Data Classification and Management in Very Large Data Warehouse

고려대학교 전산학과
DB & Fuzzy System 연구실
유재학

2009-04-01 DB & Fuzzy System LAB. 유재학
CONTENTS

1. Introduction
2. Basic architecture of a data warehouse
3. Problem specification and possible approaches
4. Historical data classification scheme
5. Data storage and access mechanism for historical data
6. Conclusions
1. Introduction

- A data warehouse mainly deals with historical data and specialized retrieval
- Multi-dimensional Grid file: storing and supporting efficient management of historical data
- The structure of operational data is often complex → normalized, it is not always meaningfully presented to the end user
- The data warehouse integrates: meaningful, accurate, intelligible for analysis.
2. Basic architecture of a data warehouse

- The data after the adequate ETT (Extraction, Transportation, Transformation processes) is stage in the data warehouse.

- High-level architecture is divided into two parts:
  - back room
  - front room

**Figure 1. Basic architecture of a data warehouse**

Operational system → Operational data store → Data warehouse → Query interface
3. Problem specification and possible approaches

- The data stored in a data warehouse in historical in nature
- Historical data can include sales by product, region and time
- Depending upon the age, data may become “stale” (unusable)

<table>
<thead>
<tr>
<th>Sales year</th>
<th>Item ID</th>
<th>Sales Region</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td>11111</td>
<td>California</td>
<td>$1000</td>
</tr>
<tr>
<td>1982</td>
<td>22222</td>
<td>Arizona</td>
<td>$1234</td>
</tr>
<tr>
<td>1987</td>
<td>15671</td>
<td>Washington</td>
<td>$2500</td>
</tr>
<tr>
<td>1990</td>
<td>51675</td>
<td>Missouri</td>
<td>$5555</td>
</tr>
<tr>
<td>1991</td>
<td>33333</td>
<td>Ohio</td>
<td>$9999</td>
</tr>
</tbody>
</table>

- Problem of example
  : the data ‘Sales Year’ is how to efficiently deal with historical data and present a workable scheme
4. Historical data classification scheme

- Data categorization of warehouse data is necessary to manage them efficiently

- Highly performance and ease of management dependent

![Figure 2. Historical data general classification](image)

**Figure 2. Historical data general classification**
4.1 Time-wise classification of historical data

1) Real-Time data: very detailed, accessed in read/write mode, Consider as the creation-time of the data, \( Current \ time \ < \ t_i \ < \ 2 \) years.

2) Post-Decade data: this data ranges from two to ten years, read mode only, \( (2+t_i) \ < \ t_i \leq 10 \) years.

3) Pre-Decade or Time-worn data: \( 10+t_i \ < \ t_i \leq 99 \) years.
- Table 2 summarizes these categories.

<table>
<thead>
<tr>
<th>Data category</th>
<th>Time expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-time data</td>
<td>Present &lt; $t_i$ &lt; 2 years</td>
</tr>
<tr>
<td>Post-decade data</td>
<td>$(2+ t_i) &lt; t_i$ &lt;= 10 years</td>
</tr>
<tr>
<td>Pre-decade</td>
<td>$10 + t_i &lt; t_i$ &lt;= 99 years</td>
</tr>
<tr>
<td>Century data</td>
<td>$10 + t_i &lt; t_i$ &lt;= n years</td>
</tr>
<tr>
<td>Millennium data</td>
<td>Ancient or behind the age</td>
</tr>
</tbody>
</table>

**Table 2. Historical Data categories**

- Processing frequency of real-time data is the highest

![Figure 3. Variation of processing frequency with creation time for historical data](image-url)
4.2 **Content-wise classified historical data**

- Contrary to the Time-Wise data, the Content-wise classification focuses only on the content of the data
- Creation-Time is not considered
- example) World War I, atomic bombing

4.3 **Content & time-wise classified historical data**

- In this both creation-time and data-content
- example) a database log-file created in the year 2000.
5. Data storage and access mechanism for historical data

- Data warehouse environment deals with multi-dimensional data.

- Conventional data access mechanisms
  : hashing, sequentially allocated files, tree-structure

- Grid file is a file structures that support multi-dimensional data.

**Figure 4. Grid file structure**
5.1 **Metrics defined in the algorithm**

- Range metric: defined as the total year span, denoted as $R$.
- Range/Granularity Threshold metric: minimum threshold value to which the time splits are allowed, denoted as $T$.
- Frequency metric: access rate $H$, data with access frequency less than the lower bound $L$ are stale data.
- Active Region metric: denotes a sub-set of high access frequency region.
5.2 Grid splitting algorithm for historical data management

1. Compute the grid region, linear cell and bucket for the record
   If Record fits in bucket
     Insert record
   Else
     If Case 1
       Split buckets and do not change the linear scale, Insert the record into the bucket, Reassign the pointers, together
       If (R/2 < T)
         Terminate algorithm (Record is not inserted)
       Else
         If (Access rate < L) Check average accesses for past $t_i$ years
         If average close to L then Go To Stale data
         Else If average is close to H then
           Center split the linear scale, split Bucket, Insert record, and Reassign pointers to buckets
If (Access rate between L and H)
  If average for \( t_i \) years < L
Then Go to Stale data
  If average Access rate between L and H for \( t_i \) years then
  Center split the linear scale, split buckets, Insert record,
  Reassign pointers to buckets
End of algorithm
Check for next record

Stale data:
Do not split the linear scale and the bucket.
 Archive this data with a suitable archive policy
 Creates empty buckets and space merging case: Merge this linear cell with
 its adjacent cell so as to optimize the storage space of the grid directory
 structure and the buckets
 Realign the bucket pointers to the new grid structure
6. Conclusions

- Data warehouse always grow large-scale

- **Advantage**
  : Multi-dimensional grid file is used to identify and archive stable data and to optimize storage
  : Data architecture improves the usefulness of any warehouse

- **Disadvantage**
  : Preprocessing (must be some kind of garbage collection)
  : Limited (historical data classification)
  : Real-time data → Clustering or Classification