BIG DATA: TECHNOLOGIES AND APPLICATIONS

4.1 Data Warehouses, OLAP, and Business Intelligence

Il-Yeol Song, Ph.D.
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104

DREXEL UNIVERSITY College of Computing & Informatics

Terminology

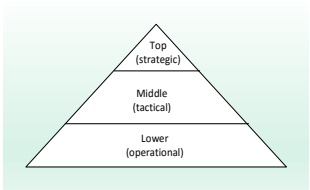
- **OLTP** (Online Transaction Processing):
  - Daily business operation with query and update processing
- **Data Warehouse (DW):**
  - An enterprise-wide data repository of historical data for decision support
- **Data Mart**:
  - A smaller targeted DW for a business process
- **OLAP** (Online Analytical Processing):
  - Multi-level and multi-dimensional data summarization
  - Complex query processing or report generation
  - Compare and contrast measures along dimensions

Terminology cont.

- **Dimensional Model/Star Schema**:
  - A DB schema structure for data warehouse/data mart
- **Data Cube**
  - A multi-dimensional data structure for OLAP
- **ETL**: Extraction, Transformation, and Loading
- **Data Mining**:
  - The algorithmic process of automating knowledge discovery and actionable rules
- **Business Intelligence**:
  - The processes, technologies, and tools that analyzes business data to improve business operation and to derive profitable business actions

Decision Making Hierarchy

Decision making hierarchy



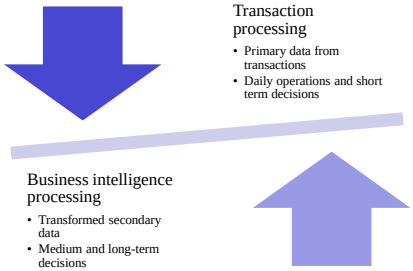
Typical decisions

Identify new markets, choose store locations

Choose suppliers, forecast sales

Resolve order delays, schedule employees

Comparison of Processing Environments

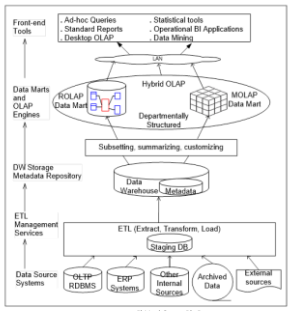


Data Warehouse

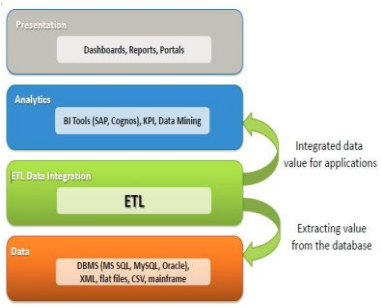
- A DW is an ***integrated repository* of data** that is put into a form that can be *easily understood*, *interpreted*, and *analyzed* by the people who need to use it ***to make decisions*.**
- Data are *extracted* from operational systems, then *cleansed*, *integrated*, *transformed*, and *aggregated*, into a read-only database that is optimized for decision-making.

A Typical DW Architecture

- A typical DW architecture consists of five layers—Data Source Systems, ETL Management Services, DW Storage and Metadata Repository, Data Marts and OLAP Engines, and Front-end Tools.



Traditional Business Intelligence Architecture

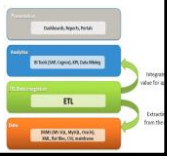


Understanding Data Warehousing

- Take all the data in your enterprise
- Integrate all the data; if necessary, include relevant data from outside
- Select, clean, and transform them
- Store the data in formats suitable for easy access and analysis for decision-making
- Analyze data for strategic decision making and business intelligence
- A decision support database is maintained separately from operational databases
- Need special data organization, access methods, implementation methods, and analysis methods

Data Warehouse Characteristics

- Essential part of infrastructure for business intelligence
- Logically centralized repository for decision making
  - Populated from operational databases and external data sources
  - Stores historical data
  - Integrated and transformed data
  - Optimized for reporting and periodic integration



Data Comparison

Characteristic	Operational Database	Data Warehouse
Currency	Current	Historical
Details level	Individual	Individual and summary
Orientation	Process	Subject
Records per request	Few	Thousands
Normalization level	Mostly normalized	Normalization relaxed
Update level	Highly volatile	Mostly refreshed (non volatile)
Data model	Relational	Relational (star schemas) and multidimensional (data cubes)

Relationship between DW/BI & Big Data

- Relational databases were dominant as OLTP systems for the last 20 years
- In mid-90's, a data warehousing became important to improve business operations and strategic decision support
  - OLAP emerged
- In early 2000, BI tops many CIO s agenda--builds on data warehousing systems to analyze data and improve operations and strategic business decisions.
- A DW marks the beginning of big data era.
- RDBs and DWs work well for gigabytes and low terabytes of structured data in batch mode, but not for petabytes with 4V characteristics
- Big Data builds on top of DW and BI

OLTP and OLAP Queries

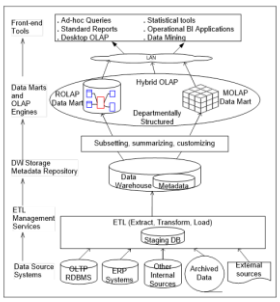
- A query that doesn't require OLAP:
  - How many shoes did we sell last month?
  - Show me the purchase history of client Bill Clinton.
- Queries that require OLAP:
  - How many size 10 shoes in red did we sell last month in the Midwest, the Northeast, the Southeast, compared with the same month last year, actual vs. budget?
  - What are the top 25 brands, by products, styles, and regions, for this period for total US based on sales dollars?
  - How much promotional expense did we spend on customers who purchased less than $100 worth of products?
  - What are sales trends per year per region per product?
  - How are our profits increasing or decreasing per product and region?
  - How much discount should we offer to boost the sales volume significantly?

OLTP and OLAP Queries cont.

- **Queries that require data mining:**
  - Identify profitable customer groups and predict how to effectively retain these customers?
  - Find the correlation between buying patterns of products of type A and those of type B.
  - Which loan applicants should we deny/approve?
  - How can we predict the risks of mortgage risks?
  - What are 10 best risks?
  - Which patients are significantly responding to our therapy?
  - What are major causes of death due to cancer A at different stages?
  - How can we automatically identify credit card fraud in real-time?

A Typical DW Architecture

- A typical DW architecture consists of five layers—Data Source Systems, ETL Management Services, DW Storage and Metadata Repository, Data Marts and OLAP Engines, and Front-end Tools.



A Typical DW Architecture cont.

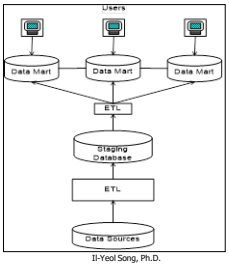
- Data Marts and OLAP Engines
  - A data mart is a small-sized DW that contains a subset of the enterprise DW or a limited volume of aggregated data for the specific analysis needs
  - An enterprise usually ends up having multiple data marts.
  - Since the data to all data marts are fed from the enterprise DW, it is very important to maintain the consistency between a data mart and the DW as well as among data marts themselves.
  - Data marts are usually implemented in one or more OLAP servers.

Major DW Architectures

- Data Mart Bus Architecture with Conformed Dimensions (by Kimball)
- Hub-and-Spoke Architecture (Centralized Architecture by Inmon)
- Federated Data Warehouse Architecture
- Virtual Data Warehouse Architecture

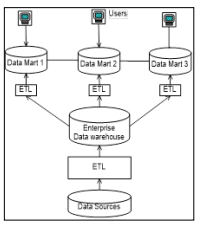
Data Mart Bus Architecture with Conformed Dimensions (Kimball)

- Multiple dimensional data marts are created that are linked with conformed dimensions and measures
- Here, an enterprise DW is a union of all the data marts together with their conformed dimensions.



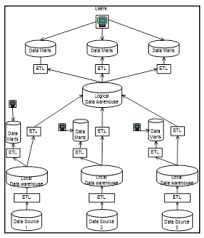
Hub-and-Spoke Architecture (Centralized Architecture, Inmon)

- A single enterprise DW, called the hub, is created with a set of dimensional data marts, called spokes, that are dependent on the enterprise DW.
- This architecture is also called the corporate information factory or the Enterprise DW architecture or a Centralized Architecture.



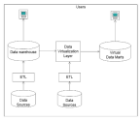
Federated Data Warehouse Architecture

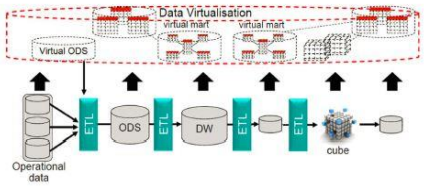
- A federated DW architecture is a variation of a distributed DW architecture, where the global DW serves as a logical DW for all local DWs.
- The logical DW provides users with a single centralized DW image of the enterprise.



Virtual Data Warehouse Architecture

- Use a data virtualization layer
- Can be built with a DW or without a DW





Database Design for Data Warehousing

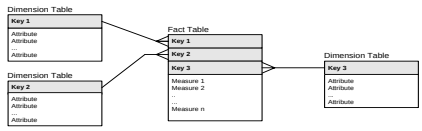
- **Dimensional Model**: A DB schema for a DW or a data mart
- Two types of Dimensional Model
  - Star Schema for **ROLAP** (Relational OLAP)
  - Data cube for **MOLAP** (Multidimensional OLAP)

Dimensional Model

- A database schema for data warehousing
- Initially developed and named by Ralph Kimball to simplify SQL queries
- Consists of a few central fact tables and many dimension tables
- Has relatively few tables and well-defined join paths
- Simplifies end-user query processing and reduce joins
- Provides a multidimensional analysis space within a RDB

Dimensional Model

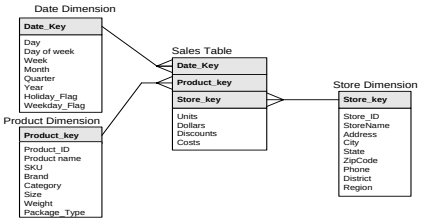
- Data can be represented as a form of *star schema* in Relational OLAP (ROLAP)
  - The center: a fact table
  - The surrounding points: dimensions
  - You are analyzing facts by different dimensions



Structure of Star Schema

Dimensional Model

- Example:



Example of Star Schema (Grocery Store)

Dimensional Model

- Fact table:
  - Stores all transactions or factual data that are to be analyzed
    - Revenue, Actuals, Budgets, Sales, Orders, Bookings, Claims
  - Typically contains numeric ***measures***
    - Balance, Units sold, Cost, Sales, Refund amount, Stock value, Frequency data
  - Stores from millions to more than billion rows
  - Hence, the size of a typical fact table is huge from hundred giga bytes to multi-terabytes

Dimensional Model

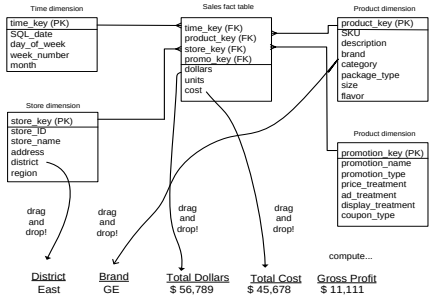
- Dimension table:
  - Dimensions are axes that are used in analyzing fact data
    - Time, Customer, Product, Promotion, Demographics, LifeStyle, Store, Markets.
  - Contains attributes about dimensions
    - Name, Brand name, Description, Location, Order date, Color, Size
  - Supports grouping, browsing, constraining
  - Provides the entry points into the DW

Dimensional Model

- A Rough Comparison among ERD, Relational model, and Star schema

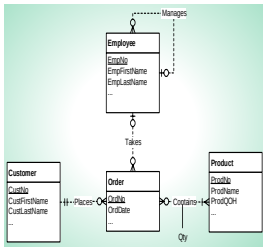
ERD	Relational model	Star schema
Entity	Relation	Dimension
Relationship	Relation or FK	Fact or Dimension

Why Star Schema ?

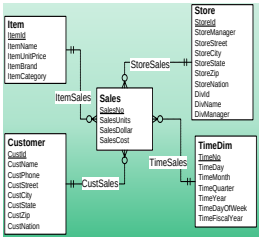


Schema Comparison

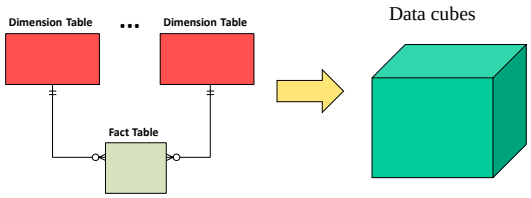
Operational database



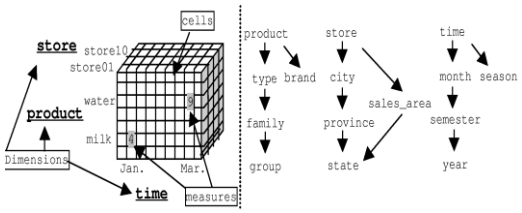
Data warehouse



Multidimensional Data Representations



Data Cube



Data Cube (Multidimensional Cube) and dimensional hierarchies

Example: Olap Usage of an Automobile Marketer

The Story

An automobile marketer wants to improve business activity. Therefore he wants to view sales figures from different perspectives.

The Data Needs

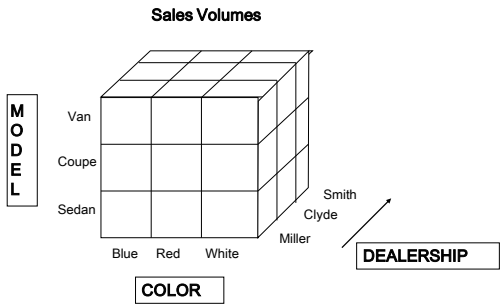
- Sales by model
- Sales by dealership
- Sales by color
- Sales over time
- etc.

A Question

What is the trend in sales volumes over a period of time for a specific model and color across a specific group of dealerships ?

Adopted from Teradata University Network presentation on OLAP.

Example: The Multidimensional View of the Data



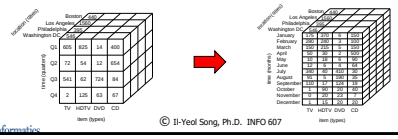
Adopted from Teradata University Network presentation on OLAP.

Typical OLAP Operations

- Drill down
- Roll up
- Slice
- Dice
- Pivot
- Drill-across
- Cube
- Cross-tab

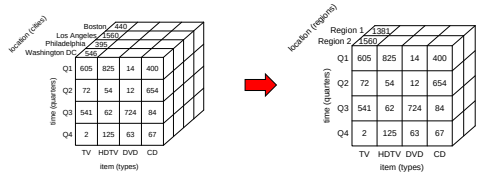
Typical MOLAP Operations

- Drill-down
  - View the data in a more specialized level within a dimension
  - Summarizing at a lower-level of a dimension hierarchy or introducing a new dimension
  - Drilling down is adding new headers from the dimension tables.
  - Example
    - Show me total sales by quarter
    - Show me total sales for each month
    - Show me total sales of each month by department
    - Show me total sales of each month by department and by product type
  - Example: Drill down by time dimension from quarters to months



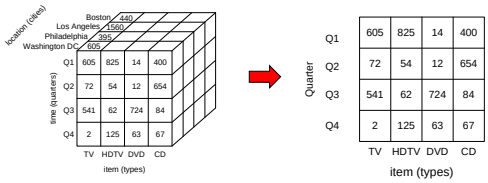
Typical MOLAP Operations

- Roll-up (drill-up)
  - Converse of drill-down
  - Summarizing at a higher level of a dimension hierarchy or by dimension reduction
  - Rolling up is removing row headers.
  - Example: Roll up by location dimension from cities to regions



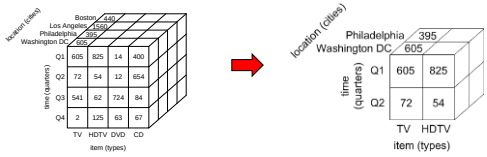
Typical MOLAP Operations

- Slice
  - Perform a selection by setting one or more dimensions to specific values Keep a subset of dimensions for selected values.
  - Reduces the dimensionality of the cube
  - Example of Slice for Location = "Washington DC"



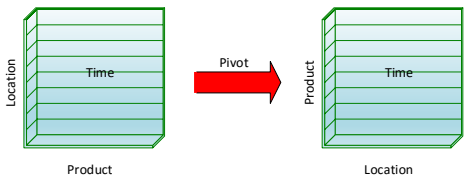
Typical MOLAP Operations

- Dice
  - Perform a selection by replacing a dimension with a subset of values of the dimension, defining a subcube
  - Reduces the number of member values of one or more dimensions
  - Example of Dice
    - Location = ("Philadelphia" or "Washington DC") and
    - Time = ("Q1" or "Q2") and
    - item = ("TV" or "HDTV")



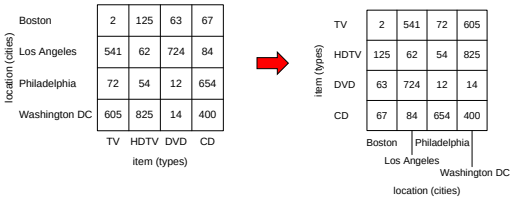
Pivot Operator

- Rotate or rearrange dimensions in a data cube



Typical MOLAP Operations

- Pivot (Rotate)
  - Rotates the data axis to see the data in different perspectives
  - See the data grouped by different dimensions
  - Example: Pivot between location dimension and item dimension



Typical MOLAP Operations

- Drill-across
  - A query that needs to access more than one fact table, linked by common dimensions.
  - Combines cubes that share one or more dimensions
- Drill-through
  - Drill down to the bottom level of a data cube down to its back-end relational tables

Typical MOLAP Operations

- Cross-tab
  - A matrix report with X and Y axis

CROSS TAB

	1998		1999		By City
	HDTV	DVD	HDTV	DVD	
DC	50	160	80	180	470
NY	70	240	90	300	700
BOSTON	60	200	100	400	760
By Product	180	600	270	880	1930

SUM

ROLAP Functions: ROLLUP

- Example: **GROUP BY ROLLUP** (Zip, Month) produces the union of
  - GROUP BY Zip, Month
  - GROUP BY Zip
  - Grand total
- Example of ROLLUP Operator
  - SQL> select d.year, d.month_name month, s.comp_num,
  - 2 sum(s.gross_sales) as tot_sales,
  - 3 sum(s.discount_amt) as tot_discountt,
  - 4 sum(s.tax_amount) as tot_tax,
  - 5 sum(s.gross_sales-s.discount_amt-s.tax_amount) as tot_net_sales
  - 6 from daily_sales s, report_date d
  - 7 where s.transaction_date = d.calendar_date
  - 8 **group by ROLLUP** (d.year, d.month_name, s.comp_num);

```
YEAR  MONTH   COMP_NUM  TOT_SALES  TOT_DISCOUNT  TOT_TAX  TOT_NET_SALES
----  ------  --------  ---------  ------------  -------  -------------
1999  APRIL   100         1250.38        131.29   206.99         912.10
1999  APRIL   200         2877.50        265.50   372.49        2239.51
1999  APRIL               4127.88        396.79   579.48        3151.61
1999                      4127.88        396.79   579.48        3151.61
                          4127.88        396.79   579.48        3151.61
7 rows selected.
```

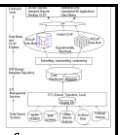
ETL (Extraction, Transformation, and Loading)

- ETL or ETIL (Extraction, Transformation, Integration, and Loading) is a key part of a DW system
- Why ETL is so complex?
  - The number and variety of data sources
  - The complexity of data transformation rules
  - The complexity of data integration
  - The availability of skill sets
  - Challenging to enterprises
    - With no data element standards
    - With no redundancy control
    - That have gone through merge and acquisition

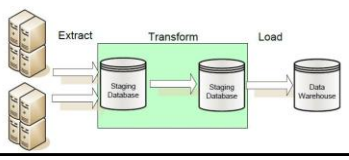
ETL & Data Integration Statistics

- Could take 60% to 80% of a DW implementation effort (Gartner, 2002)
- Tools typically cost $150K to $1 million without annual maintenance
- According to The Data Warehouse Institute, it takes an organization nearly eight weeks to add a new data source to their data warehouses.

Data Staging Area



- An area where ETL is processed
- Clean, transform, combine, de-duplicate, household, archive, and prepare source data for use in the data warehouse.
- Exists between the source system and the DW system
- Allocate additional storage to the staging area



Examples of Data Cleansing (Scrubbing) and Transformation

- **Attributes**
  - **Missing value**
    - Unknown or Not existing?
  - **Inconsistent coding**
    - Ex: gender (M, F) or (m, f), or (male, female), (1, 0)
    - IBM or I.B.M. or I.B.M, International Business Machines
    - "Mr." or "Mister"
  - **Misspellings or short names**
    - Ex: Phila, Philadelphia, *Phladelphia*, *Philadelphiaa*
  - **Misfedded values**
    - Ex: City = 'Pennsylvania"

Parsing Example

Raw input in source file
Aimee Christina Parker, Prod. Mgr.
Microsoft
One Microsoft Way
Redmond, WA

Parsed data in target file	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	One Microsoft Way
City	Redmond
State	WA

Correction Example

Parsed data	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	One Microsoft Way
City	Redmond
State	WA

Corrected data	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	15580 NE 31st St.
City	Redmond
State	WA
Postal Code	98052
Country	USA

Standardization Example

Corrected data	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Prod. Mgr.
Firm	Microsoft
Street	15580 NE 31st St.
City	Redmond
State	Washington
Postal Code	98052
Country	USA

Standardized data	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Product Manager
Firm	Microsoft Corporation
Street	15580 NE 31st Street
City	Redmond
State	WA
Postal Code	98052
Country	USA

Matching Example

Source 1	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Product Manager
Firm	Microsoft Corporation
Street	15580 NE 31st Street
City	Redmond
State	WA
Postal Code	98052
Country	USA

Source 2	
First name	Aimee
Middle name	C.
Last name	Parker-Lewis
Job title	Prod. Mgr.
Firm	Microsoft
Street	16517 78th Place NE
City	Bothell
State	WA
Postal Code	98020
Country	USA

Entity Matching Applications

Marketing

Law enforcement

Fraud detection

Health care

Data Mining

- Discover significant, implicit patterns
  - Target promotions
  - Change mix and collocation of items
- Requires large volumes of transaction data including sensor data and social media interactions
- Important tools for business intelligence

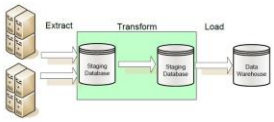
Market Shares and Trends

- Major vendors: Teradata, Oracle, IBM, Microsoft, SAP
- Large projected market growth
- Trends
  - Real time load and analysis
    - From ETL to ELT
  - Increased storage and analysis of social interactions
  - Increased usage of cloud services and appliances
    - Amazon Redshift Cloud Service

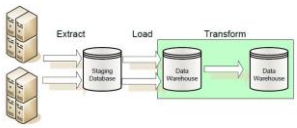
From ETL to ELT

Transformations take place in the target database after data loading.

- ETL

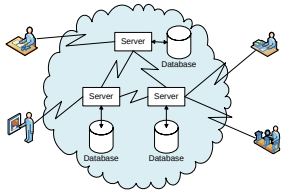


- ELT



Source: http://www.dataacademy.com/files/ETL-vs-ELT-White-Paper.pdf

Cloud Influence



- Reduces local expertise to procure technology and manage a data warehouse
- Economies of scale
- Improved scalability
- Higher variable costs but lower fixed costs

Salary Trends (USA)

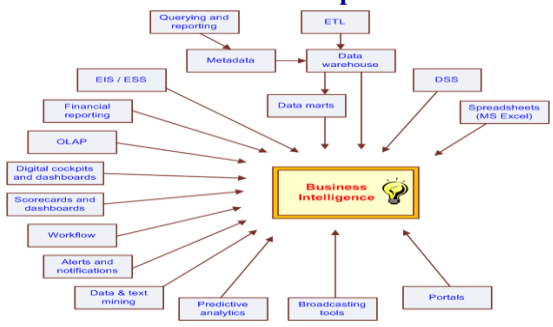
Job Title	2013	2014	% Change
DB manager	$101,750 – $140,750	$107,750 – $149,000	5.9%
DB developer	$80,500 – $128,250	$92,000 – $134,500	5.5%
Data analyst	$64,250 – $96,000	$67,750 – $101,000	5.3%
DW manager	$108,750 – $145,750	$115,250 – $154,250	5.9%
DW analyst	$93,500 – $126,500	$99,000 – $133,750	5.8%
BI analyst	$94,250 – $132,500	$101,250 – $142,250	7.4%

Salary trends from Robert Half Salary Survey
http://www.roberthalf.com/technology/it-salary-center?lobId=roberthalftechnology

Business Intelligence

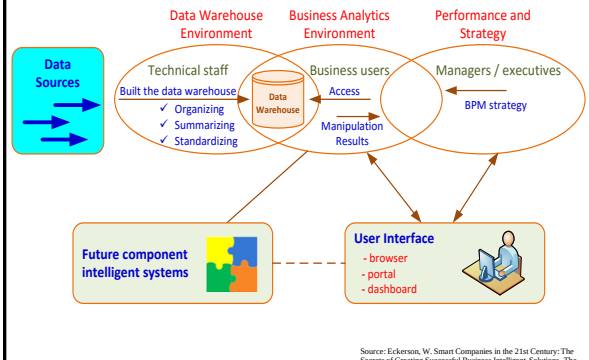
- *Business Intelligence* (BI) is an umbrella term that combines architectures, databases, analytical tools, applications, and methodologies.
- The Data Warehousing Institute (TDWI 2002) working definition of business intelligence:
  - "The processes, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action. Business intelligence encompasses data warehousing, business analytic tools and content/knowledge management"

Evolution of BI Capabilities



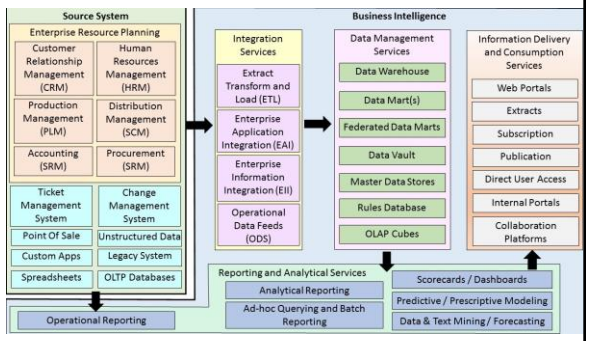
From SHARDA, RAMESH; DELEN, DURSUN; TURBAN, EFRAIM. BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, 10th Edition, © 2015. Used by permission of Pearson Education, Inc., New York, NY. All Rights Reserved.

A High-Level BI Architecture



Source: Eckerson, W. Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions. The Data Warehousing Institute, Seattle, WA, 2003, p. 32

Detailed level BI Architecture



Based on: Pant, P. Business intelligence (BI): How to build successful BI Strategy, Deloitte Consulting LLP. 2009.

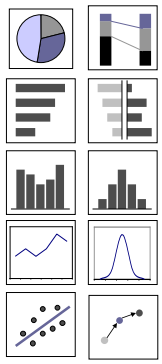
Business Reports Building Blocks

Attributes
Descriptive information providing business context and defining summarization levels for calculations

Metrics
Quantitative business measures

Category	Subcategory	Metrics	Units Sold	Profit	Profit Margin	Revenue
Books	Art & Architecture		37	$180	24.97%	$720
	Business		37	$137	24.54%	$558
	Literature		53	$87	20.61%	$423
	Books - Miscellaneous		54	$84	19.06%	$439
	Science & Technology		43	$302	24.70%	$1,222
	Sports & Health		35	$107	24.13%	$444
Electronics	Audio Equipment		21	$1,349	19.40%	$6,950
	Cameras		14	$1,337	21.32%	$6,270
	Computers		28	$560	18.89%	$2,967
	Electronics - Miscellaneous		14	$1,205	20.58%	$5,854
	TV's		31	$1,430	19.90%	$7,184
	Video Equipment		18	$1,773	20.27%	$8,750
Movies	Action		87	$101	8.99%	$1,124
	Comedy		88	$87	7.69%	$1,133
	Drama		74	$104	9.05%	$1,146
	Horror		92	$118	9.94%	$1,192
	Kids / Family		73	$100	8.93%	$1,122
	Special Interests		44	$84	9.92%	$843

In Most Cases, One of Five Basic Chart Types Provides the Most Effective Data Presentation



Pie Chart	5%	Recommended usage frequency
Bar Chart	25%	Recommended usage frequency
Column Chart	50%	Recommended usage frequency (combined)
Line Chart		
Dot Chart	10%	Recommended usage frequency

Source: Adapted from Gene Zelazny (2001). *Say It With Charts*

The Three Layers of Information in Dashboards

Monitoring. Graphical, abstracted data to monitor key performance metrics.
Analysis. Summarized dimensional data to analyze the root cause of problems and ability to drill down to lower grain data
Management. Detailed operational data that identify what actions to take to resolve a problem.

Review

- What is the difference between OLTP vs OLAP?
- What are important characteristics of a data warehouse?
- What are major differences between OLTP and DW?
- What does the ETL stand for?
- What is the difference between data warehouse and data mart?
- What is the relationship between DW and BI?
- Explain drill-down, rollup, slice, dice, and pivot.
- What is the dimensional model?
- What is the star schema? What are two components of the star schema?
- Why ETL takes so much time?
- What is dashboard used in BI?
- Why Big Data takes ELT, rather than ETL
- Do you agree with the following statement "Big Data builds on top of DW and BI"?