作者：蓦\_然 链接：<https://www.nowcoder.com/discuss/929016?page=1>

##

### Hadoop（约5.5w字）

#### Hadoop基础

1. **介绍下Hadoop**

HDFS、MAPREDUCE、YARN

1. **Hadoop的特点**
2. **说下Hadoop生态圈组件及其作用**
3. **Hadoop主要分哪几个部分?他们有什么作用?**
4. **Hadoop 1.x，2x，3.x的区别**
5. **Hadoop集群工作时启动哪些进程?它们有什么作用?**
6. **在集群计算的时候，什么是集群的主要瓶颈**
7. **搭建Hadoop集群的xml文件有哪些?**
8. **Hadoop的checkpoint流程**
9. **Hadoop的默认块大小是多少?为什么要设置这么大?**
10. **Block划分的原因**
11. **Hadoop常见的压缩算法?**
12. **Hadoop作业提交到YARN的流程?**
13. **Hadoop的Combiner的作用**
14. **Hadoop序列化和反序列化**
15. **Hadoop的运行模式**
16. **Hadoop小文件处理问题**
17. **Hadoop为什么要从2.x升级到3.x?**
18. **Hadoop的优缺点**

#### HDFS部分

1. **HDFS文件写入和读取流程**
2. **HDFS组成架构**
3. **介绍下HDFS，说下HDFS优缺点，以及使用场景**
4. **HDFS作用**
5. **HDFS的容错机制**
6. **HDFS的存储机制**
7. **HDFS的副本机制**
8. **HDFS的常见数据格式，列式存储格式和行存储格式异同点，列式存储优点有哪些?**
9. **HDFS如何保证数据不丢失?**
10. **HDFS NameNode高可用如何实现?需要哪些角色?**
11. **HDFS的文件结构?**
12. **HDFS的默认副本数?为什么是这个数量?如果想修改副本数怎么修改?**
13. **介绍下HDFS的Block**
14. **HDFS的块默认大小，64M和128M是在哪个版本更换的?怎么修改默认块大小?**
15. **HDFS的block为什么是128M?增大或减小有什么影响?**
16. **HDFS HA怎么实现?是个什么架构?**
17. **导入大文件到HDFS时如何自定义分片?**
18. **HDFS的mapper和reducer的个数如何确定?reducer的个数依据是什么?**
19. **HDSF通过那个中间组件去存储数据**
20. **HDFS跨节点怎么进行数据迁移**
21. **HDFS的数据-致性靠什么保证?**
22. **HDFS怎么保证数据安全**
23. **HDFS中向DataNode写数据失败了怎么办**
24. **Hadoop2.xHDFS快照**
25. **HDFS文件存储的方式?**
26. **HDFS写数据过程，写的过程中有哪些故障，分别会怎么处理?**
27. **NameNode存数据吗?**
28. **使用NameNode的好处**
29. **HDFS中DataNode怎么存储数据的**
30. **直接将数据文件上传到HDFS的表目录中，如何在表中查询到该数据?**

#### MapReduce部分

1. **介绍下MapReduce**
2. **MapReduce优缺点**
3. **MapReduce架构**
4. **MapReduce工作原理**
5. **MapReduce哪个阶段最费时间**
6. **MapReduce中的Combine是干嘛的?有什么好外?**
7. **MapReduce为什么一定要有环型缓冲区**
8. **MapReduce为什么一定要有Shuffle过程**
9. **MapReduce的Shuffle过程及其优化**
10. **Reduce怎么知道去哪里拉Map结果集?**
11. **Reduce阶段都发生了什么，有没有进行分组**
12. **MapReduce Shuffle的排序算法**
13. **shuffle为什么要排序?**
14. **说一下map是怎么到reduce的?**
15. **说一下你了解的用哪几种shuffle机制?**
16. **MapReduce的数据处理过程**
17. **mapjoin的原理(实现)?应用场景?**
18. **reducejoin如何执行(原理)**
19. **MapReduce为什么不能产生过多小文件**
20. **MapReduce分区及作用**
21. **ReduceTask数量和分区数量关系**
22. **Map的分片有多大**
23. **MapReduce join两个表的流程?**
24. **手撕一段简单的MapReduce程序**
25. **reduce任务什么时候开始?**
26. **MapReduce的reduce使用的是什么排序?**
27. **MapReduce怎么确定MapTask的数量?**
28. **Map数量由什么决定**
29. **MapReduce的map进程和reducer进程的ivm垃圾回收器怎么选择可以提高吞吐量?**
30. **MapReduce的task数目划分**
31. **MapReduce作业执行的过程中，中间的数据会存在什么地方?不会存在内存中么?**
32. **Mapper端进行combiner之后，除了速度会提升，那从Mapper端到Reduece端的数据量会怎么变?**
33. **map输出的数据如何超出它的小文件内存之后，是落地到磁盘还是落地到HDFS中?**
34. **Map到Reduce默认的分区机制是什么?**
35. **结合wordcount述说MapReduce，具体各个流程，map怎么做，reduce怎么做**
36. **MapReduce数据倾斜产生的原因及其解决方案**
37. **Map Join为什么能解决数据倾斜**
38. **MapReduce运行过程中会发生OOM，OOM发生的位置？**
39. **MapReduce用了几次排序，分别是什么？**
40. **MapReduce压缩方式**
41. **MapReduce中怎么处理一个大文件**

#### YARN部分

1. **介绍下YARN**
2. **YARN有几个模块**
3. **YARN工作机制**
4. **YARN有什么优势，能解决什么问题?**
5. **YARN容错机制**
6. **YARN高可用**
7. **YARN调度器**
8. **YARN中Container是如何启动的?**
9. **YARN的改进之处，Hadoop3.x相对于Hadoop 2.x?**
10. **YARN监控**

### Zookeeper（约2.6w字）

1. **介绍下Zookeeper是什么?**
2. **Zookeeper有什么作用?优缺点?有什么应用场景?**
3. **Zookeeper的选举策略，leader和follower的区别?**
4. **介绍下Zookeeper选举算法**
5. **Zookeeper的节点类型有哪些?分别作用是什么?**
6. **Zookeeper的节点数怎么设置比较好?**
7. **Zookeeper架构**
8. **Zookeeper的功能有哪些**
9. **Zookeeper的数据结构(树)?基于它实现的分布式锁?基于它实现的Master选举?基于它的集群管理? Zookeeper的注册(watch)机制使用场景?**
10. **介绍下Zookeeper消息的发布订阅功能**
11. **Zookeeper的分布式锁实现方式?**
12. **Zookeeper怎么保证一致性的**
13. **Zookeeper的zab协议(原子广播协议)?**
14. **ZAB是以什么算法为基础的?ZAB流程?**
15. **Zookeeper的通知机制**
16. **Zookeeper脑裂问题**
17. **Zookeeper的Paxos算法**
18. **Zookeeper的协议有哪些?**
19. **Zookeeper如何保证数据的一致性?**
20. **Zookeeper的数据存储在什么地方?**
21. **Zookeeper从三台扩容到七台怎么做?**

### Hive（约3.3w字）

1. **说下为什么要使用Hive?Hive的优缺点?Hive的作用是什么?**
2. **说下Hive是什么?跟数据仓库区别?**
3. **Hive架构**
4. **Hive内部表和外部表的区别?**
5. **为什么内部表的删除，就会将数据全部删除，而外部表只删除表结构?为什么用外部表更好?**
6. **Hive建表语句?创建表时使用什么分隔符?**
7. **Hive删除语句外部表删除的是什么?**
8. **Hive数据倾斜以及解决方案**
9. **Hive如果不用参数调优，在map和reduce端应该做什么**
10. **Hive的用户自定义函数实现步骤与流程**
11. **Hive的三种自定义函数是什么?实现步骤与流程?它们之间的区别?作用是什么?**
12. **Hive的cluster by、sort bydistribute by、orderby区别?**
13. **Hive分区和分桶的区别**
14. **Hive的执行流程**
15. **Hive SQL转化为MR的过程?**
16. **Hive SQL优化处理**
17. **Hive的存储引擎和计算引擎**
18. **Hive的文件存储格式都有哪些**
19. **Hive中如何调整Mapper和Reducer的数目**
20. **介绍下知道的Hive窗口函数，举一些例子**
21. **Hive的count的用法**
22. **Hive的union和unionall的区别**
23. **Hive的join操作原理，leftjoin、right join、inner join、outer join的异同?**
24. **Hive如何优化join操作**
25. **Hive的mapjoin**
26. **Hive语句的运行机制，例如包含where、having、group by、orderby，整个的执行过程?**
27. **Hive使用的时候会将数据同步到HDFS，小文件问题怎么解决的?**
28. **Hive Shuffle的具体过程**
29. **Hive有哪些保存元数据的方式，都有什么特点?**
30. **Hive SOL实现查询用户连续登陆，讲讲思路**
31. **Hive的开窗函数有哪些**
32. **Hive存储数据吗**
33. **Hive的SOL转换为MapReduce的过程?**
34. **Hive的函数:UDF、UDAF、UDTF的区别?**
35. **UDF是怎么在Hive里执行的**
36. **Hive优化**
37. **row\_number，rank，dense\_rank的区别**
38. **Hive count(distinct)有几个reduce，海量数据会有什么问题**
39. **HQL：行转列、列转行**
40. **一条HQL从代码到执行的过程**
41. **了解Hive SQL吗？讲讲分析函数？**
42. **分析函数中加Order By和不加Order By的区别？**
43. **Hive优化方法**
44. **Hive里metastore是干嘛的**
45. **HiveServer2是什么？**
46. **Hive表字段换类型怎么办**
47. **parquet文件优势**

### Flume（约0.5w字）

1. **介绍下Flume**
2. **Flume架构**
3. **Flume有哪些Source**
4. **说下Flume事务机制**
5. **介绍下Flume采集数据的原理？底层实现？**
6. **Flume如何保证数据的可靠性**
7. **Flume传输数据时如何保证数据一致性（可靠性）**
8. **Flume拦截器**
9. **如何监控消费型Flume的消费情况**
10. **Kafka和Flume是如何对接的？**
11. **为什么要使用Flume进行数据采集**

### Kafka（约5.5w字）

1. **介绍下Kafka，Kafka的作用?Kafka的组件?适用场景?**
2. **Kafka作为消息队列，它可解决什么样的问题?**
3. **说下Kafka架构**
4. **说下Kafka的特点，优缺点**
5. **Kafka相比于其它消息组件有什么好处?**
6. **Kafka生产者与消费者**
7. **Kafka分区容错性**
8. **Kafka的消费端的数据一致性**
9. **Kafka的leader挂掉之后处理方法**
10. **说下Kafka的ISR机制**
11. **Kafka的选举机制**
12. **Kafka的ISR、OSR和ACK介绍，ACK分别有几种值?**
13. **Kafka的工作原理?**
14. **Kafka怎么保证数据不丢失，不重复?**
15. **Kafka分区策略**
16. **Kafka如何尽可能保证数据可靠性?**
17. **Kafka数据丢失怎么处理?**
18. **Kafka如何保证全局有序?**
19. **牛产者消费者模式与发布订阅模式有何异同?**
20. **Kafka的消费者组是如何消费数据的**
21. **Kafka的offset管理**
22. **Kafka为什么同一个消费者组的消费者不能消费相同的分区?**
23. **如果有一条offset对应的数据，消费完成之后，手动提交失败，如何处理?**
24. **正在消费一条数据，Kafka挂了，重启以后，消费的offset是哪一个**
25. **Kafka支持什么语义，怎么实现ExactlyOnce?**
26. **Kafka的消费者和消费者组有什么区别?为什么需要消费者组?**
27. **Kafka producer的写入数据过程?**
28. **Kafka producer的ack设署**
29. **Kafka的ack机制，解决了什么问题?**
30. **Kafka读取消息是推还是拉的模式?有什么好?**
31. **Kafka如何实现高吞吐的原理?**
32. **说下Kafka中的Partition?**
33. **Kafka是如何进行数据备份的?**
34. **Kafka里面存的数据格式是什么样的?**
35. **Kafka是如何清理过期文件的?**
36. **Kafka的一条message中包含了哪些信息?**
37. **Kafka如何保证数据的ExactlyOnce?**
38. **Kafka消费者怎么保证ExactlyOnce**
39. **Kafka监控实现?**
40. **Kafka中的数据能彻底删除吗?**
41. **Kafka复制机制?**
42. **Kafka分区多副本机制?**
43. **Kafka分区分配算法**
44. **Kafka蓄水池机制**
45. **Kafka如何实现享等性?**
46. **Kafka的offset存在哪?**
47. **Kafka中如何保证数据一致性?**
48. **Kafka新旧API区别**
49. **Kafka消息在磁盘上的组织方式**
50. **Kafka在哪些地方会有选举过程，使用什么工具支持选举?**
51. **Kafka搭建过程要配置什么参数?**
52. **Kafka的单播和多播**
53. **Kafka的高水位和Leader Epoch**
54. **Kafka的分区器、拦截器、序列化器?**
55. **Kafka连接Spark Streaming的几种方式**
56. **Kafka的生成者客户端有几个线程?**
57. **Kafka怎么防止脑裂**
58. **Kafka高可用体现在哪里**
59. **Zookeeper在Kafka的作用**

### HBase（约2.8w字）

1. **介绍下HBase**
2. **HBase优缺点**
3. **说下HBase原理**
4. **介绍下HBase架构**
5. **HBase读写数据流程**
6. **HBase的读写缓存**
7. **在删除HBase中的一个数据的时候，它什么时候真正的进行删除呢?当你进行删除操作，它是立马就把数据删除掉了吗?**
8. **HBase中的二级索引**
9. **HBase的RegionServer宕机以后怎么恢复的?**
10. **HBase的一个region由哪些东西组成?**
11. **HBase高可用怎么实现的?**
12. **为什么HBase适合写多读少业务?**
13. **列式数据库的适用场景和优势?列式存储的特点?**
14. **HBase的rowkey设计原则**
15. **HBase的rowkey为什么不能超过一定的长度?为什么要唯一?rowkey太长会影响Hfile的存储是吧?**
16. **HBase的RowKey设置讲究有什么原因**
17. **HBase的大合并、小合并是什么?**
18. **HBase和关系型数据库(传统数据库)的区别(优点)?**
19. **HBase数据结构**
20. **HBase为什么随机查询很快?**
21. **HBase的LSM结构**
22. **HBase的Get和Scan的区别和联系?**
23. **HBase数据的存储结构(底层存储结构)**
24. **HBase数据compact流程?**
25. **HBase的预分区**
26. **HBase的热点问题**
27. **HBase的memstore冲刷条件**
28. **HBase的MVCC**
29. **HBase的大合并与小合并，大合并是如何做的?为什么要大合并**
30. **既然HBase底层数据是存储在HDFS上，为什么不直接使用HDFS，而还要用HBase**
31. **HBase和Phoenix的区别**
32. **HBase支持SQL操作吗**
33. **HBase适合读多写少还是写多读少**
34. **HBase表设计**
35. **Region分配**
36. **HBase的Region切分**

### Spark（约9.8w字）

1. **Spark的任务执行流程**
2. **Spark的作业运行流程是怎么样的?**
3. **Spark的特点**
4. **Spark源码中的任务调度**
5. **Spark作业调度**
6. **Spark的架构**
7. **Spark的使用场景**
8. **Spark on standalone模型、YARN架构模型(画架构图)**
9. **Spark的yarn-cluster涉及的参数有哪些?**
10. **Spark提交job的流程**
11. **Spark的阶段划分**
12. **Spark处理数据的具体流程说下**
13. **Sparkjoin的分类**
14. **Spark map join的实现原理**
15. **介绍下Spark Shuffle及其优缺点**
16. **什么情况下会产生Spark Shuffle?**
17. **为什么要Spark Shuffle?**
18. **Spark为什么快?**
19. **Spark为什么适合迭代处理?**
20. **Spark数据倾斜问题，如何定位，解决方案**
21. **Spark的stage如何划分?在源码中是怎么判断属于Shuffle Map Stage或Result Stage的?**
22. **Spark join在什么情况下会变成窄依赖?**
23. **Spark的内存模型?**
24. **Spark分哪几个部分(模块)?分别有什么作用(做什么，自己用过哪些，做过什么)?**
25. **RDD的宽依赖和窄依赖，举例一些算子**
26. **Spark SQL的GroupBy会造成窄依赖吗?**
27. **GroupBy是行动算子吗**
28. **Spark的宽依赖和窄依赖，为什么要这么划分?**
29. **说下Spark中的Transform和Action，为什么Spark要把操作分为Transform和Action?常用的列举一些，说下算子原理**
30. **Spark的哪些算子会有shuffle过程?**
31. **Spark有了RDD，为什么还要有Dataform和DataSet?**
32. **Spark的RDD、DataFrame、DataSet、DataStream区别?**
33. **Spark的Job、Stage、Task分别介绍下，如何划分?**
34. **Application、job、Stage、task之间的关系**
35. **Stage内部逻辑**
36. **为什么要根据宽依赖划分Stage?为**
37. **什么要划分Stage**
38. **Stage的数量等于什么**
39. **对RDD、DAG和Task的理解**
40. **DAG为什么适合Spark?**
41. **介绍下Spark的DAG以及它的生成过程**
42. **DAGScheduler如何划分?干了什么活?**
43. **Spark容错机制?**
44. **RDD的容错**
45. **Executor内存分配?**
46. **Spark的batchsize，怎么解决小文件合并问题?**
47. **Spark参数(性能)调优**
48. **介绍一下Spark怎么基于内存计算的**
49. **说下什么是RDD(对RDD的理解)?RDD有哪些特点?说下知道的RDD算子**
50. **RDD底层原理**
51. **RDD属性**
52. **RDD的缓存级别?**
53. **Spark广播变量的实现和原理?**
54. **reduceByKey和groupByKey的区别和作用?**
55. **reduceByKey和reduce的区别?**
56. **使用reduceByKey出现数据倾斜怎么办?**
57. **Spark SQL的执行原理?**
58. **Spark SQL的优化?**
59. **说下Spark checkpoint**
60. **Spark SQL与DataFrame的使用?**
61. **Sparksql自定义函数?怎么创建DataFrame?**
62. **HashPartitioner和RangePartitioner的实现**
63. **Spark的水塘抽样**
64. **DAGScheduler、TaskScheduler、SchedulerBackend实现原理**
65. **介绍下Sparkclient提交application后，接下来的流程?**
66. **Spark的几种部署方式**
67. **在Yarn-client情况下，Driver此时在哪**
68. **Spark的cluster模式有什么好处**
69. **Driver怎么管理executor**
70. **Spark的map和flatmap的区别?**
71. **Spark的cache和persist的区别?它们是transformaiton算子还是action算子?**
72. **Saprk Streaming从Kafka中读取数据两种方式?**
73. **Spark Streaming的工作原理?**
74. **Spark Streaming的DStream和DStreamGraph的区别?**
75. **Spark输出文件的个数，如何合并小文件?**
76. **Spark的driver是怎么驱动作业流程的?**
77. **Spark SQL的劣势?**
78. **介绍下Spark Streaming和Structed Streaming**
79. **Spark为什么比Hadoop速度快?**
80. **DAG划分Spark源码实现?**
81. **Spark Streaming的双流join的过程，怎么做的?**
82. **Spark的Block管理**
83. **Spark怎么保证数据不丢失**
84. **Spark SQL如何使用UDF?**
85. **Spark温度二次排序**
86. **Spark实现wordcount**
87. **Spark Streaming怎么实现数据持久化保存?**
88. **Spark SQL读取文件，内存不够使用，如何处理?**
89. **Spark的lazy体现在哪里?**
90. **Spark中的并行度等于什么**
91. **Spark运行时并行度的设署**
92. **Spark SQL的数据倾斜**
93. **Spark的exactly-once**
94. **Spark的RDD和partition的联系**
95. **park 3.0特性**
96. **Spark计算的灵活性体现在哪里**

### Flink（约4.7w字）

1. **Flink架构**
2. **Flink的窗口了解哪些，都有什么区别，有哪几种?如何定义?**
3. **Flink窗口函数，时间语义相关的问题**
4. **介绍下Flink的watermark(水位线)，watermark需要实现哪个实现类，在何处定义?有什么作用?**
5. **Flink的窗口(实现)机制**
6. **说下Flink的CEP**
7. **说一说Flink的Checkpoint机制**
8. **Flink的Checkpoint底层如何实现的?savepoint和checkpoint有什么区别?**
9. **Flink的Checkpoint流程**
10. **Flink Checkpoint的作用**
11. **Flink中Checkpoint超时原因**
12. **Flink的ExactlyOnce语义怎么保证?**
13. **Flink的端到端ExactlyOnce**
14. **Flink的水印(Watermark)，有哪几种?**
15. **Flink的时间语义**
16. **Flink相比于其它流式处理框架的优点?**
17. **Flink和Spark的区别?什么情况下使用Flink?有什么优点?**
18. **Flink backPressure反压机制，指标监控你是怎么做的?**
19. **Flink如何保证一致性?**
20. **Flink支持JobMaster的HA啊?原理是怎么样的?**
21. **如何确定Flink任务的合理并行度?**
22. **Flink任务如何实现端到端一致?**
23. **Flink如何处理背(反)压?**
24. **Flink解决数据延迟的问题**
25. **Flink消费kafka分区的数据时flink件务并行度之间的关系**
26. **使用flink-client消费kafka数据还是使用flink-connector消费**
27. **如何动态修改Flink的配置，前提是Flink不能重启**
28. **Flink流批一体解释一下**
29. **说一下Flink的check和barrier**
30. **说一下Flink状态机制**
31. **Flink广播流**
32. **Flink实时topN**
33. **在实习中一般都怎么用Flink**
34. **Savepoint知道是什么吗**
35. **为什么用Flink不用别的微批考虑过吗**
36. **解释一下啥叫背压**
37. **Flink分布式快照**
38. **Flink SQL解析过程**
39. **Flink on YARN模式**
40. **Flink如何保证数据不丢失**

### 数仓（约3.6w字）

1. **介绍下数据仓库**
2. **数仓的基本原理**
3. **数仓架构**
4. **数据仓库分层(层级划分)，每层做什么?分层的好处?**
5. **数据分层是根据什么?**
6. **数仓分层的原则与思路**
7. **数仓建模常用模型吗?区别、优缺点?**
8. **星型模型和雪花模型的区别?应用场景?优劣对比**
9. **数仓建模有哪些方式?**
10. **数仓建模的流程?**
11. **维度建模的步骤，如何确定这些维度的**
12. **维度建模和范式建模区别**
13. **维度表和事实表的区别?**
14. **什么是ER模型?**
15. **OLAP、OLTP解释(区别)三范式是什么，举些例子**
16. **维度设计过程，事实设计过程**
17. **维度设计中有整合和拆分，有哪些方法，并详细说明**
18. **事实表设计分几种，每一种都是如何在业务中使用**
19. **单事务事实表、多事务事实表区别与作用**
20. **说下一致性维度、一致性事实、总线矩阵**
21. **从ODS层到DW层的ETL，做了哪些工作?**
22. **数据仓库与(传统)数据库的区别?**
23. **数据质量是怎么保证的，有哪些方法保证**
24. **怎么衡量数仓的数据质量，有哪些指标**
25. **增量表、全量表和拉链表**

### 数据库（约3.9w字）

1. **数据库中的事务是什么，MySQL中是怎么实现的**
2. **MySQL事务的特性?**
3. **数据库事务的隔离级别?解决了什么问题?默认事务隔离级别?**
4. **脏读，幻读，不可重复读的定义**
5. **MySQL怎么实现可重复读?**
6. **数据库第三范式和第四范式区别?**
7. **MySQL的存储引擎?**
8. **数据库有哪些锁?**
9. **说下悲观锁、乐观锁**
10. **分布式数据库是什么?**
11. **死锁产生的条件是什么?如何预防死锁?**
12. **介绍下数据库的ioin(内连接，外连接，全连接)，内连接和外连接(左，右连接)的区别**
13. **MySQL的join过程**
14. **MySQL有哪些存储引擎?**
15. **数据库中存储引擎MvlSAM与InnoDB的区别**
16. **Mylsam适用于什么场景?**
17. **InnoDB和Mvlsam针对读写场景?**
18. **MySQL Innodb实现了哪个隔离级别?**
19. **InnoDB数据引擎的特点**
20. **InnoDB用什么索引**
21. **Hash索引缺点**
22. **数据库索引的类型，各有什么优缺点?**
23. **MySQL的索引有哪些?索引如何优化?**
24. **有哪些数据结构可以作为索引呢?**
25. **B树与B+树的区别?**
26. **为什么使用B+树作为索引结构?**
27. **不使用B+树，可以用那个数据类型实现一个索引结构**
28. **介绍下MySQL的联合索引联合索使用原则**
29. **数据库有必要建索引吗?**
30. **MySQL缺点?**
31. **什么是脏读?怎么解决?**
32. **为什么要有三大范式，建数据库时一定要遵循吗?**
33. **数据库一般对哪些列建立索引?索引的数据结构?**
34. **MySOL中索引的建立需要考虑哪些问题**
35. **关系型数据库与非关系型数据库区别**
36. **MySQL与Redis区别**
37. **列式数据库和行式数据库优劣比对**
38. **除了UTF-8还有什么编码格式**
39. **布隆过滤器的基本原理是什么？局限性是什么？使用什么方法可以增加删除的功能？**
40. **你在哪些场景下使用了布隆过滤器？**
41. **SQL慢查询的解决方案（优化）？**
42. **聚簇索引、非聚簇索引说一下**
43. **哈希索引和B+相比的优势和劣势？**
44. **MVCC知道吗？**

## 三、结语

**从上述面试题可以大概看出来，Hadoop、Spark被问得最多，其次是Hive、Kafka、Flink这些，Flume相对来说，基本上很少看到**