D1A - Database & Data Warehouse’s Approaches, Architectures, Components

Strategic Information

Complexity of Business

- Global environment
- Information drives the bottom line
- Strategic information can bring competitive advantage

Who Need Strategic Information

- Executive and manager levels
- Those who formulate business strategy, establish goals, set objectives and monitor results.

What Is Strategic Information

Information that helps executive and manager levels people to make decision about formulation and execution of business strategy and objectives.

- It is not daily operational business information.

5 Characteristics of Strategic Information

- Integrated: an single enterprise-level view of information
- Data Integrity: accurate and conformed to business rules
- Accessible: can be easily, intuitively access, can response quickly to query.
- Credible: values are credible and trusted
- Timely: data must have correct timestamps.

Difficulty in Data Integration To Get Strategic Information

- Many incompatible source databases
- Every department has their own best database (local optimum)
- Data quality may be poor
- Data in wrong format for analysis
- 导致: enterprise cannot get benefit form its data and technology assets and cannot get strategic information form its data.

Good New To Integrate Source Database

Technology is developing: computing technology, human-machine interactivity, data availability, network speed, hardware cost.

Operational vs. Decisional Databases

Operational Database (Transactional Database)

Focus on supporting daily operations.

Decisional Database (Data Warehouse)

Focus on getting information at higher level, have larger scope.

	Operational	Decisional
Data content	Currently-valid data only	Historical, summarized data
Data structure	Optimized to write and read	Optimized to response to query
Access frequency	Very high	Medium
Access type	Write, delete, read	Read only
Usage is	Predictable, repetitive	Ad hoc, unpredictable,
Response time	Sub-seconds	Seconds to minutes
User	Many (front-line staff)	Few (manager-level people)

What Is a Data Warehouse

A data warehouse is a kind of database that has following 9 features:

- It is designed for analysis tasks and provide strategic information.
- It take many (internal and external) source databases and make them into an integrated one.
- It store both current-valid and historical data.
- It should be easy and intuitive to use.
- It is read frequently (read-intensive).
- It allows user to run query and get response quickly.
- It allows user to initiate reports.
- It is helpful to analyse for a long time.
- It is stable and will be updated at fixed frequency.

Date Warehouse Defining Features

Meet Management Needs

- Data warehouse should support management needs.
- Data warehouse should has a simple and easy-to-navigate structure.
- Data warehouse should response to query quickly.

Data Is Subject-Oriented

- The data warehouse is focused around a particular subject (like sale, customer, employee)
- Data is integrated from different applications (like sale, customer, product, branch source database)
- Data in the data warehouse can across different applications (like sale, customer, product, branch)

Data Is Integrated

- Data come from different sources (different file layout, file naming rule, format, language)
- Data must be convert to a common format.
- Data form various sources must be validated before loading them into data warehouse (to ensure data quality)
- Data quality is important to the credibility可靠度 of data warehouse.

Data Is Time-Variant

- Data warehouse contains both current and historical data to support analysis tasks and decision-making.
- Historical data is store as snapshots of current data with timestamps in data warehouse.
- Storing historical data allows:
    - Analyse the history
    - Find the relationship between past data and current data
    - Forecast the future

Data is Non-Volatile (Stable)

- Data warehouse is not used to store daily operational data, it is stable.
- Data warehouse update periodically.
    - Can be secondly, hourly, daily, weekly…
    - Different data item can have different updating frequency.
- User can read only
- Updating is down by automatically ETL process and periodically by DW administrators.

Data Has Granularity

- The finer granularity, the more detailed data.
- The smallest level of granularity of data called “the grain” of the data.
- Data is most detailed in operational database.
- Summarized data can be calculated from detailed data.

Bottom-Up Approach vs. Top-Down Approach

Here, data warehouse is composed by data marts. There are 2 approach to build a data warehouse: top-down (data warehouse → data marts) and bottom-up (data marts → data warehouse)

Data Mart vs. Data Warehouse

Date Mart is a smaller-scope version of data warehouse, used by a department (like marketing) only. Date Mart contains historical data too.

Data Mart	Data Warehouse
The start of bottom-up appraoch	The start of Top down approach

Data Mart	Data Warehouse
Department-wide (only support query from marketing department	Enterprise-wide
Oriented to single business subject	Oriented to multiple subject (collection of all Data Marts)
Integrate subset of source data	Integrate all source data
Hold summarized data	Hold detailed data
Allow user to access specific set of data only (faster response time)	Allow user to access large group of records
Build by dimensional modeling using a star schema	Build by ER modelling

Reason To Build Data Mart

- Easy access to frequently needed data
- Faster query response time than full data warehouse
- Easy to build than full data warehouse
- Lower cost than full data warehouse
- Potential users can be defined more clearly than full data warehouse
- Contain essential data for a particular business only, so data mart is less clutter.
- Provide collective view by a group of users

Bottom-Up Approach

我们在Assignment中使用的approach!!需要使用dimensional modelling!!

(1) Kimball's approach
(2) Data marts are build first (using dimensional modelling)
(3) Data warehouse is a collection of conformed data marts, data marts are unionized by **conformed dimensions**
(4) Data marts store data at grain level (then data can be summarized for analysis as need)
(5) 产生较少tables, 需要较少join
(6) Denormalized, 有redundancy
(7) End-user 易用

Conformed Dimension

A conformed dimension is a dimension table which is linked to more than one fact tables (data marts).

在我们的Assignment中,如果我们只画了Sale fact table以及它周围的dimension tables, 那么可以说我们建立了一个data mart, 因为它只能回答关于Sales的问题.

如果我们画了Sale和EmployeeSale两个fact table, 那我们就是建立了一个data warehouse, 因为这是两个data mart的union, 而且它可以回答关于Sales之外的问题,比如说human resource department needs performance of salespersons.

Advantage

- Faster and easier to build
- Faster to see return on investment
- Less risk to failure
- Less starting IT expertise
- Incremental development
- Project team can learn and grow

Disadvantage

- Narrow view of each data mart
- Redundant data in different data marts
- May has inconsistent data in different data marts
- Possible hard-to-manage interface
- May be complex to union different data marts

Top-Down Approach

(1) Inmon's approach
(2) Data warehouse is build first (ER modelling)

(3) Data warehouse feed department-wide data marts

(4) Enterprise-wide data warehouse store grain level of data

(5) 产生很多tables, 需要很多join

(6) Normalized, 没有redundancy

(7) End-user 难用

Advantage

- Enterprise view of data
- Can provide **an appropriately architected structure** (not just a union)
- Central storage of data
- Centralized rules and control
- Still possible to see quick results if development is iterative.

Disadvantage

- Take longer time than a data mart
- Higher risk to failure
- Require good IT skills
- Require good cross-functional skills
- High cost and may have to wait a long time for return

Hybrid Practical Approach

- We can combine 2 approaches.
  - Still consider the big picture approach, but build conformed data marts by priority.

Steps

1. First plan the entire data warehouse and define **requirements at enterprise level.**
2. Choose an **architecture** for the entire data warehouse.
3. Conform data content
4. Implement the entire data warehouse one **supermart** by one supermart.

有Architecture结构连起来的Data Marts可以叫做Supermarts. 散落的Data Marts就只能叫Data Marts.

Types of Architecture of Data Warehouse

Independent Data Mart

- Common to used when departments develop their own data marts
- But these data marts are independent and don't provide a single version of truth (data marts之间完全没有关系!!)
- These data marts have inconsistent data, inconsistent definitions and inconsistent standards
- It is difficult to analyse across these data marts.

Federated

- This architecture is used when enterprise already has complex legacy systems and it is hard to rebuild. So just integrate them.
- keep existing decision-support systems
- Data may be **physically or logically integrated** (shared key fields, global metadata)
- however, there is NO overall data warehouse

Data Mart Bus (With Linked Dimension Table)

我们在Assignment中使用的architecture!!

- Kimball's conformed supermarts approach
- Firstly **determine requirements for specific business subjec**t (like sale, shipment, billing…)
- Build your first data mart (fact table), then second (fact table)…using **dimensional modelling**
- The resulting collection of data marts will provide an enterprise-view of data by **conformed dimension** (tables) among various data marts.
- This architecture adopts an enhanced **bottom-up** data warehouse develop approach

Centralized

- Firstly determine requirements for **entire enterprise**.
- Only a single centralized database (3NF) (只一个表!!)
- **Faster and easier** to implement
- Analytics down on the centralized data warehouse
- There is no dependent data marts.

Hub and Spoke

- Inman’s Corporate Information Factory approach
- Firstly determine requirements for **entire enterprise**.
- A centralized data warehouse (3NF) + additional dependent data marts
- Data content in these additional dependent data marts is obtained from the centralized data warehouse (spoke中的数据是从hub中拿来的!!)
- Dependent data marts may be developed for various purposes, like for a department-inside analysis, for a business process, for data mining…
- Most query done at the dependent data marts, few query down at the centralized data warehouse (大部分query在spoke中完成!!)

10 Factors Affecting Architecture Selection

Information Interdependence Between Department

If high, choose enterprise-wide architecture (hub and spoke or centralized)

Urgency of Need for Data Warehouse

If urgent, choose faster architecture (centralized, independent data mart, federated, data mart bus)
总之不能选 hub-and-spoke

Resource

IT people, business people, budget
资源不足, 选independent data mart 或者 centralized architecture.
资源充足,选 hub-and-spoke 或者 data marts bus

Strategic View of Data Warehouse Before Implementation

The extent of implementing a data warehouse was viewed as an important strategic decision-making-supporting solution.
View DW as short-term solution, then choose independent data mart.
View DW as long-term strategic solution, then choose hun-and-spoke or data marts bus.

Compatibility With Existing Legacy System

Prefer to build a new data warehouse which is compatible with existing systems.
如果不在乎compatibility, choose data mart bus

Perceived Ability of IT Staff

Some architecture is harder to implement and some is easier.
如果人笨,选简单的(independent data mart, federated)
如果人聪明,选复杂的 hun-and-spoke or data marts bus.

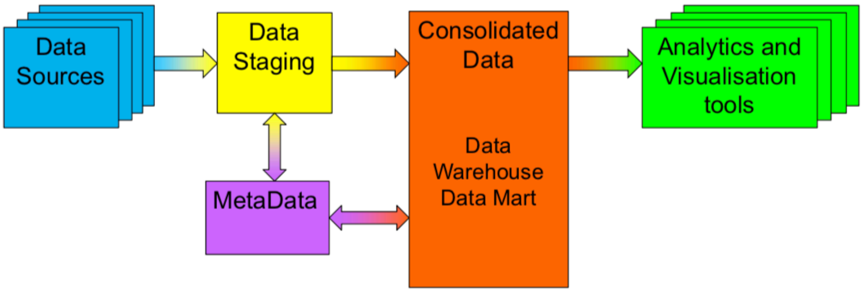
Technical Issue

- Ability to integrate metadata
- Scalability in number of users, volume of data, complexity of query
- Ability to maintain historical data
- Ability to adapt to changes

Social/Political Factor

- Expert influence: a consultant may recommend an architecture that he implemented successfully in the past.

6 Data Warehouse Components (Building Block)



1. Source Data

Production Data

- Data from operational systems (may from more than one operational system)
- Be careful of:
  - Meaning of data
  - Difference between similar data

Internal Data

- Inside-department data storage (different department has different data format)
- Make ETL process more complex

Archived Data

- 也就是historical data (operational system only contain current data, but data warehouse needs historical data)
- Be careful of:
  - Different archival methods and legacy systems. (may in disk or tape)

External Data

- Data from external environment
  - Industry information
  - News
  - Market trend
  - Change of laws and policy
- Helpful for executive level
- Usually is not conformed to the format of internal data (need convert)
- Need to organize transmission of data to the enterprise

2. Data Staging

- Now we already have data, then need to store it, fix it, and put it into the data warehouse

Date Staging Area

Need a space to temporarily store data while play and prepare data

ETL Process

- Extraction
  - Get data from sources into staging area
  - Can store extracted data in a DBMS or flat file
- Transformation
  - Transform data into correct and conformed format of the planned data warehouse
  - Fix data quality issue
  - Be careful of date type, characters set…
- Loading
  - initial loading is the 1st time of loading data into the data warehouse
  - Subsequent loading load new information into the data warehouse

3. MetaData

- MetaDate is data of data
- It can:
  - Define field of DW
  - Comment on field of DW
  - Identify the source of data
  - List update history of data

    - List responsible people
- It is used in:
    - Build DW
    - Using DW
    - Updating DW

3 Types of MetaData

- **Operational metadata**
    - The information about source operational database.
- **ETL 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.

Importance of MetaData

- It is the glue that connect all parts of a data warehouse
- Give developer information about the data content and data warehouse structure
- Give end-user information about the data content in business terms.

4. Data Storage

- The space where we store the data warehouse data
- Data warehouse is read-only to users
- Data warehouse is write-only to ETL process
- Usually data in data warehouse is stored in relational database/ multidimensional database/ NOSQL database. Can be combination of above)

5. Information Delivery

- Users may have different skill level and different complexity level of query
- Combine prefabricated reports, and build it your own query and report.
- Lots of methods to delivery information (list, diagram, vis applications…)

6. Management and Control

…

Technical Architecture (Cover 6 DW Components)

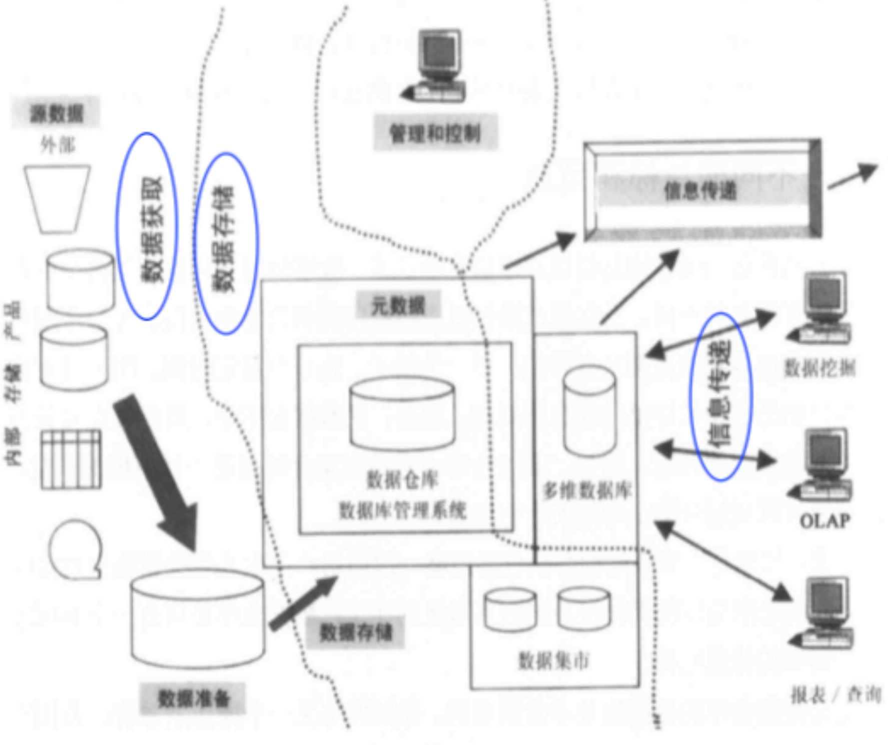


图 7-1　三个主要领域中的体系结构组成部分

1. Data Acquisition

- Data acquisition is very time-consuming, it takes 40%-60% of the total time to build a DW

How To Get Data Into DW

1. Extraction
2. Move to staging space
3. Transformation
4. Loading

Data Flow

1. Begin at extraction from data source
2. Pause at staging
3. Continue when loading data into DW

Data Source

- Need to be temporarily store at staging area for processing
- If it needs to be combine with other data, there may be mid-step to store it before staging
- Can be enterprise operational data
- Can from legacy system
- Can from enterprise resource plan system
- Can be outside data

Data Staging Area

- Where the extracted data is played
- Data is store at finest/lowest grain in staging area
- Data can be in files OR databases

Function of Data Extraction

- Select data source
- Generate auto extracted files from data source
- Create medium file to store the data that to be combined with outer data before staging area
- Reformat data form outside or departmental files
- Resolve data inconsistance

Function of Data Transformation

- Merge data
- Clean and de-duplicate
- Denormalize extracted data
- Convert data type
- Calculate some attribute as needed
- Aggregate data as needed
- Resolve missing data
- Integrate data

Function of Data Staging

- Backup and recovery the staging area
- Sort and merge files
- Create file as input to dimension table
- Create staging database
- Preserve audit trail for data item in DW
- Manage PK and FK of loading table
- Extract loading file from staging area

2. Data Storage

- Care about how to **load** data from staging area into DW
- Data coming form staging area is “finished product”

Data Flow

- Data move from staging area to data mart (then data warehouse) OR data warehouse (then data mart), depend on architecture used.
- **2 type of loading data flow**
  - Incremental updates
  - Full load

Function of Data Storage

- Load data for full refresh of DW tables
- Periodic incremental update data in DW
- Loading into multiple tables at detailed and summarized levels
- Optimize the loading process

- Provide auto services for loading into DW
- Backup and recovery for the DW
- Provide security
- Monitor and fine-tune the database
- Periodically archive data from DW

3. Information Delivery

- Care about the way of providing information to users
- Flexibility is important (some requirement is simple and some is complex)

Data Flow

- Start at DW or data marts
- End at users
- **4 types of information delivery data flow:**
  - Partial DW data
  - Full DW data
  - Analysis running on DW
  - Analysis running on different machine (eg. data mining run on other machine)

Function of Information Delivery

- Control information access for security
- Monitor user access to improve service
- Allow user to browse DW content
- Auto reformat query for better execution
- generate self-service report for users
- Store query and its result for future use
- Provide multiple level of data granularity
- Enable data feed to other analytical application, like data mining

(Management and Control)

- Control data movement to staging area
- Control ETL process
- Control loading data into DW
- Control back up some part of DW
- Monitor growth of data
- Monitor response time
- Do recovery when things fail
- Can be an auto tool OR run by DW administrator

(Overall Characteristics of DW Architectural Area)

Different Objectives and Scope

- Architecture supports information acquisition
- Scoping is hard

Data Content

- Data is read-only
- Data volume is large

Complex Query & Quick Response

- Analysis session can be long
- DW architecture must support variation in providing analysis
- Data can be near red time and analysis may change as data changes

Flexible

- All requirements might not be pre-known
- May have additional requirements
- May be able to adapt to changing business environment

Metadata-Driven

- Must be able to track data through metadata
- Equivalent metadata is not stored in operational system

Date Warehousing for…

Supply Chain

Simulate and optimize supply chain flows (reduce inventory and stock-outs)

Customer Selection

Identify customers who can provide greatest profit

Loyalty and Service

Retain customer loyalty and increase the likelihood that people want to buy the product or service.

Pricing

Identify the price which will maximize profit

Human Capital

Find the employees with best performance for a particular task.

Product and Service Quality

Detect quality problem, solve or minimize them.

Financial Performance

Better understanding the driver of financial performance, and understanding the effect of non-financial factors.

Research and Development

Improve quality and efficacy, improve safety of product and services (if applicable)