# Java编程

## 基础

### Jdk和Jre和JVM的区别？

### Java语言有哪些特点？

### 什么是字节码？采用字节码的最大好处是什么？

### 什么是Java程序的主类？应用程序和小程序主类有何不同？

### Java应用程序和小程序之间有哪些差别？

### Java和C++的区别？

### Oracle JDK和OpenJDK的对比？

### Java有哪些数据类型？

### Switch 是否能作用在byte上，是否能作用在long上，是否能作用在String上？

### 用最有效率的方法计算2乘以8

### Math.round(11.5)等于多少？Math.round(-11.5)等于多少？

### float f = 3.4是否正确？

### short s1 = 1; s1 = s1 + 1;有错吗?short s1 = 1; s1 += 1;有错吗？

### Java语言采用何种编码方案？有什么特点？

### 访问修饰符 public,private,protected,以及不写（默认）时的区别?

### &和&&的区别?

### Java为什么没有goto？

### final有什么用？

### final、finally和finalize的区别？

### this关键字的用法？

### this和 super的区别？

### static 详解/存在的主要意义/独特之处？

### static 应用场景？

### break，continue，return的区别和作用？

## 关键字

## String

### String是最基本的数据类型吗？

### 字符型常量和字符串常量的区别？

### 什么是字符串常量池？

### String有哪些特性？

### String为什么是不可变的？

### String为什么设计为不可变的？

### String真的是不可变的吗？

### 是否可以继承String类？

### String str = “i”和String str = new String(“1”)是一样的吗？

### String s = new String(“1”)创建了几个字符串对象？

### 如何将字符串反转？

### 数组有没有length方法？字符串有没有length方法？

### String类的常用方法都有哪些？

### 在使用HashMap时，用String做key有什么好处？

### String和StringBuffer的区别？

### StringBuilder和StringBuffer的区别？

### HashTable和HashMap、ArrayList和Vector、ArrayList和LinkedList的区别？

## 类和接口

抽象类和接口的对比，区别？

### 自动装箱和拆箱

### int和Integer的区别？

### Integer a = 127 和 Integer b = 127，a == b？

### 抽象类和接口的对比？

### 普通类和抽象类有哪些区别？

### 抽象类能用final修饰吗？

### 创建一个对象用什么关键字？

### 对象实例与对象引用有何不同？

### 成员变量和局部变量的区别有哪些？

### 在Java中定义一个不做事且没有参数的构造方法的作用？

### 一个类的构造方法的作用是什么？若一个类没有声明构造方法，该程序能正常执行吗？

### 构造方法有哪些特性？

### 静态变量和实例变量的区别？

### 静态变量和普通变量的区别？

### 静态方法和实例方法的区别？

### 静态方法内调用非静态成员为什么是非法的？

### 什么是方法的返回值？返回值的作用是什么？

### 什么是内部类？

### 内部类的分类有哪些？

### 内部类的优点？

### 内部类有哪些应用场景？

### 局部内部类和匿名内部类访问局部变量时，为什么变量必须加上final？

### 构造器（constructor）是否可被重写（override）？

### 重载（Overload）和重写（Override）的区别。重载的方法能否根据返回类型进行区分？

### == 和 equals 的区别是什么？

### HashSet如何检查重复

### 两个对象的 hashCode() 相同，则 equals() 也一定为 true，对吗？

### hashCode和equals方法的关系

### 面试官可能会问你：“你重写过 hashcode 和 equals 么，为什么重写equals时必须重写hashCode方法？”

### hashCode()介绍

### 为什么要有 hashCode

### hashCode()与equals()的相关规定？

### 对象的相等与指向他们的引用相等，两者有什么不同？

### 当一个对象被当作参数传递到一个方法后，此方法可改变这个对象的属性，并

### 可返回变化后的结果，那么这里到底是值传递还是引用传递？

### 为什么 Java 中只有值传递？

### 值传递和引用传递有什么区别？

### JDK中常用的包有哪些？

### import java和import javax有什么区别？

## 面向对象

### 面向对象和面向过程的区别？

### 面向搞对象的三大特征？

### 什么是多态？Java语言是如何实现多态的？

### 面向对象的五大基本原则是什么？

## IO流

java中IO流分几种？

BIO，NIO，AIO有什么区别？

Files的常用方法有哪些？

## 反射

什么是反射机制？

反射机制的优缺点？

反射机制的应用场景有哪些？

Java获取反射的三种方法？

# Java内存模型

# Java并发编程

## 什么是乐观锁和悲观锁？

### 什么是可重入锁？

### ReentrantLock详解？

# Linux

# JVM

## 垃圾回收

## Java是如何实现的平台无关？

## Java是编译型还是解释型？

## 什么是编译和反编译？

## JIT优化技术是什么？

## 对JDK进程执行kill -9有什么影响？

## Java中的对象一定在堆上分配内存吗？

## Jvm的堆是如何分代的？为什么分代？

## 新生代如果只有一个Eden和一个Survivor可以吗？

## YoungGC和FullGC的触发条件是什么？

## 什么是Stop The World？

## JVM有哪些垃圾回收算法？

## JVM中堆内存中各部分内存区域的大小比值？

## JVM中什么时候需要调整Survivor区占新生代的比例？

## JVM中什么时候需要调整新生代占全部堆内存的比例？

## JVM中什么时候需要调整老年代占全部堆内存的比例？

## JVM如何判断对象是否存活？

## 什么是三色标记算法？

## Java中的4种引用有什么区别？

## 新生代和老年代的GC算法？

## 新生代和老年代的垃圾回收器有何区别？

## G1和CMS有什么区别？

## 为什么G1从JDK9后成为默认的垃圾回收器？

## JDK11中新出的ZGC有什么新特点？

## Java8和Java11中的GC有什么区别？

## 类的生命周期是怎么样的？

## Java中类加载的过程是什么样的？

## 什么是双亲委派？如何破坏？

## 如何判断JVM中类和其它类是不是同一个类？

## JVM如何保证对象分配内存的过程的线程安全？

## 虚拟机中的堆一定是线程共享的吗？

## 常见的JVM工具有哪些？

## 有哪些常用的JVM启动参数？

## 哪些语言有GC机制？

## 一个对象的结构是什么样的？

## JVM是如何创建对象的？

## 字符串常量池是如何实现的？

## 什么是方法区（元空间）？是如何实现的？

## JVM中一次完整的GC流程是如何实现的？

## JVM为什么要把堆和栈区分出来？

## 运行时常量池和字符串常量池的关系是什么？

## 什么是堆外内存？如何使用堆外内存？

## FullGC多久一次算正常？

## 什么是跨代引用？有什么问题？

## 内存泄漏和内存溢出的区别是什么？

## 内存泄漏都有哪些情况？

## 破坏双亲委派之后，能重写String类吗？

## OutOfMemory和 StackOverflow的区别是什么？

## 什么是Class常量池？和运行时常量池的关系是什么？

## Java发生了OOM一定会导致JVM退出吗？

## 什么是safe point,有啥用？

## JDK1.8和1.9中类加载器有哪些不同？

## 什么是逃逸分析？

## 什么是AOT编译？和JIT有什么区别？

## 一个Java进程占用的内存有哪些部分？

## 说一说JVM的并发回收和并行回收？

## 为什么初始标记和重新标记需要STW？而并发标记不需要？

## 介绍下CMS的垃圾回收流程

## Java一定是和平台无关的吗？

## 什么是STW？有什么影响？

## 什么情况下会导致JVM退出？

## Java 类加载的过程？

# Scala编程

## 基础

### scala和java的区别？

### scala语言有什么特点？

### 什么是函数式编程？有什么优点？

### scala中的闭包是什么？

### scala中的纯函数是什么？有什么好处？

### scala中的模式匹配是什么？

### case class和class的区别？

### scala中的隐式转换是什么？

### scala中伴生类和伴生对象是什么？

### scala中的柯里化是什么？

### scala的尾递归是什么？

### 变长参数跟list作为参数的区别？

### 如何实现单例模式？

### trait特质是什么

### scala的高阶函数是什么？

### 隐式参数使用的优先顺序是什么？

### Option类型的定义和使用场景？

### Unit类型是什么？

### Scala类型系统中Nil, Null, None, Nothing四个类型的区别？

# Scala并发编程

# **Hadoop**

## **基础**

### Hadoop是什么？

### Hadoop HA的整体架构？

### Hadoop的配置文件有哪些？

### Hadoop常用命令？

文本, 信件

描述已自动生成

### MapReduce程序编写实例(word count)

### combiner 和 partition 的作用？

### 如果数据误删，如何抢救?

### 如何理解Hadoop中的数据倾斜现象？

### Hadoop2中block size默认是多大？

### Hadoop3中block size默认是多大？为什么变大了？

### Safe Mode是什么？

### Hadoop启动流程？

### Hadoop 框架中怎么来优化？

## **HDFS**

### HDFS的组织架构？

### HDFS读写流程？

### HDFS 存储的机制?

### NameNode在启动的时候会做哪些操作？

### 如何检查NameNode是否正常运行？重启NameNode的命令是什么？

### 避免NameNode故障导致集群宕机的解决方法是什么？

### HDFS在读取文件的时候，如果其中一个块突然损坏了怎么办？

### HDFS在上传文件的时候，如果其中一个DataNode突然挂掉了怎么办？

### Secondary NameNode的工作机制是怎样的？

### Secondary NameNode不能恢复NameNode的全部数据，那如何保证NameNode数据存储安全？

### 在NameNode HA中，会出现脑裂问题吗？怎么解决脑裂？

### 小文件过多会有什么危害，如何避免？

### NameNode与Secondary NameNode的区别与联系？

### Secondary NameNode的目的是什么？

### DataNode的工作机制？

### NameNode和DataNode的心跳机制?

### DataNode 在什么情况下不会备份?

### 一个 DataNode 宕机,怎么一个流程恢复？

### namenode 宕机,怎么解决？

先分析宕机后的损失，宕机后直接导致 client 无法访问，内存中的元数据丢失，但是硬盘中的元数据  
应该还存在，如果只是节点挂了，重启即可，如果是机器挂了，重启机器后看节点是否能重启，不能  
重启就要找到原因修复了。但是最终的解决方案应该是在设计集群的初期就考虑到这个问题，做  
namenode 的 HA。  
3. 一个 datanode 宕机,怎么一个流程恢复  
Datanode 宕机了后，如果是短暂的宕机，可以实现写好脚本监控，将它启动起来。如果是长时间宕机  
了，那么 datanode 上的数据应该已经被备份到其他机器了，那这台 datanode 就是一台新的 datanode  
了，删除他的所有数据文件和状态文件，重新启动。

## **MapReduce**

### MapReduce原理？

### MR中Map Task的工作机制？

### MR中Reduce Task的工作机制？

### 请说下MR中Shuffle阶段？

### Shuffle阶段的数据压缩机制？

### 在写MR时，什么情况下可以使用规约（combiner）？

### 用 MapReduce 来实现下面需求？

现在有 10 个文件夹,每个文件夹都有 1000000 个 URL，现在让你找出 top1000000url。

### Hadoop 中 Combiner 的作用?

### combiner 出现在哪个过程？

### 用 MapReduce 怎么处理数据倾斜问题？

### 我们开发 job 时，是否可以去掉 reduce 阶段？

### MapReduce 的 map 数量 和 reduce 数量怎么确定, 怎么配置?

### 一个Hadoop环境，整合了HBase和Hive，是否有必要给HDFS和HBASE都分别配置压缩策略？请给出对压缩策略的建议？

### 常见的压缩格式？

## Yarn

### YARN的集群架构？

### YARN工作原理?

### YARN的任务提交流程是怎样的?

### YARN的3种资源调度模型?

# Spark

## 基础

### 为什么Spark比Hadoop的MapReduce快？

1. Shuffle  
   Hadoop map reduce的中间数据必须落盘，spark并不是必须的，shuffle中间数据可以保留在内存中。spark的map reduce中间是宽依赖或者窄依赖，如果是窄依赖，则不需要shuffle。  
   Hadoop map reduce 的中间过程必须排序，spark的shuffle在sort merge join时需要排序，在broadcast hash join、shuffle hash join、shuffle nested loop join和Broadcast Nested Loop Join 时并不需要排序。
2. Directed Acyclic Graph (Dag， [有向无环图](https://www.zhihu.com/search?q=%E6%9C%89%E5%90%91%E6%97%A0%E7%8E%AF%E5%9B%BE&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2977285638%7D))  
   每一次hadoop map reduce的map开始必须从磁盘读取，reduce结果必须写入磁盘。  
   对于spark来说，多个map reduce过程被合并成一个dag，每个stage的数据读取不一定需要从磁盘中读取，除最后的运行结果不一定要保存到磁盘。
3. cache  
   Spark支持将需要反复用到的数据给Cache到内存中，减少数据加载耗时，所以Spark跑机器学习算法比较好（需要对数据进行反复迭代）。
4. 调度方式  
   在不断启停新旧task时候，进程的启动和停止需要做很多初始化工作，因此采用进程的方式会降低执行效率。**当一个机器上多个map task需要读取同一份字典来进行数据过滤或者提取的时候，需要将字典加载到每个map task进程中，会造成资源重复加载，存在浪费内存资源的问题。**Spark通过复用线程池中的线程来减少启动、关闭task所需要的开销。
5. 增加并行度  
   由于将中间结果写到磁盘与从磁盘读取中间结果属于不同的环节，Hadoop将它们简单地通过串行执行衔接起来。Spark把不同的环节抽象为Stage，允许多个Stage既可以串行执行，又可以并行执行。
6. 避免重新计算  
   当Stage中某个分区的Task执行失败后，会重新对此Stage调度，但在重新调度的时候会过滤已经执行成功的分区任务，所以不会造成重复计算和资源浪费。
7. 灵活的内存管理策略  
   Spark将内存分为**堆上的存储内存**、**堆外的存储内存**、**堆上的执行内存**、**堆外的执行内存**4个部分。Spark既提供了**执行内存和存储内存之间固定边界的实现**，**又提供了执行内存和存储内存之间“软”边界的实现**。Spark默认使用“软”边界的实现，执行内存或存储内存中的任意一方在资源不足时都可以借用另一方的内存，最大限度地提高资源的利用率，减少对资源的浪费。Spark由于对内存使用的偏好，内存资源的多寡和使用率就显得尤为重要，为此Spark的内存管理器提供的Tungsten实现了一种与操作系统的内存Page非常相似的数据结构，用于直接操作操作系统内存，节省了创建的Java对象在堆中占用的内存，使得Spark对内存的使用效率更加接近硬件。Spark会给每个Task分配一个配套的任务内存管理器，对Task粒度的内存进行管理。Task的内存可以被多个内部的消费者消费，任务内存管理器对每个消费者进行Task内存的分配与管理，因此Spark对内存有着更细粒度的管理（大概说了一下）。  
   Tungsten Project通过使用更加紧凑和高效的基于列存储的数据处理方式以及支持多种高效的数据结构，对Spark进行了数据结构上的优化，从而提高了Spark的内存处理效率。
8. 检查点支持  
   Spark的RDD之间维护了血缘关系（lineage），一旦某个RDD失败了，则可以由父RDD重建。虽然lineage可用于错误后RDD的恢复，但对于很长的lineage来说，恢复过程非常耗时。如果应用启用了检查点，那么在Stage中的Task都执行成功后，SparkContext将把RDD计算的结果保存到检查点，这样当某个RDD执行失败后，再由父RDD重建时就不需要重新计算，而直接从检查点恢复数据。
9. 更多的内置函数  
   更多来自底层的优化。
10. 列裁剪（大概说一下，了解它的功能）  
    列裁剪（Column Pruning）是 Spark SQL 中一种常用的查询优化技术，它可以从查询中删除不必要的列，只获取必要的数据列，从而减少数据传输和处理的时间和资源，提高查询效率。列裁剪主要是通过过滤掉查询中不需要的列，使得扫描的数据量减少。如果数据存储格式为列式存储（如 Parquet 格式），列裁剪可以进一步映射下推，将列过滤操作下推到数据扫描器中执行，从而进一步减少IO操作。
11. 谓词下推（大概说一下，了解它的功能）  
    谓词下推是 Spark SQL 查询优化的一种技术，它可以将过滤条件尽可能地下推到最底层执行，以减少每一层交互的数据量，从而提升性能。具体来说，谓词下推可以在查询计划中自动将过滤条件下推到关系表的扫描器中，在扫描器中进行数据过滤，减少数据在网络中的传输量和磁盘IO操作，从而实现加速查询的效果。
12. Whole-Stage Code Generation（大概说一下，了解它的功能）  
    Spark中的全阶段代码生成是一种针对SQL和DataFrame等高级API的优化技术，它能够将一整个物理执行计划中的多个操作融合到一起，产生高效的本地机器代码执行（大概说说一下，细节还不懂）。

具体来说，全阶段代码生成通过以下三个步骤实现：

(1) 将物理执行计划中的多个操作合并为一个大的操作。

(2) 使用代码生成技术把这个大操作转换成本地机器代码，然后将代码编译成Java Class对象

(3) 在运行时直接调用生成的Java Class对象对数据进行处理，从而避免了反复地序列化/反序列 化数据, 避免了虚拟函数调 用带来的性能损失。

由于全阶段代码生成技术能够减少序列化和反序列化数据的次数，并且直接生成可执行的本地机器代码，因此它能够极大地提升Spark高级API的计算效率。目前，全阶段代码生成技术已经成为Spark中一个非常重要的优化技术。

13. Hadoop YARN中的ApplicationMaster申请到Container后，具体任务需要利用NodeManager从HDFS的不同节点下载任务所需的资源（如Jar包），这也增加了磁盘I/O。Spark将应用程序上传的资源文件缓冲到Driver本地文件服务的内存中，当Executor执行任务时直接从Driver的内存中读取，也节省了大量的磁盘I/O。

### Spark常用函数/算子?

### 什么是宽依赖和窄依赖？



# Flink

## 基础

### Flink常用函数？

# Kafka

## 基础

### Kafka常用函数？

# Hive

## 基础

### Hive创建表的语句？

### hive 内部表和外部表的区别？

### 生产环境中为什么建议使用外部表？

### Hive常用函数？

### Redis,传统数据库,HBase, Hive 每个之间的区别?

### Hive底层可以使用什么计算引擎？

### hive 有哪些保存元数据的方式，各有什么特点？

### 简述Hive中的虚拟列的作用是什么？使用它的注意事项

# Elasticsearch

# Zookeeper

# CDH-6.3.2

# HBase

## 基础

### HBase 的 row key 怎么创建好？列族怎么创建比较好？

### HBase数据库对行键的设计要求是什么？

### 简述HBase Filter的实现原理是什么？结合实际项目经验，写出几个使用Filter的案例

### 简述HBase性能优化思路

### ROWKEY的后缀匹配怎么实现？例如ROWKEY是yyyyMMDD-UserID形式，如以UserID为条件查询数据，怎样实现？

# Clickhouse

# Redis

## AOF是什么？

## RDB是什么？

## 主从同步？

# MySQL

## 基础

### Mysql存储引擎有哪些？

### MyISAM和Innodb的区别？

### 数据库事务的四个特性及含义？（ACID的含义？）

### 视图的作用，视图可以更改么？

### drop, delete 与 truncate 的区别？

### 索引的工作原理及其种类？

### 连接的种类？

# MongoDB

# Maven

# ELK

# Flume

# Sqoop

# Kerberos