Chapter 1

DATAWARE HOUSE
Basic Concepts

Contents

- What Is a Data Warehouse?
- Differences between Operational Database Systems and Data Warehouses, need of Data Warehouse
- Data Warehousing: A Multi-tier Architecture
- Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse
- Online Analytical Processing,
- Characteristics of OLAP, OLAP Tools,
- OLAP Data Modelling, OLAP Tools and the Internet
- Difference between OLAP and OLTP

Data, Data everywhere yet …

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences
- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

So What Is a Data Warehouse?

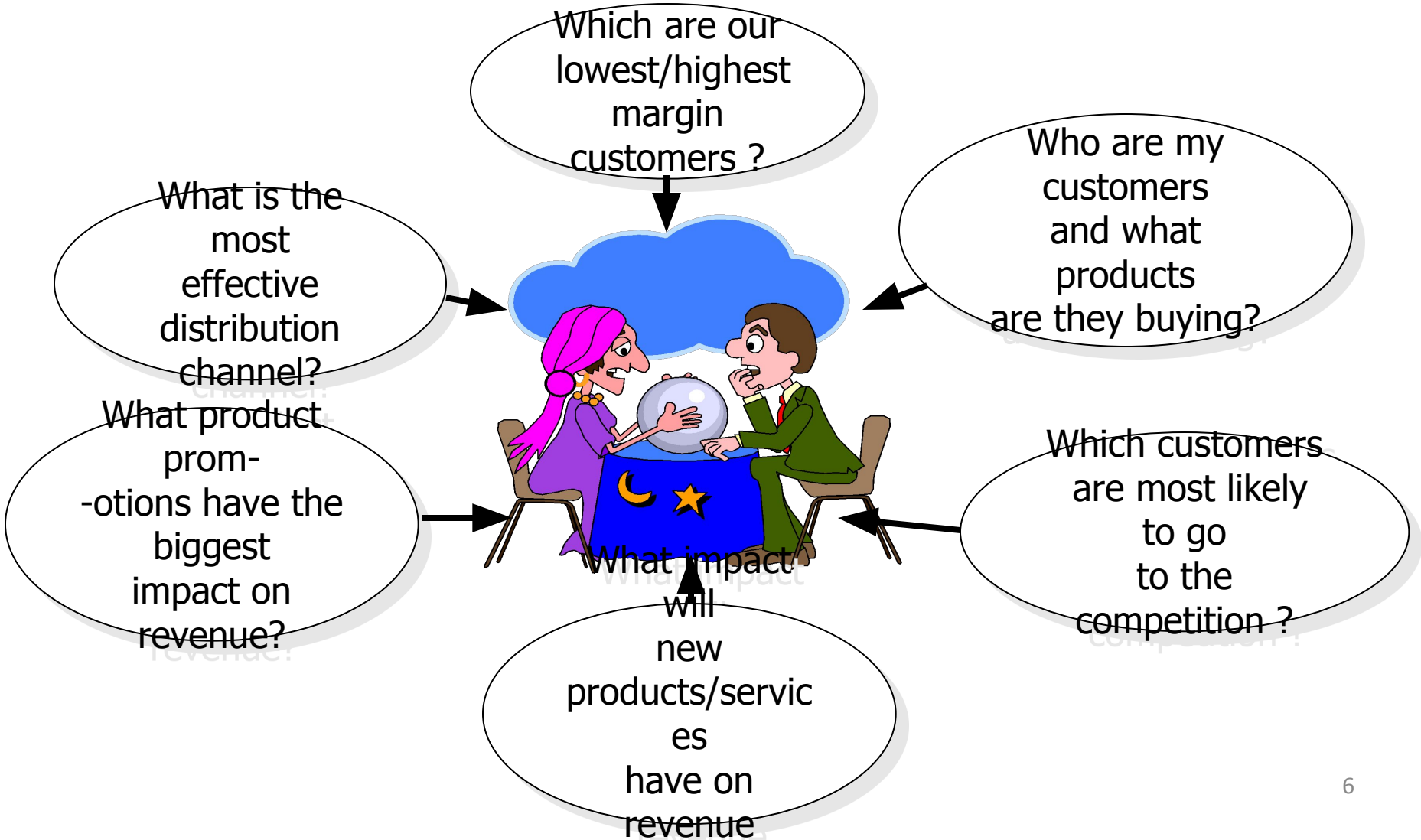
Definition: A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context. [Barry Devlin]

- By comparison: an OLTP (on-line transaction processor) or operational system is used to deal with the everyday running of one aspect of an enterprise.
- OLTP systems are usually designed independently of each other and it is difficult for them to share information.

Why Do We Need Data Warehouses?

- Consolidation of information resources
- Improved query performance
- Separate research and decision support functions from the operational systems
- Foundation for data mining, data visualization, advanced reporting and OLAP tools

Why Data Warehousing?



What Is a Data Warehouse Used for?

- Knowledge discovery
  - Making consolidated reports
  - Finding relationships and correlations
  - Data mining
  - Examples
    - Banks identifying credit risks
    - Insurance companies searching for fraud
    - Medical research

How Do Data Warehouses Differ From Operational Systems?

- Goals
- Structure
- Size
- Performance optimization
- Technologies used

What is a Data Warehouse?
A Practitioners Viewpoint

"A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."

-- Barry Devlin, *IBM Consultant*

What is a Data Warehouse?

An Alternative Viewpoint

"A DW is a

- subject-oriented,
- integrated,
- time-varying,
- non-volatile

collection of data that is used primarily in organizational decision making."

-- W.H. Inmon, Building the Data Warehouse, 1992

A Data Warehouse is...

- Stored collection of diverse data
  - A solution to data integration problem
  - Single repository of information
- Subject-oriented
  - Organized by subject, not by application
  - Used for analysis, data mining, etc.
- Optimized differently from transaction-oriented db
- User interface aimed at executive

… Cont'd

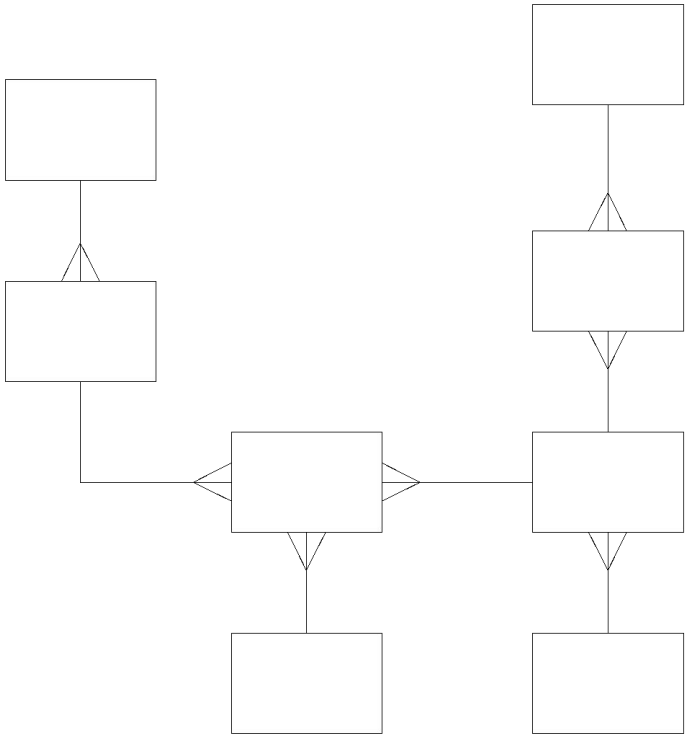
- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at Sainsbury's
  - Complete client histories at insurance firm
  - LSE financial information and portfolios

Comparison Chart of Database Types

Data warehouse	Operational system
Subject oriented	Transaction oriented
Large (hundreds of GB up to several TB)	Small (MB up to several GB)
Historic data	Current data
De-normalized table structure (few tables, many columns per table)	Normalized table structure (many tables, few columns per table)
Batch updates	Continuous updates
Usually very complex queries	Simple to complex queries

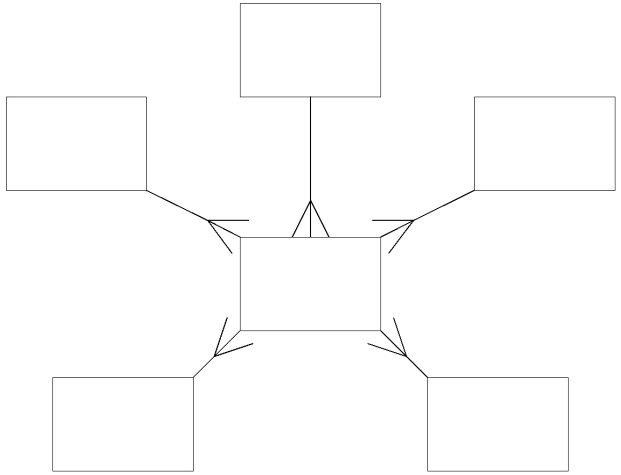
Design Differences

Operational System



ER Diagram

Data Warehouse



Star Schema

Data Warehouses, Data Marts, and Operational Data Stores

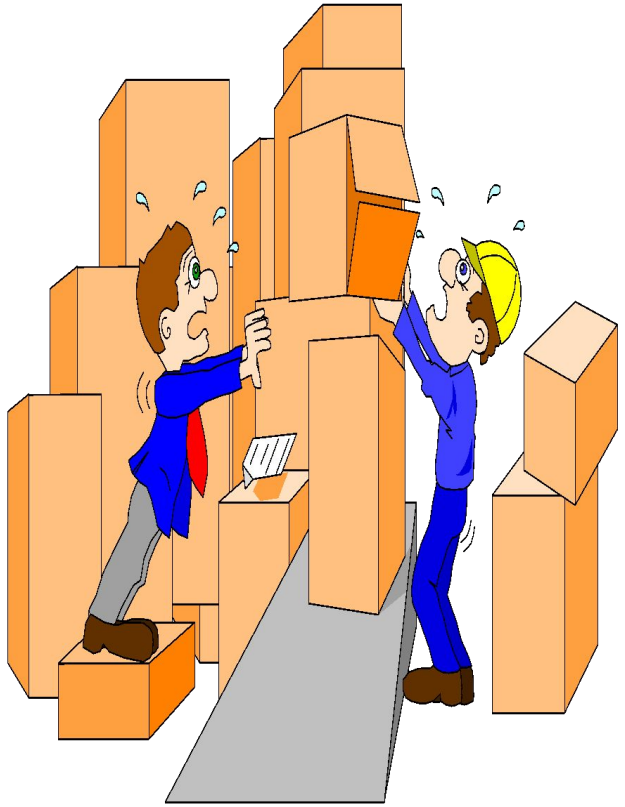
- Data Warehouse – The queryable source of data in the enterprise. It is comprised of the union of all of its constituent data marts.
- Data Mart – A logical subset of the complete data warehouse. Often viewed as a restriction of the data warehouse to a single business process or to a group of related business processes targeted toward a particular business group.
- Operational Data Store (ODS) – A point of integration for operational systems that developed independent of each other. Since an ODS supports day to day operations, it needs to be continually updated.

Decision Support

- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgements

What are the users saying...

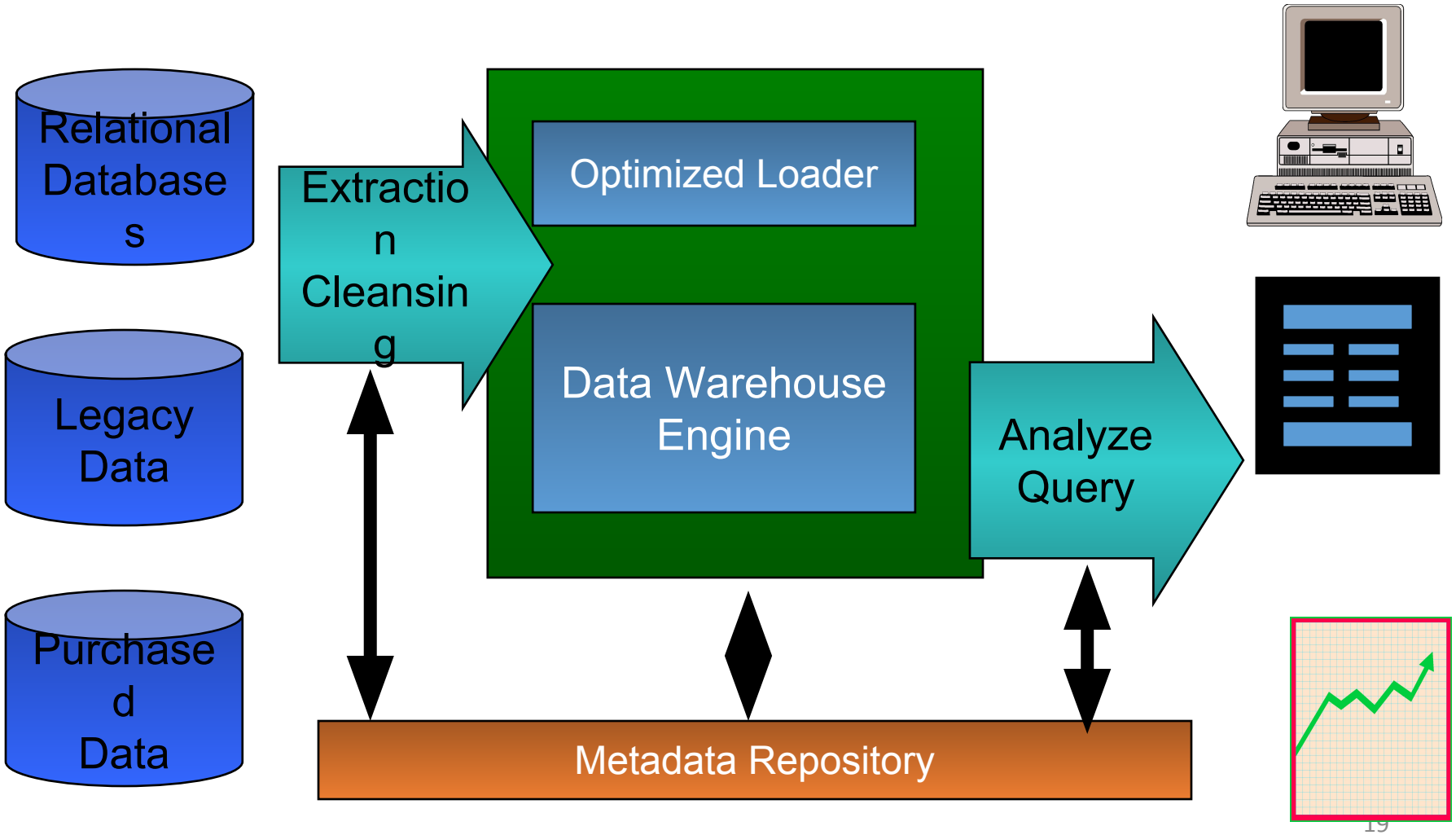
- Data should be integrated across the enterprise
- Summary data had a real value to the organization
- Historical data held the key to understanding data over time
- What-if capabilities are required



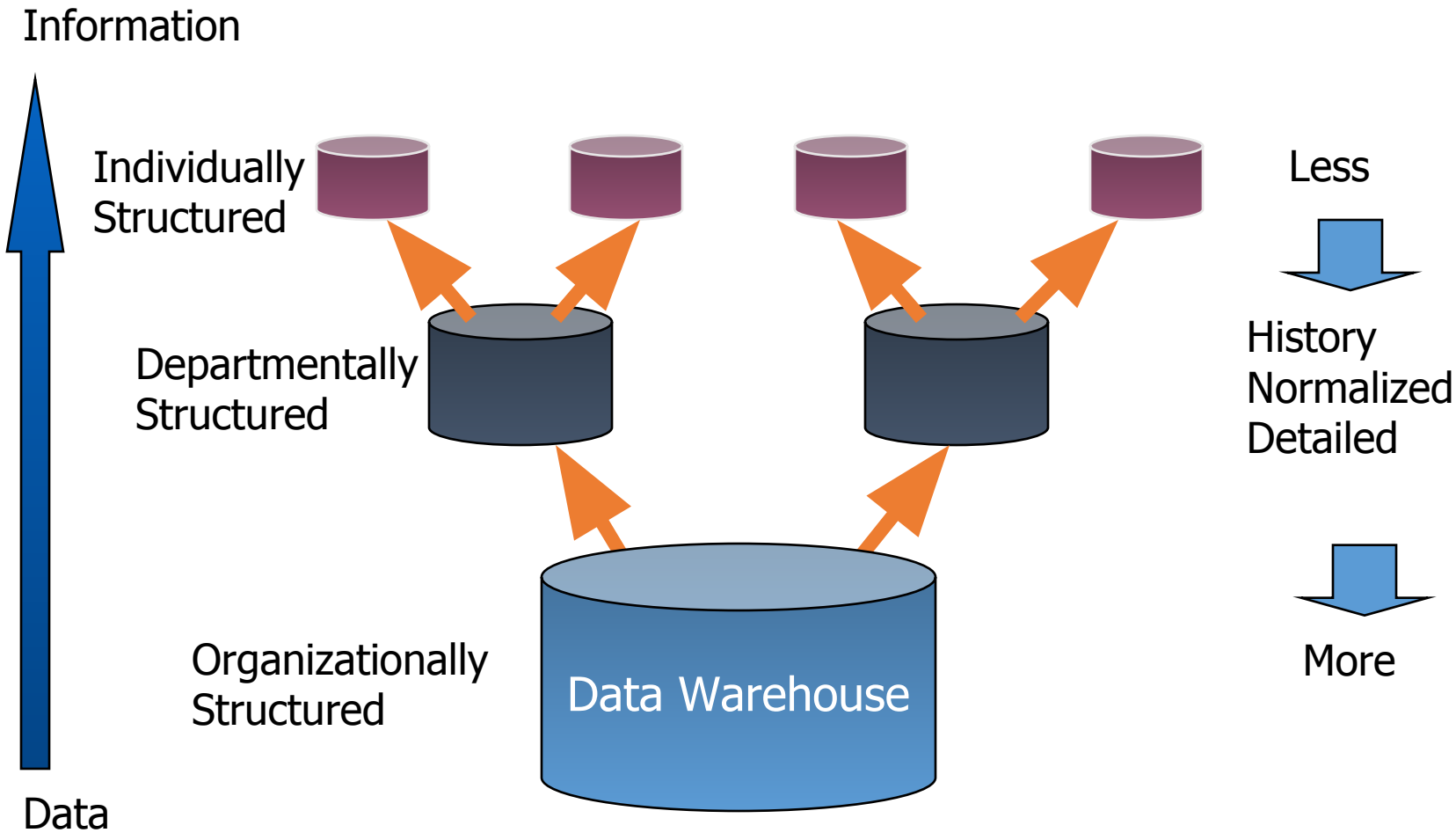
Data Warehousing -- It is a process

- Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible
- A decision support database maintained separately from the organization's operational database

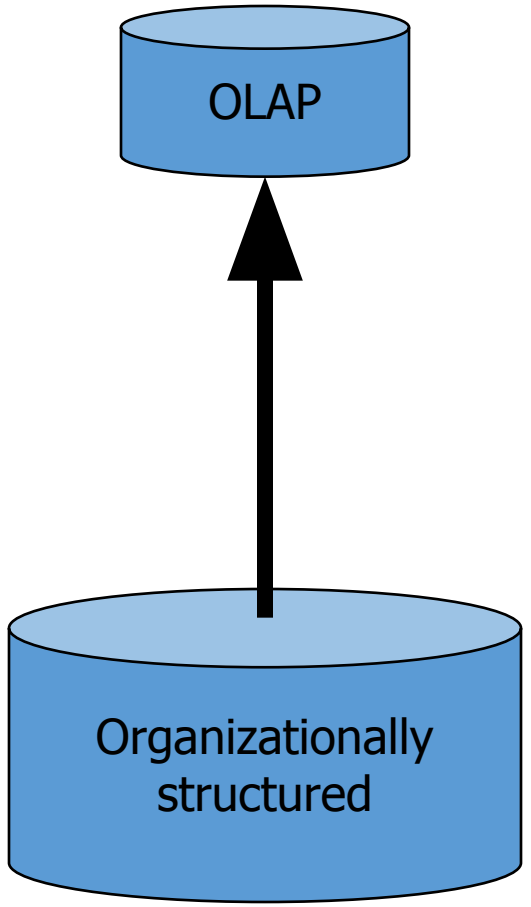
Data Warehouse Architecture



From the Data Warehouse to Data Marts

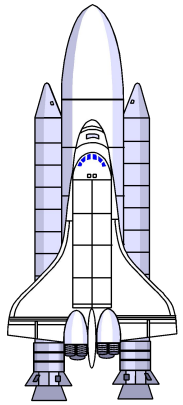


Users have different views of Data



Tourists: Browse information harvested by farmers

Farmers: Harvest information from known access paths



Explorers: Seek out the unknown and previously unsuspected rewards hiding in the detailed data

Wal*Mart Case Study

- Founded by Sam Walton
- One the largest Super Market Chains in the US

- Wal*Mart: 2000+ Retail Stores
- SAM's Clubs 100+Wholesalers Stores

    - This case study is from Felipe Carino’s (NCR Teradata) presentation made at Stanford Database Seminar

Old Retail Paradigm

- Wal*Mart
  - Inventory Management
  - Merchandise Accounts Payable
  - Purchasing
  - Supplier Promotions: National, Region, Store Level
- Suppliers
  - Accept Orders
  - Promote Products
  - Provide special Incentives
  - Monitor and Track The Incentives
  - Bill and Collect Receivables
  - Estimate Retailer Demands

New (Just-In-Time) Retail Paradigm

- No more deals
- Shelf-Pass Through (POS Application)
  - One Unit Price
    - Suppliers paid once a week on ACTUAL items sold
  - Wal*Mart Manager
    - Daily Inventory Restock
    - Suppliers (sometimes SameDay) ship to Wal*Mart
- Warehouse-Pass Through
  - Stock some Large Items
    - Delivery may come from supplier
  - Distribution Center
    - Supplier's merchandise unloaded directly onto Wal*Mart Trucks

Information as a Strategic Weapon

- Daily Summary of all Sales Information
- Regional Analysis of all Stores in a logical area
- Specific Product Sales
- Specific Supplies Sales
- Trend Analysis, etc.
- Wal*Mart uses information when negotiating with
  - Suppliers
  - Advertisers etc.

Why Separate Data Warehouse?

- **Performance**
  - Op dbs designed & tuned for known txs & workloads.
  - Complex OLAP queries would degrade perf. for op txs.
  - Special data organization, access & implementation methods needed for multidimensional views & queries.

- **Function**
  - Missing data: Decision support requires historical data, which op dbs do not typically maintain.
  - Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources: op dbs, external sources.
  - Data quality: Different sources typically use inconsistent data representations, codes, and formats which have to be reconciled.

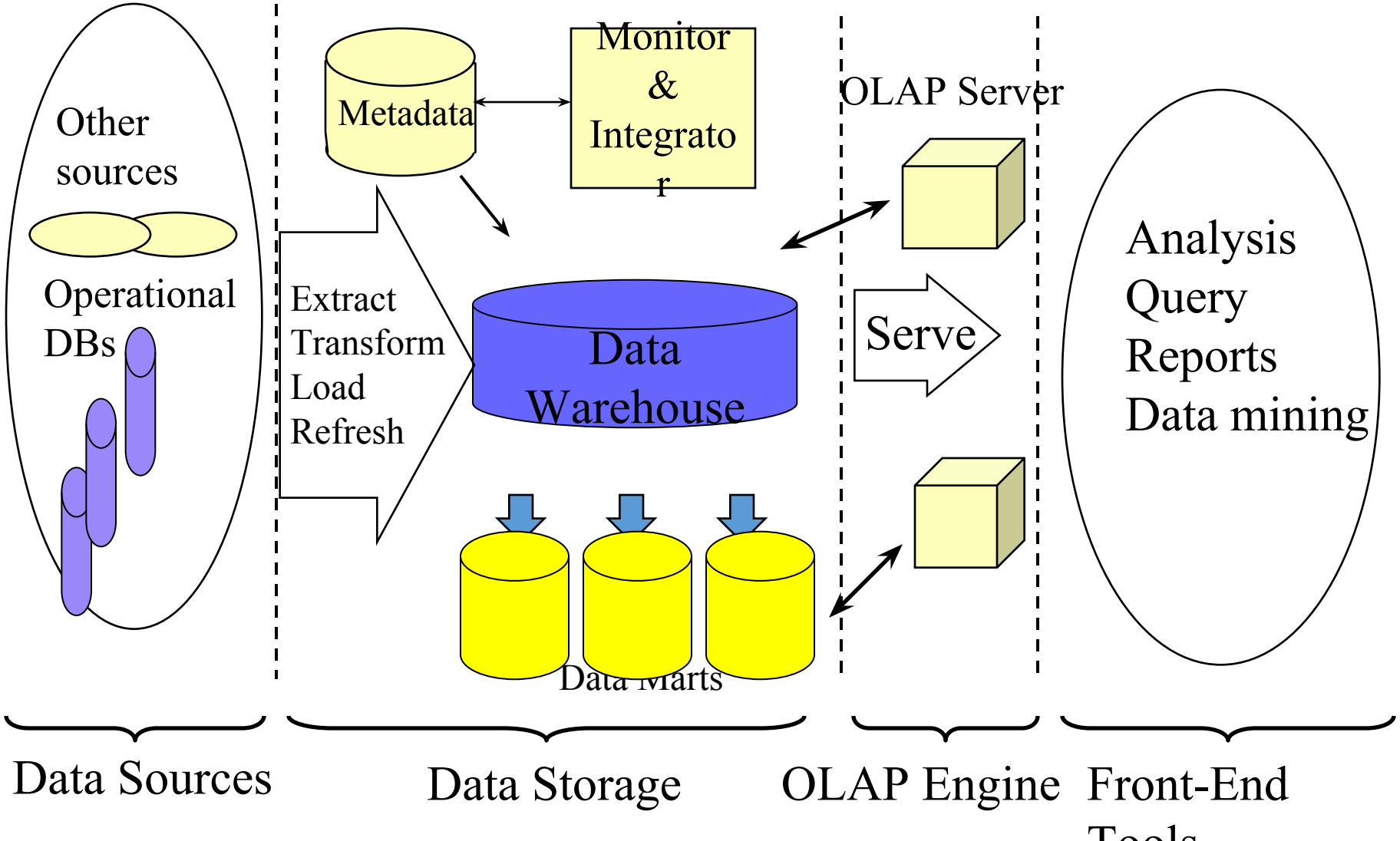
OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why a Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

Data Warehouse: A Multi-Tiered Architecture



Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

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
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

What is OLAP?

- **Online Analytical Processing (OLAP)** is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view.
- Analysts frequently need to group, aggregate and join data. These operations in relational databases are resource intensive. With OLAP data can be pre-calculated and pre-aggregated, making analysis faster.
- OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy. OLAP stands for Online Analytical Processing.

How does it work?

- A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.
- The extracted data is cleaned and transformed. Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

Basic analytical operations of OLAP

Four types of analytical operations in OLAP are:

- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice: *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

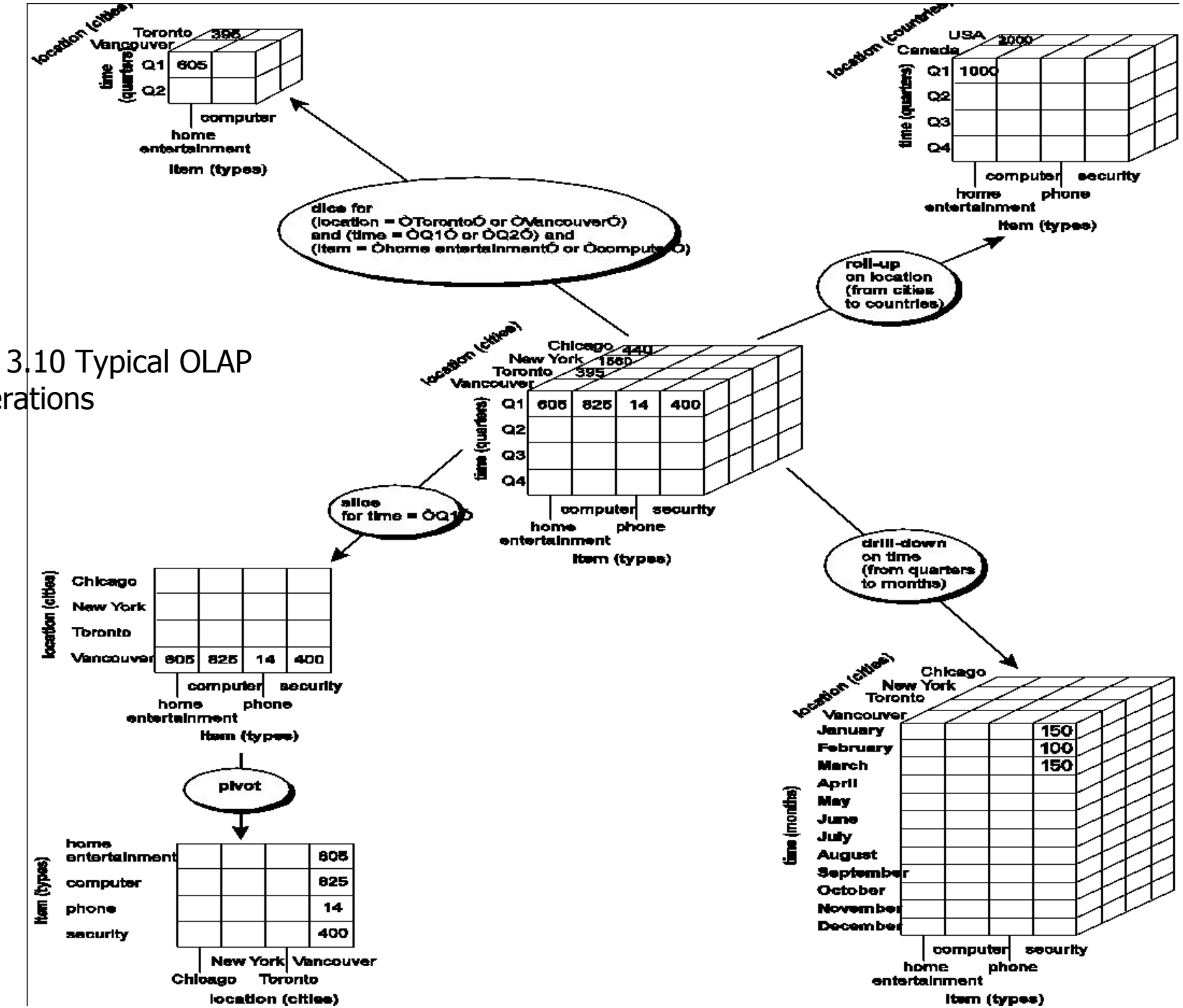


Fig. 3.10 Typical OLAP Operations

Advantages of OLAP

- OLAP is a platform for all type of business includes planning, budgeting, reporting, and analysis.
- Information and calculations are consistent in an OLAP cube. This is a crucial benefit.
- Quickly create and analyze "What if" scenarios
- Easily search OLAP database for broad or specific terms.
- OLAP provides the building blocks for business modeling tools, Data mining tools, performance reporting tools.
- Allows users to do slice and dice cube data all by various dimensions, measures, and filters.
- It is good for analyzing time series.
- Finding some clusters and outliers is easy with OLAP.
- It is a powerful visualization online analytical process system which provides faster response times

Disadvantages of OLAP

- OLAP requires organizing data into a star or snowflake schema. These schemas are complicated to implement and administer
- You cannot have large number of dimensions in a single OLAP cube
- Transactional data cannot be accessed with OLAP system.
- Any modification in an OLAP cube needs a full update of the cube. This is a time-consuming process

OLAP systems share four main characteristics:

- They use multidimensional data analysis techniques.
- They provide advanced database support.
- They provide easy-to-use end-user interfaces.
- They support the client/server architecture.

Top 10 Best Analytical Processing (OLAP) Tools

#1) Xplenty

- Availability: Licensed tool.
- Xplenty is a complete toolkit for building data pipelines. It provides features to integrate, process, and prepare data for business intelligence. It has coding, low-code, and no-code capabilities.
- No-code and the low code option will let anyone create ETL pipelines. Its API component will provide advanced customization and flexibility.

#2) IBM Cognos

- Availability: Proprietary License
- IBM Cognos is an integrated, web-based analytical processing system owned by IBM. It contains toolkit to perform analysis, reporting and score carding along with the provision to monitor metrics.
- It also contains numerous inbuilt components to meet various information requirements in an organization.

#3) Micro Strategy

- Availability: Licensed
- MicroStrategy is a Washington-based company that provides services on BI and mobile software worldwide.
- MicroStrategy Analytics enables companies/organizations to analyze large volumes of data and distribute the business specific insight throughout the organization securely.

Top 10 Best Analytical Processing (OLAP) Tools

#4) Palo OLAP Server

Availability: Open source

Palo is an MOLAP- multidimensional online analytical processing server typically used as a BI tool for various purposes like controlling and budgeting etc. Palo is a product of Jedox AG.

It has spreadsheet software as its user interface. Palo allows different users to share a centralized database that acts as a single source of truth.This type of flexibility to handle complex data models enables users to have a deeper insight into statistics.

#5) Apache Kylin

Availability: Open source

Apache Kylin is a multidimensional open-source analytics engine. It is designed to provide SQL interface and MOLAP in synchronous with Hadoop to support large data sets.

#6) icCube

Availability: Licensed

Switzerland-based company icCube owns a business intelligence software of the same name.

It sells an online analytical processing server that is implemented in Java as per J2EE standards. It is an in-memory OLAP server and it is compatible to work with any data source that holds its data in tabular form

Top 10 Best Analytical Processing (OLAP) Tools

7) Pentaho BI

Availability: Open source

Pentaho is a powerful open source tool that provides key BI features like OLAP services, data integration, data mining, extraction-transfer-load (ETL), reporting and dashboard capabilities.

Pentaho is built on Java platform that can work with Windows, Linux and Mac operating systems.

#8) Mondrian

Availability: Open source

Mondrian is a very interactive tool with outstanding features and strengths like its ability to work with categorical data, large data as well as geographical data. It is a general purpose data visualization tool. It consists of interlinked plots and queries.

Difference between OLAP and OLTP

- What is OLAP?
- OLAP stands for Online analytical processing. It includes software tools that help in analyzing data mainly for business decisions. This particular OLAP system enables users to examine database reports from various database systems at one time.
- What is OLTP?
- OLTP stands for online transaction processing. It encourages transaction-oriented applications in a 3-tier architecture. It also helps in managing the day-to-day transactions of an organization. ATM center is a prominent example of OLTP.

Difference between OLAP and OLTP cntd.

S.No.	OLAP	OLTP
1	OLAP stands for Online analytical processing.	OLTP stands for online transaction processing.
2	It includes software tools that help in analyzing data mainly for business decisions.	It helps in managing online database modification.
3	It utilizes the data warehouse.	It utilizes traditional approaches of DBMS.
4	It is popular as an online database query management system.	It is popular as an online database modifying system.
5	OLAP employs the data warehouse.	OLTP employs traditional DBMS.
6	It holds old data from various Databases.	It holds current operational data.
7	Here the tables are not normalized.	Here, the tables are normalized.
8	It allows only read and hardly write operations.	It allows both read and write operations.
9	Here, the complex queries are involved.	Here, the queries are simple.