业务方面

1. 在处理大数据过程中，如何保证得到期望值

保证在数据采集的时候不丢失数据，这个尤为重要，如果在数据采集的时候就已经不准确，后面很难达到期望值

在数据处理的时候不丢失数据，例如sparkstreaming处理kafka数据的时候，要保证数据不丢失，这个尤为重要

前两步中，如果无法保证数据的完整性，那么就要通过离线计算进行数据的校对，这样才能保证我们能够得到期望值

2. 你感觉数仓建设中最重要的是什么

数仓建设中，最重要的是数据准确性，数据的真正价值在于数据驱动决策，通过数据指导运营，在一个不准确的数据驱动下，得到的一定是错误的数据分析，影响的是公司的业务发展决策，最终导致公司的策略调控失败。

3. 数据仓库建模怎么做的

数仓建设中最常用模型–Kimball维度建模详解

4. 数据质量怎么监控

单表数据量监控

一张表的记录数在一个已知的范围内，或者上下浮动不会超过某个阈值

SQL结果：var 数据量 = select count（\*）from 表 where 时间等过滤条件

报警触发条件设置：如果数据量不在[数值下限, 数值上限]， 则触发报警

同比增加：如果((本周的数据量 -上周的数据量)/上周的数据量\*100)不在 [比例下线，比例上限]，则触发报警

环比增加：如果((今天的数据量 - 昨天的数据量)/昨天的数据量\*100)不在 [比例下线，比例上限]，则触发报警

报警触发条件设置一定要有。如果没有配置的阈值，不能做监控

日活、周活、月活、留存（日周月）、转化率（日、周、月）GMV（日、周、月）

复购率（日周月）

单表空值检测

某个字段为空的记录数在一个范围内，或者占总量的百分比在某个阈值范围内

目标字段：选择要监控的字段，不能选“无”

SQL结果：var 异常数据量 = select count(\*) from 表 where 目标字段 is null

单次检测：如果(异常数据量)不在[数值下限, 数值上限]，则触发报警

单表重复值检测

一个或多个字段是否满足某些规则

目标字段：第一步先正常统计条数；select count(\*) form 表；

第二步，去重统计；select count(\*) from 表 group by 某个字段

第一步的值和第二步不的值做减法，看是否在上下线阀值之内

单次检测：如果(异常数据量)不在[数值下限, 数值上限]， 则触发报警

跨表数据量对比

主要针对同步流程，监控两张表的数据量是否一致

SQL结果：count(本表) - count(关联表)

阈值配置与“空值检测”相同

5. 数据分析方法论了解过哪些？

数据商业分析的目标是利用大数据为所有职场人员做出迅捷，高质，高效的决策提供可规模化的解决方案。商业分析是创造价值的数据科学。

数据商业分析中会存在很多判断：

观察数据当前发生了什么？

比如想知道线上渠道A、B各自带来了多少流量，新上线的产品有多少用户喜欢，新注册流中注册的人数有多少。这些都需要通过数据来展示结果。

理解为什么发生？

我们需要知道渠道A为什么比渠道B好，这些是要通过数据去发现的。也许某个关键字带来的流量转化率比其他都要低，这时可以通过信息、知识、数据沉淀出发生的原因是什么。

预测未来会发生什么？

在对渠道A、B有了判断之后，根据以往的知识预测未来会发生什么。在投放渠道C、D的时候，猜测渠道C比渠道D好，当上线新的注册流、新的优化，可以知道哪一个节点比较容易出问题，这些都是通过数据进行预测的过程。

商业决策

所有工作中最有意义的还是商业决策，通过数据来判断应该做什么。这是商业分析最终的目的。