三次信息化浪潮 标志 解决的问题

第一次信息化浪潮:1980年前后 个人计算机 信息处理

第二次信息化浪潮:1995年前后 互联网 信息传输

第三次信息化浪潮:2010年前后 大数据、云计算和物联网 信息爆炸

信息科技为大数据时代的提供了技术支撑

1. 存储设备容量不断增加
2. CPU处理能力大幅提升
3. 网络带宽不断增加

大数据发展的三个重要阶段

第一阶段:萌芽期（20世纪90年代-21世纪初）

第二阶段:成熟期（21世纪前10年）

书三阶段:大规模应用期（2010年以后）

四种范式

1.第一种范式:实验科学

2.第二种范式:理论科学

3.第三种范式:计算科学

4.第四种范式:数据密集型科学

**Hadoop生态系统**

HDFS：Hadoop分布式文件系统

HBase：分布式的列式数据库

MapReduce：分布式计算框架

Hive：一个基于Hadoop的数据仓库工具

Pig：数据流处理

Mahout：数据挖掘库

Zookeeper：分布式协作服务

Flume：日志收集

Sqoop：数据库ETL

Ambari：安装、部署、配置和管理工具

Hadoop特性

高可靠性；高效性；高可扩展性；高容错性；成本低；支持多种编程语言；运行在Linux平台上

Hadoop的基本安装配置步骤：

创建Hadoop用户

更新apt和安装Vim编辑器

安装SSH和配置SSH无密码登录

安装Java环境

安装单机Hadoop

Hadoop伪分布式安装

名称结点：负责管理分布式文件系统的命名空间，即FsImage和EditLog

数据结点：是分布式文件系统HDFS的工作结点，负责数据的存储和读取

第二名称结点：是HDFS架构的一个重要组成部分，具有两个方面的功能:首先，它可以完成EditLog 与 FsImage的合并操作,减小EditLog文件大小,缩短名称节点重启时间;其次，它可以作为名称节点的“检查点”，保存名称节点中的元数据信息。

UMP系统中的角色：

Controller服务器

Web控制台

Proxy服务器

Agent服务器

日志分析服务器

信息统计服务器

愚公系统

依赖的开发组件：Mnesia、RabbitMQ、Zookeeper、LVS

**UMP的系统功能：**

1. 容灾
2. 读写分离
3. 分库分表
4. 资源管理
5. 资源调度
6. 资源隔离
7. 数据安全

大规模数据处理包括分布式存储和分布式计算两个核心环节

大数据处理主要包括三个类型：

复杂的批量数据处理（小时级）

基于历史数据的交互式查询（分钟级、秒级）

基于实时数据流的数据处理（毫秒级、秒级）

Spark生态系统的组成及其功能：

Spark Core：包含Spark的基本功能，如内存计算、任务调度、部署模式、故障恢复、存储管理等

Spark SQL：允许开发人员直接处理RDD，同时也可查询Hive、HBase等外部数据源

Spark Streaming：支持高吞吐量、可容错处理的实时流数据处理

Structured Streaming：是一种基于Spark SQL引擎构建的、可扩展且容错的流处理引擎

MLib（机器学习）：提供了常用机器学习算法的实现，包括聚类、分类、回归、协同过滤等

GraphX:是Spark中用于图计算的API

RDD概念：是一个分布式对象集合，本质是一个只读的分区记录集合每个RDD可以分成多个分区，每个分区就是一个数据片段。RDD提供了一个高度受限的共享内存模型

RDD特性：高效的容错性、中间结果持久化到内存、存放的数据可以是Java对象

**HDFS采用抽象的块概念可以带来以下几个明显的好处：**

1. 支持大规模文件存储
2. 简化系统设计
3. 适合数据备份

**NoSQL的四大类型：**

**键值数据库**

**列族数据库**

**文档数据库**

**图数据库**

NoSQL的特点：灵活的可扩展性、灵活的数据模型、与云计算紧密融合

**NoSQL的三大基石：CAP、BASE、最终一致性**

**CAP:** C (Consistency):一致性；A( Availability):可用性；P ( Tolerance of Network Partition):分区容忍性（**最多同时满足两个**）

**BASE——一个数据库事物具有ACID四性：**A ( Atomicity): 原子性、C(Consistency): 一致性、I ( Isolation): 隔离性、D (Durability): 持久性

**BASE的基本含义：基本可用、软状态、最终一致性**

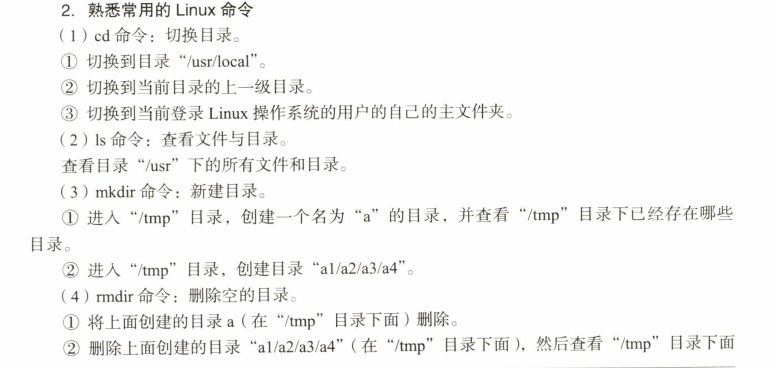
**最终一致性：：因果一致性、“读写之所写”一致性、会话一致性、单调读一致性、单调写一致性**

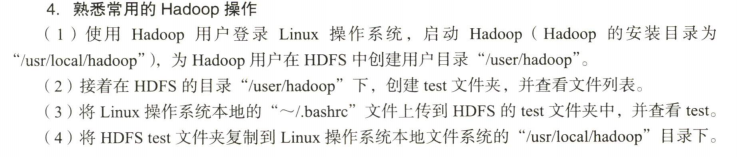
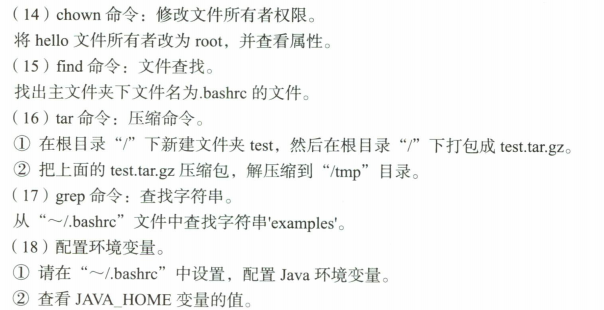
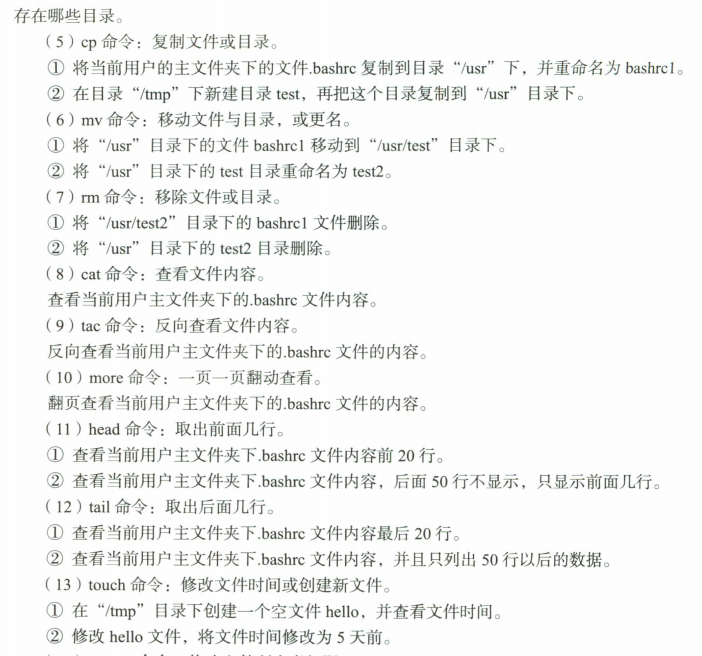
**FLink的优势**

1. 同时支持高吞吐，低延迟，高性能
2. 同时支持流处理和批处理
3. 高度灵活的流式窗口
4. 支持有状态计算
5. 具有良好的容错性
6. 具有独立的内存管理
7. 支持迭代和增量迭代

**FLink应用情景**

1. 事件驱动型应用
2. 数据分析应用
3. 数据流水线应用





HDFS常用命令

