DATA WAREHOUSING

TOPICS

- Definition
- Types
- Components
- Architecture
- Database Design
- OLAP
- Metadata repository

OLTP VS. WAREHOUSING

- Organized by transactions vs. Organized by particular subject
- More number of users vs. less
- Accesses few records vs. entire table
- Smaller database vs. Large database
- Normalised data structure vs. Unnormalized
- Continuous update vs. periodic update

DEFINITION

- A datawarehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of managements decision making process.
- It is the process whereby organizations extract value from their informational assets through use of special stores called data warehouses

TYPES

- Operational Data Store: Operational data mirror. Eg: Item in stock.
- Enterprise data warehouse: Historical analysis, Complex pattern analysis.
- Data Marts

USES OF A DATAWAREHOUSE

- Presentation of standard reports and graphs
- For dimensional analysis
- Data mining

ADVANTAGES

- Lowers cost of information access
- Improves customer responsiveness
- Identifies hidden business opportunities
- Strategic decision making

ROADMAP TO DATAWAREHOUSING

- Data extracted, transformed and cleaned
- Stored in a database - RDBMS, MDD
- Query and Reporting systems
- Executive Information System and Decision Support System

DATA EXTRACTION AND LOAD

- Find sources of data : Tables, files, documents, commercial databases, emails, Internet
- Bad data Quality: Same name but different things, Different Units
- Tool to clean data - Apertus
- Tool to convert codes, aggregate and calculate derived values - SAS
- Data Reengineering tools

METADATA

- Database that describes various aspects of data in the warehouse
- Administrative Metadata: Source database and contents, Transformations required, History of Migrated data
- End User Metadata:
  - Definition of warehouse data
  - Descriptions of it
  - Consolidation Hierarchy

STORAGE

- Relational databases
- MDD

  Measurements are numbers that quantify the business process Dimensions are attributes that describe measurements

INFORMATION ANALYSIS & DELIVERY

- Speed up retrieval using query optimizers and bitmap indices
- Adhoc query - Simple query and analysis functions
- Managed Query - Business layer between end users and database
- Multidimensional - OLAP - support complex analysis of dimensional data

INFORMATION ANALYSIS & DELIVERY

- EIS/DSS

  Packaged queries and reports

  Preplanned analytical functions Answer specific questions

- Alerts

  Specific indicators

MANAGING THE DATA WAREHOUSE

- Data - Size storage needs
  Security
  Backups
  Tracking
- Process- Monitoring update process like
  changes in source, quality of data
  Accurate and upto date

TOOLS

- Data Extraction - SAS
- Data Cleaning - Apertus, Trillium
- Data Storage - ORACLE, SYBASE
- Optimizers - Advanced Parallel Optimizer
  Bitmap Indices
  Star Index

TOOLS

- Development tools to create applications

  IBM Visualizer, ORACLE CDE

- Relational OLAP

  Informix Metacube

ARCHITECTURE

- Rehosting Mainframe Applications Moving to lower cost microprocessors Tools - Micro Focus COBOL Lowers Cost

No transparent Access to data

ARCHITECTURE

- Mainframe as server 2-tier approach Front end client & back end server Power Builder, VB - Front end tools Minimal investment in extra hardware Data inconsistency hidden Fat Client Cannot be used if number of end users increase

ARCHITECTURE

- Enterprise Information Architecture 3 tier

  Source data on

  host computer Database servers like

  ORACLE, Essbase(MDD)

  Front-end tools - DSS/EIS

RDBMS

- RDBMS provide rapid response to queries Bitmap index
  Index structures

- Functionality added to conventional RDBMS like data extraction and replication

MDD

- Decision support environment
- Supports iterative queries
- Extensions to SQL - for high performance data warehousing
- Performance degrades as size increases
- Inability to incrementally load
- Loading is slow
- No agreed upon model

MDD

- No standard access method like SQL
- Minor changes require complete reorganization

DATA ACCESS TOOLS

- Simple relational query tools - Esperent
- DSS/EIS - EXPRESS used by financial specialists

DATABASE DESIGN

- Simple
- Data must be clean
- Query processing must be fast
- Fast loading

STAR SCHEMA

- Consists of a group of tables that describe the dimensions of the business arranged logically around a huge central table that contains all the accumulated facts and figures of the business.
- The smaller, outer tables are points of the star, the larger table the center from which the points radiate.

STAR SCHEMA

- Fact Table -Sales, Orders, Budget, Shipment Real values (numeric)
- Dimension Table -Period, Market, Product Character data
- Summary/Aggregate data

STAR SCHEMA

- Data you can trust

  Referrential Integrity

- Query Speed

  Fact table - Primary key

  Dimension table - all columns Query optimizer which understands star schema

STAR SCHEMA

- Load Processing Must be done offline Issue if aggregate data is stored

VARIATIONS OF STAR SCHEMA

- Outboard tables
- Fact table families
- Multistar fact table

OLAP

- Front end tool for MDD
- Slice Report
- Pivot Report
- Alert-reporting
- Time-based
- Exception reporting

WIDE OLAP

- Generating (synthesizing) information as well as using it, and storing this additional information by updating the data source
- Modeling capabilities, including a calculation engine for deriving results and creating aggregations, consolidations and complex calculations
- Forecasting, trend analysis, optimization, statistical analysis

RELATIONAL OLAP

- Has a powerful SQL-generator
- Generates SQL optimized for the target database
- Rapidly changing dimensions

MDD OLAP

- Row level calculations
- Financial functions, currency conversions, interest calculations

METADATA

- User Oriented

  Definition of attributes

- System oriented

  and field edit procedure names

Record

USES OF METADATA

- Map source system data to data warehouse tables
- Generate data extract, transform, and load procedures for import jobs
- Help users discover what data are in the data warehouse
- Help users structure queries to access data they need

DESCRIBING THE DATA WAREHOUSE

- I/P - O/P object
  File/Table
  Archive Period
- Relationship
- Data element - Name, Defn., Type
- Relationship Member - Role, Participation Constraint
- Field Assignment

EXTRACT JOBS

- Wholesale replace
- Wholesale append
- Update replace
- Update append

PLANNING

- Interviews
- Data quality
- Data Access
- Timeliness and history
- Data sources
- Decide on Architecture

DEVELOPMENT PROCESS

- Project Initiation
- Develop Enterprise Info. Architecture
- Design Data Warehouse Database
- Transform data
- Manage Metadata
- Develop User-Interface
- Manage Production

EVOLUTION

- Support the current DW baseline
- Enhance current baseline capabilities
- Define new business requirements
- Implement new baseline

MISTAKES

- Starting with the wrong sponsorship chain
- Setting expectations that cannot be met
- Believing that DW design is the same as Transactional Database Design
- Believing the Performance, Capacity Promises
- Believing that Once the Data Warehouse Is Up and Running Problems are finished

- NSWCDD - ORACLE on UNIX
- Harris Semiconductor IYM
  with Alarms, INGRES