Document History

Date	Course Version No.	Software Version No.	Developer / SME	Reviewer(s)	Approver	Change Record Remarks
	0.1D	NA				Content Creation
Jan-2009	0.1	NA	BI CDI team			Review
16-Apr-2009	2.0	NA	Priya Rane			Material Revamp
04-Feb-2010		NA	CLS Team			Review
31-July-2012		NA	Coordinators			Change to IGATE Format
June 2016	2.1	NA	Swati Rao	Rajita Dhumal	Mahima Sharma	Material Revamp as per Integrated ToC for I & D LoT

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2

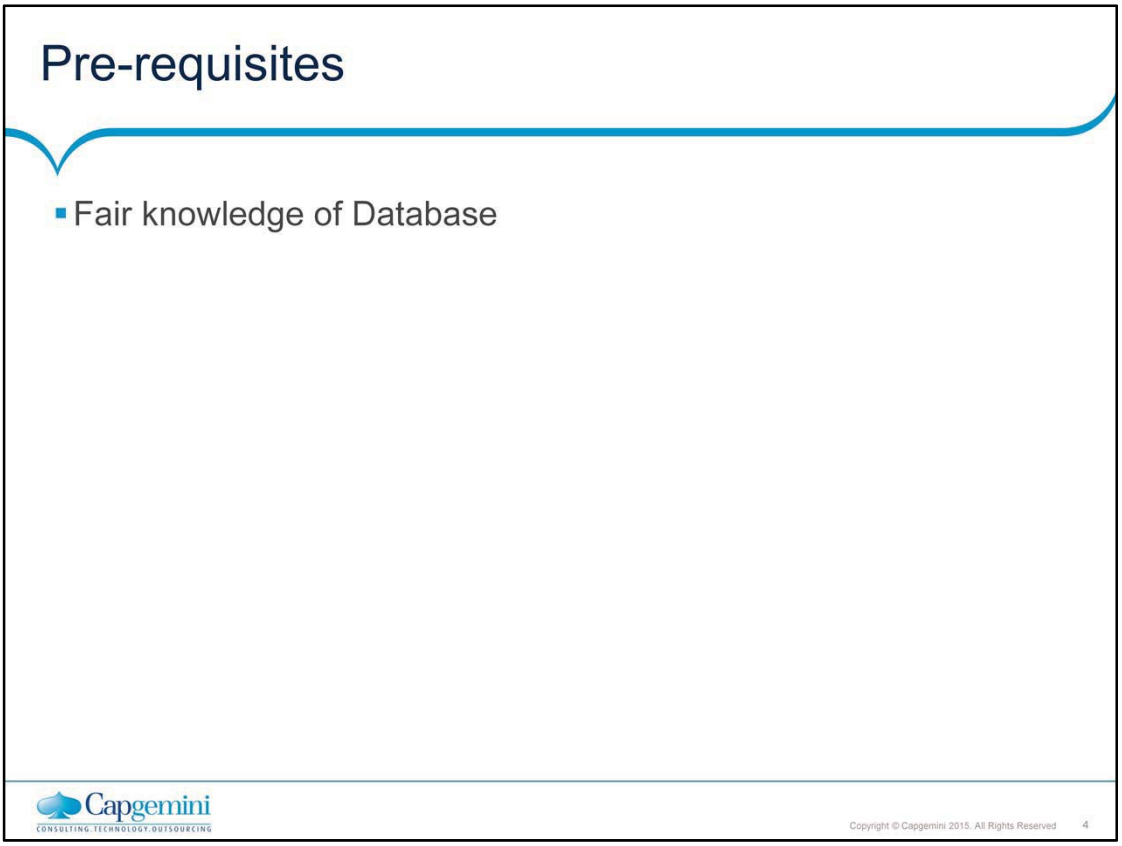
Course Goals and Non Goals

- Course Goals
  - At the end of this program, participants gain an understanding of basic concepts in Data warehousing.

- Course Non Goals
  - Implementation of dimensional modeling is not the part of this course.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 3


Pre-requisites
Fair knowledge of Database
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
4

Intended Audience

- Software Engineers and Senior Software Engineers

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 5

Day Wise Schedule

- Day 1
  - Lesson 1: Business Intelligence
  - Lesson 2: General concept of Data Warehouse
  - Lesson 3: Dimensional modeling
  - Lesson 4: ETL and Metadata
  - Lesson 5: Online Analytical Processing (OLAP)
  - Lesson 6: Data Mining
  - Lesson 7:Best Practices for Building Data Warehouse
  - Lesson 8: Case Studies

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 6

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 7

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 8

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 9

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Table of Contents

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 11

References

- Student material:
  - Class Book (presentation slides with notes)
- Book:
  - The Data Warehousing Toolkit – Ralph Kimball
  - Introduction to Database Systems – C.J. Date
  - Advanced Data Warehouse – IBM
- Web-site:
  - http://www.datawarehouse.org
  - http://etl-tools.info/

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Next Step Courses (if applicable)

- BI related tool training

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13

Other Parallel Technology Areas

- NA

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 14

Data Warehousing Concepts

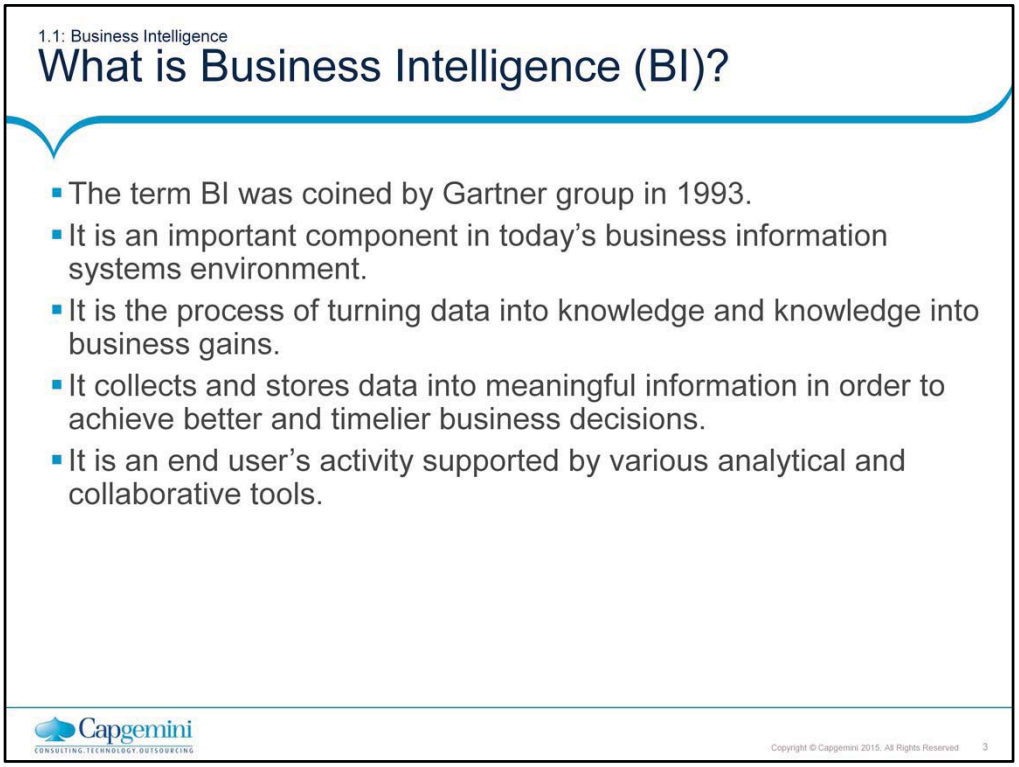
Lesson 1: Business Intelligence

Lesson Objectives

- In this lesson, you will learn:
  - What is Business Intelligence?
  - Need of Business Intelligence
  - Terms used in Business Intelligence
  - Components of Business Intelligence

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2



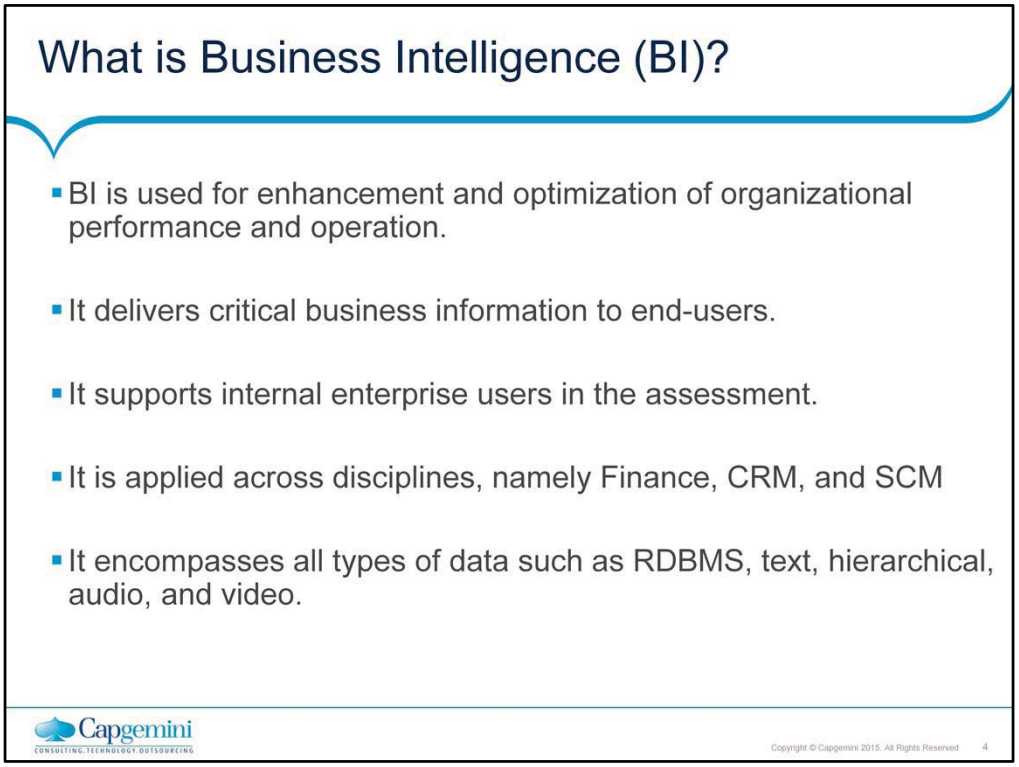
Business Intelligence:

- **Business Intelligence (BI)** is the process of getting useful information from data.

> BI is an important component in today's business information systems environment.

- As the business environment has become increasingly competitive, the need to use corporate data as a strategic resource has intensified. However, most of the organizations in technology based businesses are **data rich** and are **information poor**. Much of the essential information that is needed to anticipate changing market conditions and customer preferences is locked in various transactional systems, spread sheets, and log files. So without the ability to deliver the right information to the right people at the right time, companies cannot stay competitive in this fast changing economy. So the **BI value proposition** is a term for the ability to navigate complex sales channels by maximizing knowledge about the customer base and developing strategies that leverage that knowledge from decision to action.

- **BI applications** are decision support tools that enable real-time, interactive access, analysis, and manipulation of mission-critical corporate information.



Business Intelligence:

- Business Intelligence gives answers to the questions such as given below:
  - Who are my top ten customers?
  - How effective was my last sales campaign?
  - Who is my best sales person by volume, and by dollar revenue, per region, during the last week of each month? How does that compare with last year?
  - How much more intelligent can you make your business processes?
  - How much more insight can you gain into your business?
  - How much more integrated can your business processes be?
  - How much more interactive can your business be with customers, partners, employees and managers?
  - BI solutions answer all these questions.

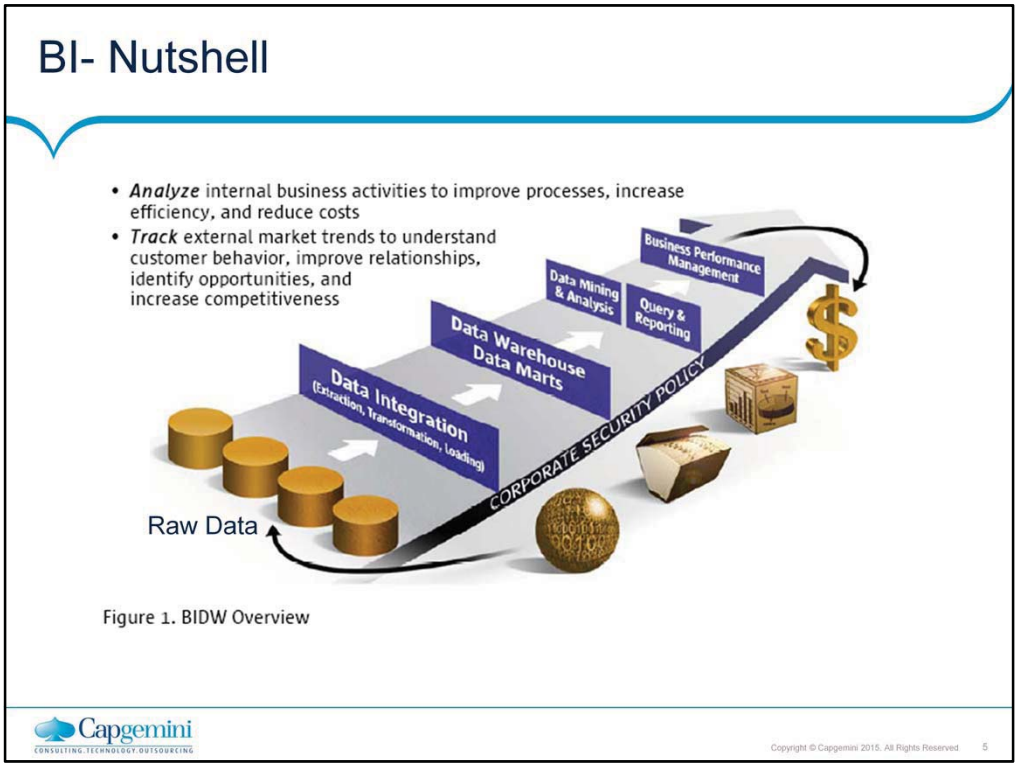


Figure 1 illustrates the major components of a BI system and the process of generating business results from raw data (the operational data that is used to run the business).

The BIDW process can be broken down into the following steps:
• Raw data is stored: Raw data is typically stored, retrieved, and updated by an organization's on-line transaction
processing (OLTP) system. Additional data that feeds into the data warehouse may include
external and legacy data that is useful to analyze the business.
• Information is cleansed and optimized: The information is then cleansed (for example, all duplicate items are removed) and optimized for decision support applications (i.e. structured for queries and analysis vs. structured for transactions). It is usually "read only" (meaning no updates allowed) and stored on separate systems to lessen the impact on the operational systems.

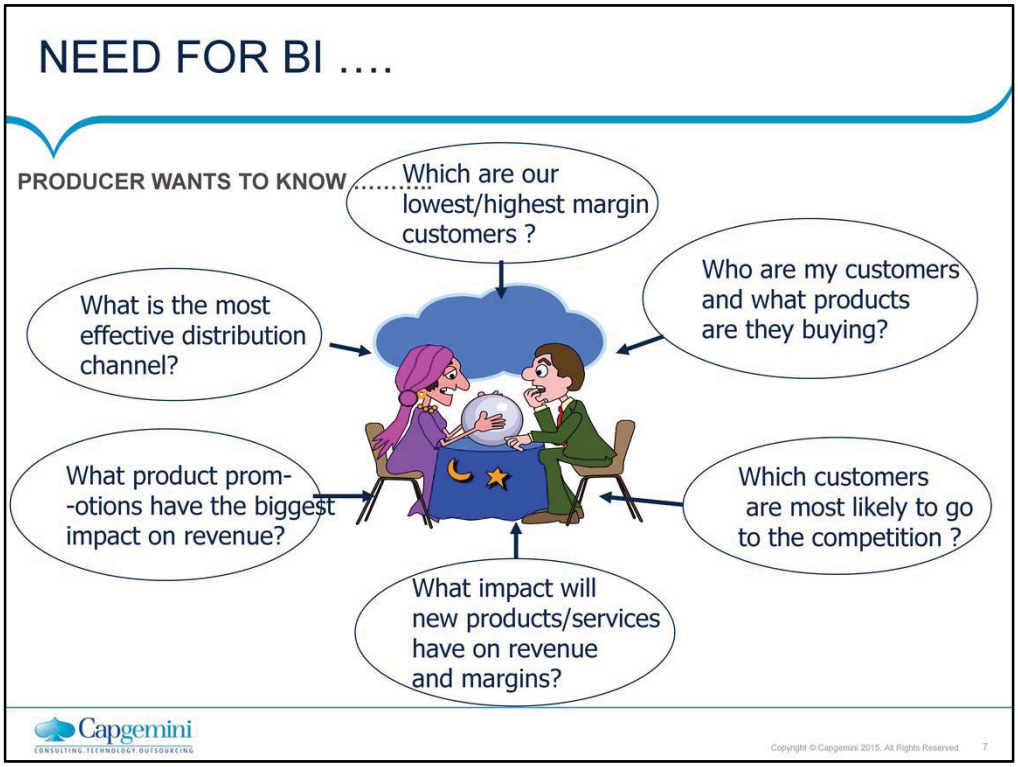
• Data mining, query and analytical tools generate intelligence: Various data mining, query and analytical tools generate the intelligence that enables companies to spot trends, enhance business relationships, and create new opportunities.

• Organizations use intelligence to make strategic business decisions: With this intelligence, organizations can make effective decisions, and create strategies and programs for competitive advantage.

• The system is regulated by an overall corporate security policy o Information in a data warehouse is typically confidential and critical to a company's business operations. Consequently, access to all functions and contents of a data
warehouse environment must be secure from both external as well as internal threats and should be regulated by an overall, corporate security policy.

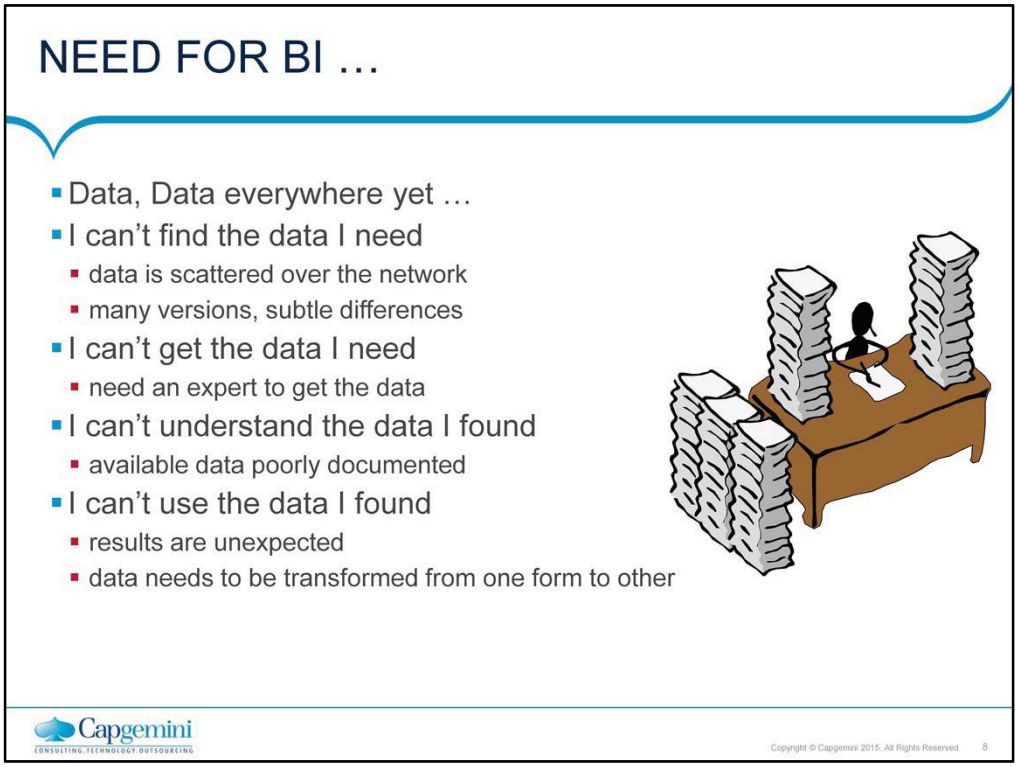
• Business performance management applications track results: A well-run BIDW operation also includes Business Performance Management (BPM) applications, which help track the results of the decisions made and the performance of the programs created.

integration.intelligence.insight



If u see some of the top management questions such as
what is the most effective distribution chennel? Or
who are our customers and what products are they buying.

These questions will help top management to take long range decision. Hence, regular operational system does not answer above questions because of regular database consists of only current information. So if at all we need to answer above question we need to have huge amount of historical and detailed data in the database which is integrated from several sources.



In most of the organization either it is big or small some of the problems are common such as;

May not be able to find required data since the data which is maintained are geographically scattered and whatever they are available it is in different version or format.
eg: The data might be maintained in excell sheet. Or The data might be maintained using any database such as oracle, db-2 or sql etc or It might be in just word documents.

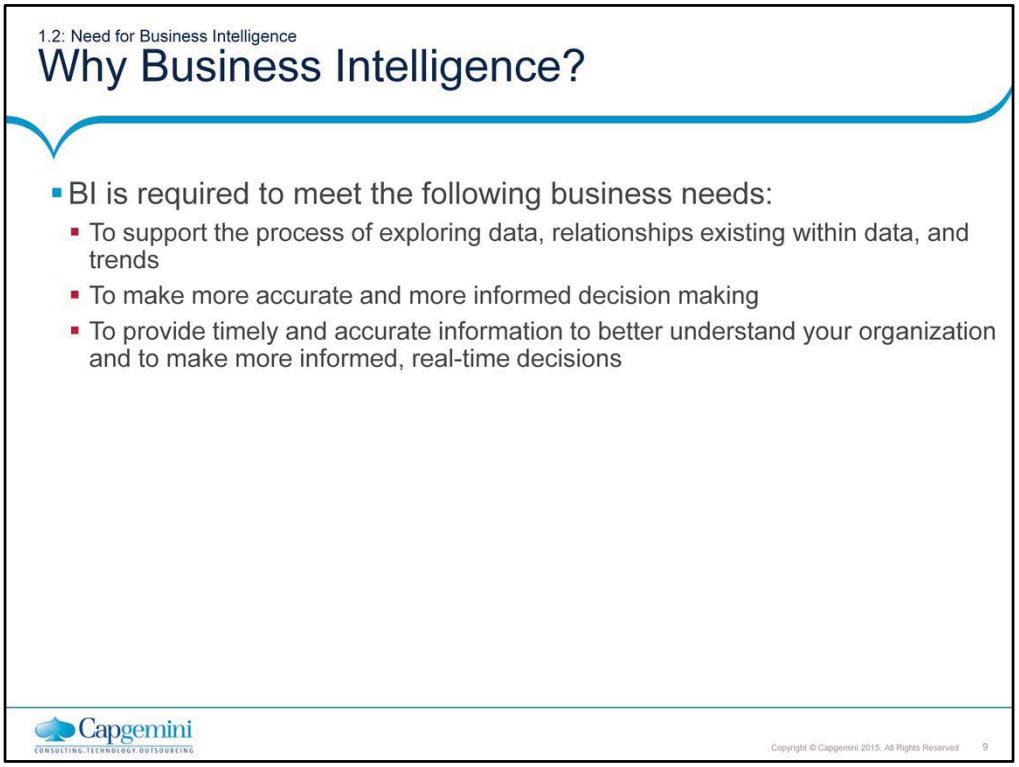
2. May not be able to analyze the data due to lack of expertise in the organization

3. May not be able to fetch properly because of data is maintained poorly (Might be maintained in unstructured way eg. using note pad or word document)

4. The result of all problems leads unexpected results hence these unstructured data, maintained in different formats need to transform into single format so that it will help top management to take strategic decision.

In order to resolve above problems the only solution is DataWarehouse. The Data warehouse is called single version of truth in which different sources of data is captured and stored in a single place.

For this, The Data Warehouse Motivation is
Huge amounts of data need to be summarized in various forms to enable data creators and data users to get quick overviews and dig into details as needed with high performance and flexibility



Need for Business Intelligence:

- **BI exhibits the following utility features:**
  - BI is a general term for applications, platforms, tools and technologies that support the process of exploring data, relationships existing within data, and trends.
  - BI is important in helping organizations to stay ahead of the competition by providing the means for quicker, more accurate, and more informed decision making.
  - BI provides timely and accurate information to better understand the organization and to make more informed, real-time decisions.

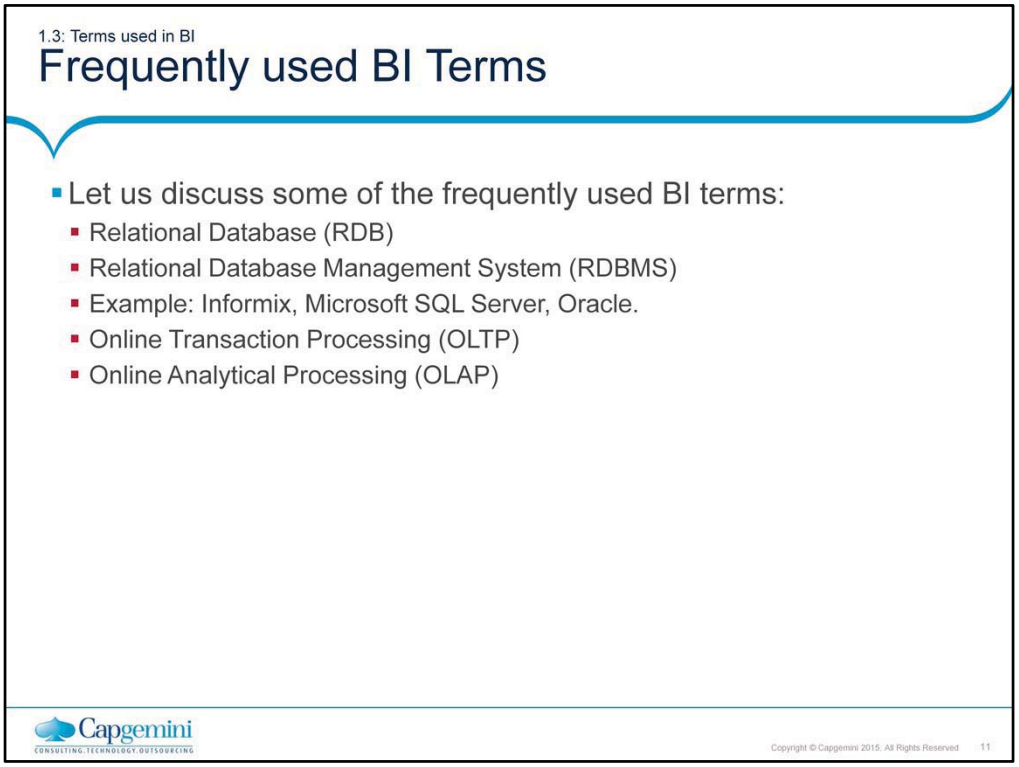
- **But why do you need Business Intelligence?**
  - For many years, database vendors have focused on getting data into a database. The emphasis has led to great achievements in **online transaction processing** and capacity. Many companies have accumulated data that can be measured in gigabytes, terabytes, and even petabytes.
  - **Transactional data,** which is the data that is used to run the business, is good for keeping track of what is happening in an organization. However, it is not well suited to finding out why things are happening or predicting future performance.
  - Hence there arises a strong need for BI applications.

Why Business Intelligence?

- Data Analysis is a huge and crucial part of Business Intelligence.
- Many organizations need to know the overall performance and the way its business is functioning.
- BI is used to gather past as well as present data.
- Modern BI systems are capable of managing large amount of unstructured data.

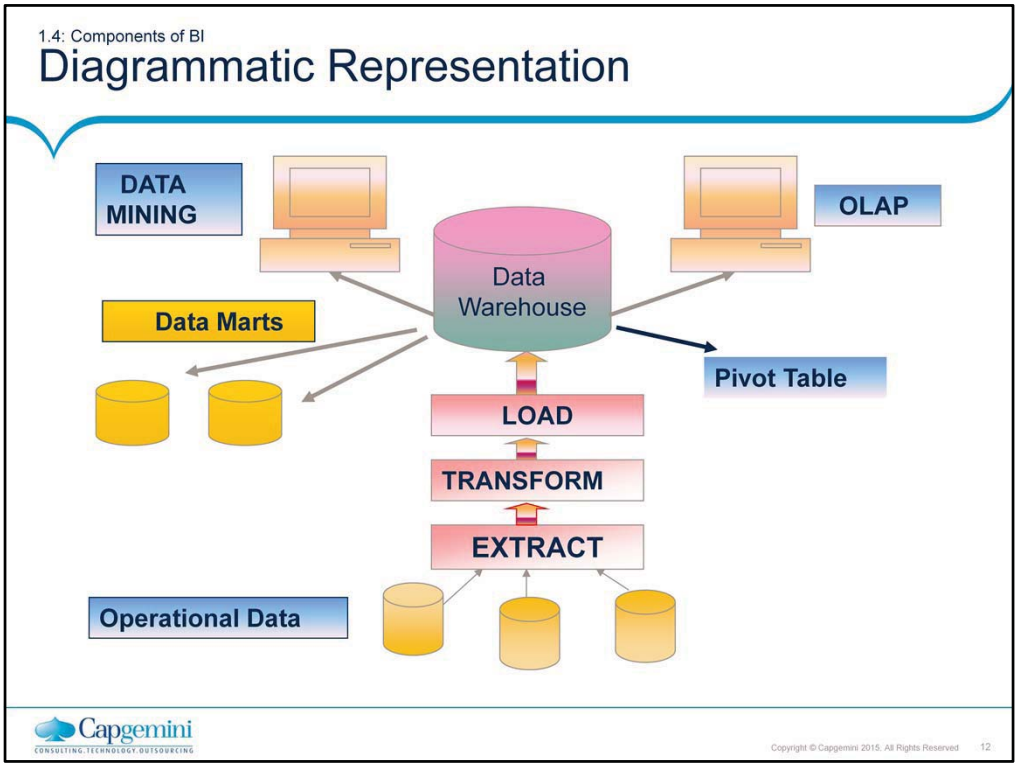
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 10



Terms used in BI:

- **Relational Database (RDB):**
  - It is a database that conforms to the relational model.
- **Relational Database Management System (RDBMS):**
  - It refers to the software used to create a RDB.
  - Example: Informix, Microsoft SQL Server, Oracle
- **Online Transaction Processing (OLTP):**
  - OLTP is a process which is used for day to day transaction processing.
  - **Example:** Operational systems, High volume data collection
- **Online Analytical Processing (OLAP):**
  - This processing method provides fast access to shared multidimensional data.
  - It is used to generically refer to software and applications that provide users with the ability to store and access data multi-dimensionally.



Components of BI:

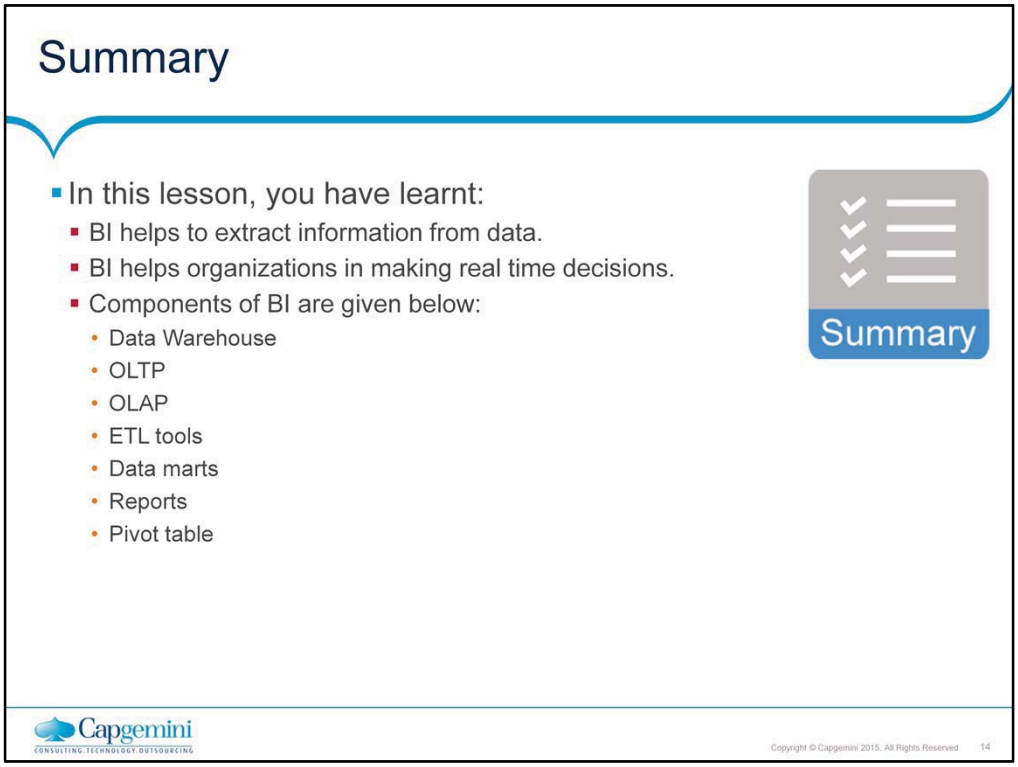
Following are the various components of BI:

1. **Operational Data:** Typically data is sourced from transaction processing systems. It is also called as Data Source. Typically data is sourced from transaction processing systems (Manufacturing, ERP, Sales). Example: Customer, Inventory, Credit, Sales, Operation and External are the data source.
2. **ETL Tools:**
   - **Extract:** It is the process of pulling the data from external and operational data sources in order to source data for the data warehouse.
   - Transform: It is the process that converts data to the format required by data warehouse. It cleanses data to ensure accuracy. It validates primary keys against defined owner. It converts to different numbering schema.
   - Load: It is the process that loads data to data warehouse. It follows guidelines as outlined by the data warehouse.
3. **DWH:** Data Warehouse integrates and aggregates data from various operational and external database maintained by different Business Units.
4. **Data Mart:** Data mart is a repository of data collection from operational data source and other sources that are designed to serve a particular community of knowledge workers.
5. **Reports:** A report presents the data in a format understandable by the end user.

Components of BI:

Following are the various components of BI (contd.):

6. **OLAP:** OLAP is a category of software technology that enables the users to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information.
7. **Pivot Table:** A pivot table is the simplest tool to aggregate data by creating a dimension for each field and grouping the same values in a field. A pivot table is a data summarization tool found in data visualization programs such as spreadsheets. It allows you to reorganize and summarize selected columns and rows of data in a spreadsheet or database.


Summary
In this lesson, you have learnt:
BI helps to extract information from data.
BI helps organizations in making real time decisions.
Components of BI are given below:
Data Warehouse
OLTP
OLAP
ETL tools
Data marts
Reports
Pivot table
Summary
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
14

Review Questions

- Question 1: This a huge and crucial part of Business Intelligence.
  - Option 1: Data collection
  - Option 2: Data analysis
  - Option 3: Data availability

- Question 2: OLAP Analysis is not the part of BI presentation.
  - True / False
- Question 3: ___ operation converts data to format required by data warehouse.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 15

Review Question: Match the Following

1. pulling the data from external and operational data sources	A. OLAP Analysis
2. part of BI presentation	B. RDB
3. Software for relational database	C. Extract
	D. OLTP
	E. RDBMS

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 16

Data Warehousing Concepts

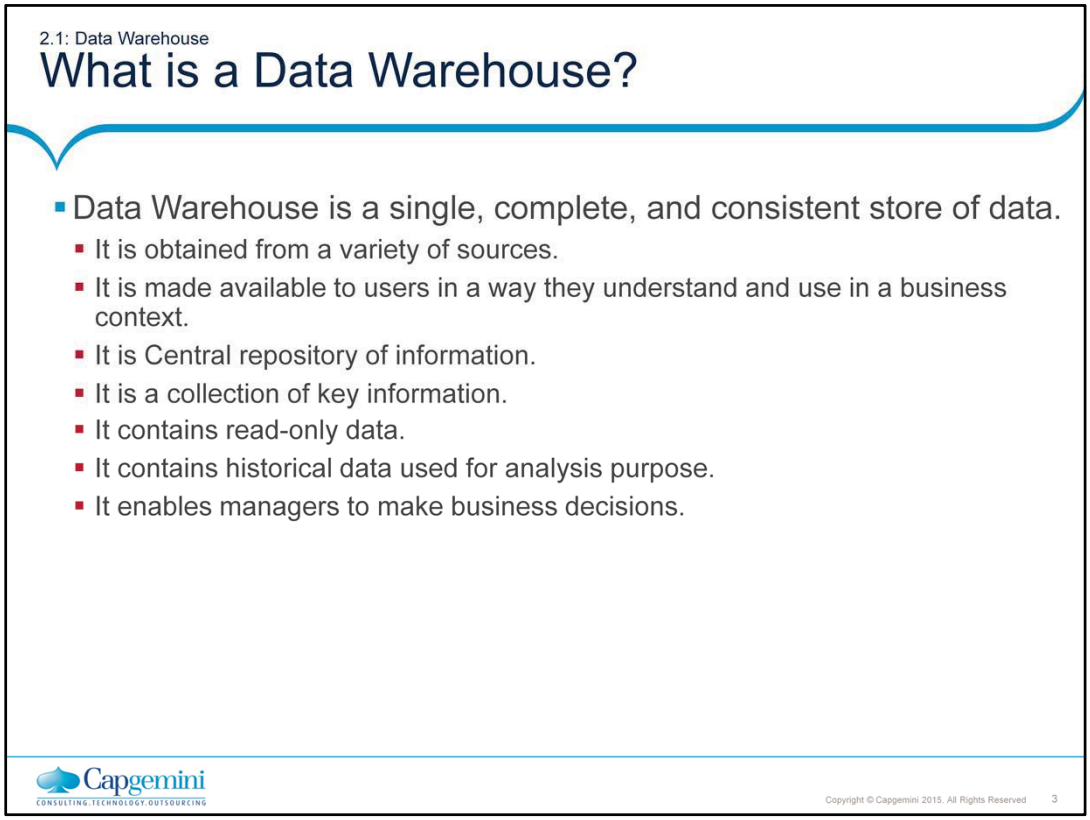
Lesson 2: General Concept of Data Warehouse

Lesson Objectives

- In this lesson, you will learn:
  - What is a Data Warehouse?
  - History of Data Warehouse
  - Need Of Data Warehouse
  - Data Warehouse Architecture
  - Data Warehouse Components
  - Features of Data warehouse
  - Data Mart
  - Application areas

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

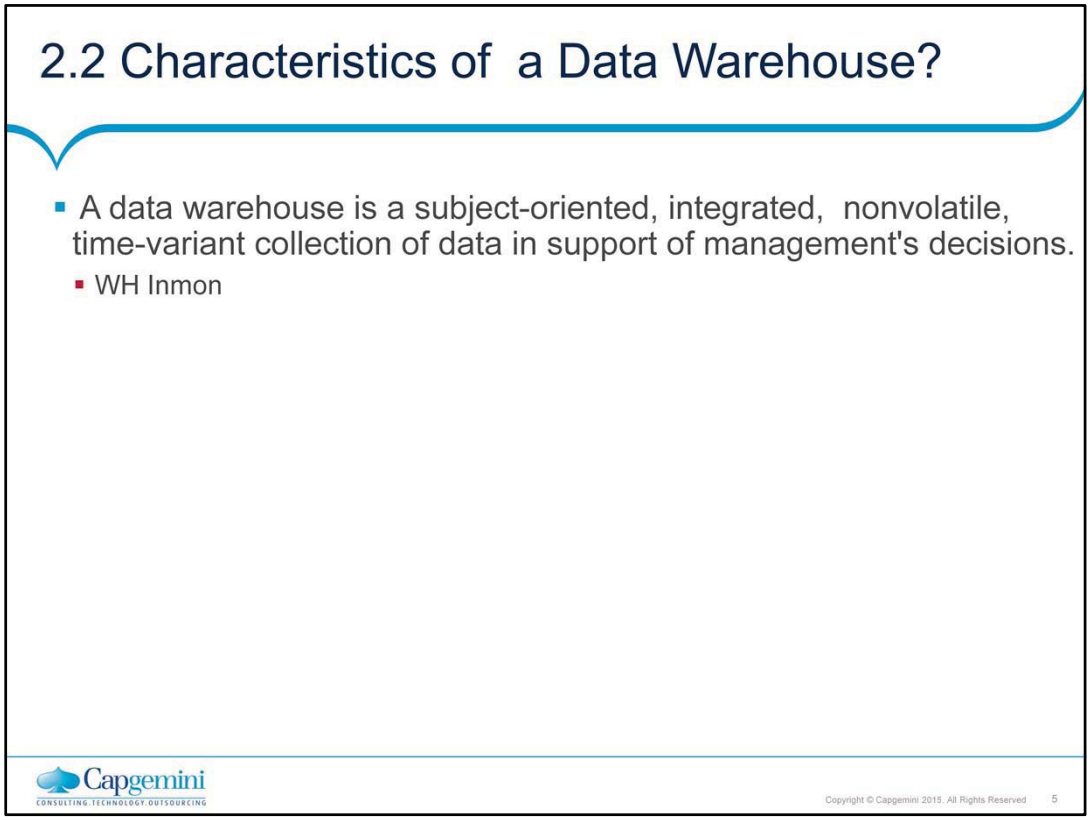
 2



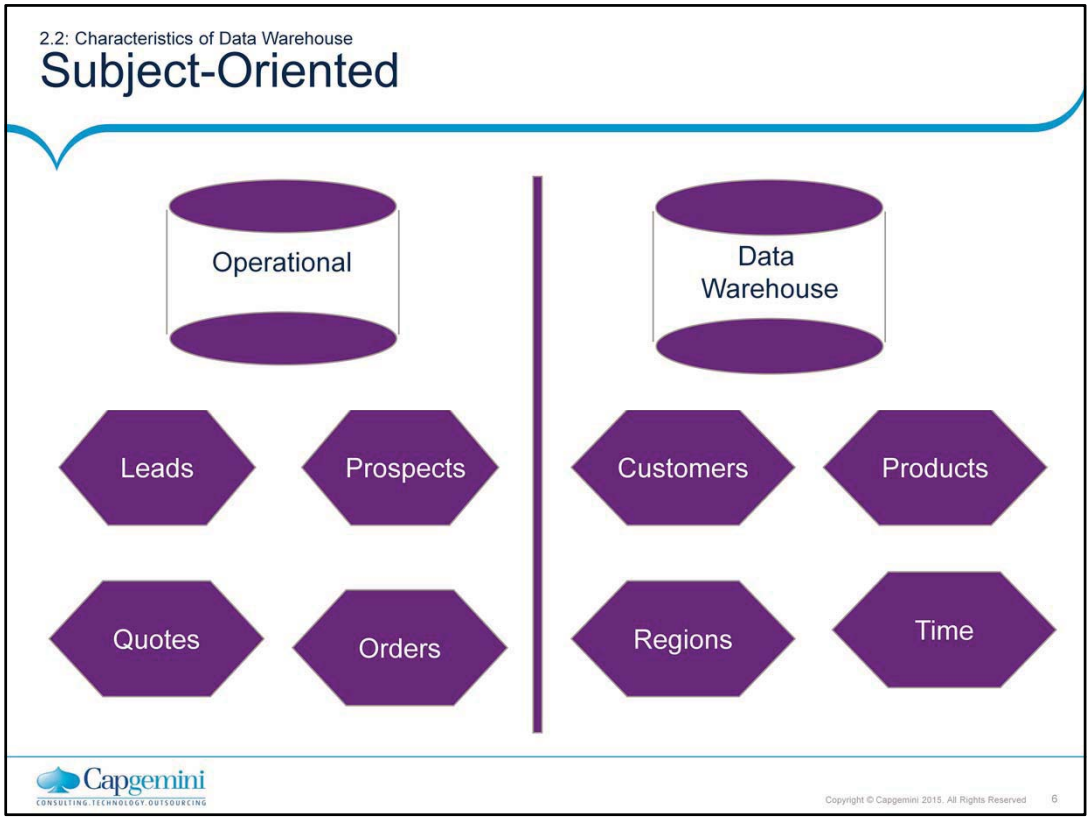
Data Warehouse:
A Data Warehouse is collection key of pieces of information to manage and direct the business for profitability.
A Data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way that they can understand and use it in a business context. It is nothing but a database or a data store. It is a database, so data has to be structured. The data is logically and physically transformed from multiple source applications to align with the business structure. It requires more historical data than that is generally maintained in operational database.
Data is non- changing. It does not get updated. Data is never erased, so it is called non-volatile. Data Warehouse is designed for the analysis of non-volatile data.
Data Warehouse integrates and aggregates data from various operational and external databases maintained by different Business Units.
It is a process that managers use to load the warehouse query that makes information available. It enables people to make informed decisions. It is maintained for a long time period.
A Data Warehouse is a central repository of information with appropriate tools.

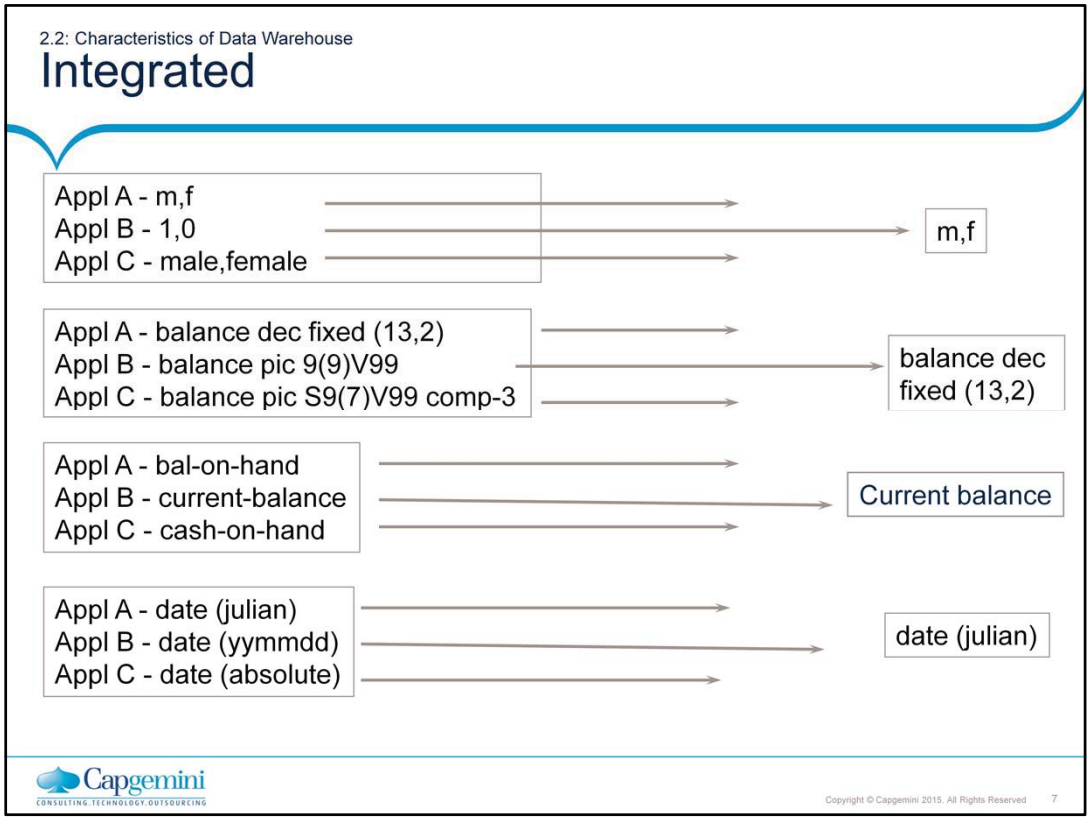
Data Warehouse (contd.):

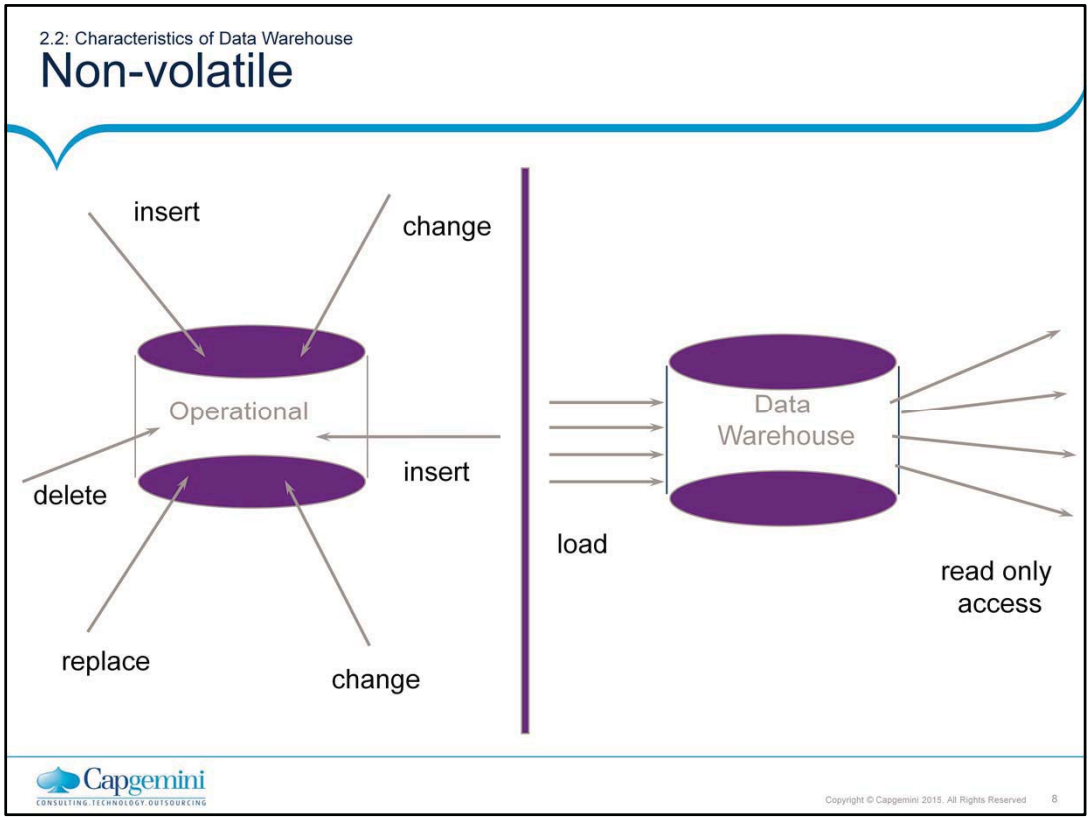
- A Data Warehouse can also be defined as a **structured**, **extensible environment** designed for the analysis of non-volatile data, which is logically and physically transformed from multiple source applications to align with business structure, updated and maintained for a long time period, expressed in simple business terms.
- A Data Warehouse is used by different people in different fields. Companies use Data Warehouses to store information for marketing, sales, and manufacturing to help managers to get the feel of the data and run the business more effectively.
- A **database application** is a piece of software, which provides a user interface for users to add, delete, query, and update data, updates is called an on-line transaction processing (OLTP) application. An application that issues queries to the **read-only database** is called a **Decision Support System (DSS)**.

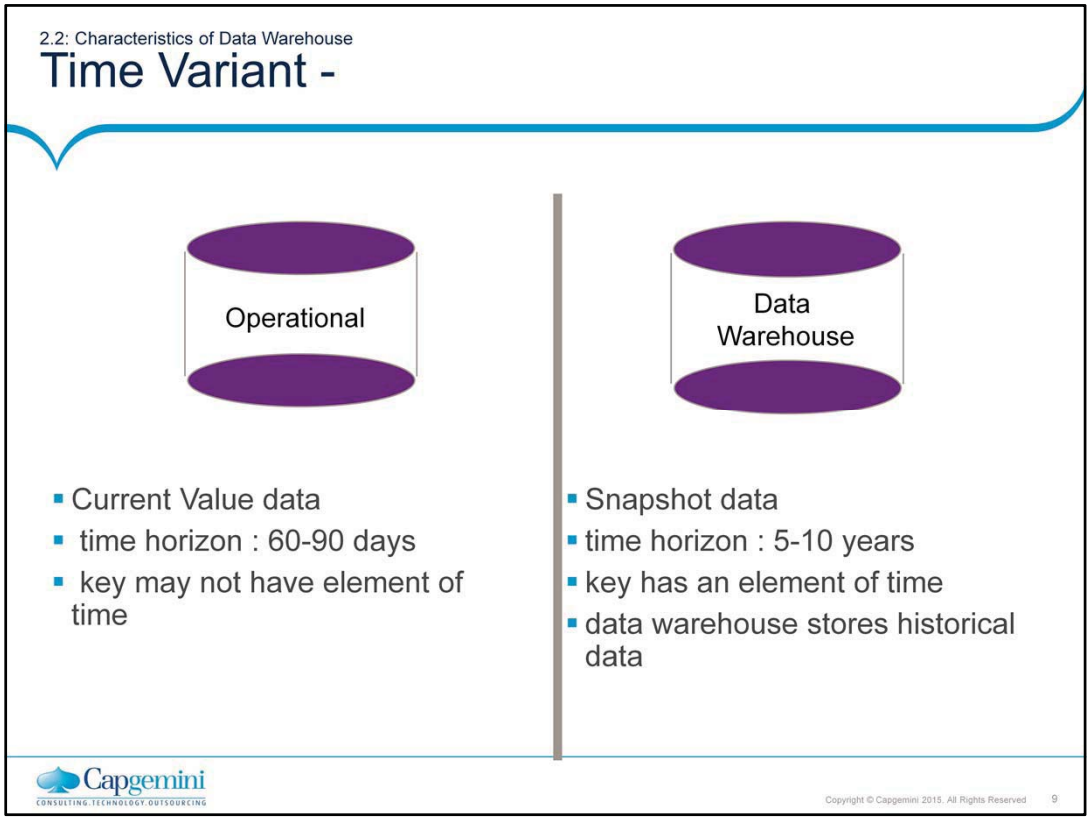


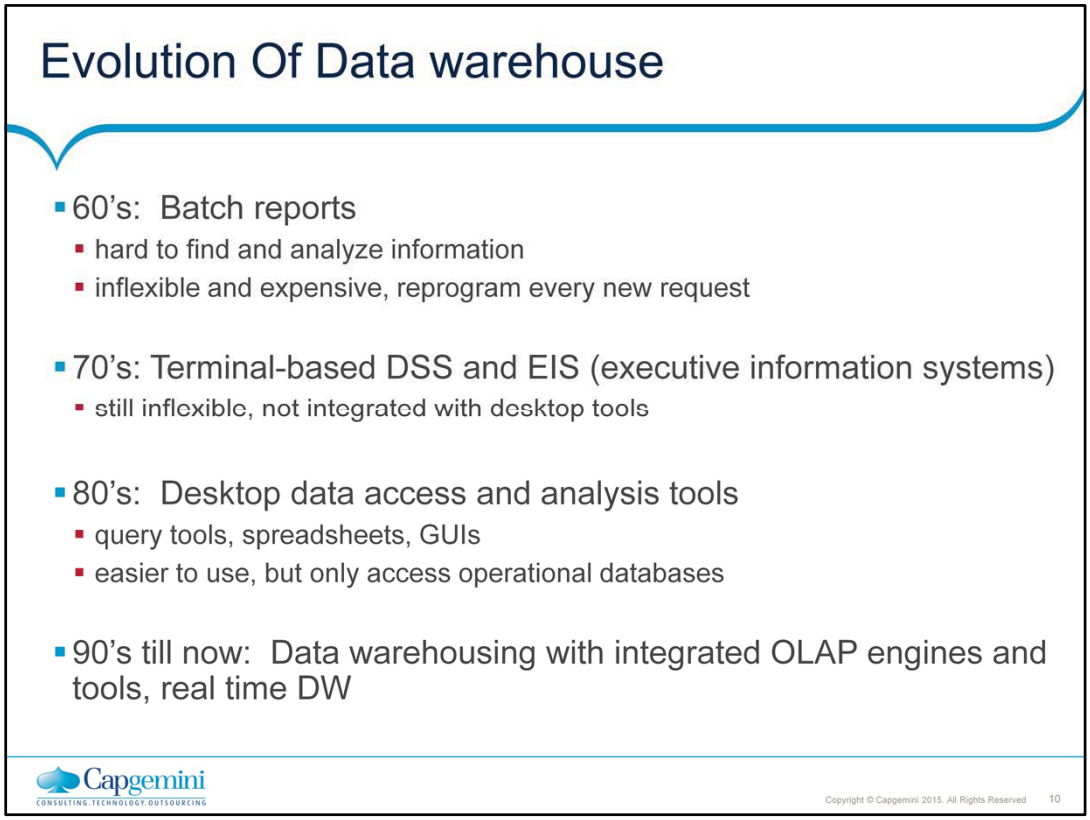
Historical : The data is continuously collected from sources and loaded in the warehouse. The previously loaded data is not deleted for long period of time. This results in building historical data in the warehouse.
Subject Oriented: we mean data grouped into a particular business area instead of the business as a whole.
Integrated: It means, collecting and merging data from various sources. These sources could be disparate in nature.
Time-variant: It means that all data in the data warehouse is identified with a particular time period.
Non-volatile: It means, data that is loaded in the warehouse is based on business transactions in the past, hence it is not expected to change over time


2.2: Characteristics of Data Warehouse
Subject-Oriented
Operational
Data Warehouse
Leads
Prospects
Customers
Products
Quotes
Orders
Regions
Time
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
6


2.2: Characteristics of Data Warehouse
Integrated
Appl A - m,f
Appl B - 1,0
Appl C - male,female
m,f
Appl A - balance dec fixed (13,2)
Appl B - balance pic 9(9)V99
Appl C - balance pic S9(7)V99 comp-3
balance dec fixed (13,2)
Appl A - bal-on-hand
Appl B - current-balance
Appl C - cash-on-hand
Current balance
Appl A - date (julian)
Appl B - date (yymmdd)
Appl C - date (absolute)
date (julian)
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
7


2.2: Characteristics of Data Warehouse
Non-volatile
insert
change
Operational
delete
insert
replace
change
load
Data
Warehouse
read only
access
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
8


2.2: Characteristics of Data Warehouse
Time Variant -
Operational
Data Warehouse
Current Value data
time horizon : 60-90 days
key may not have element of time
Snapshot data
time horizon : 5-10 years
key has an element of time
data warehouse stores historical data
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
9



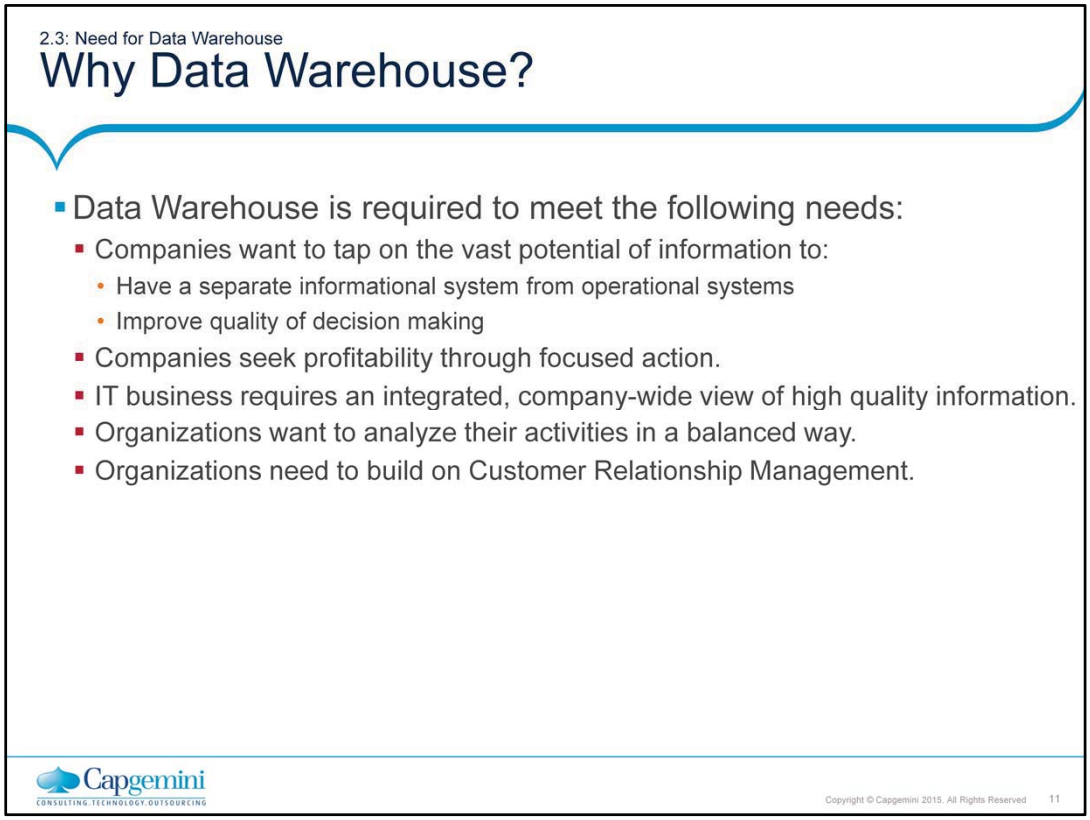
Data warehousing is known as decision support system. If we look at the historical development of Decision Support System (DSS) this is how DSS has been developed since 1960.

In the late 1960 there were only batch reports where data used to process batch wise. The analysis of information was very difficult so this resulted in slow decision making. Also it was not proved flexible.It was very expensive since it need to be re-programmed for every new request.

In the year 1970's though it was improved bit. Here, the provision has made to process on line but it was supporting stand alone (ie. terminal based DSS) hence it was difficult to obtain integrated information.

In the year 1980 there was tremendous improvement in the terminal based DSS. Hence, an attempt is made to include query tools and spreadsheet tools. This resulted an effective decision since most of the query raised by top management is adhoc in nature and they also need pictorial representation of data so that it will help them to take an effective and quick decision.

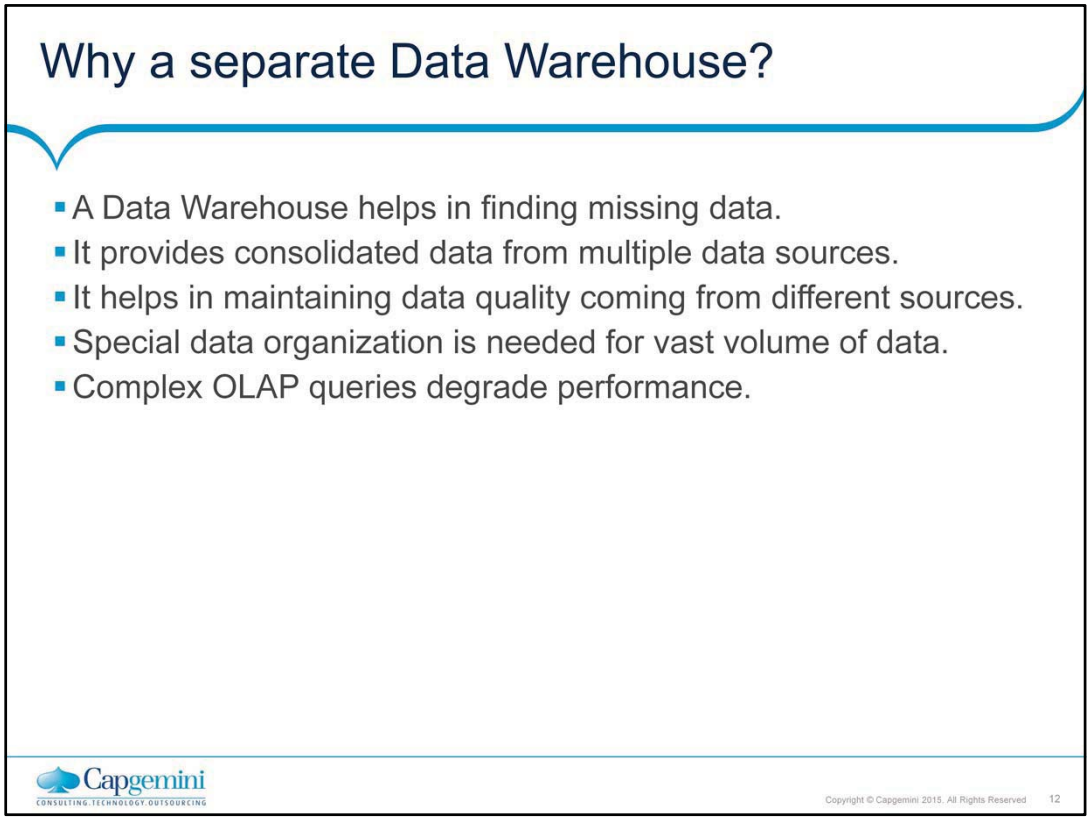
But if we look at present trend where one can obtain information in integrated manner hence information can be accessed at any point in time. Information can be accessed within no time.



Need for Data Warehouse:
The Informational Systems department must separate informal systems from operational systems in order to dramatically improve performance in managing company data. Operational Data systems are typically fragmented and are inconsistent. They are distributed over a variety of incompatible hardware and software platforms.
IT professionals, in turn, must ensure that the enterprise's IT infrastructure properly supports a myriad set of requirements from different business users, each of whom has different and constantly changing needs.

Example: One file containing customer data may be located on a UNIX based server running an Oracle DBMS, while another is located on IBM main frame running the DB2 DBMS.

Organizations want to analyze the activities in a balanced way.
Customer Relationship Management is a building block of organizations. Organizations, in all sectors, are realizing that there is value in having a total picture of their interactions with customers across all touch points like for a bank, these touch points include ATM, electronic funds transfers, investment portfolio management, and loans.



Why a separate Data Warehouse?
Functions of a Data Warehouse:
A Data Warehouse is typically used for data consolidation and enforcing uniform data quality.

Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources, namely operational databases, external sources.
Data quality: Different sources typically use inconsistent data representations, codes, and formats that have to be reconciled.

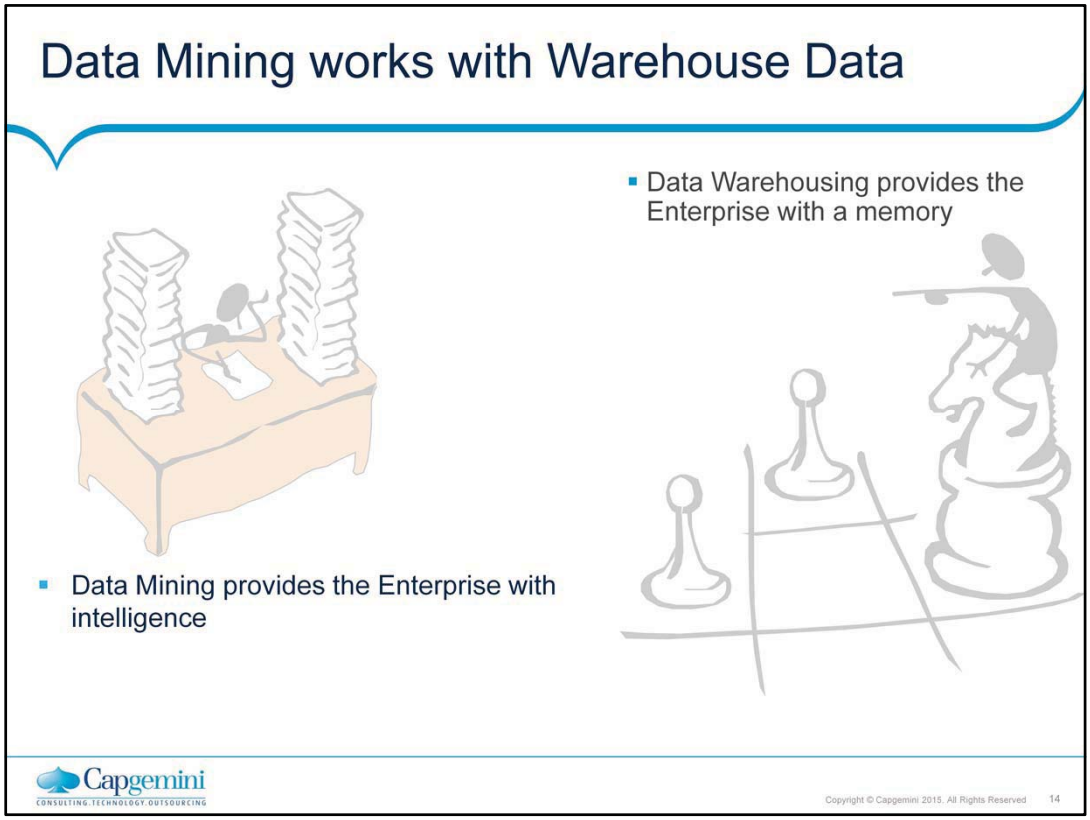
2.4: Data Warehouse Architecture

What is Data Warehouse Architecture?

- Data Warehouse Architecture is a description of the components and services of the Data Warehouse.
  - It provides the mechanism to achieve enterprise integration to support business.
  - It provides an organizing framework that will improve data sharing.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13



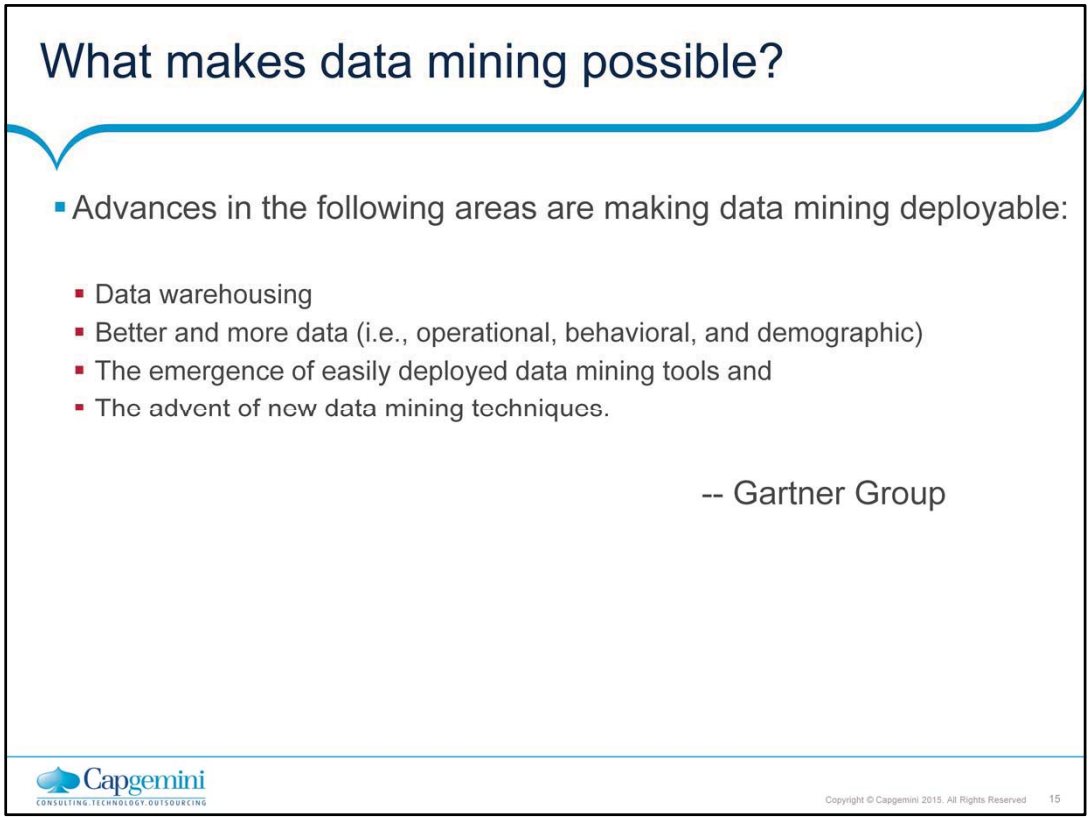
Data Mining is one of the new research avenue which is closely works with data warehousing. Data mining is the process of extracting some relevant and useful data from an already available data store .

Data warehouse is acting as enterprises memory since it holds huge amount of data from enterprise wide where as data mining adds an intelligent to the enterprise memory.

Data Mining helps to extract hidden data from the data warehouse. So, We can define data mining in general - It is process of extracting hidden data from the data warehouse which is previously unknown and potentially useful for decision making process.

There are number of Data Mining tools are available in the market such as SAS, Intelligent miner weka etc.

And there are number of Data Mining applications such as fraud detection, risk analysis , churn predication and so on and so forth in various sectors such as banking, telecommunication and insurance etc.



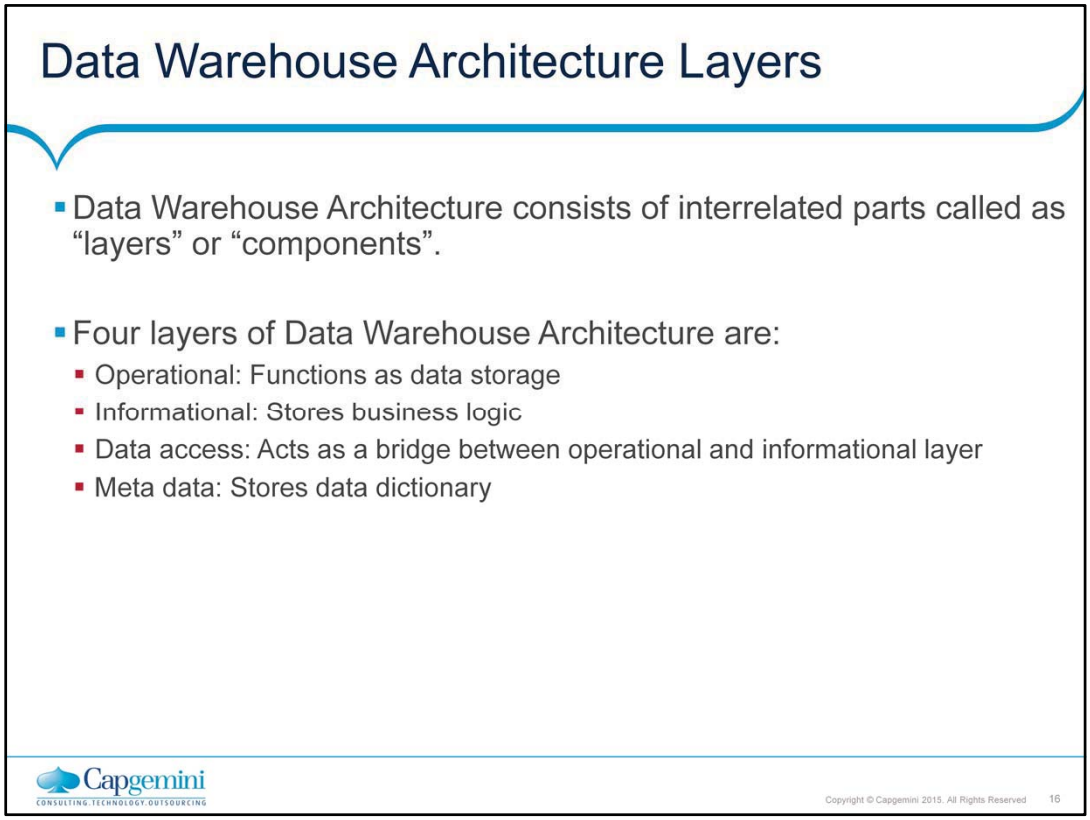
As per Gartner Group Survey Data Mining made possible to work with data warehouse since ;

Technology is supporting for holding huge amount of data in the data warehouse. The data warehouse is growing day by day because of technology is supporting to enter data into database eg. Entering data through bar coder, through e-commerce site etc.

The data is coming from various sources is more cleaned format. And more scrubbing tools are available in the market to clean the data based on requirement.

In early days there were only few statistical techniques to analyze data. Now days we can get many advanced statistical techniques such as artificial neural network etc are used for analyzing complex data. These advanced techniques helps to retrieve data more efficiently.

As a result Data Mining made all possible to work with data warehouse very effectively.



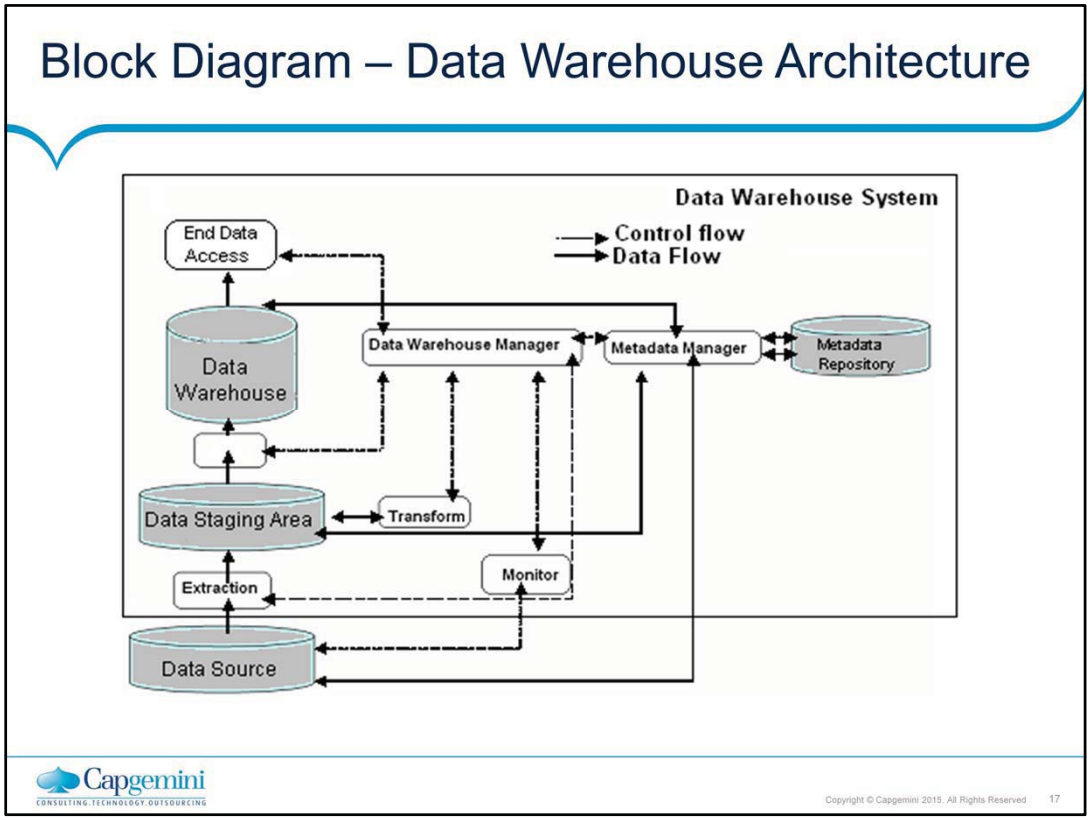
Data Warehouse Architecture:
It consists of four interrelated layers:

Operational: It is the data source for Data Warehouse. It is also called as Internal / Physical layer. It takes care of how data is stored physically on disk.

Informational: It performs data extraction for conducting analysis and reporting. It is also known as External / Logical layer. It is concerned with the way data is presented to the end user.

Data access: It is an interface between Operational and Informational layer. It is also known as Conceptual layer.

Meta data: It serves as a data dictionary for Data Warehouse.



Data Warehouse Architecture:

Let us go through the different aspects of the Data Warehouse Architecture:

Data Staging Area: You need to clean and process your operational data before putting it into the warehouse. You can do this programmatically, although most data warehouses use a Staging Area instead.

The Data Warehouse Staging Area is a temporary location where data from source systems is copied.

A staging area is mainly required in a Data Warehousing Architecture for timing reasons. In short, all required data must be available before data can be integrated into the Data Warehouse.

Metadata: It provides a guide for warehouse users to understand DW.

End User Access Tools: High performance is achieved by pre-planning the requirement for joins, summations, and periodic reports by end users.

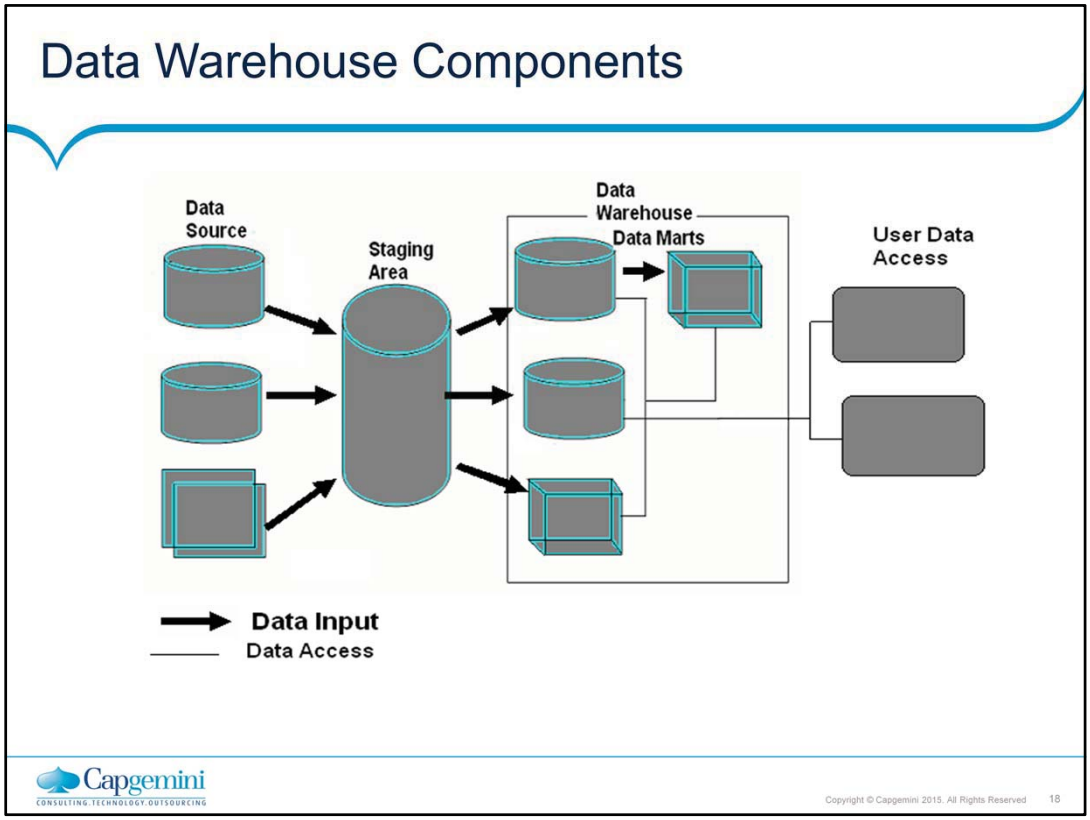
Data Warehouse Manager: It performs all operations associated with the management of the data in the warehouse.

First, at the Data Access layer, the Data Source contains information.

At the Operational layer, the data is extracted from Data Source and put into Data Staging Area.

Metadata Repository stores the guidelines about Data Warehouse. With the help of transformation techniques, the Data Warehouse Manager and Metadata Manager load data into Data Warehouse.

Finally, in the Informational layer, with the help of external view of database, the end user accesses the data.



Data Warehouse Components:

There are various components of Data Warehouse:

Data Source: Typically data is sourced from transaction processing systems (Manufacturing, ERP, Sales).

Data often resides in heterogeneous databases.

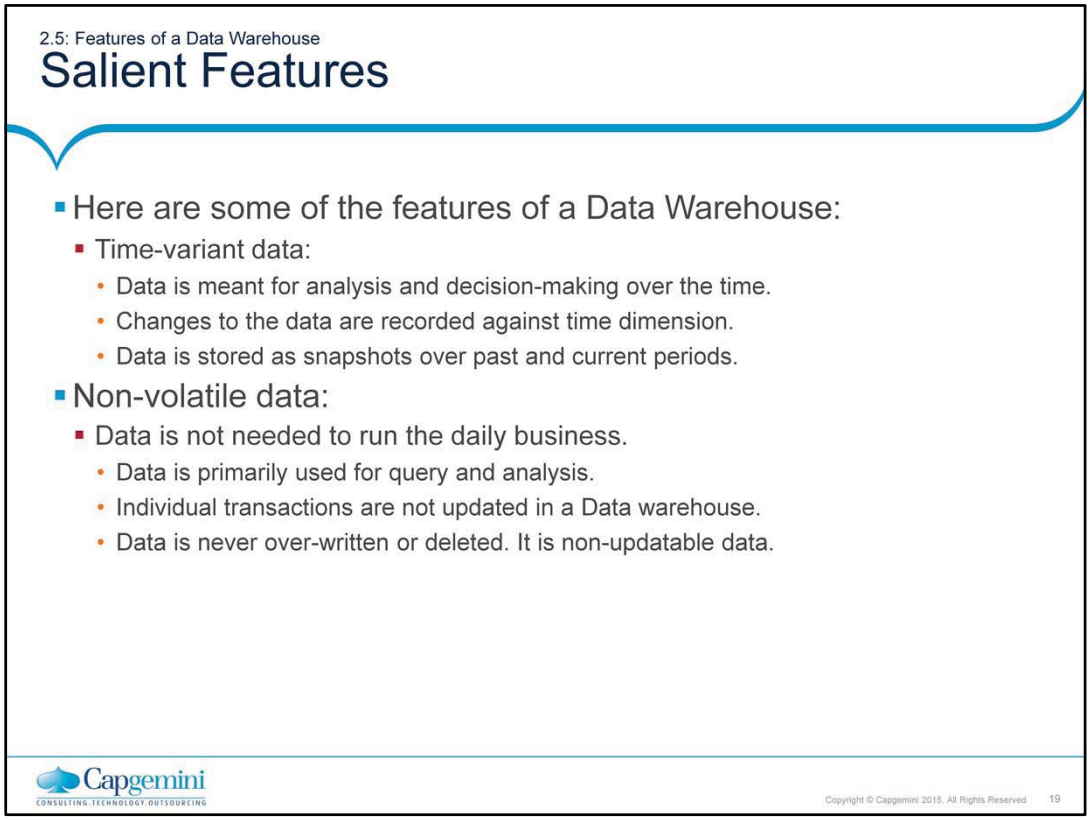
It comprises of different relational data (ORACLE, DB2, SQL Server, etc.).

Data could be on Mainframe (VSAM, IMS).

Data Staging Area: You need to clean and process your operational data before putting it into the warehouse. You can do this programmatically, although most data warehouses use a Staging Area instead.

Data Marts: You may want to customize your warehouse's architecture for different groups within your organization. You can do this by adding data marts, which are systems designed for a particular line of business.

End User Access Tools: High performance is achieved by pre-planning the requirement for joins, summations and periodic reports by end users.



Features of a Data Warehouse:
Here are some of the features of a Data Warehouse:
Time-variant data:

Data in the Data warehouse contains a time dimension so that it may be used to study trends and changes.
This nature of data:

Allows for analysis of the past.
Relates information to the present.
Enables forecast of the future.

Non-volatile data:

Data in the Data warehouse is loaded and refreshed from operational systems. However, it cannot be updated by end users.

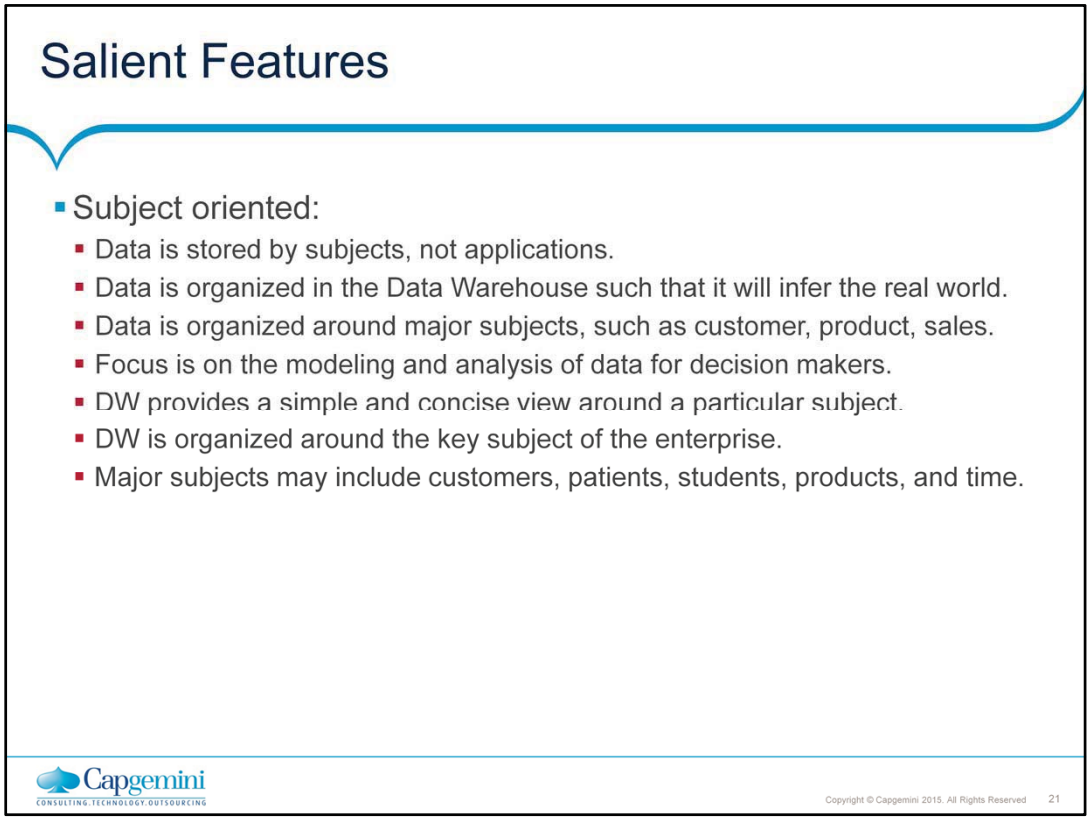
Non-volatile data is not needed to run the daily business.
Non-volatile data is primarily used for query and analysis.
Individual transactions are not updated in a data warehouse.
Data is never over written or deleted.
Data warehouse consists of only non-updatable data.

Salient Features

- Data granularity:
  - It refers to the level of detail.
  - It is inversely proportional to the amount of data stored.
  - Data is summarized at different levels.
  - Many Data warehouses have at least two levels of granularity.
  - Summarized data is stored.
  - It reduces storage costs.
  - It reduces CPU usage.
  - It increases performance since smaller number of records have to be processed.
  - Design is around traditional high level reporting needs.
  - Tradeoff with volume of data to be stored and detailed usage of data.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 20



Features of Data Warehouse (contd.):

Data Warehouse is subject-oriented:

Focus is on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.

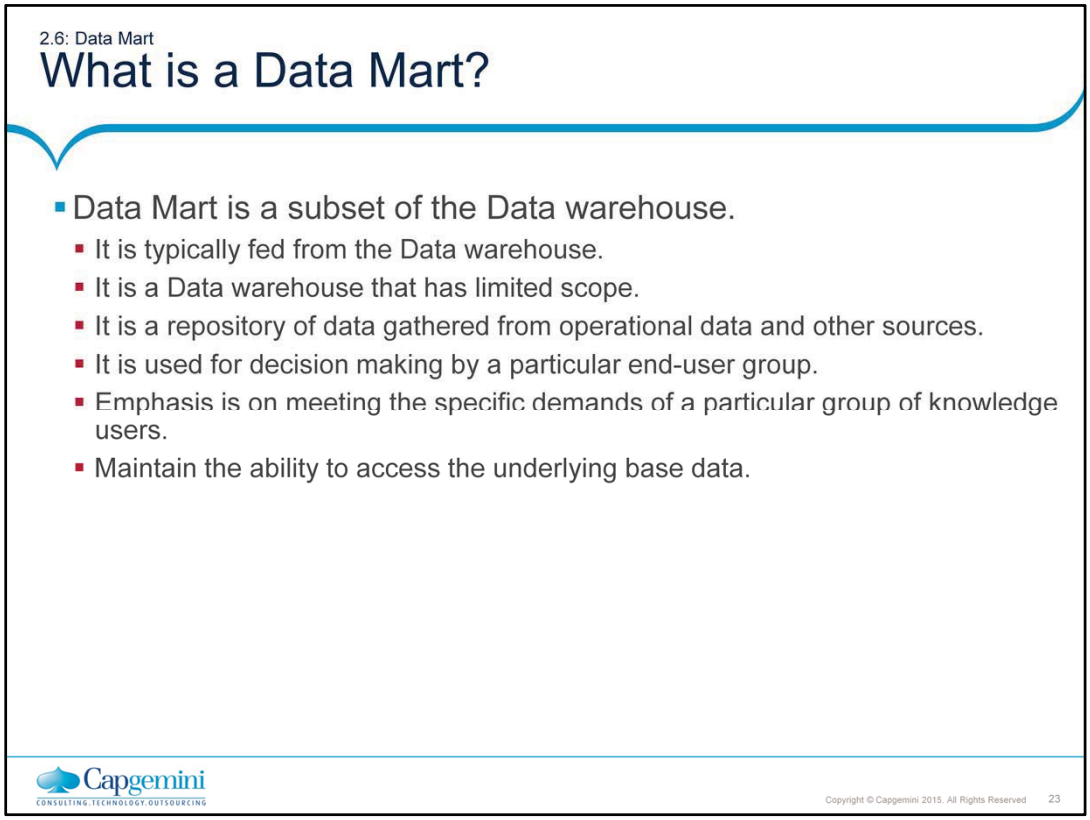
DW provides a simple and concise view around particular subject issues by excluding data that is not useful in the decision support process.

Salient Features

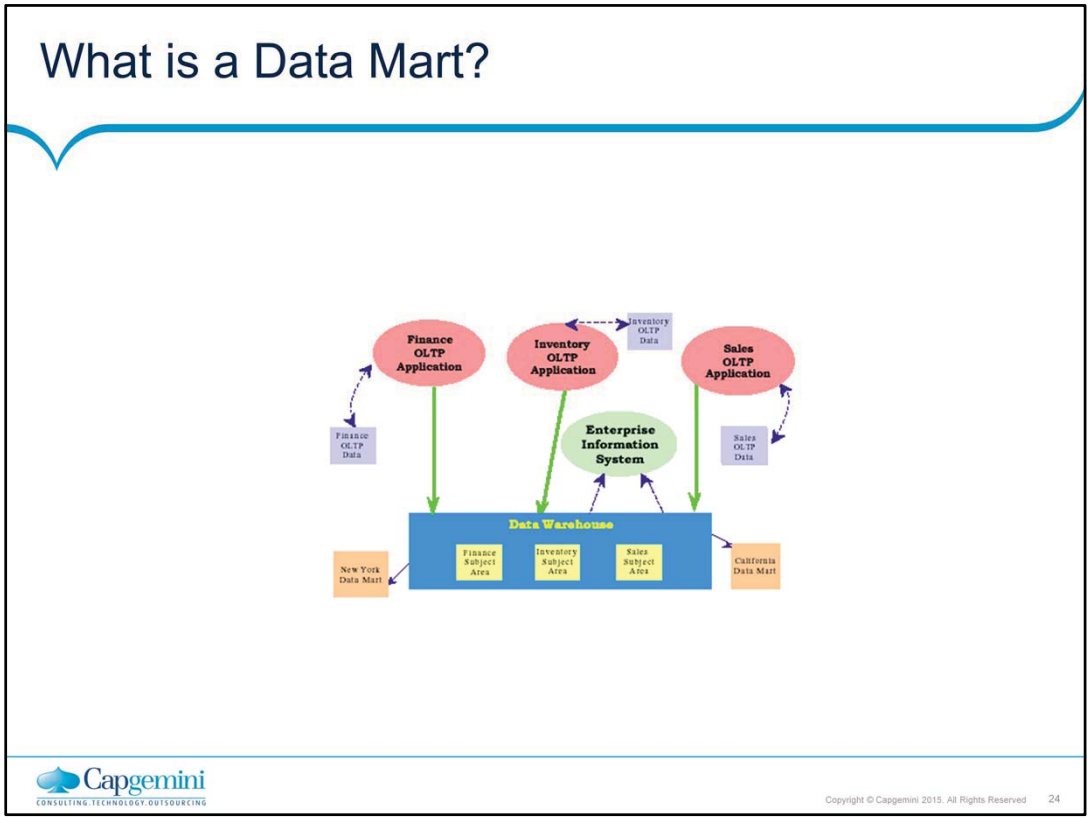
- Integrated data:
  - Data is pulled form various databases from all applications.
  - Operational platforms and operating systems for the data could be different.
  - Data has to undergo a process of transformation, consolidation, and integration.
  - Data inconsistencies are removed, standardization is achieved.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

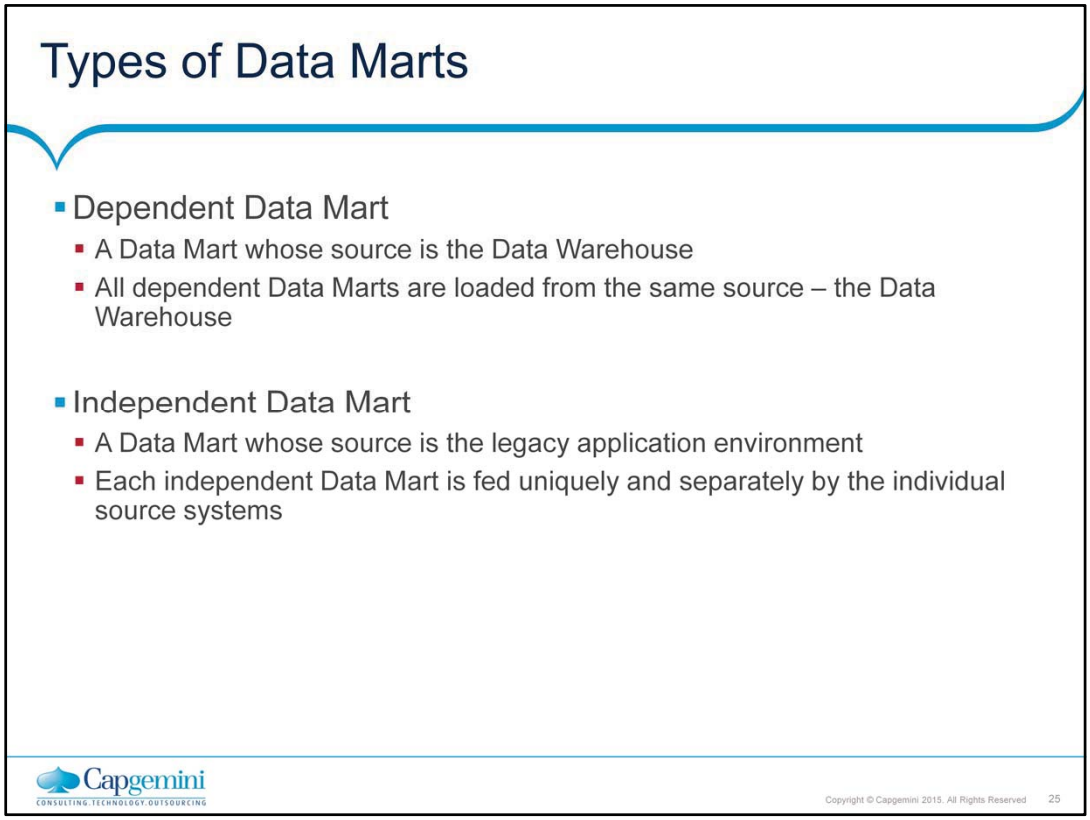
 22



Data Mart:
Data Mart is a logical subset of a Data Warehouse that may make it simpler for users to access key corporate data. A Data Mart has a smaller data model, users only need a piece of data from the data warehouse.
A Data Mart is a repository of data gathered from operational data and other sources. It is designed to serve a particular community of knowledge workers.
In scope, the data may derive from an enterprise-wide database or data warehouse or be more specialized. The emphasis of a Data Mart is on meeting the specific demands of a particular group of knowledge users in terms of analysis, content, presentation, and ease-of-use. Users of a Data Mart can expect to have data presented in terms that are familiar.
In practice, the terms Data Mart and Data warehouse, each tend to imply the presence of the other in some form. However, most writers using the terms seem to agree that the design of a Data Mart tends to start from the analysis of user needs. Similarly, the design of a Data warehouse tends to start from an analysis of the data that already exists and the manner in which it can be collected in such a way that the data can be used later.
A Data warehouse is a central aggregation of data (which can be distributed physically). Whereas a Data Mart is a data repository that may or may not derive from a Data warehouse and emphasizes on the ease of access and usability for a particular design purpose. In general, a Data warehouse tends to be a strategic but somewhat unfinished concept. A Data Mart tends to be tactical and aimed at meeting an immediate need.
In short, we need one large and complete Data warehouse that provides information to more focused, department-specific, and efficient Data Marts.
Data Mart may derive from an enterprise-wide database or data warehouse or be more specialized.



Data Mart (contd.):
Data Mart is a Data warehouse that is limited in scope.
The emphasis of a data mart is on meeting the specific demands of a particular group of knowledge users in terms of analysis, content, presentation, and ease-of-use.
Users of a Data Mart can expect to have data presented in terms that are familiar.
It is important to maintain the ability to access the underlying base data to enable drilldown analysis as necessary. The only difference between a Data Warehouse and a Data Mart is the scope. One can define the Data Warehouse from various Data Marts. On the other hand, one can define Data Marts from the Data Warehouse.



The main difference between independent and dependent data marts is how you populate the data mart; that is, how you get data out of the sources and into the data mart. This step, called the Extraction-Transformation-and Loading (ETL) process, involves moving data from operational systems, filtering it, and loading it into the data mart.

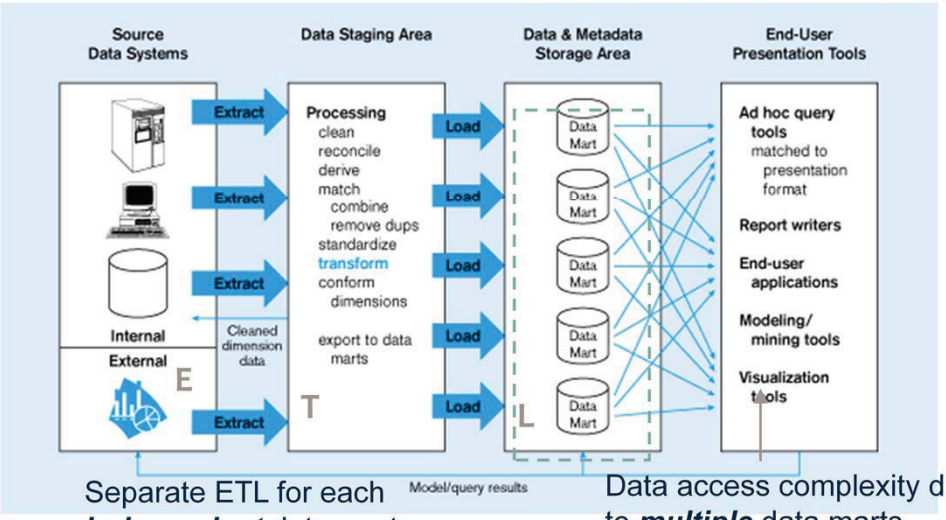
With dependent data marts, this process is somewhat simplified because formatted and summarized (clean) data has already been loaded into the central data warehouse. The ETL process for dependent data marts is mostly a process of identifying the right subset of data relevant to the chosen data mart subject and moving a copy of it, perhaps in a summarized form.

In Independent data mart, each data mart is sourced directly from the operational systems. One must deal with all aspects of the ETL process, much as you do with a central data warehouse. The number of sources is likely to be fewer and the amount of data associated with the data mart is less than the warehouse, given your focus on a single subject.

The motivations behind the creation of these two types of data marts are also typically different. Dependent data marts are usually built to achieve improved performance and availability, better control, and lower telecommunication costs resulting from local access of data relevant to a specific department. The creation of independent data marts is often driven by the need to have a solution within a shorter time.

Independent Data Mart

Data marts:
Mini-warehouses, limited in scope



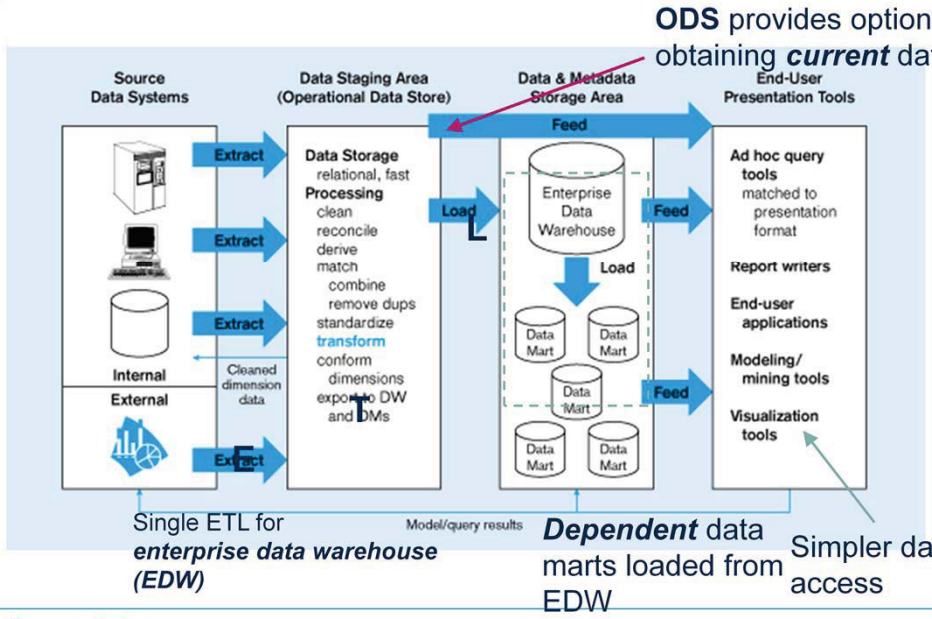
Separate ETL for each ***independent*** data mart

Data access complexity due to ***multiple*** data marts

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 26

Dependent data mart with operational data store



2.7: Data Warehouse Application Areas

Industry-wise Application

Industry	Application
Finance	Credit Card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record analysis
Transport	Logistics management
Consumer goods	Promotion analysis
Data Service providers	Value added data
Utilities	Power usage analysis

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 28

Summary

- In this lesson, you have learnt:
  - Data Warehouse stores historical data.
  - Data Mart emphasizes on meeting the specific demands of a particular group of knowledge users.
  - Features of Data Warehouse are:
    - Time variant data
    - Non volatile data
    - Data granularity
    - Subject oriented
    - Integrated data

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 29

Review Question

- Question 1: _____ is a subset of data warehouse.
- Question 2: Data Mart is a structure for corporate view of data.
  - True/ False
- Question 3: ___ is used for decision making by a particular end-user group.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 30

Data Warehousing Concepts

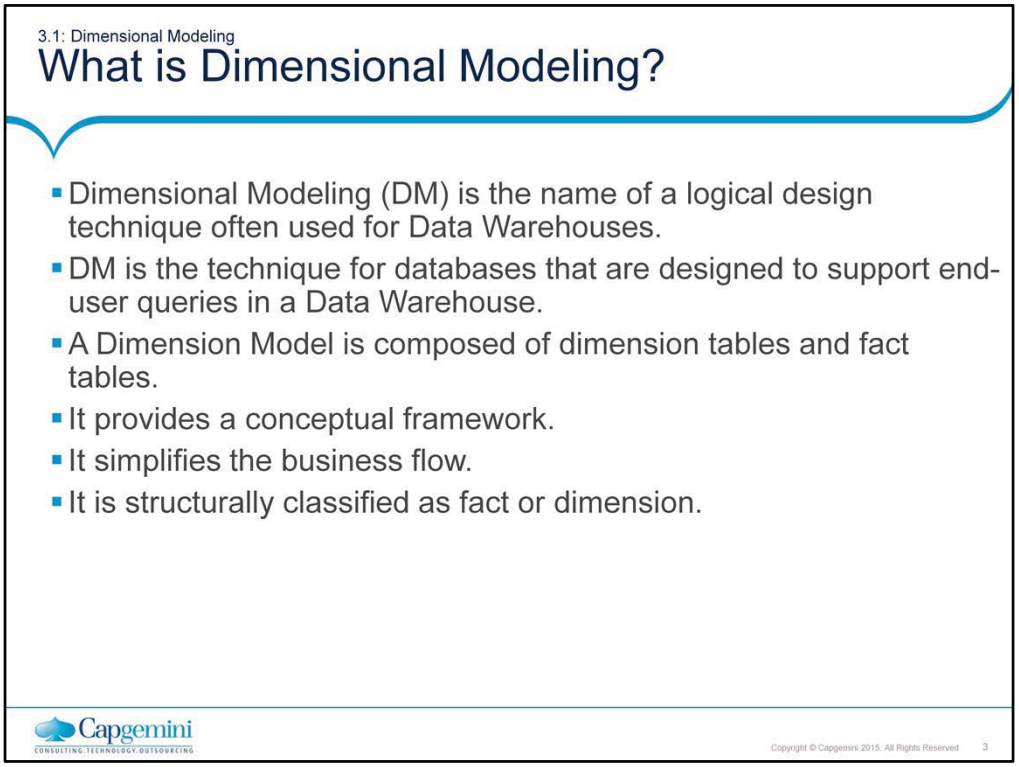
Lesson 3: Dimensional Modeling

Lesson Objectives

- In this lesson, you will learn:
  - What is Dimensional modeling ?
  - Facts and Dimension tables
  - Database schema
  - Schema Design for Modeling

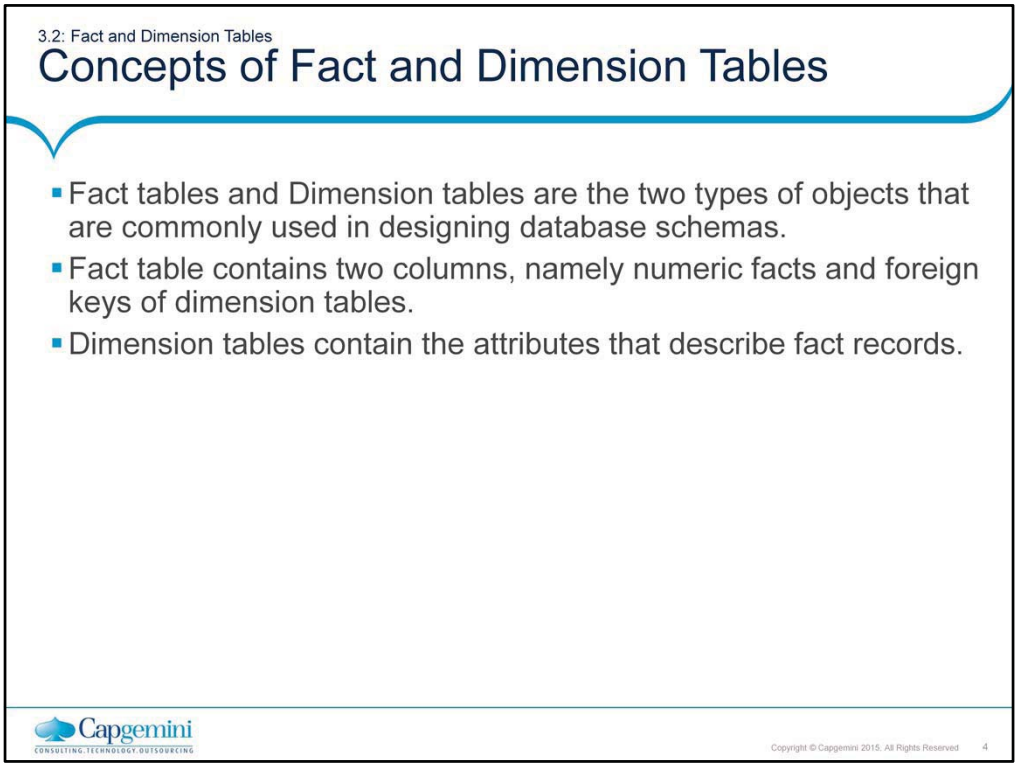
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2



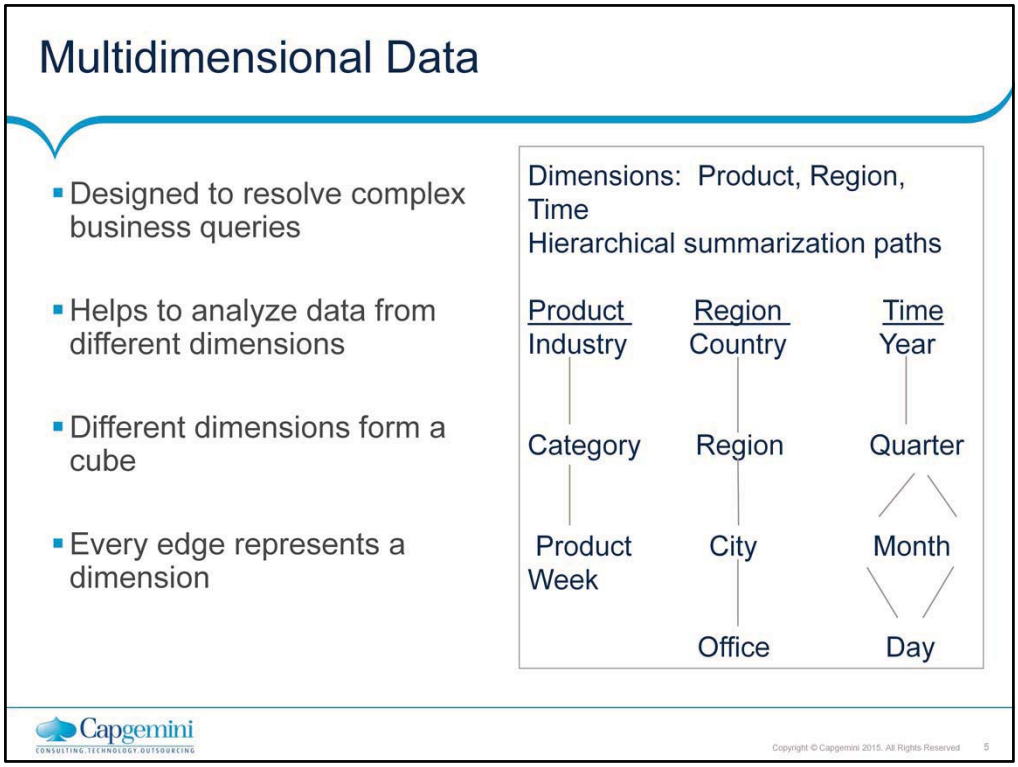
Dimensional Modeling:

- **Dimensional Modeling** has the characteristic for organizing data roughly into base facts and dimensions of those facts.
- Dimensional Modeling provides the Conceptual Framework. It is basically used for faster query performance for the business users. **Facts** are basically organization's business processes. They are usually numeric values. **Dimension** is a context that describes the fact.
- Every organization has Dimensional Modeling for its business processes, and it consists of **fact tables** and **dimensional tables**. It helps business users in easily understanding the typical system model.
- Dimensional Modeling represents the complexities of the business process in a simple manner. **Understandability** and **Query performance** are two major reasons for which dimensional modeling is accepted widely in the industry.
- Dimensional Modeling is a logical design technique that allows to retrieve the data with high-performance.



Fact and Dimension Tables:

- A **fact table** has two types of columns.
  - The first column type contains numeric facts (often called measurements).
  - The other column type contains the foreign keys of dimension tables.
- A **fact table** contains multiple foreign keys.
- Each pair of primary key in dimension and foreign key of fact table contains the measurements.
- A **Dimension table** contains the attributes that describe fact records. Some dimension table attributes provide descriptive information and other attributes (primary key) are used to join with fact tables.
  Example: A customer dimension table contains two attributes, namely customer id (Primary key) and customer description. So we will use the primary key attribute customer id to join with fact tables.
- However, **dimensional** and **fact modeling** is not of the highest Normal Form, but makes use of a key of performance indicators. Dimensions can strive to be in Boyce Codd (BCNF) 3rd Normal Form. Whereas Fact tables may be in 1st Normal Form, having only a primary key being unique.

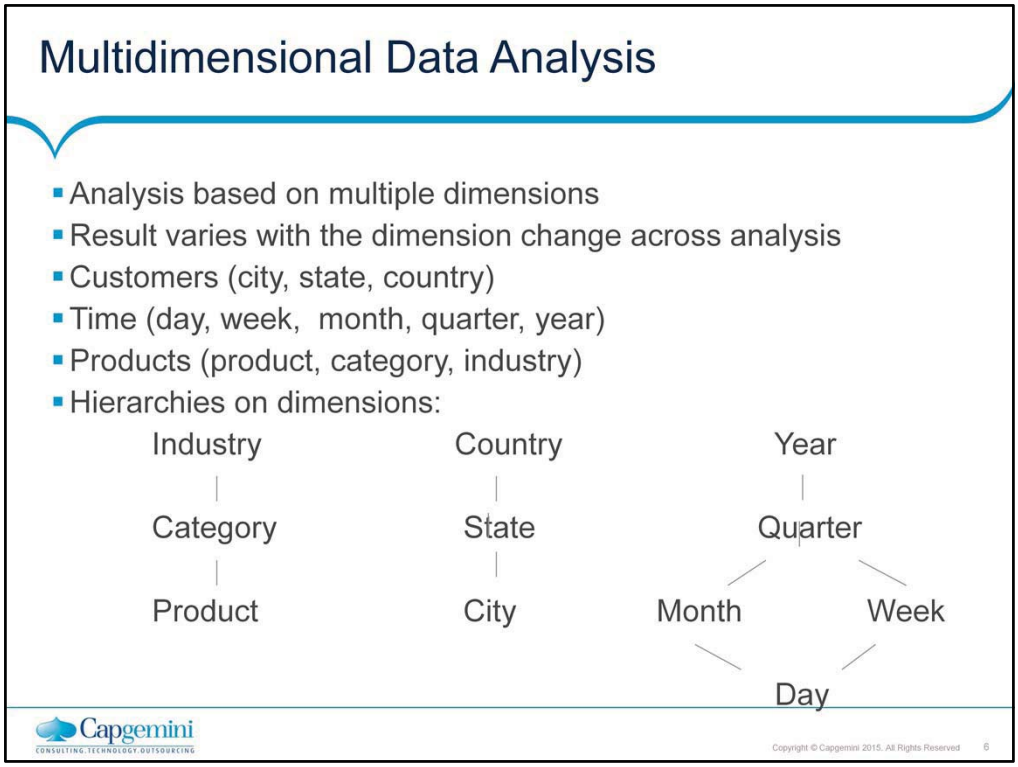


Multidimensional Data :

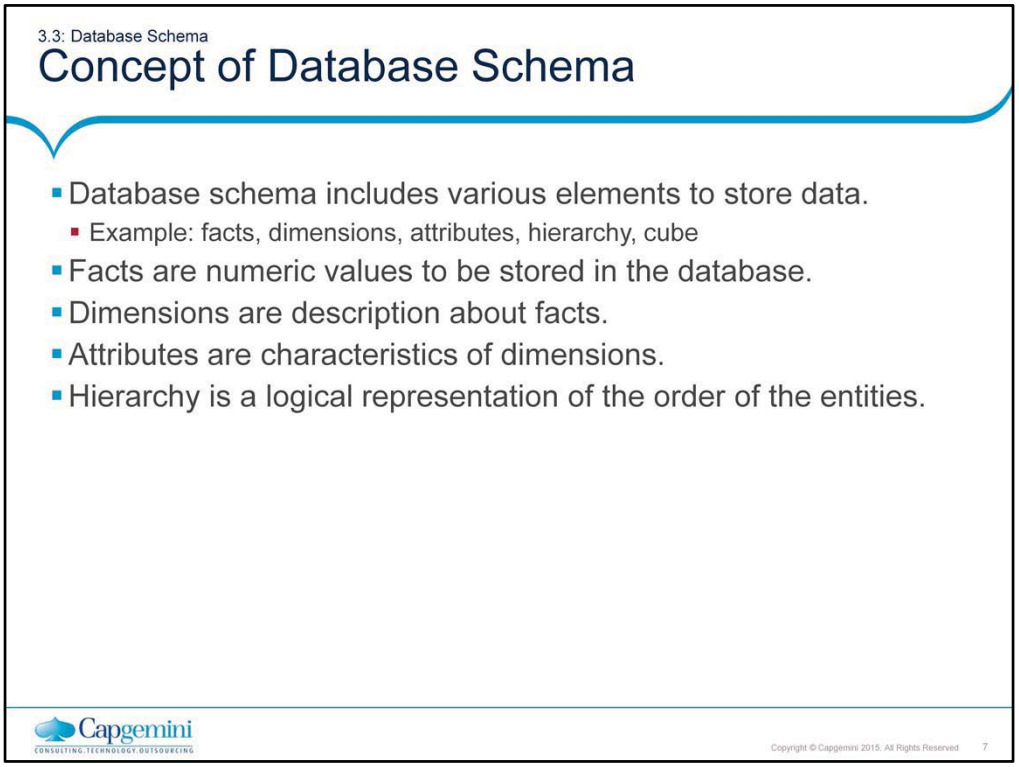
The multidimensional data model is the integral part of On line analytical Processing. The multidimensional data model is designed to resolve the complex queries.

In the logical multidimensional model, a cube represents the measures with same shape. In a cube every edge represents a dimension. Members of Dimension are aligned on the edges and divide the cube shape into cells in which stored the data values. It is basically used for developing data mart.

In above cube three edges represent the three dimension table Product, Region and time.



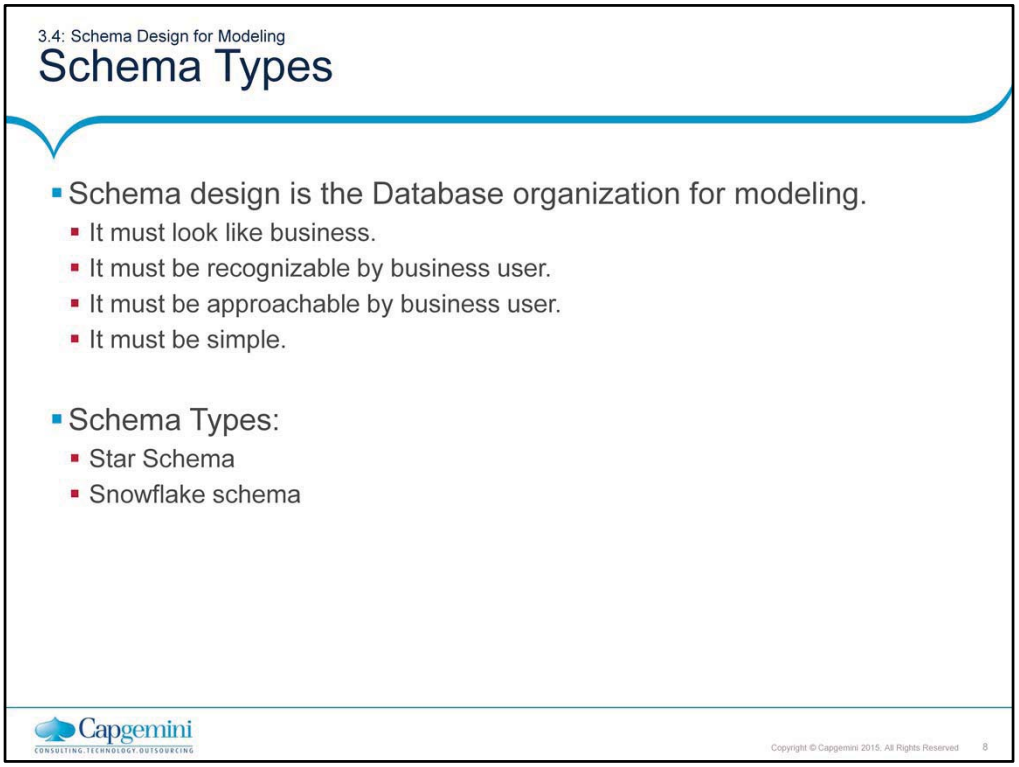
Multidimensional Data Analysis is the analysis of data based on dimensions. It includes analysis of a particular data with respect to different and multiple dimensions. The value varies when there is a change in the dimensions across the analysis. It changes in terms of context one wishes to analyze data.
For E.g. Analysis of Product by City, Transactions for last 3 years.



Fact and Dimension Tables:

Database Schema:

- **Database schema** is a set of facts in multi-dimensional data. A fact has a measure dimension quantity that is analyzed, for example, number of visas.
- It has a set of dimensions on which data is analyzed, for example, country, consulate, date of issue for a visa. Each dimension has a set of attributes
- **Example:** "Visa" dimension has visa date, visa type, and visa category
- Attributes of a dimension may be related by partial order, or Hierarchy: for example, post > county > region.



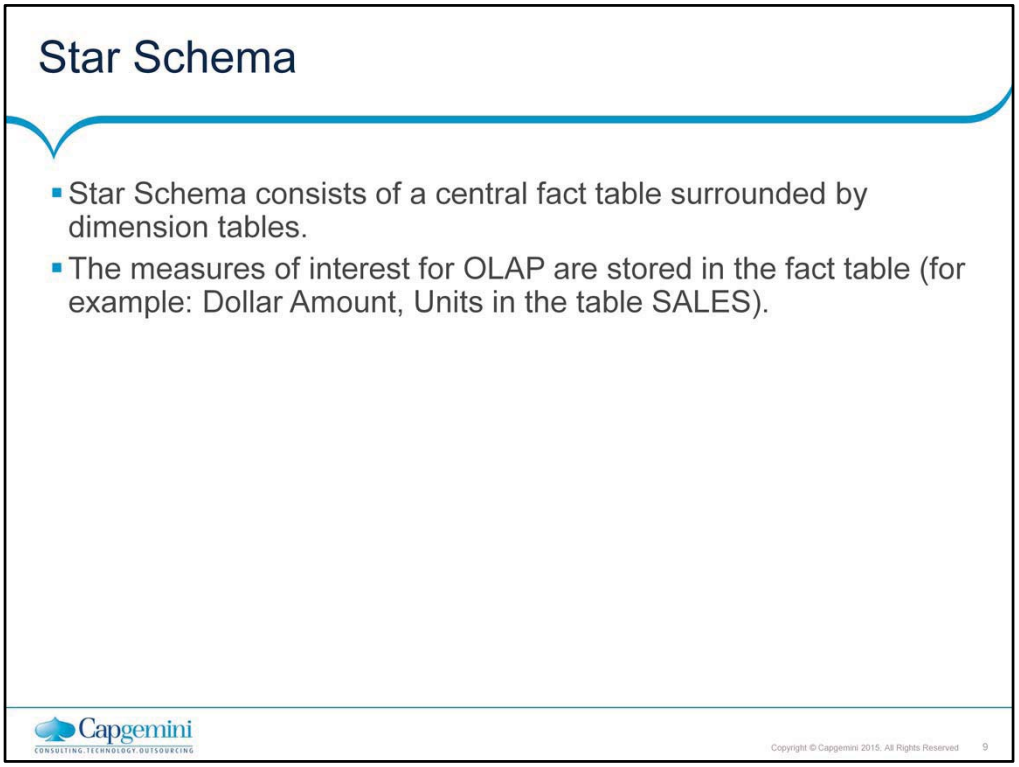
Fact and Dimension Tables:

Schema Design for Modeling:

- **Schema design** is the organization of database for modeling.
- The design shows how the model will be implemented in a system. It must be kept simple and familiar with the business context. It must be easily understood by business user. It should be designed in such a way that the business users can fully understand it in terms of facts, measures, dimensions, and hierarchies.

Schema Types:

- **Star Schema-Fact and Dimension tables:** Star schema has all multi-leveled dimensions that are flattened.
- **Snowflake Schema:** It has dimensional hierarchy directly by normalizing tables. In Snowflake schema, at least one multi-leveled dimension is kept separate.

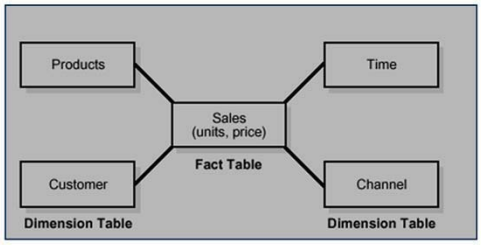


Fact and Dimension Tables:
Schema Types (contd.):

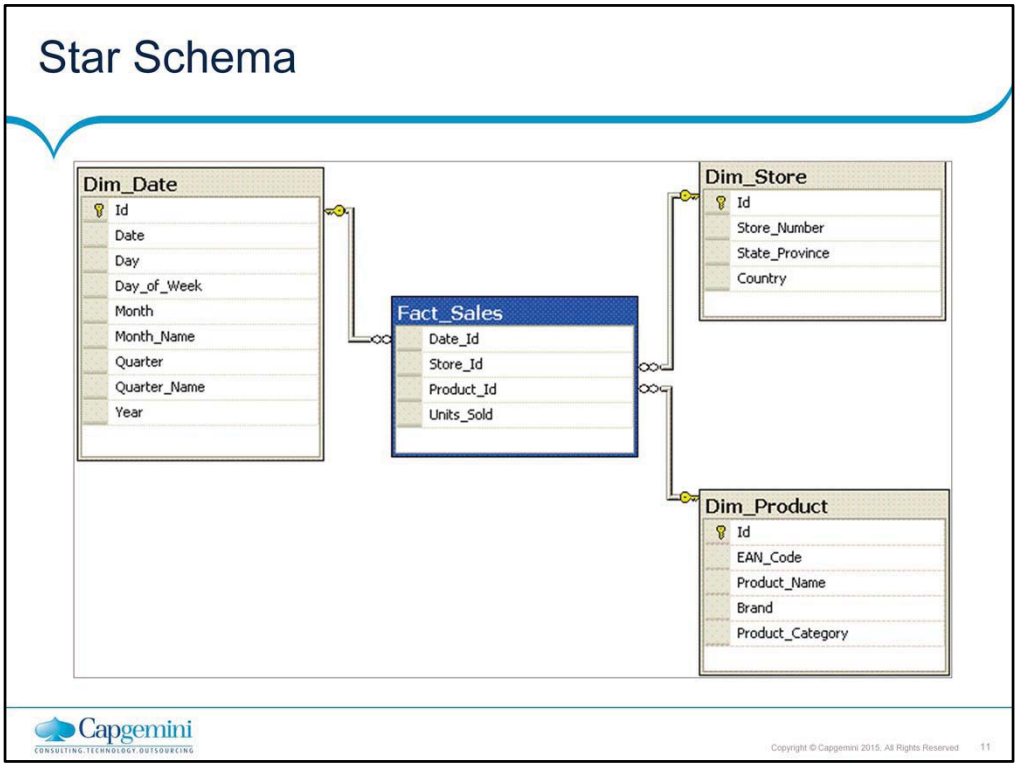
Star Schema (contd.):

- For each dimension of the multidimensional model there exists a dimension table (for example: Geography, Product, Time, Account) with all the levels of aggregation and the extra properties of these levels.
- It consists of Fact table.
- It consists Compound primary key.
- Star schema focuses on two major advantages, namely:
  - Ease of use
  - Efficient performance

Star Schema - Sample



Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

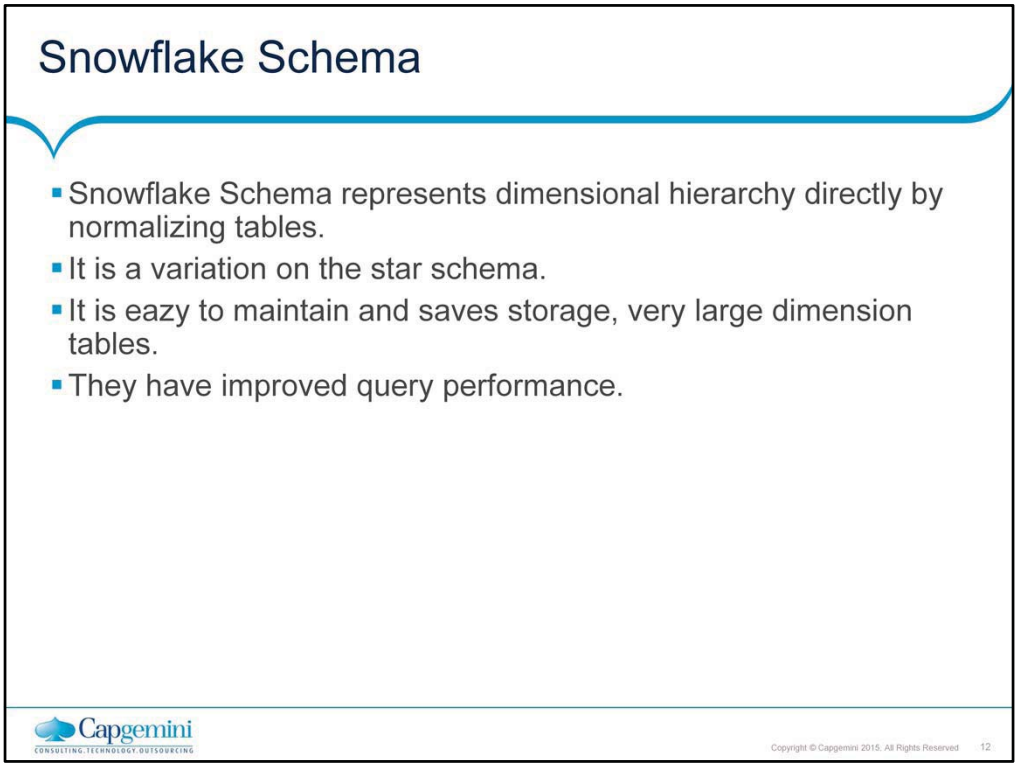


Fact and Dimension Tables:
Schema Types (contd.):

Star Schema:

- **Star schema** is commonly used by relational databases. The performance can be improved by using this design rather than traditional join operations. A **Star schema** is a database design that contains a central table, called a **fact table**, which is in relationship with many tables called **dimension tables**. This schema design resembles a star, thus the name is Star Schema. It is a very simple programmatic approach. It is very similar way in which a user thinks about a system, hence it is simple. It is easier to use. It is very efficient in the performance. It is best suited for **MOLAP application tools**. Typically, most of the fact tables in a star schema are in database Third Normal Form, while dimensional tables are de-normalized (Second Normal Form). Despite the fact that the Star schema is the simplest Data warehouse architecture, it is most commonly used in the Data warehouse implementations about 90-95%, across the world today.

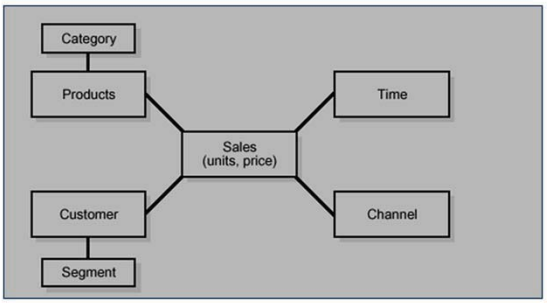
  Example:
  - Fact Table: Fact_Sales table
  - Dimension table: Dim_Date, Dim_Store, Dim_Product



Fact and Dimension Tables:
Schema Types (contd.):
Snowflake Schema:

- It is more complex than Star schema design. The main difference is that dimensional tables in a snowflake schema are normalized, so they have a typical relational database design.
- Snowflake schemas are generally used when a dimensional table becomes very big and when a Star schema cannot represent the complexity of a data structure. For example, if a PRODUCT dimension table contains millions of rows, then the use of Snowflake schemas should significantly improve performance by moving out some data to other table (with REGION for instance). The data redundancy is eliminated. The problem is that the more normalized the dimension table is, the more complicated SQL joins must be issued to query them. This is because in order for a query to be answered, many tables need to be joined.

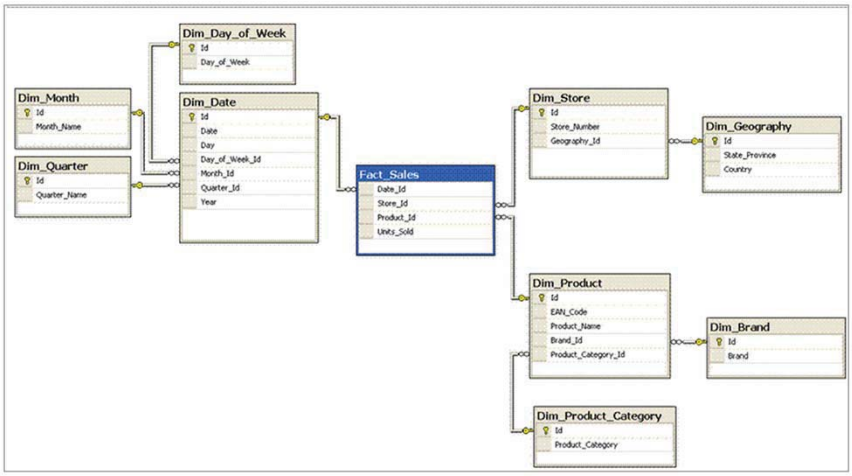
Snowflake - Sample



Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13

Snowflake Schema



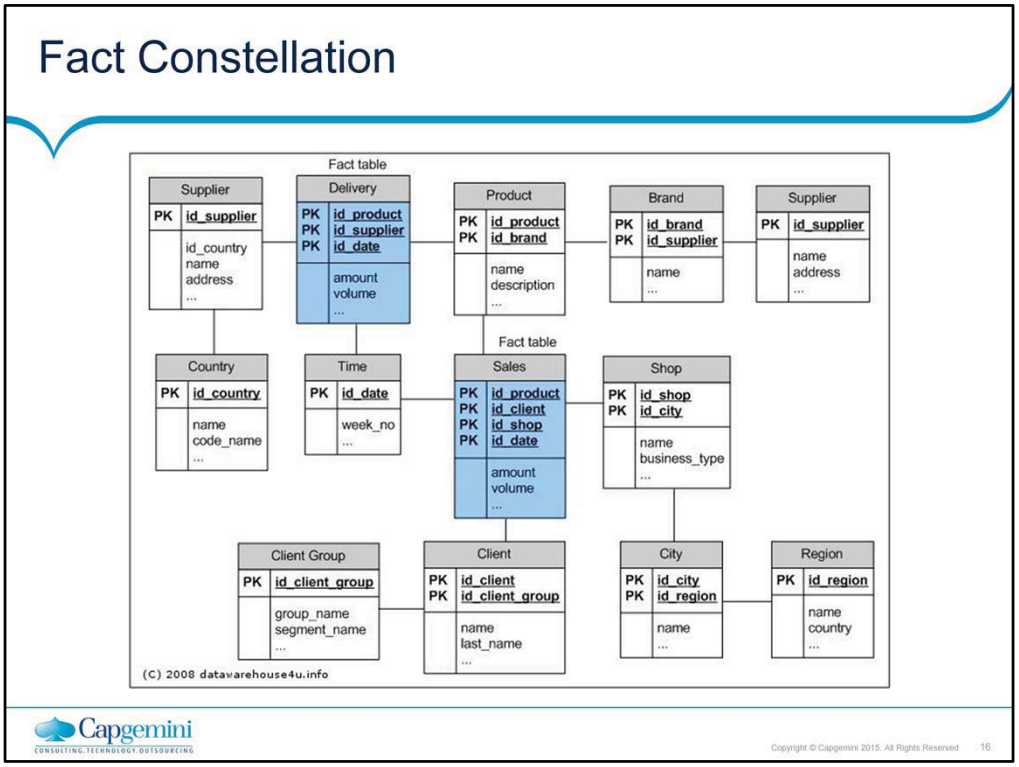
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Fact Constellation

- Multiple fact tables share dimension tables.
- This schema is viewed as collection of stars hence called galaxy schema or fact constellation.
- Sophisticated application requires such schema

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 15



In fact constellation, there are many fact table sharing the same dimension tables.

This examples illustrates a fact constellation in which the fact tables sales and shipping are sharing the dimension tables such as time,product ……..

Summary

- In this lesson, you have learnt:
  - Dimensional Modeling represents the complexities of the business process in a simple manner.
  - The schema types are star schema and snowflake schema

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 17

Summary

- Database schema has various elements, such as:
  - Fact
  - Dimension
  - Attributes
  - Hierarchy
  - Cube
- Schema design is the organization of database.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 18

Review Question

- Question 1: _____are description about facts.
- Question 2: _____ in Snowflake Shema are normalized into multiple tables.
- Question 3: ___ is the name of a logical design technique often used for Data Warehouses.

Data Warehousing Concepts

Lesson 4: ETL and Metadata

Lesson Objectives

- In this lesson, you will learn:
  - ETL Process
  - Metadata used in ETL
  - Metadata in Data Warehousing
  - Simple Warehouse Model

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2

4.1: Extract Transform and Load (ETL) Process

Concept of ETL

- The Data Warehouse always has enterprise data. Data comes from various sources, such as Spreadsheets, Mail lists, and Databases.
- The required data is extracted, transformed to suit information needs and finally loaded at a central location.
- This is done by ETL (Extract Transform and Load) process.
  - Extract: Data extraction and staging
  - Transform: Convert to format required by data warehouse
  - Load: Load data to data warehouse

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 3

ETL Process

- Extraction
  - Data extraction from various source (Heterogeneous systems)
  - Different data representations, formats
    - e.g: RDBMS, Flat files, IMS, VSAM
  - Data to be converted to a common format for transformation process
  - Extracts the data from data source and keeps in staging.
  - Data comes from an operational source or archive systems which are the primary sources of data for the Data warehouse.
  - It minimizes impact on production data sources

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 4

ETL Process

- Transformation
  - Various sets of business rules and functions are applied on extracted data before the data gets loaded to Data warehouse
  - One or more of the following steps may be involved in the transformation process
    - Selecting only certain columns to load
    - cleansing the data to remove duplicates and enforce consistency
    - Translating coded values (e.g., if the source system stores 1 for male and 2 for female, It may be translated as M for male and F for female in data warehouse)
    - Encoding free-form values (e.g., mapping "Male" and "1" and "Mr" into M)
    - Deriving a new calculated value
    - Joining together data from multiple sources (e.g., lookup, merge, etc.)

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

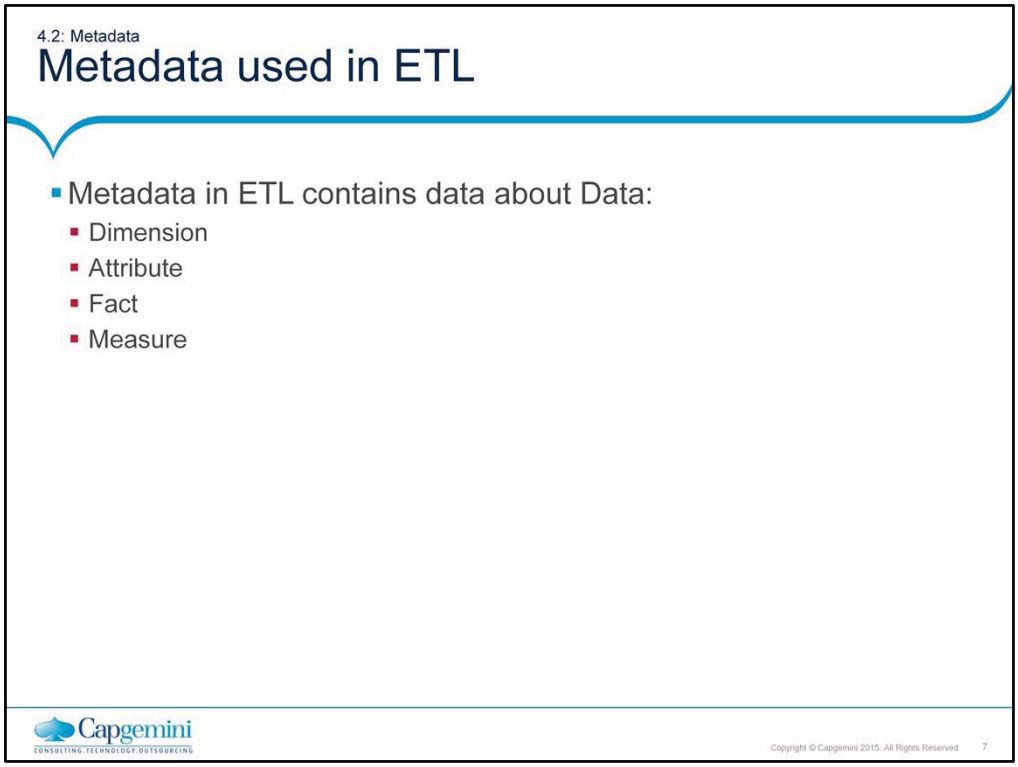
 5

ETL –Load Process

- Transformed data loaded to Data warehouse
- Load Dimensions and then Fact
- Indexes to be dropped before loading and recreated after loading the Data Warehouse
- Load cycle (Daily, weekly Monthly…)/Schedules
- Bulk Loads
- Full Refresh
- Incremental Loads

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 6

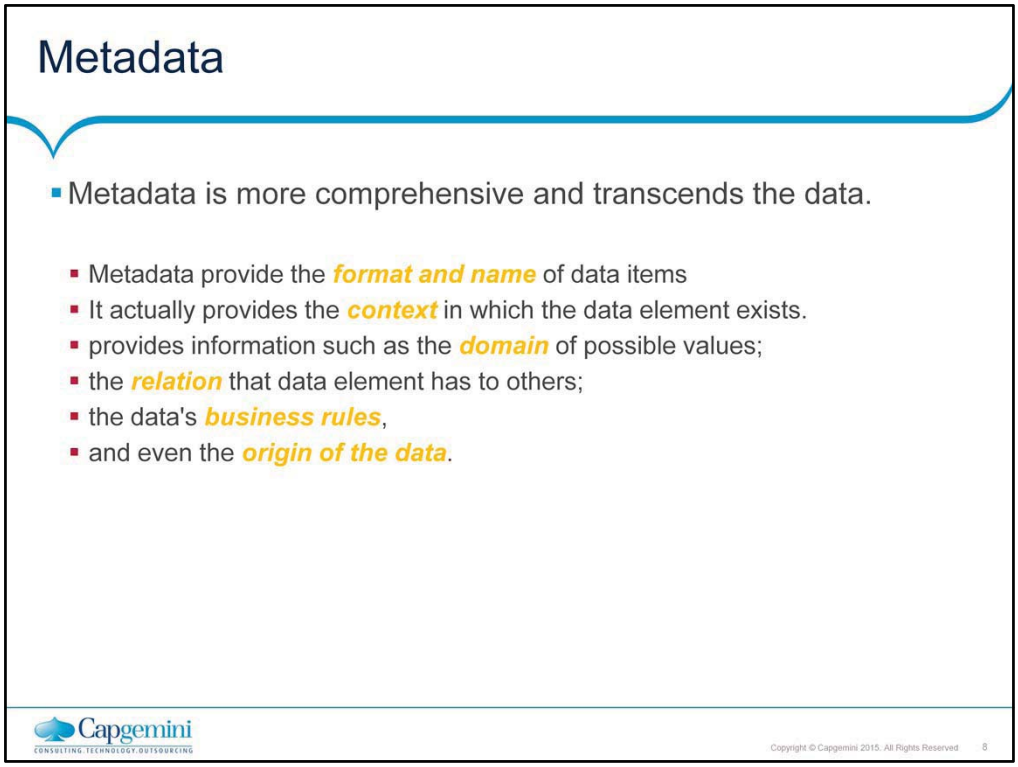


Metadata:

Metadata is the data about Data.

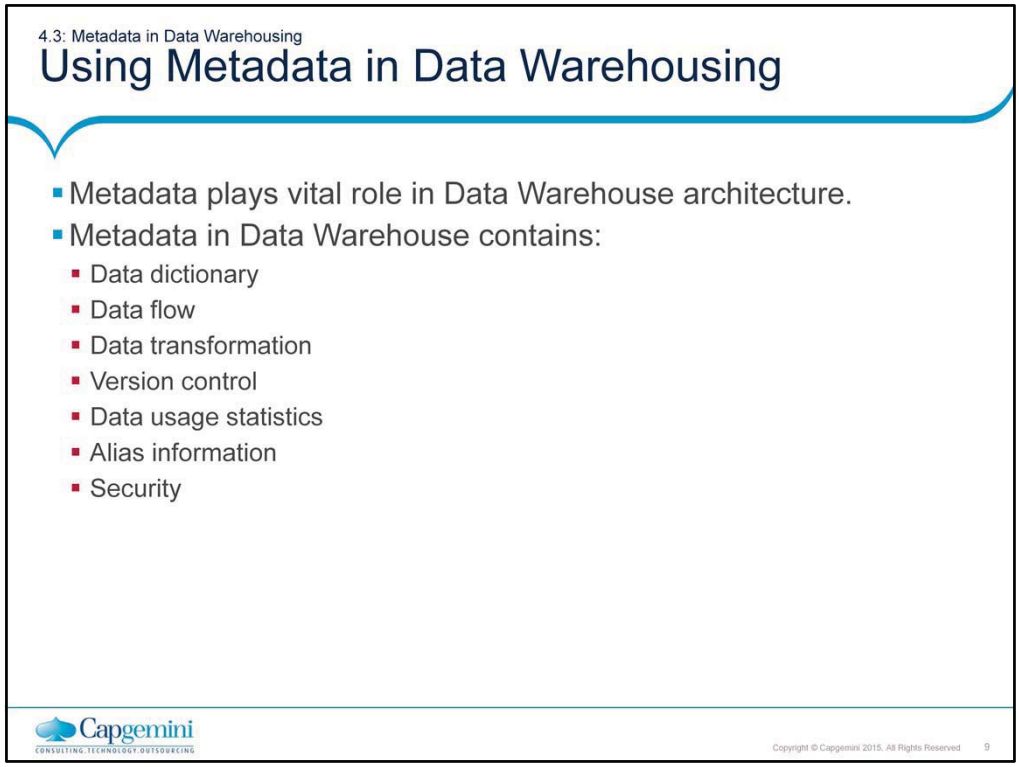
- **Dimension:** It is a perspective that can be used to analyze the data. Dimensions become more useful when there are many descriptive attributes that can be used for analyzing the data.
- **Attribute:** It is often used to describe the extended Dimension.
  Example: Customer, Item, Date, Fact
- **Fact:** It is the raw enumerable piece of information about the transaction. It is always a numeric value (usually aggregatable) about the transaction.
  Examples: Quantity, Unit Price, Count
- **Measure:** Measure can be the product of one or more fact tables. Measure can be the result of any formula which is derived from Relational Database or Business Intelligences Tool analytical engine. It is the product of one or more Facts.
  Examples: Quantity, Unit Price, Count, Quantity * Unit Price, Average (Unit Price) and Minimum (Quantity).

  For example, if a customer is an attribute from your databases table, then customer metadata can give the information about the customer like name, address will be the dimensions, whereas telephone number can act as fact, etc. Age can considered as measure, since it can be calculated from DOB-Current date.



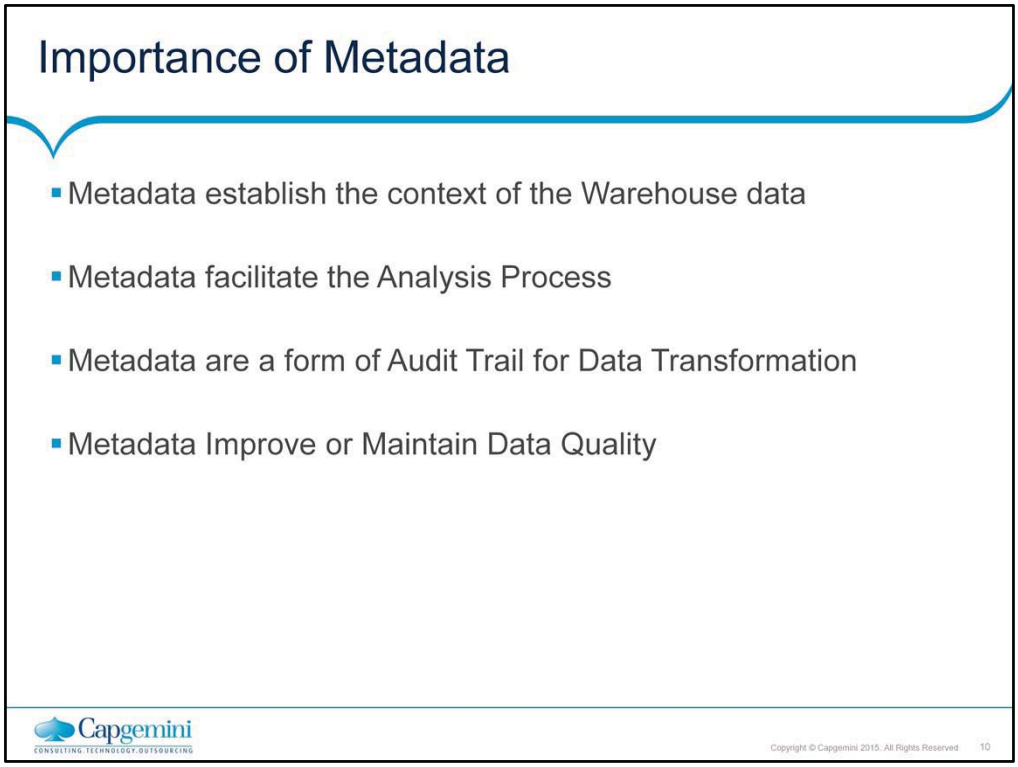
Metadata is the high level core internal document of the source code which runs as the lifeblood for a data warehouse.

Metadata not only describe the format and name but it provides details about the context I,e what is the need of the data item and what are the values that the data item can have, the relationship between the data elements ie whether the data element is found on other locations and how they are inter-linked to each other. Apart from the technical details It also holds the business rule. The origin of the data is so critical that the end user might like to trace back to the origin of the data which end user sees through the OLAP tools.



Metadata in Data Warehousing:

- Metadata is the blood of the Data Warehouse. It is the information that describes the system. Metadata plays a vital role in Data Warehouse architecture. It provides the information to the application to control warehouse activities. A single change in the metadata repository affects the entire architecture.
- Metadata in Data Warehousing:
  - **Data dictionary:** It contains definitions of the databases and relationship between data elements.
  - **Data flow:** It contains direction and frequency of data feed.
  - **Data transformation:** It contains transformations required when data is moved.
  - **Version control:** It records changes to stored metadata.
  - **Data usage statistics:** It is a profile of data in the warehouse.
  - **Alias information:** It contains alias names for a field.
  - **Security:** It contains the names of the data access authorized people.



Importance of Metadata

Metadata establish the context of the Warehouse data
Metadata helps data warehouse administrators and users locate and understand data items, both in the source systems and in the warehouse data structures.

E.g.: The date 02/05/2010 could mean either May 2, 2010 or February 5, 2010 depending on the date convention used. Metadata describing the format of this date field could help determine the definite and unambiguous meaning of the data item.

Metadata facilitate the Analysis Process
Metadata must provide data warehouse end-users with the information they need to easily perform the analysis steps. It should thus allow users to quickly locate data that are in the warehouse.

Metadata should allow analysts to interpret data correctly by providing information about data formats and data definitions.

Metadata are a form of Audit Trail for Data Transformation
Metadata document the transformation of source data into warehouse data. Hence warehouse metadata must be capable of explaining how a particular piece of warehouse data was derived from the operational systems.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

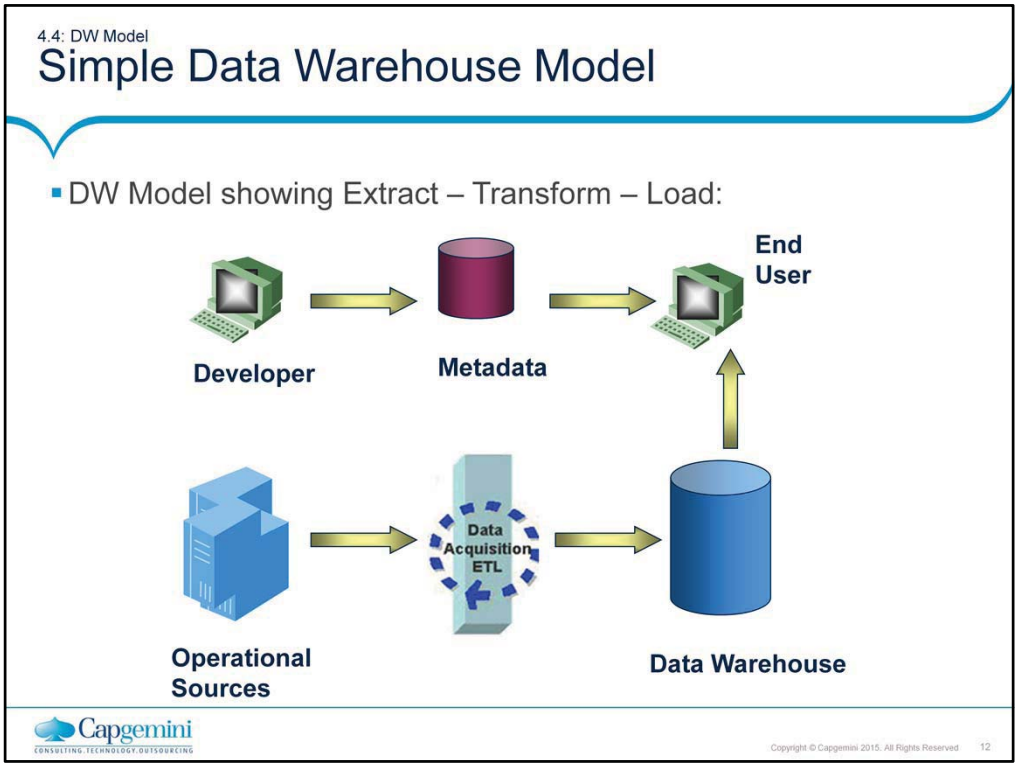
Metadata Improve or Maintain Data Quality
Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on an as needed basis.

All business rules governing the transformation of data to new values or new formats are also documented as metadata.

Metadata Improve or Maintain Data Quality
Metadata can improve or maintain warehouse data quality through the definition of valid values for individual warehouse data items. Using a data quality tool prior to actual loading into the warehouse, the warehouse load images can be reviewed to check for compliance with valid values for key data items. Data errors are quickly highlighted for correction.

Metadata can be used as the basis for any error-correction processing that should be done if a data error is found. Error-correction rules are documented in the metadata repository and executed by program code on a need basis.



DW Model:

- A Data Warehouse setup typically comprises of the following end points:
  - **Developer:** The developer puts business rules for data transformation into the metadata repository.
  - **Metadata:** It indicates about the data is available in the warehouse and where the data is located.
  - **Data Warehouse:** Data Warehouse integrates and aggregates data from various operational and external databases maintained by different Business Units.
  - **Operational Sources:** It can comprise of Customer Database, Sales Database, and Product Database.
  - **End User:** High performance is achieved by pre-planning the requirement for joins, summations, and periodic reports by end users.

Summary

- In this lesson, you have learnt:
  - ETL Process
  - Metadata used in ETL
  - Metadata in Data Warehousing
  - Simple Warehouse Model

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13

Review Questions

- Question 1: Metadata contains the following:
  - Option 1: Data Dictionary
  - Option 2: Data Flow
  - Option 3: Data Mart

- Question 2: Multidimensional data represents business complexities
  - True/ False

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 14

Review Question: Match the Following

1. Puts business rules	A. Data dictionary
2. Product of one or more fact tables	B. Measure
3. Direction and frequency of data feed	C. End user
	D. Data flow
	E. Developer

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 15

Data Warehousing Concepts

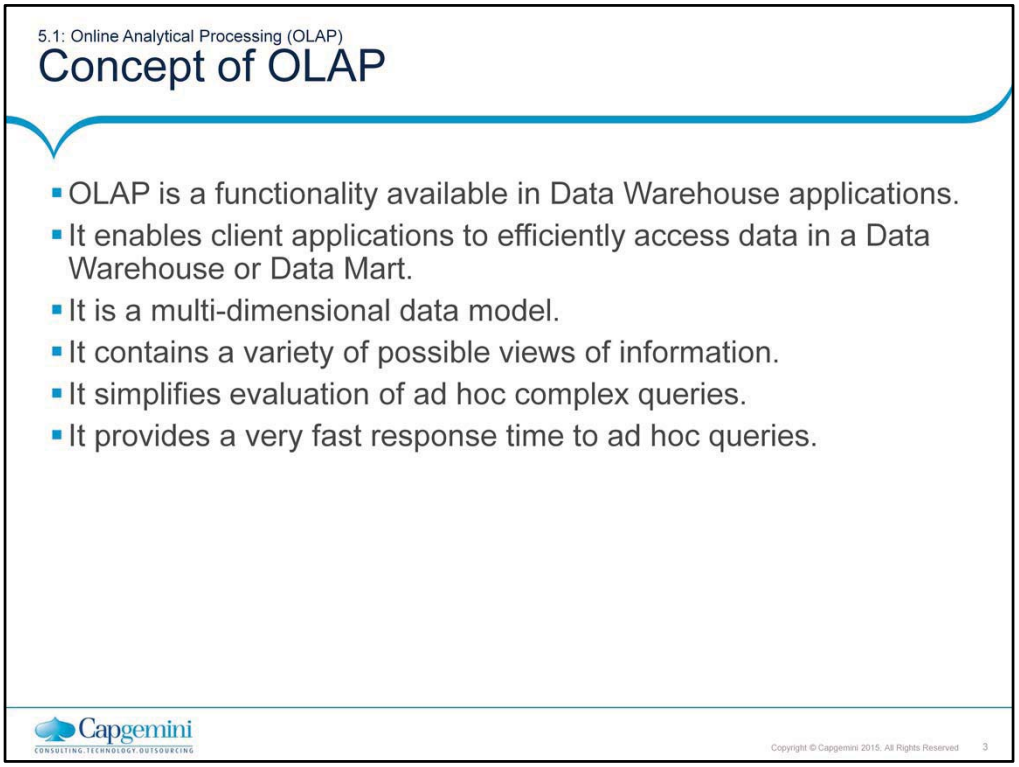
Lesson 5: Online Analytical Processing (OLAP)

Lesson Objectives

- In this lesson, you will learn about:
  - The concept of Online Analytical Processing
  - Need for Separate Operational and Informational Systems
  - Nature of OLAP analysis
  - Types Of OLAP
  - OLAP Service Tools
  - OLTP and OLAP
  - Operational versus Informational Systems

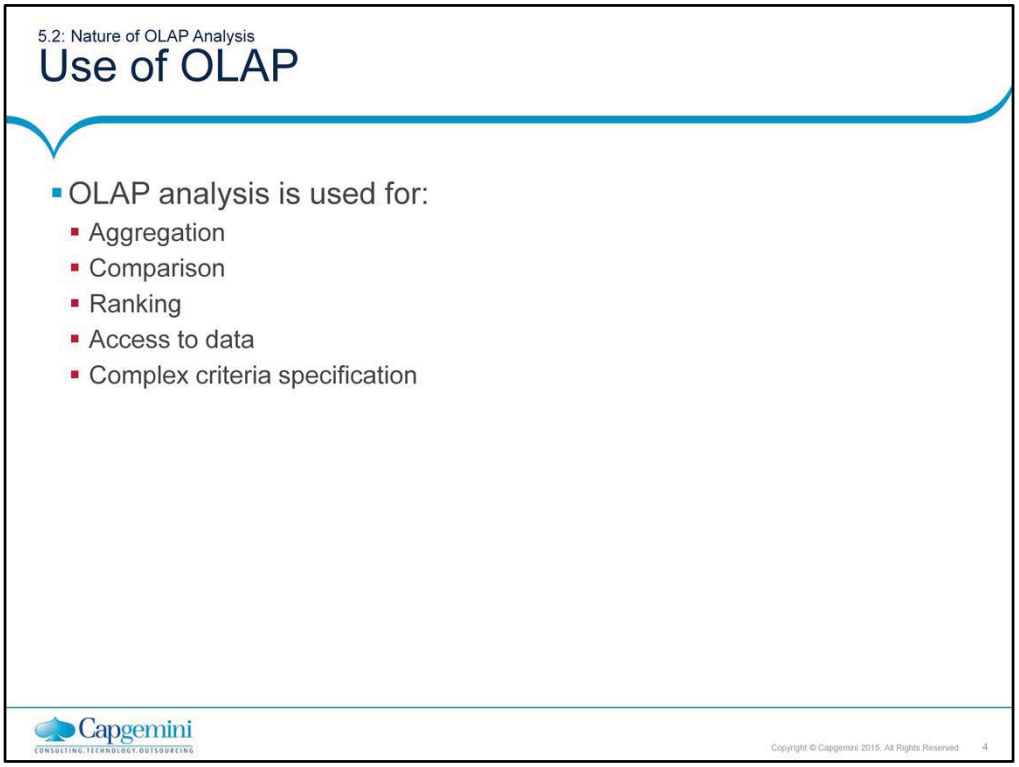
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2



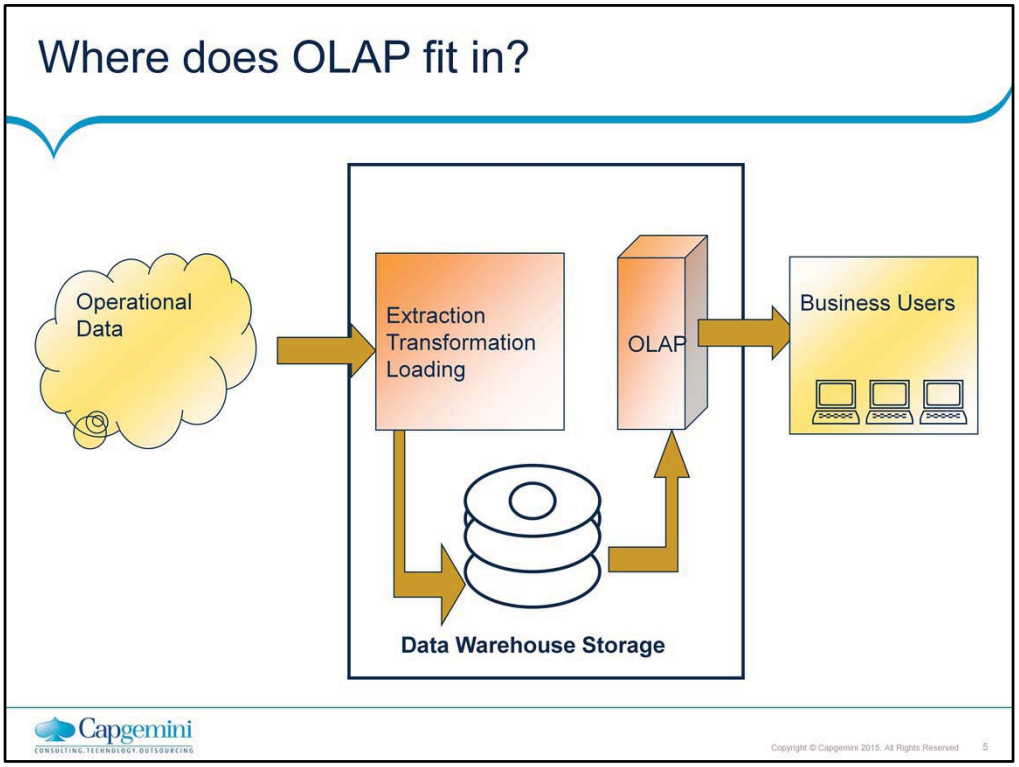
Online Analytical Processing (OLAP):

- **OLAP** is a category of software technology. It enables the users to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information, which has been transformed from raw data, to reflect the real dimensionality of the business.
- It provides benefits like **pre-aggregation** of frequently required data, enabling a very fast response time to ad hoc queries. It gives a **multi-dimensional data model** that makes it easy to select, navigate, and explore the data.
- OLAP systems enable managers and analysts to rapidly and easily examine key performance data. OLAP systems allow comparison and trend analysis even on very large volumes. OLAP allows users to view data from various perspectives. It is fast and easy because some aggregations are computed in advance.



<u>Nature of OLAP Analysis</u>:

- The nature of OLAP analysis varies with the multiple ways of using it.
  - It is used as the aggregation for summing up the data.
  - It provides the easy way of comparison.
  - It allows us to rank the data such that you will be able to find the top most and lower most values in the analysis.
  - It also helps in accessing the data in detailed way.
  - It allows to perform complex specification on the criteria.
  - It represents the data in a more simpler way such that it is easily visualized in terms of graphical presentation.
- OLAP Analysis is used for:
  - Aggregation: (total sales, percent-to-total)
  - Comparison: Budget versus Expenses
  - Ranking: Top 10, quartile analysis
  - Access to detailed and aggregate data
  - Complex criteria specification



Nature of OLAP Analysis:
Where does OLAP fit in?

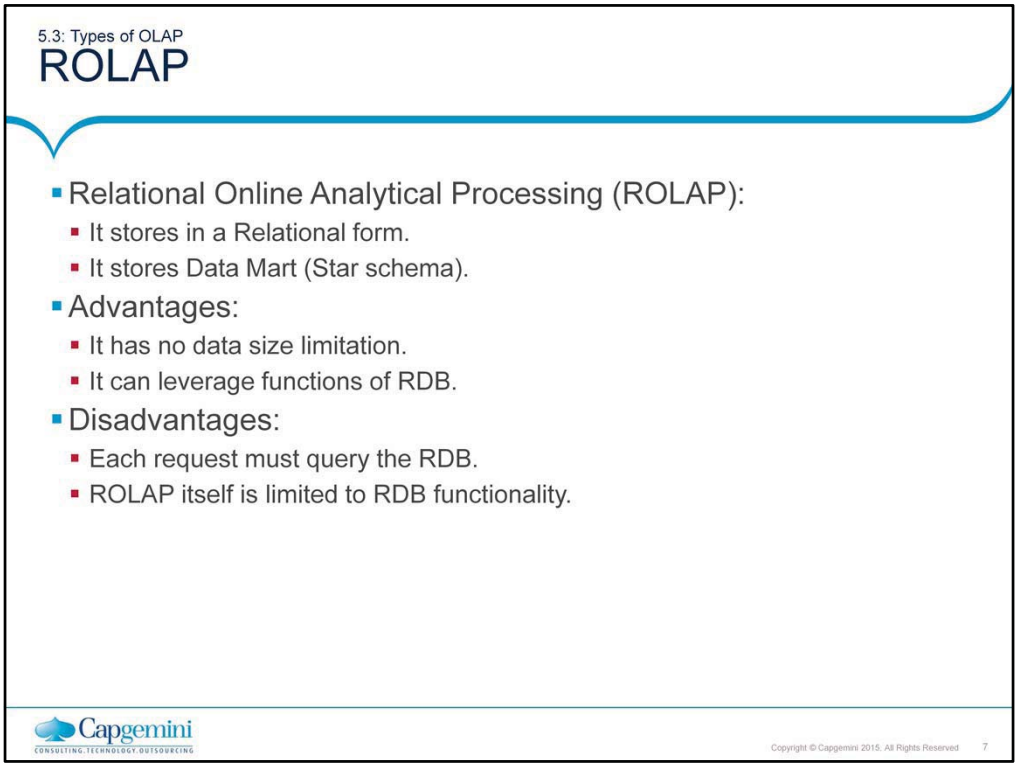
- Data from various sources goes through the ETL process and is integrated into a Data Warehouse. Subsequently, OLAP is used to analyze the data in the Data Warehouse. OLAP focuses on meeting end-user's analytical requirements.
  - **Operational Data:** It is the Customer Database, for example, Sales Database and Product Database.
  - **End User:** High performance is achieved by pre-planning the requirement for joins, summations and periodic reports by end users.
  - **Extract Transform and Load (ETL):**
    - **Extract:** It extracts data from data source and keeps it in staging.
    - **Transform:** It converts data into format required by Data Warehouse.
    - **Load:** It loads data to Data Warehouse.
  - **DWH:** Data Warehouse integrates and aggregates data from various operational and external data bases maintained by different Business Units.
  - **OLAP:** It has been in use to process and record transactions that create new data and update existing information in databases.
  - **Business User:** High performance is achieved by pre-planning the requirement and putting business rules by business users.

OLAP Models

- OLAP models are of different types
- The processing in all these different types is the same:
  - Online Analytical processing
- The storage methods are different in different models
- Different OLAP Models are
  - ROLAP
  - MOLAP
  - HOLAP

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 6



<u>Types of OLAP</u>:

Relational Online Analytical Processing (ROLAP):

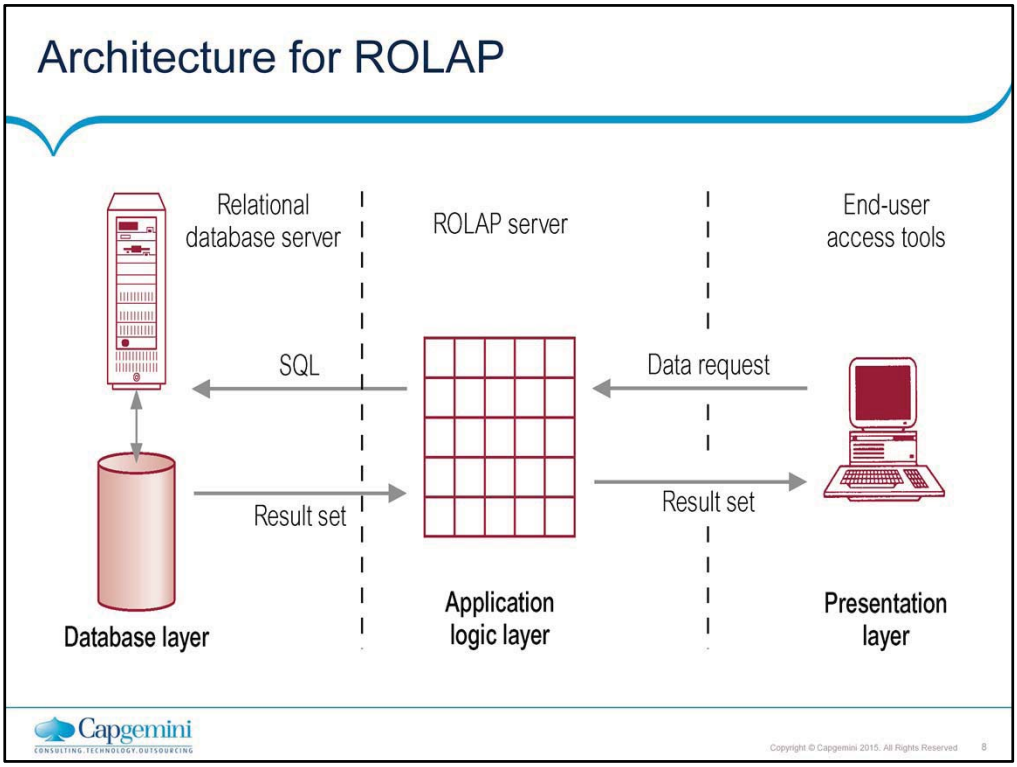
- The data stored in the relational database gives the appearance of traditional OLAP's slicing and dicing functionality.
- In Relational OLAP (ROLAP), Relational DBMS stores Data Mart (Star schema).

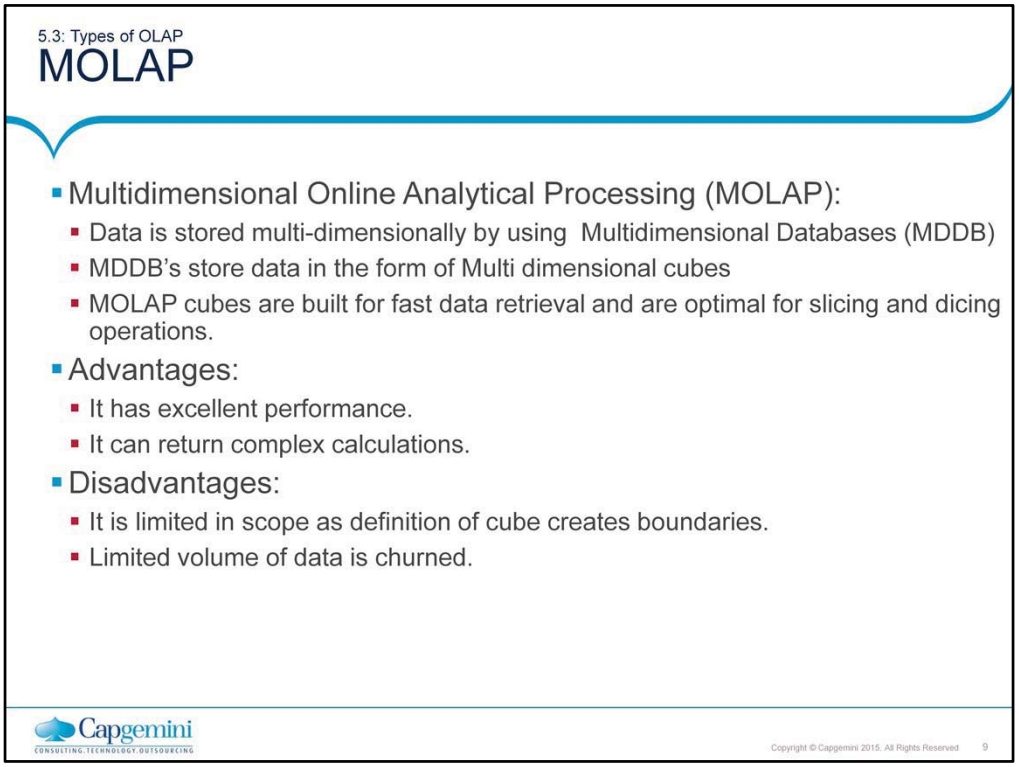
Advantage:

- ROLAP itself places no limitation on data amount.
- Relational database already comes with a host of functionalities. ROLAP technologies, can leverage these functionalities since they sit on top of the relational database.

Disadvantage:

- Each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database. The query time can be long if the underlying data size is large.
- ROLAP itself is limited to RDB functionality.


Architecture for ROLAP
Relational database server
ROLAP server
End-user access tools
SQL
Data request
Result set
Result set
Database layer
Application logic layer
Presentation layer
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 8



Types of OLAP:

Multidimensional Online Analytical Processing (MOLAP):

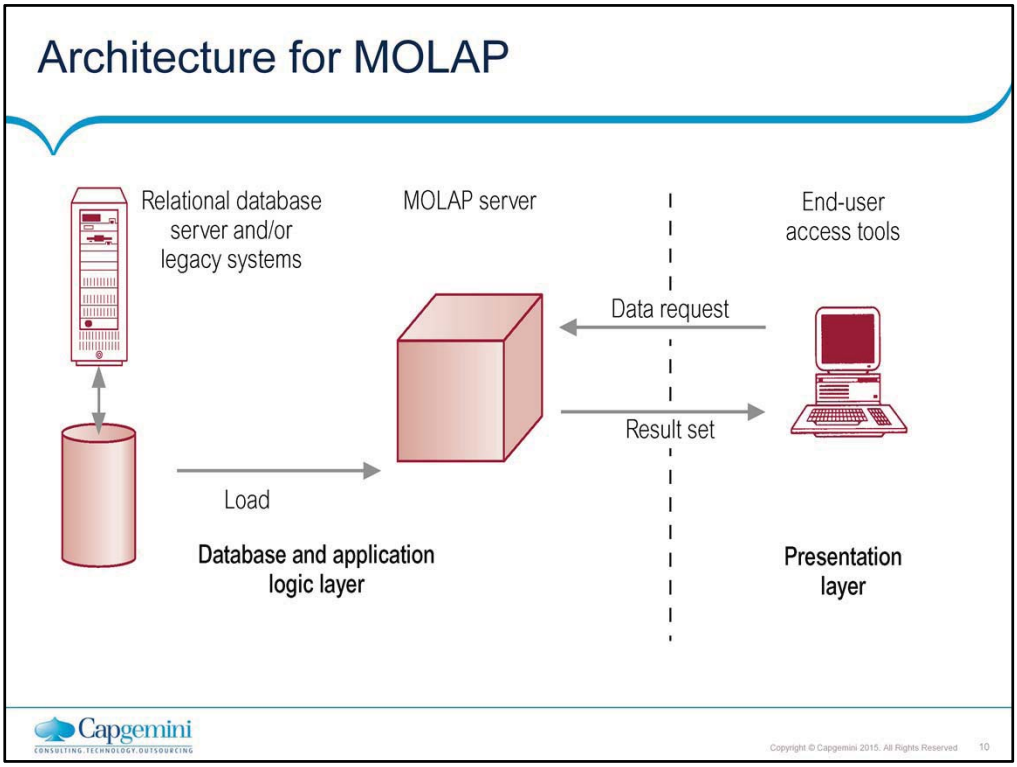
- In MOLAP, the data is stored in a multi-dimensional cube. The storage is not in the relational database, but in proprietary formats.
- MOLAP storage structure is Array-based.

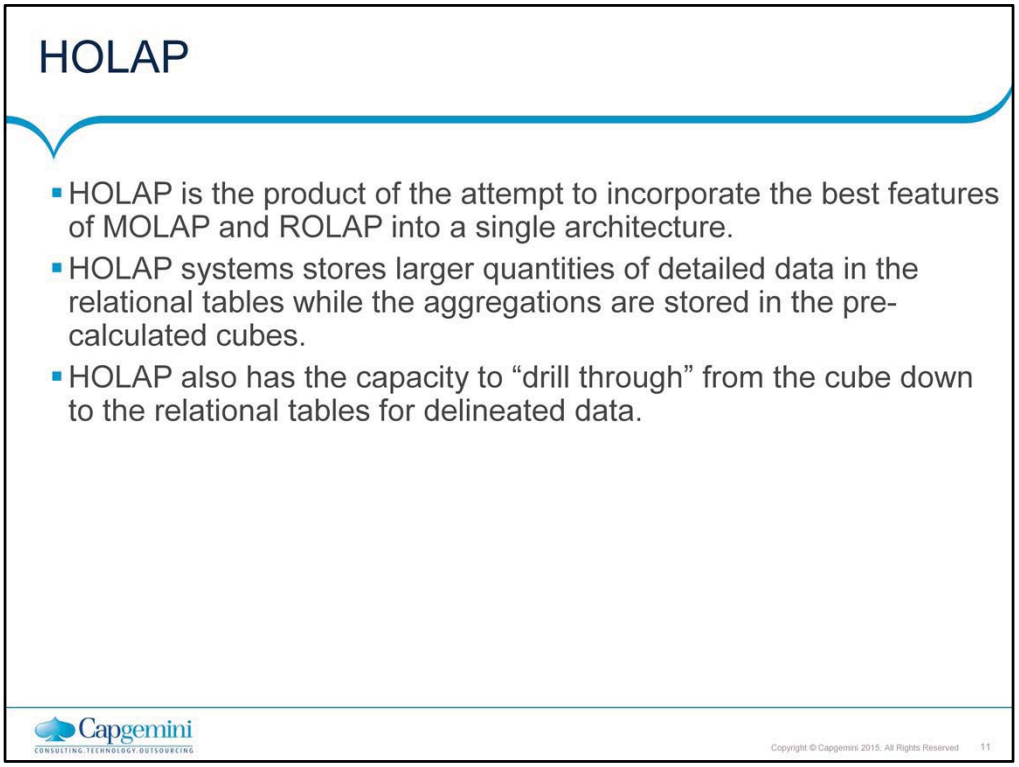
Advantage:

- MOLAP cubes are built for fast data retrieval, and is optimal for slicing and dicing operations.
- All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, they return quickly, as well.

Disadvantage:

- In case of MOLAP, scope is limited as definition of cube creates boundaries.
- It is not possible to include a large amount of data in the cube itself. This is because all calculations are performed when the cube is built. Only summary-level information will be included in the cube itself.

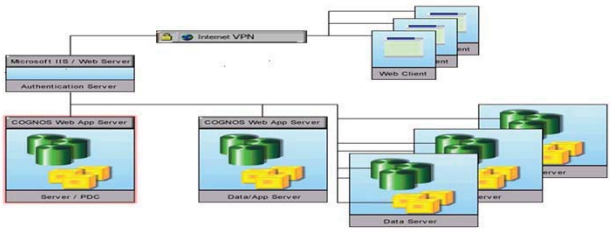

Architecture for MOLAP
Relational database server and/or legacy systems
MOLAP server
End-user access tools
Data request
Result set
Load
Database and application logic layer
Presentation layer
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved.
10



This tool tried to bridge the technology gap of both products by enabling access or use to both multidimensional database (MDDB) and Relational Database Management System (RDBMS) data stores.

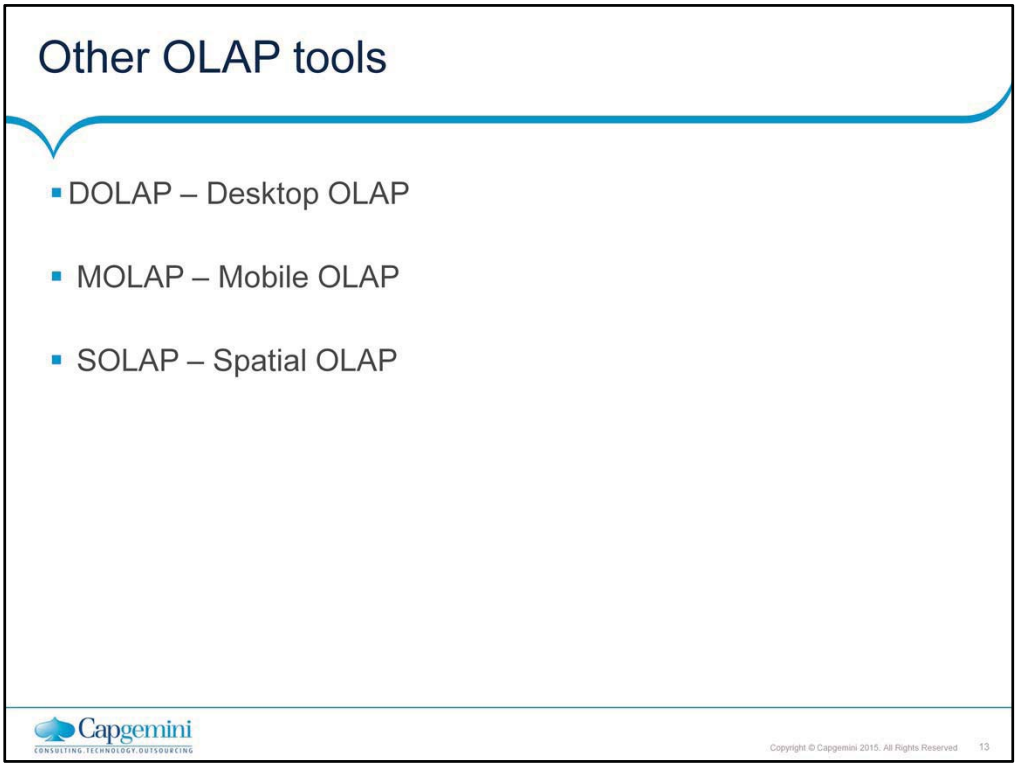
WOLAP Architecture

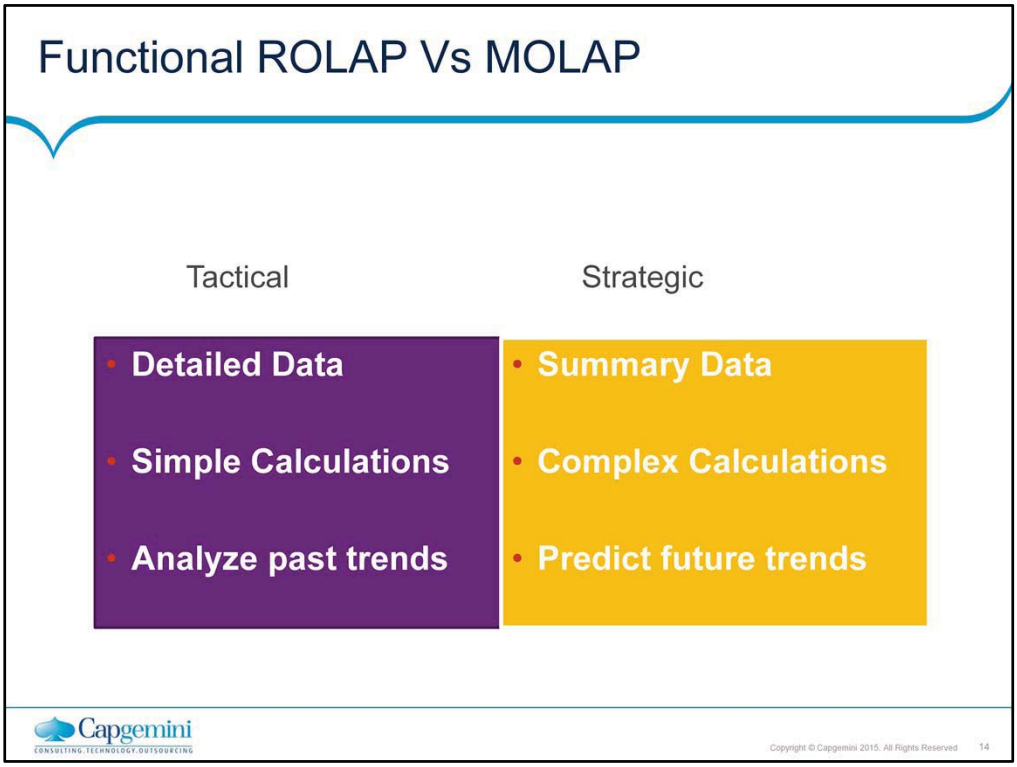
- Web Based Online Analytical Processing: WOLAP, or Web-enabled OLAP, which uses a browser to deliver OLAP. The combination is powerful, say many business managers. The delivery capability of the Web, coupled with the business intelligence tool of OLAP, will allow a broader number of business analysts to benefit from the software.
- Example :

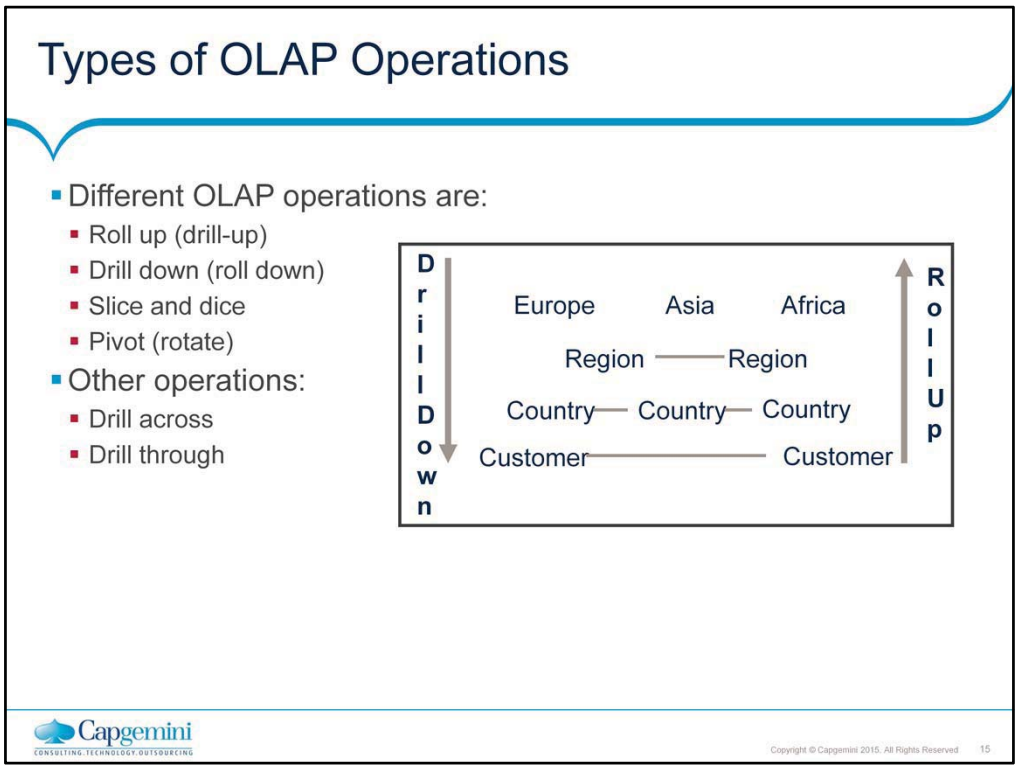


Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 12


Other OLAP tools
DOLAP – Desktop OLAP
MOLAP – Mobile OLAP
SOLAP – Spatial OLAP
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 13


Functional ROLAP Vs MOLAP
Tactical
Strategic
Detailed Data
Simple Calculations
Analyze past trends
Summary Data
Complex Calculations
Predict future trends
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved. 14

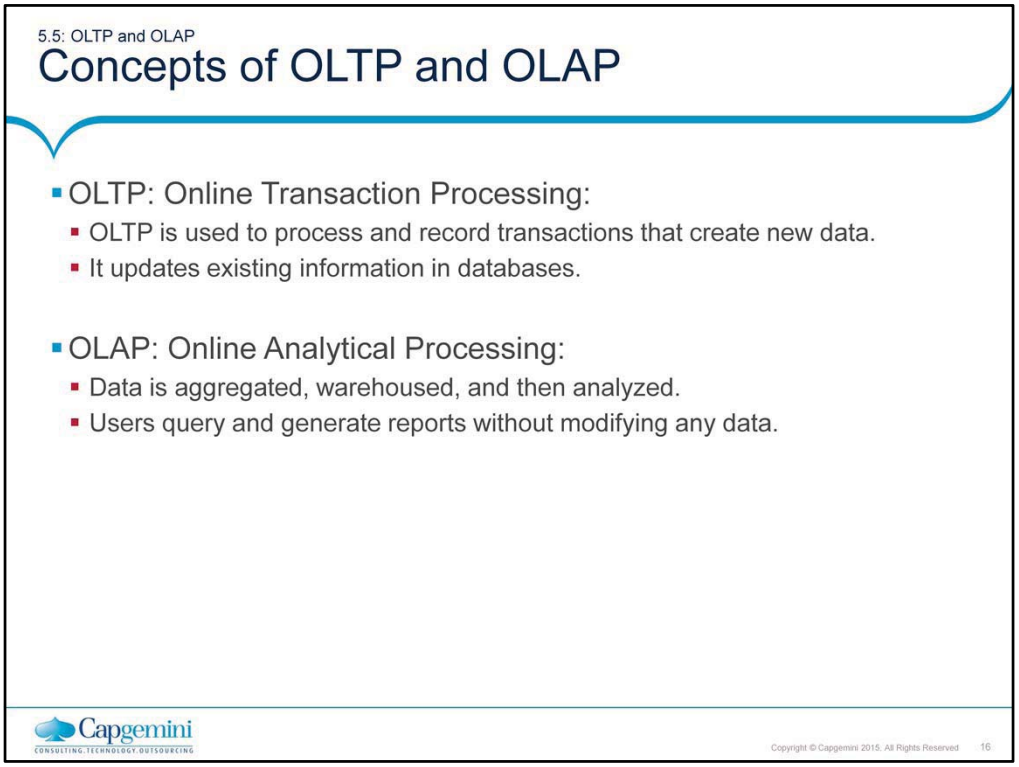


Types Of OLAP Operations:

- **Aggregation / Consolidation / Roll up (drill-up):** Roll Up operation is used to summarize data by climbing up hierarchy or by dimension reduction. This allows you to aggregate data from lower level details to the parent level. For example, the total revenue generated by a particular product type will be the rolled up value of the revenue generated by descendants, that is, products that belong to that particular product type.
- **Drill down (roll down):** Roll down operation is the reverse of roll-up, that is, a drill down from higher level summary to lower level summary or detailed data, or introducing new dimensions. It allows you to view data from a top-level to a detailed view by going down the hierarchy. For example, we can view the Sales data for the Year and drill down from the year level to a quarterly view and further down to a monthly view.
- **Slice and dice:** It is a general term for viewing data from any angle.
- **Pivot (rotate):** Rotate operation is used for reorienting the cube, visualization, 3D to series of 2D planes.

Other operations:

- **Drill Across:** It involves (across) more than one fact table.
- **Drill Through:** It involves drilling through the bottom level of the cube to its back-end relational.



OLTP and OLAP:

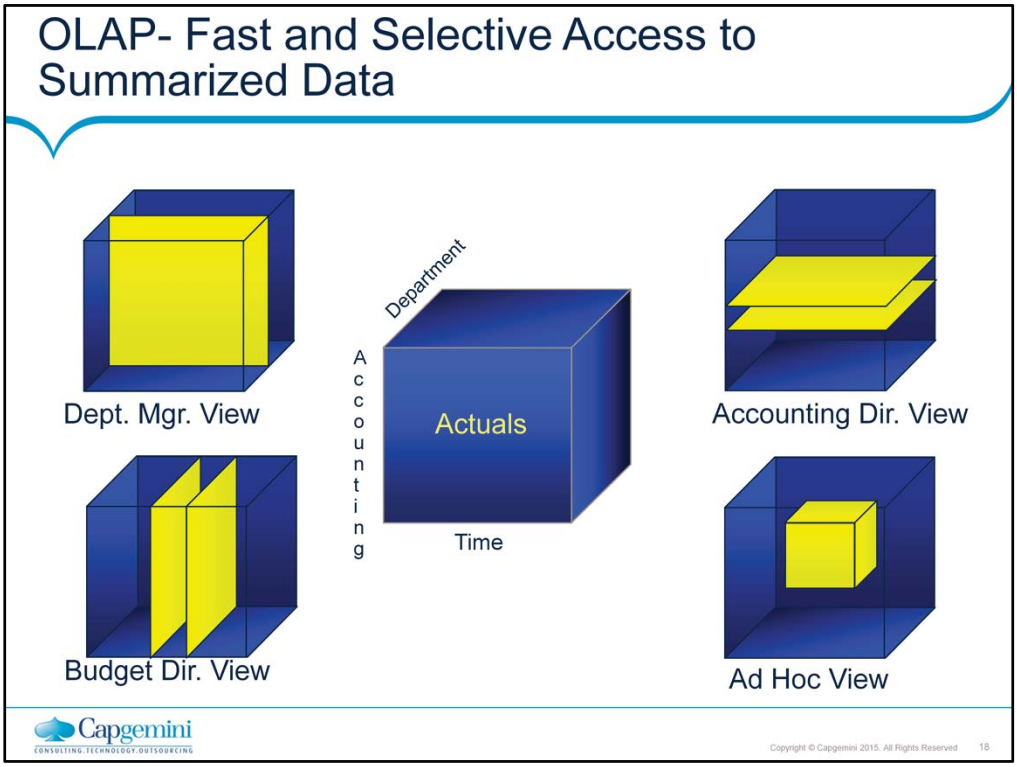
- OLTP is basically operational data, wherein data is frequently changing.
- For example, you can consider an online Railway Reservation system. A passenger books a ticket for two people. This becomes an operational data. S/He can also change the number of passengers travelling online since OLTP data is frequently updated.
- On the other hand, OLAP data is non-operational data, wherein data is read-only data. It used for analytical purpose.
- For example, suppose one wants to see, the number of trains running on the previous day. This becomes analysis of data that is not updatable.

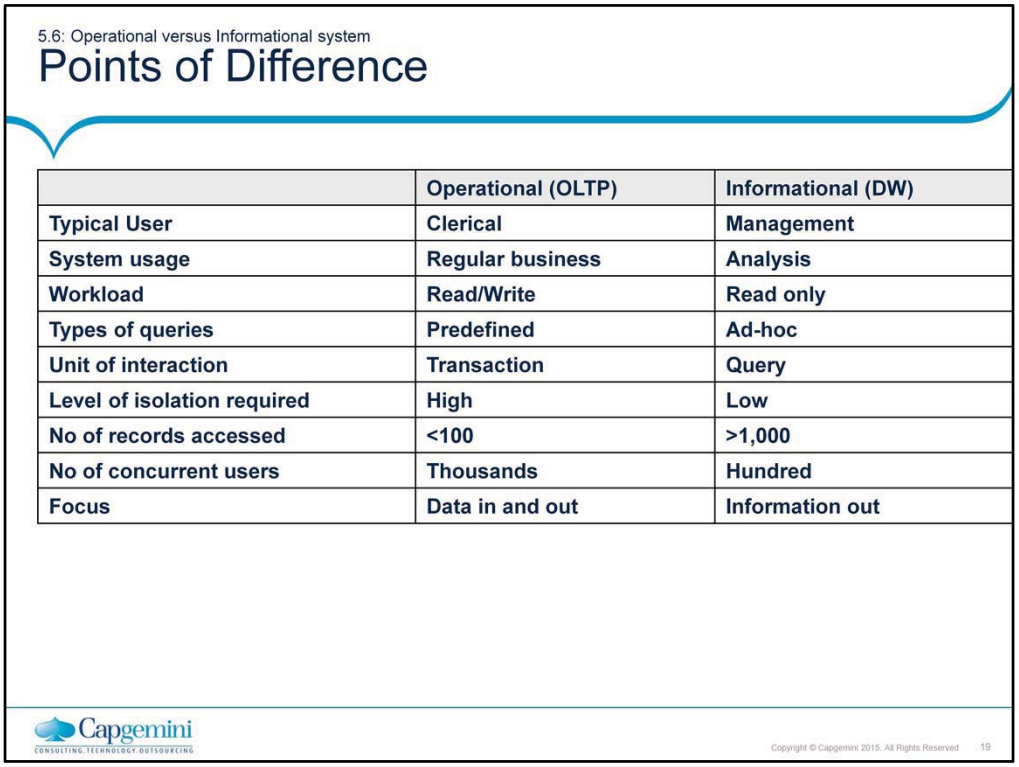
OLAP Functional Requirements

- Fast Access and Calculations
  - Speed is critical to maintain an analyst's train of thought.
  - An analyst needs to navigate throughout the data which requires aggregations, or roll-ups.
- Powerful Analytical Capabilities
  - There is more complicated calculations to OLAP than simple aggregations, or roll-ups.
- Flexibility
  - Viewing: graphs, charts, row or columns
  - Definitions: format of numbers, name changes
  - Analysis: Sales analyze data differently than marketing
  - Interfaces: section wise, report looks

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

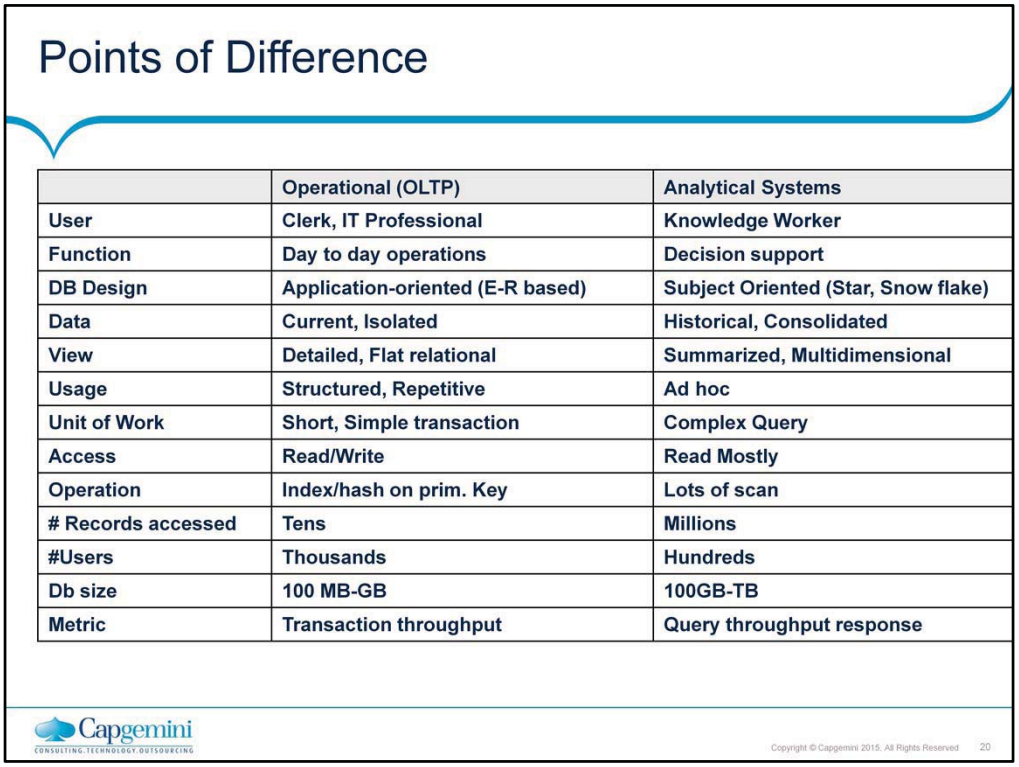
 17


OLAP- Fast and Selective Access to Summarized Data
Dept. Mgr. View
Department
Accounting
Actuals
Time
Accounting Dir. View
Budget Dir. View
Ad Hoc View
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
18



	Operational (OLTP)	Informational (DW)
Typical User	Clerical	Management
System usage	Regular business	Analysis
Workload	Read/Write	Read only
Types of queries	Predefined	Ad-hoc
Unit of interaction	Transaction	Query
Level of isolation required	High	Low
No of records accessed	<100	>1,000
No of concurrent users	Thousands	Hundred
Focus	Data in and out	Information out

Note: The above table shows the comparison between **OLTP** and **Data Warehouse**.



Points of Difference

	Operational (OLTP)	Analytical Systems
User	Clerk, IT Professional	Knowledge Worker
Function	Day to day operations	Decision support
DB Design	Application-oriented (E-R based)	Subject Oriented (Star, Snow flake)
Data	Current, Isolated	Historical, Consolidated
View	Detailed, Flat relational	Summarized, Multidimensional
Usage	Structured, Repetitive	Ad hoc
Unit of Work	Short, Simple transaction	Complex Query
Access	Read/Write	Read Mostly
Operation	Index/hash on prim. Key	Lots of scan
# Records accessed	Tens	Millions
#Users	Thousands	Hundreds
Db size	100 MB-GB	100GB-TB
Metric	Transaction throughput	Query throughput response

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 20

Note: Above table depicts the comparison between **Operational** and **Analytical systems**.

Summary

- In this lesson, you have learnt:
  - OLAP allows users to view data from various perspectives.
  - The nature of OLAP analysis varies in multiple ways of using it.
    - Aggregation
    - Comparison
    - Ranking
    - Access to data
    - Complex criteria specification

Summary

- OLAP is used to analyze the data in the Data Warehouse.
- Different types of OLAP are:
  - MOALP
  - HOLAP
  - ROLAP

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 22

Review Question

- Question 1: OLAP analysis is used for:
  - Option 1: Retrieving data
  - Option 2: Updating data
  - Option 3: Summarizing data
- Question 2: OLAP makes use of multidimensional data model.
  - True/ False
- Question 3: ___ OLAP operation helps for viewing data from any angle.

Data Warehousing Concepts

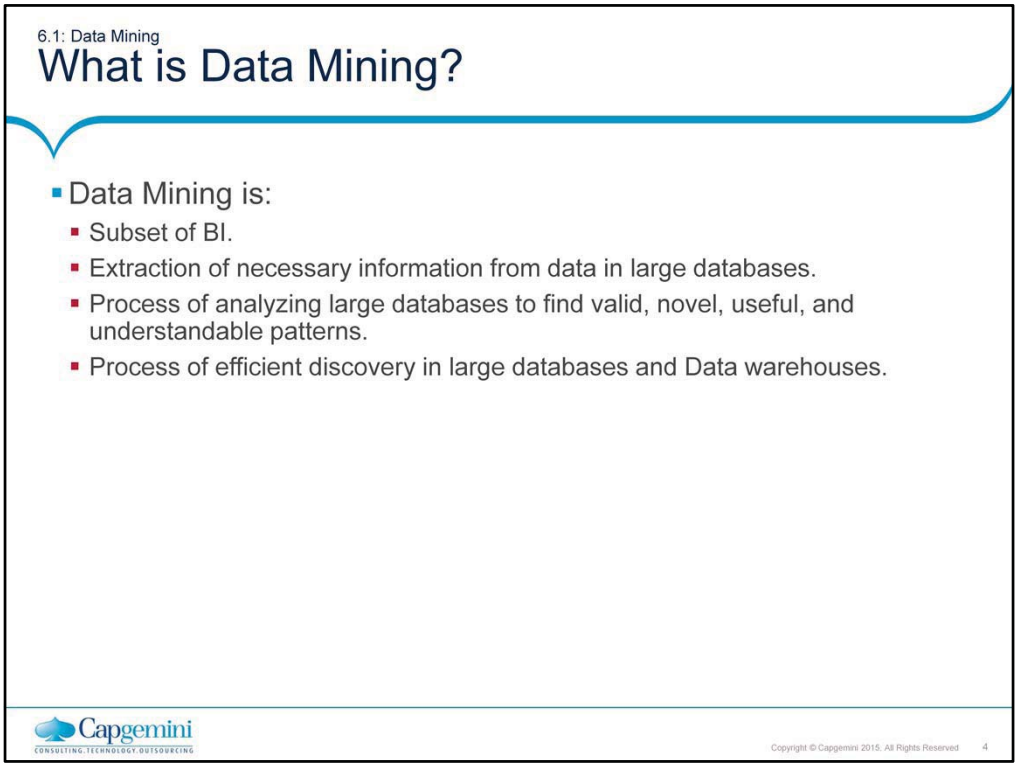
Lesson 6: Data Mining

Lesson Objectives

- In this lesson, you will learn about:
  - Online Analytical Processing
  - Data Mining
  - The Knowledge Discovery Process
  - Why Use Data Mining Today?
  - Data Mining Usage
  - Data Mining and Business Intelligence
  - Types of Data used in Data Mining
  - Data Mining Applications

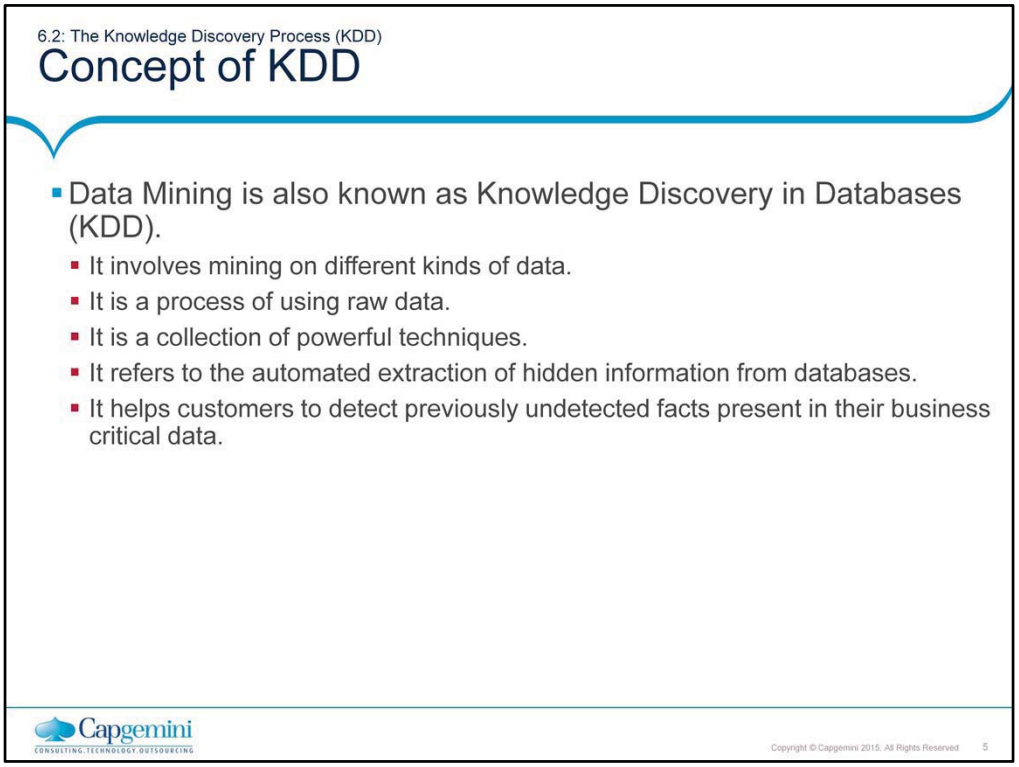
Lesson Objectives

- In this lesson, you will learn about (contd.):
  - Data Mining Products
  - Mining market



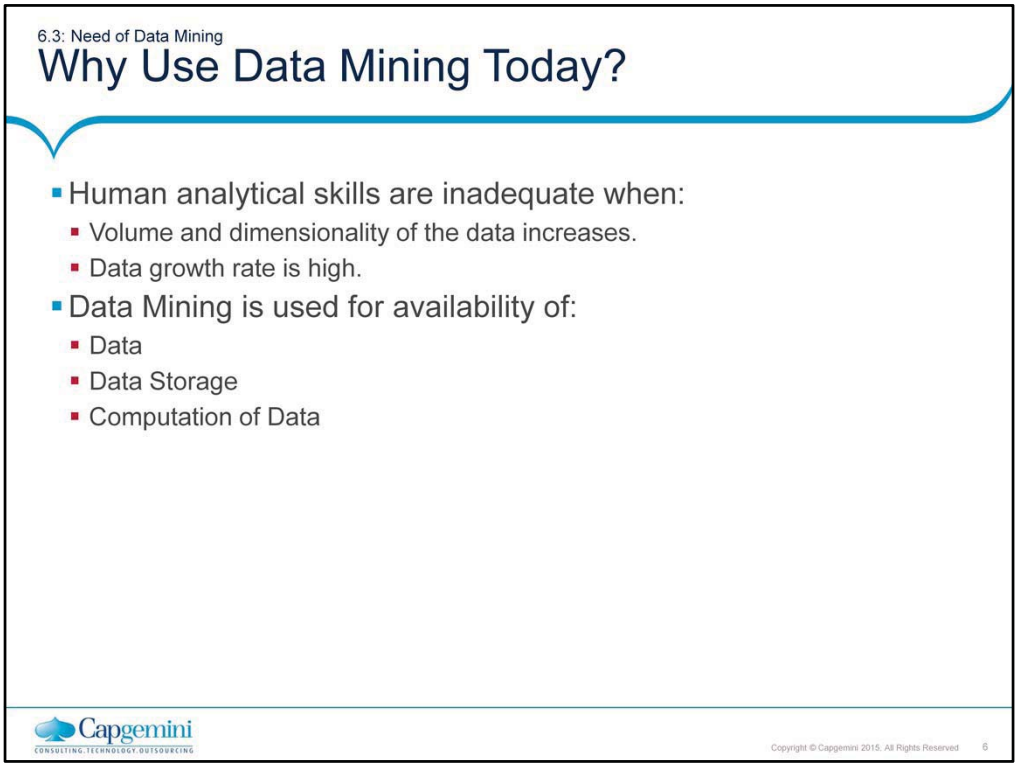
Data Mining:

- **Data mining** is the way of analyzing data by exploring large databases. It helps in understanding the business by extracting necessary information from the databases. It allows you to understand the pattern and helps in predicting the behavior of it.
- Data mining helps in increasing the business and forecasting the chunks related to it at early stages. It includes finding patterns that are suitable for the organization.



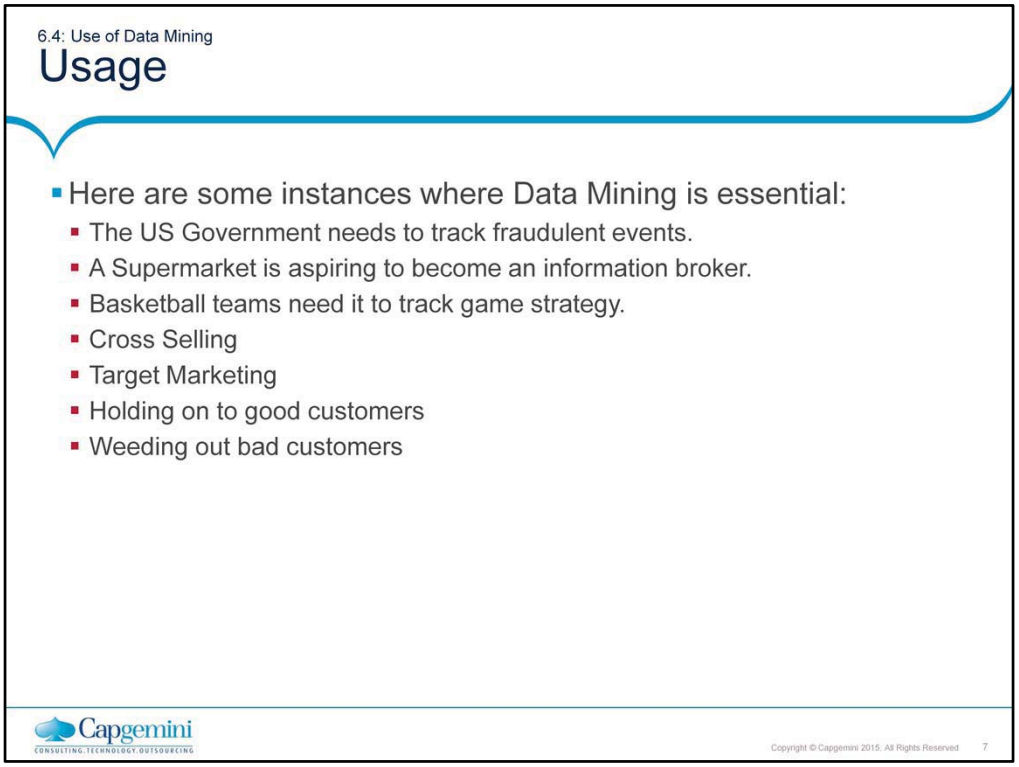
The Knowledge Discovery Process (KDD):

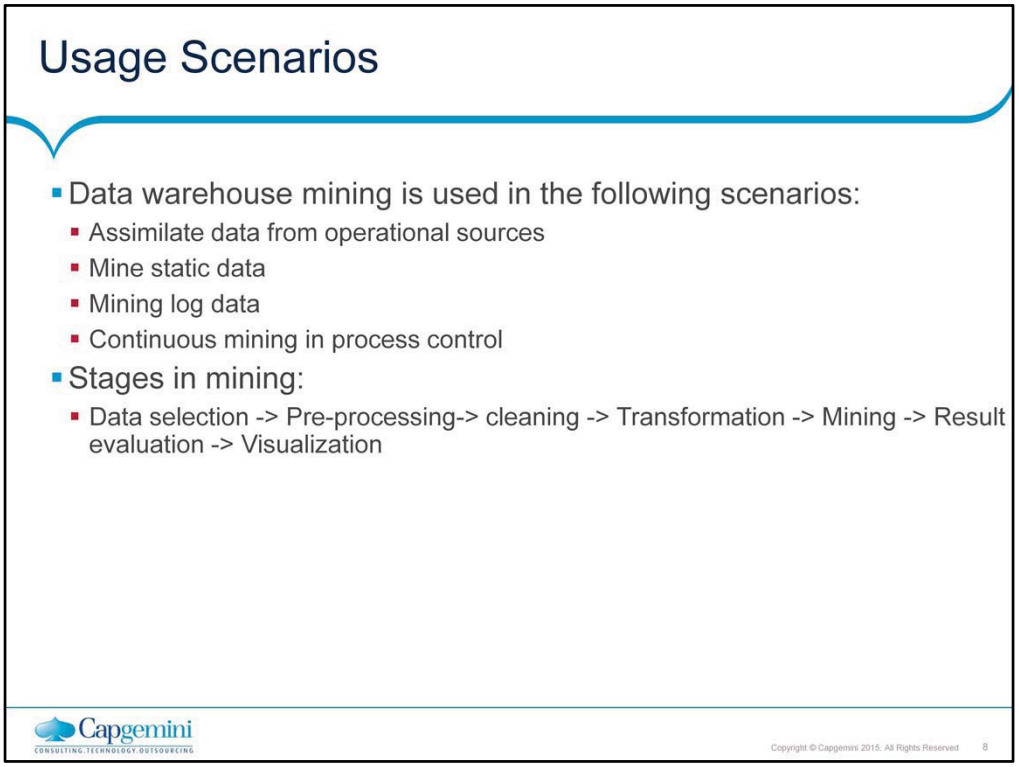
- **Data Mining** involves mining on different kinds of data such as Relational databases, Data warehouses,Transactional databases, Advanced DB systems and information repositories, Object-oriented and object-based databases, Text databases and multimedia databases, Heterogeneous and legacy databases. Data mining is the process of using raw data to infer important business information. It is a collection of powerful techniques for analyzing large amounts of data. Data mining tools can access data directly in the Data Warehouse.
- The advantage of mining is that no separate copy of data is needed for data mining. Data may not be organized in a way that is efficient for the tool. Data Mining is done by running a software that examines a database and looks for patterns in the data. Data Mining will not tell users about patterns in data that users may not have thought about. Data mining is used to try and mine key information from a Data warehouse to find **patterns in data**. Data mining allows organizations to collect information and make themselves more productive and beat their competitors.



Need for Data Mining:

- Data mining is essential because of the following utilities:
  - Data mining helps to identify why customers buy certain products.
  - Data mining provides the ideas for very direct marketing.
  - Data mining provides the ideas for shelf placement.
  - It helps for training of employees versus employee retention.
  - It helps to identify employee benefits.


6.4: Use of Data Mining
Usage
Here are some instances where Data Mining is essential:
The US Government needs to track fraudulent events.
A Supermarket is aspiring to become an information broker.
Basketball teams need it to track game strategy.
Cross Selling
Target Marketing
Holding on to good customers
Weeding out bad customers
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
7



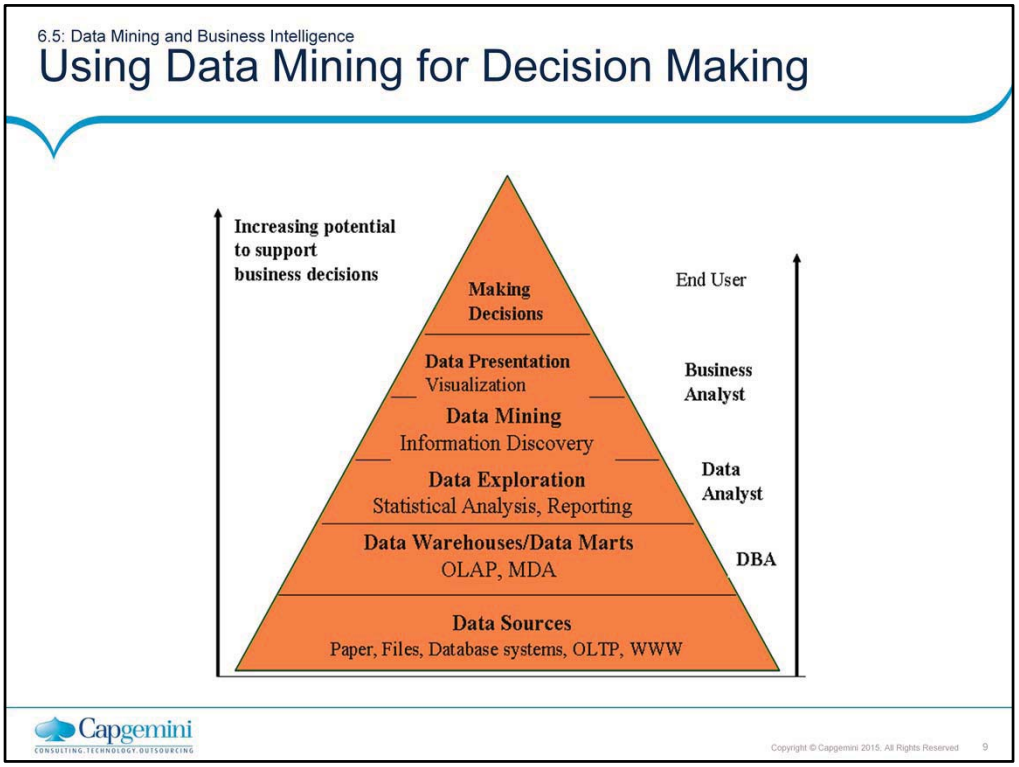
Use of Data Mining:

Usage scenarios:

- Data warehouse mining assimilates data from operational sources.
- Data warehouse mining mines static data.
- Mining log data.
- Continuous mining in process control.

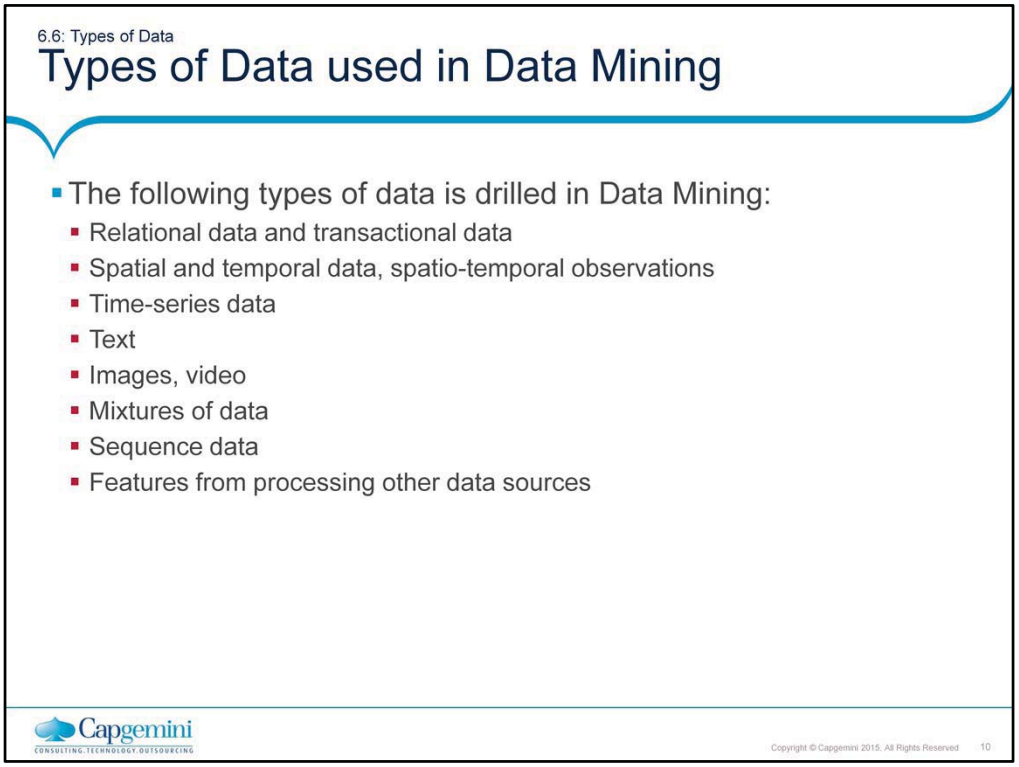
Stages in mining:

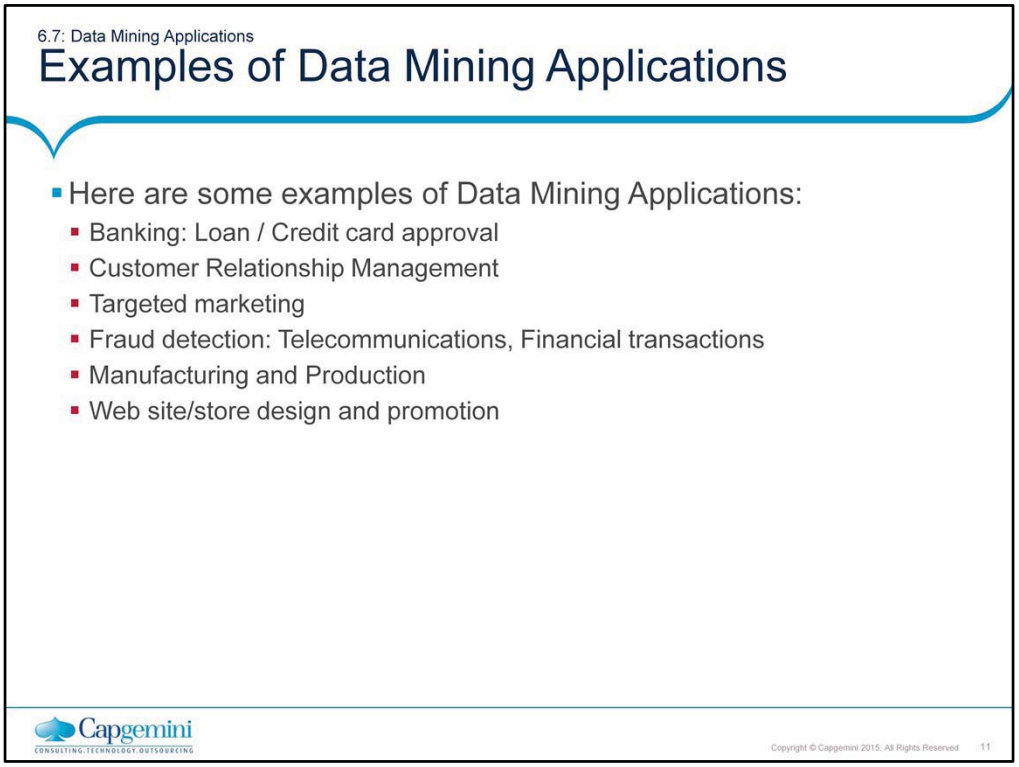
1. Data selection
2. Pre-processing: cleaning
3. Transformation
4. Mining
5. Result evaluation
6. Visualization



Data Mining and Business Intelligence:

- Data Mining has grown drastically in many businesses. Data Mining has become very popular since it helps in increasing organization's profit and achieving the target.
- When Data mining gets involved in Business Intelligence, it actually helps in understanding the functionality of the organization. It helps in increasing the potential for supporting the business decisions. It makes the data visible in a visual form to the business analysts. It helps in exploring data in terms of reporting and statistical analysis.
- Data Mining along with Business Intelligence takes the following steps in logical progression:
  - **Data Source:** Typically data is sourced from transaction processing systems (Manufacturing, ERP, Sales).
  - **Data Marts (OLAP & MDA)/DBA:** You may want to customize your warehouse's architecture for different groups within your organization. You can do this by adding data marts, which are systems designed for a particular line of business.
  - **End User/Making Decision:** The principle purpose of Data warehousing is to provide information to the business user for strategic decision making.

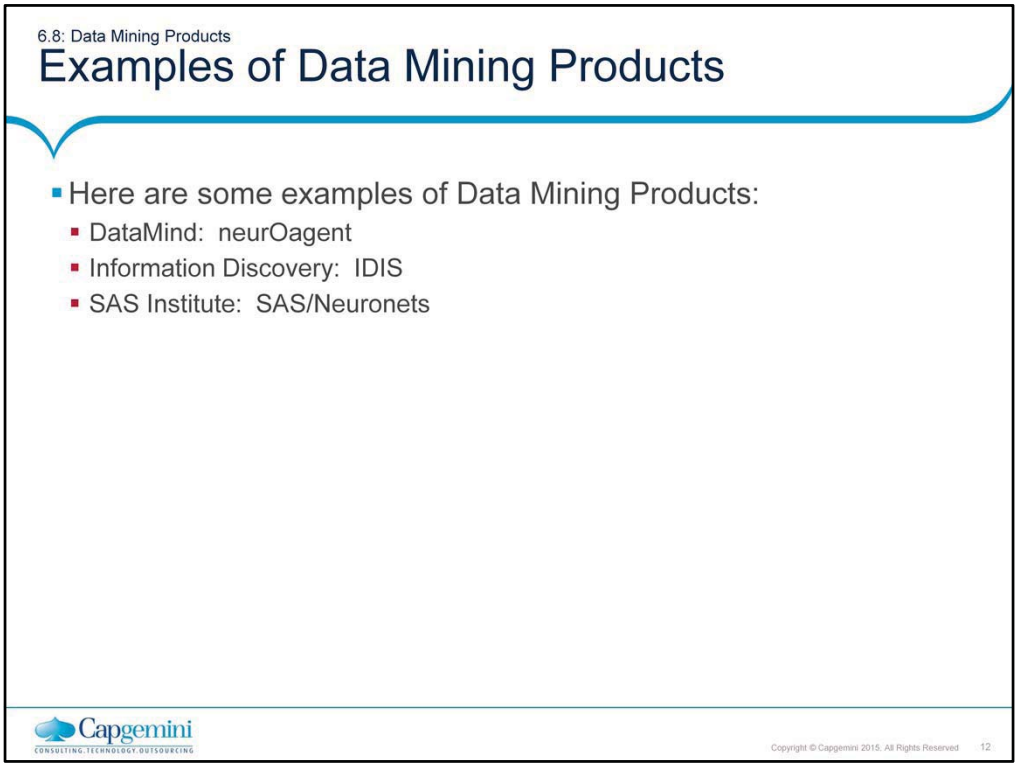

6.6: Types of Data
Types of Data used in Data Mining
The following types of data is drilled in Data Mining:
Relational data and transactional data
Spatial and temporal data, spatio-temporal observations
Time-series data
Text
Images, video
Mixtures of data
Sequence data
Features from processing other data sources
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
10



Data Mining Applications:

Let us discuss some examples of Data Mining Applications:

- Banking: loan/credit card approval:
  - Predict good customers based on old customers
- Customer Relationship Management:
  - Identify those who are likely to leave for a competitor.
- Targeted marketing:
  - Identify likely responders to promotions.
- Fraud detection: Telecommunications, Financial transactions
  - From an online stream of events, identify fraudulent events.
- Manufacturing and production:
  - Automatically adjust knobs when process parameter changes.
- Web site/store design and promotion:
  - Find affinity of visitor to pages and modify layout.


6.8: Data Mining Products
Examples of Data Mining Products
Here are some examples of Data Mining Products:
DataMind: neurOagent
Information Discovery: IDIS
SAS Institute: SAS/Neuronets
Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING
Copyright © Capgemini 2015. All Rights Reserved
12

6.9: Data Mining Market

Mining Market and Vendors

- There are around 20 to 30 mining tool vendors.
- Major tool players:
  - Clementine
  - IBM's Intelligent Miner
  - SGI's MineSet
  - SAS's Enterprise Miner
- Many embedded products:
  - Fraud detection
  - Electronic commerce applications
  - Health care
  - Customer Relationship Management: Epiphany

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 13

Summary

- In this lesson, you have learnt:
  - Data Mining is the way of analyzing data by exploring the large databases.
  - Data Mining is used to mine key information from a data warehouse.
  - It helps in exploring data in terms of reporting and statistical analysis.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Review Question

- Question 1: Data exploration for statistical analysis is done by:
  - Option 1: DBA
  - Option 2: Business analyst
  - Option 3: Data analyst

- Question 2: Data Mining is a subset of DW.
  - True/ False
- Question 3: Mining is also known as ___.

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 15

Review - Match the Following

1. End user	A. Data mining
2. Business analyst	B. Data warehouse
3. Data analyst	C. Data presentation
	D. Making decisions
	E. Data exploration

Capgemini
CONSULTING. TECHNOLOGY. OUTSOURCING

 16

Data Warehousing Concepts

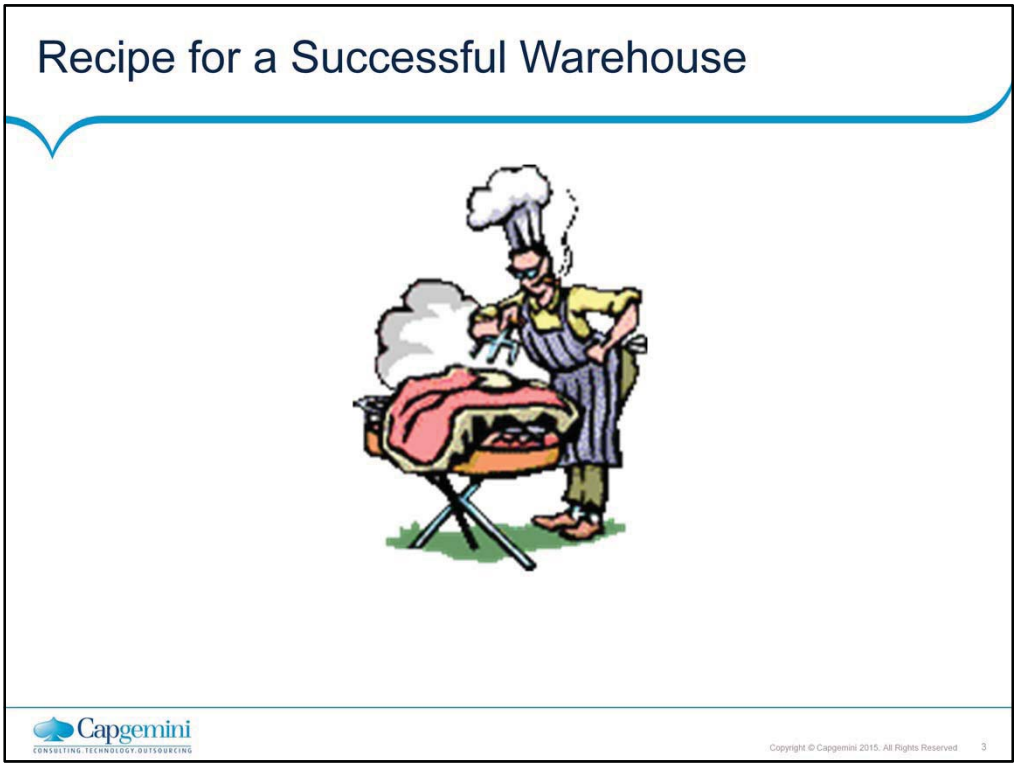
Lesson 7: Best Practices for Building Data Warehouse

Lesson Objectives

- In this lesson, you will learn:
  - Requirement for successful Data Warehouse
  - Data warehouse pitfalls
  - Popular BI DW tools and suits
  - Trends in BIDW

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 2



From day one establish that warehousing is a joint user/builder project
Most of the Warehouse projects will fail if the builders get specs from the users, go off for 6 months, and then come back with the 'finished' project. Warehouses are iterative! (Hear I put the word iterative means there are lots of mistakes in the projects.) Builders and users working with each other will not reduce the number of iterations, but it will reduce the size of them.

Establish that maintaining data quality will be an ONGOING joint user/builder responsibility
Organizations undertaking warehousing efforts almost continually discover data problems. Best to establish right up front that this project is going to require some additional ongoing responsibility.

Train the users one step at a time
Typically users are trained once. In several days they learn both the basics and intermediate and sometimes advanced aspects of using a tool. Slow down! Consider providing training initially in the minimum needed for the user to get something useful from the tool. Then let the user use the tool for a while (meaning several days, weeks, or months). Having basic training and some hands on experience, the user will have a much better context with which to grasp the next level. Also, once the basics and the next level are learned, keep training the users! After a year using the tool, schedule advanced training.

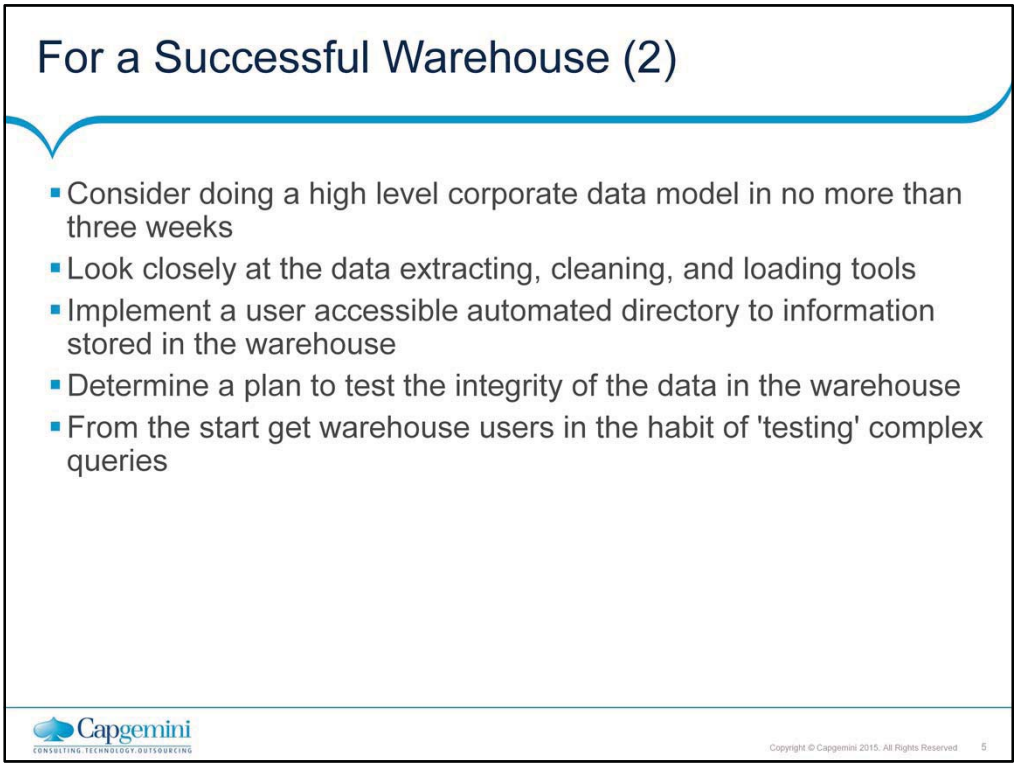
Train the users about the data stored in the data warehouse
Users often need more training about the stored data than about the tools used to access the data. One should not assume the data are self-explanatory or that any metadata you may provide will answer any questions. Note that users are often used to seeing data in canned reports and seeing data in its "raw" form can be confusing.

For a Successful Warehouse (1)

- From day one establish that warehousing is a joint user/builder project
- Establish that maintaining data quality will be an ONGOING joint user/builder responsibility
- Train the users one step at a time
- Train the users about the data stored in the data warehouse

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 4



Consider doing a high level corporate data model / data warehouse architecture "exercise" in three weeks
Actually, the key point regarding time is to "time-box" the exercise into a relatively short time. After about three weeks, the marginal benefits from additional time devoted to these types of exercises rapidly decrease. - The corporate model is going to identify, at a high level, subjects and relationships and most importantly, what are the chunks of information that it makes sense to deliver in different projects. The architecture part of the exercise to determine the dimensions, definitions of derived data, attribute names, and information sources that you will attempt to use consistently in your data warehousing efforts. The exercise also consists of coming to an agreement as to how to keep the corporate model up-to-date and how to make sure future data warehousing efforts pay attention to the architectural principles.

Implement a user accessible automated directory to information stored in the warehouse

The majority of successful warehousing efforts I have seen included providing some means for the warehouse user to locate stored information. Most of the times this involved building a separate database with directory information. And most of the time, a pretty simple database sufficed for initial use.

Once you know what raw data you want to feed into the data, request that data

If you have done some reading on data warehouse development you probably have read that figuring out the process of extracting, transforming, and loading (ETL) usually takes the majority of the time in initial data warehouse development. In project management lingo, figuring out ETL is usually on the critical path. - If you know what raw data you need, request it as soon as you know it. You are probably going to have to ask one of the programmers of the legacy feeder systems to initially get this data for you.

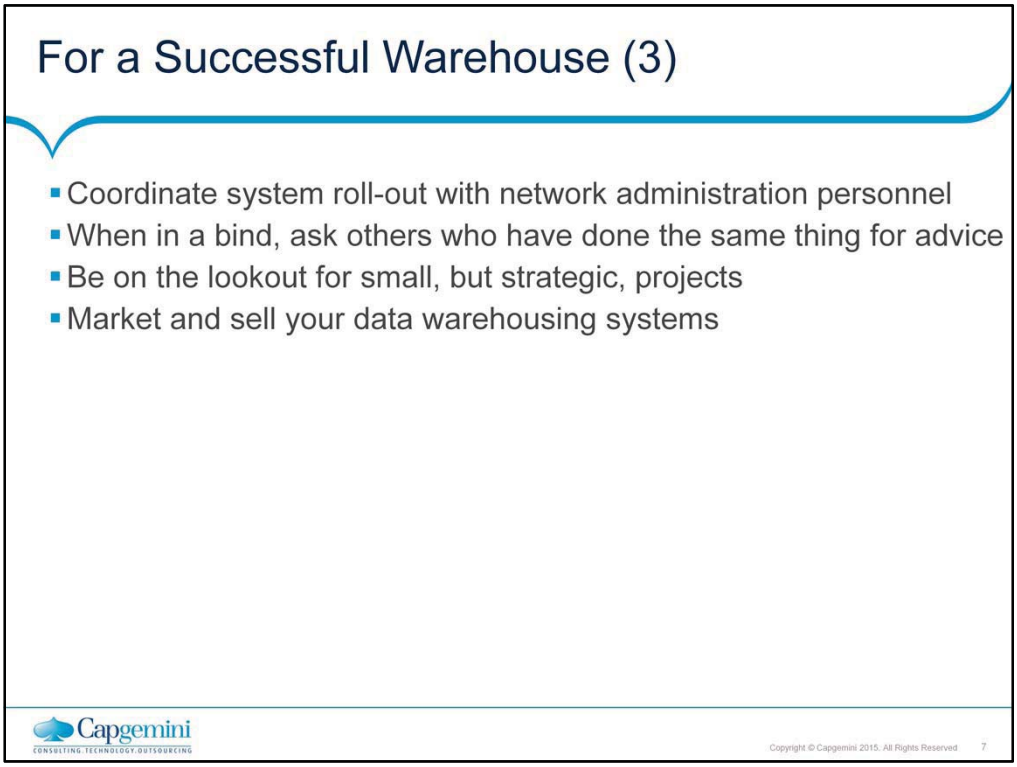
For reasons of politics, overwork, and just plain lack of knowledge of how data are physically stored in a system, the feeder system programmer often can take a while to get you that data.

Determine a plan to test the integrity of the data in the warehouse

Do not underestimate the importance of user faith in the integrity of the warehouse data. Huge warehouse efforts quickly go sour if after system roll-out users find multiple mistakes. A good investment of time in the initial stages of a warehouse project is for the builder and user to jointly determine what checks will be made on the warehouse data during development and what checks need to be made on an ongoing basis. The checks including tying warehouse data controls back to controls in feeder systems, checking the correctness of aggregation logic, testing whether classifications codes were assigned correctly.

From the start get warehouse users in the habit of 'testing' complex queries

Many people will assume that the query result is correct. At the very least, get the user in the habit of eyeballing the query or report to check if several records that should be included are, in fact, included and that several records that should not be included are, in fact, not included.



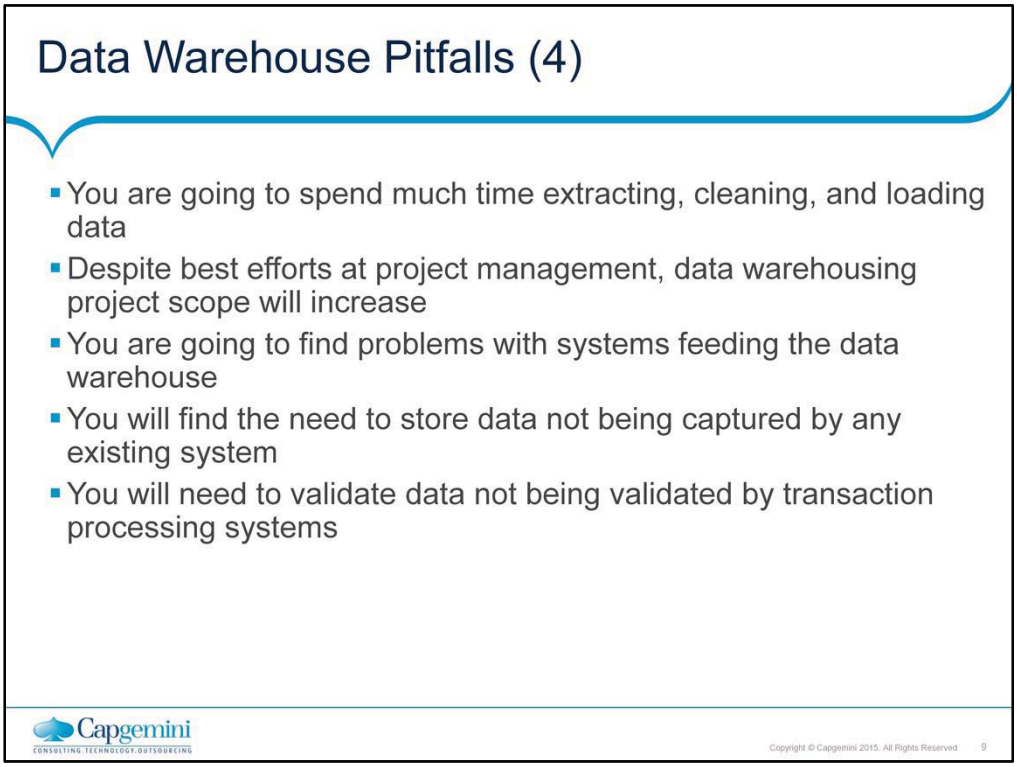
Coordinate system roll-out with network administration personnel
Use of data warehousing systems can bring about some strange spikes in network activity. If you keep network administration people informed of the roll-out schedule, chances are they will monitor network activity for you and be ready to make adjustments to the network as necessary.

Have a good grasp of desktop databases and spreadsheets
Even if you are dealing with a 100 TB database, there are so many little tasks to be done in a data warehousing project where knowledge of these tools will be helpful. Skillful use of these tools during development can be a huge productivity enhancer.

Be prepared to support beginning users immediately and at any time
We developers often greatly underestimate users' hesitation to begin using the data warehouse. This hesitation could be because of user fear of technology or user fear that they will not get Information System support. So, the first point is to be available to help when the user wants to try to use the data warehouse the first time. Users also may want to use the data warehouse for the first time during the weekend or at 6:00 in the morning or 8:00 at night. The distractions are less at those times. If you want to make that beginning user as a committed customer of your data warehouse, you better be available to support the user when he starts out whatever the day or the hour.

Maintain the audit trail to the feeder systems
That is, make it as easy as possible to tie the data in the data warehouse to the feeder systems. Your users have to trust the numbers in the data warehouse. You owe this to the users in order to maintain their trust.

Market and sell your data warehousing systems
For the most part, use of data warehousing systems is optional. This means you have to identify the potential users of the systems, help them understand what are the benefits of the system, and then make them want to keep coming back to use the system.



You are going to spend much time extracting, cleaning, and loading data
The usual figure quoted is that approximately 80% of the time building a data warehouse will be spent on this type of work. (No one has ever explained how this percentage was obtained though.) Suffice it to say, though, the amount of time on these tasks is often grossly underestimated. Note that this point is about extracting and cleaning and loading. Though by now many people are aware the cleaning the data is complex, extracting data and loading data are equally, if not more, complex.

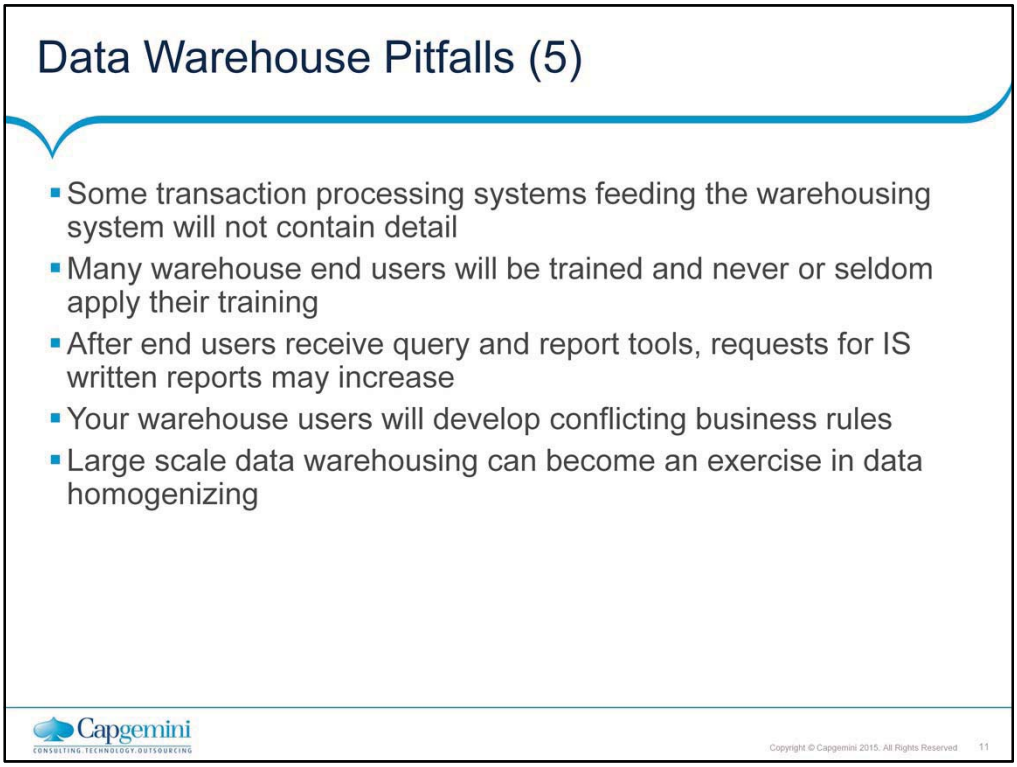
Despite best efforts at project management, data warehousing project scope will increase
To paraphrase data warehousing author W. H. Inmon, traditional projects start with requirements and end with data. Data warehousing projects start with data and end with requirements. Once warehouse users see what they can do with 2000's technology, they will want much more. (Which is fine!)

One piece of advice for the warehouse builder is never to ask the warehouse user what information he wants. Rather, ask what information he wants next.

You are going to find problems with systems feeding the data warehouse
Problems that have gone undetected for years will pop up. You are going to have to make a decision on whether to fix the problem in what you thought was the 'read-only' data warehouse or fix the transaction processing system

You will find the need to store data not being captured by any existing system
A very common problem is to find the need to store data that are not kept in any transaction processing system. For example, when building sales reporting data warehouses, there is often a need to include information on off-invoice adjustments not recorded in an order entry system. In this case the data warehouse developer faces the possibility of modifying the transaction processing system or building a system dedicated to capturing the missing information.

You will need to validate data not being validated by transaction processing systems
Typically once data are in warehouse many inconsistencies are found with fields containing 'descriptive' information. For example, many times no controls are put on customer names. Therefore, you could have 'DEC', 'Digital' and, 'Digital Equipment' in your database. This is going to cause problems for a warehouse user who expects to perform an ad hoc query selecting on customer name. The warehouse developer, again, may have to modify the transaction processing systems or develop (or buy) some data scrubbing technology.



Some transaction processing systems feeding the warehousing system will not contain detail
This problem is often encountered in customer or product oriented warehousing systems. Often it is found that a system which contains information that the designer would like to feed into the warehousing system does not contain information down to the product or customer level. By the way, this is what some people label a 'granularity' problem.

You will under budget for the resources skilled in the feeder system platforms
In addition to understanding the feeder system data, you may find it advantageous to build some of the "cleaning" logic on the feeder system platform if that platform is a mainframe. Often cleaning involves a great deal of sort/merging - tasks at which mainframe utilities often excel. Also, you may find that you want to build aggregates on the mainframe because aggregation also involves substantial sorting.

Many warehouse end users will be trained and never or seldom apply their training
I once read a study that claimed that only one quarter of the people who get training in a query tool actually become heavy users of the tool.

After end users receive query and report tools, requests for IS written reports may increase
This phenomenon was seen with many of the information centers of the 1980s. It comes about because the query and report tools allow the user the users to gain a much better appreciation of what technology could do. However, for many reasons the users are unable to use the new tools themselves to realize the potential. By the way, if this happens do some honest research on why. Granted there are many reports that are so complex that Information Systems expertise is going to be required no matter what tool the end user has. However, many times this phenomenon points to training needs.

Your warehouse users will develop conflicting business rules
Many warehouse tools allow users to perform calculations. The tools will allow users to perform the same calculation differently. For instance, suppose you are summarizing beverage sales by flavor category. Also suppose that the flavor category includes cherry and cola. If you have a cherry cola brand there is a chance that two users will classify the brand in different categories. You will find that there are means to incorporate some of the business rules in your warehouse. However, the number of possible business rules is so large that you will not be able to incorporate all rules

Your warehouse users may not know how to use data
After many years of using whatever reports have been thrown in their faces, the users may not know what data to use their newfangled decision support tools to retrieve. To use a phrase from pop sociology, the users have been "culturally conditioned" to use what they are given and to never ask for more.

Large scale data warehousing can become an exercise in data homogenizing
Data have quirks! Sometimes when we developers combine detailed data for different subjects, in our efforts to make everything 'fit' we can take the life out of the data. For instance, if your company sells dog food and auto tires, you want to be careful if you are building a sales data warehouse for both lines of business. You have to make a judgment call as to whether these businesses fit the same logical and/or physical model.

Data Warehouse Pitfalls (6)

- 'Overhead' can eat up great amounts of disk space
- The time it takes to load the warehouse will expand to the amount of the time in the available window... and then some
- Assigning security cannot be done with a transaction processing system mindset
- You are building a HIGH maintenance system
- You will fail if you concentrate on resource optimization to the neglect of project, data, and customer management issues and an understanding of what adds value to the customer

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 13

'Overhead' can eat up great amounts of disk space
A popular way to design a decision support relational databases is with star or snowflake schemas. Persons taking this approach usually also build aggregate fact tables. If there are many dimensions to the data, be aware that the combination of the aggregate tables and indexes to the fact tables and aggregate fact tables can eat up many times more space than the raw data. If you are using multidimensional databases, be aware that certain products pre-calculate and store summarized data. As with star/snowflake schemas, storage of this calculated data can eat up far more storage than the raw data.

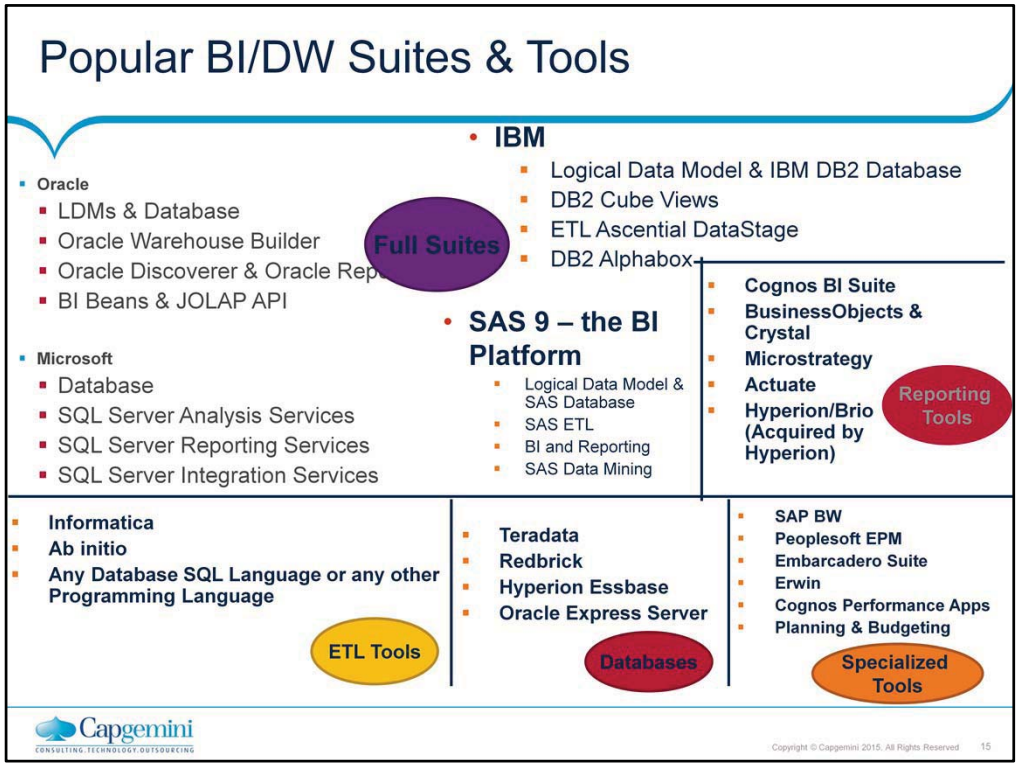
The time it takes to load the warehouse will expand to the amount of the time in the available window... and then some
You will do yourself well by understanding the different ways to approach updating the warehouse. Before you decide that you can do complete refreshes, be aware that "There's all day Sunday to load the database!" have been famous last words of more than a handful of warehouse developers.

You are going to have a tough problem with security - especially if you make your data warehouse Web-accessible
You are going to face a paradox - the more accessible you make your data warehouse (and by accessible, I don't just mean making it Web accessible - I mean architecting it in a way that people want to use it), the greater security risk you are exposing yourself too. Frankly, restricting people to "need to know" does not cut it in the organization on the 2000s. But, on the other hand, exposing information to theft from anyplace in the globe is not too great for job security either.

The data warehouse data you do not reconcile with the feeder systems will cause the problems
For certain data warehouse data you are going to think that there is no logical way that data in the feeder systems can be reconciled with what are in the warehouse. Then, when a user looks at a report and tells you "I think there is a problem", it will be with the unreconciled data. Unfortunately, you will then discover there is a way, albeit roundabout, to reconcile the data.

You are building a HIGH maintenance system
Reorganizations, product introductions, new pricing schemes, new customers, changes in production systems, etc. are going to affect the warehouse. If the warehouse is going to stay 'current' (and being current will be a big selling point of the warehouse), changes to the warehouse have to be made fast.

You will fail if you concentrate on resource optimization to the neglect of project, data, and customer management issues and an understanding of what adds value to the customer
If you provide a system that is fast and technically elegant but adds little value or has suspect data, you will probably lose your customer from day one and will have a tough time getting him back. For the most part, use of data warehousing systems is optional. The customer has to want to use the system.



There are lot of BI tools in the market. The Organization like Oracle, Microsoft , IBM and SAS providing tools which provides end to end solutions that includes Designing , Profiling, MetaData, ETL , Database and Reporting solutions.

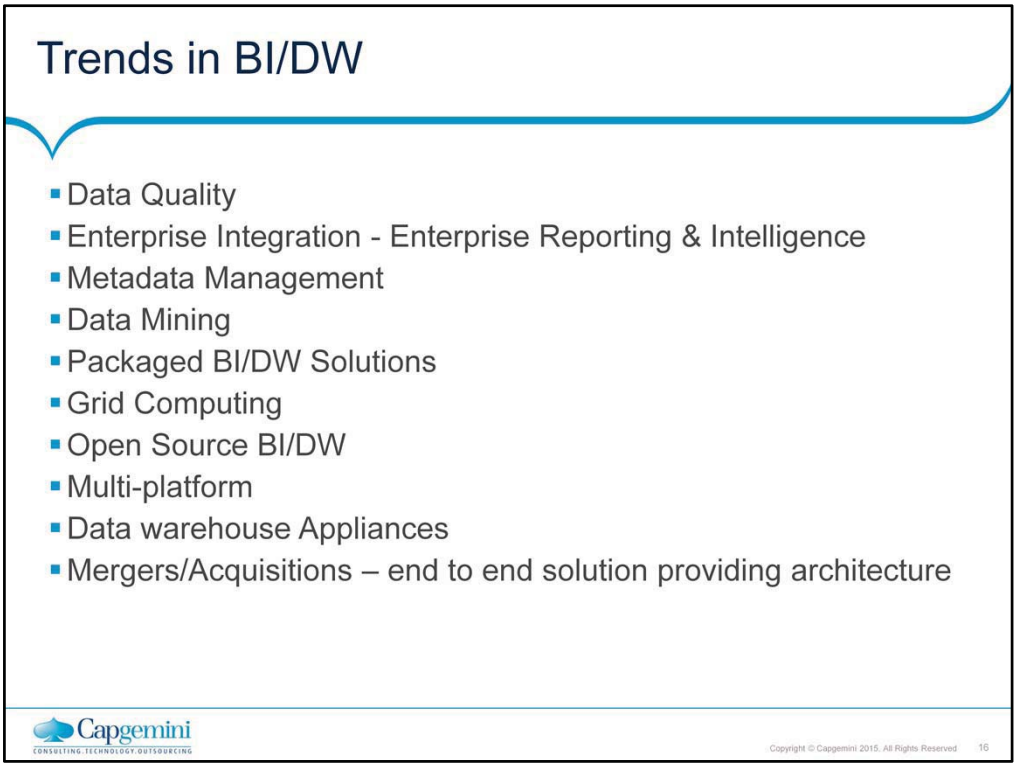
There are exclusive ETL tools such as Informatica, DataStage, Business Object Data Integrator, OWB , Abinitio which provides Extraction , Transformation and Loading solutions and handles huge volumes of data.

There are exclusive Database tools like Teradata, Redbrick etc which provides database solutions to hold huge amount of data.

There are exclusive Reporting tools like cognos, Business Objects XI, Actuate etc which provides Reporting solutions for various users view. And also comfortable with drill down, roll up, drill across, slice, dice operations.

In addition, there are some specilized tools which can used for specific purpose for instance Erwin would be used for designing database etc.

In every tools lot of enhancements are taken place and most tools supports for SOA (Service Oriented Architecture), Data Integration , Data Quality and Cloude Computing.



In this presentation we can discuss about emerging Trends in Business Intelligence and Data Warehouse. Following are the Emerging Area where BIDW is playing vital role.
Data Quality – As Industry needs quality data to take decision hence most of the tools are providing solution to provide quality data for instances in DataStage Quality stage has embaded in DataStage 8x, In BOBJ Data Services they provides transformers which supports for Quality.

Most of the tools supports and maintain **Metadata** and keep track of every metadata it maintains. Tools provides to generate impact analysis report for metadata.

Data Mining is equally playing vital role in industry like Insurance, Telecommunication, Banking etc. There are lot of tools uses different algorithms.

Many organization provides end to end solutions including software and **hardware appliances**.

Every tools are enhancing new features to work with **multiple platforms**, multiple database and multiple architecture.

As data is growing in million billions of records as a result performing complex operations through single computing system may not be sufficient hence most of the tools are providing **Grid computing facility** in where Data can be routed across multiple computing systems to perform complex tasks.

Open Source BIDW are also emerging trends and so that one customize the tool based on their requirements.

Summary

- In this lesson, you have learnt:
  - Precautions to be taken for successful data warehouse
  - Tools available for data warehouse
  - Data warehouse trends

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

 17