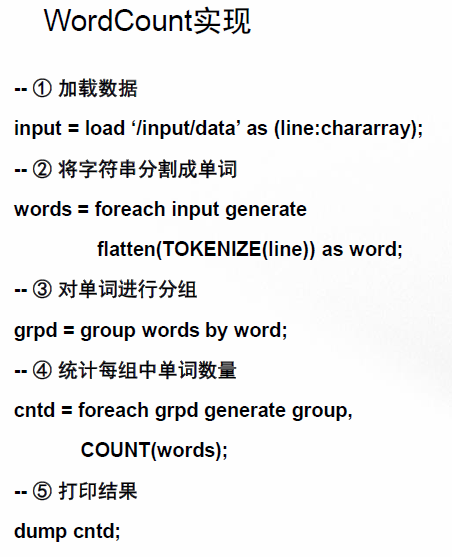
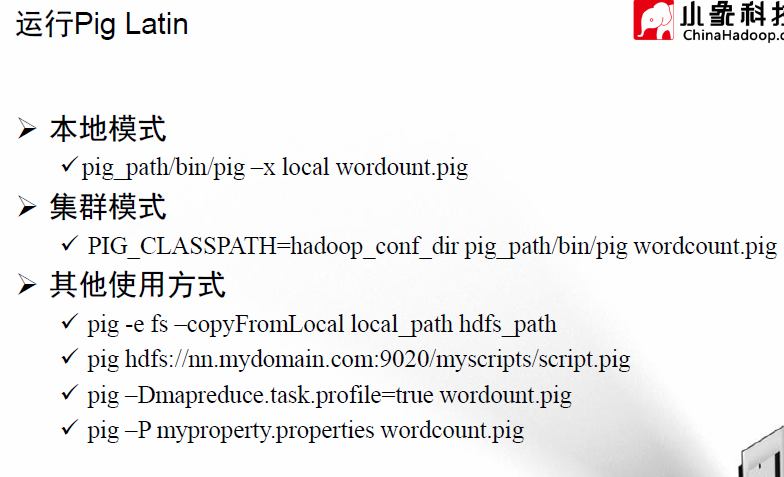
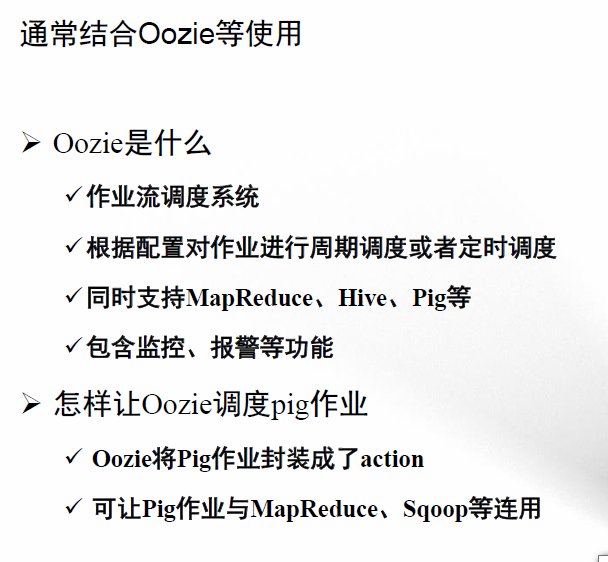
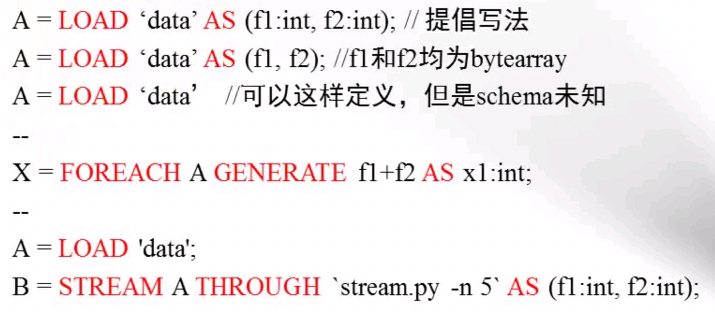
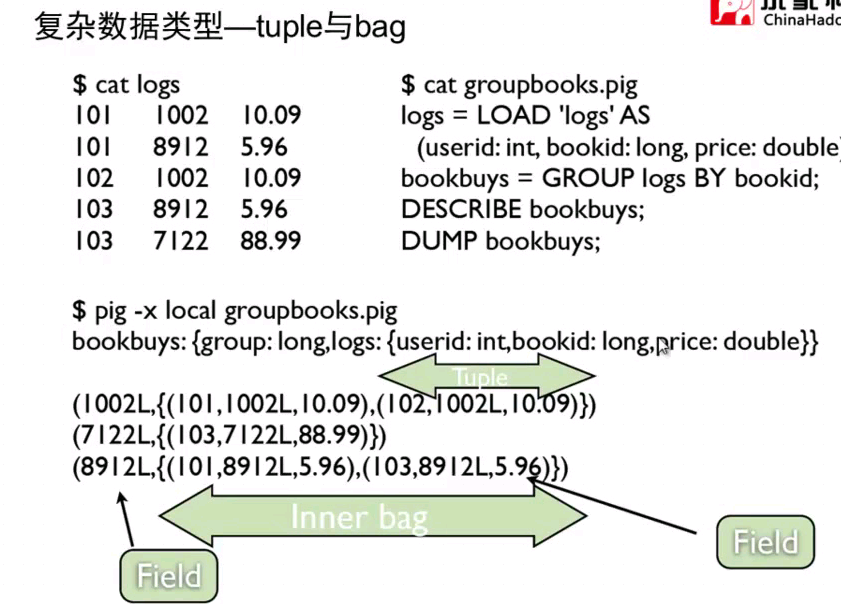
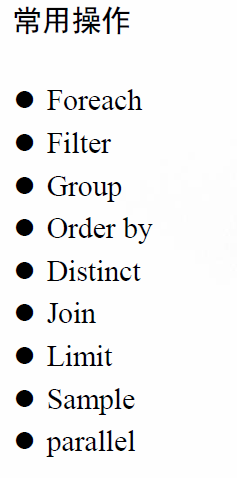
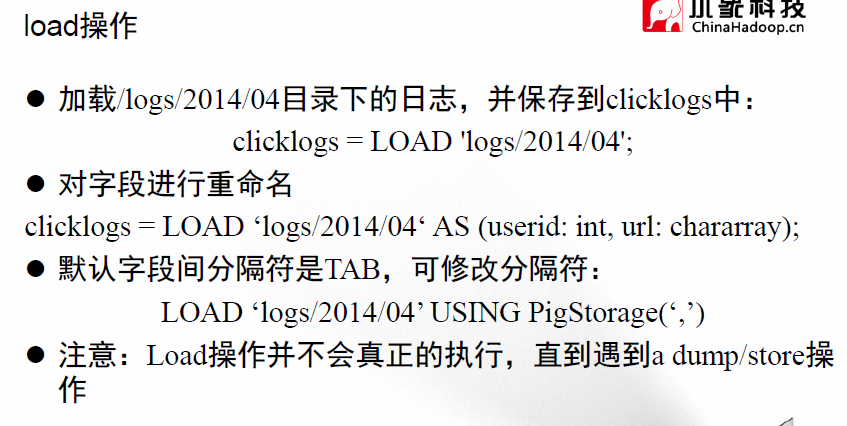
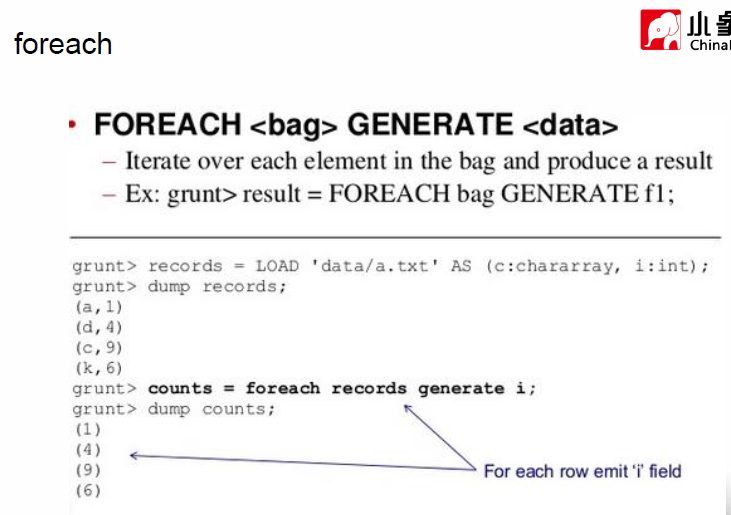
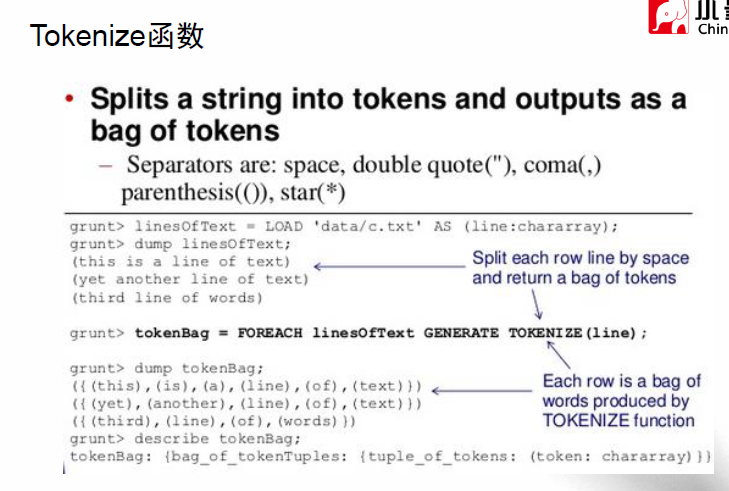
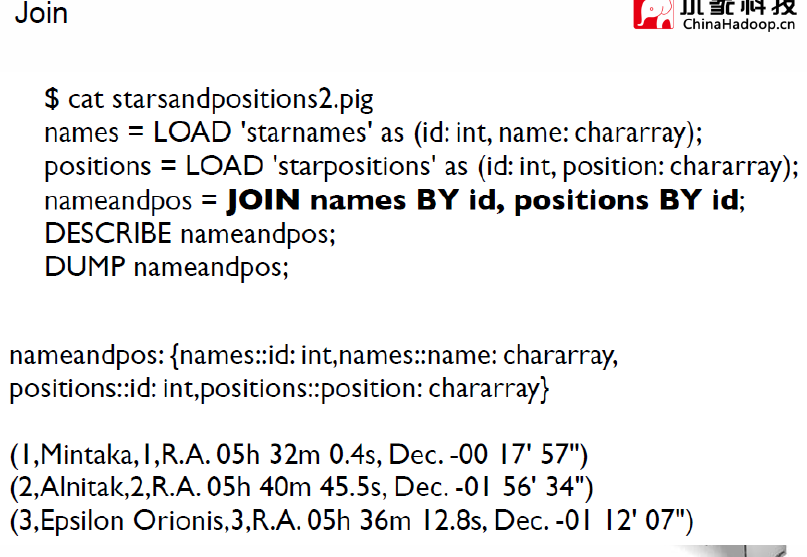
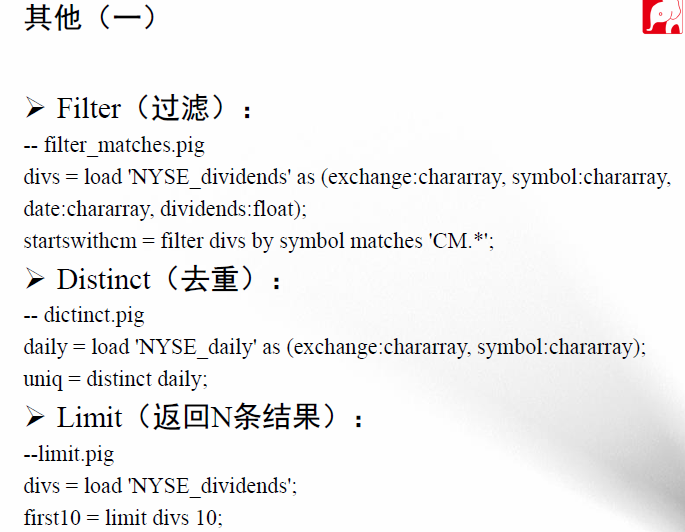
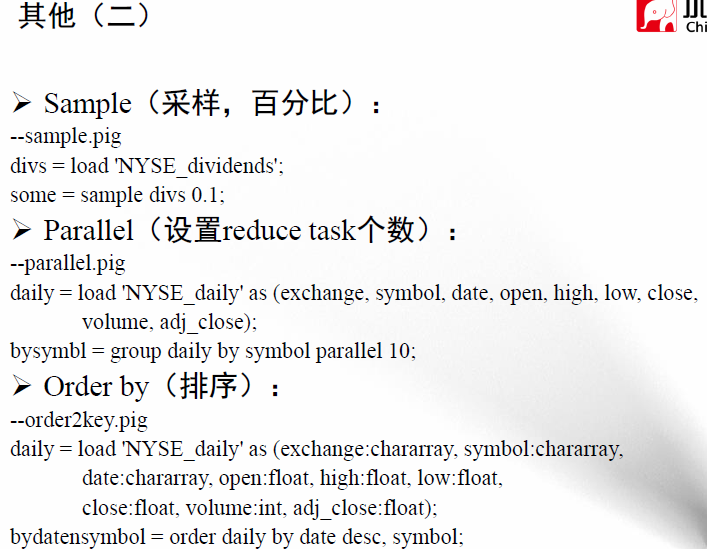
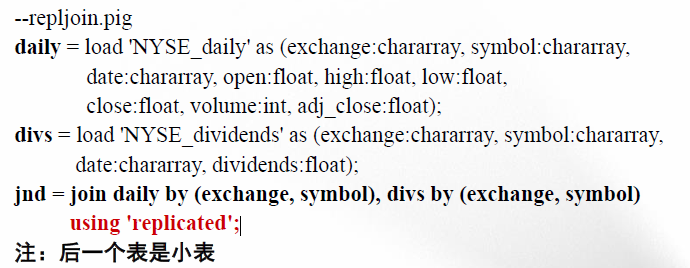
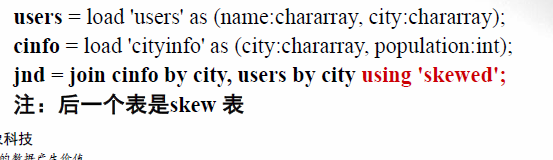
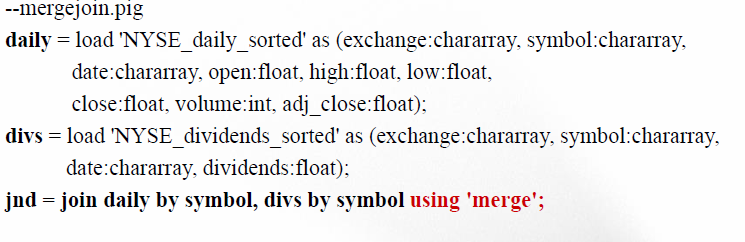
Pig是什么？

1. Hadoop上的数据流引擎（由Yahoo开源）
   1. 利用HDFS存储数据
   2. 利用MapReduce处理数据
2. 使用Pig Latin语言表达数据流
   1. Pig Latin是一种新的数据流语言
   2. Pig将Pig Latin语句转化为MapReduce作业
   3. Pig Latin比MapReduce程序更易编写
3. 直接动机：让MapReduce用起来更简单
   1. 与Hive一致
4. Pig与Hive异同
   1. 相同点
      1. 运行在Hadoop之上；
      2. 设计动机是为用户提供一种更简单的Hadoop上的数据分析方式；
      3. 解决相同问题的两个工具（Yahoo！vs Facebook）
   2. 不同点
      1. Hive要求待处理的数据必须有schema，而pig则无此要求；
      2. Hive有server需要安装，pig无server无需要；
      3. 编程语言不通，SQL与Pig Latin
         1. SQL：得到什么样的结果，Pig Latin：如何处理数据
         2. SQL：过程简化语言，PIG Latin：数据流语言。
5. Pig使用
   1. 
   2. 
   3. 
   4. Pig Latin：
      1. 数据模型：
         1. 可通过LOAD、FOREACH和STREAM三个操作符将数据模式化
         2. 可为每个迷失致命每个字段的类型，也可以不指定名称，默认是bytearray
         3. 
         4. 
         5. 
         6. 
         7. 
         8. 
         9. 
         10. 
         11. 
         12. 
         13. 
         14. 
         15. 
   5. 调优：
      1. Fragment-replicate join或者shuffle join或者map-side join(一个大表一个小表)
      2. 
      3. Skew data join（一个表某个key特别多）：
      4. 
      5. 两个有序表join：
      6. 
   6. Pig参数调优
      1. Pig.cachedbag.memuseage: Map task或者reduce task中，bag数据占用内存达到JVM内存的一定百分比后，数据将被溢写到磁盘上，默认值是0.1；
      2. Mapred.compress.map.output: 是否对Map Task中间结果进行压缩，如果设置为true，可进一步使用mapred.map.output.compression.codec.设置采用的压缩器；
      3. pig.noSplitCombination: 是否将不及一个block大小的多个小文件交由一个Task处理，默认是false，注意，这会降低数据本地性。
      4. 其他参数调优：MapReduce级别的buffer调优，包括io.sort.mb等