VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

Populating a Data Warehouse

Lecturer : Dr Pavle Mogin

SWEN 432
Advanced Database Design and Implementation

Plan for Populating a DW

- Data extracting
- Data cleaning
- Data loading
- Data refreshing

- *Readings* :
  - *S. Chaudhuri, U. Dayal:*

    An Overview of Datawarehousing and OLAP Technology

Some Terminology

- ***Source*** data are data from operational databases and external sources
- ***Base*** data are Data Warehouse fact table or dimension table data
- ***Derived*** data is Data Warehouse data produced by materializing views and building auxiliary access structures (indexes)

Populating and Updating a DW

- Data Warehousing systems use a variety of software tools for:
  - data extraction,
  - data cleaning,
  - DW loading, and
  - DW refreshing
- All these tools have the goal to provide data of high quality for the decision making purposes
- ETL = Extraction, Transformation, and Loading

Data Extraction

- Data from operational databases and external sources are extracted using gateways
- A ***gateway*** is an application program interface that allows a client program to generate SQL statements to be executed at a server
- Common examples of gateways are:
  - Open Database Connectivity (ODBC),
  - Object Loading and Embedding for Databases (OLE), and
  - Java Database Connectivity (JDBC)

Data Cleaning

- Since a Data Warehouse is built using data from various sources, there is a high probability of errors and anomalies in data
- Most frequent errors are:
  - inconsistent field lengths,
  - inconsistent attribute names,
  - inconsistent value assignment,
  - missing entries, and
  - violated integrity constraints
- ETL software transforms, cleans, and discovers violation of constraints in input data

Data Cleaning Tools

- ***Data migration*** tools allow simple data transformation rules to be specified:
  - Replace *Surname* by *Last_Name*
  - Convert *pound* to *kg*
- ***Data scrubbing*** tools are more sophisticated, they use domain specific knowledge (business rules of the real system) to clean data from various sources
  - Use functional dependency *ProductID* $\rightarrow$*Prod_Name* to clean product data from production and marketing databases,
  - Convert country code number part of a telephone number into country name (e.g. 64 into New Zealand)
  - Fill in missing *Address* data
- ***Data auditing*** tools are used to scan data and discover strange patterns (data mining)
  - Products that have been never sold
  - Exceptionally large attribute values (although within limits allowed)

Building Time Dimension

- Operational data rarely contain precise time data
- In OLTP databases time is not so important as in OLAP databases
- At most, operational data records contain the date
- So, to satisfy need for a unique and immutable *Timeid* and build a time attribute hierarchy, the time dimension is built before Data Warehouse data loading

Data Loading

- Before loading data some additional data preprocessing has to be done:
  - sorting,
  - summarization,
  - aggregation,
  - building indexes, and
  - building materialized views
- The load utilities have to deal with very large volumes of data during small time slots (a night)
- Sequential loads would take weeks (or more), so pipelined and parallel loads are exploited instead

Loading Data

- Doing a full load has advantage of using the current version of a Data Warehouse for queries during the time the load is in progress
- But doing a full load can last to long
- To reduce the amount of data, incremental loading during refresh is used instead
  - Only the updated operational tuples influence data to be inserted
- But the incremental load conflicts with ongoing queries
  - To avoid conflicts, incremental loading is performed as a sequence of transactions that commit periodically

Data Refreshing

- DW refreshing is an alternative to full load
- Refreshing a Data Warehouse consists in propagating updates on source data (operational and external data) to corresponding updates of base and derived data (in the Data Warehouse)
- The Data Warehouse refresh policy has to concern two issues:
  - Frequency, and
  - Procedures

  of data refreshing

Data Refreshing Frequency

- Data refreshing frequency depends on user needs and OLTP traffic
- Usually, a DW is refreshed periodically (daily or weekly)
- But, if users need current data, it is necessary to propagate every relevant update from OLTP data to OLAP data
- Also, if the OLTP update traffic is high and the DW refreshment frequency low, data volumes during refreshment may overwhelm the refreshment utility
- So, OLTP update traffic also influences refreshment policy (high traffic leads to frequent updates)

Data Refreshing Procedures

- Generally, DW refreshing is made using one of the following two techniques:
  - Data Shipping and
  - Transaction shipping
- Both techniques suppose that the operational DBMS supports replication servers that incrementally propagate updates from a primary database to replicas
- If the operational database system is a legacy one, and does not support replication, extracting the whole source database can be the only choice

Data Replication

- To reduce data transfer cost and enhance availability, distributed databases store copies of data on every location where data is in high demand
- Data copies are called data ***replicas*** or ***snapshots***
- There are two kinds of data replication:
  - Synchronous and
  - Asynchronous

Data Replication

- In ***synchronous*** replication, when source data is updated, all its data replicas have to be updated before the transaction commits
- In ***asynchronous*** replication, the source updates are propagated to replicas periodically (sometimes even long after transaction commits)
- A Data Warehouse is considered as an asynchronous replica

Data and Transaction Shipping

- In ***data shipping*** a table in the DW is treated as a remote snapshot of a table in the source database
  - Whenever the source table changes, a mechanism called *After_row* trigger is used to update a snapshot log file, and
  - A refresh procedure is set up to propagate the update to the DW (at some time)
- In ***transaction shipping*** a regular transaction log file is used instead of the trigger and the snapshot log file:
  - At the source site the transaction log is checked for updates that influence the replicated tables
  - The log records containing appropriate changes are transferred to replication servers

Data versus Transaction Shipping

- Data shipping is more appropriate when operational databases and The Data Warehouse are from different vendors, since transaction log files are not standardized
- Transaction shipping is more appropriate for homogeneous systems, since the problem of interpreting the contents of the transaction log file is not present
- Transaction shipping requires less resources of the operational database server

Maintenance of Materialized Views

- Making a view consistent with its (DW) base tables is called ***view refreshing***
- If the cost of algorithms for view refreshing is proportional to the change of the view, they are sad to be ***incremental***
- A view maintenance policy is a decision about when a view has to be refreshed

View Updates

- View update can be:
  - ***Immediate*** (within the same transaction that updates the base tables), or
  - ***Deferred*** (a period of time after the base tables are updated)
- Deferred update can be done either:
  - During the time a view is used for a query evaluation for the first time after the update of base tables,
  - Periodically, in regular time intervals, or
  - Forced, after a certain number of base table updates

View Refreshing and Aggregates

- A special consideration is needed when aggregate views are refreshed
- Views containing distributive aggregates are refreshed without any problem
- Views containing algebraic aggregates are easily refreshed if they contain all other necessary data (most frequently it is count);
- Views containing holistic aggregates are hard to refresh, they are rather build every time from scratch

Summary

- Data extraction is done using gateways (ODBC and JDBC)
- Data cleaning software reconciles inconsistent:
  - Field length,
  - Attribute names,
  - Value assignment,
  - Missing entries

  and discovers integrity violations and suspicious data patterns
- Data loading can be:
  - Either full, or
  - Incremental

Summary (continued)

- Incremental load is done during data refreshing
- Data refreshing consist of updating Data Warehouse base and derived data by data inferred from changes made to source data
- Data refreshing techniques:
  - Data shipping
  - Transaction shipping