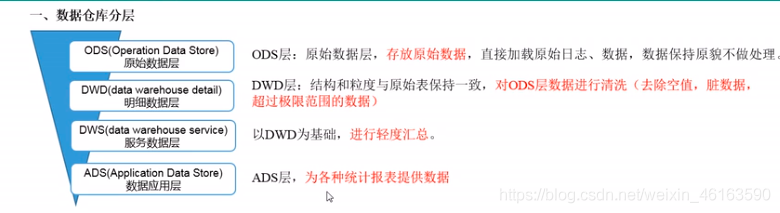
# 数仓面试题

## 数仓分层



## 二、为什么要对数仓进行分层？

**1. 把复杂问题简单化**

将一个复杂的任务分解成多个步骤来完成，每一层只处理单一的步骤，比较简单、并且方便定位问题。

**2. 减少重复开发**

规范数据分层，通过中间层数据，能够极大地减少重复计算，增加一次计算结果的复用性。

**3. 隔离原始数据**

不论是数据的异常还是数据的敏感性，使真实数据与统计数据解耦开

## 三、说说对数据仓库的理解

数据仓库是面向主题进行组织的，数据是集成的、不可更新的、随时间变化的。

数据仓库经历了这样三个阶段：简单报表阶段、数据集市阶段、数据仓库阶段。

## 四、数据库和数据仓库的区别

**数据库**：是一种逻辑概念，用来存放数据的仓库，通过数据库软件来实现。数据库由很多表组成，表是二维的，一张表里可以有很多字段。

**数据仓库**：是数据库概念的升级。从逻辑上理解，数据库和数据仓库没有区别，都是通过数据库软件实现存放数据的地方，但是从数据量来讲数据仓库比数据库更庞大些。数据仓库的主要作用用于数据挖掘、数据分析，辅助领导来做决策的。

区别：

实际上二者之间讲的是OLTP和OLAP

操作型处理 OLTP：它是针对具体业务在数据库联机的日常操作，通常对少数记录进行查询、修改。

分析型处理 OLAP：一般针对某些主题的历史数据进行分析，支持管理决策。

## 五、为什么叫星型模型和雪花模型？

星型模型是：多维表的数据关系，它由一个事实表和一组维表组成，每个维作为主键

雪花模型是：当一个或多个维没有直接连接到事实表上，而是通过其他维表连接到事实表上的时候，其图解就像雪花模型连接在一起、

使用场景：

雪花模型使得维度分析更加容易，比如“针对特定的广告主，有哪些客户或者公司是在线的?”

星形模型用来做指标分析更适合，比如“给定的一个客户他们的收入是多少?”

## 六、什么是缓慢变化维？

缓慢变化维

1. 缓慢变化维在数据仓库中，有个概念叫缓慢变化维，经常简写为SCD，

2. 缓慢变化维的提出是因为现实世界中，维度的属性并不是静态的，它会随时间的流失发生缓慢的变化。

这种随着时间变化的我们称之为缓慢变化维。

## 七、数仓三层架构，具体每层作用？

1. **数据访问层**：主要是对非原始数据（数据库或者文本文件等存放数据的形式）的操作层，而不是指原始数据，也就是说，是对数据库的操作，而不是数据，具体为业务逻辑层或表示层提供数据服务。

2. **业务逻辑层**：主要是针对具体的问题的操作，也可以理解成对数据层的操作，对数据业务逻辑处理，如果说数据层是积木，那逻辑层就是对这些积木的搭建。

3. **界面层**：主要表示WEB方式，也可以表示成WINFORM方式，WEB方式也可以表现成：aspx，如果逻辑层相当强大和完善，无论表现层如何定义和更改，逻辑层都能完善地提供服务。

## 八、拉链表知道么？

拉链表就是随着时间变迁产生历史数据。

拉链表的含义：就是记录历史。记录一个事务从开始一直到当前状态的所有变化信息。

## 九、in exists 的区别 not in 和 not exists的区别？

in 是一个集合运算符.

a in {a,c,d,s,d…}

这个运算中,前面是一个元素,后面是一个集合,集合中的元素类型是和前面的元素一样的。

而exists是一个存在判断,如果后面的查询中有结果,则exists为真,否则为假。

总之：

尽量使用not exist ，避免使用not in

not in 会默认调用子查询

not exist 会调用关联子查询

## 十、增量处理的时候，源数据库删除的数据是怎么处理

目前想到两种思路：面试当时没有理清楚题目，现在来回答下。

1.写kafka API消费掉原始数据

2.如果是大数据架构的可以使用sqoop、DataX进去同步至HDFS 或者别的数据仓库。

## 十一、数据集市与数据仓库概念

**数据集市**：则是一种微型的数据仓库，通常有更少的数据，更少的主题区域，更少的历史数据，部门级的，一般只能为某个局部范围内的管理人员服务。

**数据仓库**：是企业级的，能为整个企业各个部门的运行提供决策支持手段。

## 十二、表的分类

* 实体表：一般是指现实存在的业务对象
* 维度表：一般是指对应一些业务状态，也可称为码表
* 事务型事实表：一般指随着业务发生不断产生数据，特点是一旦发生不会再变化
* 周期型事实表：一般指随着业务发生不断产生变化（更新，新增）的数据，如订单

## 十三、同步策略

* 全量表：存储完整的数据
* 增量表：存储新增加的数据
* 新增及变化表：存储新增加的数据和变化的数据
* 拉链表：对新增及变化表做定期合并
* 实体表（用户，商品，商家）：每日全量
* 维度表（订单状态，审批状态，商品分类）：每日全量

事务型实时表（交易流水，操作日志，出库入库记录）：数据量大且不变，每日增量表，每日创建一个分区存储

周期型事实表（订单，请假等）：用每日新增和变化表，制作一张拉链表。

## 十四、范式理论

目的：

（1）：减少数据冗余，尽量让每个数据只出现一次；

（2）保证数据一致性；

## 十五、三范式区分

（1）第一范式：属性不可切割，所有关系数据库等最基本要求

（2）第二范式：不能存在部分函数依赖

（3）低三范式：不能存在传递函数依赖

## 十六、关系建模与维度建模

* 关系建模：为了保证数据的一致性和避免冗余，所以大部分业务系统的表都是遵循低三范式
* 维度建模：
  + 星型模型：区别主要在于维度的层级，标准
  + 雪花模型
  + 星座模型

oltp（随机延时读写）

olap（批量导入）

## 十七、数据仓库的输入数据源和输出系统分别是什么？

输入系统：埋点产生的用户行为数据、JavaEE后台产生的业务数据。

输出系统：报表系统、用户画像系统、推荐系统

## 十八、数据仓库的基本架构是什么？

(数据源，ETL, data stage，ODS，data warehouse,datamart,OLAP等等)

### 数据仓库系统体系结构

1. 数据源-> 2.ETL -> 3.数据仓库存储与管理-> 4.OLAP -> 5.BI工具

数据源：是数据仓库系统的数据源泉,通常包括企业各类信息,包括存放于RDBMS中的各种业务处理数据和各类文档数据；各类法律法规、市场信息和竞争对手的信息等等；

数据的存储与管理：数据的存储和管理是整个数据仓库的核心，是关键。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。从数据仓库的技术特点着手分析，来决定采用什么产品和技术来建立数据仓库，然后针对现有各业务系统的数据，进行抽取、清理，并有效集成，按照主题进行组织。数据仓库按照数据的覆盖范围可以分为企业级数据仓库和部门级数据仓库（通常称为数据集市）。

### OLAP服务器

对需要的数据进行有效集成，按多维模型予以组织，以便进行多角度、多层次的分析，并发现趋势。其具体实现可以分为：ROLAP（关系型在线分析处理）、MOLAP（多维在线分析处理）和HOLAP（混合型线上分析处理）。ROLAP基本数据和聚合数据均存放在RDBMS之中；MOLAP基本数据和聚合数据均存放于多维数据库中；HOLAP基本数据存放于RDBMS之中，聚合数据存放于多维数据库中。

前端工具：主要包括各查询工具、数据分析工具、数据挖掘工具、种报表工具以及各种基于数据仓库或数据集市的应用开发工具。

数据分析工具主要针对OLAP服务器。报表工具、数据挖掘工具主要针对数据仓库。

## 十九、构建企业级数据仓库五步法:

### 1. 确定主题

即确定数据分析或前端展现的主题(例:某年某月某地区的啤酒销售情况)。主题要体现出某一方面的各分析角度(维度)和统计数值型数据(量度)之间的关系,确定主题时要综合考虑.

### 2. 确定量度

确定主题后，需要考虑分析的技术指标(例:年销售额等等)。它们一般为数据值型数据，其中有些度量值不可以汇总；些可以汇总起来，以便为分析者提供有用的信息。量度是要统计的指标，必须事先选择恰当，基于不同的量度可以进行复杂关键性指标(KPI)的设计和计算。

### 3. 确定事实数据粒度

确定量度之后，需要考虑该量度的汇总情况和不同维度下量度的聚合情况.例如在业务系统中数据最小记录到秒，而在将来分析需求中，时间只要精确到天就可以了，在ETL处理过程中，按天来汇总数据,些时数据仓库中量度的粒度就是”天”。如果不能确认将来的分析需求中是否要精确的秒，那么，我们要遵循”最小粒度原则”,在数据仓库中的事实表中保留每一秒的数据，从而在后续建立多维分析模型(CUBE)的时候,会对数据提前进行汇总，保障产生分析结果的效率。

### 4. 确定维度

维度是分析的各个角度.例:我们希望按照时间，或者按照地区，或者按照产品进行分析。那么这里的时间，地区，产品就是相应的维度。基于不同的维度，可以看到各个量度汇总的情况，也可以基于所有的维度进行交叉分析。

维度的层次(Hierarchy)和级别(Level)。例:在时间维度上，按照”度-季度-月”形成了一个层次，其中”年” ,”季度” ,”月”成为了这个层次的3个级别。我们可以将“产品大类-产品子类-产品”划为一个层次，其中包含“产品大类”、“产品子类”、“产品”三个级别。

我们可以将3个级别设置成一张数据表中的3个字段,比如时间维度；我们也可以使用三张表，分别保存产品大类，产品子类，产品三部分数据,比如产品维度。

建立维度表时要充分使用代理键.代理键是数据值型的ID号码(每张表的第一个字段)，它唯一标识了第一维度成员。在聚合时，数值型字段的匹配和比较，join效率高。同时代理键在缓慢变化维中，起到了对新数据与历史数据的标识作用。

### 5. 创建事实表

在确定好事实数据和维度后，将考虑加载事实表。业务系统的的一笔笔生产，交易记录就是将要建立的事实表的原始数据.

我们的做法是将原始表与维度表进行关联，生成事实表。关联时有为空的数据时(数据源脏)，需要使用外连接，连接后将各维度的代理键取出放于事实表中，事实表除了各维度代理键外，还有各度量数据，不应该存在描述性信息。

事实表中的记录条数据都比较多，要为其设置复合主键各蛇引，以实现数据的完整性和基于数据仓库的查询性能优化。

## 二十、元数据

描述数据及其环境的数据。两方面用途:

首先,元数据能提供基于用户的信息,如记录数据项的业务描述信息的元数据能帮助用户使用数据。

其次,元数据能支持系统对数据的管理和维护,如关于数据项存储方法的元数据能支持系统以最有效的方式访问数据。

元数据机制主要支持以下五类系统管理功能:

（１）描述哪些数据在数据仓库中；

（２）定义要进入数据仓库中的数据和从数据仓库中产生的数据；

（３）记录根据业务事件发生而随之进行的数据抽取工作时间安排；

（４）记录并检测系统数据一致性的要求和执行情况；

（５）衡量数据质量。

ODS: Operational Data Store

ODS为企业提供即时的，操作型的，集成的数据集合，具有面向主题性，集成性，动态性，即时性，明细性等特点

ODS作为数据库到数据仓库的一种过渡形式，与数据仓库在物理结构上不同，能提供高性能的响应时间,ODS设计采用混合设计方式。

ODS中的数据是”实时值”,而数据仓库的数据却是”历史值”，一般ODS中储存的数据不超过一个月，而数据仓库为10年或更多.

Data Mart

为了特定的应用目的或应用范围，而从数据仓库中独立出来的一部分数据，也可称为部门数据或主题数据（subjectarea）。在数据仓库的实施过程中往往可以从一个部门的数据集市着手，以后再用几个数据集市组成一个完整的数据仓库。需要注意的就是在实施不同的数据集市时，同一含义的字段定义一定要相容，这样再以后实施数据仓库时才不会造成大麻烦。

DDS(decision-support system)决策支持系统：

用于支持管理决策的系统。通常，DSS包括以启发的方式对大量的数据单元进行的分析，通常不涉及数据更新。

## 二十一、什么叫OLAP？用途是什么？

联机分析处理，On-Line Analysis Processing即从数据仓库中抽取详细数据的一个子集并经过必要的聚集，存储到OLAP存储器中供前端分析工具读取。

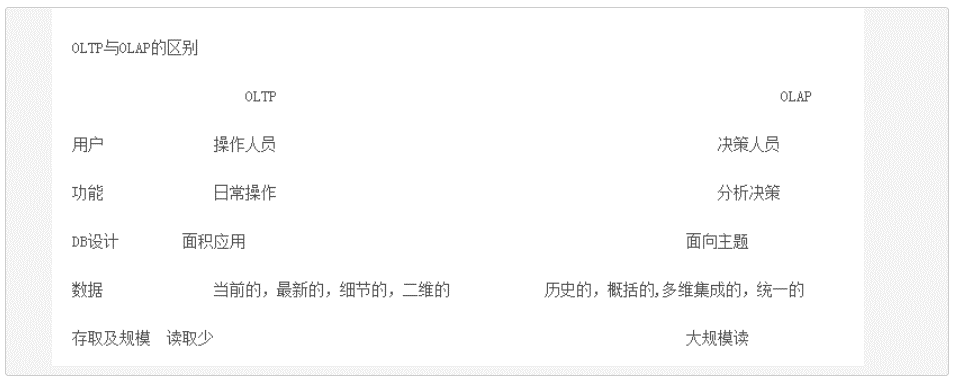
OLAP系统按照数据存储格式可以分为关系OLAP（RelationalOLAP，简称ROLAP）、多维OLAP（MultidimensionalOLAP，简称MOLAP）和混合型OLAP（HybridOLAP，简称HOLAP）三种类型。

ROLAP将分析要用的多维数据存储在关系数据库中，并根据应用的需要有选择的定义一批实视图也存储在关系数据库中。

MOLAP将OLAP分析所要用到的多维数据物理上存储为多维数组的形式，形成“立方体”的结构。

HOLAP能把MOLAP和ROLAP两种结构的优点有机的结合起来，能满足用户各种复杂的分析请求。

## 二十二、OLTP与OLAP的区别



## 二十三、事实表

事实表是包含大量数据值的一种结构。事实数据表可能代表某次银行交易,包含一个顾客的来访次数，并且这些数字信息可以汇总，以提供给有关单位作为历史的数据。

每个数据仓库都包含一个或者多个事实数据表。事实数据表只能包含数字度量字段和使事实表与维度表中对应项的相关索引字段.，该索引包含作为外键的所有相关性维度表的主键。

事实数据表中的“度量值”有两中：一种是可以累计的度量值，另一种是非累计的度量值。用户可以通过累计度量值获得汇总信息。

## 二十四、维度表

用来描述事实表的某个重要方面，维度表中包含事实表中事实记录的特性：有些特性提供描述性信息，有些特性指定如何汇总事实数据表数据，以便为分析者提供有用的信息，维度表包含帮助汇总数据的特性的层次结构

## 二十五、缓慢变化维

在实际情况下，维度的属性并不是静态的，它会随着时间的流失发生缓慢的变化。

处理方法:

1. 新信息直接覆盖旧信息；

2. 保存多条记录，并添加字段加以区分（用y,n;0,1,2或用时间来区别新旧记录）；

3. 另外建表保存历史记录；

4. 混合模式；

## 二十六、退化维

一般来说事实表中的外键都对应一个维表，维的信息主要存放在维表中。但是退化维仅仅是事实表中的一列，这个维的相关信息都在这一列中，没有维表与之相关联。比如：发票号，序列号等等。

**那么退化维有什么作用呢？**

1、退化维具有普通维的各种操作，比如：上卷，切片，切块等

(上卷汇总，下钻明细；切片，切块:对二维数据进行切片，三维数据进行切块，可得到所需要的数据)

2、如果存在退化维，那么在ETL的过程将会变得容易。

3、它可以让group by等操作变得更快

## 二十七、粒度:（granularity）

是指数据仓库的数据单位中保存数据的细化或综合程度的级别，细化程度越高，粒度就越小。

## 二十八、钻取

首先从某一个汇总数据出发，查看组成该数据的各个成员数据。

KPI（Key Performance Indication）关键业绩指标用来衡量业绩好坏比如销售这个主题,销售增长率、销售净利润就是一个KPI。

## 二十九、E T L

extract/transformation/load寻找数据，整合数据，并将它们装入数据仓库的过程。

ETL是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程，目的是将企业中的分散、零乱、标准不统一的数据整合到一起，为企业的决策提供分析的依据。

工作流--->抽取--->清洗，转换--->加载--->数据流--->业务系统ODS数据仓库

### 1. 抽取

**方法有三种：**

1）利用工具,例如datastage，informatic，OWB，DTS，SISS。

2）利用存储过程。

3）前两种工具结合.

**抽取前的调研准备工作：**

1）弄清数据是从哪几个业务系统中来，各个业务系统的数据库服务器运行什么DBMS。

2）是否存在手工数据，手工数据量有多大。

3）是否存在非结构化的数据。

**抽取中的数据处理方法:**

业务系统服务器与DW的DBMS相同时，在DW数据仓库服务器与原业条系统之间建立直接的链接关系就可以写select语句直接访问。

业务系统服务器与DW的DBMS不同时,对不能建立直接链接的话，可以将源数据导入.txt文件，在导入ODS中,或通过程序接口来完成。

对于文件类型数据源(.txt.xls)利用数据库工程将这个数据导入指定的数据库，如(oracle的SQL\*LOADER,db2的import)

**如何实现增量抽取？**

业务系统会记录业务发生的时间，我们可以用来做增量的标志，每次抽取之前首先判断ODS中记录最大的时间，然后根据这个时间去业务系统取大于这个时间所有的记录。

### 2. 清洗与转换

**清洗：**

数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。

清洗的数据种类: 1）不完整数据；2）错误数据；3）重复的数据；

**转换：**

1）不一致数据转换:编码转换(m,f;男女)；字段转换(balance,bal)；度量单位的转换(cm,m)

2）数据粒度的转换;业务系统数据存储非常明细的数据，而数据仓库中数据是用分析的，不需要非常明细，会将业务系统数据按照数据仓库粒度进行聚合。

3）商务规则的计算。不同企业有不同的业务规则，不同的数据指标,在ETL过程，将这些数据计算好之后存储在数据仓库中，供分析使用(比如KPI)。

### 3. 加载经过前两步处理后的数据可直接加载入数据仓库

用过什么ETL工具（informatica，ssis，owb，datastage），以及该工具简单讲述特点。

　DataStage是一套专门对多种操作数据源的数据抽取、转换和维护过程进行简化和自动化，并将其输入数据集市或数据仓库目标数据库的集成工具。

它有四个组件:Administrator:用来管理project和环境变量。Manager:用于job，表定义，的引导，引出。Designer:用来设计job。Direct:用运查看job运行日志。

## 三十、星形模型与雪花模型的区别？

1. 星星的中心是一个大的事实表，发散出来的是维度表，每一个维度表用一个PK-FK连接到事实表，维度表之间彼此并不关联。一个事实表又包括一些度量值和维度。

2. 雪花模型通过规范维度表来减少冗余度，也就是说，维度表数据已经被分组成一个个的表而不是使用一个大表。例如产品表被分成了产品大类和产品小类两个表。尽管这样做可以节省了空间，但是却增加了维度表的数量和关联的外键的个数。这就导致了更复杂的查询并降低了数据库的效率

## 三十一、什么叫查找表，为什么使用替代键？

(其实目的和上面一样，从基础表到缓慢维度表的过程中的一种实现途径)

替代键（alternate key）可以是数据表内不作为主键的其他任何列，只要该键对该数据表唯一即可。换句话说，在唯一列内不允许出现数据重复的现象。

## 三十二、数据仓库项目最重要或需要注意的是什么，以及如何处理？

数据质量，主要是数据源数据质量分析，数据清洗转换，当然也可以定量分析

数据仓库有两个重要目的,一是数据集成,二是服务BI

数据准确性是数据仓库的基本要求,而效率是项目事实的前提,数据质量、运行效率和扩展性是数据仓库项目设计、实施高明与否的三大标志;

**代理键：**

在关系型数据库设计中，是在当资料表中的候选键都不适合当主键时，例如资料太长，或是意义层面太多，就会用一个attribute来当代理主键，此主键可能是用流水号，来代替可辨识唯一值的主键。

在数据仓库领域有一个概念叫Surrogate key，中文一般翻译为“代理关键字”。代理关键字一般是指维度表中使用顺序分配的整数值作为主键，也称为“代理键”。代理关键字用于维度表和事实表的连接。可以避免通过主键的值就可以了解一些业务信息。

## 三十三、如何保证数据的准确性？

元数据的建设与管理是其中重要的一个环节。元数据建设的目标是打通数据从接入到加工，再到数据消费的整个链路，规范元数据体系与模型，提供统一的元数据服务出口，保障元数据产出的稳定性和质量。首先梳理清楚元仓底层数据，对元数据做分类，如计算元数据、存储元数据、质量元数据等，减少数据重复建设，保障数据的唯一性。

另外，要丰富表和字段使用说明，方便使用和理解。根据元仓底层数据构建元仓中间层，建设元数据基础宽表，也就是元数据中间层，打通从数据产生到消费整个链路。

当然，也可在粒度、规范等方面展开，见仁见智。

## 三十四、如何做数据治理？数据资产管理呢？

在明确数据治理是数据管理的一部分之后，下一个问题就是定义数据管理。治理相对容易界定，它是用来明确相关角色、工作责任和工作流程的，确保数据资产能长期有序地、可持续地得到管理。

而数据管理则是一个更为广泛的定义，它与任何时间采集和应用数据的可重复流程的方方面面都紧密相关。

其实在数仓的整个链路中数据治理的理念是渗入其中的，在ETL过程中开发人员会对数据清洗这其实就是治理的一部分，再加上后期数据资产的管理和落定都有数据治理的渗入。

## 三十五、如何控制数据质量？

1. 数据质量保证原则：完整性，准确性，数据质量，及时性，一致性

2. 数据质量方法：数据资产等级的划定

3. 数据加工过程卡点校验

4. 风险点监控：针对在线或者离线数据的监控

5. 质量衡量：故障等级的划定以及数据质量的事件的记录