长亮 平安项目 甲方面试

数据处理的整体流程，如何理解？

大数据处理的基本流程可以分为数据抽取与集成、数据分析和数据解释（展示）等步骤

 大数据的处理流程可以定义为，在合适工具的辅助下，对广泛异构的数据源进行抽取和集成，将结果按照一定的标准进行统一存储，然后利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识，并利用恰当的方式将结果展现给终端用户。

Ods分几层？

在项目内主要负责什么？

根据需求关联事实表，维度表，和进行指标的计算

接到一个业务需求之后如何实现？介绍详细开发流程？

需求是客户先提到需求分析师那里，然后需求分析师会对需求进行拆解和澄清，确认取数逻辑，然后我们会开会讨论如何实现，包括可行性分析，设计好模型表，然后我们进行开发的。

整体开发流程主要从 需求阶段🡪设计阶段🡪开发阶段🡪测试阶段🡪上线阶段

 需求阶段，产品经理给我们交接完需求，项目经理会根据需求给我们分配开发任务；

 设计阶段，我们会根据需求完成，表结构的设计，整理出数据字典；

 开发阶段，我们会根据自己负责的开发任务进行开发，每天项目经理会跟我们开晨会，过一个每个人的开发进度，开发完成以后，我们会直接做UT测试（单元测试），转测之前我们项目经理会组织我们进行代码评审，评审通过以后提交测试验证；

 测试阶段，主要是改BUG测试发现问题我们修复，然后跟踪缺陷闭环；

上线阶段，参与上线支持，项目组的全体成员都参与，上线当天有负责生产环境部署的同事进行版本上线的工作，有问题我们及时定位解决，没问题就在公司休息，等上线结束。

在学什么新技术?

最近准备看 Linux命令与shell脚本编程大全，hadoop大数据分析与挖掘实战，

1、自我介绍，项目介绍

2、宽表是谁设计的

DM层的开发人员设计的

事实表、维表，然后两种表聚合成一个宽表，其实这个宽表就是指标表。

宽表从字面意义上讲就是字段比较多的数据库表。

通常是指业务主题相关的指标、维度、属性关联在一起的一张数据库表。

需要维度和指标！而不是机械得认为把维度拆多了，就叫宽表！

3、项目分了几层

用的是传统数仓，三层 ODS 贴源层，DW数仓层，DM集市层

以下问题参考

<https://blog.csdn.net/dannistang/article/details/80597776?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522161478689416780271517581%2522%252C%2522scm%2522%253A%252220140713.130102334.pc%255Fall.%2522%257D&request_id=161478689416780271517581&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~first_rank_v2~rank_v29-3-80597776.first_rank_v2_pc_rank_v29&utm_term=KETTLE%E8%B0%83%E7%94%A8Store+Procedure>

4、ADM分了几层

问题没懂，体系结构开发方法？

5、共性加工层在哪里

在加工汇总层(ADM)层

大数据主流架构

SDM层（共性加工层）：按照业务的共同属性对数据进行整合。以订单支付为例，一个卖家

只有一条实时成交金额。

6、ODS和FDM有什么区别

基础数据层（FDM：FOUNDATION DATA MODLE）。来源于标准化的各源系统提供的业务基础数据。基础数据层的数据粒度为细节级数据。存储从源系统抽取的业务基础数据。这些数据经过清洗、标准化，主要是操作型数据和参数数据。基础数据层按业务主题进行归类、整合。

7、ODS是大宽表吗

8、ODS和ADM的区别

加工汇总层（ADM：ANALYSIS DATA MODLE）。在基础数据的上进行加工汇总形成的指标数据存储分析型和加工汇总型数据。来源于标准化的各源系统的汇总、报表数据，是基础数据经过加工按一定维度汇总的指标，或分析数据。加工汇总层的数据需求来源于应用的一些共同性指标，可以是一些中间数据，这些指标的存在，可以大大提高应用系统的处理效率。

9、项目用了哪些技术

10、调度工具用什么，如何实现

Kettle，Control-M

11、做过什么优化

12、同层不能依赖是否就重复加工了

13、相同字段下沉依据是什么

14、拉链表如何设计及实现

15、分区是如何设计

16、基于Hive的优化，查询速度慢如何排查问题，多大的数据量不先进内存，是否还有其他原因，资源不足在哪里看

17、是否熟悉关系型数据库

1.hive的优化有哪些？

2.对于表分区查询的时候必须 where 表分区字段，必须指定表分区限制

3.join的时候必须用 on 关联不能用 where关联，用 where会进行笛卡尔积

4.join优化 ：如果有小表我们开启join优化，小表放左边。

5.处理数据去重时，应该使用group by做分组去重，不要用distinction count（）避免数据倾斜。

6.对于经常查询的一些字段可以建立索引，但查询索引字段时不能对索引字段使用substr或拼接函数，这样会导致查询不走索引。

7.查询时select的字段个数也会影响查询效率，select的字段数越少，速度越快。

8.hive脚本报错了怎么处理？

9.一般就是hql会报错，通过jobid查看后台日志，比如字段类型,字段长度，函数写错了等等，通过修改，测试来解决。

10如果存储过程中用到临时表，必须在存储过程的开始和接收时先删除临时表。

11你常用的hive函数有哪些？

sum，count,avg,round,to\_date,coalesce,from\_unixtime,month,year,trunc

12hive日期取月初，季初，年初等函数有哪些？

trunc（‘date’‘YY’）取年初，trunc（‘date’‘MM’）取月初。

add\_months（‘date’1）增加一个月

date\_add('date'1)增加一天

13说下你之前的项目经验及项目内容

14分区的类型有哪些？

普通分区，使用日期etl\_dt字段作为分区；

组合分区，使用日期加状态作为组合的分行；