**语言类：**

Java基础（数组、排序（冒泡、快读排序 二叉树 多线程、数据库、）

Linux常用的操作命令，Shell脚本编程

**数据分析**

Hadoop 搭建集群的步骤

集群结构

hadoop运行原理（包括HDFS（存储机制）和Mapreduce两部分。）

Hadoop集群常见问题及解决方案

hadoop怎样实现二级排序

hadoop优化问题

如何确认hadoop集群的健康状况

Hadoop 生态圈中各种框架的运用场景

熟悉Yarn的资源调度原理

Spark

原理

集群常见问题解决方案

性能调优

数据分析

Spark的常用算子

**s**park中的数据倾斜的现象、原因、后果及如何解决

spark中的宽窄依赖

spark 如何防止内存溢出

spark sql 如何使用？

sparkStreaming和storm的区别

## hadoop 与 storm、spark 的比较？

Spark 与mr的区别？mr遇到哪些问题？

**数据管理：**

Hbase：海量数据的存储检索

内部机制 优化

hbase 怎么预分区？

Hdfs和hbase区别联系？

Etl：

数据仓库hive (存储、数据清洗)

hive有哪些方式保存元数据，各有哪些优点

hive内部表和外部表的区别 表操作

hive 跟 hbase 的区别是？如何整合的？

数据清洗（遇到哪些问题）过滤的方法？

mapreduce

数据挖掘mapreduce定时调用及监控

请列出你在工作中使用过的开发mapreduce的语言

用mapreduce怎么处理数据倾斜问题

.MapReduce优化

mapreduce的大致流程

mapreduce的原理

举一个例子说明mapreduce是怎么运行的

Mysql

Redis：对数据进行缓存

数据传输工具：

Flume、

Kafka：（工作原理、核心组件）

flume 是如何导入数据到 kafka?

kafka整合sparkStreaming问题与flume整合sparkStreaming问题两者优缺点

如何处理海量数据保障执行速度？

如何预防数据丢失？一旦发生又如何解决？

Sqoop：工作原理，实现数据整合；

数据的压缩前后差别？

搭集群、数据仓库、数据清洗、数据传输、数据分析、数据管理