Hadoop开源软件及生态系统：方向hadoop的运维，hadoop的开发按用户规格或为开源软件做二次开发。

云计算与大数据：狭义的云计算与广义的云计算；三层模型；

Hadoop的起源：Doug Cutting，google核心技术，

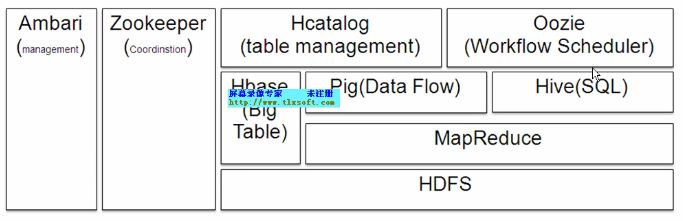


Google vs Hadoop

Hadoop的特点：开源社区的支持，分布式文件系统的备份恢复机制以及MapReduce的任务监控保证了分布式处理的可靠性，并且它的框架可以运行在任何普通的pc上，不论是存储的可扩展还是计算的可扩展都是hadoop的设计根本，分布式文件系统的高效数据交互实现，以及MapReduce结合的localdata的处理模式，为高效处理海量信息做了基础准备。

Hadoop架构简介：hadoop的内核：hdfs组件，MapReduce组件，Common组件，common组件是hadoop基础，提供了一些hadoop io，压缩，rpc通信，序列化等功能，同时，common组件可以利用jni方法调用c/c++编写的native库，加速数据压缩，数据校验等；hdfs采用流式数据访问机制，可以用来存储超大文件，hdfs集群拥有两种节点，名称节点namenode，数据节点datanode，名称节点在内存中保存着文件数据块的映像信息和整个文件系统的名字空间，数据节点负责存储和读取数据文件。Hdfs组件，mapreduce组件（jobtracker-tasktracker-maptask，reducetask，word count应用）；mapreduce的执行过程。

Hadoop生态系统：



Hadoop发行版：Cloudera CDH，Hortonworks HDP，intel Distribution，IBM BigInsight。解决繁琐的依赖关系等。

Hadoop版本选择：hadoop 1.0,2.0，其中1.0包含0.20.x，0.21.x, 0.22.x，其中0.20.x最后演化为1.0.x，后两者则加上了NameNode HA等重大特性。Hadoop2.0版本分别为0.23.x，2.x，他们不同于hadoop1.0，是一套全新的架构，含有HDFS Federation和YARN两个系统，相比于0.23.x，2.x增加了NameNode HA，Wire-compatibility的特性。