背诵Sheet

第2章 大数据处理架构Hadoop

Hadoop为用户提供了_____的分布式基础架构

Hadoop是基于_____开发的，具有_____特性

Hadoop的核心是_____和_____

Hadoop的特性是？（四高一低二L）

Hadoop在企业中的应用架构?

Hadoop的项目结构/生态系统中各个组件?

NameNode作用是？DataNode作用是？JobTracker作用是？TaskTracker作用是？SecondaryNameNode作用是?

第3章 HDFS

HDFS要实现的目标是？局限性是？（简大跨流廉+小多低）

HDFS采用抽象的块的概念可以带来的好处是？（大备简）

HDFS主要组件的功能NameNode？DataNode?

请阐述HDFS中名称节点启动的过程?

请阐述HDFS中名称节点运行期间EditLog不断变大的问题?

Secondary NameNode是怎么做到对名称节点进行备份的?

HDFS的命名空间包含_____、_____和_____

请阐述HDFS中客户端如何使用?

HDFS体系结构中的局限性？（用隔限瓶）

HDFS中多副本冗余保存有哪些优点？（靠快查）

请阐述HDFS中数据的存放策略和数据读取策略?

HDFS是如何实现较高的容错性的？（三个出错）

第4章 分布式数据库HBase

HBase是一个_____、_____、_____、_____的分布式数据库，主要用来存储_____和_____的松散数据。

Hadoop生态系统中HBase与其他部分的关系?

HBase和BigTable的底层技术对应关系?

关系数据库已经流行多年，并且Hadoop已经有了HDFS和MapReduce，为什么还要HBase？（随实大停）

请阐述HBase与传统关系数据库有什么不同？（类型、操作、存储、索引、维护、扩展）

Master主服务器的作用是?（增删改查均调移）Region服务器的作用是？客户端是如何进行读写操作的?

请阐述Region的定位过程?

请阐述Region服务器的工作原理?

请阐述Hlog的工作原理?

请阐述当Region出错时，Hbase如何保持容错性?

HBase实际应用中如何进行性能优化?

请阐述用Coprocessor构建二级索引的原因，优点和缺点。

第5章 NoSQL简介

NoSQL数据库的特点？（云扩型）

NoSQL兴起的原因？（多并扩）

请阐述MySQL集群有哪些不足？（复扩移复）

请分析NoSQL和关系数据库之间的优势和劣势。

请阐述键值数据库的典型应用和优点、缺点。

请阐述列族数据库的典型应用和优点、缺点。

请阐述文档数据库的典型应用和优点、缺点。

请阐述图数据库的典型应用和优点、缺点。

CAP,BASE,ACID理论分别说的是?

请阐述如何实现各种类型的最终一致性?

请阐述MongoDB的主要特点。

请阐述MongoDB的组成。（数据库、文档、集合）

第6章 云数据库

云计算是什么？云计算的八大优势是什么?

云数据库的特性有哪些?

为什么说云数据库是个性化数据存储需求的理想选择?

UMP系统架构设计遵循了怎么样的原则？（用一弹弹）

请阐述UMP系统架构是如何达到容灾的效果？主从切换过程是？主库重新上线的流程是?

分库分表时，系统如何处理用户查询?

请阐述UMP系统架构如何进行资源隔离?

UMP系统架构都有哪些机制以保证数据安全?

请阐述一个典型的Hadoop作业执行时，AWS具体的操作流程。

第7章 MapReduce

Mapreduce的设计理念是_____ 。

请阐述MapReduce的体系结构:

Client的功能？JobTracker的功能？TaskTracker的功能？“Slot”？“Split”？“Task”?

请阐述MapReduce中的Shuffle过程。

请列出一些MapReduce的可能的应用场景。

第8章 Hadoop架构再探讨

Hadoop1.0的核心组件的不足

HDFS的架构改进：过去？现在？两个新框架

MapReduce的架构改进：过去？现在？一个新框架

Pig：是什么？解决了？功能有？应用场景？

Spark：是什么？解决了？缺点是？

OOzie：是什么？解决了？缺点是？

Tez：是什么？解决了？核心思想是？优化体现在？与Impala、Dremel、Drill的区别？

Kafka：是什么？解决了？

HDFS HA：目的是？架构是？

HDFS Federation：HDFS 1.0存在的问题是？能够解决？缺点是？

HDFS Federation的架构

HDFS Federation对于HDFS1.0的优势：三个优势

MapReduce1.0的缺陷：四个缺陷

YARN体系结构：总体架构？集群部署方式？

ResourceManager：四个作用？有什么用？调度器的作用？容器是什么？应用程序管理器的作用？

NodeManager:三个角色？在框架中的作用？

ApplicationMaster：两种任务？工作流程？

请阐述YARN的工作流程。

YARN与MapReduce1.0框架的对比优势：五个优势

请阐述YARN的发展目标。

“一个框架一个集群”的问题

YARN的架构优势：三个优势

Spark：为什么MapReduce无法胜任实时处理？

第9章 数据仓库分析工具Hive

数据仓库是一个_____、_____、_____、_____的数据集合。

请阐述传统数据仓库的挑战。

为什么说Hive非常适用于数据仓库?

请阐述Hive与Hadoop生态系统中其他组件的关系。（HDFS，MapReduce，Pig，HBase）

请从多个角度对Hive与传统数据库进行对比。（插入、更新、索引、分区、执行延迟、扩展性）

Hive系统都由哪些模块组成?

请阐述join转化成MapReduce任务的具体过程?

请阐述group by转换成MapReduce任务的具体过程?

请阐述Hive中SQL查询转换成MapReduce作业的过程?

请阐述Hive查询的具体执行过程?

请阐述Hive HA的原理和在报表中心上的应用流程?

Impala可以直接与_____和_____进行交互，所以可以用SQL语句查询，所以可以用于_____。Hive底层执行使用的是_____，所以主要用于_____。Impala和hive采用相同的_____、_____和_____。

请阐述Impala的系统架构:

Impalad的功能？State Store的功能？CLI的功能?

HDFS NN记录了什么?

Impalad进程主要包含_____、_____和_____三个模块与_____运行在同一个节点上完全分布运行在_____。

请阐述Impala执行查询的具体过程?

请对Impala和Hive进行比较，它们有哪些不同点，有哪些相同点?

第10章 Spark

Spark具有哪些主要特点？（快易通多)

Scala具有哪些特性？（并简兼交）

具体说说Spark是怎么优于Hadoop的?

在实际应用中大数据处理包括哪三种类型？同时部署三种软件会带来什么问题?

Spark设计遵循"____"的理念。

RDD是什么？Job是什么？Stage是什么？DAG是什么？Executor是什么?

请阐述Spark运行的基本流程。

请阐述Spark运行架构有哪些特点?

请阐述RDD的典型执行过程。这样的过程能够带来什么优点?

请阐述Spark采用RDD之后能够实现高效计算的原因。

请阐述Spark如何划分Stage?

运用Spark架构部署有什么优点?

Spark Streaming可以实现毫秒级的流计算吗？Storm呢?

为什么Spark架构不能完全取代Hadoop?

不同计算框架统一运行在YARN中有什么好处?

第11章 Storm

流数据具有什么特征？（多快大整倒）

请阐述批量计算和流计算的特征。

流计算的概念?

流计算系统应达到的要求（高海实分易靠）

传统的数据处理过程：隐含了两个前提

流计算处理过程的三个阶段：

数据实时采集：架构的三部分?

数据实时计算

实时查询服务

请阐述流处理系统与传统的数据处理系统的不同。

流计算适合哪些场合?

Storm具有哪些主要特点?

请阐述Spark架构中各个组件的特点：Streams？Spout？Bolts？Topology？Stream Groupings?

Storm和Hadoop架构组件功能对应关系?Topology?Nimbus?Supervisor?Spout/Bolt?

请阐述Storm集群是如何采用主从模式进行工作的?

请阐述Storm的工作流程。

Spark Streaming的基本原理是?

请分析对比Spark Streaming和Storm。

第12章 Flink

Flink可以同时支持____和____

Flink的主要特性有?

请阐述流处理架构。

为什么说Flink是理想的流计算框架?

Flink有什么优势?

Flink是怎么支持事件驱动型应用的?

Flink是怎么支持数据分析应用的?

Flink是怎么支持数据流水线应用的?

Flink的核心组件栈是怎么样的?

Flink是如何采用主从模式进行工作的?

第13章 图计算

传统图计算解决方案有什么不足之处?

每个超步包括哪三个组件?

采用消息传递模型主要基于什么原因?

请阐述Pregel的计算过程。这个算法什么时候可以结束如何确定?

请阐述一个Pregel用户程序的执行过程。

请阐述Pregel是如何实现容错性的?

第15章 大数据应用

请阐述常用的推荐算法。

一个推荐系统包括哪些模块?

请阐述UserCF的特点和步骤。

请阐述ItemCF的特点和步骤。

请阐述UserCF和ItemCF之间的对比。