IS660J

Lecture 1

Professor K.M. Burns

Why Data Warehouses?

Data, simple facts, and figures, describe an enterprise and its competitive environment.

Data Warehouses can help organizations to safeguard their data assets and make data more accessible to decision makers.

True Value → Communication and collaboration among individuals

History

- Prehistoric - recording information on cave walls
- 3000 BC – Egyptians – formal means of written communication
- 105 AD – Chinese - Ts'ai Lun invented paper
- 400 AD – Chinese - Wei Tan invented durable ink
- 1442 – Germans – Johannes Guttenberg invented printing press
- 1876 – Americans – Thomas Edison invented telephone
- Post WWII –
  - 1945 – Hardware Era – ENIAC
  - 1975 – Software Era – Altair 8800
    - 1980s – PC – Productivity Tools
  - 1990s – Content Era - Internet

Content Era - Trends

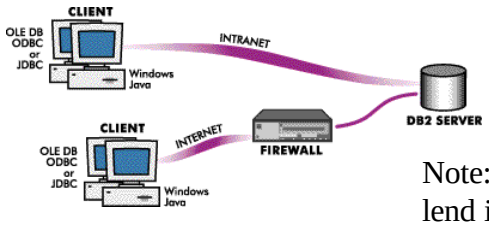
Information management and information dissemination

- Shift from OLTP to Data Warehousing, OLAP and Data Mining
- Shift from Client/Server system architectures to n-tiered architectures

Two-tier Architectures

Client:

Presentation and, some or all, Logic Processing.



Note: Does not lend itself well to the WWW

Client-centric: Example: US and European headquarters could install ODBC or OLE DB middleware on clients and use a query tool, or a custom Visual Basic application, to access the common financial database.

Three-tier Architectures

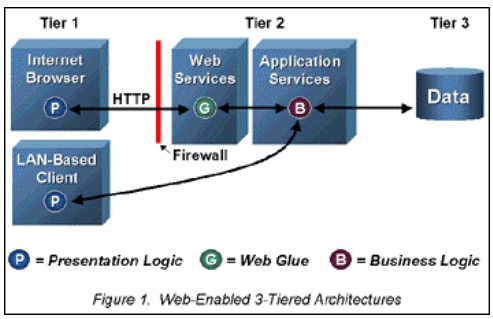


Figure 1. Web-Enabled 3-Tiered Architectures

As shown in Figure 1, three-tiered architectures enable the presentation, business logic, and data elements of applications to be cleanly separated and run on different machines connected by a network. What makes three-tiered architecture so important in Web environments is that applications can have browser-based user interfaces that access business logic and data components that reside safely behind corporate firewalls. Popular two-tiered approaches haven't mapped well to the Web, because they expose corporate data and application logic to tampering by the user.

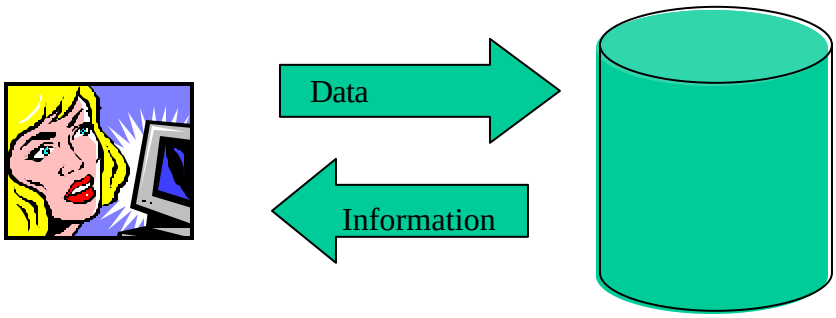
Distributed/Collaborative Enterprise Architecture (n-tiered architectures)

The distributed/collaborative enterprise architecture allows a business to analyze its internal processes in new ways that are defined by changing business opportunities instead of by preconceived systems design (such as monolithic data processing applications). In this architectural design, an object model represents all aspects of the business; what is known, what the business does, what are the constraints, and what are the interactions and the relationships. A business model is used to integrate and migrate parts of legacy systems to meet the new business profile. Distributed/collaborative enterprise builds its new business applications on top of distributed business models and distributed computing technology. Applications are built from standard interfaces with "plug and play" components. At the core of this infrastructure is an off-the-shelf, standards-based, distributed object computing, messaging communication component such as an Object Request Broker (ORB) that meets Common Object Request Broker Architecture (CORBA) standards.

Intellectual Capital

- Financial Capital → Balance Sheets
- Intellectual Capital → ???
  - Competitive mandate organizations that do not stay intellectually ahead of their competitors may pay the ultimate price – Failure.

Data versus Information



Structured Content

- Alphanumeric data stored in rows and columns of a database
- In contrast, most of the data types on the internet represent unstructured content (image, audio, text)

OLAP

Online Analytical Processing provides tools needed to access and analyze data. Capabilities:

- Query and reporting (simplifies SQL code generation)
- Multidimensional analysis
- Statistical analysis (data reduction)
- Data mining (a type of data exploration)

Initial questions are generally simple, but their complexity increases as the preceding questions are answered.

E.F. Codd – paper in 1993 - 12 rules that differentiate OLAP

Financial Data Warehouses

- Monitor business performance in financial terms → rapid response (spreadsheets)
- Snapshots of financial history (ie. revenues and expenses); Scorecard
- Relatively small number of facts – numeric performance measures
- Periodic updates – monthly, quarterly

Marketing Data Warehouses

- Evaluate business performance of a product or service from multiple perspectives
- Competitive intelligence – analyze the impact of each vendor's marketing activities on the sales of products or services
- Analysis can be very complex and queries are generally unpredictable
- Frequently updated

Behavioral Data Warehouses

- Database marketing or relationship modeling → attracting new clients and/or building loyalty with existing customers
- Detailed customer data/profiles → large in size
- Frequency of updates varies with application → design varies

Planning and Implementing a Data Warehouse

- Scope and Objectives
  - Executive Management – what are the critical success factors for the enterprise?
- User Information Requirements
  - Inclusive project teams
- Budget and timing considerations
- Project plan
  - Phased implementations with early deliverables

Planning and Implementing a Data Warehouse

- Applications Strategy (flows directly from CSFs and User Requirements)
- Database Strategy (tightly coupled with Applications strategy)
  - Scalability – size, number of users, complexity
- Deployment Strategy
  - Platforms; User Interfaces; Collaboration tools; Software Agents

Conclusion

A data warehouse has several processes that require several technology components. Batch and transaction processing data first has to be extracted from operational databases and then cleaned up to remove redundant data, fill in blank and missing fields and organized into consistent formats. The data is then loaded into a relational database. Business analysts can then dig into the data using data access and reporting software including On-Line Analytical Processing (OLAP) tools, statistical modeling tools, geographic information systems (GIS) and data mining tools.