大数据归类

hadoop

概念题

简述hadoop生态组件
mapreduce原理
对mapreduce的了解
hadoop任务调度，进程调度
mapReduce的过程
hadoop的事务怎么操作
请描述HDFS存储的机制
请详细比较Hadoop和传统SQL数据库
请用MapReduce如何实现两张表连接，有哪些方法
请描述MapReduce架构中combiner,partition作用
.reduce的数目为什么默认是一个
mapper reducer 数量如何确定
一个datanode死掉，会怎么样？如果这个datanode 之后恢复了，然后会怎么样
说一下HDFS的全称
设置map个数是在哪个配置文件里
数据倾斜，什么时候出现2次mapreduce
HA配置过程

业务题

mapreduce怎么同时读2个文件
对mapreduce进行过哪些调优
mapreduce分析top项
Map中的有3个key,1个key是另外两个key的和，如何操作
编写一个mapreduce
一般会给需求
写过什么mapreduce
用mapreduce简述一下实现最热商品（一天商品访问量）、会员用户活跃度（一天登陆次数）、会员访问时长（一天内）等多个模块之间各个模块（前10个）。
top排序（前10个）。

算法题

给10亿条记录，key好像是100个字节，value是800个字节长度，计算出前100个Top值
给定a、b两个文件，各存放50亿个url,每个url各占64字节，内存限制是4G，找出a、b文件共同的url？
在hadoop开发工程中主要用过哪些算法

HIVE

概念题

Hive有几种交互方式
Hive是怎么从本地装载数据到一个分区表中的
说下Hive中的matastore表
hive中导致数据倾斜的原因有哪些及解决方案
谈谈hive 和 hbase 的区别
Hive中内部表与外部表的区别
怎么对hive进行优化的
Hive中sql语句与MySql中sql语句的区别
ROW_NUMBER使用的场景，有没有出现什么问题，如何解决问题的
UDAF的编写
hive和hbase交互，写sql语句分析的具体过程
sortby，orderby，distinctby 区别
hive哪个版本的distinct有bug,需要改写其他形式
如何干预负载均衡

业务题

用HQL两种方法简述一下实现最热商品（一天商品访问量）、会员用户活跃度（一天登陆次数）、会员访问时长（一天内）等多个模块top前各个模块的（前10）。
文件各个模块的
UDTF解析IP地址可行？怎么进行解析的
手写sql 批量统计会话时长
Hive语句实现WordCount 假设数据存放在Hadoop下，路径为：/home/hadoop/worddata里面全是一些单词
设定一个场景，解决数据倾斜

大部分面试要求手写SQL语句处理较为复杂的业务

HBase

概念题

rowkey的设计
HBase的优化
HBase分区表的了解
mapreduce与Hbase集成
HBase读写
Hbase的表设计
为什么不直接使用Hive，还要从Hbase读取
java api与hbase集成增删改
Hbase的内部机制是啥么
HBase宕机如何处理
HBase和MongoDB在设计上的区别

业务题

HBase的rowkey设计的不合理，导致现在的Region特别大，该如何处理
HBase上的数据经过Hive处理后如何再放入HBase中
面试会问关于自己公司的表的字段数量，rowkey设计，需要结合实际

Spark

概念题

SparkSQL与Hive的区别
spark的快速计算是怎么实现的
spark如何与sql交流
Spark提交任务如何划分task
spark如何注册临时表
对spark的了解
Spark reducebykey和groupbykey的区别
spark streaming优化
sparkstreaming处理的数据来源，处理结果放在哪
RDD的了解
spark为什么比mapreduce快
Spark与Hadoop的优缺点
sparkstreaming如何和接收kafka的数据

业务题

描述spark项目的流程，数据获取来源及分析过程
写一个spark的程序（非wordcount）
spark streaming的结果给spark sql继续分析
批量计算页面时间Session
spark统计热门商品top10

SCALA

概念题

从java的map转化成scala的集合
scala的变长数组和不可变长数组
scala中的循环（1-10）
scala中string数组 组成一个整形数组
SCALA的map和flatmap

业务题

用scala写正则表达式过滤日志邮箱
scala编mapreduce

协作框架

flume如何将数据写入HDFS

flume有几种source，几种channel
zookeeper的作用
kafka的原理
sqoop的导入语句
sqoop如果遇到数据量过大如何处理

其他

公司主营业务，大数据的服务对象
集群大小，节点个数
公司大数据团队人数
使用的管理工具，出过什么问题
每日数据量多大
公司的数据表有多少字段
项目来源
日志格式
所负责的模块一天多少语句，运行多久