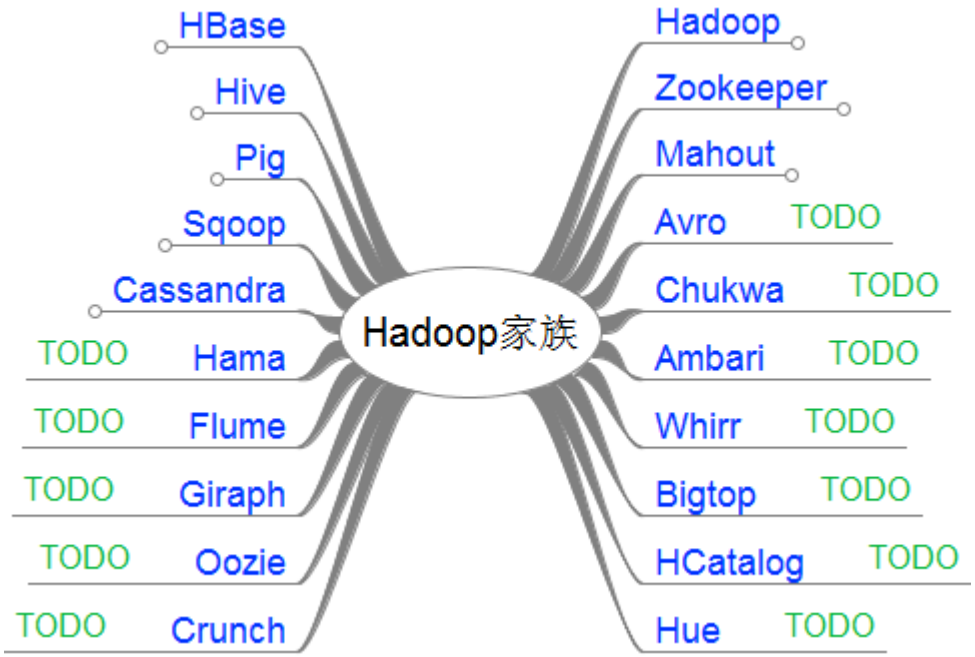
Day4 3月2日

软件51 庞建业 2151601012

Hadoop

Cloudera QuickStarts

学习资料 安装好的镜像文件




```
开发环境：
1、Jdk(Windows和Linux)
2、Eclipse(sts-3.5.1.RELEASE)(SVN插件) (Maven, Nexus可选)
3、Tomcat (可选)
4、VisualSVN
   (配置管理：日志、周报、项目文档、项目源代码)
5、Vmplayer
6、CentOs,Ubuntu的安装文件
7、Hadoop
8、MySQL (可选)
9、PowerDesigner (可选)
10、putty(管理工具)
```

数据之美

数据可视化是个非常宽泛的领域，大体可以分为`“信息图Infographic”`和`“可视化Data visualization”`两个方向。信息图主要用于网站、期刊、杂志等媒体传播渠道，社会传播属性较强。数据可视化主要用于商业数据分析场合，旨在清晰、简洁展示数据，进行商业分析和决策支持。

NOW

Running for 17s

UNITED STATES	27%
TURKEY	10%
SPAIN	10%
MALAYSIA	7%
MEXICO	7%

TODAY

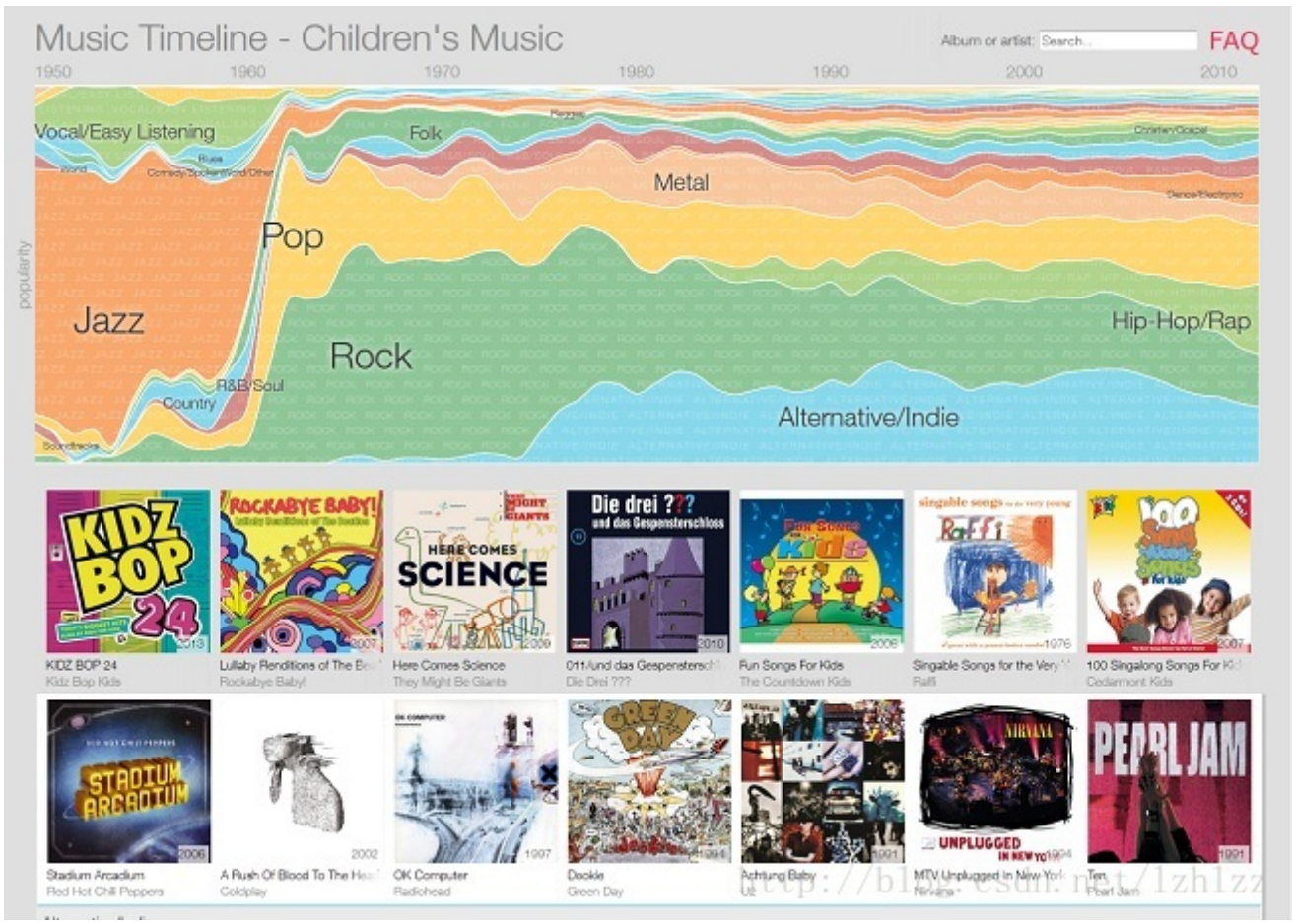
17th Aug 2014

waiting...	
waiting...	
waiting...	
waiting...	
waiting...	

HISTORY

Since 1st Nov 2010

UNITED STATES	27%
BRAZIL	20%
INDONESIA	10%
UNITED KINGDOM	6%
JAPAN	3%



大数据

数据管理 数据分析 数据计算

关系数据库 数据导入到hadoop中

有`数据仓库` `nosql`等

MapReduce spark计算框架

计算完后数据放到hadoop中 mysql二维数据库 数据展示

腾讯: `蓝鲸Paas` `flink/Druid`数据采集框架

> IaaS（Infrastructure as a Service）：商品是基础设施 PaaS（Platform as a Service，PaaS）：商品是平台 SaaS（Software as a Service，SaaS）：商品是最终消费品

```
Platform as a Service: 平台即服务，是面向软件开发者的服务，云计算平台提供硬件，OS，编程语言，开发库，部署工具，帮助软件开发者更快的开发软件服务．比如Google的GAE．
SaaS: 软件即服务，是面向软件消费者的，用户无需安装，通过标准的Internet工具(比如浏览器)，即可使用云计算平台提供的软件，比如Salesforce的CRM管理系统，Google的Gmail．
```

模型版本hadoop3: hadoop.apache.org/docs/current/

`单节点安装模式` 一个虚拟机name node命名节点 底下很多data node 数据计算管理 相当于服务 resource manage资源管理 运算计算在很多节点完成

`单机模式`是所有进程在一个机子上完成

使用JPS查看JAVA进程

JAVA_HOME 在虚拟机上

export JAVA_HOME=/jre1.8/

echo可以看到JAVA路径

ssh免秘钥登录

三台机子间做文件拷贝

网址 分布式环境搭建

Hadoop 环境搭建

a）首先完成SSH无密码登录

b）安装JDK，要确认里面有JPS工具

c）下载hadoop安装

在Master机器上，/hadoop/sbin/start-all.sh

启动节点 客户端查看 http 端口访问

硬盘 内存 cpu 存储+计算 传统计算分布式计算 随着计算机数量增加 到一定程度算力下降

为了提高并行运算速度 提高存储效率 共同利用计算机内存

```
（1）配置机器 hostname
Hostname Hadoop-Master
hostname Hadoop-Slave1
hostname Hadoop-Slave2
修改hosts文件关联关系
实验ping通
（2）ssh免密码验证配置
首先在Master机器配置
1-1.进去.ssh文件：[root@Hadoop-Master ~]#cd ~/.ssh 如果没有该目录，先执行一次ssh localhost，不要手动创建，不然配置好还要输入密码
1-2.生成秘钥 ssh-keygen:ssh-keygen -t rsa,一路狂按回车键就可以了,最终生成（id_rsa,id_rsa.pub两个文件）
1-3.生成authorized_keys文件：[spark@S1PA11 .ssh]$ cat id_rsa.pub >>authorized_keys
1-4.在另两台台机器Slave1、Slave2也生成公钥和秘钥
1-5.将Slave1机器的id_rsa.pub文件copy到Master机器：[root@Slave1 .ssh]#scp id_rsa.pub root@Hadoop-Master:~/.ssh/id_rsa.pub_s1
1-6.将Slave2机器的id_rsa.pub文件copy到Master机器：[root@Slave1 .ssh]#scp id_rsa.pub root@Hadoop-Master:~/.ssh/id_rsa.pub_s2
1-7.此切换到Master机器合并authorized_keys;
 [root@Hadoop-Master .ssh]# cat id_rsa.pub_s1>> authorized_keys
 [root@Hadoop-Master .ssh]# cat id_rsa.pub_s2>> authorized_keys
1-8.将authorized_keyscopy到Slave1、Slave2机器：
 [root@Hadoop-Master.ssh]# scp authorized_keys root@Hadoop-Slave1:~/.ssh/
 [root@Hadoop-Master.ssh]# scp authorized_keys root@Hadoop-Slave2:~/.ssh/
1-9.现在将各台 .ssh/文件夹权限改为700，authorized_keys文件权限改为600（or 644）
chmod 700 ~/.ssh
chmod 600 ~/.ssh/authorized_keys
1-10.验证ssh
[root@Hadoop-Master .ssh]# ssh Hadoop-Slave1
Welcome to aliyun Elastic Compute Service!
[root@Hadoop-Slave1 ~]# exit
logout
Connection to Hadoop-Slave1 closed.
[root@Hadoop-Master .ssh]# ssh Hadoop-Slave2
Welcome to aliyun Elastic Compute Service!
[root@Hadoop-Slave2 ~]# exit
logout
Connection to Hadoop-Slave2 closed.
```

> 数据转换成知识 知识是有用的知识

数据汇总 数据仓库

（1）ETL 数据抽取： 抽取 转换 装载

时间 业务 区域 BI多维分析

（2）建模 影响因素

（3）维度 主题 大的为维度 小的为主题

主题由多个维度推算过来

（4）挖掘 抽取 类脑模拟

RAID 冗余磁盘阵列存储

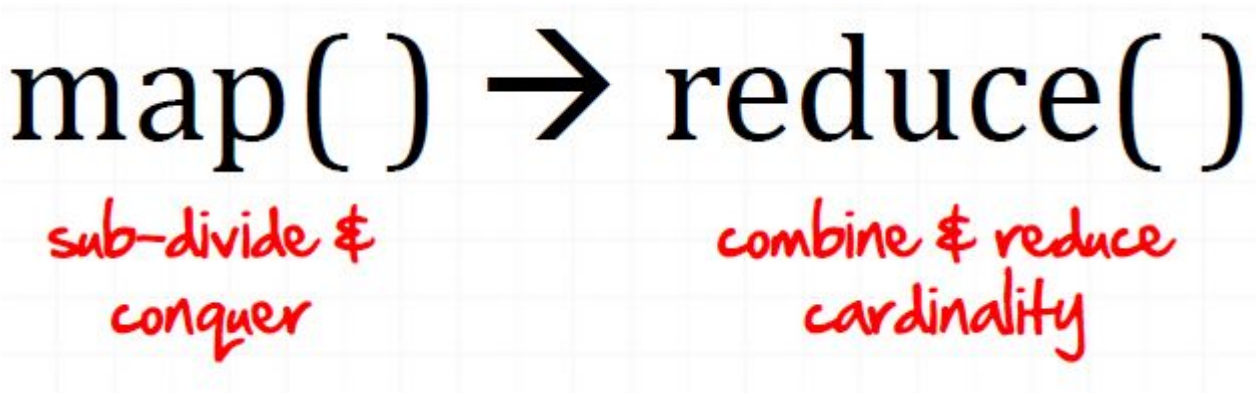
```
以前win98 FAT32 -> NTFS -> DFS（google）-> HDFS
```

hadoop中需要关注的 HDFS/MapReduce

HDFS

传统的文件系统是单机的，不能横跨不同的机器。**HDFS**（Hadoop Distributed FileSystem）的设计本质上是为了大量的数据能横跨成百上千台机器，但是看到的是一个文件系统而不是很多文件系统。比如说要获取/hdfs/tmp/file1的数据，引用的是一个文件路径，但是实际的数据存放在很多不同的机器上。作为用户，不需要知道这些，就好比在单机上你不关心文件分散在什么磁道什么扇区一样。HDFS管理这些数据。

```
HDFS，在由普通PC组成的集群上提供高可靠的文件存储，通过将块保存多个副本的办法解决服务器或硬盘坏掉的问题。
```



文件系统最主要解决存储问题： **存储效率高 存储大**

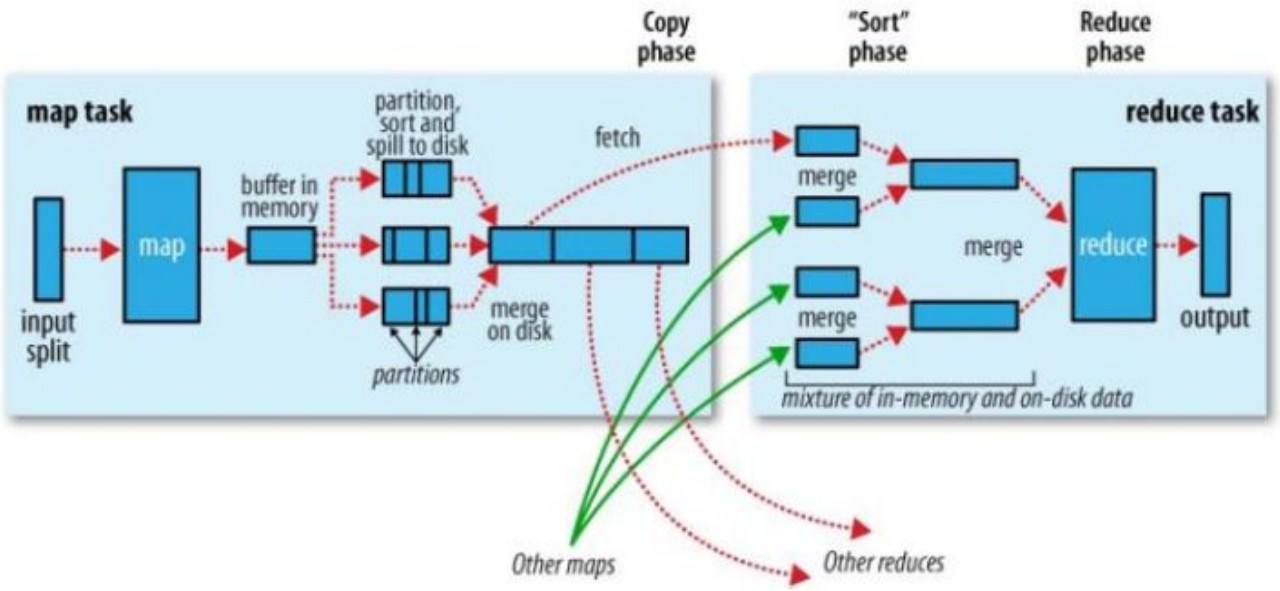
MapReduce

```
MapReduce，通过简单的Mapper和Reducer的抽象提供一个编程模型，可以在一个由几十台上百台的PC组成的不可靠集群上并发地，分布式地处理大量的数据集，而把并发、分布式（如机器间通信）和故障恢复等计算细节隐藏起来。而Mapper和Reducer的抽象，又是各种各样的复杂数据处理都可以分解为的基本元素。这样，复杂的数据处理可以分解为由多个Job（包含一个Mapper和一个Reducer）组成的有向无环图（DAG），然后每个Mapper和Reducer放到Hadoop集群上执行，就可以得出结果。
```

在MapReduce中，Shuffle是一个非常重要的过程，正是有了看不见的Shuffle过程，才可以使在MapReduce之上写数据处理的开发者完全感知不到分布式和并发的存在。

MapReduce Model:

Map，Shuffle，Reduce三个阶段：
Map阶段中，每台机器先处理本机上的数据，像图1中各个机器计算本机的文件中关键字的个数。
各个机器处理完自己的数据后，我们再把他们的结果汇总，这就是Shuffle阶段，像刚才的例子，机器A，B，C，D.....从1-n所有机器上取出Map的结果，并按关键字组合。
最后，进行最后一步处理，这就是Reduce，我们刚才的例子中就是对每一个搜索关键字统计出现总次数。



Spark

Apache Spark是一个新兴的大数据处理的引擎，主要特点是提供了一个集群的分布式内存抽象，以支持需要工作集的应用。

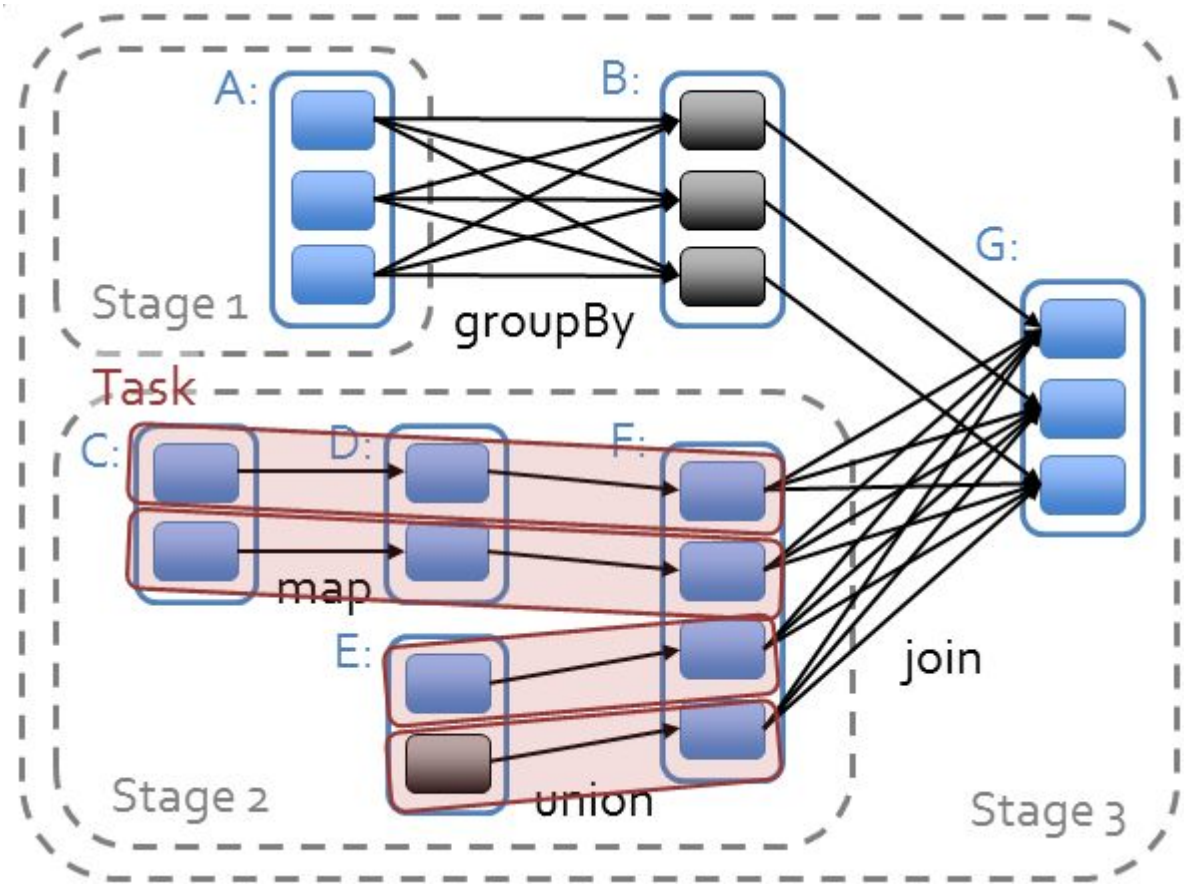
这个抽象就是RDD（Resilient Distributed Dataset），RDD就是一个不可变的带分区的记录集合，RDD也是Spark中的编程模型。Spark提供了RDD上的两类操作，转换和动作。转换是用来定义一个新的RDD，包括map, flatMap, filter, union, sample, join, groupByKey, cogroup, ReduceByKey, cros, sortByKey, mapValues等，动作是返回一个结果，包括collect, reduce, count, save, lookupKey。

Spark WordCount API

```
val spark = new SparkContext(master, appName, [sparkHome], [jars])
val file = spark.textFile("hdfs://...")
val counts = file.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```

生成的DAG

```
scala> counts.toDebugString
res0: String =
MapPartitionsRDD[7] at reduceByKey at <console>:14 (1 partitions)
  ShuffledRDD[6] at reduceByKey at <console>:14 (1 partitions)
    MapPartitionsRDD[5] at reduceByKey at <console>:14 (1 partitions)
      MappedRDD[4] at map at <console>:14 (1 partitions)
        FlatMappedRDD[3] at flatMap at <console>:14 (1 partitions)
          MappedRDD[1] at textFile at <console>:12 (1 partitions)
            HadoopRDD[0] at textFile at <console>:12 (1 partitions)
```



Spark对于有向无环图Job进行调度，确定阶段（Stage），分区（Partition），流水线（Pipeline），任务（Task）和缓存（Cache），进行优化，并在Spark集群上运行Job。RDD之间的依赖分为宽依赖（依赖多个分区）和窄依赖（只依赖一个分区），在确定阶段时，需要根据宽依赖划分阶段。根据分区划分任务。

云计算

> 云计算（cloud computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法。

服务 SAAS PAAS

硬件 电脑 带宽 底层 -SAAS

租服务 软件应用系统 平台即服务 开发系统-PAAS

上层做平台服务 多租户应用

1）存储层：

存储层是云存储最基础的部分。存储设备可以是FC光纤通道存储设备，可以是NAS和 iSCSI等IP存储设备，也可以是 SCSI或SAS等 DAS存储设备。云存储中的存储设备往往数量庞大且分布多不同地域，彼此之间通过广域网、互联网或者 FC光纤通道网络连接在一起。

存储设备之上是一个统一存储设备管理系统，可以实现存储设备的逻辑虚拟化管理、多链路冗余管理，以及硬件设备的状态监控和故障维护。

2）基础管理层：

基础管理层是云存储最核心的部分，也是云存储中最难以实现的部分。基础管理层通过集群、分布式文件系统和网格计算等技术，实现云存储中多个存储设备之间的协同工作，使多个的存储设备可以对外提供同一种服务，并提供更大更强更好的数据访问性能。

CDN内容分发系统、数据加密技术保证云存储中的数据不会被未授权的用户所访问，同时，通过各种数据备份和容灾技术和措施可以保证云存储中的数据不会丢失，保证云存储自身的安全和稳定。

3）应用接口层：

应用接口层是云存储最灵活多变的部分。不同的云存储运营单位可以根据实际业务类型，开发不同的应用服务接口，提供不同的应用服务。比如视频监控应用平台、IPTV和视频点播应用平台、网络硬盘引用平台，远程数据备份应用平台等。

4）访问层：

任何一个授权用户都可以通过标准的公用应用接口来登录云存储系统，享受云存储服务。云存储运营单位不同，云存储提供的访问类型和访问手段也不同。

大数据4V特征

> 在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》 中大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。大数据的4V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。

体量 Volume 非结构化数据的超大规模和增长

多样性 Variety 大数据的异构和多样性

价值密度 Vaule 大量的不相关信息

速度 Velocity 实时分析而非批量式分析

实时分析 数据多样性

计算速度

“大量化(Volume)、多样化(Variety)、快速化(Velocity)、价值密度低（Value）”就是“大数据”的显著特征，或者说，只有具备这些特点的数据，才是大数据。

数据

结构化 非结构化 半结构化

多样性 快速 海量

数据操作

挖掘 分析 检索 扩充

大数据产品 管理 终端