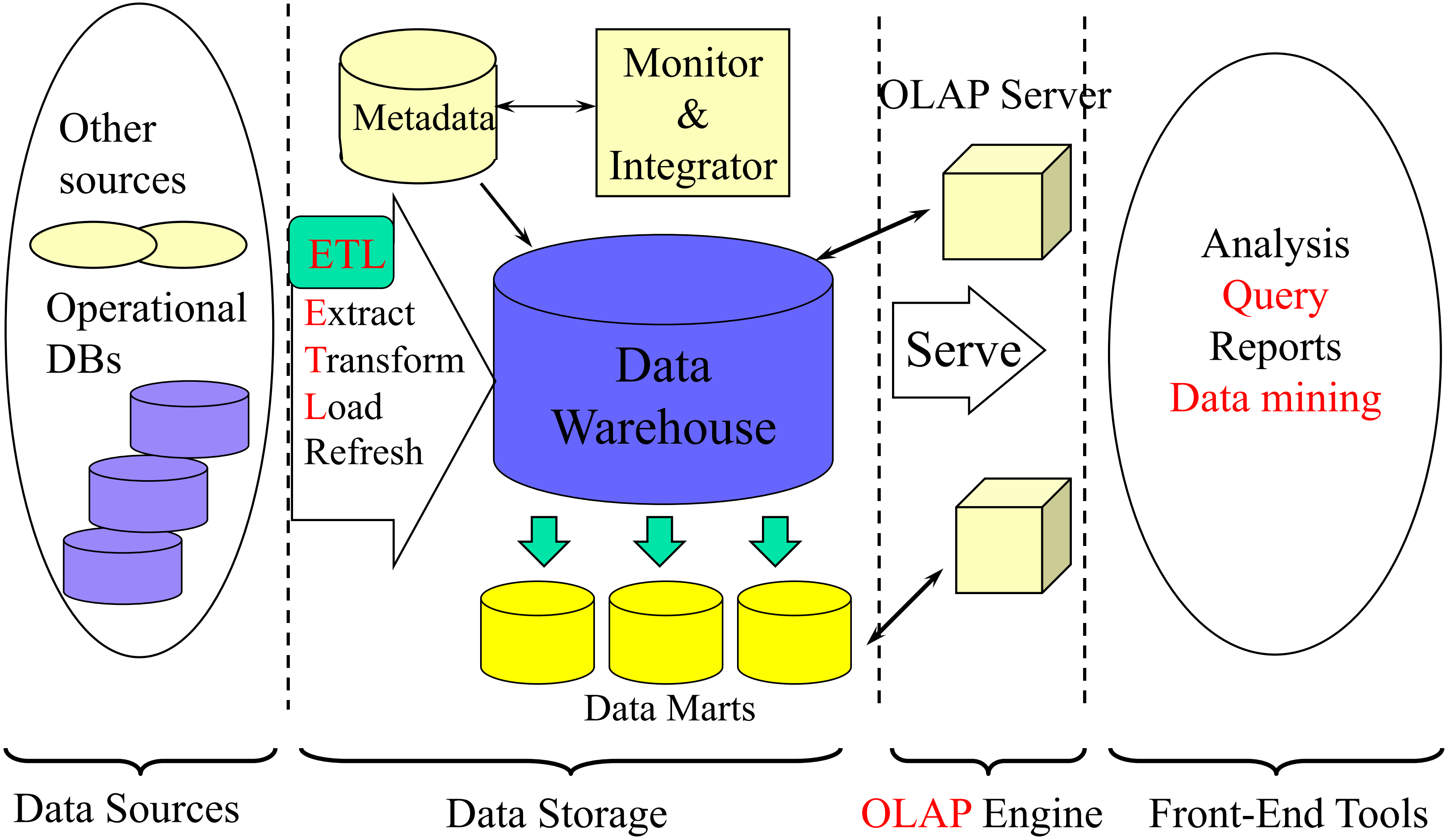
Data Warehousing (Cont.)

- DW Modeling: Data Cube and OLAP
- DW Tools and Terms
- DW Design and Usage
- DW Maintenance Issues
- Summary

Data Warehouse Terms and Jargons

- Data warehouse modelling
  - Multidimensional data modeling: Data cube and OLAP
- Data warehouse architecture
  - Multi-Tiered framework
- Data warehouse models
  - Enterprise warehouse, Data marts, Virtual warehouse
- Data warehouse backend tools
  - Extraction, Transformation, and Loading (ETL)
  - Metadata repository
- Data Warehouse Platforms

Data Warehouse Architecture: A Multi-Tiered Framework



Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
  - provides decision-making support to different departments of an enterprise.
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users (e.g. accounting, finance, sales, purchases, or inventory, etc.). Its scope is confined to specific, selected groups.
- Operational Data Store
  - A set of views over operational databases
  - It is used when an organization's reporting requirements cannot be met by OLTP system.

DW Backend Tools:
Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Loading**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

Metadata Repository

Meta data is the data defining warehouse objects. It stores:

- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data definition, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

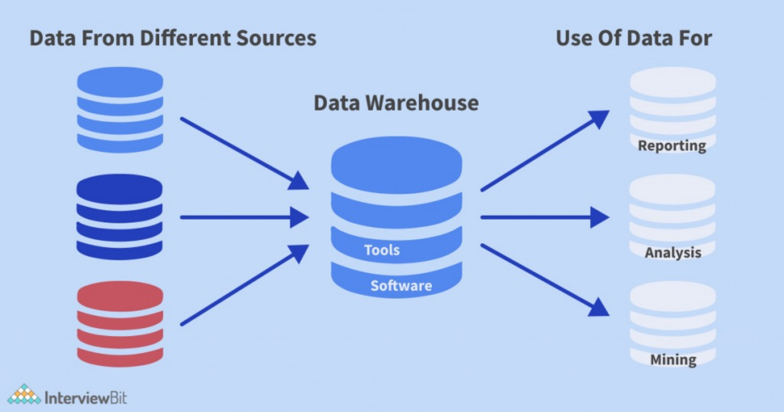
Data Warehouse Platforms

Recall that

- A data warehouse is a platform where data can be integrated precisely to generate meaningful business insights.
- With data warehousing, businesses can
  - quickly access information,
  - speed up query response times, and
  - gain new insights into big data.
- Having your data organized in a single location allows you to perform analysis and reporting at a variety of aggregate levels.

Data Warehouse Platforms

- Data warehouses have become significantly cheaper for businesses due to the emergence of cloud technology.
- There are various actions that can be done on data using the Data Warehouse tools, including:
  - Cleaning data and separating it from junk or duplicate data.
  - Extraction, transformation, and loading (ETL) of the data from various formats of sources into a single common format at the destination.
  - Querying data from warehouse in order to fetch, update, delete or analyze different combinations of data.
  - Creating reports for analysis and business decision-making processes.



Data Warehouse Platforms

Amazon Redshift

- The Amazon Redshift platform is a fully managed cloud-based data warehouse designed to store and analyze large-scale data using SQL queries with Business Intelligence (BI) tools like Tableau, Microsoft Power BI, etc.
- It is a simple, cost-effective tool and is considered a very critical part of Amazon Web Services, one of the most popular cloud computing platforms.
- Data analysis can be done with the system in a matter of seconds, which makes it ideal for high-speed data analysis.

Data Warehouse Platforms

Microsoft Azure

- In Microsoft Azure, there is an analytical data warehouse called SQL Data Warehouse (SQL DW), which is scalable for petabytes and built according to SQL Server.
- Through all of the over 200 product and cloud services, highly scalable and efficient applications can be built, run, and managed across multiple cloud networks using AI (Artificial Intelligence) and Machine Learning.

Azure

Data Warehouse Platforms

Google BigQuery

- BigQuery is a cost-effective data warehousing tool with built-in machine learning capabilities that allows scalable analysis over petabytes of data. This is a Platform as a Service that makes it easy to query big datasets using super-fast SQL queries.

Snowflake

- Snowflake is a cloud-based Data Warehouse Tool that provides a faster, easier-to-use, and more flexible framework than other data warehouses. It offers a complete SaaS (Software as a Service) architecture. Snowflake simplifies data processing by letting users work (data blending, analysis, and transformations) with varied forms of data structures (structured and semi-structured) using a single language, SQL.

Also, **PostgreSQL**, **Teradata** and others…

Data Warehouse

Design and Usage

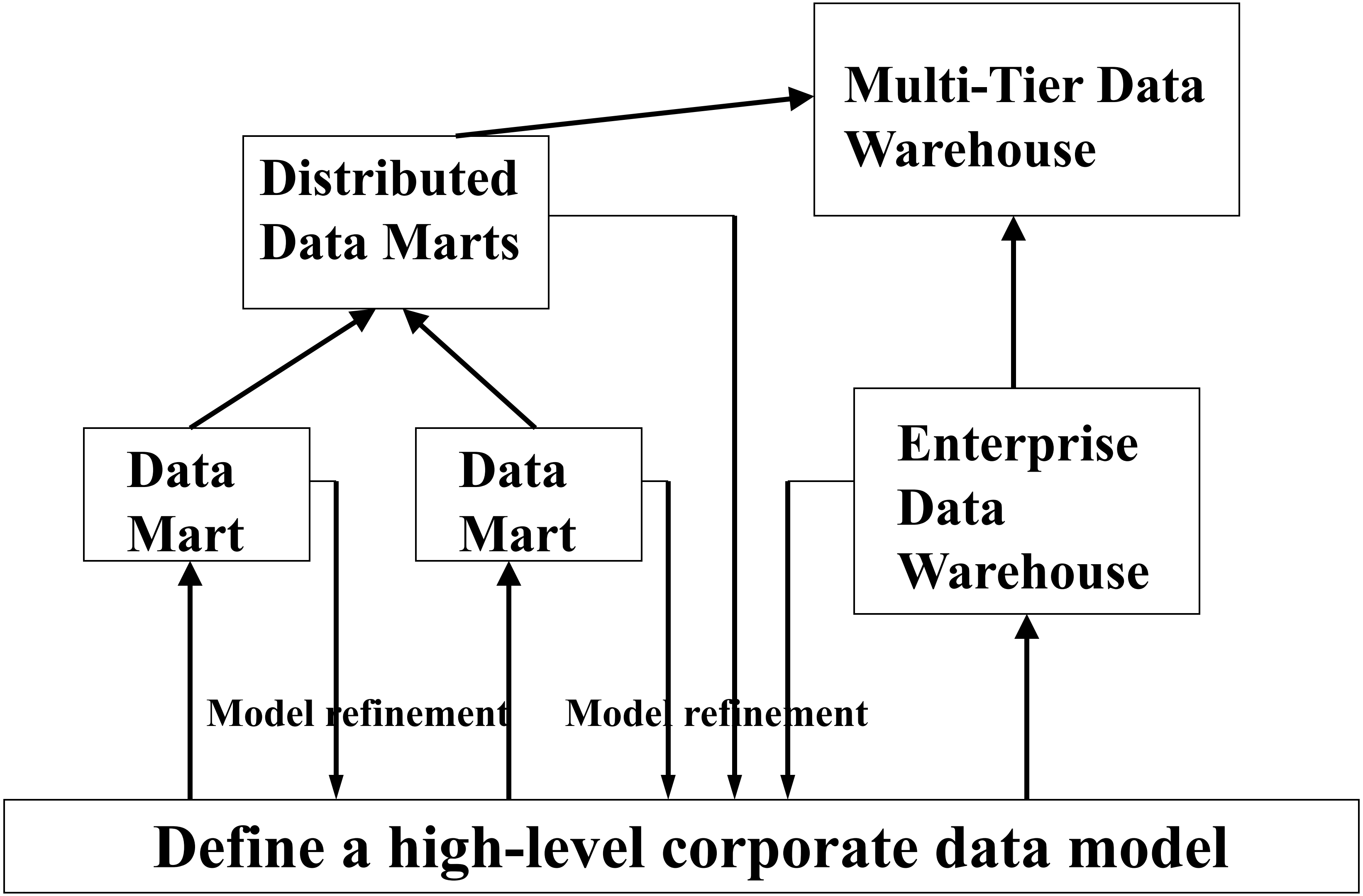
Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse Design Process

- **Top-down, bottom-up approaches or a combination** of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- **From software engineering point of view**
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- **Typical data warehouse design process**
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

Data Warehouse Development: A Recommended Approach


Multi-Tier Data Warehouse
Distributed Data Marts
Data Mart
Data Mart
Enterprise Data Warehouse
Model refinement
Model refinement
Define a high-level corporate data model

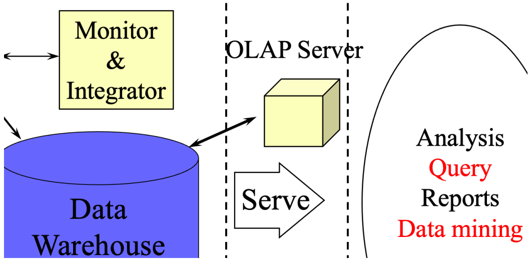
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

- Why online analytical mining?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks

OLAP Server Architectures



- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Data Warehouse
Maintenance Issues

Views, OLAP and Materialization

- Views are frequently used in decision support systems to allow data analyst to consider just his/her part of the business
- Decision support queries are typically aggregate queries over very large fact tables
- To allow fast answers, view materialization is a viable alternative
- The DW itself is considered to be a (materialized) view of the operational databases and external data sources
- When deciding which views to materialize, one should consider the following issues:
    - How many queries potentially can be speeded up?
    - How many space will be required to store the views?, and
    - How will the views influence the DW maintenance (update)?

Choosing Views to Materialize

- The choice of views to materialize is complex, because the range of views that can be used for a query evaluation is very broad
- On the other hand, materialized views strongly influence the storage occupancy and DW maintenance time
- It is the goal to materialize a small, carefully chosen set of views that can be used to evaluate the majority of important queries
- Conversely, once the set of materialized views is determined, the query processor has to choose one of them to evaluate a given query
- ROLAP Engines offer automatic advise to a DW/DB administrator on which materialized views to build and which ones to drop
- That advising is done on the base of statistical data gathered during a Data Warehouse querying

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 said 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 table), and
  - Deferred (some times after the base tables are updated)
- Deferred update can be done:
  - During the time view is used for a query evaluation for the first time after base table update,
  - Periodically, in regular time intervals
  - 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
- Views containing holistic aggregates are hard to refresh, they are rather built from scratch, again

To summarize (view materialization)

- OLAP queries are typically aggregate queries over very large fact tables
- To allow fast answers, view materialization is a viable alternative
- Materialized views strongly influence the storage occupancy and DW maintenance time
- It is the goal to materialize a small, carefully chosen set of views that can be used to evaluate most of the important queries

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 purpose

Typically referred as ETL (Extract, Transform, and Load)

Data Extraction

- Data from operational databases and external sources are extracted by 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 Tools

Data cleaning tools transforms, cleans, and discovers violation of constraints in input data

- *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 (rules of behavior in the real system) to do cleaning of data from various sources
  - Use functional dependency *ProductID->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. 852 into Hong Kong)
  - 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)

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 (night)
- Sequential loads would take weeks (or more), so pipelined and parallel 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 too long
- To reduce the amount of data, incremental loading during refresh is used instead
- Only the updated operational tuples influence data to be inserted

Data Refreshing

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

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 is low, data volumes during refreshment may overwhelm the refreshment utility
- So, OLTP update traffic also influence 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 support 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

To summarize

- Data extraction is done by means of gateways
- Data cleaning software reconciles inconsistency and discovers integrity violations and suspicious data patterns
- Data loading can be either full, or incremental
- Some terminology:
  - Source data are data from operational DBs and external sources
  - Base data are DW fact table or dimension table data
  - Derived data is DW data produced by materializing views and building auxiliary access structures (indexes)

Part II's Summary

- Many cloud-based DW platforms exist!
- Data Warehouse Architecture, Design, and Usage
  - Multi-tiered architecture
  - Business analysis design framework
  - Information processing, analytical processing, data mining, OLAM (Online Analytical Mining)
- Maintenance Issues
  - ETL tasks
  - Partial vs. full vs. no materialization
  - Populating DW

Acknowledgement

- Slides/Materials of
  - J. Han et al.’s DM: Concepts and Techniques textbook
  - https://web2.utc.edu/~djy471/
- Photos from Internet

References

- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.