## 1、数据仓库是什么

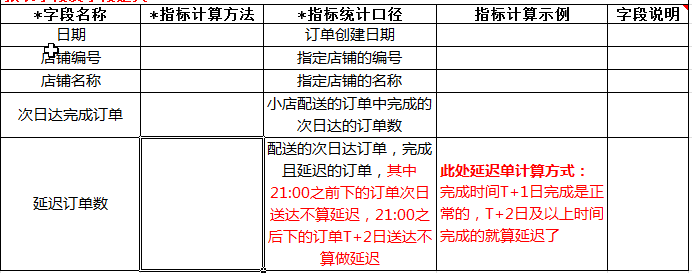
能干什么？

1、年度销售目标的指定，需要根据以往的历史报表进行决策，不能拍脑袋。

2、如何优化业务流程

案例1：

一个电商网站订单的完成包括：浏览、下单、支付、物流，其中物流环节可能和中通、申通、韵达等快递公司合作。快递公司每派送一个订单，都会有订单派送的确认时间，可以根据订单派送时间来分析哪个快递公司比较快捷高效，从而选择与哪些快递公司合作，剔除哪些快递公司，增加用户友好型。



案例2：

互联网中国需要对APP进行推广，考核的主要目标是下载安装，有些第三方渠道会对这些数据造假，比如某个渠道在凌晨批量下载，点赞操作，操作步骤一致。通过数据分析，分析出应用的名称和安装时间，来判断一个渠道的是否优质、是否作假。

## 2、数据仓库的特点

1、数据仓库是面向主题的，比如商品主题，订单主题。（领导关注那些方面）

传统数据库面向应用，提供什么功能。数据仓库面向分析，提供那些主题的分析。

从规模来讲依次是，数据仓库、数据集市、数据报表。

2、数据仓库是集成的，数据源是分散的，来自不同的应用。数据仓库中的综合数据，不能从源数据中直接得到，一般会经过etl过程（数据抽取、数据转换、数据加载）。数据抽取一般会定时的进行抽取，避免对业务系统造成影响，一般叫做T-1抽取、T+1抽取。

目前企业对数据的实时性要求越来越高，比如实时监控一个实时的活动效果，并根据效果进行不同策略的营销手段，保持活动的效果。

3、数据仓库是不更新的，数据仓库反应的是一段相当长的时间内的数据内容，主要的操作集中在数据查询上。 一般数据结果计算出来之后，特别是明细数据，会存放在关系数据库中，因为主流的报表工具都支持数据库。

对数据库的查询，最基本的操作是创建索引，比如300万的数据根据手机号查询需要十几秒，创建btree索引之后，需要几十毫秒。

4、数据仓库中的数据是随着时间而变化的。

## 3、数据仓库的发展历程

第一阶段：简单报表阶段

解决日常工作中业务人员需要的报表，为领导生成简单的汇总数据

大数据库+前段报表的形式

第二阶段：数据集市阶段

按照不同部门、不同业务人员的需要，进行一定的数据采集，整理，并进行多维度报表的展现，能够提供对特定业务指导。

业务部门对数据的口径不一致，产生的汇总数据不一致。对大领导来看，就需要一个标准的口径。

第三阶段：数据仓库阶段

对整个企业的数据进行采集，整理，并且能够按照各个业务部门的需要，提供跨部门的，完全一致的业务报表数据，能够通过数据仓库生成对业务具有指导性的数据，同为为领导决策提供全面的数据支持。

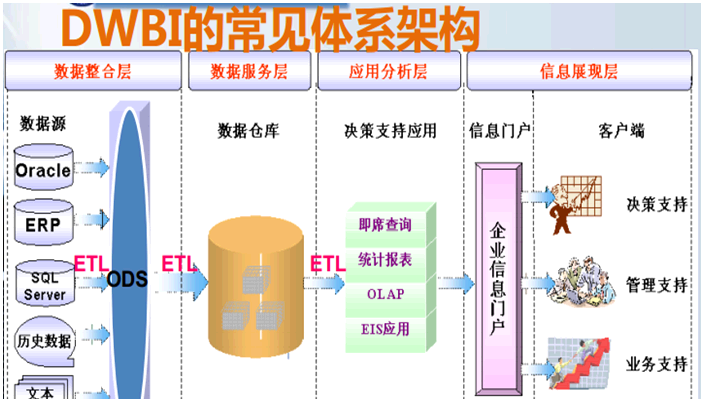
数据仓库和数据集市的区别，在于数据模型的支持。沉默用户的计算，没有沉默字段的标识，需要些复杂的sql，有沉默字段的话，一个简单的sql就能搞定。

其他：城市商品表

## 4、数据库与数据仓库的区别

1、数据仓库是集成的，数据库为单一的业务提供服务。

2、BI结构：数据整合层、数据服务层、应用分析层、信息展现层



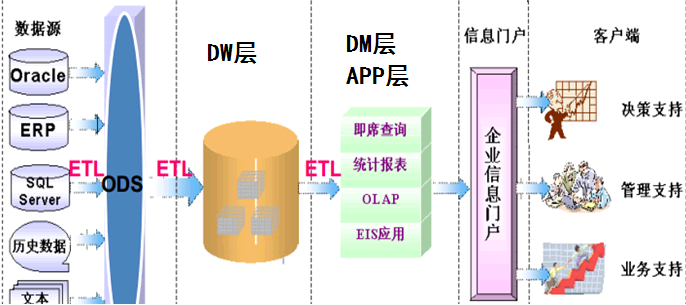
3、数据层库结构

ODS(临时存储层)，一般都是贴源设计、业务数据库是什么，ODS层就是什么

PDW(数据仓库层)，将年月日，拆分成年、月、日字段，一般采用Int类型;通过ODS层到DW层的etl脚本对数据进行数据清洗，进行设计。分部门根据业务需求进行设计。如果没有业务需求，就根据源系统的表结构和自己建模经验去处理。

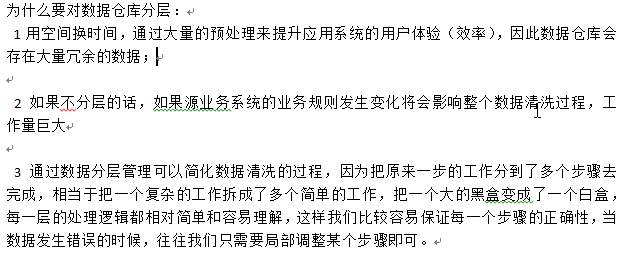
DM(数据集市层)，维度建模，星形模型，雪花模型。需要什么数据就去拉取什么数据。

APP（应用层），报表展现，需要的数据，与DM层处于同一级别。



4、ODS层分为增量更新或者全量更新；PDW层一致的、准确的、干净的数据，一般遵循数据库三范式设计。

11、为什么数据仓库需要分层？



## 5、数据质量检查

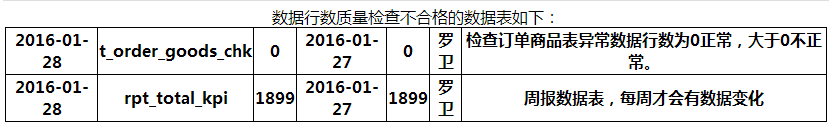
保证报表数据的正确性、稳定性，通过告警机制尽可能快的发现异常、尽可能快的解决问题。

出错的次数太多之后，领导会对你失去信心，该辞职了。

检查方法：

1. 数据行数据的比较
2. 行数有变化，但是指标有变化。对领导关系的重点指标进行筛选。





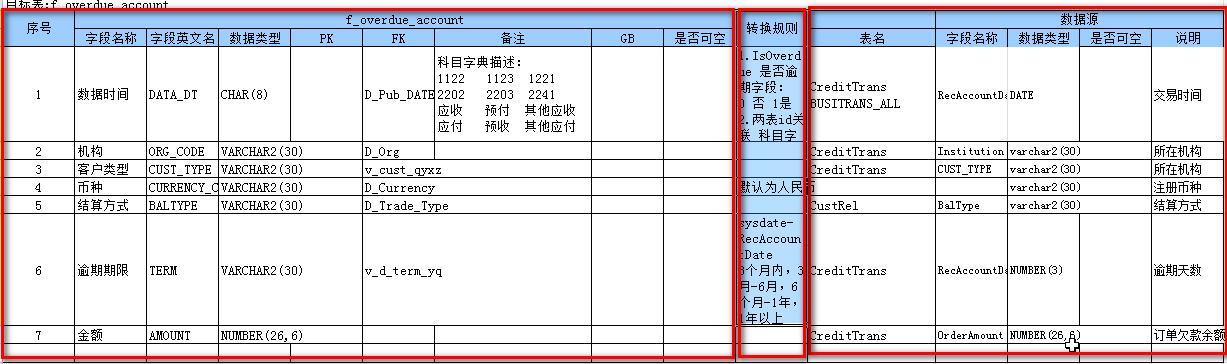
在领导发现问题之前，解决问题。

告警

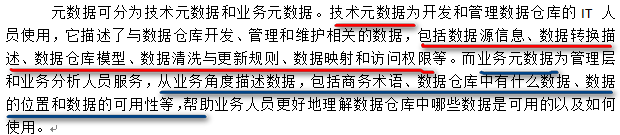


## 6、元数据管理

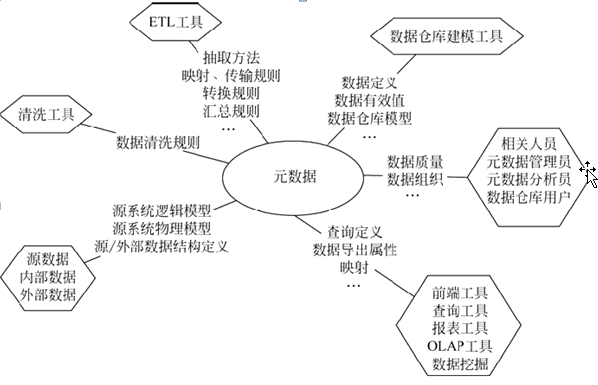
元数据：数据的数据，记录数据从哪里到哪里去，中间如何转化的。



元数据分为技术元数据和业务元数据

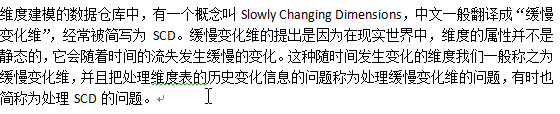


元数据中的数据都有哪些？



## 7、数据仓库命名规范

## 8、缓慢变化维



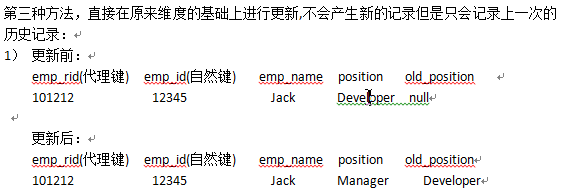
如何解决缓慢变化维带来的影响？



**上图中position有变化**

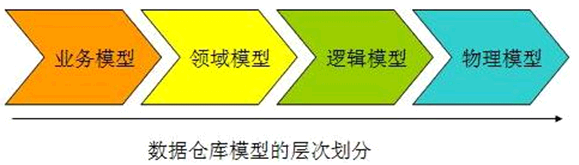


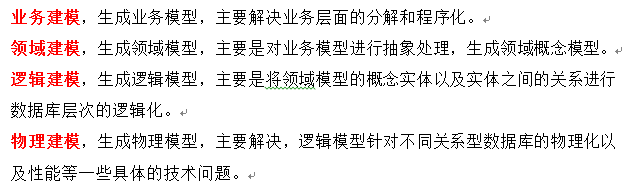
**上图中多了一条记录**

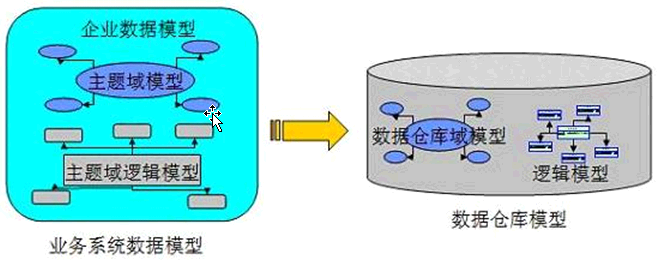


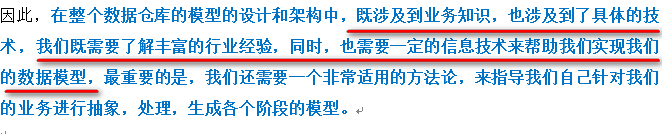
**多一个字段，用来存放以前的position**

## 9、数据仓库建模

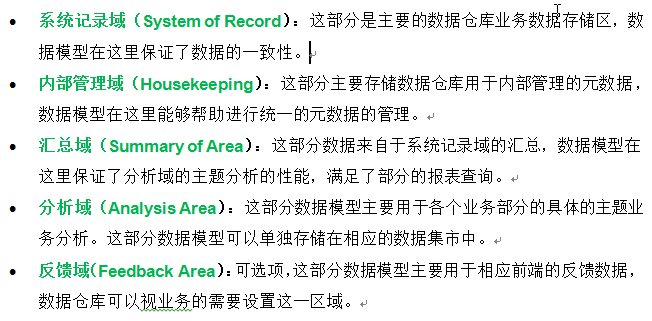
****

****

****

****

## 10、数据仓库五大核心模块



## 11、实体建模和维度建模

## 12、O2O业务建模案例

