Tashfeen Abbasi i22-2041 DS-D

(Repo of organization electronically stored data) **Data Warehouse** (OLAP) (Historical data)

Lecture no 1

What is data warehouse? (collective data repository)

- Very large database.
- Start around 1TB → several petabytes
- Wal-mart data warehouse is 101 TB.
- Several servers & needs impressive amount of computing power.
- Snapshots → Copies of data at particular moment of time.
- Contains snapshot of operational data.
- Obtained through data cleansing.
- Bill Inmon → data warehouse is
  - Subject oriented
  - Non-volatile
  - Integrated
  - Time Varient.
- Subject Oriented → data in data warehouse is organized so all data elements relating to same real-world event are linked together.
- Integrated → DW contains data from most or all organizational's operational systems and data is made consistent.
- Non-Volatile →
  - Data in DW is never over-written or deleted. Once committed, data is static, read-only
  - Data is loaded not updated.
  - Changes occur, new snapshot is written.
- Time Varient →
  - Reports are produced showing changes over time.
  - operational systems → 60-90 days span of time (stores & analyze)
- DW → 5-10 years stores and analyze

- Datawarehouse → • resides on computers.
  • Run on DBMS such as Oracle, Micro soft SQL Server.
  • retain data for long period of time
- Pessimistic locking → Transaction locks data before making Changes.

- OLTP
  - operational data
  - for data entry & retrival
  - reflects only current state of data.

OLAP (provide timely, accurate and understandable information)
- represent front-end analytics
- decision oriented.
- Many Flavors → ROLAP, DOLAP, MOLAP.

- Operational DB
  - Mostly updates
  - small transactions
  - MB-TB of data
  - Raw data
  - related to users
  - up-to date data.

DW
- Mostly reads
- long complex queries
- GB-PB of data
- summerized data
- Decision makers
- Slightly out dated.

- Complex queries can be solved by DW b/c DW tables are rearranged and pre-aggregated
- DW is Base repo for front-end analytics
  - OLAP
  - KDD (Knowledge discovery in database) (Data mining)
  - Data visualization
  - Reporting.

↓ Construct models of data in question.

Decision Support System.

- Users of DW are called DSS analysts (Buissness person)
  - Explorative line of work.

Lecture no 2

Lifecycle of DW & Basis Architecture.

- Classical SDLC
  - Building softwares
  - Design & analysis
  - Programming & testing
  - Implementation.
  - Step-by-step process.

OLTP

DW SDLC / CLDS
- data integration
- Program against data
- design DSS system
- Analyze results.

OLAP

- CLDS (data driven development life cycle)
  ↓
  Spiral development methodology.

- Operating a DW → • Monitoring • Extraction • Transforming • loading • analyzing

- Monitoring → • caring of data sources
  - Detect data modification
  - decides which data modification should be processed in next step.

  ↓ techniques are

  Monitering techniques → • Replication mechanism
  - Protocol based mechanism
  - Application based mechanism
  - Active mechanism - Event Condition action (ECA)

- Extracting → • read data → selected from monitering phase and insert in data structures of workplace.
  - large data (use compression)

  ↓ depends on hardware and software.

  - Time-point for extraction → • Periodical
    - on request
    - Event driven
    - Immediate.

- Transforming → Normalization
  Date handling
  Measurements units & scaling
  Save calculated values
  Aggregation.
  Data cleaning
- Loading → • takes place during weekends or nights.
  - Intial loading → initilized DW
  - periodical loading → update DW
  - Initial loading → performed by
    - partioning
    - parallelization
    - incremental actualization
- Analyzing → • Data access
  - OLAP
  - Data mining

- Architecture → • Basic architecture.
  - Storage Structures
  - Tier Architectures
  - Distributed DW
  - DW data modeling

• Basic Architecture →

(1) • Data staging area
→ Storage + process area (ETL process)
→ represent everything that happens b/w • operational source system
• data presentation area.
→ architectural requirement for data staging area i) off limits to business users
2) does not provide query & presentation service.

(2) Data presentation area → • where data is organized, stored and made available for queries, and other analytical processes.