D3A - Metadata & ETL

Metadata

3 Types Of Metadata

Operational Metadata

The information about source operational database.

Extraction and Transformation Metadata

extraction time, update frequency, extraction method, business rules for extraction…

End-User Metadata (Metadata For End-User Read)

Map of the data warehouse (what's this, what is the meaning of…, how to find…) should use business term.

Metadata Example

- Alias (Client, Account)
- Definition (A person or Organization that purchases goods or services from the company)
- Table structure (columns of Customer table, and their data types)
- Remarks (the Customer dimension includes current and past customers)
- Source system (Sale department, marketing department)
- Create date (Feb 3, 2015)
- Last update date
- Update cycle (weekly)
- Last full refresh date (Dec 31, 2016)
- Full refresh cycle (every 6 month)
- Data quality reviewed date
- Last de-duplication date
- Planned archival cycle (every 6 month)
- Responsible user (Slim Dusty)

谁需要 Metadata

- User needs metadata to know the definition of table in DW, know what is available in the DW
- Staff working for ETL needs to know data sources and the transformation of the table
- DBA needs to know the structure of DW, load/update cycle…
- Metadata is needed in administering of DW

<table>
<tr><th></th><th>IT Professionals</th><th>Power Users</th><th>Casual Users</th></tr>
<tr><td rowspan="2">Information Discovery</td><td colspan="2">Databases, tables, columns</td><td>List of predefined queries and reports, business views</td></tr>
<tr><td>Server platforms</td><td></td><td></td></tr>
<tr><td rowspan="4">Meaning of Data</td><td>Data structures</td><td></td><td>Business terms, filters, data sources, conversion, data owners</td></tr>
<tr><td></td><td colspan="2">Business terms</td></tr>
<tr><td colspan="3">Data definitions</td></tr>
<tr><td colspan="2">Data mapping, cleansing functions, transformation rules</td><td></td></tr>
<tr><td>Information Access</td><td>Program code in SQL, 3GL, 4GL, front end applications, security</td><td>Query toolsets, database access for complex analysis</td><td>Authorisation requests, information retrieval to desktop applications like spreadsheets</td></tr>
</table>

Providing Metadata

- Metadata should be a roadmap of DW
- Some DW metadata comes from source system metadata, and it is added to the ETL process
- Must have process to:
  - Standardize metadata across systems
  - Revise metadata across systems
  - Exchange metadata across systems
  - Allow users to query metadata

Source of Metadata

From Source System

- Data modelling of source system 数据模型
- Data dictionary of source file

- Document of source file
- Layout of source file
- Program specification

In Data Extraction Process

- Data on the source platform　要取哪些数据
- Layout and definition of the selected source data　数据的布局
- Field definition of the source data table　每列的定义
- Rules for standardizing data type and length　如何统一确定每列的数据类型
- Data extraction schedule　抽取数据的日程表
- Extraction method for future changes　未来新数据怎么抽取
- Process of merge future data to existing data　未来新数据怎么和老数据合并

In Data Transformation Process

- Specification of mapping extracted data to staging area　取到的数据怎么放进来
- Conversion rule
- Default value for missing data
- Business rule for validity checking
- Arrange rule
- Audit trail details

In Data Loading Process

- Specification of mapping data from staging area to DW
- Rule for manage keys in DW
- Audit trail for data staging to loading
- Schedule for future loading

About Data Storage

- Data modelling of DW
- Subject of data marts
- Table and column definition
- Physical file

About Information Delivery

- List of query
- List of report
- List of tool
- Data model for OLAP
- Schedule for data load into OLAP

Manage Metadata

- Every software tool has its own metadata, hard to make them work together
- There is not industry standard about metadata
- Sharing metadata or metadata in a system
- No widely-accepted way to transfer metadata from source system to staging area and from staging area to DW
- Version control for metadata
- Unify metadata is a huge task

Metadata Repository

- We must store both technical and business metadata

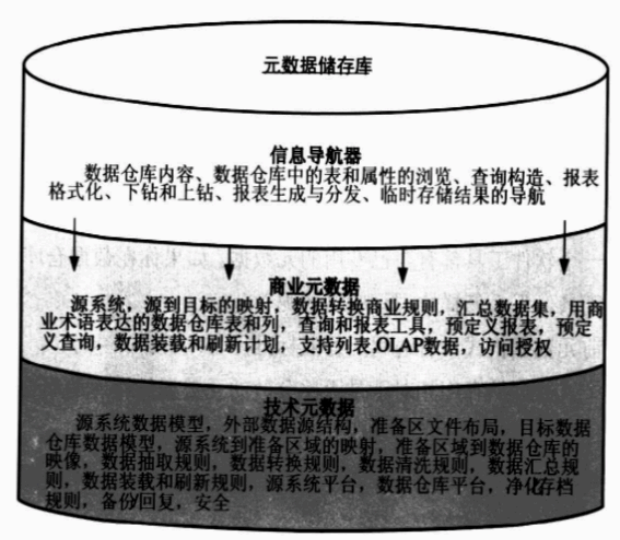

图 9-9　元数据储存库

Metadata Standard

1 main standard of metadata for DW: Common Warehouse Metamodel (CWM)

- Main purpose of CWM is to enable easy interchange of DW and business intelligence metadata between DW tools, DW platforms and DW metadata repositories in distributed environments.

ETL

Requirements 或者说 Steps

- 决定 target data
- 决定 data sources
- Prepare staging space
- Prepare mapping data from source to staging space
- 建立comprehensive data extraction rule
- 决定 data transformation and cleaning rule
- Plan for aggregation as needed
- Create procedure for data loading
- Carry out ETL for dimension tables
- Carry out ETL for fact tables

Extraction

数据在操作性系统中的储存方式

- Current value in operational system
- Periodic status in operational system

抽取方式

- **Extract static data　(抽取当前的全部源数据)**
- Extract revision data　(抽取what is changed)

Extract Revision Data　(抽取 Change)

- Immediate data extraction (当源系统中发生了新交易, 立即抽取新交易到DW)
  - **Capture from 源系统的transaction logs**　(只适用于database)
  - **Capture from database triggers**　(设置让新交易触发trigger, 只适用于database)
  - **Capture from source application**　(modify source application to support extract the change, may downgrade the source application performance, 适用于老系统)
- Deferred data extraction　(定时抽取变化,比如每天半夜检查有哪些变化)
  - **Capture based on timestamp**　(比如每天半夜抽取timestamp为前一天半夜到现在的数据变化,源系统必须有timestamp)
  - **Capture by comparing files**　(比较今天的和昨天的源数据文件, 得出变化. 适用于没有transaction log和timestamp的源数据)

Task

1. 数据源确认　(为DW中的每个数据元素找到对应额数据源)
2. 抽取方法确认　(人工?基于工具?)
3. 抽取频率
4. 抽取时间
5. 工作顺序
6. 异常处理　(missing value, error in program)

Transformation

基本操作 Basic Task

1. Selection　(在抽取后的数据上可进行再次选择)
2. Splitting/Joining
3. Conversion
4. Summarization　(DW中无法保存最低粒度的数据时,必须summarize. eg. 7-11不必在DW中储存每一笔交易的数额,只要某家店某天某个产品的总销售额就行)
5. Enrichment　(eg. address分成多段)

功能Functions (Major Transformation Tasks) 通过基本操作组合来完成

1. Format revision　(改变data type and length)

2. Decoding of fields (统一不同源系统中对同一个field的名字,比如ItemID和ProductID. 统一不同源系统中的表达方式,eg.男女用F/M还是0/1?)
3. Calculate and drive values as needed (eg. 计算cost, margin, AgeGroup)
4. Splitting single field (split name, split address)
5. Merge information (从不同的数据源中merge关于一个order的所有信息)
6. Character set conversion (convert every thing to one character set)
7. Measurements unit conversion
8. Date/Time format conversion
9. De-duplication (remove duplicate rows)
10. Create new keys for DW

2 Types of Difficulty

1. Entity identification problem (need algorithm to identify a physical customer from multiple sources)
2. Multiple sources problem (eg. Order中有单价,product中也有单价, 两个数字不一样的时候, who is correct?)

Loading

- Full loading 会花很多时间
- DW在loading时必须是offline
- 需要预先找一个DW的空闲时间去loading
- 安排test loading来预估full loading的时间
- full refresh table if more than 15%~25% rows need to be updated.

3 Loading Types

1. Initial loading (PS可以分成几个sub-loading)
2. Incremental loading
3. Full refresh (对于一个或者多个table, remove everything and load again)

4 Modes to Copy Data Into DW Tables

1. Load (remove everything and add data)
2. Append (keep all exiting data and add)
3. Destructive merge (add data, if key existed, overwrite old record)
4. Constructive merge (add data, if key existed, keep and mark the old row and add the new, 会有两个相同的key了!)

ETL Tools

- Data transformation engine (根据定义的时间间隔,执行pre-defined ETL)
- Data capture by replication (使用database的transaction log/trigger, 把transaction log/trigger表现的变化near实时copy到staging area)
- Code generator (eg. pentaho)
- 以上都不但可以进行ETL, 还能manage metadata