# 一 数据仓库

## 1.数据仓库分层

### 1.2 主题层

FDM主题层:  
团体

事件

协议

产品

财务

渠道

资产

公共

汇总

报表

## 2.遇到的困难

## 3.平时工作中如优化重复性的工作

## 4.给一个需求场景 简述如何设计开发

# 二 大数据

ORC 表：ORC的全称是(Optimized Record Columnar)

创建普通临时表：

create table if not exists test\_orc\_tmp(

  name string,

  gender string,

  cnt BIGINT

)row format delimited fields terminated by '|' stored as textfile;

创建ORC表：

drop table test\_orc;

create table if not exists test\_orc(

  name string,

  gender string,

  cnt BIGINT

)STORED AS ORC;

创建测试数据  vi orc.txt

goldenkey0|male|111

goldenkey1|male|112

goldenkey2|male|113

goldenkey3|male|114

goldenkey4|male|115

goldenkey5|male|116

goldenkey6|male|117

将测试数据导入临时普通表：

LOAD DATA INPATH '/tmp/orc.txt' OVERWRITE INTO TABLE test\_orc\_tmp;

将临时普通表的数据插入到ORC表：

INSERT INTO TABLE test\_orc SELECT \* FROM test\_orc\_tmp;

查看ORC表是否有数据：

select \* from test\_orc\_tmp;

ORC知识链接：

<https://www.jianshu.com/p/36e4f0137744>

Inceptor命令行快速入门

任务：在本地创建文本，导入到Inceptor新建的表格中。

1. 登录安装有Inceptor Server的节点，新建一个文本文件quickstart.txt,并将其放在tmp文件夹中。

vim /tmp/quickstart.txt

2. 向quickstart.txt中写入两行文字，如下图所示。



输入后按Esc键，输入:wq保存退出。

3. 命令行连接Inceptor

• InceptorServer 1

transwarp -t -h localhost

• InceptorServer 2 简单认证

beeline -u "jdbc:hive2://localhost:10000/default"

• InceptorServer 2 LDAP认证

beeline -u "jdbc:hive2://localhost:10000/default" -n <username> -p <password>

• InceptorServer 2 Kerberos认证

kinit <your\_principal>

beeline -u "jdbc:hive2://localhost:10000/default;principal=hive/node1@TDH"

4. 新建外表quickstart

CREATE EXTERNAL TABLE quickstart (a INT, b STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY

'|';

5. 将位于本地的quickstart.txt导入至表格quickstart中

LOAD DATA LOCAL INPATH '/tmp/quickstart.txt' INTO TABLE quickstart;

注意如果Inceptor使用LDAP或者Kerberos认证，您需要是这个文件的owner。

6. 查看表格内容，结果为：



3.3.2.1.1. 直接建表

语法：

CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] ①

[<database\_name>.]<table\_name>　 ②

[(<column\_name> <data\_type> [COMMENT '<column\_comment>'] ③

[, <column\_name> <data\_type> ...])]

[COMMENT '<table\_comment>'] ③

[PARTITIONED BY (<part\_key> <data\_type> [COMMENT '<partition\_comment>'] ③

[, <part\_key> <data\_type>...])] ④

[CLUSTERED BY (<col\_name> [, <col\_name>...]) ⑤

[SORTED BY (<col\_name> [ASC|DESC] [, <col\_name> [ASC|DESC]...])] ⑥

INTO <num\_buckets> BUCKETS]

[

[ROW FORMAT <row\_format>] ⑦

[STORED AS (TEXTFILE|ORC|CSVFILE)] ⑧

| STORED BY '<storage.handler.class.name>' [WITH SERDEPROPERTIES (<...>)] ⑨

]

[LOCATION '<hdfs\_path>'] ⑩

[TBLPROPERTIES ('<property\_name>'='<property\_value>', ...)]　 ⑪

① TEMPORARY 为临时表选项，EXTERNAL 为外表选项。

② 加上 <database\_name>. 则将表建在指定的数据库中。

③ 可以为列、表和分区用 COMMENT 加注释，注意注释要放在单引号中。

④ 指定分区键，具体使用方式请参考“分区表”。

⑤ 分桶子句，具体使用方式请参考“分桶表”。

⑥ 桶内排序选项，ASC 为升序，DESC 为降序。

⑦ 行格式，在建TEXT表时使用，具体细节请参考“TEXT表”部

⑧ 指定文件格式，该选项在建TEXT表，ORC表或CSV表时使用。

⑨ 指定使用的storage handler。

⑩ 指向HDFS上的一个目录。这个选项我们推荐只在建外表时使用，也就是和 EXTERNAL 选项合用（虽然

也可以在建内表时使用，但是我们　不建议 这样做）。 该路径必须是绝对路径，并且执行操作的用户必

须是这个路径指向的目录或文件的owner。如果 <hdfs\_path> 指向的目录不存在，Inceptor会尝试新建这

个目录，但是安全模式下Inceptor可能没有在指定路径新建目录的权限，所以星环科技建议尽量避免让

<hdfs\_path> 指向不存在的目录。

⑪ 表属性，由键值对表示，在建Holodesk表，ORC事务表和CSV表时使用，具体细节请参考“ORC表”，

“Holodesk表”和“CSV表”。

# 三 数据治理

# 四 Java基础

# 五 版本控制工具

$ git merge dev-20191119

# 六、面试记录

## PAJF面试

## 1. 挑选项目经历中最有价值的一个讲出来，如果让你重构，你会怎么样做才能做得更好？

### 2.数据库基础知识

文章参考地址：

<https://wenku.baidu.com/view/c73b15b17e21af45b207a850.html>

<https://www.jianshu.com/p/c65c4ec0fb3f>

<https://blog.csdn.net/dosthing/article/details/87954213>

#### （1） 数据库设计的三范式 原则及为了解决什么问题

为了建立冗余较小、结构合理的数据库，设计数据库时必须遵循一定的规则。在关系型数据库中这种规则就称为范式(英文Normal Form,缩写NF)。范式是符合某一种设计要求的总结。要想设计一个结构合理的关系型数据库，必须满足一定的范式

在实际开发中最为常见的设计范式有三个：

1）.第一范式（确保每列保持原子性）

第一范式是最基本的范式。如果数据库表中的所有字段都是不可分解的原子值，就说明了该数据库表满足了第一范式。数据表中的每一列（字段），必须是不可拆分的最小单元，也就是确保每一列的原子性，而不是集合。

实例：如以下订单表，买家地址列并不符合第一范式，需要继续拆分



上表所示的订单遵循了第一范式的要求，这样对用户使用城市进行分类的时候就非常方便，也提高了数据的性能

2）.第二范式（确保表中的每列都和主键相关）

第二范式在第一范式的基础之上更进一步。第二范式需要确保数据库中的每一列和主键相关，而不能只与主键的某一部分相关（主要针对联合主键而言）。也就是说在一个数据库表中，只能保存一种数据，不可以把多种数据保存在同一张数据库表中。

满足1NF的基础上，要求：表中的所有列，都必须依赖于主键，而不能有任何一列与主键没有关系（一个表只描述一件事情）。第二范式消除表的无关数据。

将上述订单表拆分成多张表，拆分后如下图



3）.第三范式（确保每列都和主键列直接相关，而不是间接相关）

第三范式需要确保数据表中的每一列数据都和主键直接相关，而不能间接相关，列和列之间不存在相互依赖关系。要求一个关系中不包含已在其它关系已包含的非主关键字信息。

满足2NF的基础上，任何非主属性不依赖于其他非主属性（在2NF的基础上消除依赖传递）（也表明不允许数据存在冗余的现象）。

上述表中，虽然单价和订购数量可以计算出总价，但是单价是直接和商品相关的，并不是和订单直接相关，所以不满足第三范式，需要继续拆分



**（2）数据库的事务性**

除了数据库设计的三大范式之外，事务处理也是保证数据完整性的重要手段。事务是单独的工作单元，该单元可以包含多个操作以完成一个完整的任务。锁是在多用户环境中对数据访问的限制。事务和锁确保了数据的完整性。

**事务处理**

提交commit，当所有的操作步骤都被完整执行后，称该事务被提交。

回滚rollback，由于某一操作步骤执行失败，导致所有都没有被提交，则事务必须回滚，即回到事务执行前的状态。

**事务ACID属性**

事务处理的特性，每一个事务都有他们所共有的特性，叫做ACID特性，分别是原子性atomicity、一致性consistence、隔离性Isolation、持久性Durability

1）原子性，事务的原子性表示事务执行过程中，把事务作为一个工作单元处理，一个工作单元可能包含若干个操作步骤，每个操作步骤都必须完成才算完成，若因任何原因导致其中的一个步骤操作失败，则所有操作步骤失败，前面的步骤必须回滚。

2）一致性，事务的一致性保证数据处于一致状态。如果事务开始时处于一致状态，则事务结束时也应处于一致状态，不管事务成功还是失败。

3）隔离性，事务的隔离性保证事务访问的任何数据都不会受到其他事务所做的任何改变的影响，直到该事务完成。

4）持久性，事务的持久性保证假如事务执行，则它在系统中产生的结果应该是持久的。

（3）数据库五大约束

1、主键约束(Primay Key)

唯一性，非空性

2、唯一约束(Unique)

唯一性，可以空，但只能有一个

3、检查约束(Check)

对该列数据的范围、格式的限制(如：年龄、性别等)

4、默认约束(Defaut)

该数据的默认值

5、外键约束（Foreign Key）

需要建立两表间的关系

#### （2） 数据库设计的拉链表 必须字段

本文链接：<https://blog.csdn.net/weixin_45399233/article/details/100601551>

本文链接：<https://blog.csdn.net/u012965373/article/details/81515463>

本文链接：https://blog.csdn.net/fhy36897/article/details/89553784

增量数据与历史库做成拉链表

首先介绍几个表概念：

**全量表**：每天的所有的最新状态的数据

**增量表**：每天的新增数据

**拉链表**：维护历史状态，以及最新状态数据

**流水表**：对于表中的每一个记录都会修改，可以用于反映实际记录的变更

拉链表VS流水表：

拉链表：通常是对账户信息的历史变动进行处理保留的结果，用于统计业务相关情况

流水表：每天的交易形成的历史，用户统计账户及客户的情况

在数据仓库的数据模型设计过程中，经常会遇到这样的需求：

1）. 数据量比较大；

2）.表中的部分字段会被update，如用户的地址，产品的描述信息，订单的状态等；

3）.需要查看一个时间点或者时间段的历史快照信息，比如，查看一个订单早历史某一个时间点的状态，比如，查看某一个用户在过去某一个时间内，更新过几次等等；

4）变化的频率和比例不是很大，比如，总共有1000万的会员，每天新增和发生变化的有10万左右；

5）如果对这个表每天都保留一份全量，那么每次全量中会保存很多不变的信息，对存储是极大的浪费。

拉链历史表，技能满足反应数据的历史状态，又可以最大程度的节省存储。.

拉链表案例：

1）有一张订单表，2019-06-20 这天里面有三条订单数据：



2）到了2019-06-21这天，表中有5条记录：



3）到了2019-06-22日，表中有6条记录：



这种存储方式有什么问题呢？

1.只保留一份全量，则数据和6月22日的记录一样，如果要查看6月21日订单001的状态，则无法满足。

2、每天都保留一份全量，则数据仓库中的该表共有14条记录，但好多记录都是重复保存，没有任何变化，如订单002,004，数据量大了会造成很大的存储浪费；

如果设计成历史拉链表，如下：



拉链表中字段解释：

1）、dw\_begin\_date 表示该条记录的生命周期的开始时间，dw\_end\_date表示该条记录的生命周期结束时间；

2）dw\_end\_date=’9999-12-31’表示该条记录目前处于有效状态；

3）如果查询当前所有有效的记录，则select \* from order\_his where dw\_end\_date =’9999-12--31’

4)如果查询2019-06-21的历史数据，则select \* from order\_his where dw\_begin\_date <= ‘2019-06-21’ and dw\_end\_date >= ‘2019-06-21’,这条语句会查询到以下记录：



可以看出，这样的历史拉链表，技能满足对历史数据的要求，又能很大程度的节省存储资源。

举例：













**需要参考博客，讲的很清楚，自己整理的由于篇幅限制截图不全。**



**增量抽取**

每天，从源系统订单表中，将前一天的增量数据抽取到ODS层的增量数据表。这里的增量需要通过订单表中的创建时间和修改时间来确定：



**增量刷新历史数据**

从2019-08-22开始，需要每天正常刷新前一天（2019-08-21）的增量数据到历史表。

**第一步，通过增量抽取，将2019-08-21的数据抽取到ODS：**



ODS增量表中2019-08-21的数据如下：



**第二步，通过DW历史数据(数据日期为2019-08-20)，和ODS增量数据(2019-08-21)，刷新历史表：**













#### （3） 数据库调优如何看调优的，怎么解决的，看什么字段

#### （4）给一个场景如何开发

#### （5）常用数据库语句，如何写，分组排序取最大/小

原文链接：https://blog.csdn.net/qq\_25221835/article/details/82762416

<https://blog.csdn.net/qq_25221835/article/details/82762416>

语法格式：row\_number() over(partition by 分组列 order by 排序列 desc)

在使用 row\_number() over()函数时候，over()里头的分组以及排序的执行晚于 where 、group by、  order by 的执行。

例一：

表数据：

create table TEST\_ROW\_NUMBER\_OVER(

id varchar(10) not null,

name varchar(10) null,

age varchar(10) null,

salary int null

);

select \* from TEST\_ROW\_NUMBER\_OVER t;

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(1,'a',10,8000);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(1,'a2',11,6500);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(2,'b',12,13000);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(2,'b2',13,4500);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(3,'c',14,3000);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(3,'c2',15,20000);

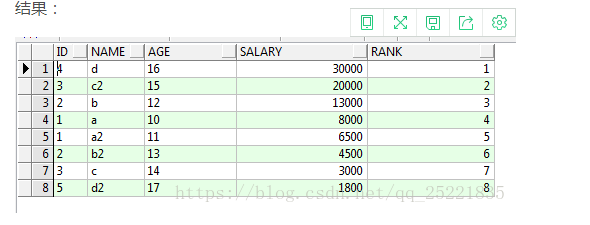
insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(4,'d',16,30000);

insert into TEST\_ROW\_NUMBER\_OVER(id,name,age,salary) values(5,'d2',17,1800);

一次排序：对查询结果进行排序（无分组）

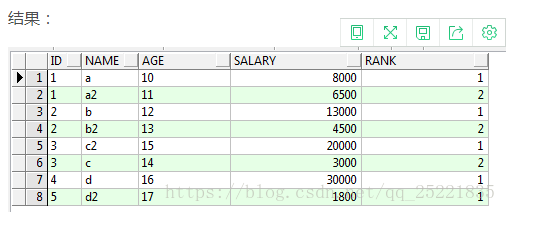
select id,name,age,salary,row\_number()over(order by salary desc) rn

from TEST\_ROW\_NUMBER\_OVER t

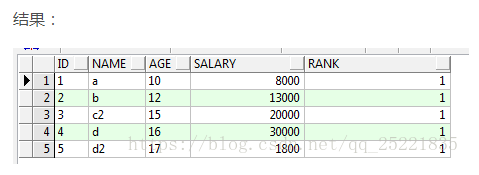


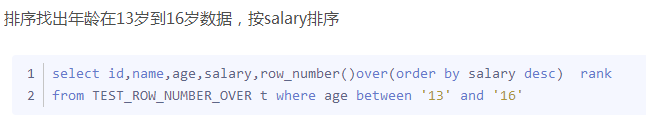
进一步排序：根据id分组排序

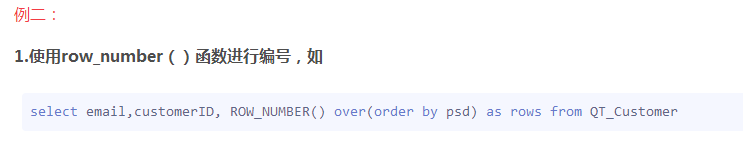
1. select id,name,age,salary,row\_number()over(partition by id order by salary desc) rank
2. from TEST\_ROW\_NUMBER\_OVER t



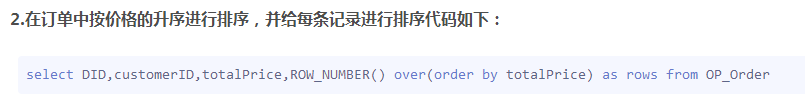
1. select \* from(select id,name,age,salary,row\_number()over(partition by id order by salary desc) rank
2. from TEST\_ROW\_NUMBER\_OVER t)
3. where rank <2



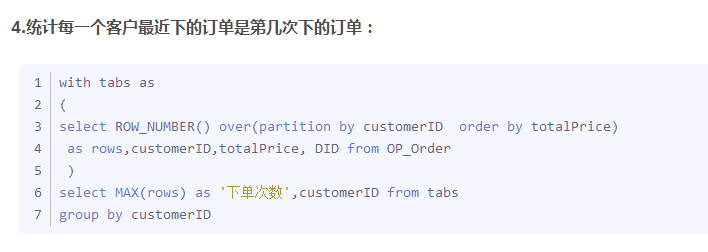




原理：先按psd进行排序，排序完后，给每条数据进行编号。







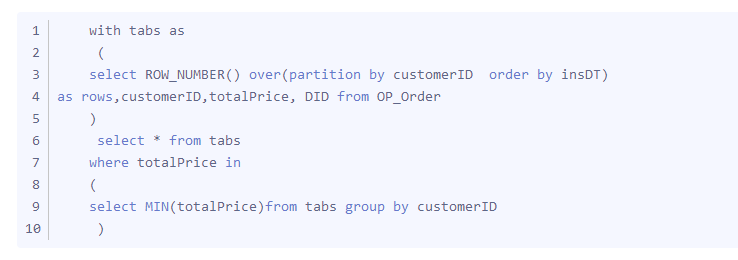
5.统计每一个客户所有的订单中购买的金额最小，而且并统计改订单中，客户是第几次购买的：

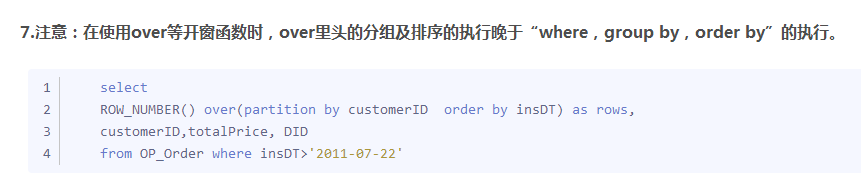
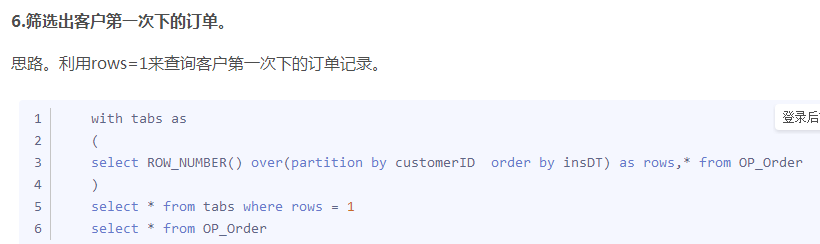
思路：利用临时表来执行这一操作。

1.先按客户进行分组，然后按客户的下单的时间进行排序，并进行编号。

2.然后利用子查询查找出每一个客户购买时的最小价格。

3.根据查找出每一个客户的最小价格来查找相应的记录。





## 3.JAVA会不会

（1）spring /springMVC 的区别，有没有接触过springboot

## 4.SHELL脚本

（1） 定时语句/命令怎么写

（2） 查找一个文件中某个字符串出现的次数

## 5.其他

（1） 平时 有没有打建过系统环境，SVN,服务器之类的，如何搭建的

## 6.为什么要换工作，平时有没有对新技术进行钻研

# 七、读书笔记

第一章 决策支持系统的发展

当数据从操作型环境传向数据仓库环境时，需要对数据进行集成。

第二章 数据仓库环境

数据仓库是一个面向主题的、集成的、非易失，随时间变化的用来支持管理人员决策的数据集合。