面试笔试题

系统抖动：

在请求分页存储管理中，从主存(DRAM)中刚刚换出(Swap Out)某一页面后(换出到Disk)，根据请求马上又换入(Swap In)该页，这种反复换出换入的现象，称为系统颠簸，也叫系统抖动。产生该现象的主要原因是置换算法选择不当。即对刚被替换出去的页，立即又要被访问。需要将它调入，因无空闲内存又要替换另一页，而后者又是即将被访问的页，于是造成了系统需花费大量的时间忙于进行这种频繁的页面交换，致使系统的实际效率很低，严重导致系统瘫痪，这种现象称为抖动现象。解决方案运用局部性原理优化置换算法。

磁盘的读写单位：

扇区是磁盘存储信息的最小物理单位，大小是硬件生产商固定的。

操作系统操作的基本单位是磁盘块，大小由操作系统决定，是虚拟的单位

ICMP协议属于网络层 IP协议的一种

多播IP用的是哪类地址：多播地址(multicast address)即组播地址，是一组主机的标示符，它已经加入到一个多播组中。在以太网中，多播地址是一个48位的标示符，命名了一组应该在这个网络中应用接收到一个分组的站点。

http协议的八种请求类型

10.170.31.1/25的同网段的地址：前25位不变 所以最大的是10.170.31.127 第25位位0 要是128 第25位是1

容器事务类型？Java事务类型：JDBC事务、容器事务、JTA事务

UDP要管对方的状态吗：无连接 不可靠 可靠性由上层应用层保证

数据倾斜，如何解决

Hdfs架构的基本原理，各个模块的基本作用

Wordcount

Rdd:

conf = new SparkConf()

sc = new SparkContext(conf)

line = sc.textFile(“”)

line.flatMap(\_.split(" ")).map((\_, 1)).reduceByKey(\_+\_).collect().foreach(println)

sc.stop()

dataframe:

data = sparkSession.read.text("src/main/resources/data.txt").as[String]

words = data.flatMap(value => value.split("\\s+"))

groupedWords = words.groupByKey(\_.toLowerCase)

counts = groupedWords.count()

counts.show()

sparksql:

from pyspark.sql import SparkSession

spark= SparkSession.builder()

.appName("WordCount")

.master("local")

.getOrCreate()

Data=spark.read.load(‘ ’),format(‘text’)

data.createOrReplaceTempView(‘data’)

result=spark.sql(“select word ,count(\*) from data group by word”)

result.show()

### 2、星飞面试题

Spring的IOC和AOP

Ioc：对象的实例化不是由代码来控制，而是由spring容器控制。当一个类要实例化另一个类时，传统使用代码的话会增加调用者与被调用者之间的耦合性，Ioc可以解决这个问题。Spring容器负责将被依赖对象赋值给调用者的成员变量，相当于调用者注入了它依赖的实例，这就是依赖注入。

Linux命令：查看集群负载，内存使用情况等；

Hadoop命令：上传文件到本地，下载文件到本地，查看文件目录等

如何预测namenode分配多大内存空间（数据块block的多少？因为namenode是存储数据块元信息的）

为什么spark比mapreducer快

Spark是基于内存的计算，而Hadoop是基于磁盘的计算。Spark多个任务之间数据通信是基于内存，而Hadoop是基于磁盘

统一的RDD抽象和操作

DAG ，它可以把整个执行过程做一个图，然后进行优化。

HBase的表结构

HBase为什么快 LSM-TREE文件组织方式、存储时表分区、部分是经过cache缓存

HBase如何查询 get查询一条和scan扫描

Hive优化：

（1）配置文件设置map和reduce个数；

（2）大表join小表；

（3）如果有一张表非常非常小，而另一张关联的表非常非常大的时候，你可以使用mapjoin此join 操作在 Map 阶段完成，不再需要Reduce，也就不需要经过Shuffle过程

（4）表关联时，关联的字段去掉null值。或者把空值key设成随机数字符串

（5）表关联时，不同类型的关联字段改成同类型的

（6）分区分桶。

内部表和外部表

（1）内部表数据由Hive自身管理，外部表数据由HDFS管理；

（2）内部表数据存储的位置是hive.metastore.warehouse.dir（默认：/user/hive/warehouse），外部表数据的存储位置由自己制定；

（3）删除内部表会直接删除元数据（metadata）及存储数据；删除外部表仅仅会删除元数据，HDFS上的文件并不会被删除；

分区表和桶表

（1）分区是文件夹，分桶是文件。

（2）对某一列进行hash，使用hashcode对 桶的个数求模取余，确定哪一条记录进入哪一个桶。

（3）在表关联时，两个字段都分桶的话，就很容易找到是哪个桶，加快关联。但两张表的桶数要是倍数或因子。

hive表的存储格式textfile，rcfile，orcfile

textfile正常文本格式，无压缩

rcfile数据以二进制格式编码，并且支持压缩，基于列存储（列值重复多）

rcfile游程编码（相同的数据不会重复存储）

orcfile游程编码、字典编码（压缩）、Bit编码

UDF UDAF UDTF