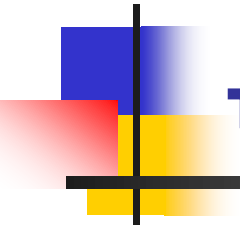
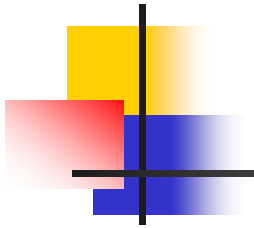


The Reality of Real-Time Business Intelligence

Divy Agrawal
Computer Science
UC Santa Barbara



The beginning

50 Years of Business Intelligence

- Vision of Business Intelligence:
  - Hans Peter Luhn in a 1958 article.
  - Predates the notions of Databases and Data Management.

- A pioneer in Information Sciences:
  - New use of the term ***thesaurus***
  - Automatic creation of literature abstracts
  - 16 digit Luhn's number widely used for credit cards and other banking instruments
  - …

Luhn's Vision

- Defined BI as:

  "… provides means for selective dissemination to each of its action points in accordance with their current requirements or desires."

- Key technologies:
  - Auto-abstracting of documents,
  - Auto-encoding of documents, and
  - Auto creation and updating of profiles

- Breadth of the vision:

  "… *business* is a collection of activities carried on … be it **science, technology, commerce, industry, law, government, defense**, et cetra."

  "… *intelligence* is also defined … as the ability to **apprehend the interrelationships** of presented **facts** in such a way as **to guide action** towards a desired goal."

The intervening years

The Early Years (1970s-1980s)

- Contrary to Luhn's overarching vision – early efforts on business information remained focused on ***database management technology***.

- With the advent of the relational model:
  - DBMS technology became pervasive and matured.
  - Widely adapted by most enterprises.
  - Online Transaction Processing became a proven paradigm for business operations.

- Consequence:
  - Massive proliferation of OLTP systems especially within a single enterprise.
  - Data-driven decision making became a norm.
  - Disparate reporting from multiple operational data sources.

Notion of "Data Warehousing" (1990s)

- Presence of multiple operational systems created a ***fractured*** view of an enterprise.
- Devlin & Murphy introduced the term ***business data warehouse*** in 1988:
  - A unified view of the enterprise primarily for integrated reporting.
- Catalysts:
  - Demand for reporting – key factors being PCs and spread-sheets.
  - Market potential – Teradata, Red-brick Systems, etc.
- Negative factors:
  - Unproven, immature, and expensive technology proposition.
  - Distinction between DBMS and DW: no clarity, ***?duplication?***
  - Fairly laborious and time-consuming data integration process
  - No clear stake-holders ➔ ***2nd Class Entity*** often resulting in adversarial atmosphere.

Data Warehousing: Current State

- Keys to success:
  - Enormous contribution of DW evangelist Ralph Kimball
  - STAR schema & Dimensional model for DW: intuitive and scalable
  - No compromise on the autonomy of operational data sources

- Persisting head-winds:
  - Since does not directly contribute to P&L:
    - ROI question still persists.
  - Not a plug & play technology:
    - Very high consulting costs.
  - Legacy of significant time and cost over-runs of most data warehousing projects.
  - Batch-oriented DW Architecture:
    - Deemed too costly just for integrated reporting.
    - Needed intuitive analytical capabilities.

Hither "Business Intelligence" (2000-)

- Gray et al. [1996] introduced the CUBE operator for roll-up and drill-down analysis of multi-dimensional data (i.e., DW Model).

- DW enterprises (Hyperion, Cognos, Analysis Services, etc.) adapted the CUBE architecture and called it:
  - ***business intelligence.***

- Problem:
  - Early BI (CUBE) technology had serious issues of scaling ➔ only accentuated the ill-repute of DW/BI technologies
  - Underlying problem: exponential explosion of data storage

Business Intelligence: Current State

- While the BI/Cube technology was still evolving – the spin doctors needed to undo the early damage.
- Hence, perhaps the term **Real-time Business Intelligence –** to convey the "criticality" of such technology to business leaders.
- Current debate: what exactly is meant by "real-time" in Business Intelligence?
  - In 2006, in this workshop, Donovan Schneider – gave numerous examples of "degree of timeliness" for a variety of analysis tasks.
  - My personal view is that the correct term should have been: ***Online Business Intelligence***.
- Assuming that – redefine the DW/BI architecture to support RTBI.

The present & the future

Real-time Business Intelligence: Required?

- Anecdotal evidence from Sam Walton

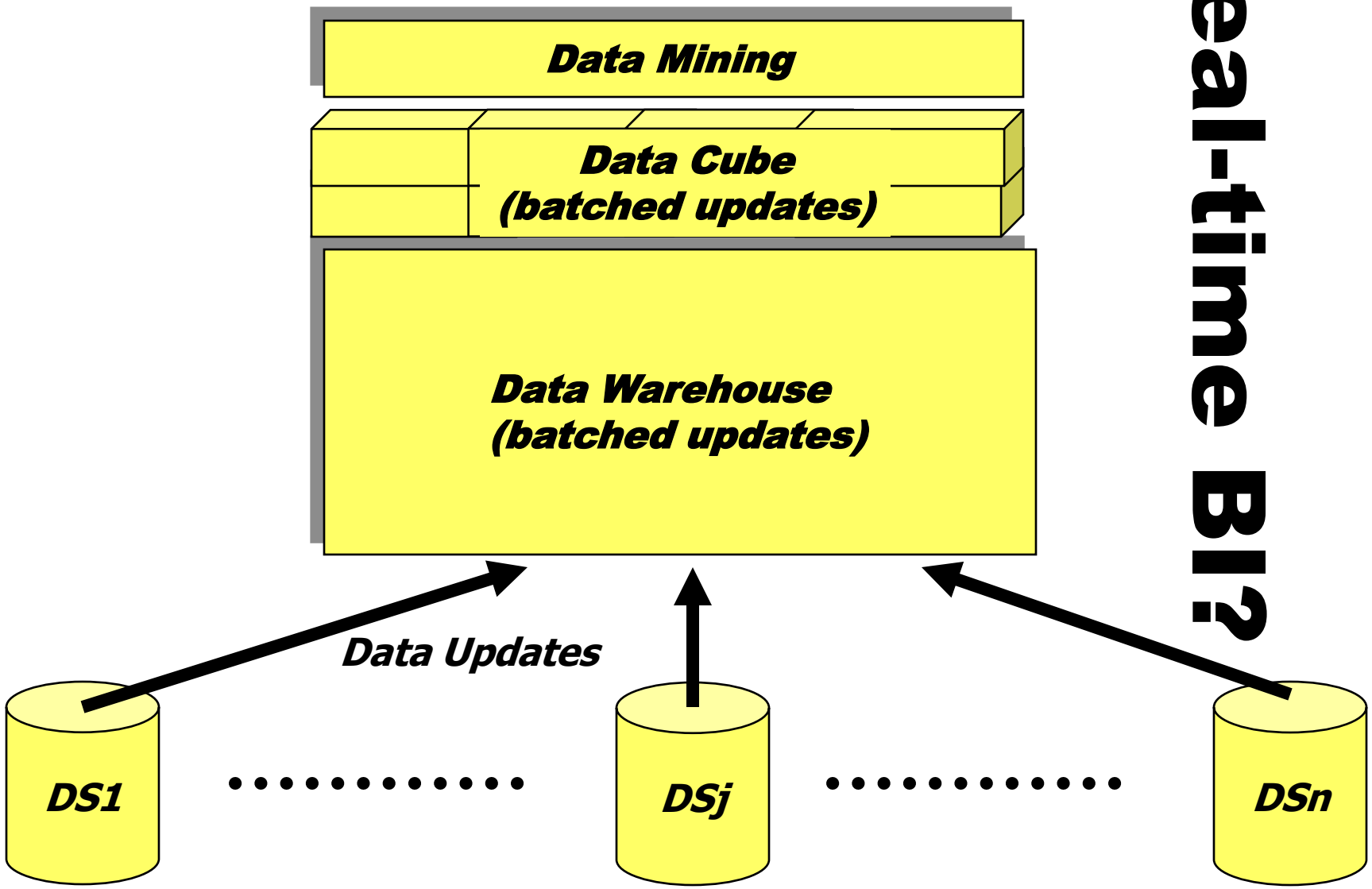
Airplane & Parking Lot Story

• Demonstrates the power of 10,000 feet view (from the airplane) versus the local view (from the parking lot).

• Numerous cases where "timeliness" of "intelligence" is extremely valuable.

➔ The case of RTBI is very-well justified.
➔ The question however is at what cost?

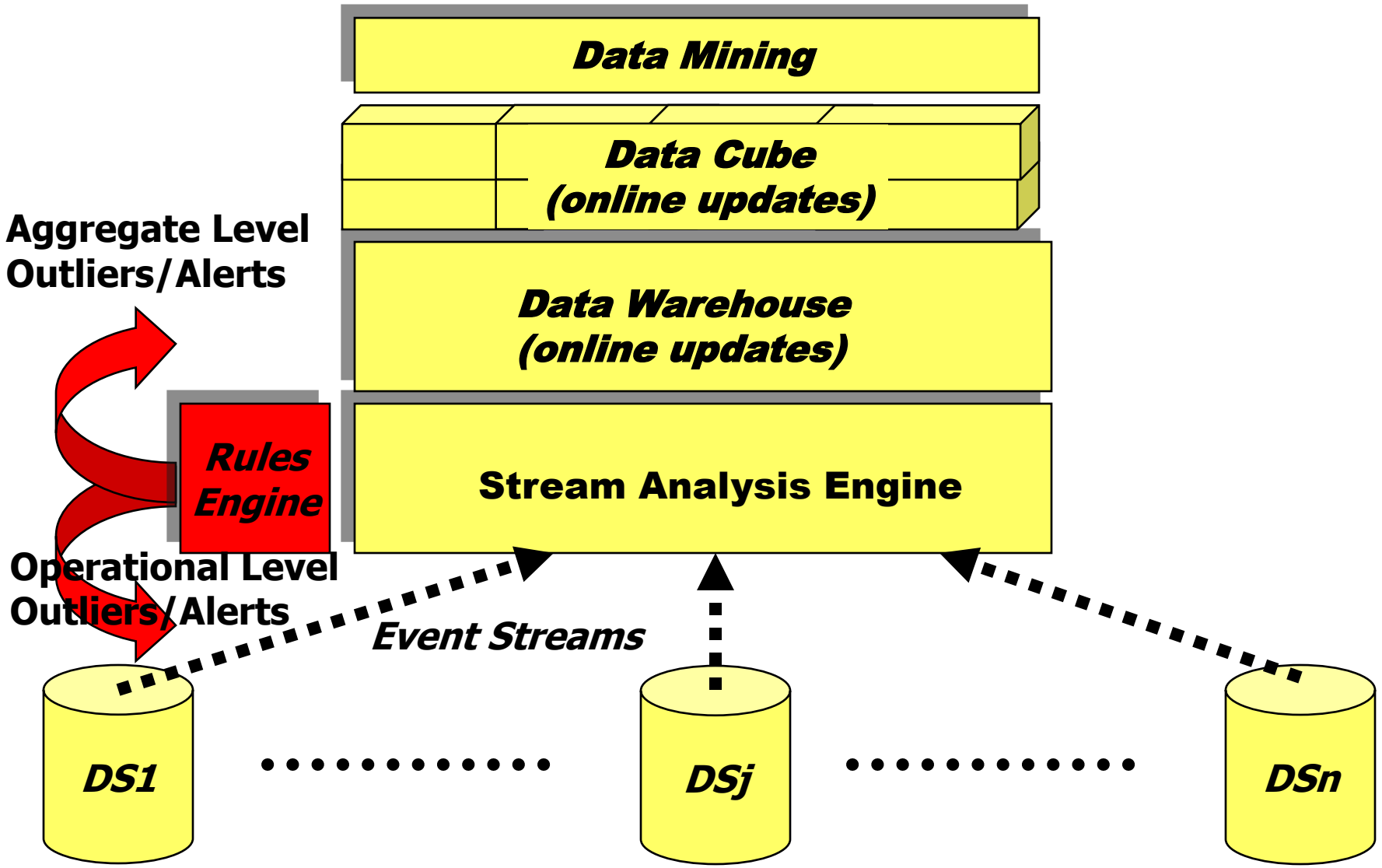
BI/DW Architecture (Revisited)



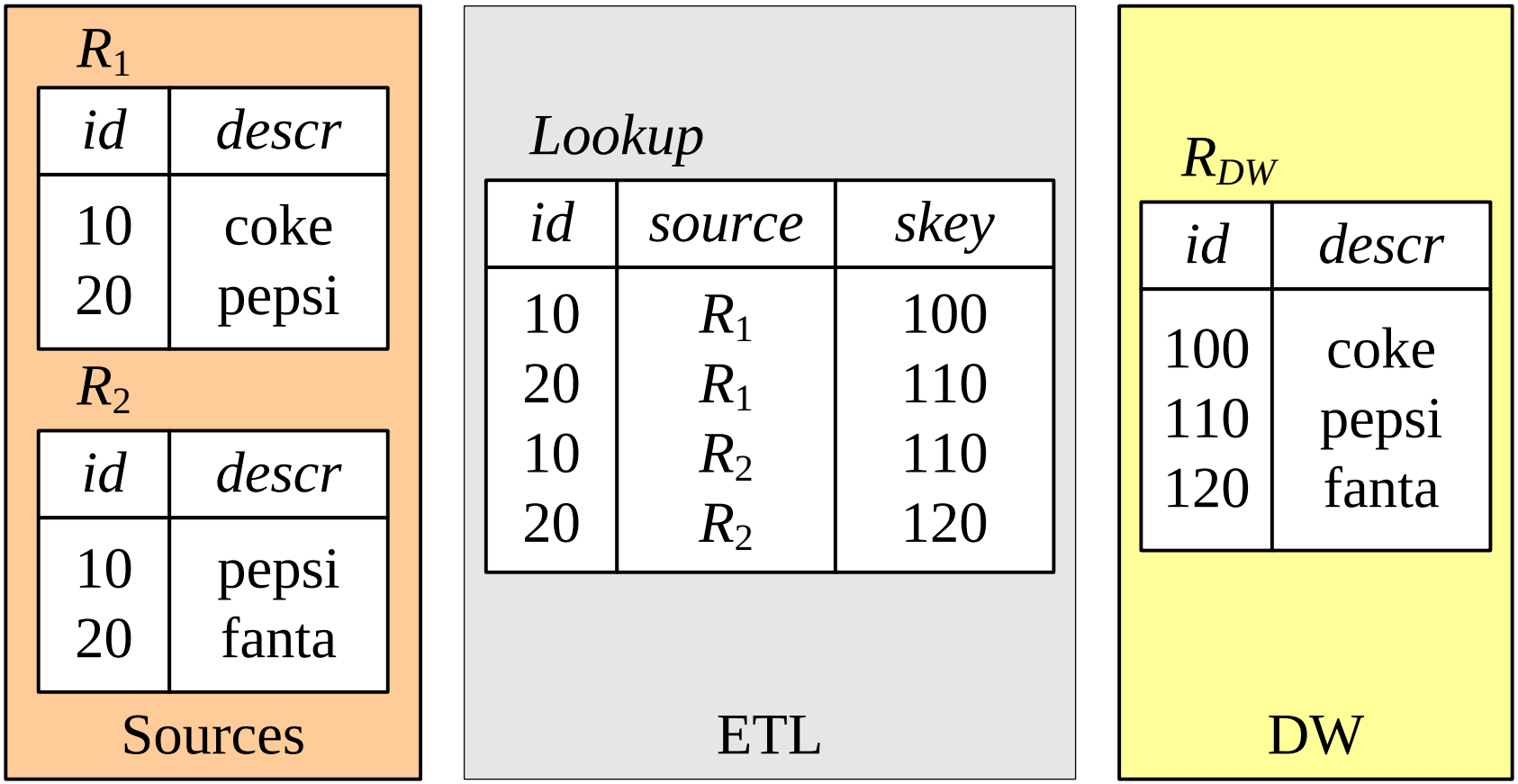
Underlying Technology Components

System	Technology Components
Database Management Systems	■ Relational Model ■ Declarative Language ■ Data Independence
Data Warehouse	■ Dimensional Model ■ Design Methodology ■ ETL Tools
Business Intelligence	■ Data Cube Model
Real-time Business Intelligence	

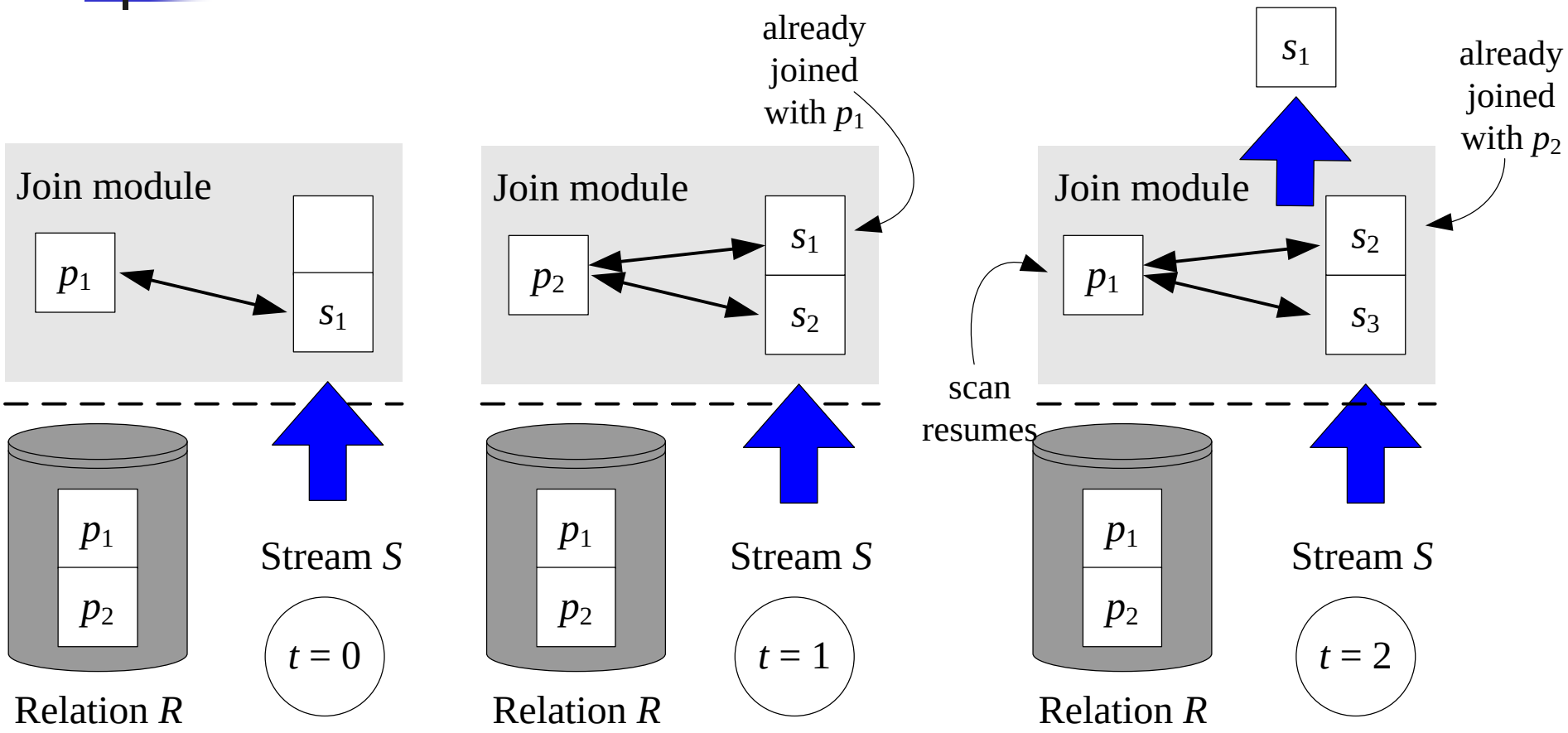
Real-time in BI/DW Architecture?



Real-Time ETL: Surrogate Key, Duplicate Elimination (R&D efforts)



Mesh-Join [Polyzotis et al.]

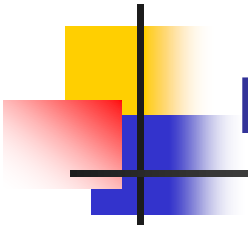


Vassiliadis & Simitsis: Near Real-time ETL (forthcoming)

Real-time Scheduling of Updates: on-going work

Enabling Real-time BI: Source Updates

- Online updates:
  - Move from periodic refresh to continuous updates
  - Example: The window of opportunity for up-sell/cross-sell of a product is while the customer is still around NOT AFTER he/she has left.
  - Tighter coupling between the operational data sources to the data warehouse:
    - In the past, operations team viewed the DW/BI as a necessary evil
    - In the current business landscape, DW/BI should be viewed as a means to survival



Enabling Real-time BI: Data Streams

- Stream Analysis & Management:
  - Event monitoring before updates incorporated in the warehouse.
  - Stream operators:
    - Heavy-hitters (frequency counting)
    - Fraud detection
    - Performance monitoring
    - Histograms and quantile summaries
    - …
  - Outlier detection for operational intelligence
  - Summary/Aggregate analysis for strategic decision-making.

Enabling Real-time BI: Data Integration

- Automated Data Integration:
  - Current approaches of integrating data from operational data sources into the data warehouse too tedious and time consuming.
  - Although this task is greatly simplified with the plethora of ETL tools that are available in the marketplace (e.g., Informatica)
  - New research for automated schema integration (e.g., Pay-as-you-go Data Spaces model)
  - Problem: uncertainty of data integration.

➔ A monumental challenge especially since the enterprises of today are highly dynamic and are constantly evolving.

Enabling Real-time BI: Analysis Language

- Declarative approach for analytical processing:
  - Current approach of analytical processing is ad-hoc and error-prone.
  - Translating business questions into analysis queries is highly manual.
  - Newer approaches are emerging:
    - MapReduce from Google significantly simplifies Web Log analysis.
    - Yahoo's PigLatin project
    - Microsoft's DRYAD project
    - Need similar efforts for other types of analysis and mining tasks (MDX?).

Enabling Real-time BI: Scaling with Large Data Volumes

- Scalability:
  - Certain queries (Temporal and Spatial Correlations) are bound to access huge amounts of data.
  - Need to rely on hardware solutions to provide scalability.
  - Emerging solutions (Parallel DBMS Technology):
    - GreenPlum
    - HP's NeoView
    - Google's GoogleFS and BigTable
    - Yahoo endorsed Hadoop
    - Cloud Computing?

RTBI: Technology Components

System	Technology Components
Database Management Systems	■ Relational Model ■ Declarative Language ■ Data Independence
Data Warehouse	■ Dimensional Model ■ Design Methodology ■ ETL Tools
Business Intelligence	■ Data Cube Model
Real-time Business Intelligence	■ Online updates ■ Stream Operators & Events ■ Next-gen MDX ■ Parallel Query Processing

Concluding Remarks

- Real-time BI (equivalently Online BI) has the immense potential for:
  - Data-driven operational decision making.
  - Data-driven feedback towards business strategy.
- Current adaptation of Real-time BI is hampered because of:
  - Lack of clarity about the underlying technology components
  - Significant costs associated with custom solutions
- Our task:
  - To clearly define the overall architecture of the next-generation Real-time BI Systems
  - Design and develop the necessary technology components.
  - Realize economies-of-scale to bring the cost factors down for a wide-scale adaptation

Hans Peter Luhn