1. flume采集的数据传输到HDFS是否回出现重复数据，如果有重复数据，如何去重？

当flume读取kafka并且用高阶api，flume写入的过程中超时，会重新消费。

Flume写入hdfs的过程中突然中断，一般都是批次写入，则这一批的数据会重新写入。

我们在遇到重复数据的时候，都是在数仓的后面统一去重处理。如果非要去重，可以在flume采集数据的时候，给每条日志生成一个id，加在最前面：

（服务器ip+整条日志的hash值）拼接一个字符串再生成一个hashcode。

再通过id来去重

1. Spark Streaming处理实时数据的时间间隔，每秒钟大概有多少条数据，数据从处理到写入hbase大概得多长时间，是否是出现数据的积压

我们的时间间隔一般是5秒。计算3分钟。一分钟1-2w左右。倒没写hbase，我们是写kafka。

数据积压这种情况有出现过，每次的时间越来越来长，写出来的数据大大落后于写入的数据。后来通过监控任务的时间，超时就杀掉任务，并且修改streaming的consumer id，启动后重新消费数据。

1. Spark Streaming如何保证处理的数据不会有丢失

源数据通过kafka的游标来记录上次读取的数据位置。

生成的RDD的时候依赖Lineage的容错机制，数据丢失可以理解为lineage的异常，则会自动重新计算对应的分区数据。

其他的还可以手动用checkpoint来对宽依赖的情况做备份。

1. 业务数据如何从mysql导入数仓，如何做增量

通过datax，走jdbc导入数据仓库。

增量通过监控mysql 的binlog导入数据仓库

1. 写过MR程序吗？MR程序如何做测试？

这个很久之前写过。我以前测试就是保存中间数据到hdfs，或是抛异常，查看日志。

1. 数仓中大概有哪些维度的数据

App的时间（时间戳，日），区域，系统版本，sdk版本，语言，定位，新闻的点击，展示,主体，图片数等

1. 数据从hdfs到数仓ods层如何做数据的校验

通过hive UDF，hive sql语句等做校验。

1. 数仓各层数据之间如何做流转

通过spark sql 或hive sql进行ETL的流转。