Survey Report on Real-time Data Warehouses

Ziyu Lin

Peking University

Oct 19, 2006
Outline

- Project introduction
- Real-Time Data Warehousing: Challenges and Solutions
- Our Research Work
- Reference
Project Background

Information evolution in data warehousing

STAGE 1
Reporting
What happened?

STAGE 2
Analysis
Why did it happen?

STAGE 3
Prediction
What will happen?

STAGE 4
Operationalize
What is happening?

STAGE 5
Activate
Active Warehousing
What do I want to happen?

- Primarily batch with predefined queries
- Increase in ad hoc queries
- Analytical modeling grows
- Continuous updates and time-sensitive queries gain importance
- Event-based triggering takes hold

Batch
Ad Hoc
Analytics
Continuous update/short queries
Event-based triggering
Project Objective

- Scalable, Real-Time and Active MPP based Data Warehouse For Telecommunications Industry

- advance the Real-Time Active Data Warehouse (RTADW) technology on a scalable, parallel database system platform

- demonstrate its applicability in tackling emerging business system challenges in the telecommunication industry.
Outline

- Project introduction
- Real-Time Data Warehousing: Challenges and Solutions
- Our Research Work
- Reference
Real-Time Data Warehousing: Challenges and Solutions

1. – enabling real-time ETL
2. – modeling real-time fact tables
3. – OLAP queries vs. changing data
4. – scalability & query contention

Department of Computer Science and Technology, Peking University, Oct 19, 2006
Challenge 1: Enabling Real-time ETL

ETL in batch mode

- almost all ETL tools and systems, whether based on off-the-shelf products or custom-coded, operate in a batch mode.

- ETL process typically involves downtime of the data warehouse

Real-time ETL

- there can't be any system downtime

- The requirements for continuous updates with no warehouse downtime are generally inconsistent with traditional ETL tools and systems.

Problem Statement
Challenge 1: Enabling Real-time ETL

1. Near real-time ETL
2. Direct trickle feed
3. Trickle & Flip
4. External Real-time Data Cache

Real-time ETL
Challenge 2: Modeling Real-time Fact Tables

- where the real-time data is stored
- how best to link it into the rest of the data model

Problem Statement
Challenge 2: Modeling Real-time Fact Tables

1. Modeling as Usual with Direct Fact Table Feed
2. Separate Real-time Partition
3. Integrated Real-time through Views
4. Modeling with an External Real-time Data Cache

Modeling Real-time Fact Tables
– OLAP and Query tools were designed to operate on top of unchanging, static historical data

– the results may be negatively influenced by data changes concurrent to query execution

– relational OLAP tools are particularly sensitive to this problem
Challenge 3: OLAP Queries vs. Changing Data

1. Use a Near Real-time Approach
2. Risk Mitigation for True Real-time
3. Use an External Real-time Data Cache

Department of Computer Science and Technology, Peking University, Oct 19, 2006
Challenge 4: Scalability & Query Contention

Problem Statement

1. The issue of query contention and scalability is the most difficult issue facing organizations deploying real-time data warehouse solutions.

1. In a real-time system, the additional burden of continuously loading and updating data further strains system resources.

1. The contention between complex selects and continuous inserts tends to severely limit scalability.
Challenge 4: Scalability & Query Contention

1. Simplify and Limit Real-time Reporting
2. Apply More Database Horsepower
3. Separate & Isolate in a Real-time Data Cache
4. Just-in-time Information Merge
5. Reverse Just-in-time Data Merge

Scalability & Query Contention
Outline

- Project introduction
- Real-Time Data Warehousing: Challenges and Solutions
- Our Research Work
- Reference
• **OLAP and Query tools:**
  - designed to operate on top of unchanging, static historical data
  - assume that the underlying data is not changing
  - the results they produce may be negatively influenced by data changes concurrent to query execution

In some cases, this can lead to inconsistent and confusing query results, which is called *internal inconsistency of report.*
Table 1: A multi-pass SQL statement

```sql
create table TEMP1(
    Category_Id LONG, DOLLARSALES DOUBLE)
insert into TEMP1
select all.[Category_Id] AS Category_Id,
    sum (all.[Tot_Dollar_Sales]) AS DOLLARSALES
from [YR_CATEGORY_SLS] all
    group by all.[Category_Id]
create table TEMP2 (ALLPRODUCTSD DOUBLE)
insert into TEMP2
select sum((all.[Tot_Dollar_Sales]) AS ALLPRODUCTSD
from [YR_CATEGORY_SLS] all
select distinct pa1.[Category_Id] AS Category_Id,
    all.[Category_Desc] AS Category_Desc,
    all.[DOLLARSALES] AS DOLLARSALES,
    (pa1.[DOLLARSALES]/pa2.[ALLPRODUCTSD]) AS DOLLARSALESC
from    [TEMP1] pa1,
            [TEMP2] pa2,
            [LU_CATEGORY] all
    where pa1.[Category_Id]=all.[Category_Id]
drop table TEMP1
drop table TEMP2
```
The description of layer-based view approach

Figure *Layer technology used in painting software*
The description of layer-based view approach

When to update already read-locked data, Layer 2 is generated above Layer 1.

Layer technology used in this paper
The architecture of real-time OLAP

Figure The architecture of real-time OLAP
The architecture of real-time OLAP

• **Data modeling**
  - no special data modeling is required
  - generally modeled identically to the data warehouse
  - typically contains only the tables that are real-time.

• **Data integrating**
  - batch loading
  - change data capture

• **Data merging**
  - JIM (Just-in-time information merging)
  - RJIM (Reverse Just-in-time information merging)
Outline

- Project introduction
- Real-Time Data Warehousing: Challenges and Solutions
- Our Research Work
- Reference
Reference


