# 数据仓库开发方法

## 1 、数据仓库风险

谈到数据仓库的开发方法，我们必须首先了解进行数据仓库的开发存在什么样的风险，它主要有三大类风险：技术风险、工程管理风险和业务风险。

### 1.1 技术风险

技术风险是一种开发人员不能使技术正确地发挥的风险，主要反映在对技术不了解，对技术不精通，不能解决开发过程中的技术问题等方面。

可以通过如下手段来减少这方面的风险：

- 经验：让有数据库或数据仓库经验的员工参加工程的开发，使用熟悉的开发工具。
- 培训：对不熟悉技术的开发人员进行培训。
- 避免使用未经证明的技术：尽量不要采用新的技术，如果一定要使用新技术，一定要先做两方面的工作，一是对新的技术的性能做出评估，看是否真是工程开发所需要的技术。二是在新技术不能达到预期要求的情况下采用一个备份的技术。
- 概念试验：也就是对工程中的关键技术进行前期的试验，确认其技术可行性。
- 结构复查：让开发团队以外的人员参与系统技术结构设计，并发表评论。

### 1.2 工程管理风险

即使开发团队采取了正确的技术，并正常地使用了它，但还是存在不能按时或按预算完成工程的开发和实施。可以通过如下的手段来克服这种风险：

- 经验：具有数据仓库构建及其开发任务方面的知识能够让你和管理数据仓库的开发工作。
- 方法：一个强大的方法会起到工程管理和工程团队路标的作用，指导开发人员如何前进。
- 具有献身精神的工程管理员：一个工程管理员的任务包括：制定工程的工作计划、给团队成员下达任务及任务完成的期限，跟踪每个任务的进度，分配工程所需的资源。要确定工程是按计划进行的。
- 需求变更控制：一定要控制好需求变更。

### 1.3 业务风险

业务风险是指工程完工后却没有人使用它。可以通过如下的手段如下处理它。

- 开发工作始终让用户参与：要避免在收集了用户需求后，开发工作就脱离了用户。因为用户在提出需求时的想法和最终在屏幕上看到的想法可能会是不一致的。
- 专注业务过程的实质：不同于业务系统，数据仓库的使用是一个可选的系统，也就是说用户可以不使用它。一定要让数据仓库真正被最终用户使用起来，能帮助他们开展工作。

**2、方法概述**

对付数据仓库风险的最好方法是采用一套成熟的方法学(methodology)，方法学可以看作一本开发数据仓库的食谱，它列出你要执行的步骤，提供一些信息帮助，为这些步骤作计划和预算。好的方法学包含人们在数据仓库构建实践中所积累的成功和失败的经验。

当使用数据仓库方法时，确信你明白其中的每一步，它会产生什么结果，以及按那样交付为什么是重要的，方法学是可以裁剪的，可增可减其中的步骤以适应具体工作的需要。

此处提供的方法学由 6 个阶段组成：

| 阶　段 | 目　　标 |
| --- | --- |
| 设想阶段 | 证明数据仓库是否正确，并定义一个数据仓库战略，通常这一步只执行一次，它会产生一系列的数据仓库版本，在方法学中的蓁步骤教师针对这些而需要执行的。 |
| 探索阶段 | 为一个功能领域获取详细的用户需求。 |
| 体系结构设计阶段 | 设计解决这些需求的技术方案 |
| 构造阶段 | 建立整个系统的原型 |
| 实现阶段 | 将系统交付用户使用 |
| 审查和反复阶段 | 收集所有反馈意见，并作出进一步的改进。 |

## 3 、第一阶段-设想阶段

设想是最初的设计阶段。

设想背后的观点是为公司决策支持和企业数据仓库(EDW)建立一个长期计划，它将产生一系列的版本，因此它通常只产生一次。

设想一般是对企业一级的数据仓库进行研究计划，没有必要为一个数据集市进行。

### 3.1 设想阶段目标

- 确定在数据仓库及其相关技术的投资是否有保证。
- 确定数据仓库及相关技术在哪些方面能最好地服务于整个公司。
- 为企业数据仓库建立一个长期的计划。

### 3.2 设想的关键交付项

设想阶段的关键交付项包括：一份设想报告、企业数据仓库的长期工作计划和一个企业数据仓库的实体级数据模型。

1、设想报告

是一个描述和总结本阶段的调查结果报告，主要有如下一些信息。

＞对企业已经 IT 结构的一个描述，包括已存在的事务处理系统及决策支持系统。

＞对用户的需求的一个描述。

＞……

2、长期工作计划

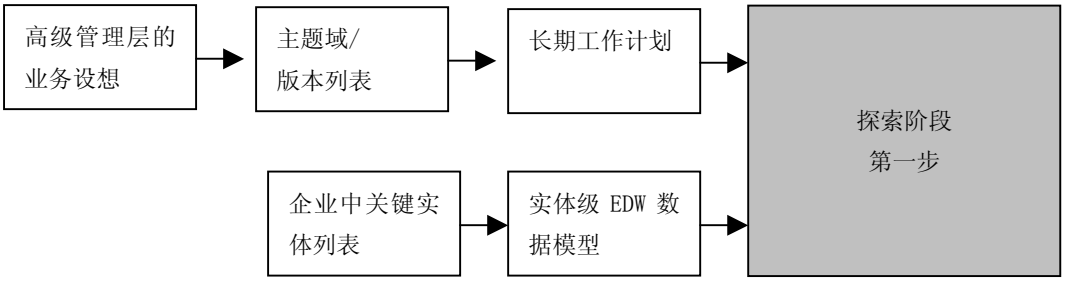
这个长期的工作计划主要是让项目涉及的企业部门了解此项目何时关系到此部门，并了解项目的大致进度计划。

3、企业数据模型

我们一般先建立企业部门的数据集市，然后建立需要为市场数据集市提供数据的部分综合层。

我们必须建立一个高度概括的企业数据模型版本，获取关键实体和它们之间的关系。

### 3.3 设想流程



## 4、第二阶段-探索阶段

探索阶段是方法学中进行详细的需求收集阶段，从而产生不同的版本。

### 4.1 探索阶段目标

- 为每个正在考虑中的数据仓库版本收集详细的需求。
- 将系统用户纳入开发过程。

### 4.2 探索阶段的关键交付项

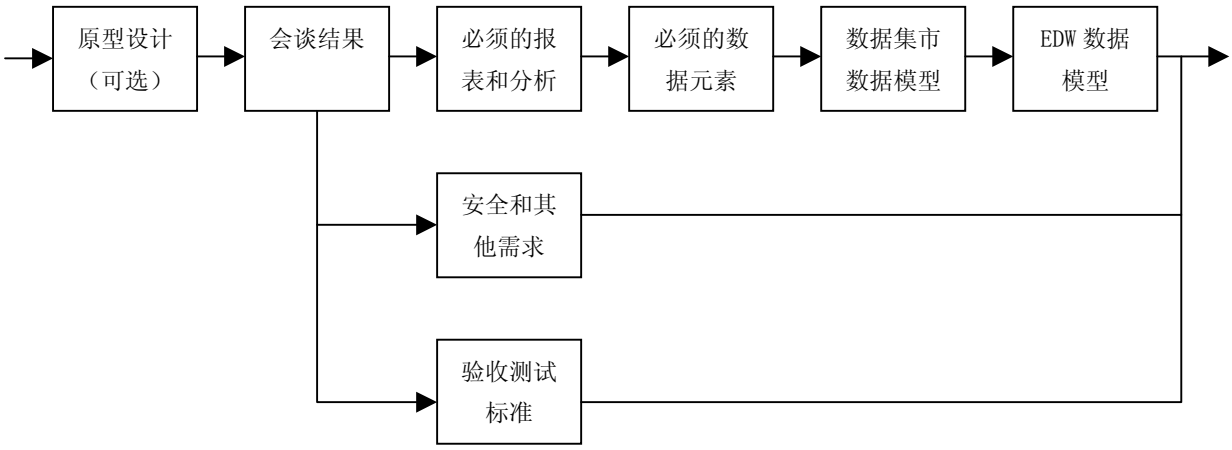
1、 报表和分析的布局：用户首先需要的是一些固定格式的报表，然后才是一些可定制的报表。通过报表详细布局帮助开发人员了解什么是用户需要的、有关的数据元素。

收集报表信息的方法：与关键用户一些，然后让他们画出他们所需要的报表草图，包括如下报表信息：报表名称、报表字段、过滤器、排序顺序、需要的小计和总计、报表的页眉和页脚、报表周期、报表用途的描述及其他相关信息。

2、 逻辑数据模型：逻辑数据模型数据集市星形模式和为星形模式提供数据的 EDW 模型部分。这些模型应包含用于建立报表所需要的每个字段，满足用户提出的分析请求。

3、 用户验收测试标准：在此阶段制定用户验收测试标准，保证后续的开发工作是有的放矢的。

### 4.3 探索阶段流程



## 5、第三阶段-体系结构设计阶段

在了解了用户的需求后，就必须进行工程的体系结构设计来满足用户的需求。

### 5.1 体系结构设计阶段目标

为数据仓库开发健壮的、高层次的、详细的设计，开发团队将在工程构造阶段按照这个设计建立数据仓库。

### 5.2 体系结构设计阶段的关键交付项

1、物理数据模型：

在探索阶段产生了包含用户需求的逻辑数据模型，需要在体系结构设计阶段转化成物理模型，具体的做法是确定逻辑数据模型在硬盘上的结构和布局，更具体的做法是指将实体变成数据库中的表、以及进行必要的重建构造、定义所需要的索引、开发磁盘上分割数据的计划。

2、字段级映射

字段级映射是重要的体系结构设计，要为物理数据仓库数据库中每个建立一个清单，然后为每个字段编制数据源及数据规则文档，最好形成一个电子表格。

不仅要映射源文件和数据库迁移到数据仓库的数据，在有数据集市的情况下，也建立从数据集市到数据仓库的数据映射关系。

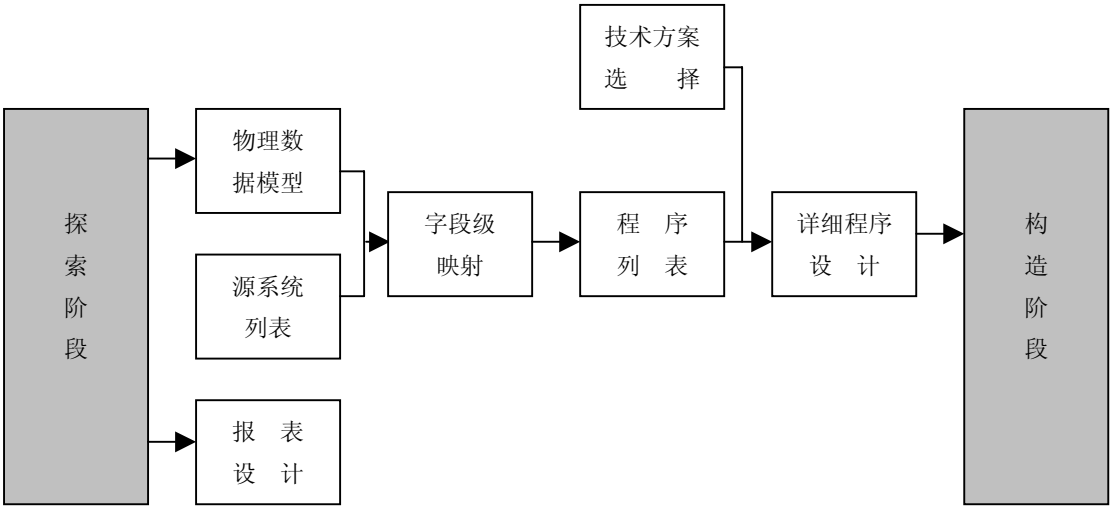
3、技术方案选择

必须在详细的程序设计开始之前完成数据仓库开发工具的选择，并形成开发团队的技术标准。

4、详细的程序设计

详细地设计构造阶段要开发的每个程序是估计工程构造阶段需要花费的人力和财力的重要方法，因此要进对每个程序进行认真而翔实的设计，并形成相应的技术文档，以指导后续的开发。

5.3 体系结构设计阶段流程



## 6 、第四阶段-构造阶段

构造阶段是数据仓库开发中最耗时、费用最大的阶段，在本阶段将实现体系结构设计阶段所有的设计。

### 6.1 构造阶段目标

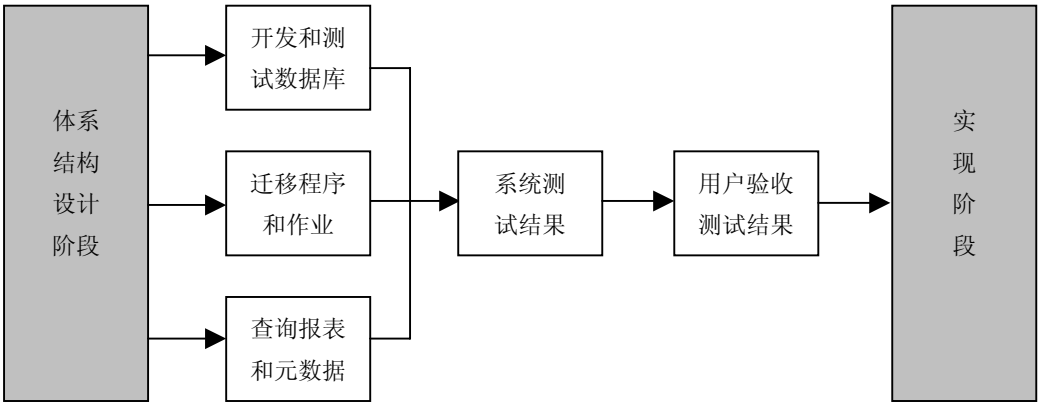
阶段的目标是建立和测试系统，从而使它成为产品，并能在实现阶段将它交付给用户。

### 6.2 构造阶段的关键交付项

构造阶段最后交付的是一个能交付给最终用户的系统，这个系统由数据库、数据迁移作业（ETL）和报表查询界面组成。

要进行充分的测试，然后纠正发现的问题，确保提交给最终用户的系统是基本稳定和正常的。

### 6.3 构造阶段流程



## 7、第五阶段-实现阶段

在系统投入使用，并将访问工具交付给用户。

一个更重要的工作是对用户进行系统培训，确保最终用户如何熟悉地掌握新系统的使用方法。

### 7.1 实现阶段的目标

两个主要目标：

- 将系统投入使用，确保使用过程中它能可靠地运行。
- 确保用户能熟练地使用和操作系统。

### 7.2 实现阶段的关键交付项

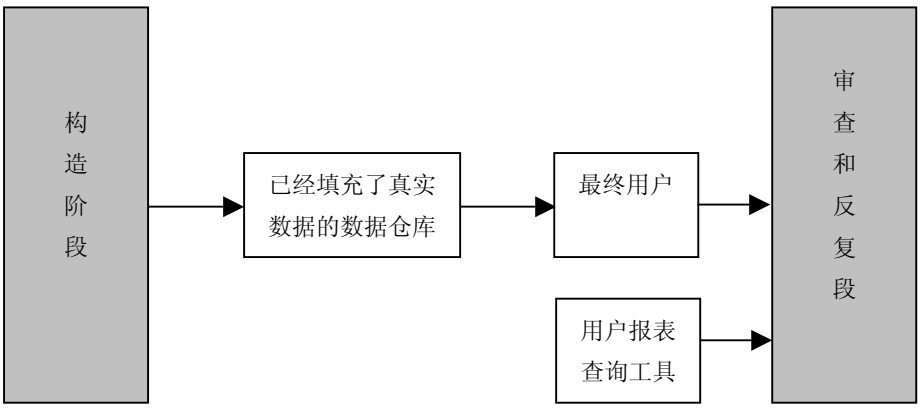
一个运行的系统和一个经过培训的、能熟练地使用系统的用户群。

对于用户培训来说，要做如下的工作：

- 编写培训文档
- 组织用户培训

关键是要让用户了解数据的含义以及让用户能够自己自定义报表，进行灵活的报表查询。

### 7.3 实现阶段流程



## 8、第六阶段-审查和反复阶段

在将产品提交给最终用户后，必须对系统运行的状况和性能情况进行监视，以进行及时的错误修改和性能调整。

### 8.1 审查和反复阶段目标

审查和反复阶段的目标是：

- 确保交付的系统的性能不断满足用户的需求。
- 为用户提供一个反馈他们对系统进行改进意见的渠道。

### 8.2 审查和反复阶段关键交付项

为了不断改进系统，开发人员必须开发许多脚本和文档工具。

### 8.3 审查和反复阶段流程