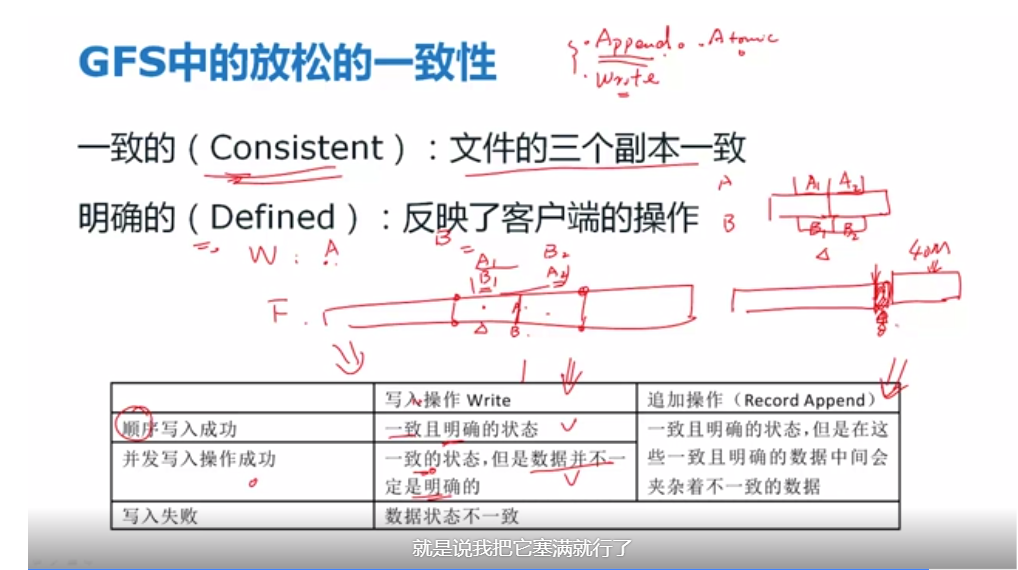


#### GFS可靠性

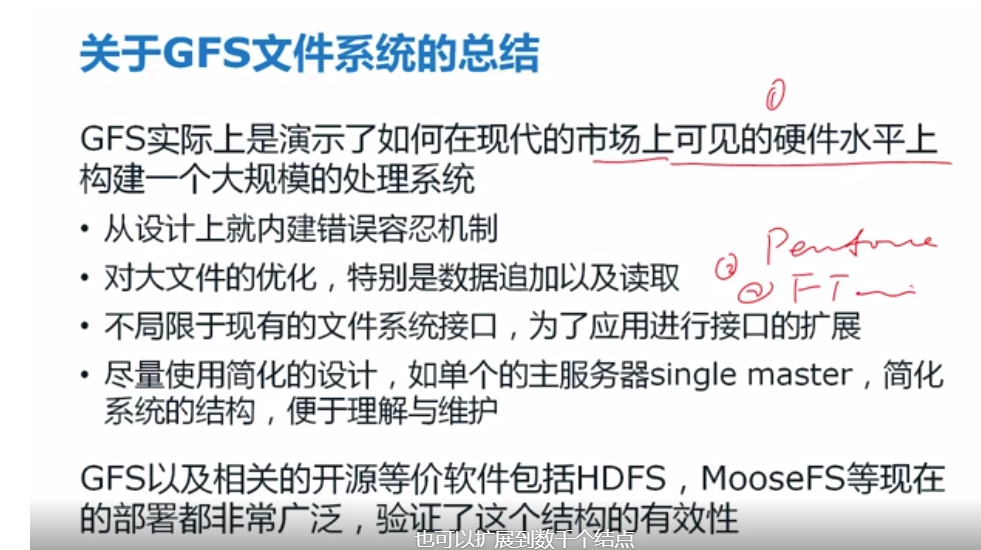


#### GFS数据块一致性：三个副本一致性



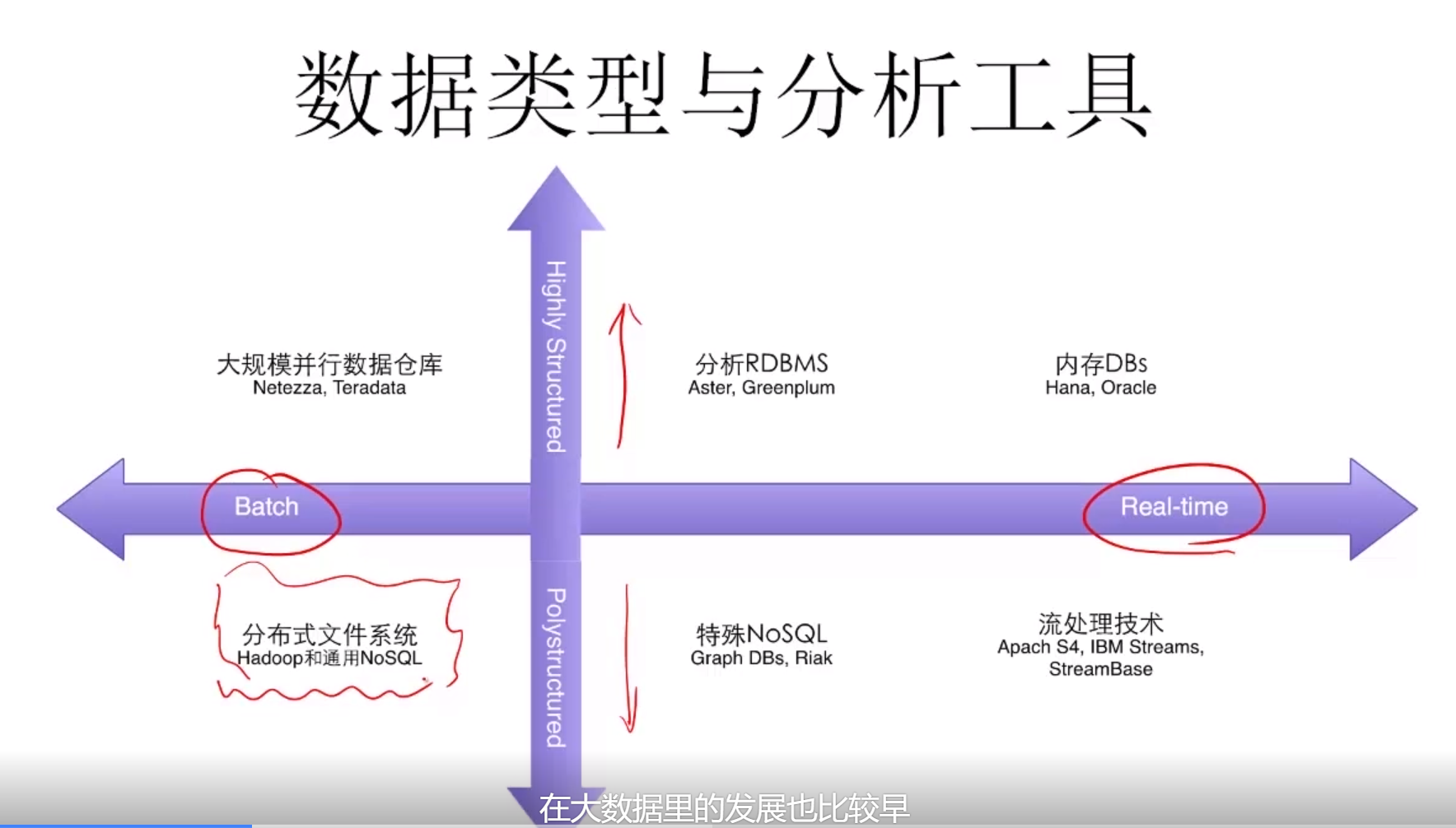


#### 小结



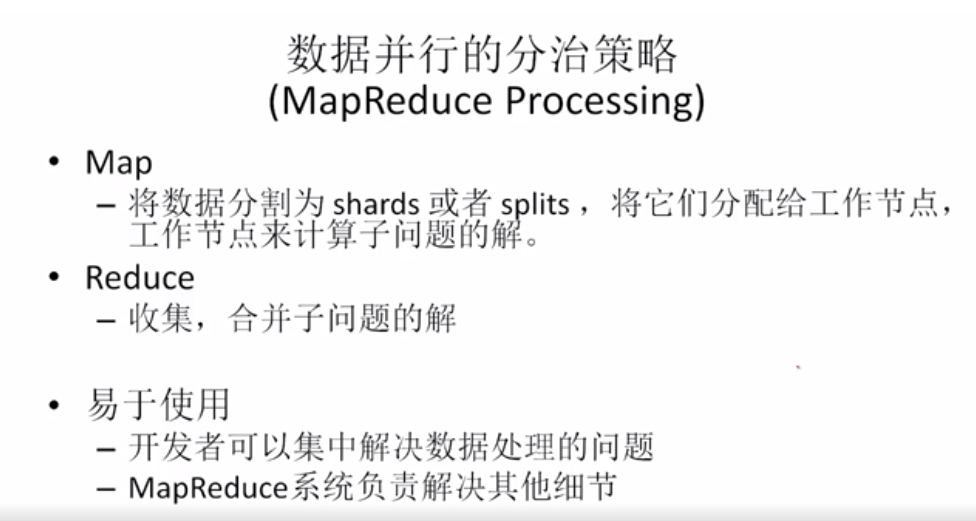
## 处理框架

### 数据类型与分析工具

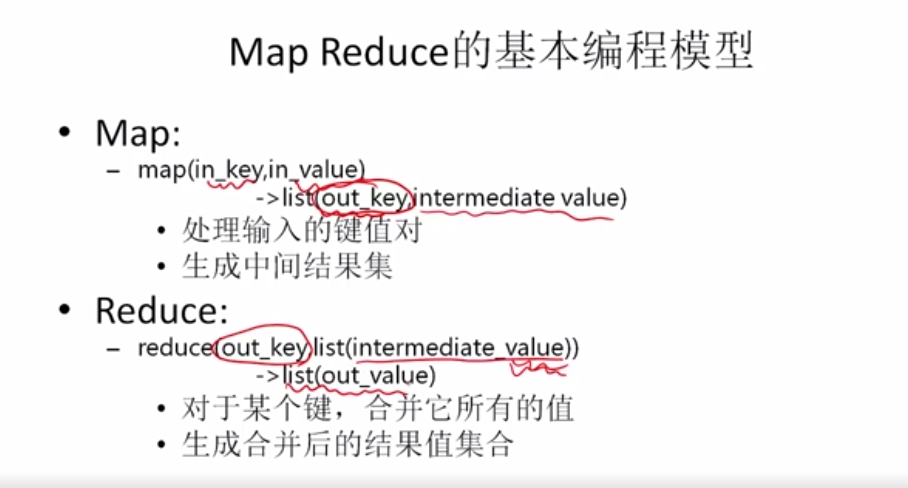


### MapReduce

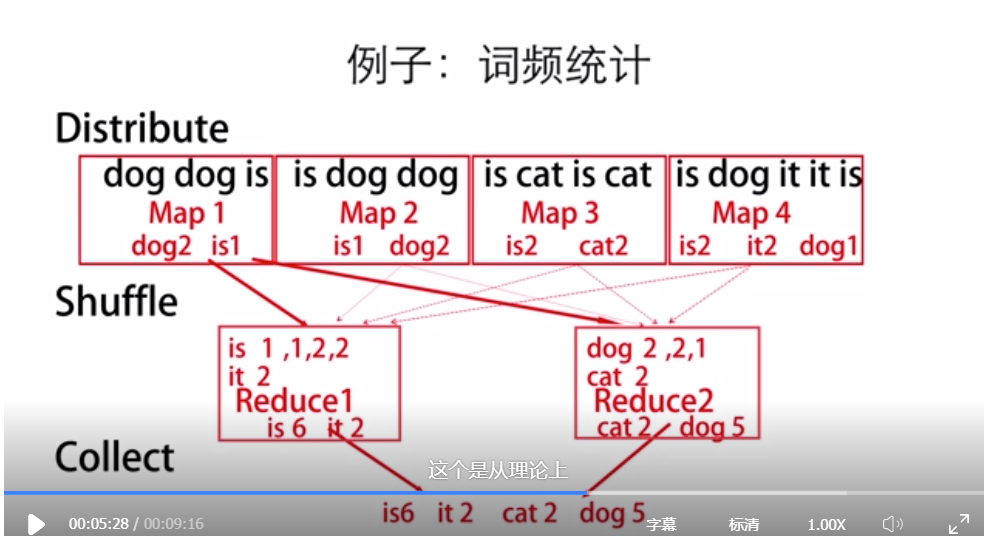
#### MapReduce分而治之策略

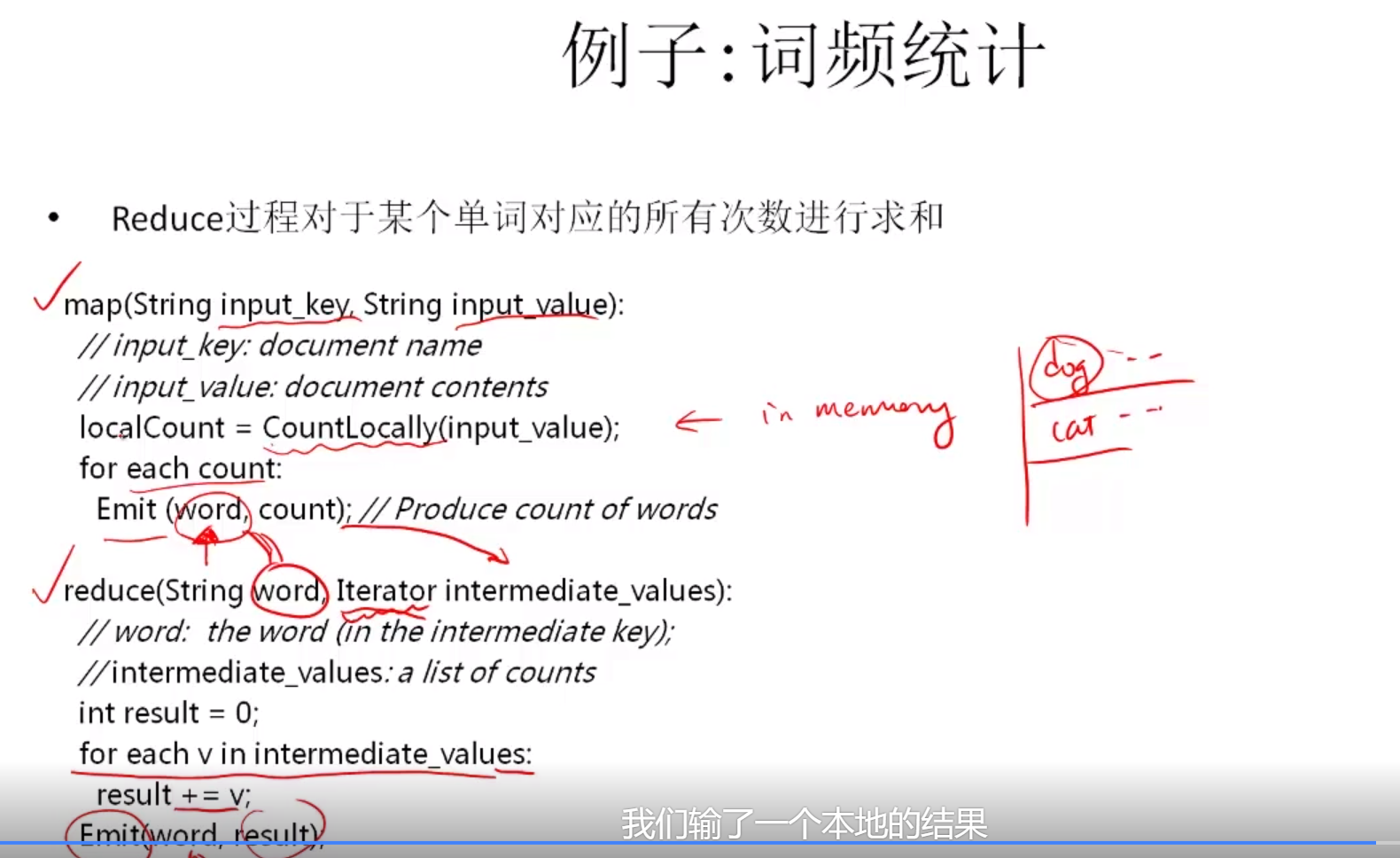


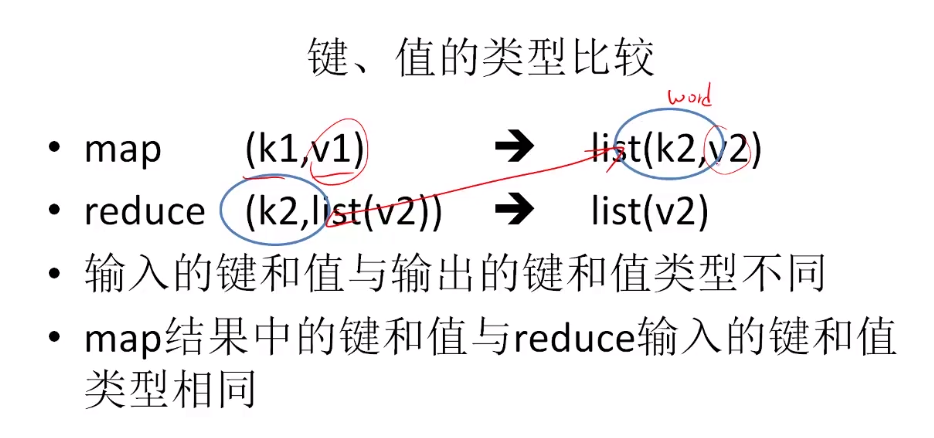
#### MapReduce基本编程模型



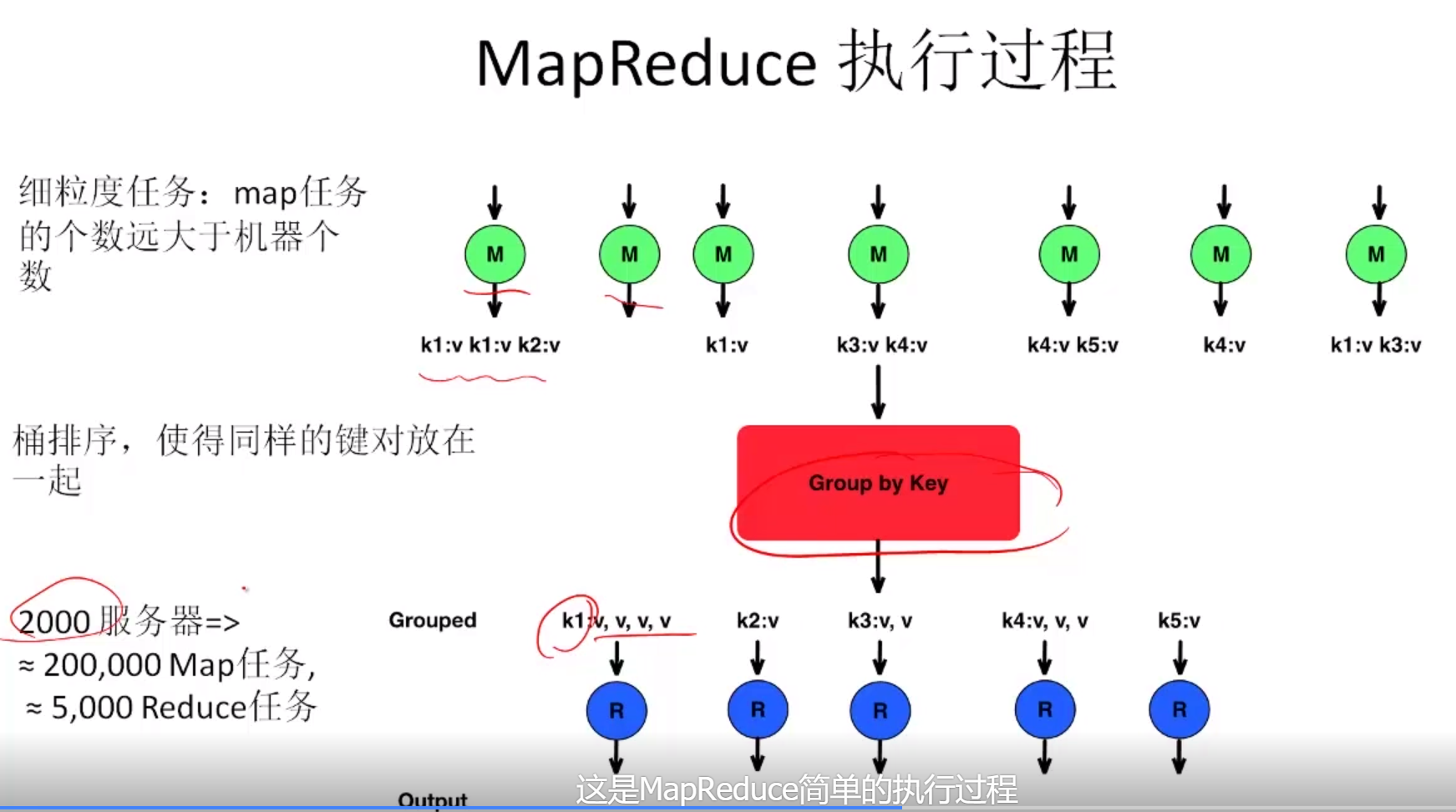
#### 词频统计示例

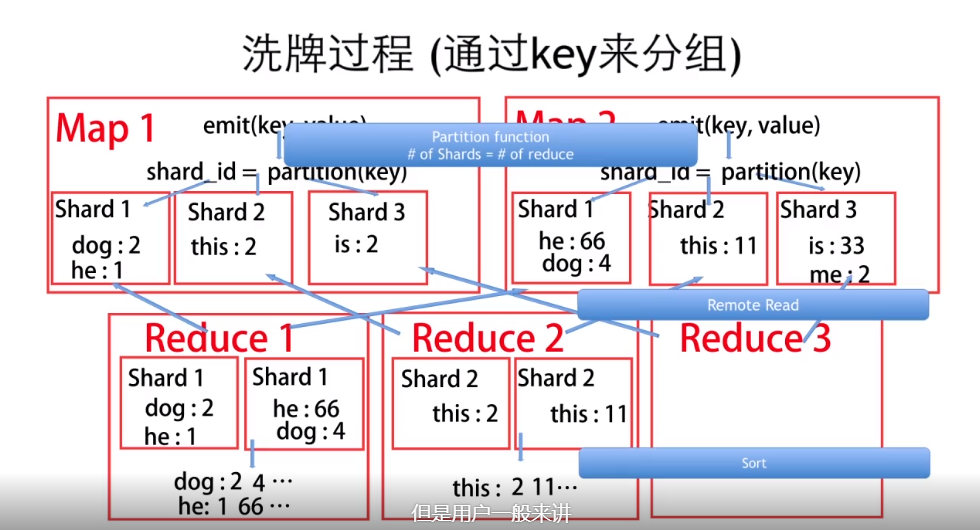




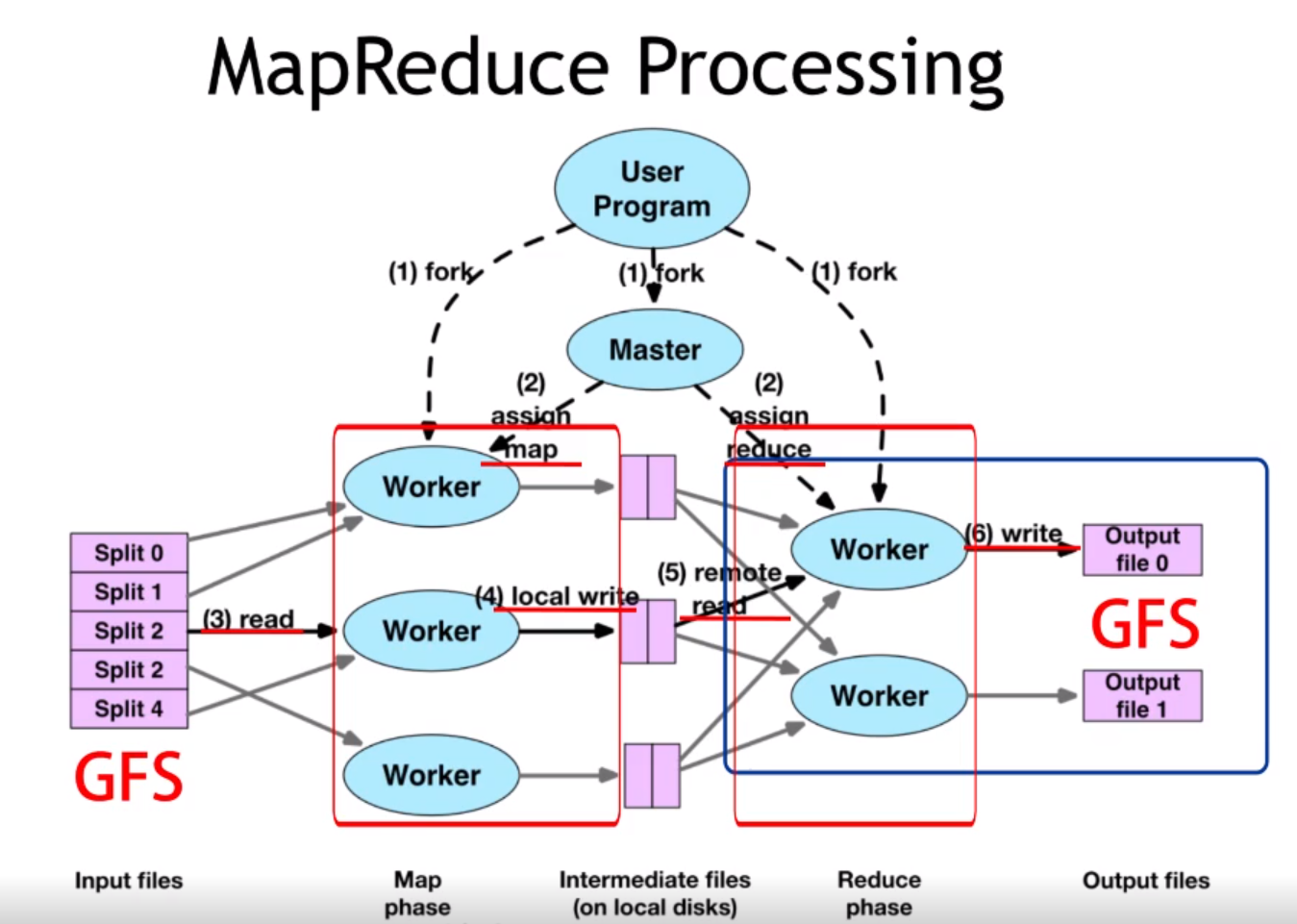


#### MapReduce执行过程

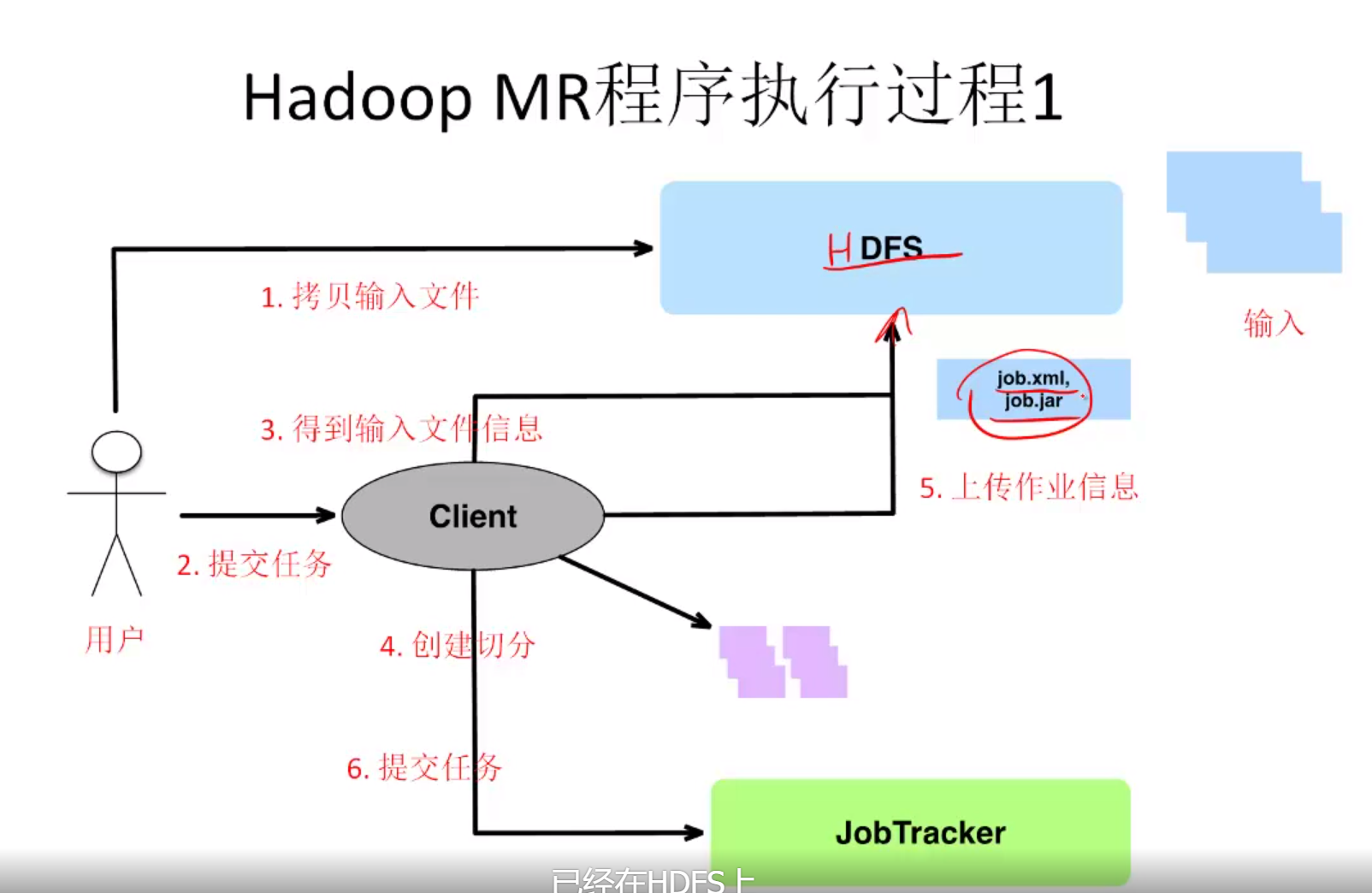


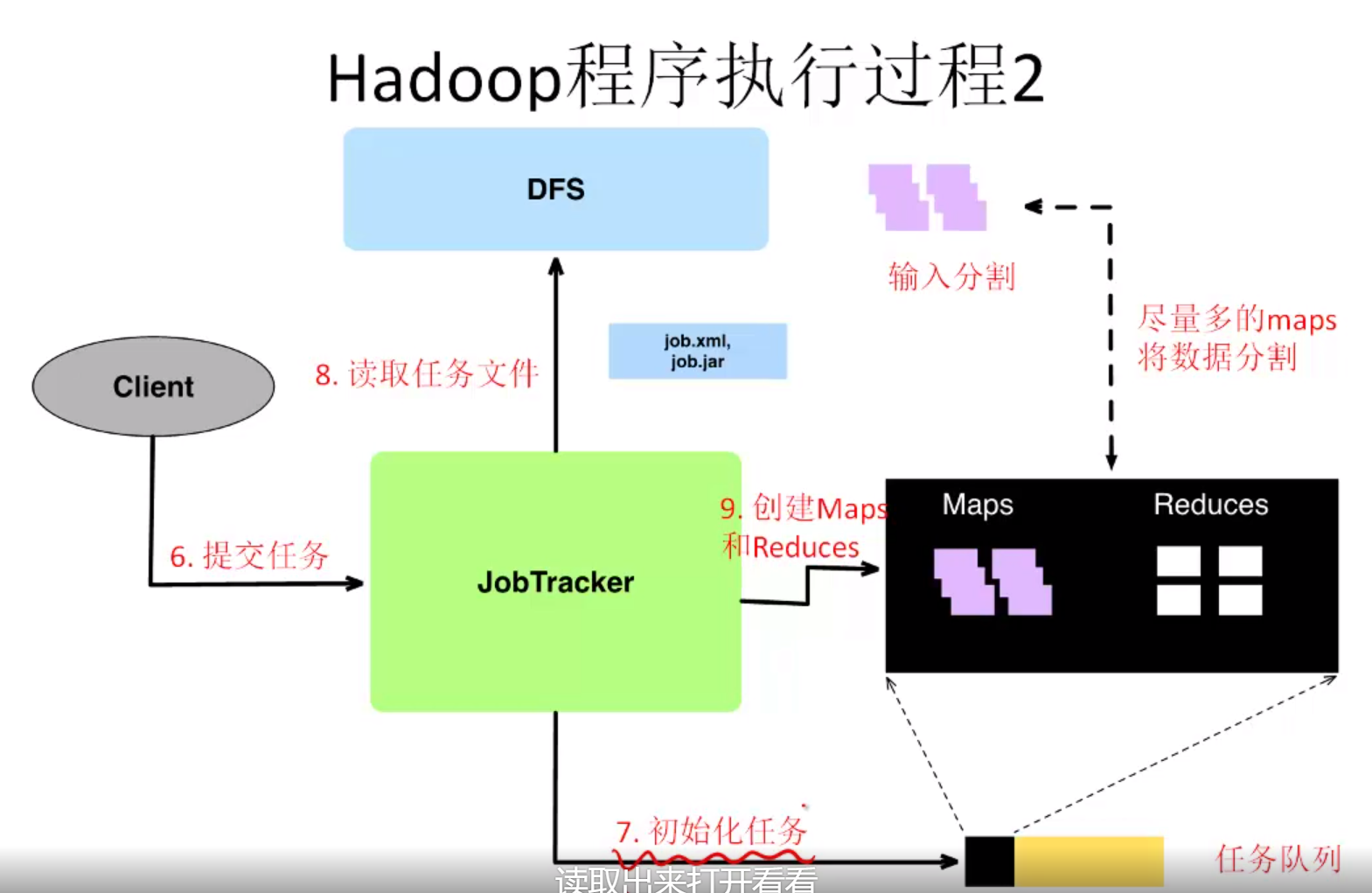


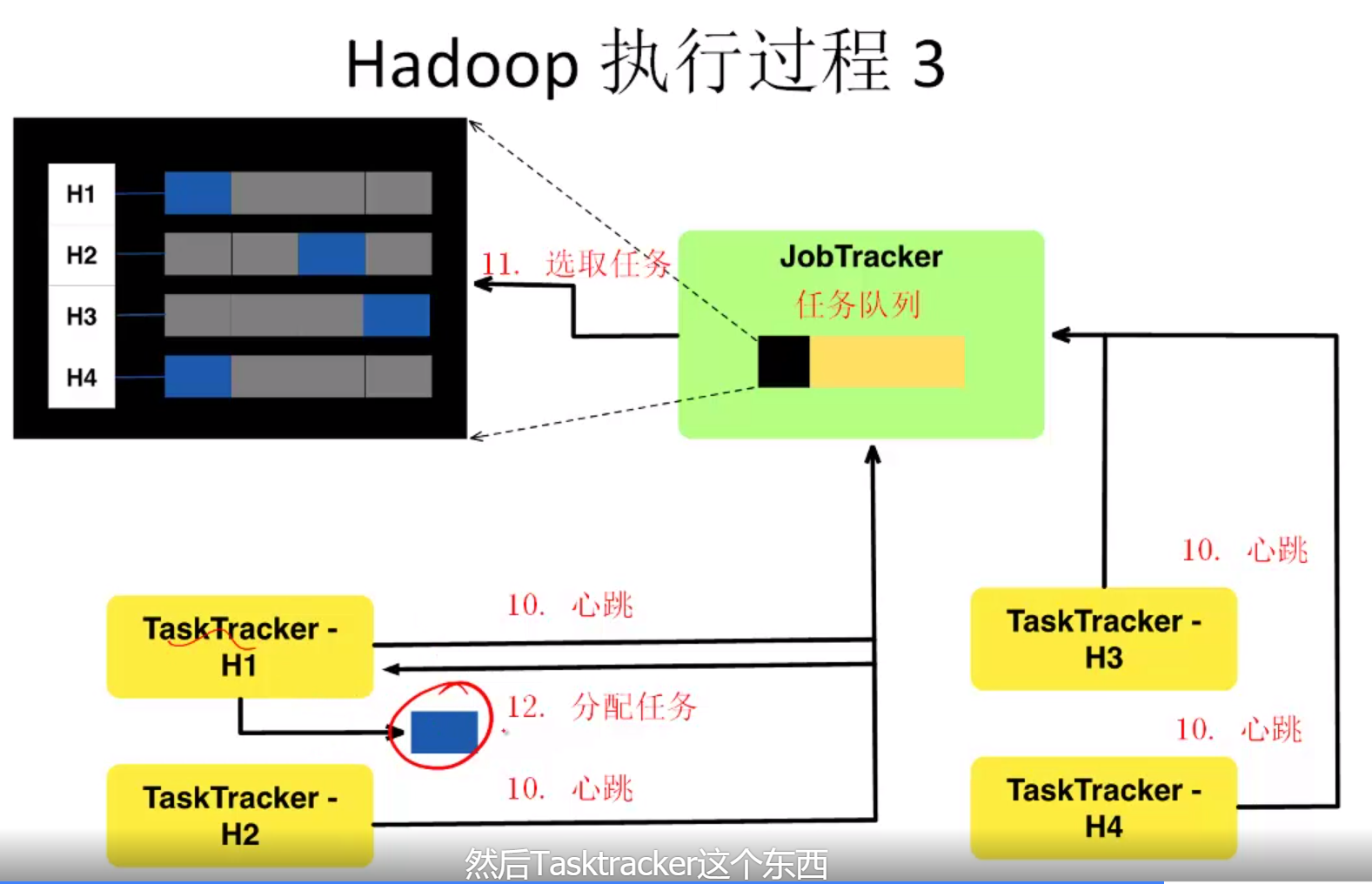
#### MapReduce数据流



### Hadoop: MapReduce的开源实现







## PySpark使用

### PySpark安装

1. Pip install 升级+清华镜像 若Pip install 失败，多试几次就好。
2. Hadoop相关 下载Hadoop binary版本，并按照参考链接配置好环境变量，下载所指出的wutils相关放入Had。
3. SQL Server JDBC驱动安装,官网下载<https://docs.microsoft.com/zh-cn/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15>，并放置在Python脚本所在的文件夹，在获取PySpark上下文时需要设置好驱动的相对路径或绝对路径。
4. 参考链接：<https://blog.csdn.net/shiheyingzhe/article/details/80714301>