云计算总结资料

旨在帮助大家迅速抓住重点、以最高效率复习，预祝大家都能收获满满！

# 第一章：云计算技术概述

## 1.1大数据时代

### 定义：

所谓大数据就是海量数据或巨量数据

### 特点：

4v+1c（这里又出现了，想必大家都再熟悉不过）

数据量大、数据种类多、处理速度快、价值密度低、高复杂度

## 1.2云计算

### 定义：

云计算是一种商业计算模型。云计算是通过网络按需提供可动态伸缩的廉价计算服务。

### 云计算设计初衷（兴起的原因）：

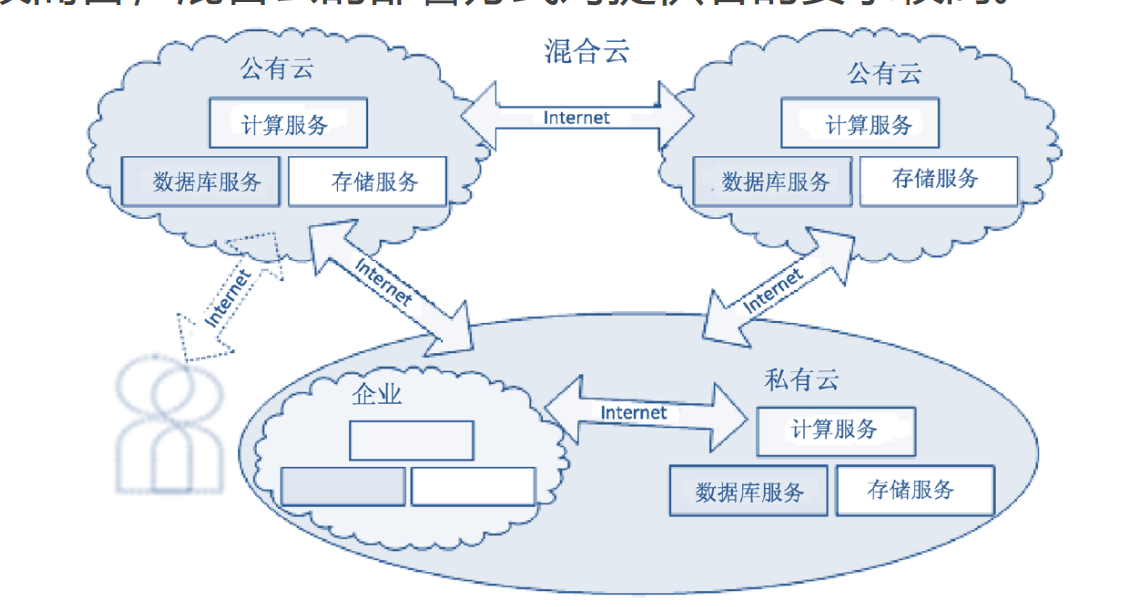
1. 个人计算机计算、存储各方面性能有限；
2. 用户不需要购买、安装、维护复杂的系统和应用软件；
3. 用户不需要担心数据的安全存储。

### 云计算的交付模式：

1. IaaS (Infrastructure as a Service，基础设施即服务)
2. PaaS（Platform as a Service，平台即服务）
3. SaaS（Software as a Service，软件即服务）

### 云计算的部署模式：

私有云、公有云、混合云



### 云计算的关键技术：

虚拟化、分布式存储、分布式计算（并行编程）、云平台管理技术等

### 云计算的优势与挑战：

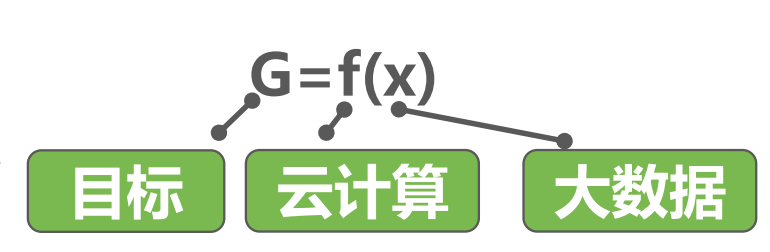
优势：

虚拟化、超高规模、高可靠性、通用性、高可伸缩性、按需服务、廉价

挑战：

服务的持续可用性、服务的安全性、服务的迁移、服务的性能

## 1.3云计算&大数据

云计算是处理大数据的手段，大数据为云计算提供数据来源。

# 第二章：虚拟化与容器技术

## 2.1虚拟化简介

### 虚拟化定义：

虚拟化就是将原本运行在真实环境上的计算机系统或组件

运行在虚拟出来的环境中



三层含义：

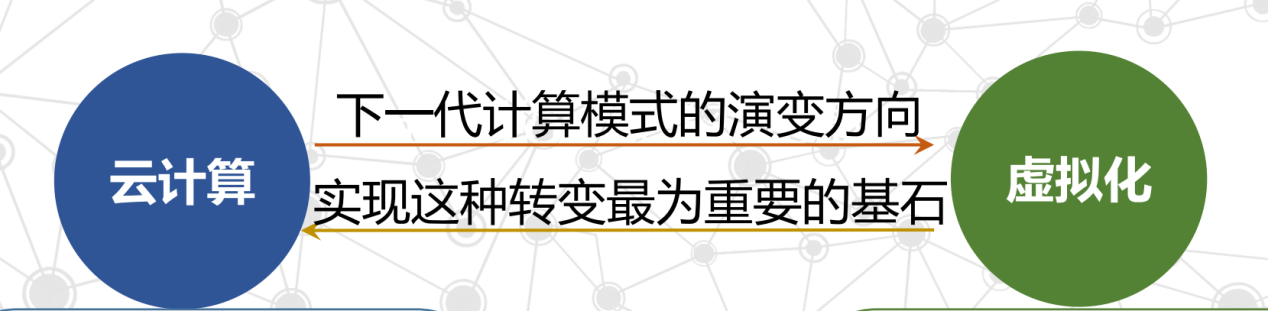
1. 虚拟化的对象是各种各样的资源

2. 经过虚拟化后的逻辑资源对用户隐藏了不必要的细节

3. 用户可以在虚拟环境中实现其在真实环境中的部分或者全

部功能

### 虚拟化&云计算



云计算必定是虚拟化的，虚拟化技术实现了硬件无差别的

封装。

## 2.2虚拟化的分类

### 2.2.1服务器虚拟化

概念：

实现操作系统与物理计算机的分离，使得在一台物理计算机上

可以同时安装和运行一个或多个虚拟的操作系统。（比如：虚拟机）

分类：（要会区分、简述两者区别）

寄居虚拟化和裸机虚拟化（原生虚拟化）







服务器虚拟化的底层实现（关键技术）：

Cpu虚拟化、内存虚拟化、IO设备虚拟化、虚拟机迁移

### 2.2.2网络虚拟化

定义：

将不同网络的硬件和软件资源结合成一个虚拟的整体。

分类：

虚拟专用网（VPN）、虚拟局域网

### 2.2.3存储虚拟化

定义：

是指将存储网络中各个分散且异构的存储设备按照一定的策

略映射成一个统一的连续编址的逻辑存储空间，称为虚拟存

储池，并将虚拟存储池的访问接口提供给应用系统。（将一个大文件存储到各个节点，用户不关心具体存到哪台物理设备，逻辑上是连续存储就行）

分类：

基于主机、基于存储设备、基于网络的存储虚拟化

### 2.2.4应用虚拟化

两层含义：

应用软件虚拟化、桌面虚拟化

## 2.3开源技术（\*）

（只需要知道有xen和kvm这么个东西就行了）

Xen

Xen 是一个开放源代码虚拟机监视器，由 剑桥大学开发。它打算在单个计算机上运 行多达100个满特征的操作系统。操作系统必须显式地修改（“移植”）以在Xen上运行。

KVM

KVM使用Linux自身的调度器进行管理，所以相对于Xen，其核源码很少，KVM的虚拟化需要硬件支持。

## 2.4Docker基础与实践

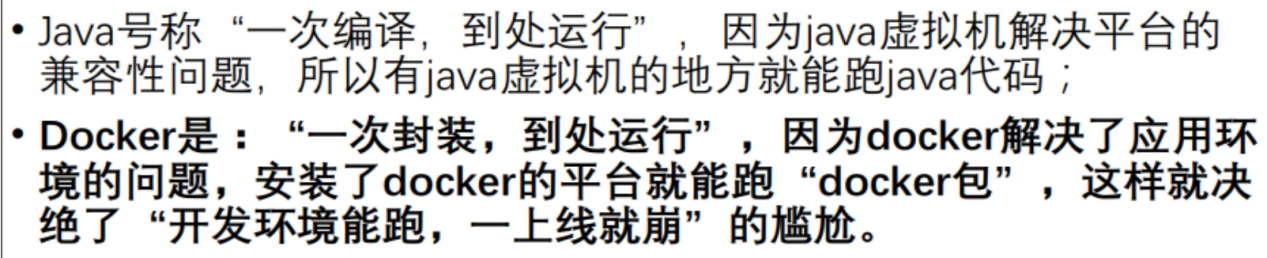
2.4.1Docker简介

Docker （本意码头工人）是一个开源的应用容器引擎，让开发

者可以打包他们的应用以及依赖包到一个可移植的容器中，然

后发布到任何 Linux 机器上。

Docker的核心思想：隔离

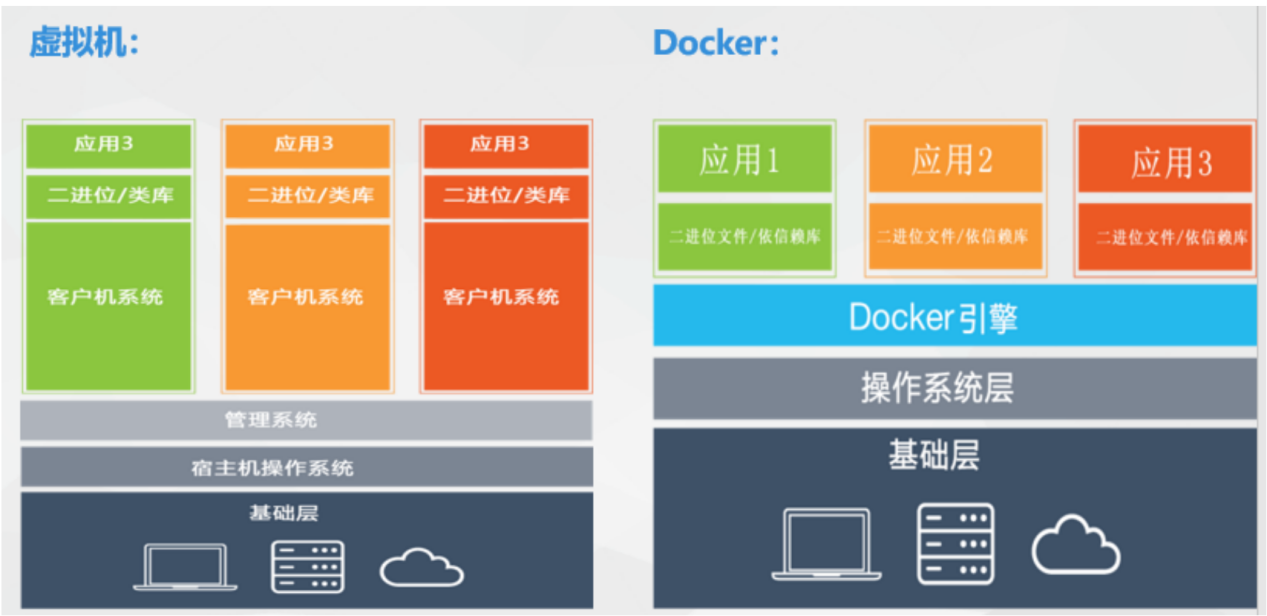


要能简述虚拟机和Docker的区别：

传统虚拟机是虚拟出一套硬件，运行一个完整的操作系统，

然后在这个系统上安装和运行软件，由Hypervisor来启动、管理客户机操作系统，耗时费力，启动慢；

Docker容器内的应用直接运行在宿主机的内核，容器没有自己的内核，通过docker守护进程管理Docker容器，比较轻便，且每个容器间是互相隔离，互不影响



正是因为Docker的明显优于传统虚拟机技术，使得Docker成为云计算技术中主流的部署方案之一。

2.4.2基本概念

Docker采用 c/s 架构

|  |  |  |  |
| --- | --- | --- | --- |
| 容器 | 对象 | 一个应用 | 类似一个简易版的linux环境 |
| 镜像 | 类 | 一个集装箱 | 包含完整的操作系统环境 |
| 仓库 | 库 | 一个超级码头 | 存放多个镜像 |

# 第三章：分布式文件系统HDFS

## 概述（\*）

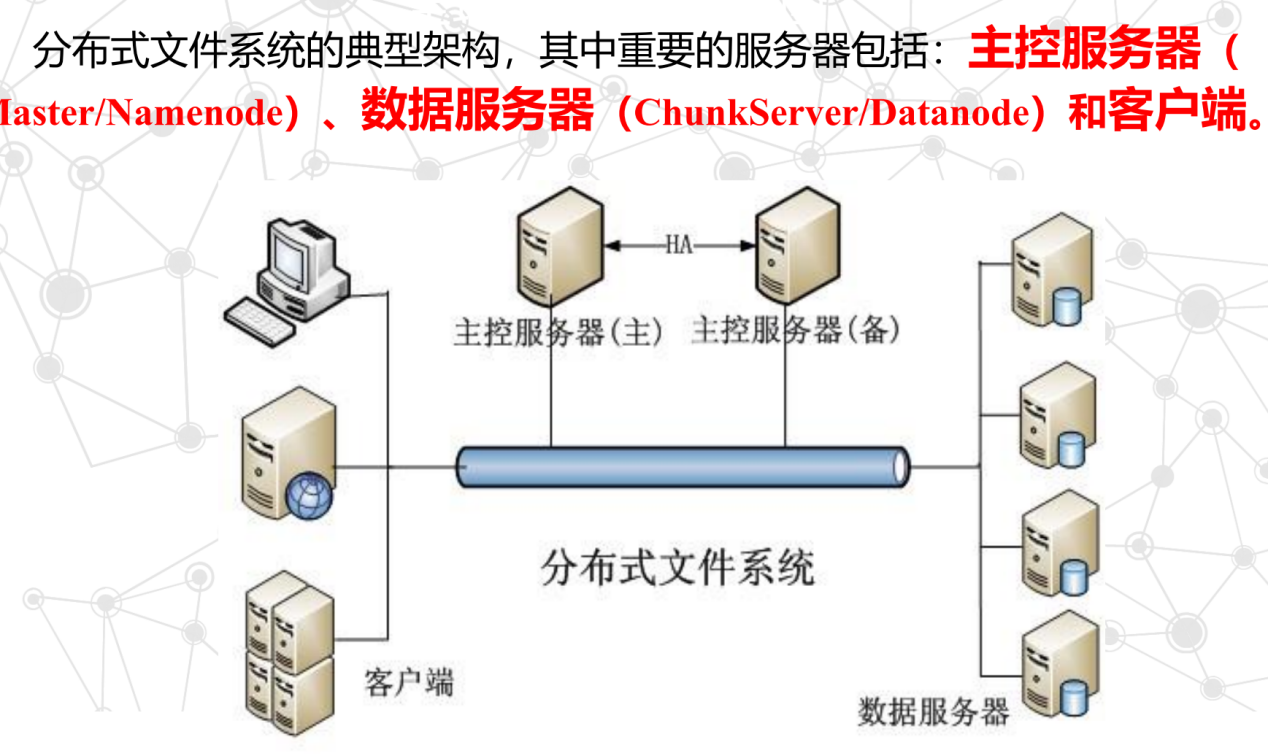
本地文件系统：用本地磁盘存储文件

分布式文件系统：将文件分布式存储在网络上的各个节点

（由于内存过大存储不下）

## 基本架构（\*）

主/从架构



## GFS

GFS（Google File System）是由Google开发并设计的一个面向大模数据处理的分布式文件系统

主/从模式

## HDFS

Hadoop Distributed File System

|  |  |  |  |
| --- | --- | --- | --- |
|  | 架构 | block大小 | 备份 |
| GFS | 主/从 | 64M | 3份 |
| HDFS | 主/从 | 128M | 3份 |

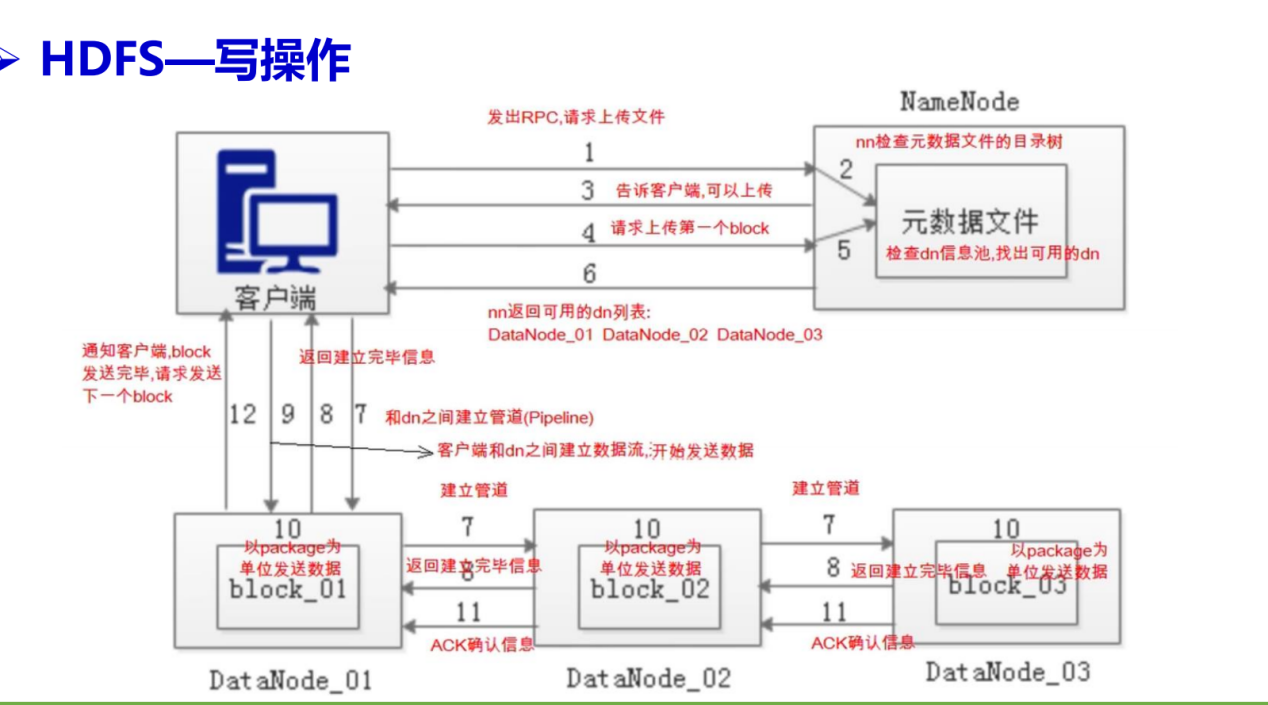
概念：

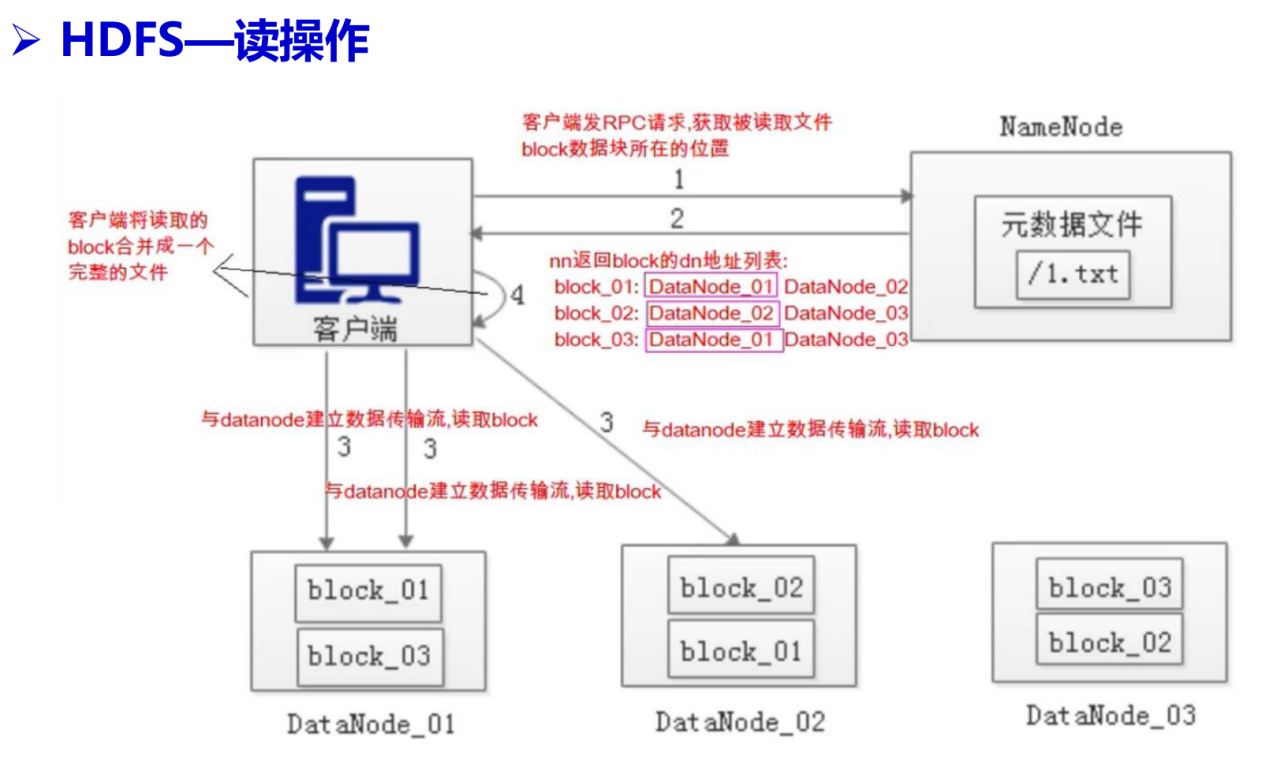
NameNode：管理文件和Block的对应关系；

DataNode：用来存放数据；

Rack：存放部署Hadoop集群的服务器

Hadoop读写过程必须掌握，要会大概叙述：





## 云存储（\*）

云存储系统由多个存储设备组成，通过集群功能、分布式文件系统或类似网格计算等功能联合起来协同工作，并通过软件或应用接口给用户提供服务

典型应用：网络磁盘、在线文档、在线网络游戏等

# 第四章：分布式计算MapReduce

## 数据密集型计算

### 概念（\*）

数据密集型计算是采用数据并行方法实现大数据量并行计算的应用，被称为大数据的核心支撑技术（TB级别的）

### 应用

充分利用数据本地性原则可提高系统吞吐量

应用于天文计算、基因工程、商业领域计算等

## 分布式数据处理（云计算的核心技术之一）

分布式数据处理的概念是建立在集中和分散这两种信息服务都能实现的原则基础上的（大数据级别）

## 并行编程模型

### 两种并行编程模型

数据并行编程模型是将相同的操作同时作用于不同的数据。

编程级别比较高，编程相对简单，但它仅适用于数据并行问

题；

消息传递编程模型是指各个并行执行的部分之间通过消息

传递来交换信息、协调步伐、控制执行。编程级别相对较低，

但灵活性和控制手段的多样性，使得消息传递编程模型有更

广泛的应用范围

### MapReduce概述

MapReduce是Hadoop核心组件之一，它是一种可用于大数

据并行处理的计算模型、框架和平台。

核心思想是“分而治之”（归并），总共有两个阶段：

✓Map阶段：负责将任务分解，即把复杂的任务分解成若干个“简单的任务”来并行处理，但前提是这些任务没有必然的依赖关系，可以单独执行任务。

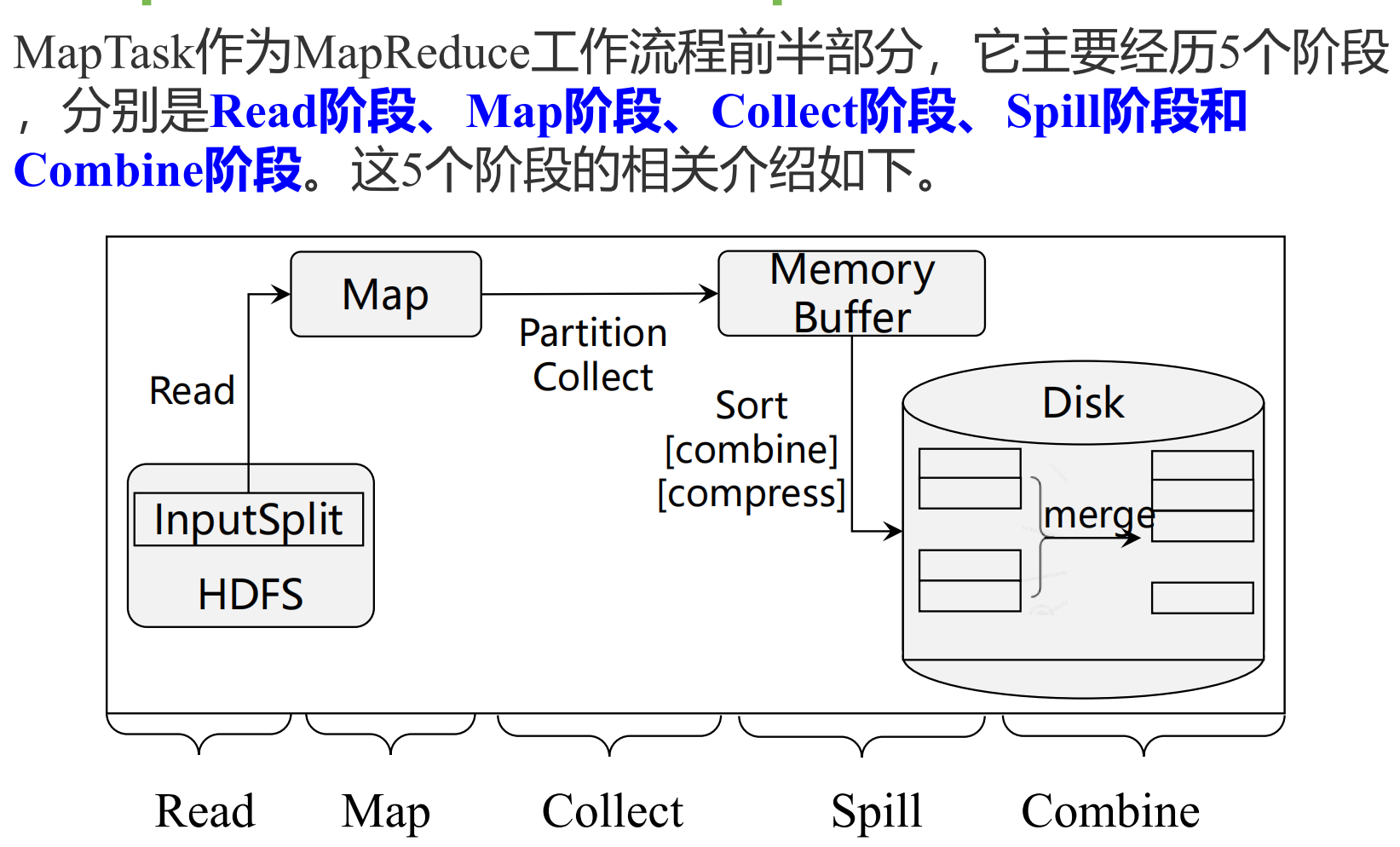
✓Reduce阶段：负责将任务合并，即把Map阶段的结果进行全局汇总。

## MapReduce工作过程（五步）

### 1.分片、格式化数据源

先将数据源文件分成大小相等的数据块，再将划分好的数据格式化为键值对，作为后续map阶段的输入（每个分区对应每个map和reduce进程）

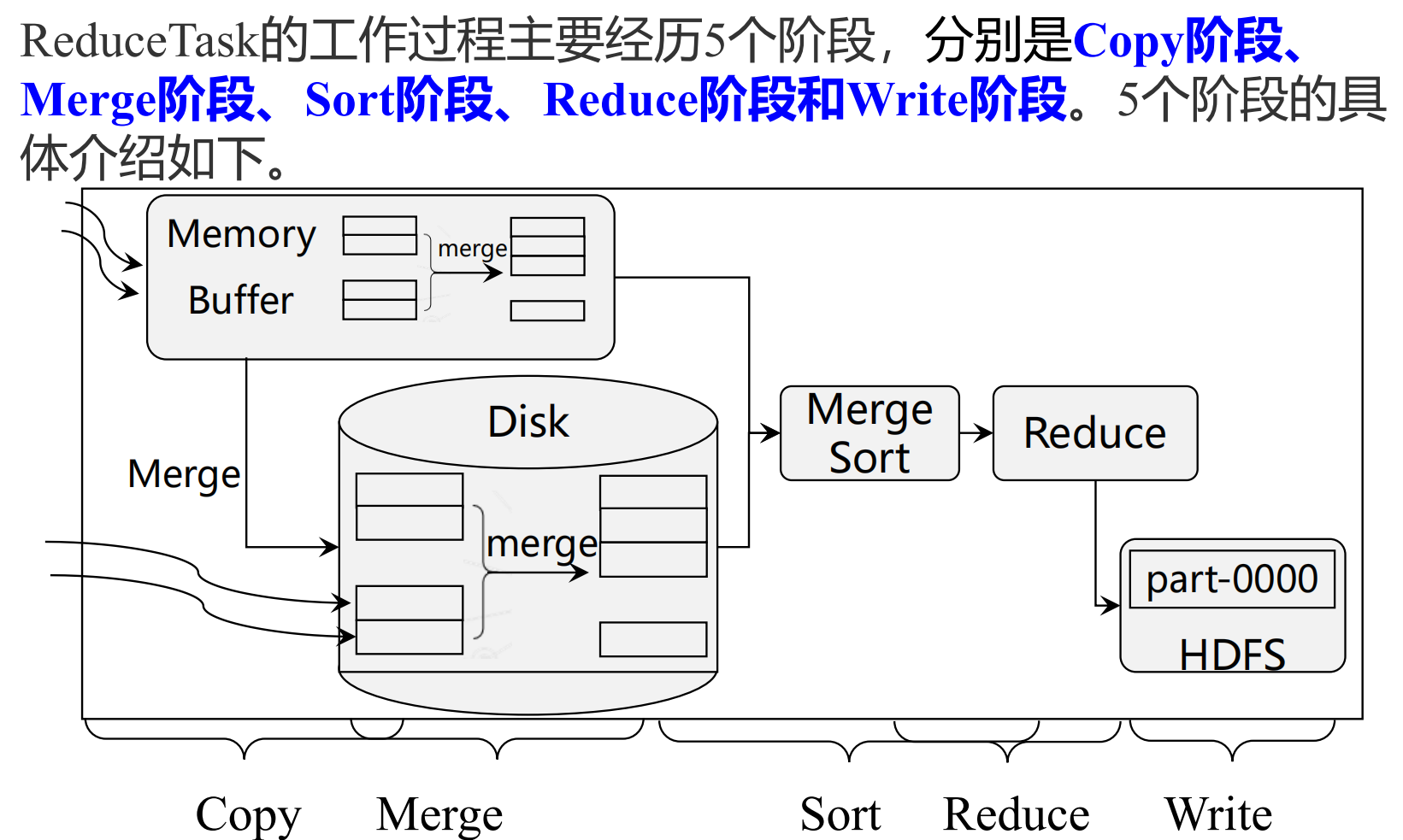
### 2.MapTask



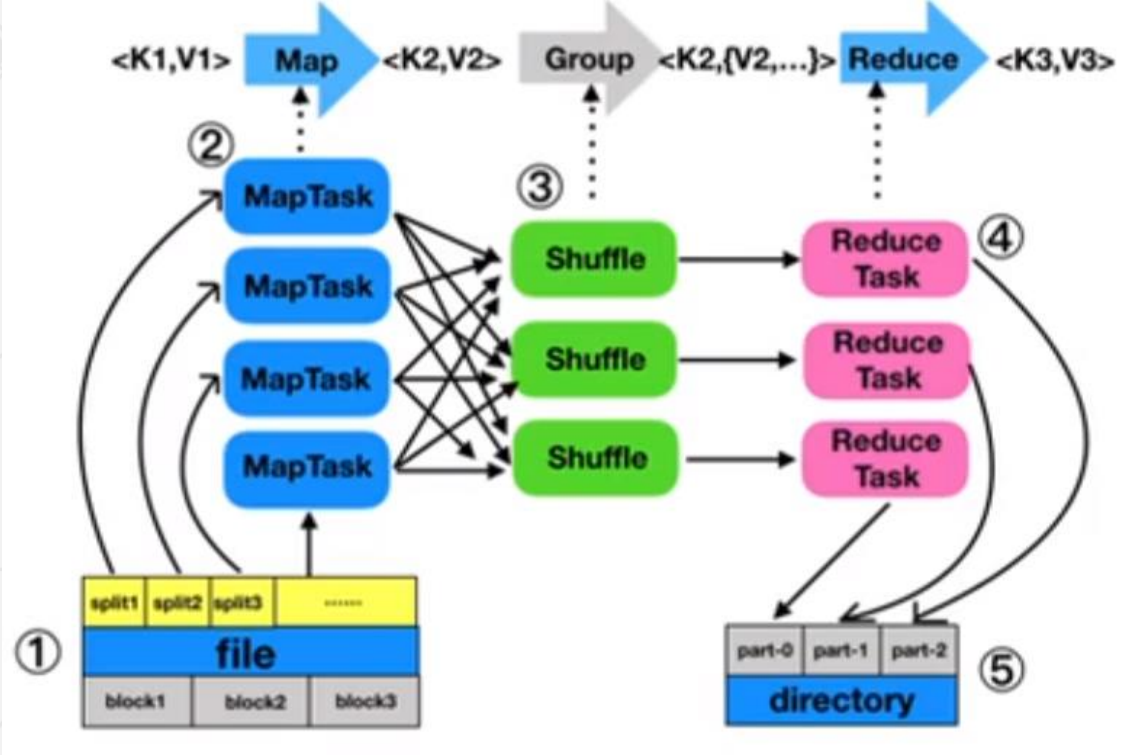
### 3.Shuffle

Shufle将MapTask输出的处理结果数据，分发给对应的ReduceTask，并在分发的过程中，对数据按key进行分区和排序。

### 4.ReduceTask



### 5.写入文件



# 第五章：分布式存储HBase

## HBase概述

### 概念：

分布式存储系统是大量普通PC服务器通过Internet互联，对

外作为一个整体提供存储服务的系统。

分布式存储系统的特性：可扩展、低成本、高性能、易使用

### 应用领域：

分布式存储系统、数据库

## NoSQL数据库

### 概念：

NoSQL（not only sql）：一切非关系型数据库的统称。特点是：易扩展、高性能、灵活、高可用

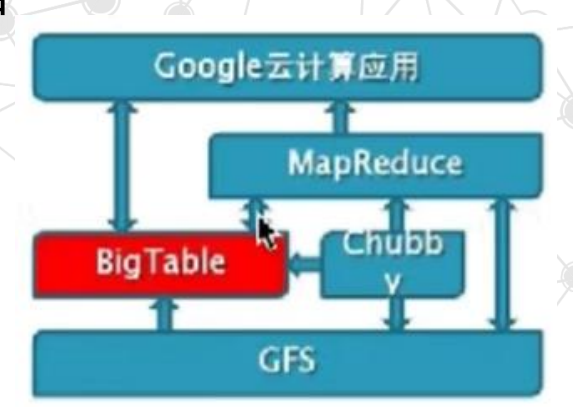
NoSQL的结构通常提供弱一致性的保证，如最终一致性（CAP定理）

### 分类

|  |  |  |
| --- | --- | --- |
| 数据库类型 | 数据模型 | 代表软件 |
| 图数据库 | 图结构 | Neo4j |
| 文档数据库 | Key对应文档 | MongoDB |
| 键值数据库 | Key-Value | Redis |
| 列祖数据库 | 以列族存储 | Cassandra |

## BigTable & HBase

### BigTable

Bigtable是Google设计的分布式结构化数据表，最初是为了存储海量的搜索数据。它常常用在mapreduce和gfs中间作为缓存，用来保存mapreduce的输入、输出。

### HBase

#### 介绍：

HBase是一个分布式的、可扩展的、面向列的、支持海量数据

存储的NoSQL数据库，是基于GFS的开源实现。

#### 对比GFS

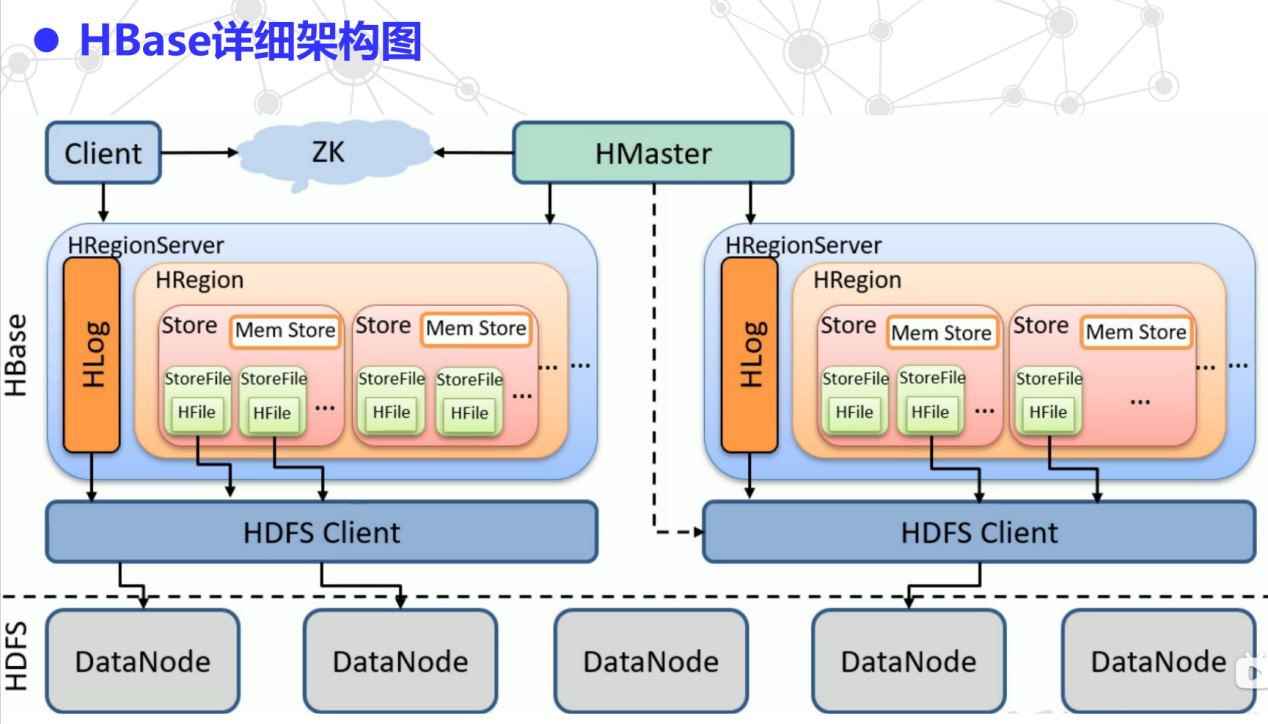
|  |  |  |
| --- | --- | --- |
| 分布式数据库 | BigTable | HBase |
| 文件存储系统 | GFS | HDFS |
| 海量数据处理 | MapReduce | Hadoop MapReduce |
| 协同服务管理 | Chubby | Zookeeper |

#### HBase框架



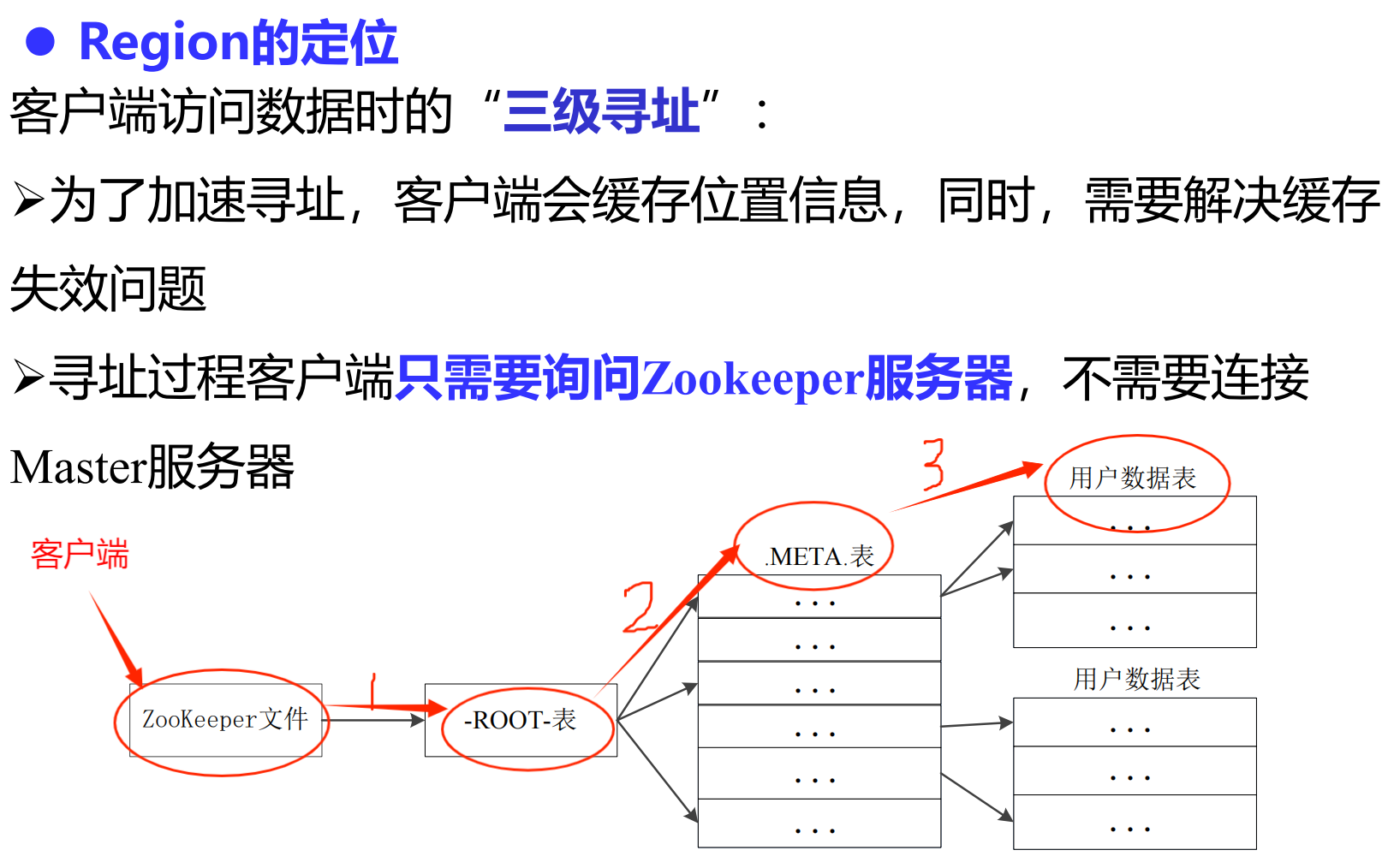
通过横向行键划分分区，通过纵向列划分列族，横纵结合划分store，最终通过思维向量{rowkey，columnFamily，column，timeStamp}准确查找到cell单元格中想要的数据

#### HBase系统架构



客户端可通过Zookeeper得到数据存储在哪个region，客户端再去访问相应的region服务器，读取数据。

其中客户端如何通过Zookeeper找到数据所在位置又有三级寻址，



# 第六章：云计算的应用（\*）

## 概述

云应用是云计算概念的子集，是云计算技术在应用层的体现。

云应用的特点：跨平台、易用、轻量

## Google云应用

Google Docs是一个基于Web的工具，它有跟Microsoft Office相

近的编辑界面，有一套简单易用的文档权限管理，而且它还记

录下所有用户对文档所做的修改。

## Amazon云应用

弹性计算云（Elastic Compute Cloud，EC2）

亚马逊AWS（Amazon Web Services (AWS) ）是亚马逊提供的是全球最全面、应用最广泛的云平台， 从全球数据中心提供超过 200 项功能齐全的服务。

## Microsoft云应用

Microsoft Azure是微软基于云计算的操作系统

## 其他

包括阿里云、华为云、腾讯云、IBM蓝云云计算等