1. 谈一下你们公司的整体架构。

属于研发部，技术总监下面有各个项目组，我们属于数据组，其他还有后端项目组，基础平台等。总监上面就是副总等级别了。其他的还有产品运营部，等。

1. 你们公司的测试环境是什么样的。

测试环境的配置是生产的一半。

1. 测试环境的数据量大概多少。

离线跟生产是同一个数据源，我们用的是分布式对象存储系统S3，所以测试也可以访问生产的s3。一般拿生产数据的一半测试。Kafka的数据也是从离线数据里面拿的，或是造数据，一般比较少。

1. 你感觉你们公司的架构合理不，有什么改进的地方。

* 目前我们所有产品的日志都写在一个kafka topic里面，后面可以分开
* 从日志到kafka的过程，flume其实有一层缓冲，可以通过自己写producer来发送消息，大大减少使用flume占用的内存和空间。
* 引入kettle做ETL的工作

1. 你们公司3年进行过多少次的项目迭代，每一个项目具体是如何迭代的。

差不过一个月会迭代一次。就产品或我们提出优化需求，然后评估时间。每周我们都会开会做下周计划和本周总结。

有时候也会调去做一些java的事情。

1. 你们公司集群的运行性能怎么样，是有什么好的改进措施。

集群是弹性的。能满足临时需要补数据的任务。性能若是6亿多的原始数据一起跑清洗，跑满整个队列，一个小时左右。

1. 你个人在工作过程中，谁给你提出需求，做完之后提交给谁，通过什么提交。

运营和产品。代码提交git，用git命令。结果生成数据表，导入Mysql，通过Hue查询，或是导出csv，给他们看

1. 你们大数据组一天要提交多少job？

整个组的话，生产环境按一个表就一个任务，然后两百多个表，加上实时的数据，差不多300多个任务。

1. 数据仓库权限问题，涉及哪些业务系统的数据，有没有设置权限管理，你的权限是什么？

我们的权限是让写入hdfs的数据只读（原始数据只读）。开放给外部都是只读权限（GUI或接口）

10. 使用什么进行ETL？有没有专门的ETL工具。

没有用专门的etl工具，我们都是手写sql脚本。（Kettle可以去了解一下）

11. 数据量非常大的时候如何更好的将数据导入Hive？

非常大的时候不需要导入hive，直接写入hdfs，用hive外部表关联。

12. 有没有其他的日志采集框架？（咱们的说的人太多了）

Logstash

13. 你们部门属于公司的第几级部门，部门的职级等级，晋升规则。

这个第几级部门不是很清楚。职级就分初级，中级，高级。晋升规则不一定，看公司效益和职位空缺。

14. 数据回溯。

我们接口层的数据一般都有保存在S3中，而且保存了2年以上的数据。可以随时回溯之前的原始数据。

15. 从业务数据库每天采集多少条业务数据，大概多大？采集流程耗时多少时间。

全表同步20多张表，大概几亿数据。时间一小时左右。

16. 每天处理多少条实时数据？

2千万左右。

17. 每天清洗完成的数据量有多大？

这个表太多了，没有统计过。但原始数据是6,7亿，清洗完成之后，出现细节表，各个埋点表，还有产品表，相当于\*3，所以是十几亿的样子。

18. 项目完后如何测试（测试使用那些监控工具），如果测试发现问题如何解决，最后如何上线

集群性能就看ganglia系统负载。

任务性能就看spark history server看应用的运行日志，用了多少executor。

Hadoop的任务就看yarn的log日志。

还有打点看实时任务的延时，性能调配一般在生产环境调节资源。

结果写入kafka或hive表中，查看数据。Spark调试用show（）或take（），hive调试用中间表。

上线的时候，将脚本打包，提交git。先发邮件抄送经理和总监，运维。通过之后跟运维一起上线。

19. 你具体每天的工作内容，具体说明每天大概做什么

新需求比如埋点或是报表来了之后，需要设计做的方案，设计完成之后跟产品讨论，再开发。

数仓的任何步骤出现问题，需要查看问题，比如日活，月活下降等。

20. 实现一个需求大概多长时间

不一定，有长有短，长的一周，短的半天也有。

21. 测试数据从哪里来，数据量多大

从生产环境直接拿。生产的数据有落地到S3，测试环境可以直接访问。

22. 项目测试时的各项测试指标

一个就是任务执行时间，这个可以看spark日志。多少次shuffle，项目的延时。还有就是读取和写入的数据量。

23. 每天离线生成的HDFS文件会有多少

这个没有具体统计过，但200多张表，hive一般一个表60多个文件，spark 200个文件，粗略估计2w多个文件。

24. Sqoop数据导出的时候一次执行多长时间，如何进行Sqoop调优

Sqoop这个没有用过