Survey Report on Real-time Data Warehouses

Ziyu Lin

Peking University

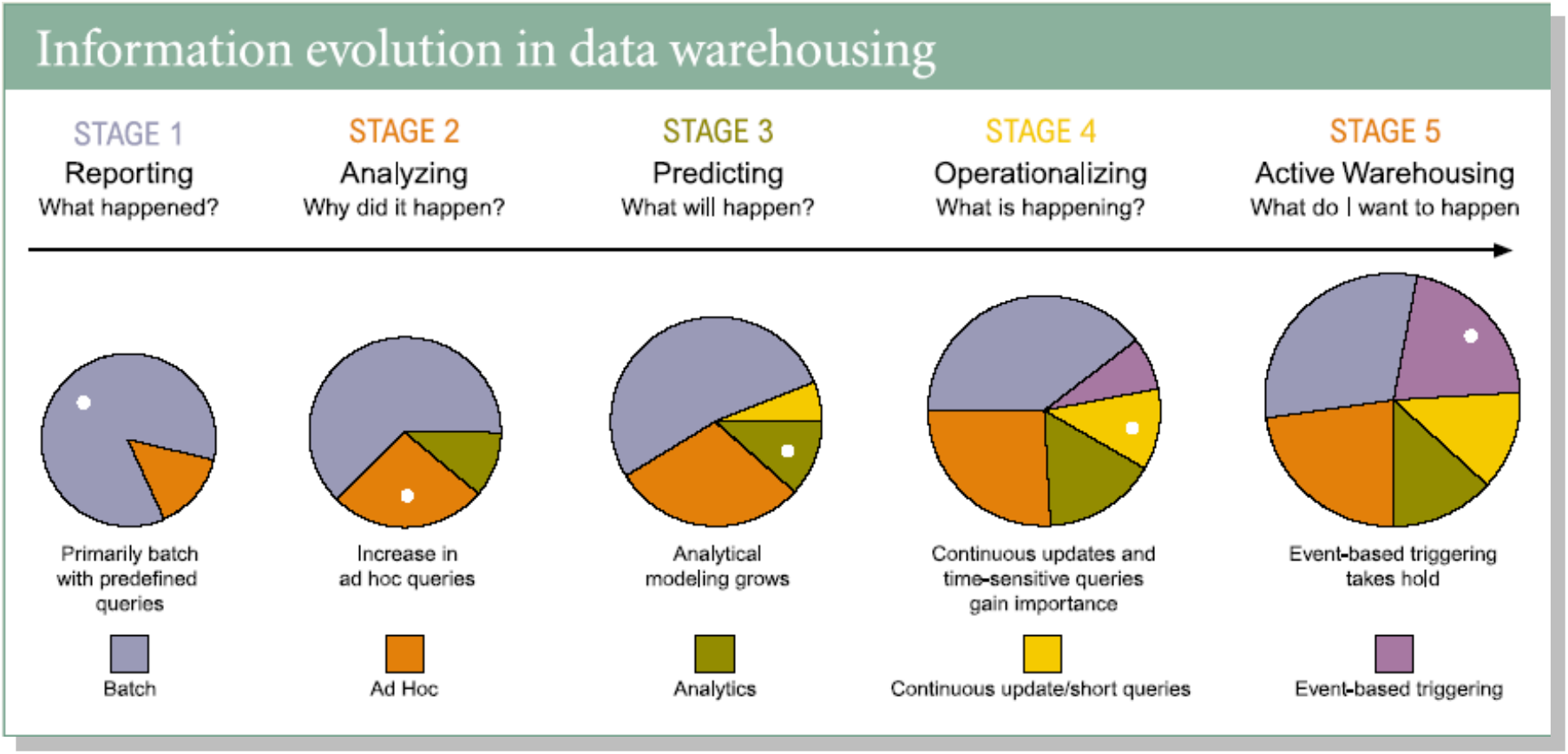
Oct 19, 2006

Outline

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

Project Background

1 Reporting → 2 Analysis → 3 Prediction → 4 Operationalize → 5 Activate



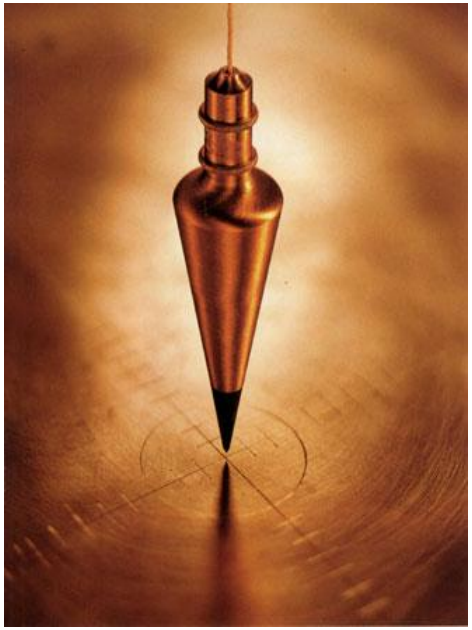
Project Objective

- **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

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

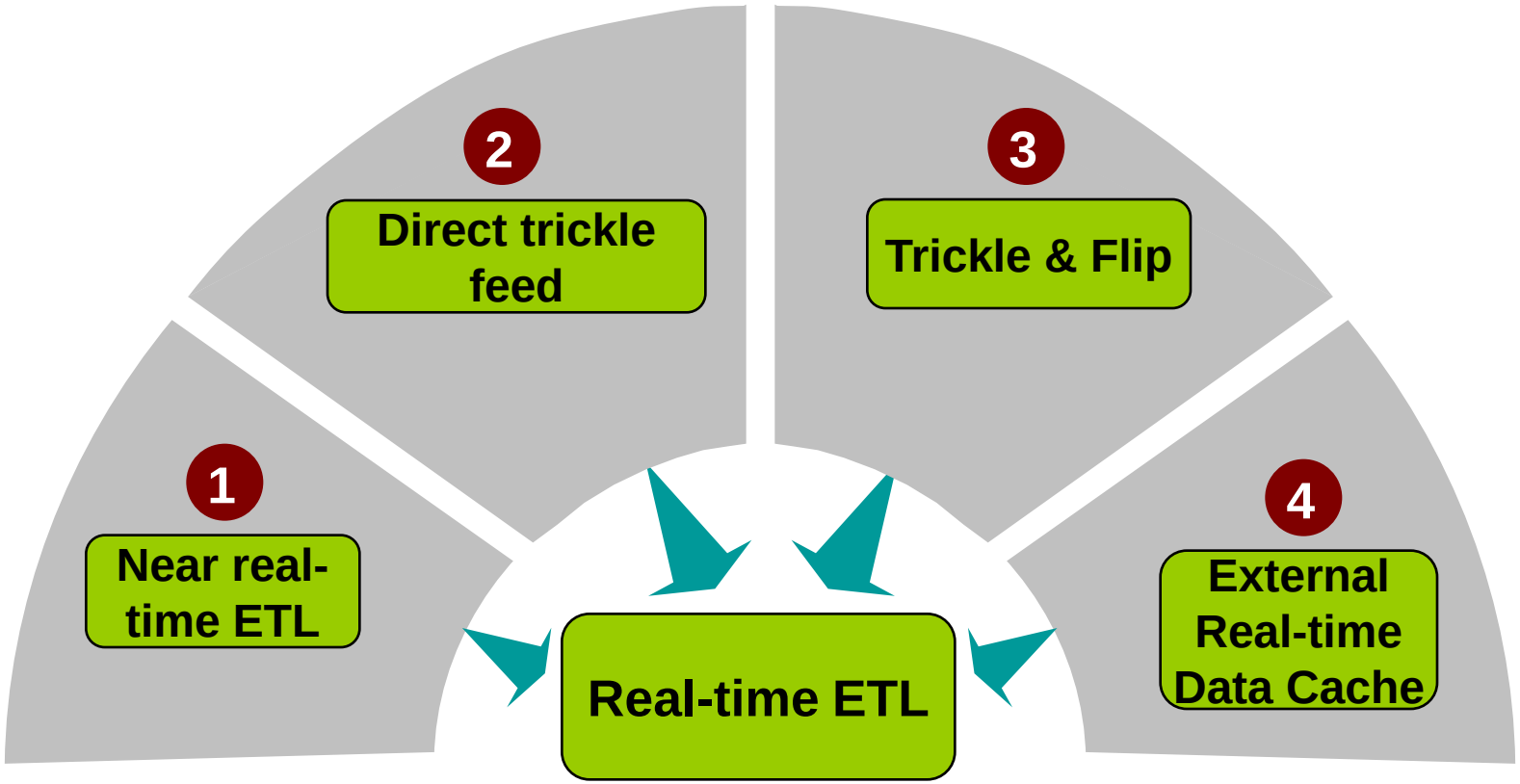
Problem Statement

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.

Challenge 1：Enabling Real-time ETL



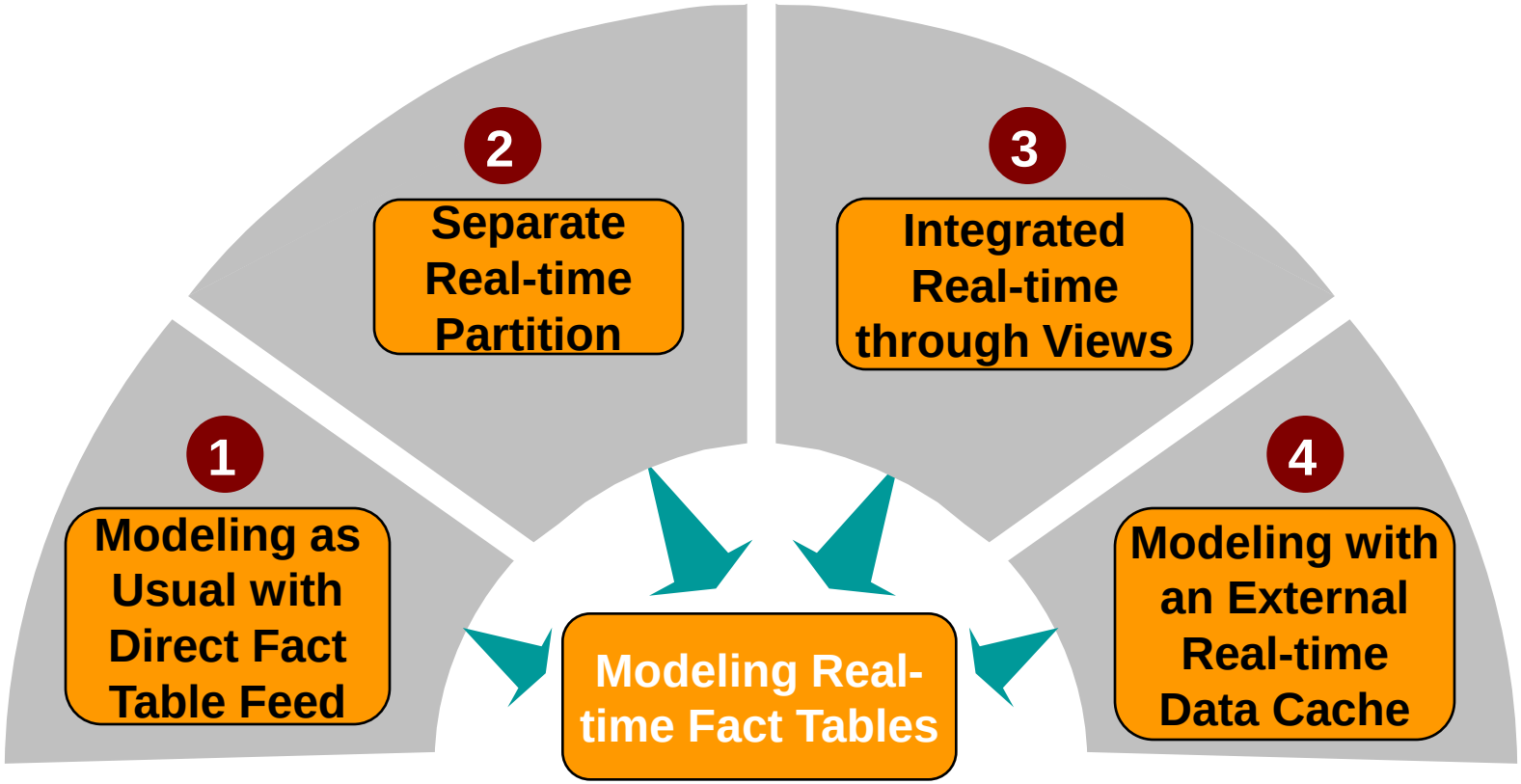
Challenge 2：Modeling Real-time Fact Tables

– 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



Challenge 3:OLAP Queries vs. Changing Data

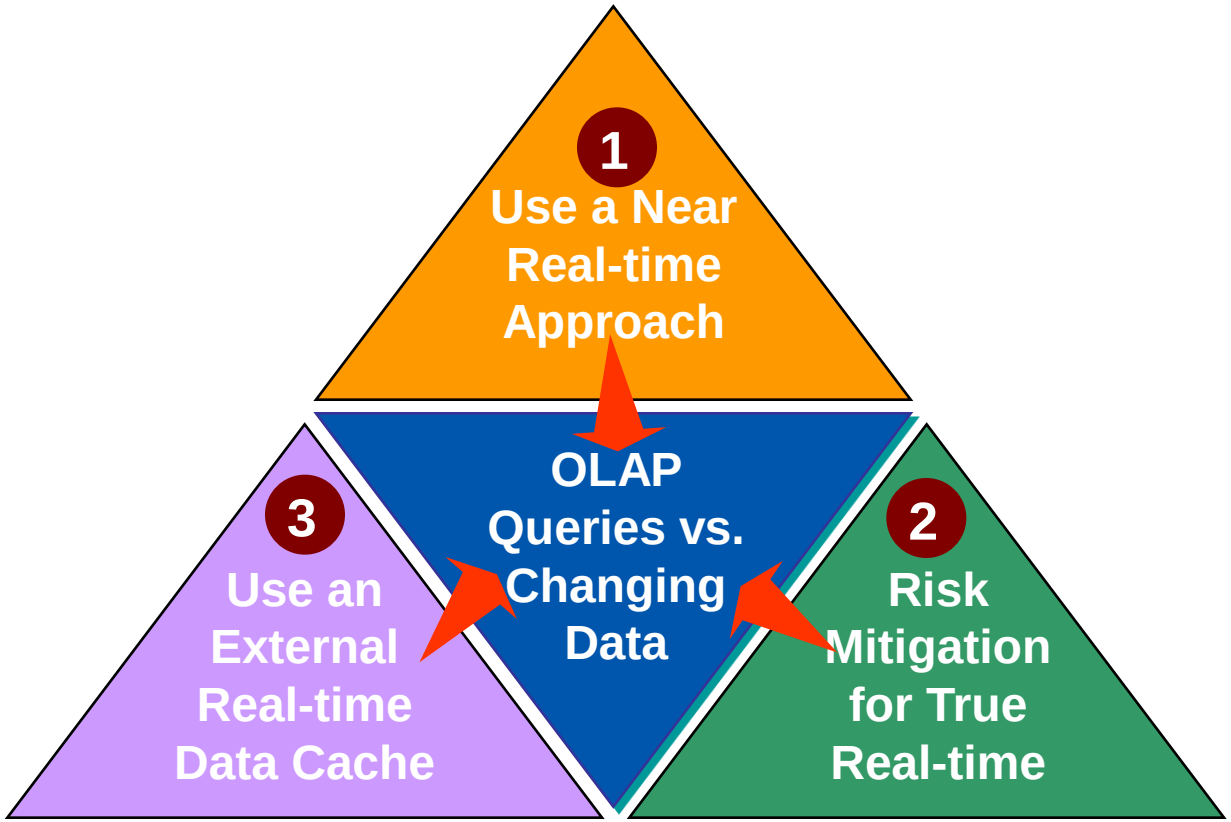
– OLAP and Query tools were designed to operate on top of unchanging, static historical data

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

– relational OLAP tools are particularly sensitive to this problem

Problem Statement

Challenge 3:OLAP Queries vs. Changing Data

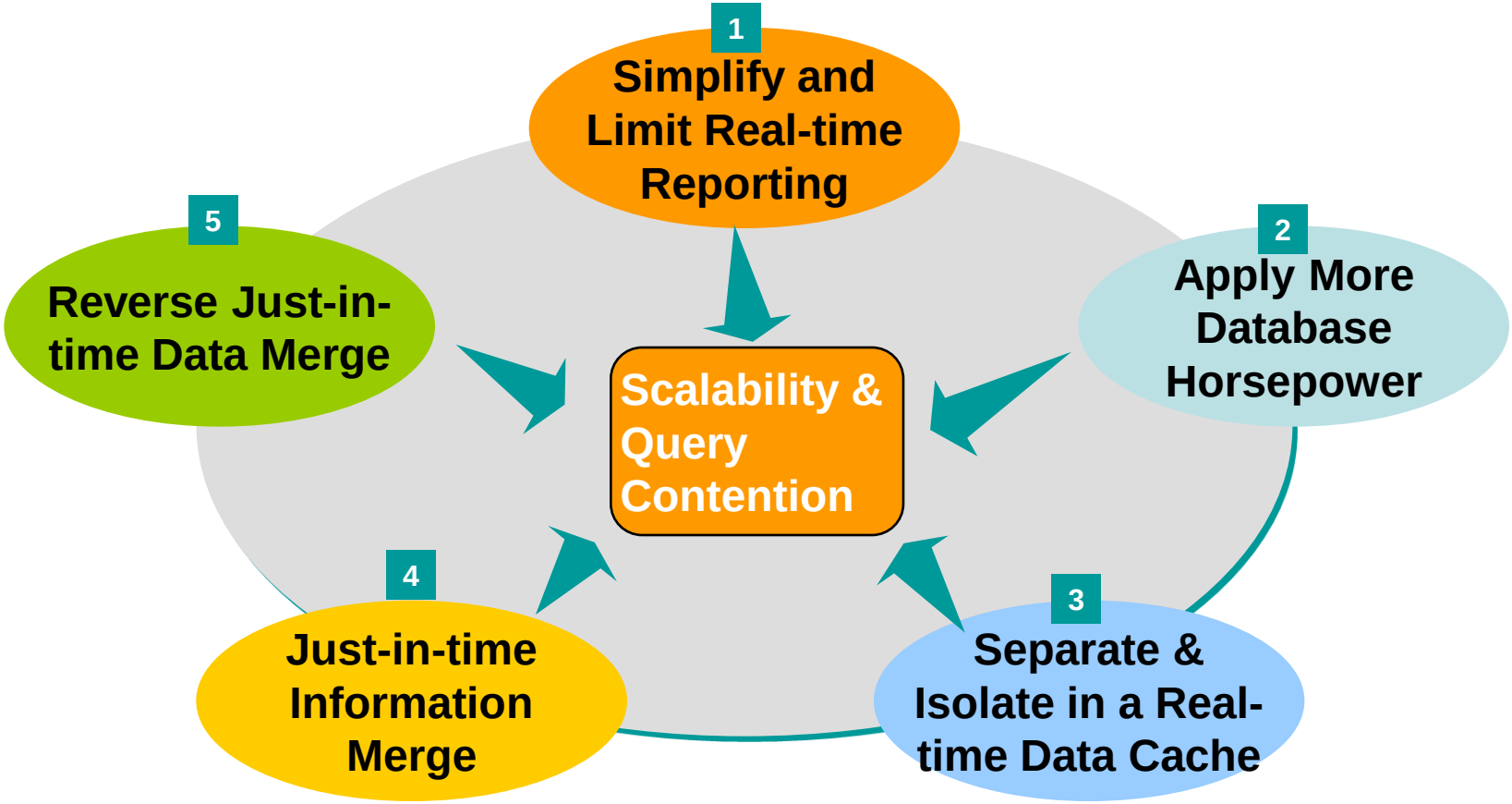


Challenge 4: Scalability & Query Contention

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



Outline

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

Introduction

- *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*.

Introduction

```
0:00    create table TEMP1(
        Category_Id LONG, DOLLARSALES DOUBLE)
0:01    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]
0:05    create table TEMP2 (ALLPRODUCTSD DOUBLE)
0:06    insert into TEMP2
        select sum((all.[Tot_Dollar_Sales]) AS
        ALLPRODUCTSD
        from [YR_CATEGORY_SLS] all
```

```
0:08    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]
0:09    drop table TEMP1
0:10    drop table TEMP2
```

Table 1: A multi-pass SQL statement

The description of layer-based view approach

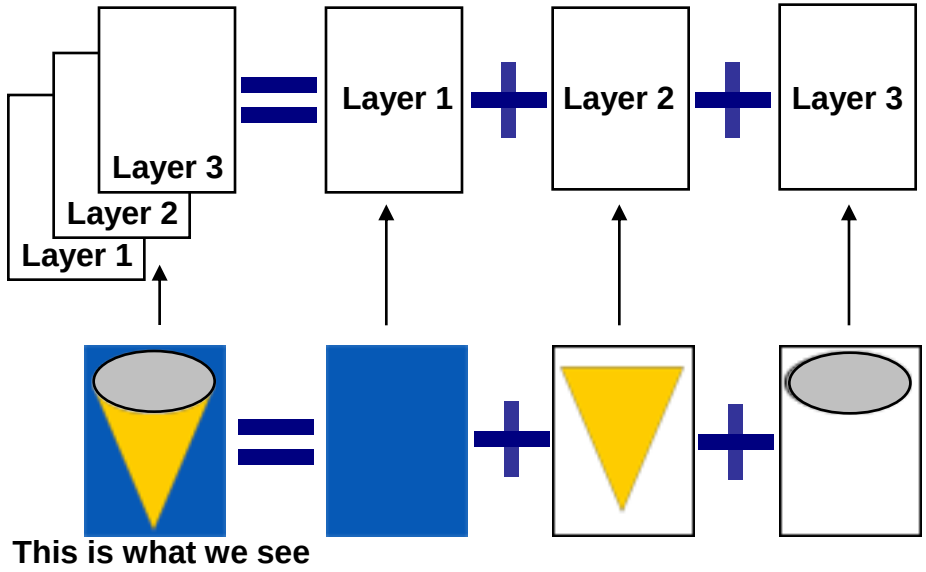


Figure *Layer technology used in painting software*

The description of layer-based view approach

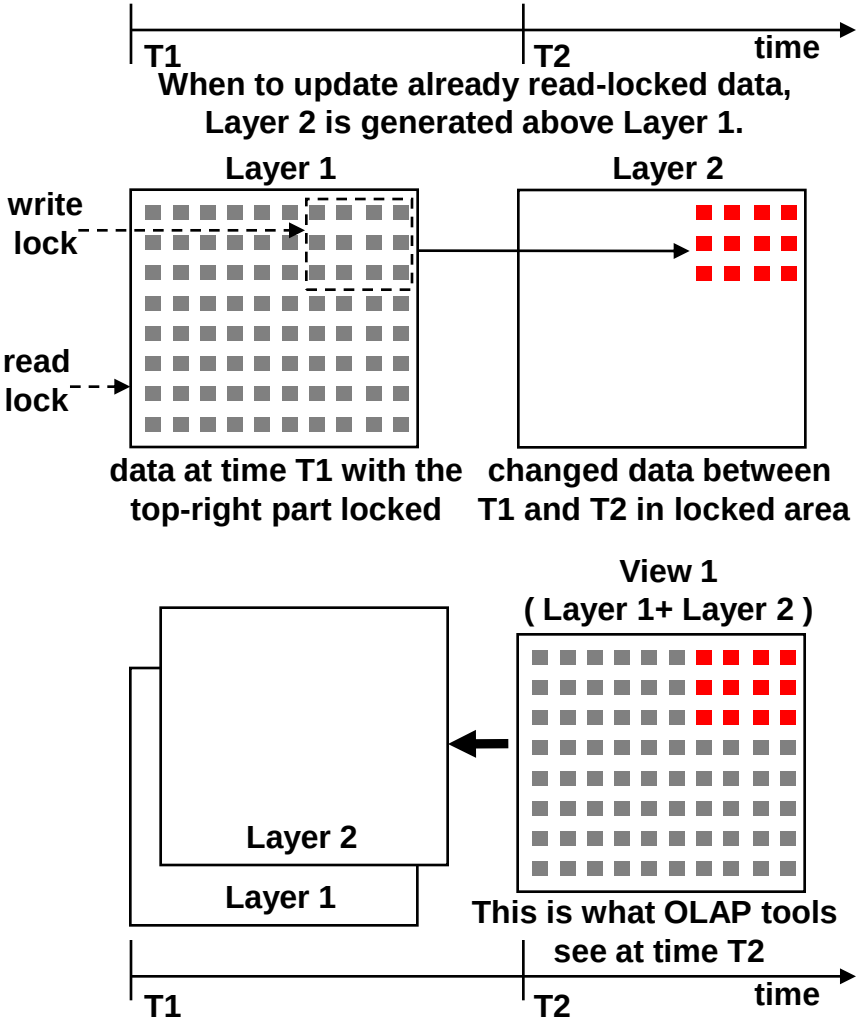


Figure *Layer technology used in this paper*

The architecture of real-time OLAP

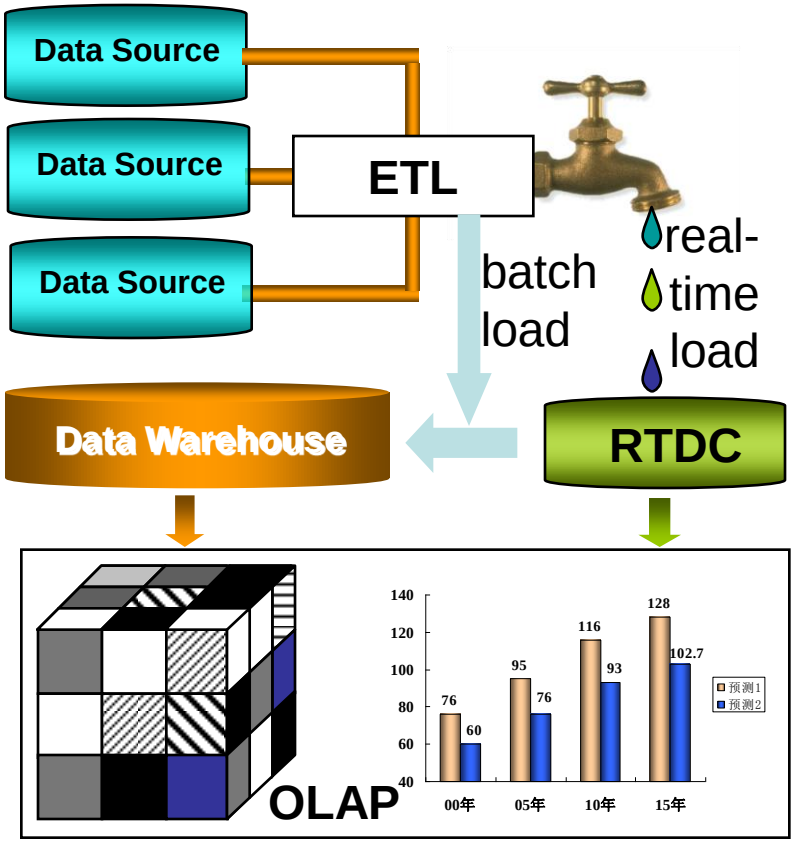


Figure *The architecture of real-time OLAP*

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

[1] Langseth, J., "Real-Time Data Warehousing: Challenges and Solutions", DSSResources.COM, 02/08/2004.
[2] Hongfei Guo, Per-Ake Larson, Raghu Ramakrishnan, Jonathan Goldstein. Relaxed currency and consistency: how to say "good enough" in SQL. In SIGMOD 2004.P:815-826.
[3] Robert M. Bruckner and A.M. Tjoa. Managing Time Consistency for Active Data Warehouse Environments. DaWaK 2001, LNCS 2114, pp. 254–263, 2001.
[4] Robert M. Bruckner and A.M. Tjoa. Capturing Delays and Valid Times in Data Warehouses—Towards Timely Consistent Analyses. Journal of Intelligent Information Systems, 19:2, 169–190, 2002
[5] Zaniolo, C., Ceri, S., Faloutsos, C., Snodgrass, R.T., Subrahmanian,V.S., and Zicari, R. Advanced DatabaseSystems. San Francisco: Morgan Kaufmann Publishers. 1997.
[6] Gregersen, H. and Jensen, C.S. Temporal Entity-Relationship Models—A Survey. IEEE Transactions on Knowledge and Data Engineering, 11(3), 464–497, 1999.

Reference

[7] Widom, J. (1995). Research Problems in Data Warehousing. In Proc. of the 5th Intl. Conference on Information and Knowledge Management (CIKM), Baltimore, Maryland (pp. 25–30). ACM-Press.
[8] Yang, J.. Temporal DataWarehousing. Ph.D. Thesis, Department of Computer Science, Stanford University. 2001.
[9] Yang, J. and Widom, J. .Maintaining Temporal Views over Nontemporal Information Sources for Data Warehousing. In Proc. of the 6th Intl. Conf. On Extending Database Technology (EDBT), Valencia, Spain, Springer LNCS Vol. 1377 (pp. 389–403). 1998.
[10] Yang, J. and Widom, J.. Temporal View Self-Maintenance. In Proc. of the 7th Intl. Conf. On Extending Database Technology (EDBT), Konstanz, Germany, Springer LNCS, Vol. 1777 (pp. 395–412). 2000.
[11] Pedersen, T.B., Jensen, C.S., and Dyreson, C.E.. A Foundation for Capturing and Querying Complex Multidimensional Data. Information Systems, 26(5), 383–423.2001.

Reference

[12] Tho, M. Nguyen; Tjoa, A. Min. “Zero-Latency Data Warehousing for Heterogeneous Data Sources and Continuous Data Streams”;Proceedings of iiWAS 2003, Fifth International Conference on Information and Web-based Applications Services, Jakarta, Indonesia; Austrian Computer Society (OCG) (2003), 3-902134-72-0; 55–64.
[13] S. S. Conn. OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis. In: Southeast Con, 2005. Proceedings. IEEE., pages 515-520, 2005.
[14] Itamar Ankorion. Change Data Capture – Efficient ETL for Real-Time BI. Article published in DM Review Magazine, January 2005 Issue.
[15] T. Thalhammer and M. Schrefl. Realizing active data warehouses with off-the-shelf database technology. software-Practice & Experience, ACM, 32(12), pages 1193-1222, 2002.

Department of Computer Science and Technology, Peking University, Oct , 2006