Practical Considerations for Real-Time Business Intelligence

Donovan Schneider

YAHOO! Strategic Data Solutions

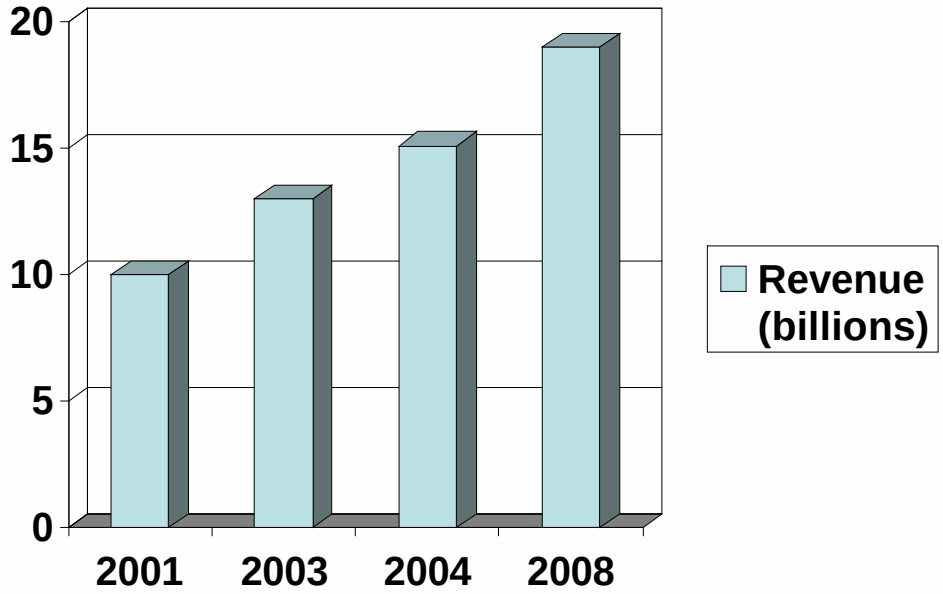
September 11, 2006

Outline

- Business Intelligence (BI) Background
- Real-Time Business Intelligence Examples
- Two Requirements of Real-Time BI
- Architecture Alternatives
- Conclusions and Open Research Challenges

Business Intelligence Market

"business analytics software market comprises tools and applications for tracking, storing, analyzing, modeling, and presenting data in support of automating decision- making and reporting processes", IDC



IDC
Analyze the Future

Progression of BI

- Transactional reporting
  - "give me my reports"
- Data warehousing/OLAP
  - "explore data to find interesting patterns/details"
- Business Performance Management (BPM)
  - "how am I tracking to business goals?"
- Guided Analytics/Business Activity Monitoring
  - "where should I look next?"
- Tactical decisions
  - "what do I do right now?"

Selected Business Intelligence Vendors

FairIsaac

ACTUATE

SPSS

streambase
WHEN NOW MEANS RIGHT NOW.

sas

COGNOS

Hyperion
The future in sight

Business Objects

IBM

SAP

Microsoft

Teradata
a division of NCR

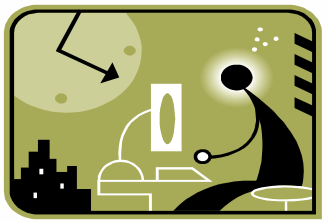
MicroStrategy
Best In Business Intelligence

ORACLE

Examples of Real-Time Business Intelligence

Real-Time Enterprise BI Applications

- Recommendations ★★★★½
  - Collaborative filtering, e.g., people who like X also like Y
  - Timeliness (freshness of data): hours
- Fraud Detection
  - Detect anomalies in credit card usage
  - Timeliness: minutes
- Call Center
  - Provide next best offer or action (cross-sell, up-sell)
  - Timeliness: minutes
- Close of books
  - Track deals at quarter close to grant/refuse contract concessions
  - Timeliness: minutes
- Defect/Incident Tracking
  - Track open/closed incidents
  - Timeliness: minutes

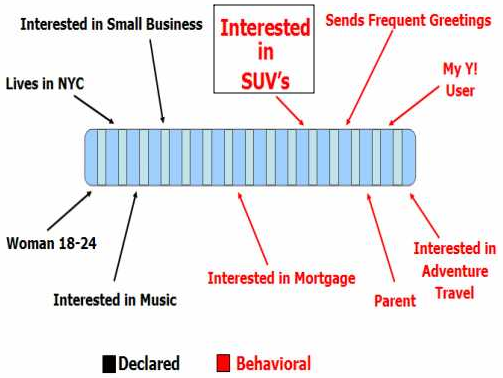


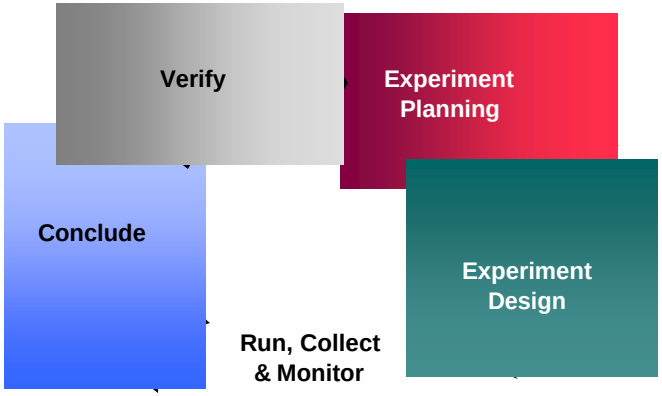
Real-Time Web Analytics Examples

- Web Page Usage
  - Analyze web page usage (page views, ad views, link views, clicks) by property, geography, user demographics, referrer, etc.
  - Timeliness: hours/next day
- Ad Campaign Effectiveness
  - Bid for search terms on Yahoo!, Google, MSN. Analyze click-thru and conversion rates
  - Timeliness: minutes

Real-Time Web Analytics Examples

- Targeting
  - Display an ad or content based on demographic profile, geographic location or behavior
  - Timeliness: minutes
- Experimentation
  - Run A/B or multivariate test on page content/layout. Analyze user engagement, click-thru rate, etc.
  - Timeliness: hours
- Search Term Analytics
  - Find most popular search terms by geography, gender, age range, etc.
  - Timeliness: hours





Business Activity Monitoring and Operational Performance Management

- Measure and monitor real-time business events within the enterprise to improve business performance
  - More than real-time alerts
  - Integrate, aggregate, correlate to improve business processes
- Examples
  - Real-time inventory analysis
  - Timeliness: minutes

Requirements for Real-Time Business Intelligence

Requirement #1: Time is Money

- Nothing is free; it costs money to reduce data latency
  - Specialized hardware (clusters, large memories, high bandwidth networks, fault tolerant)
  - Specialized software
    - highly-available, fault-tolerant, high performance
  - Integrated systems
- The business decisions to be made with reduced latency must justify the investment

Requirement #2: Actionable Data

- Context
  - Must provide contextual information to aid decision making
  - Typically requires access to detailed data
  - Typically requires access to trending/historical data
  - All silo'd solutions will ultimately fail this
- Audience
  - Provide role-specific views of data (e.g., sales rep, sales manager, district manager, executive), task-specific views, etc.
- Data
  - Data must be "clean" (normalized, conformed)
- Time
  - Timely presentation of data to decision maker

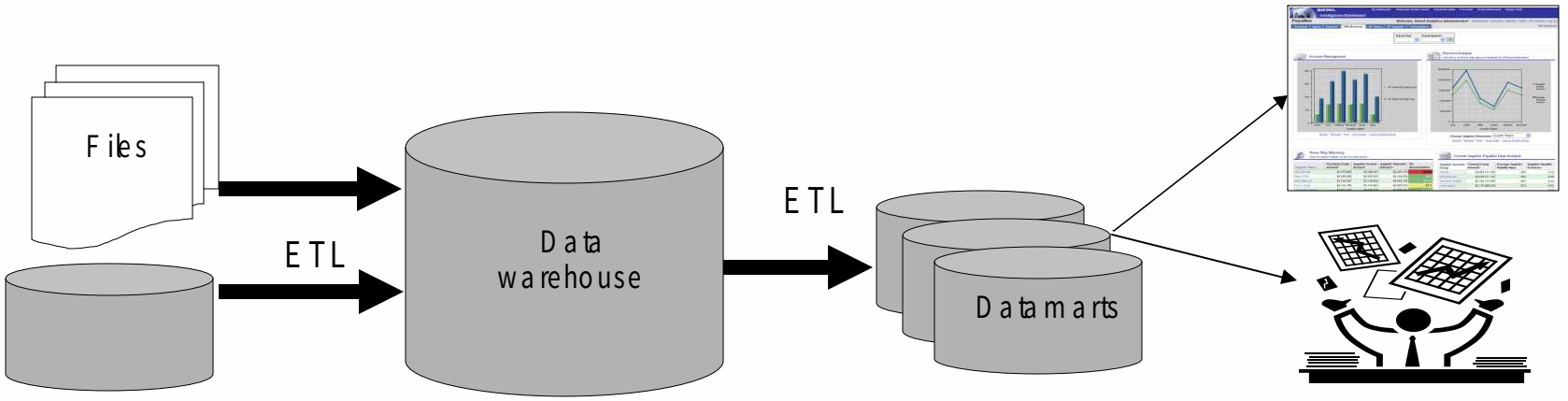
Challenges in real-time BI

- Data Scale
- Performance, Performance, Performance
  - Low latency data delivery
  - Consistent response times
  - Caching often used
- Cost
  - Performance/low-latency costs money
- High Availability
  - Servers, network, databases, middleware, applications
- Integration

Architecture Alternatives

Architecture Alternatives

- Custom Solutions
- Enterprise Data Warehouse (EDW)
- Federated system (virtual EDW)
- Streaming



Custom Solutions

- Build specialized systems to meet latency needs
- Pros
  - Optimize for specific needs
  - Initial development cost is low for data marts
  - Cheap enough that department VP can purchase
  - Can adapt quickly to meet changing business needs
- Cons
  - Lack of integration with contextual data
  - Lack of integration with detailed data
  - Multiple, competing sources of truth
  - Scalability (#of users, amount of data)
  - Lack of shared services
    - ETL processes, security, reporting, DBAs
  - Tend to proliferate across an enterprise; overall cost to company is high

Enterprise data warehouse (EDW)

- Consolidate data marts into a central data warehouse
- Pros
  - All non-production OLTP data sources in a single system
    - No multi-database joins
    - One system to administer and operate
    - Single source of truth
- Cons
  - Many EDWs fail for technical and organizational reasons
    - Departments lose control of their data
    - Departments lose agility
    - Difficult to deliver incrementally due to conforming dimensions
  - Difficult to support high volume, low latency ETL along with complex, ad-hoc decision support
  - Not real-time
  - Costs up to $50 million for a large organization
- Examples
  - Oracle/Teradata/DB2 EDW, SAP/BW

Virtual EDW

- Provide federated/virtual view(s) of enterprise data
- Pros
  - Each source system is optimized for a specific need or workload
    - Cubes, data warehousing, OLTP
  - Departments retain some control over their systems/data
  - Incremental build out (unlike an EDW), modulo conforming dimensions
- Cons
  - Requires conformed dimensions across systems
  - Some source systems restrict query access (e.g., OLTP systems will not allow large queries)
  - Security unification across disparate systems
  - Updates must be coordinated (between OLTP, DW, caches, etc.)
  - Sophistical SQL generation to optimally access data sources
  - Sophistical execution engine to compensate for data source limitations
  - High availability and problem diagnosis hampered by multiple systems
- Examples
  - Oracle Business Intelligence EE (formerly known as Siebel Analytics)
  - SAP/BW

Streaming, Business Activity Monitoring, Operational BI

- Make rapid decisions based on large volumes of data
- Pros
  - Optimized for low latency (few disk accesses)
- Cons
  - Data inconsistency (late, missing data)
  - Requires very high availability
  - Extend SQL for streaming operations
  - Integration with other sources can be a challenge
- Examples
  - Hedge funds processing ticker feeds for arbitrage
  - Fraud detection
  - Revenue alerting

Conclusions and Open Problems

Conclusions

- Some applications are demanding, and willing to pay for, very low latency access
- Most applications do not require latency in the seconds granularity
  - Delivery may be real-time (seconds), e.g., targeting, alerting, recommendations but underlying data can be less “fresh”
- Common evolution strategy is to increase frequency of ETL operations
  - Mini-batch ETL, e.g., load every 10 minutes
  - Requires fast, scalable, and high availability
    - load, clean, transform and aggregate

Research Challenges in real-time BI

- Data scalability
- User scalability
  - Decision making is reaching deeper in organizations
- Breadth of data access
  - Providing context for decision making requires accessing diverse set of data sources
- Query Language
- Cost
- Performance
  - Efficient, incremental algorithms
  - Very large dimensions (millions to hundreds of millions members)
- Ad-hoc vs. Production
  - How to support dynamic, mixed workloads
- High Availability/Fault Tolerance
  - Production decision making systems cannot fail
- Backup/Recovery