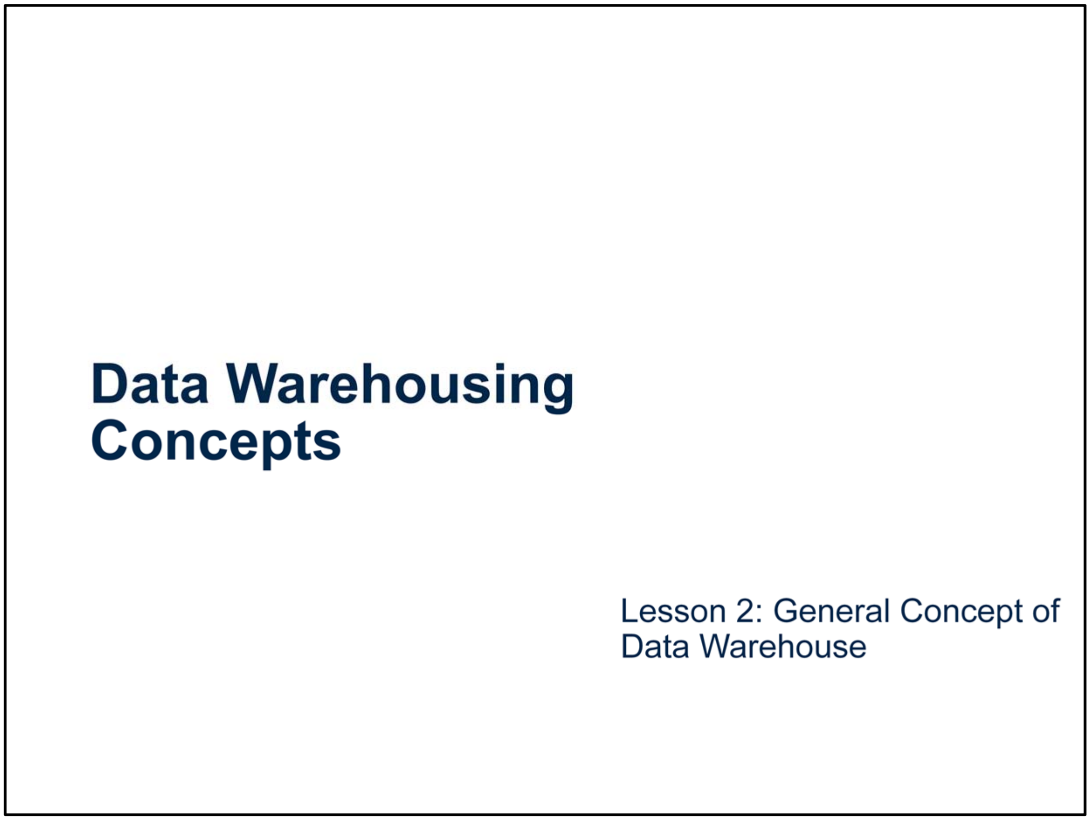
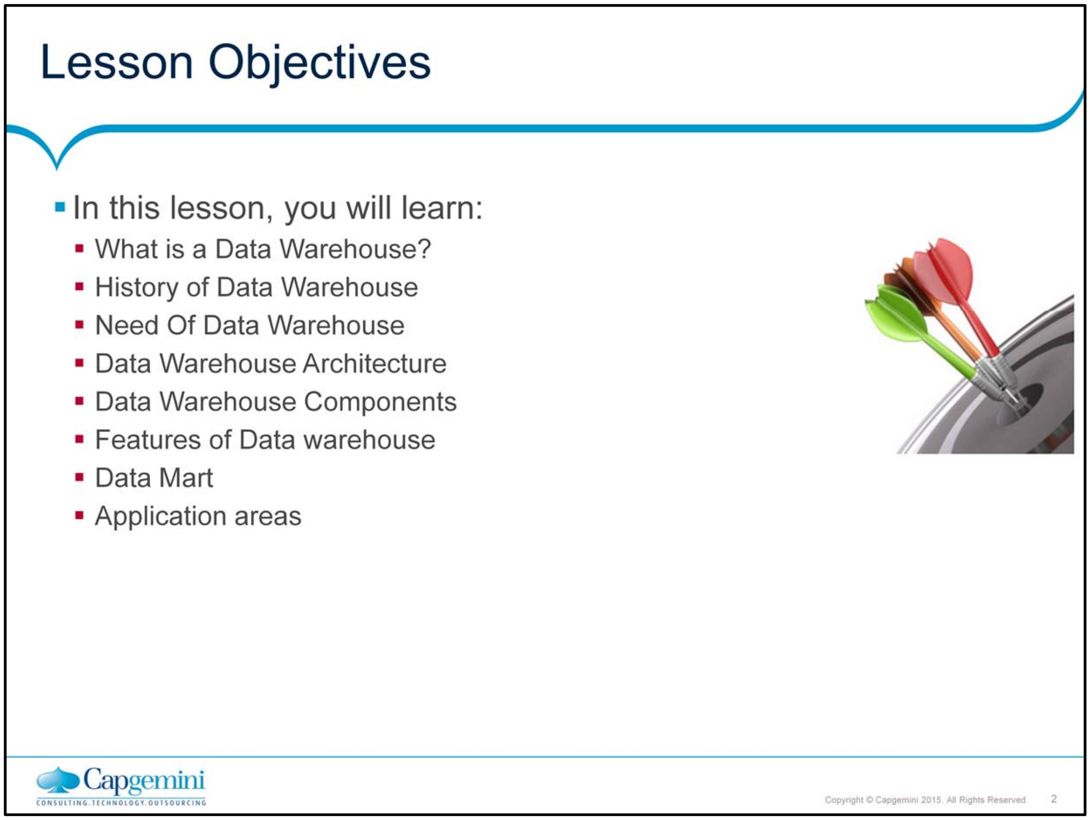


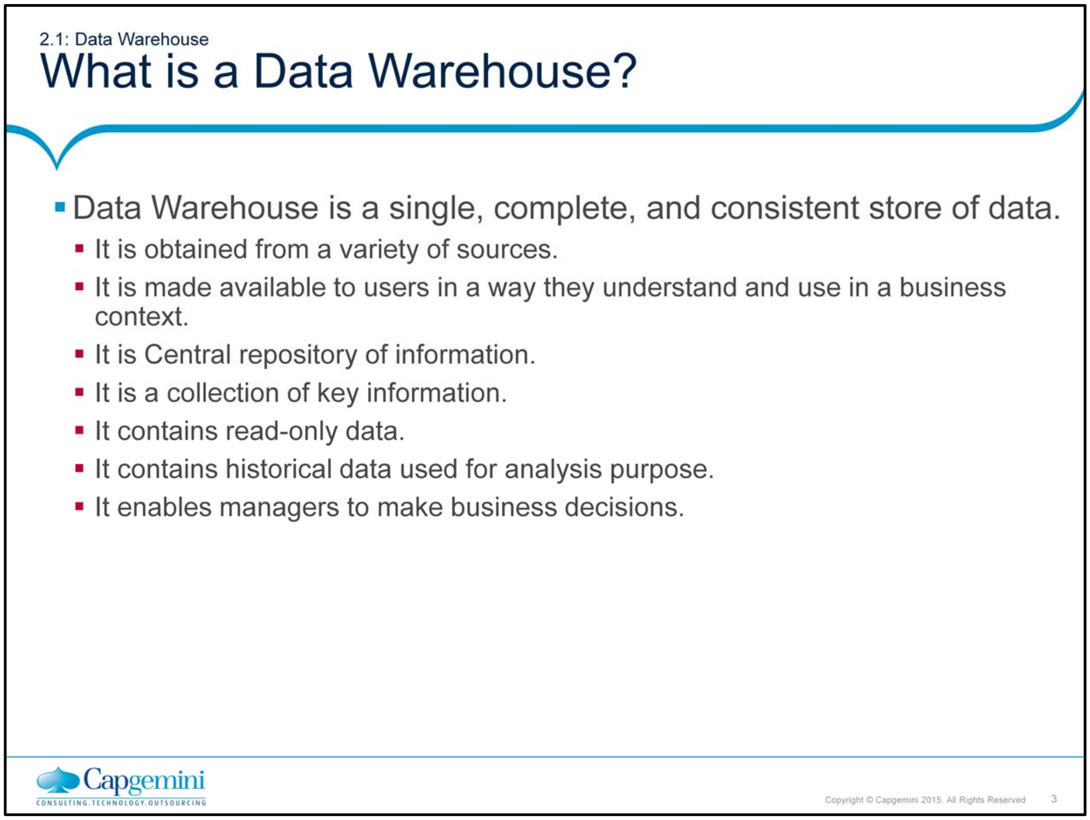
Data Warehousing Concepts

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



2.1: Data Warehouse

What is a Data Warehouse?

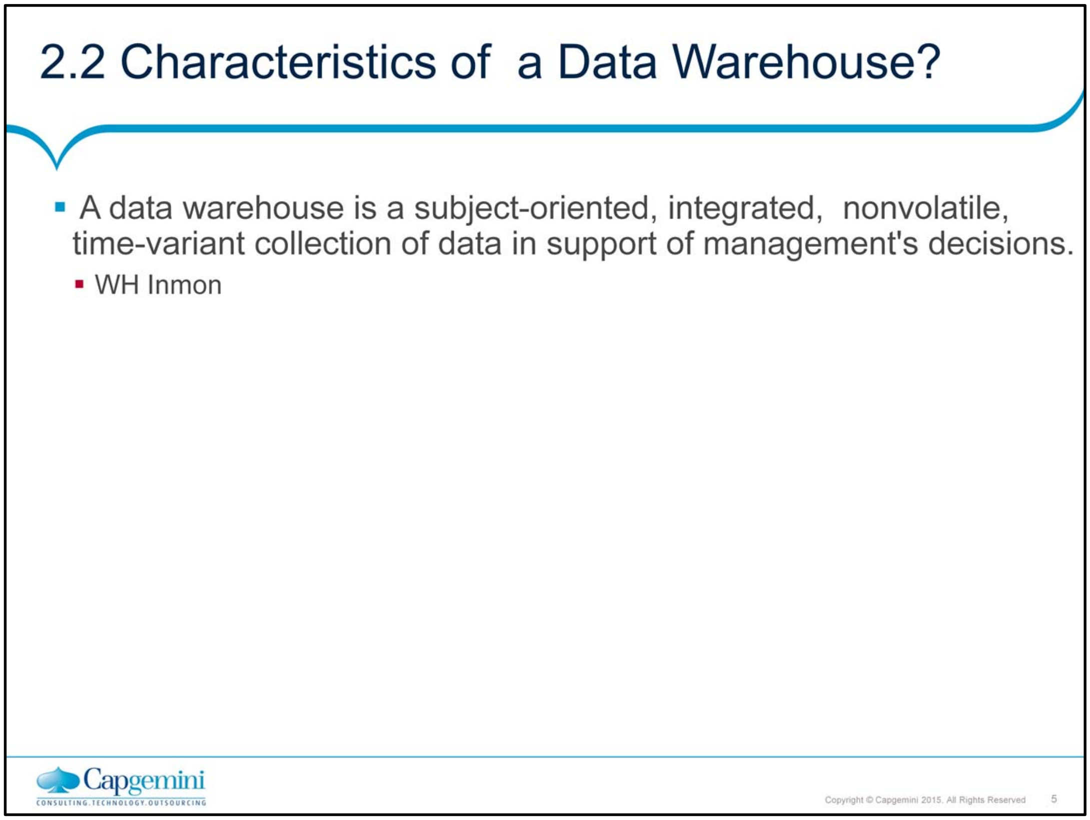
- Data Warehouse is a single, complete, and consistent store of data.
  - It is obtained from a variety of sources.
  - It is made available to users in a way they understand and use in a business context.
  - It is Central repository of information.
  - It is a collection of key information.
  - It contains read-only data.
  - It contains historical data used for analysis purpose.
  - It enables managers to make business decisions.

3

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.):

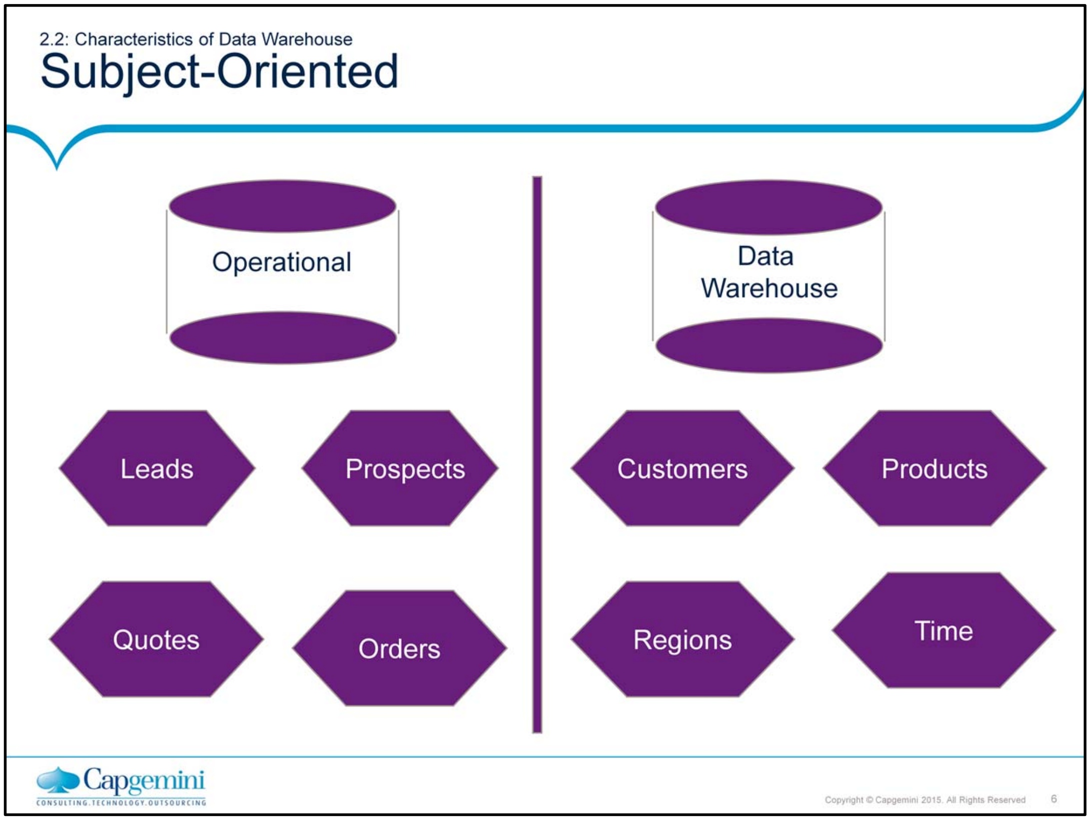
- A Data Warehouse can also be defined as a **structured**, **extensible environment** designed for the analysis of non-volatile data, which is logically and physically transformed from multiple source applications to align with business structure, updated and maintained for a long time period, expressed 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)**.



2.2 Characteristics of a Data Warehouse?

- A data warehouse is a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions.
  - WH Inmon

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

Quotes

Orders

Customers

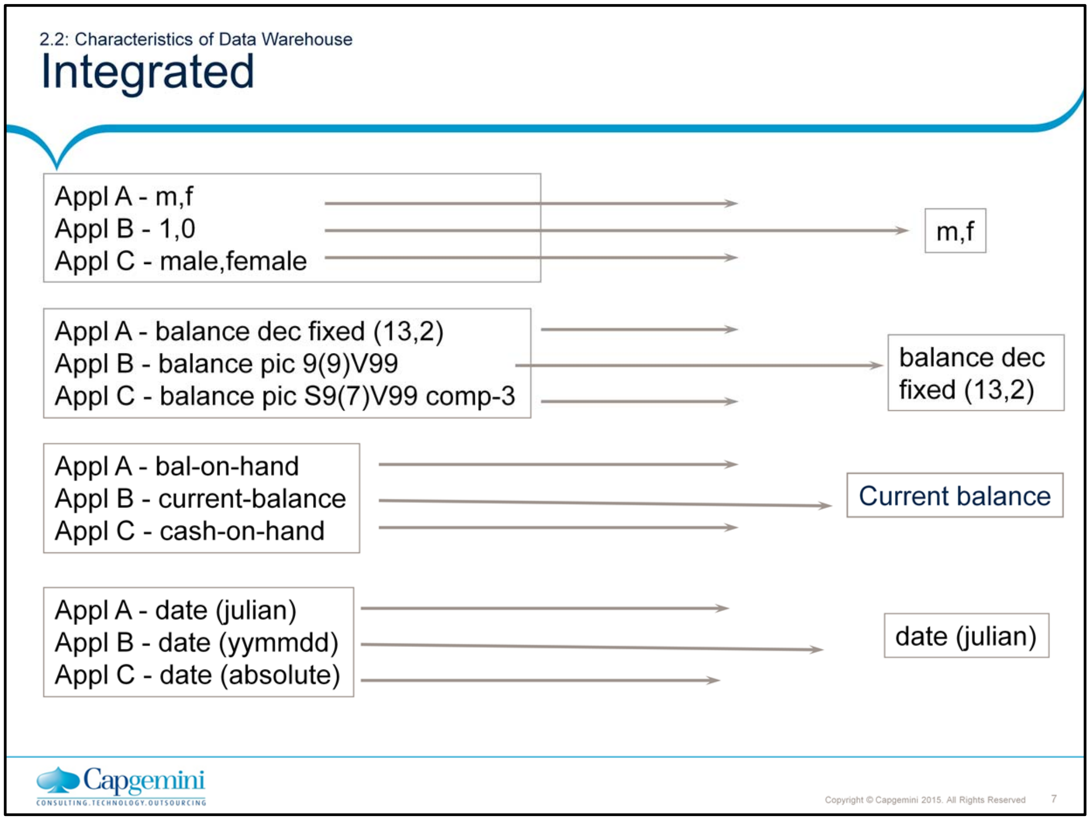
Products

Regions

Time

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

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

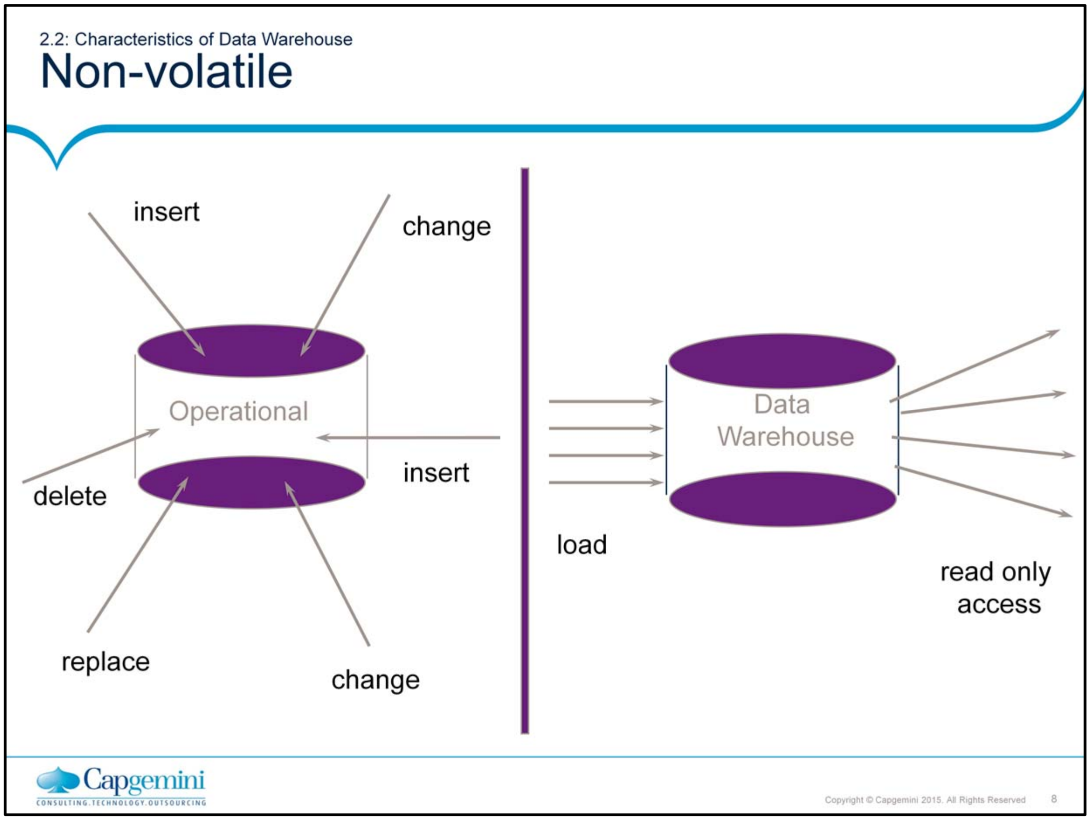
balance dec fixed (13,2)

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)



2.2: Characteristics of Data Warehouse

Non-volatile

insert

change

Operational

delete

insert

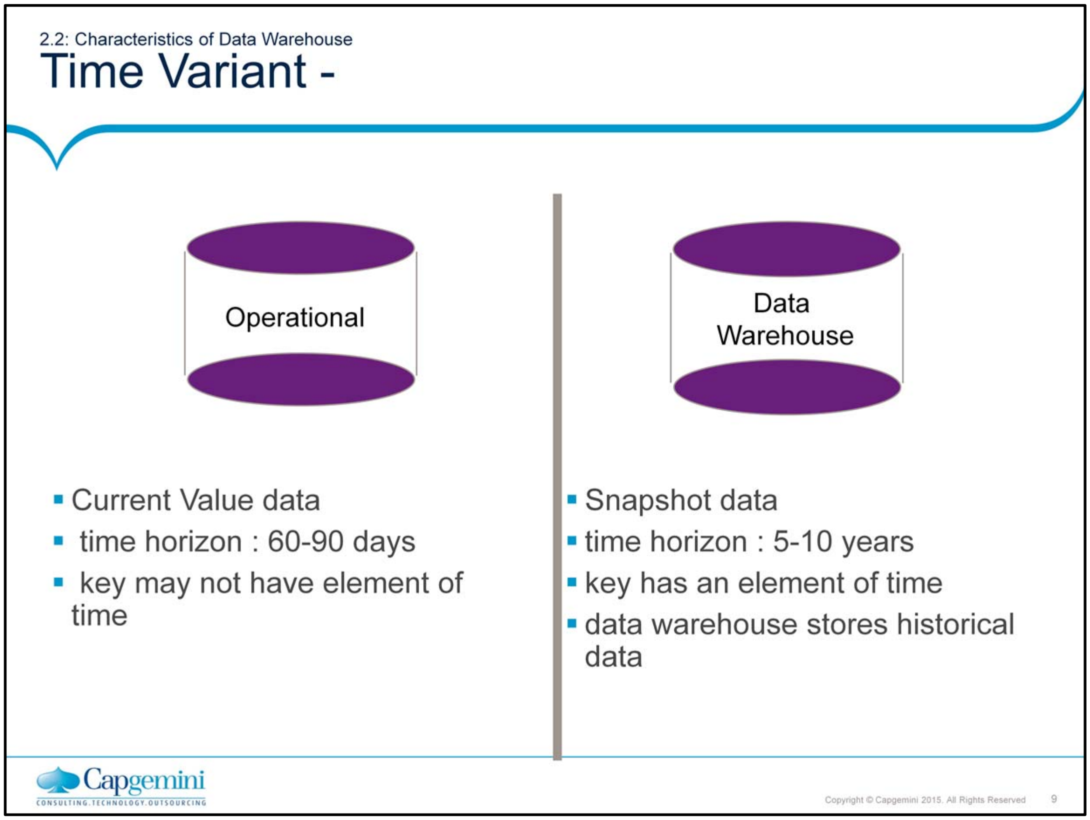
replace

change

load

Data Warehouse

read only access



2.2: Characteristics of Data Warehouse

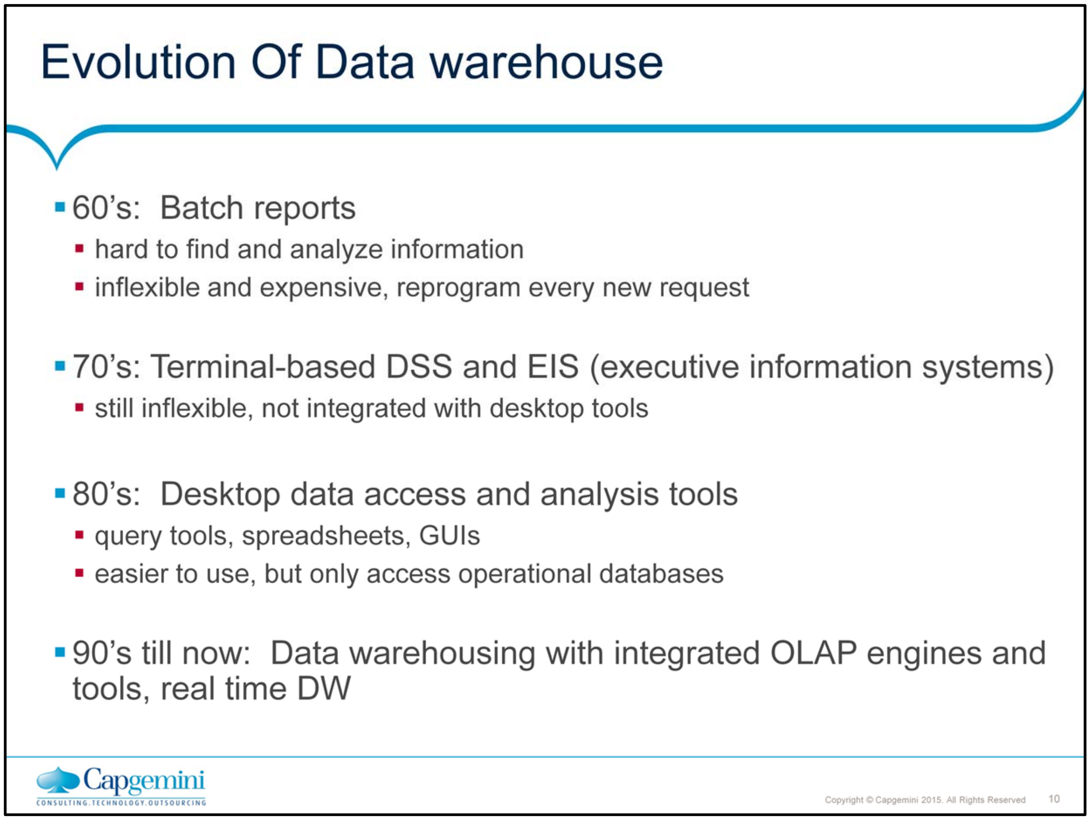
Time Variant -

Operational

- Current Value data
- time horizon : 60-90 days
- key may not have element of time

Data Warehouse

- Snapshot data
- time horizon : 5-10 years
- key has an element of time
- data warehouse stores historical data



Evolution Of Data warehouse

- 60's: Batch reports
  - hard to find and analyze information
  - inflexible and expensive, reprogram every new request
- 70's: Terminal-based DSS and EIS (executive information systems)
  - still inflexible, not integrated with desktop tools
- 80's: Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easier to use, but only access operational databases
- 90's till now: Data warehousing with integrated OLAP engines and tools, real time DW

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

10

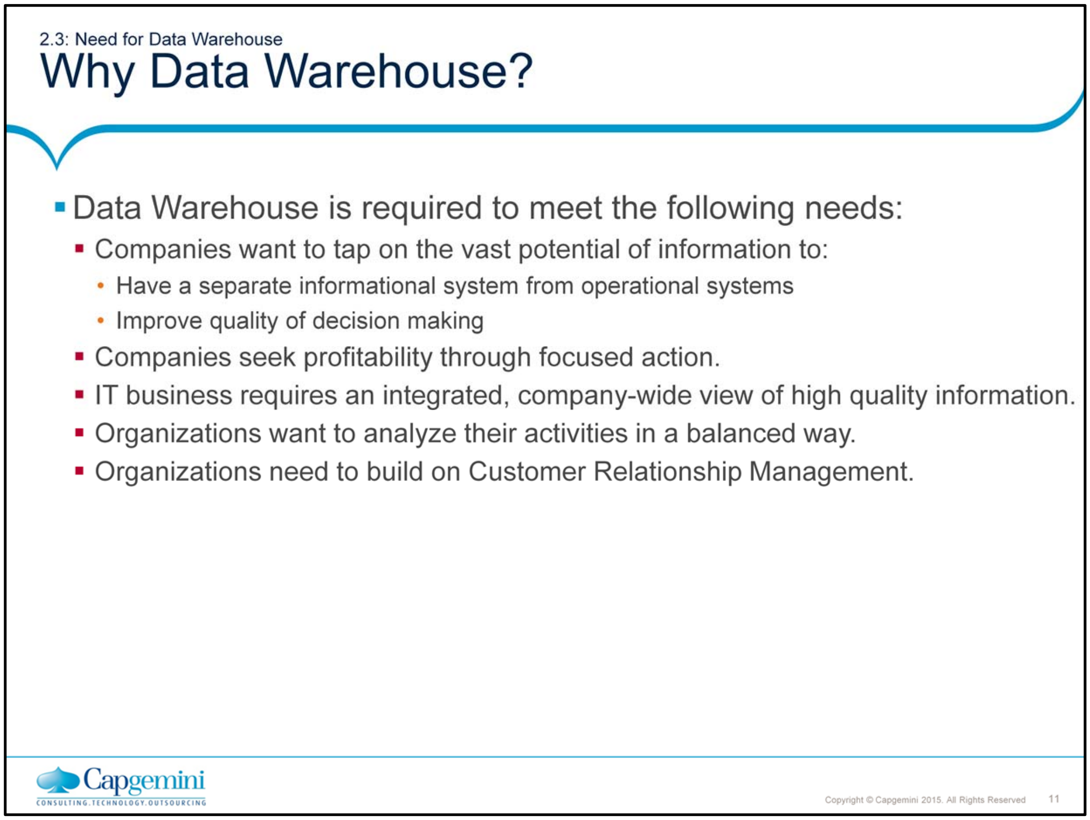
Data warehousing is known as decision support system. If we look at the historical development of Decision Support System (DSS) this is how DSS has been developed since 1960.

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.



2.3: Need for Data Warehouse

Why Data Warehouse?

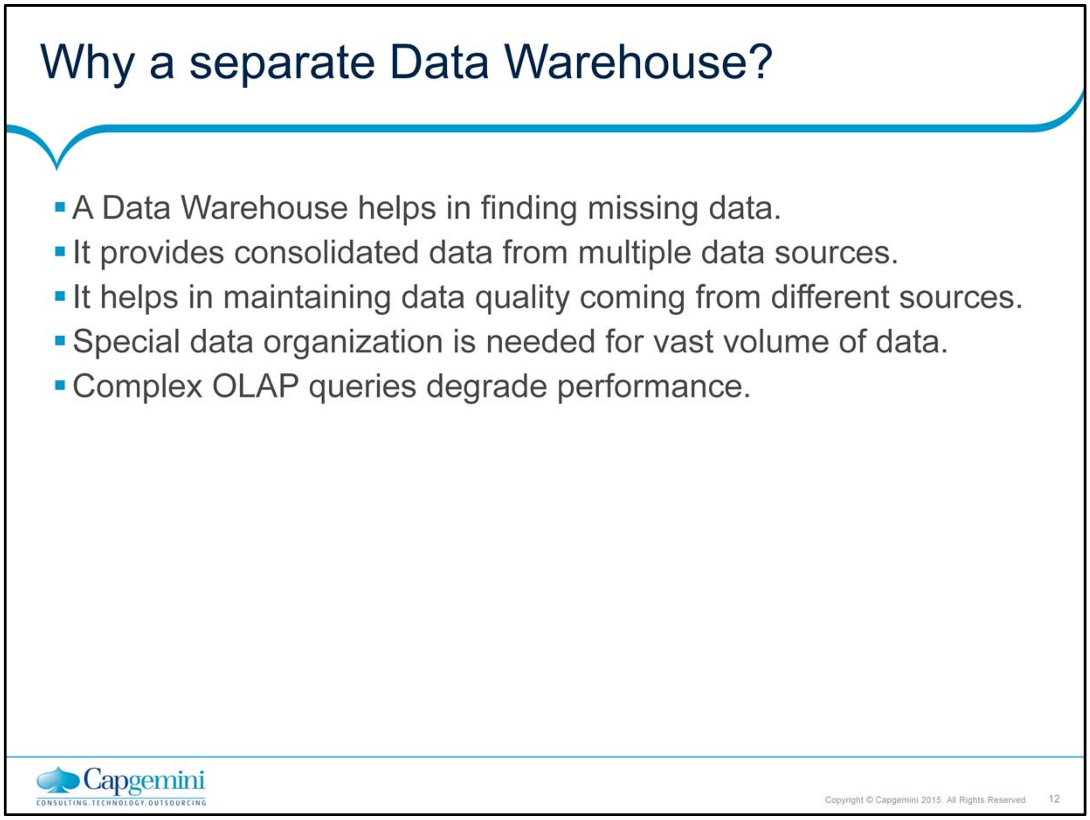
- Data Warehouse is required to meet the following needs:
  - Companies want to tap on the vast potential of information to:
    - Have a separate informational system from operational systems
    - Improve quality of decision making
  - Companies seek profitability through focused action.
  - IT business requires an integrated, company-wide view of high quality information.
  - Organizations want to analyze their activities in a balanced way.
  - Organizations need to build on Customer Relationship Management.

11

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?

- A Data Warehouse helps in finding missing data.
- It provides consolidated data from multiple data sources.
- It helps in maintaining data quality coming from different sources.
- Special data organization is needed for vast volume of data.
- Complex OLAP queries degrade performance.

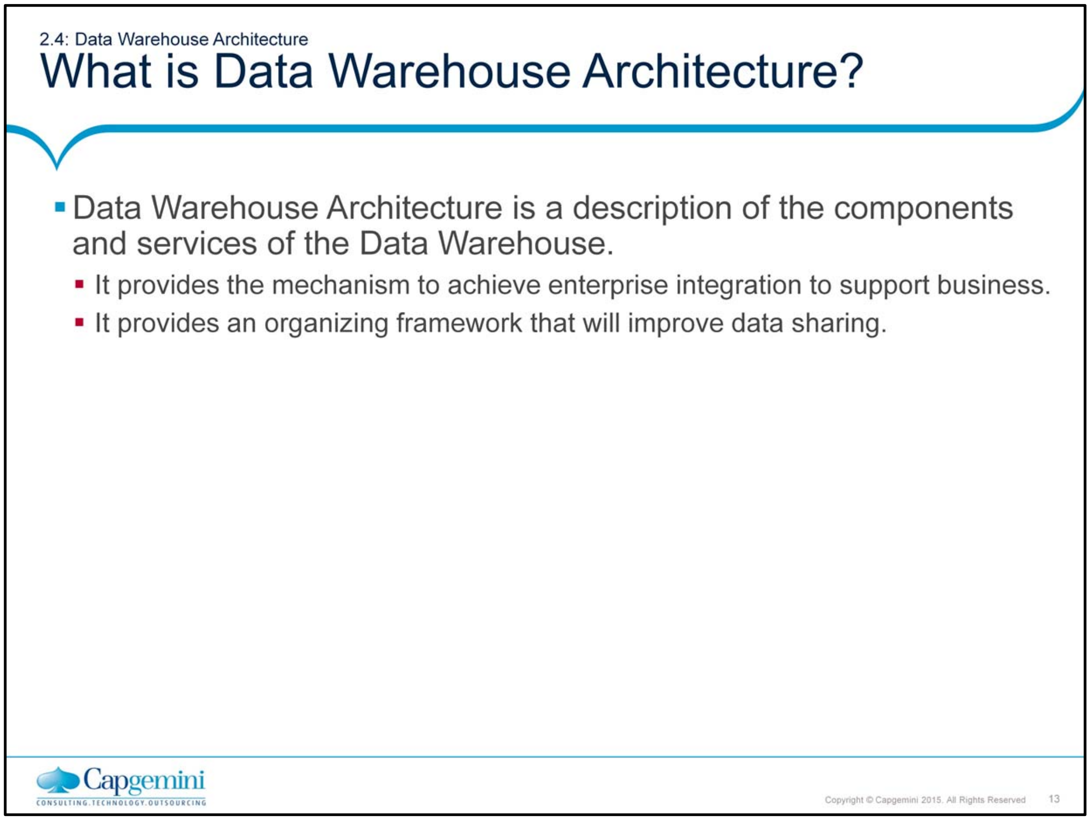
12

Why a separate Data Warehouse?

Functions of a Data Warehouse:

A Data Warehouse is typically used for data consolidation and enforcing uniform data quality.

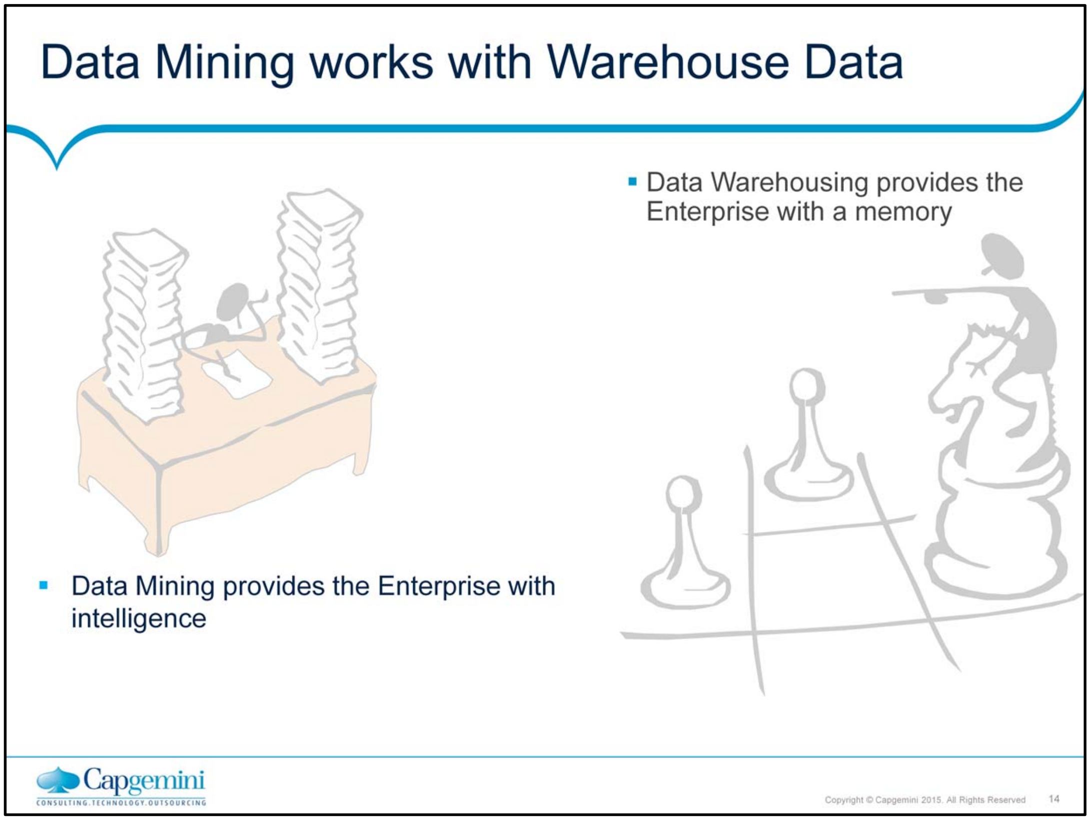
> Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources, namely operational databases, external sources.
>
> Data quality: Different sources typically use inconsistent data representations, codes, and formats that have to be reconciled.



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.



Data Mining works with Warehouse Data

- Data Warehousing provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence

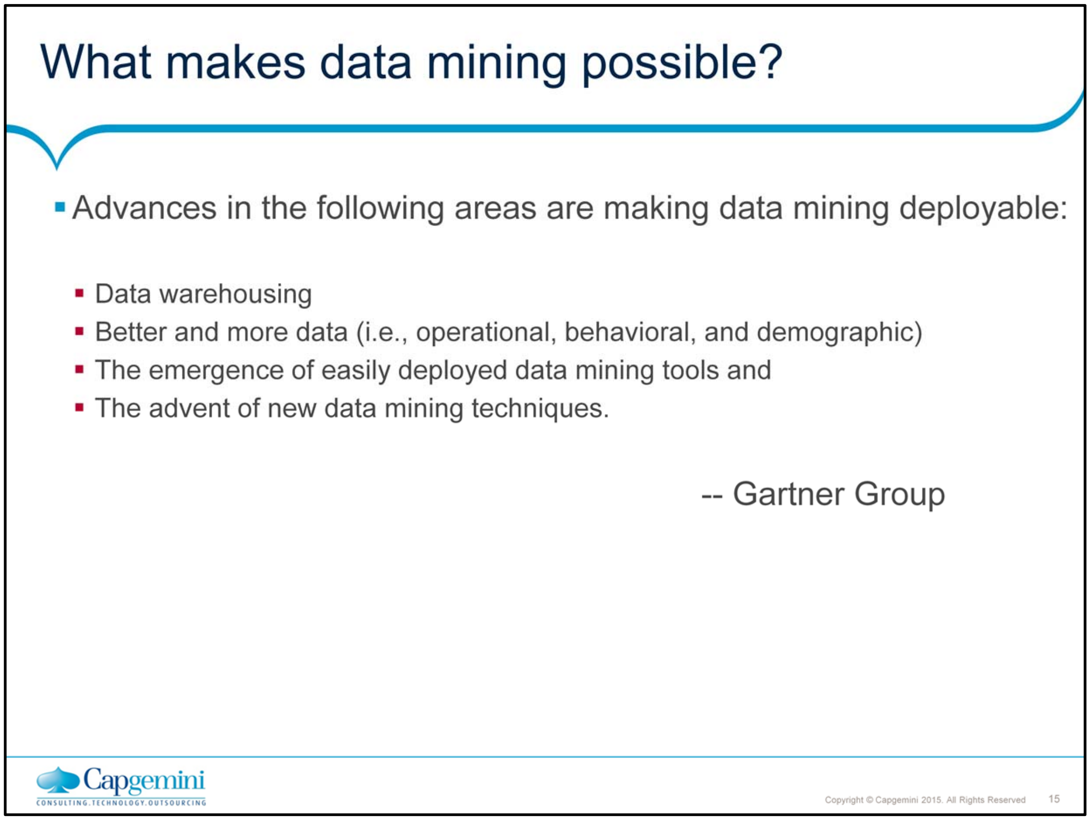
Data Mining is one of the new research avenue which is closely works with data warehousing. Data mining is the process of extracting some relevant and useful data from an already available data store .

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.



What makes data mining possible?

- Advances in the following areas are making data mining deployable:
  - Data warehousing
  - Better and more data (i.e., operational, behavioral, and demographic)
  - The emergence of easily deployed data mining tools and
  - The advent of new data mining techniques.

-- Gartner Group

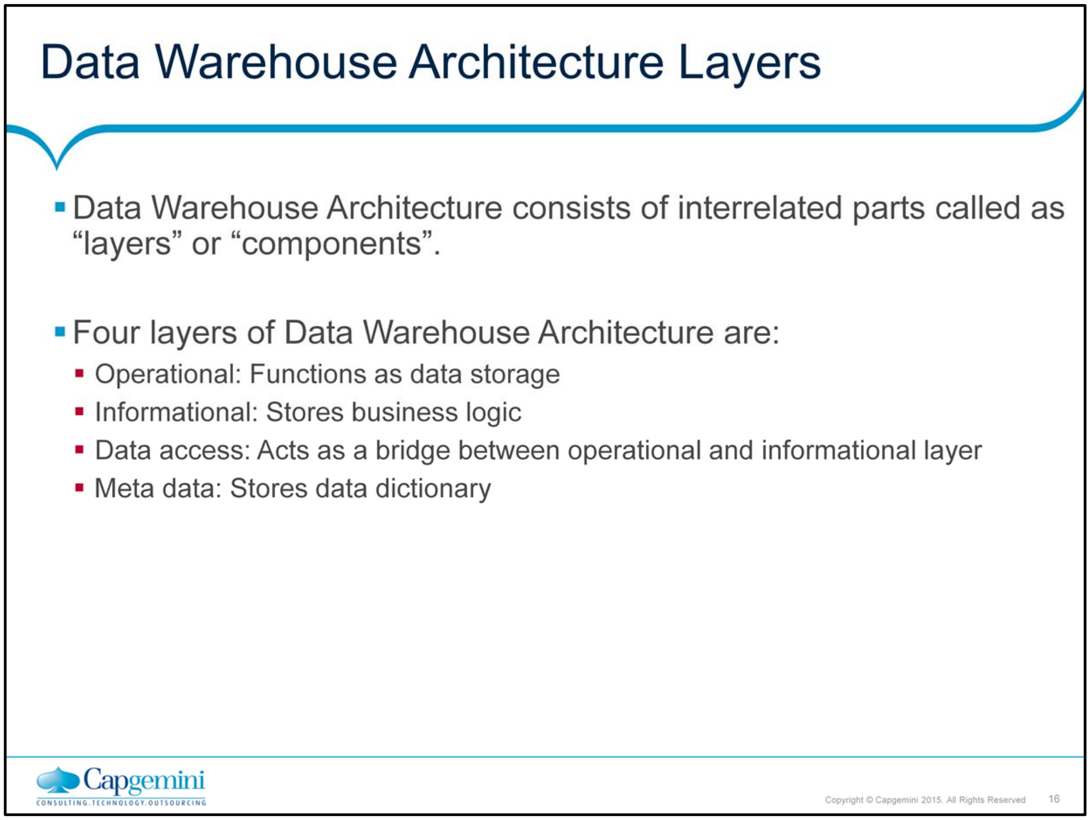
As per Gartner Group Survey Data Mining made possible to work with data warehouse since ;

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 Layers

- Data Warehouse Architecture consists of interrelated parts called as “layers” or “components”.
- Four layers of Data Warehouse Architecture are:
  - Operational: Functions as data storage
  - Informational: Stores business logic
  - Data access: Acts as a bridge between operational and informational layer
  - Meta data: Stores data dictionary

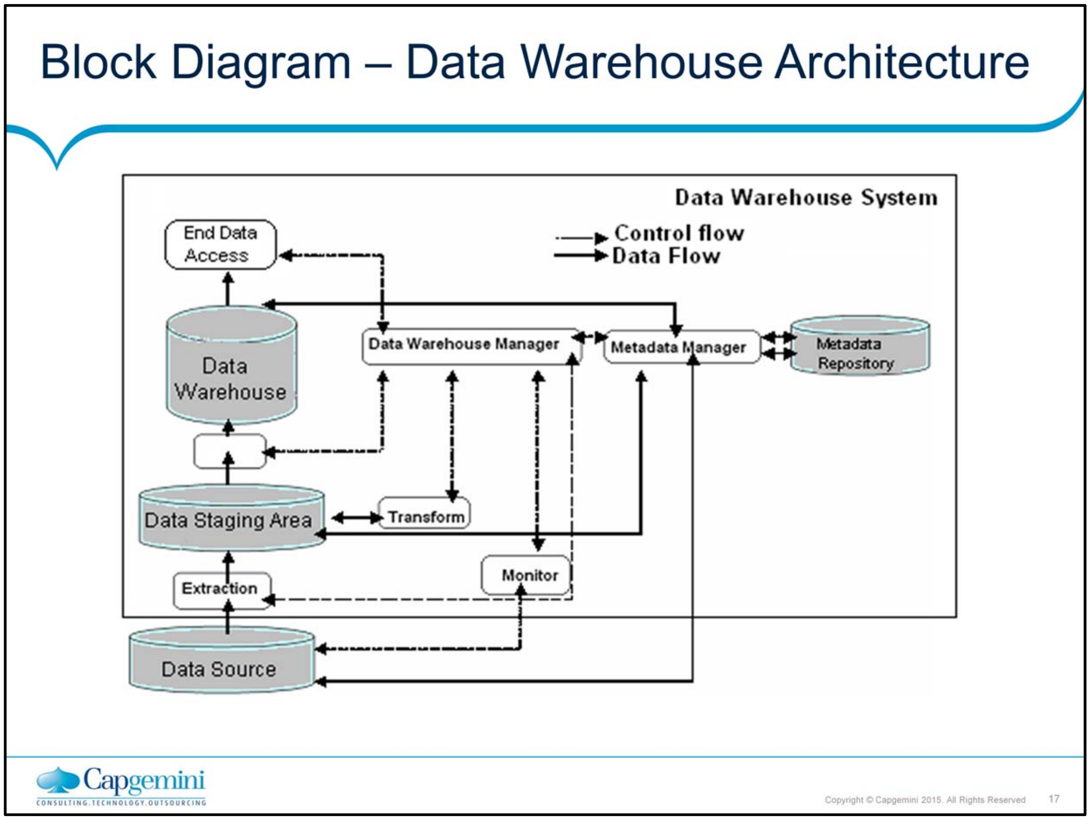
Data Warehouse Architecture:
It consists of four interrelated layers:

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.



Block Diagram – Data Warehouse Architecture

17

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.

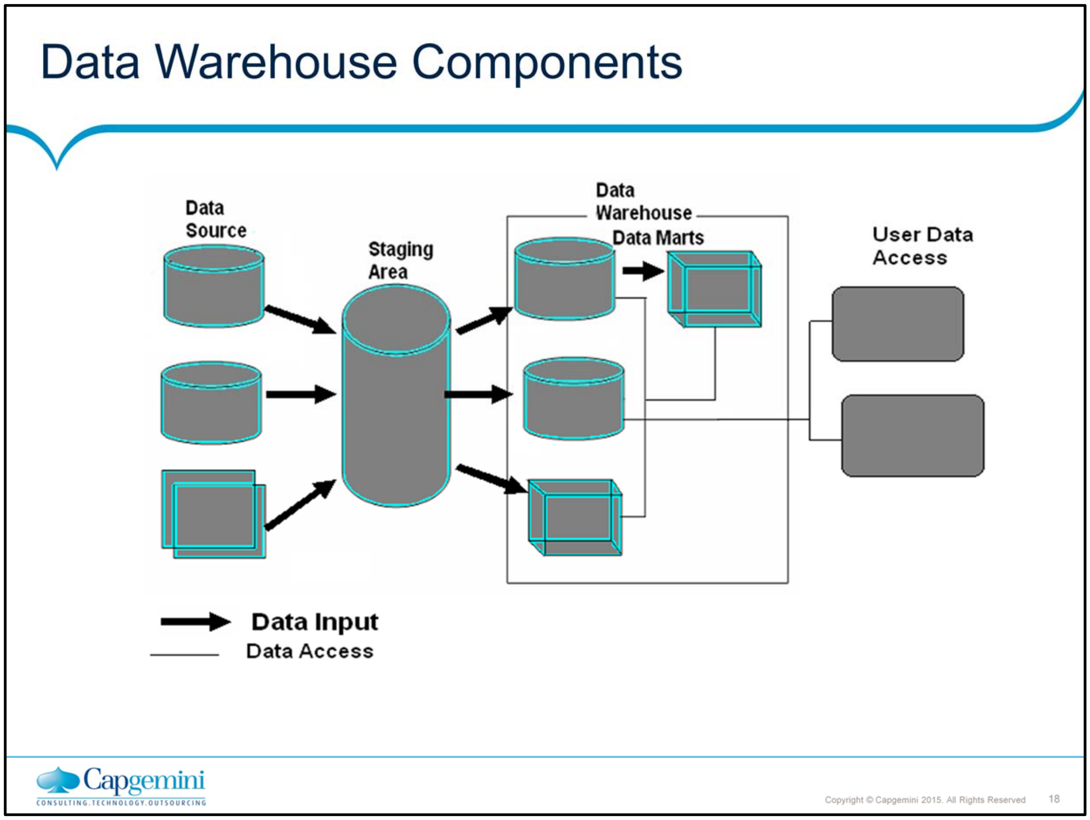
Data Warehouse Manager: It performs all operations associated with the management of the data in the warehouse.

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

18

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.

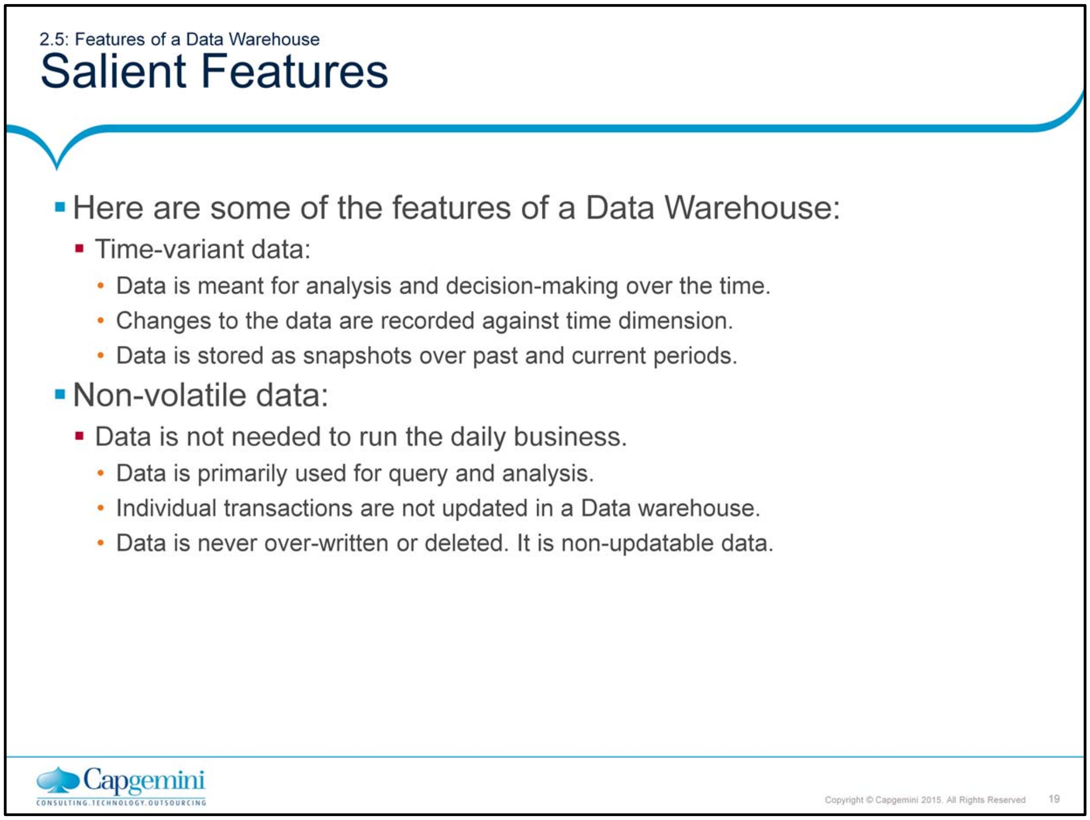
It comprises of different relational data (ORACLE, DB2, SQL Server, etc.).

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.



2.5: Features of a Data Warehouse

Salient Features

- Here are some of the features of a Data Warehouse:
  - Time-variant data:
    - Data is meant for analysis and decision-making over the time.
    - Changes to the data are recorded against time dimension.
    - Data is stored as snapshots over past and current periods.
- Non-volatile data:
  - Data is not needed to run the daily business.
    - Data is primarily used for query and analysis.
    - Individual transactions are not updated in a Data warehouse.
    - Data is never over-written or deleted. It is non-updatable data.

19

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.

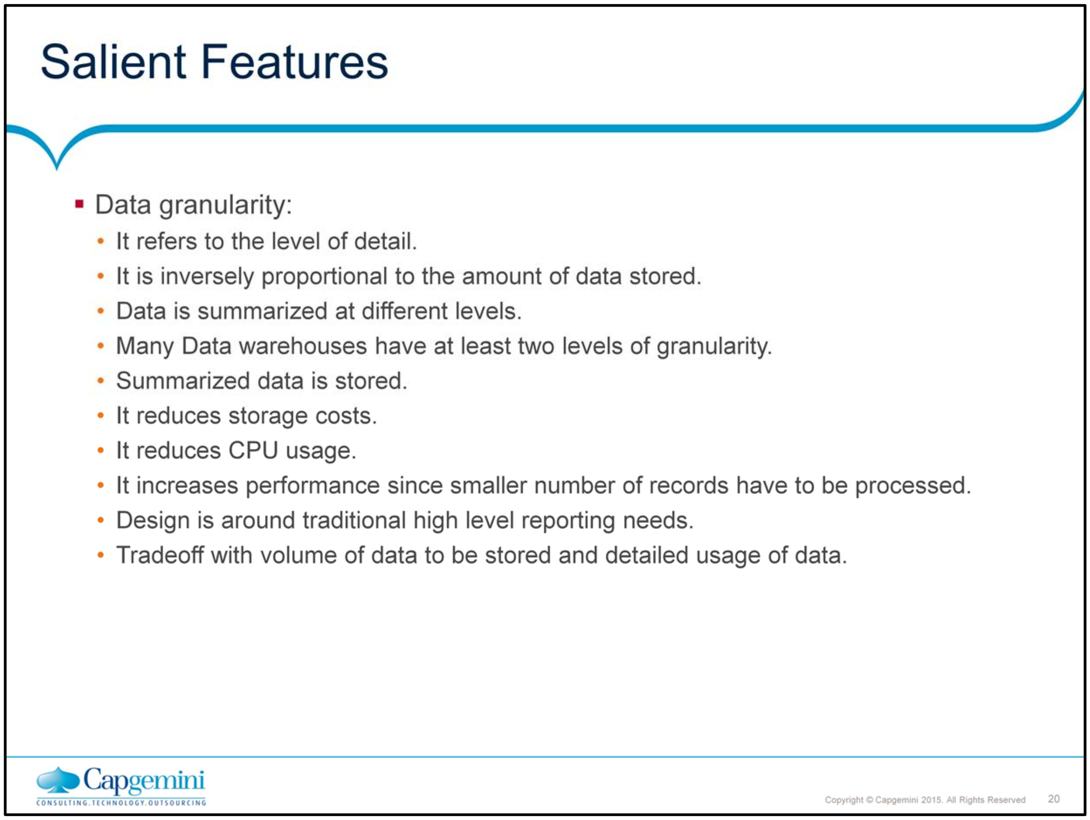
This nature of data:

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

Non-volatile data:

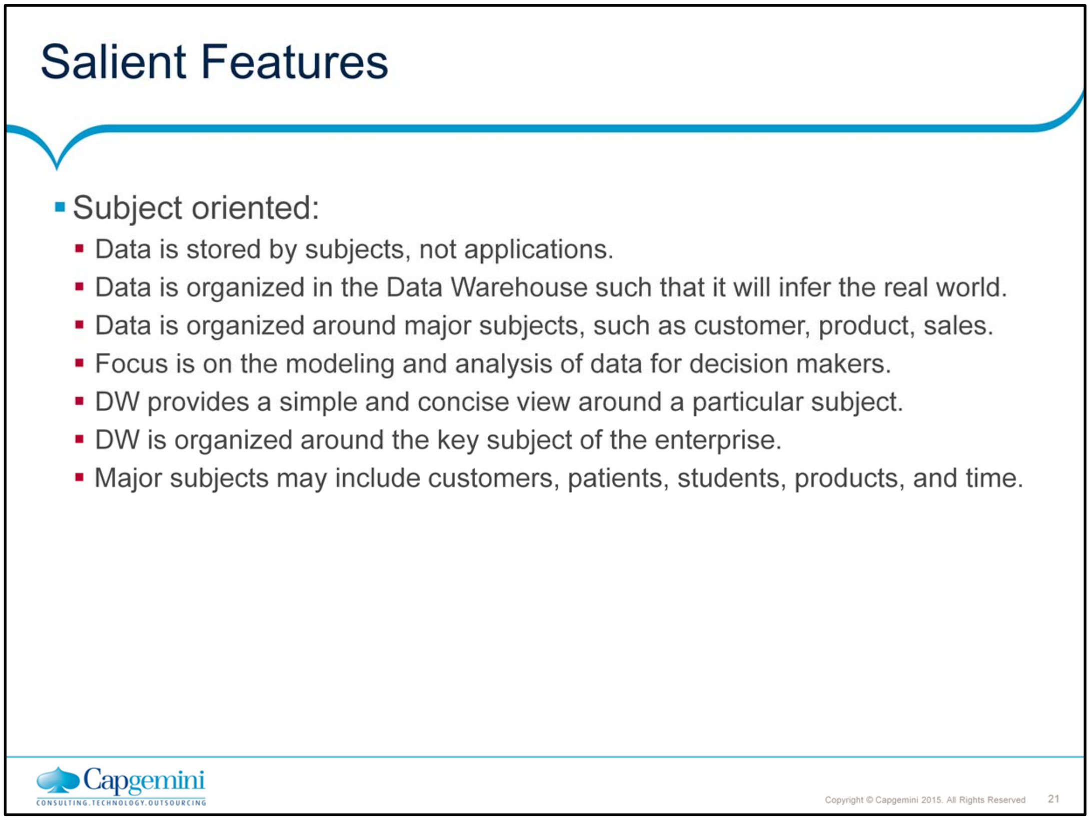
Data in the Data warehouse is loaded and refreshed from operational systems. However, it cannot be updated by end users.

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.



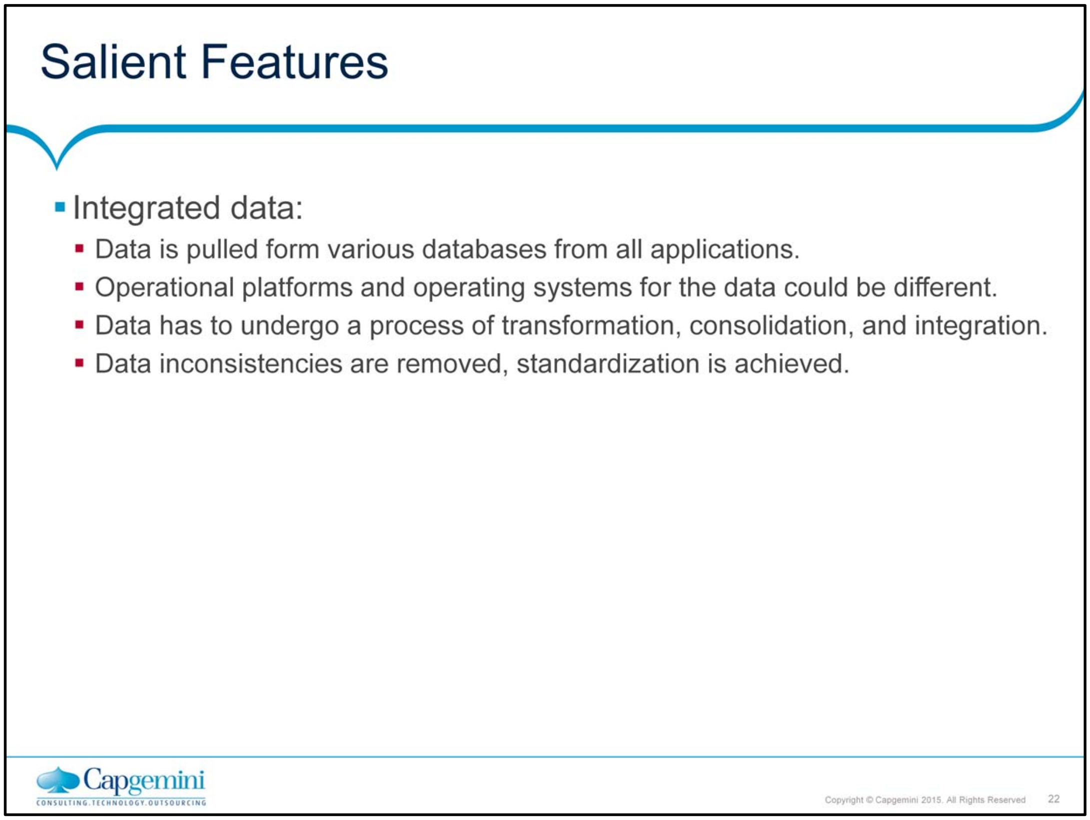
Salient Features

- Subject oriented:
  - Data is stored by subjects, not applications.
  - Data is organized in the Data Warehouse such that it will infer the real world.
  - Data is organized around major subjects, such as customer, product, sales.
  - Focus is on the modeling and analysis of data for decision makers.
  - DW provides a simple and concise view around a particular subject.
  - DW is organized around the key subject of the enterprise.
  - Major subjects may include customers, patients, students, products, and time.

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.

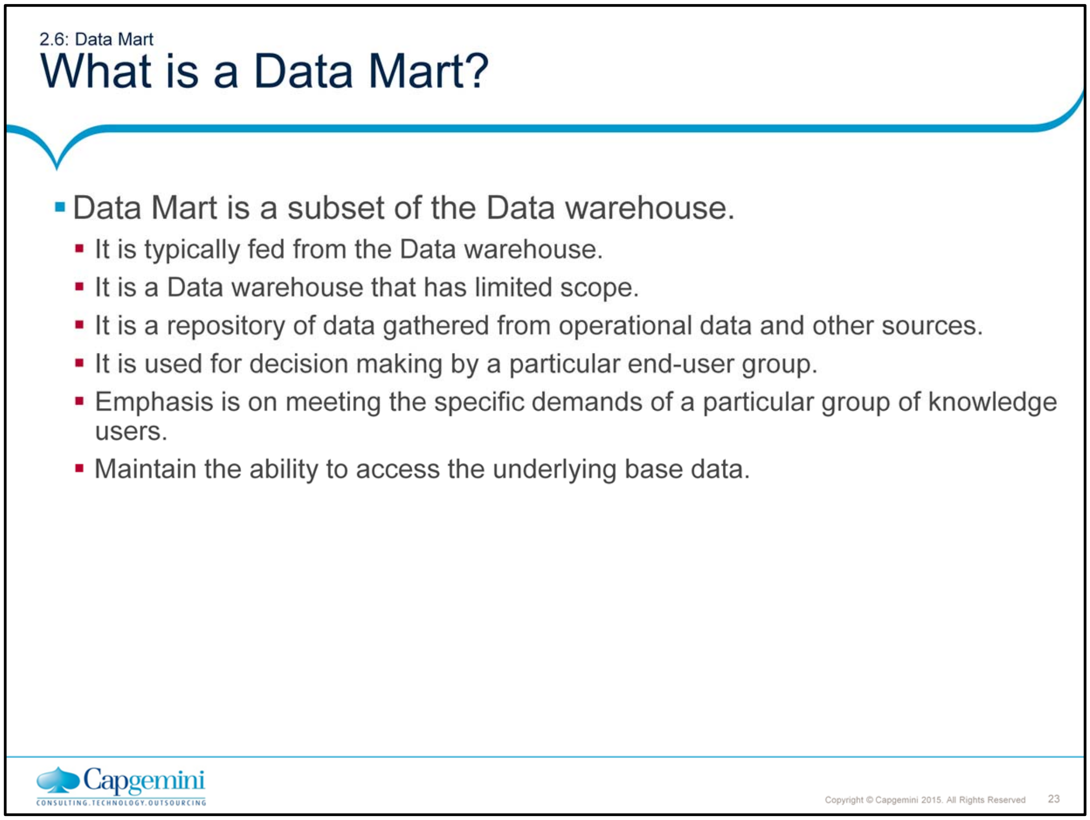
DW provides a simple and concise view around particular subject issues by excluding data that is not useful in the decision support process.



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.

22

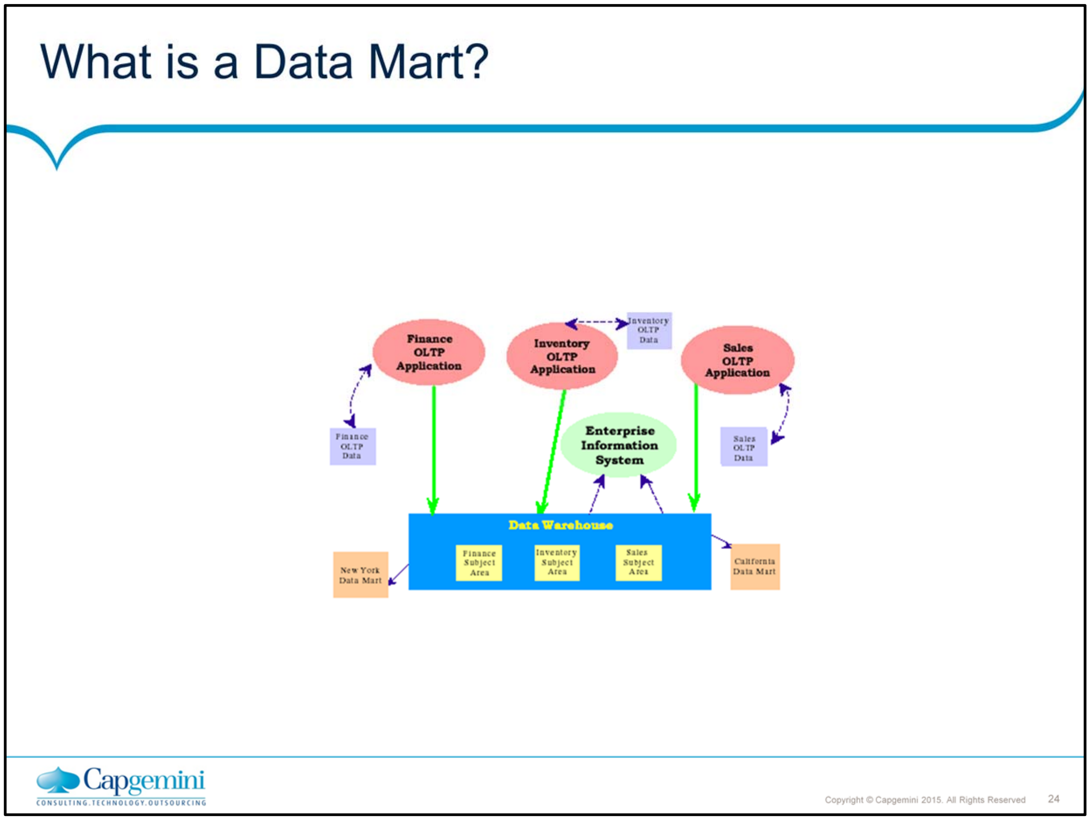


2.6: Data Mart

What is a Data Mart?

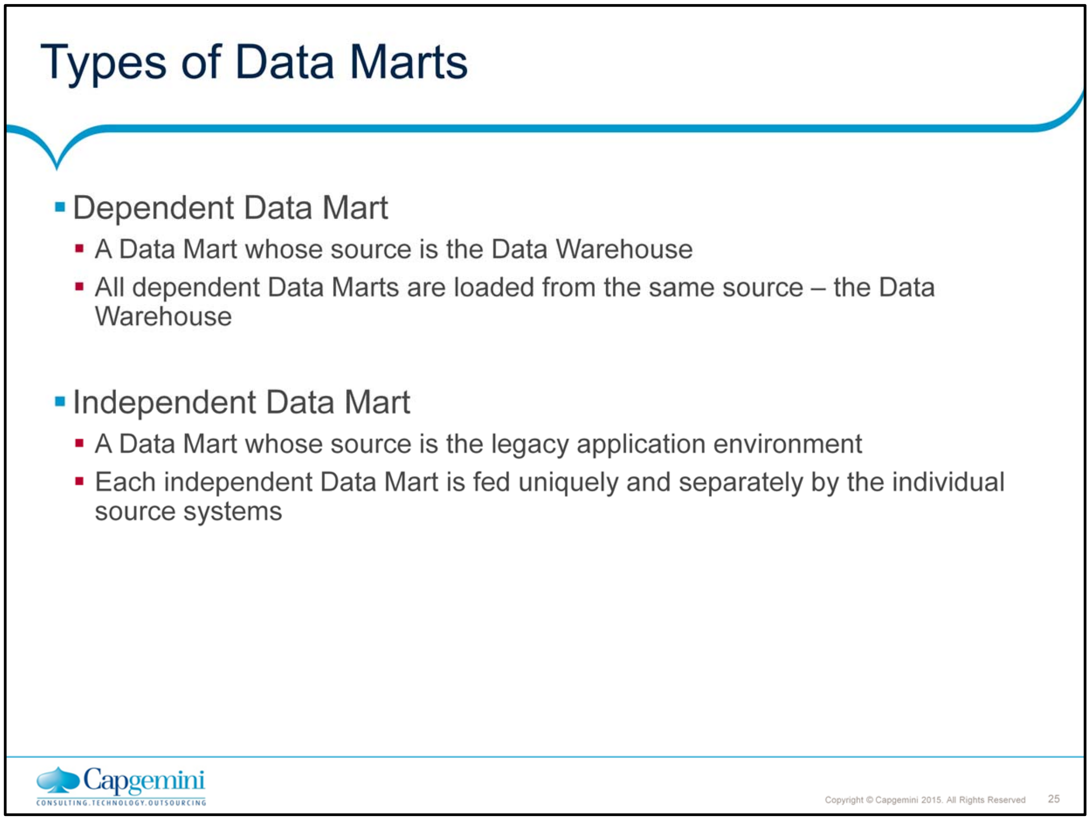
- Data Mart is a subset of the Data warehouse.
  - It is typically fed from the Data warehouse.
  - It is a Data warehouse that has limited scope.
  - It is a repository of data gathered from operational data and other sources.
  - It is used for decision making by a particular end-user group.
  - Emphasis is on meeting the specific demands of a particular group of knowledge users.
  - Maintain the ability to access the underlying base data.

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.



What is a Data Mart?

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.



Types of Data Marts

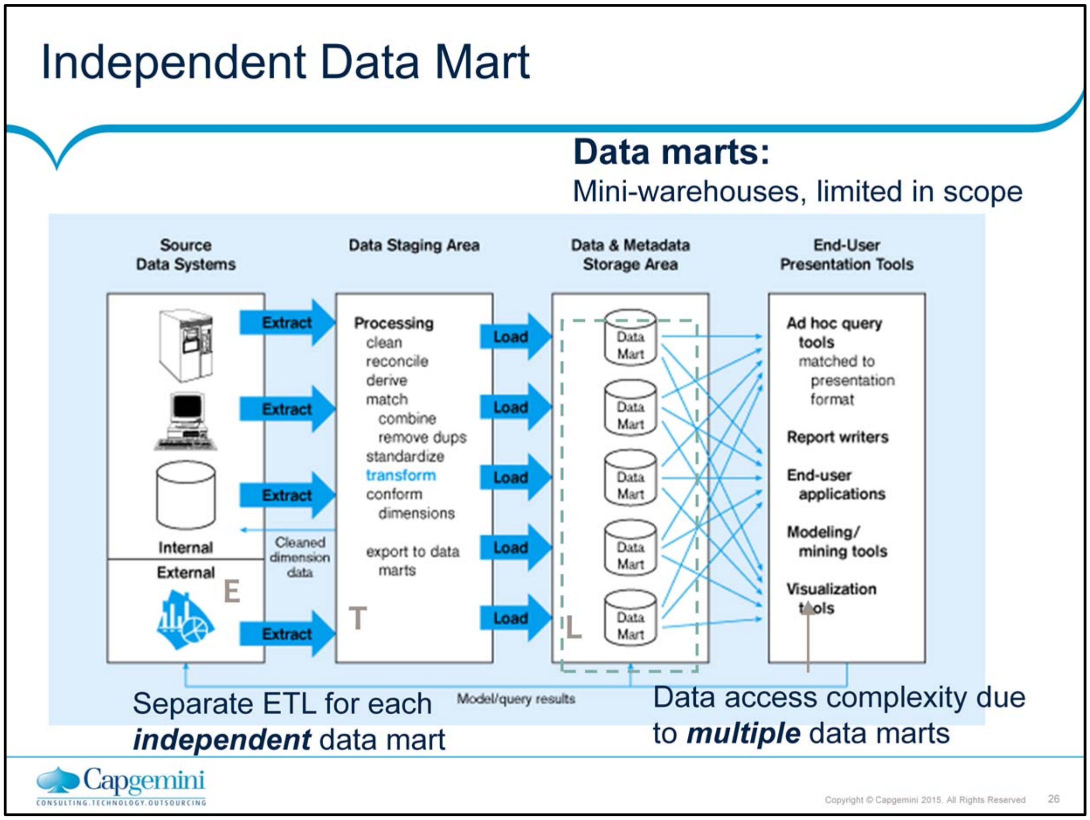
- Dependent Data Mart
  - A Data Mart whose source is the Data Warehouse
  - All dependent Data Marts are loaded from the same source – the Data Warehouse
- Independent Data Mart
  - A Data Mart whose source is the legacy application environment
  - Each independent Data Mart is fed uniquely and separately by the individual source systems

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:

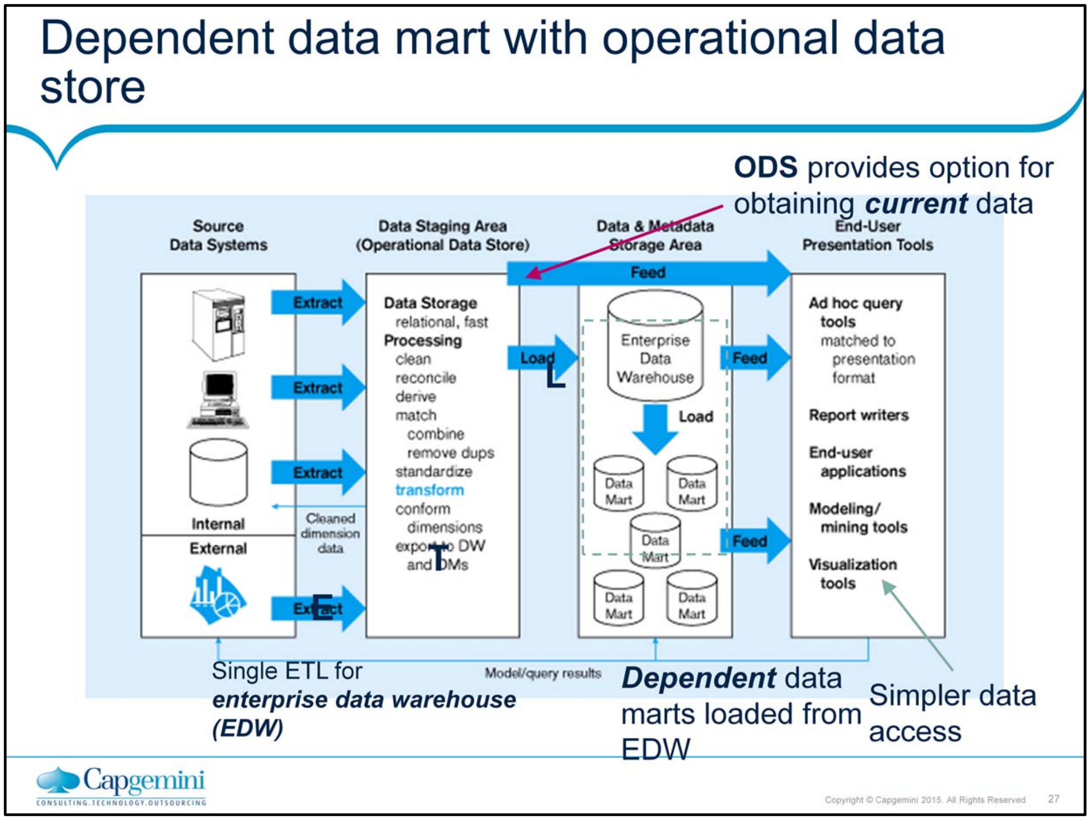
Mini-warehouses, limited in scope

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

ODS provides option for obtaining ***current*** data

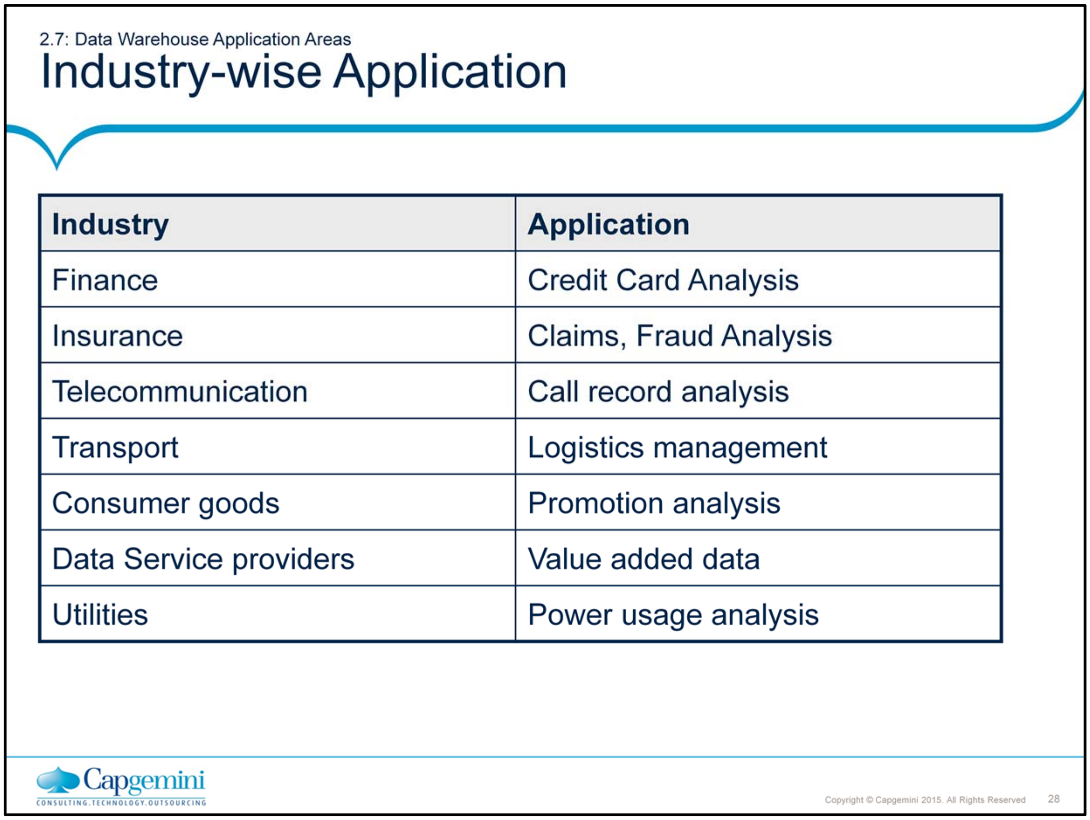
Single ETL for ***enterprise data warehouse (EDW)***

Dependent data marts loaded from EDW

Simpler data access

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

27

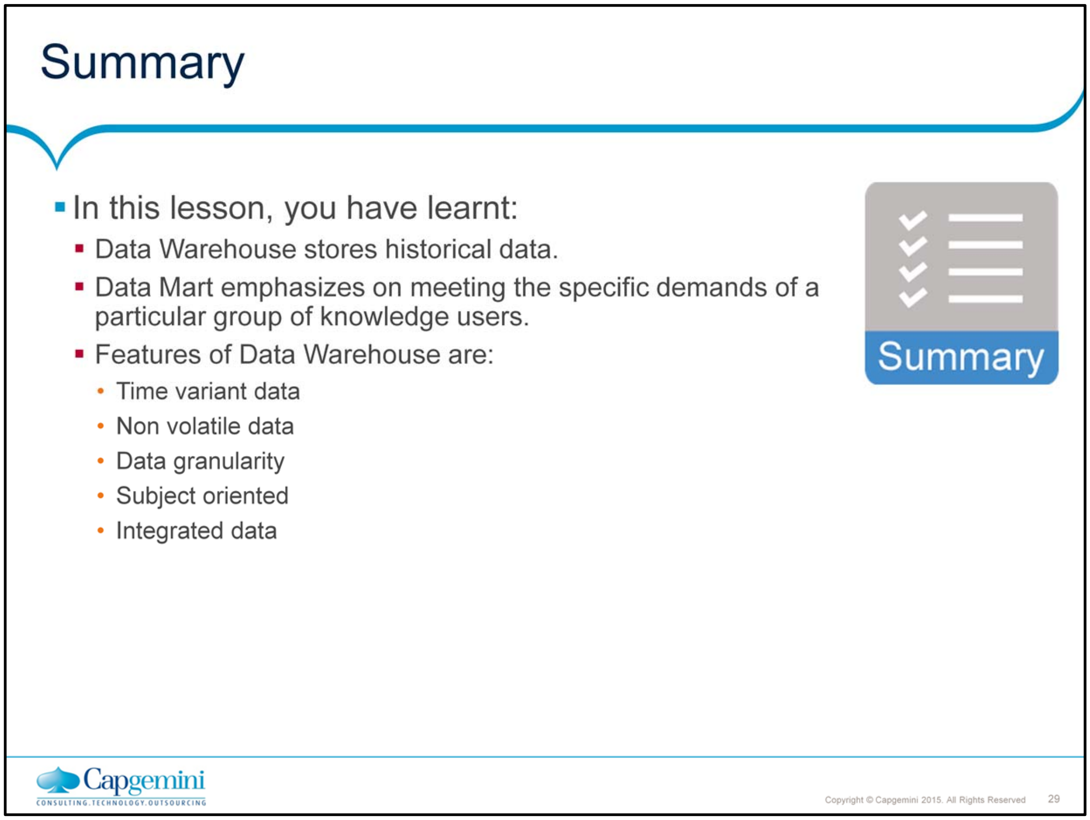


2.7: Data Warehouse Application Areas

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

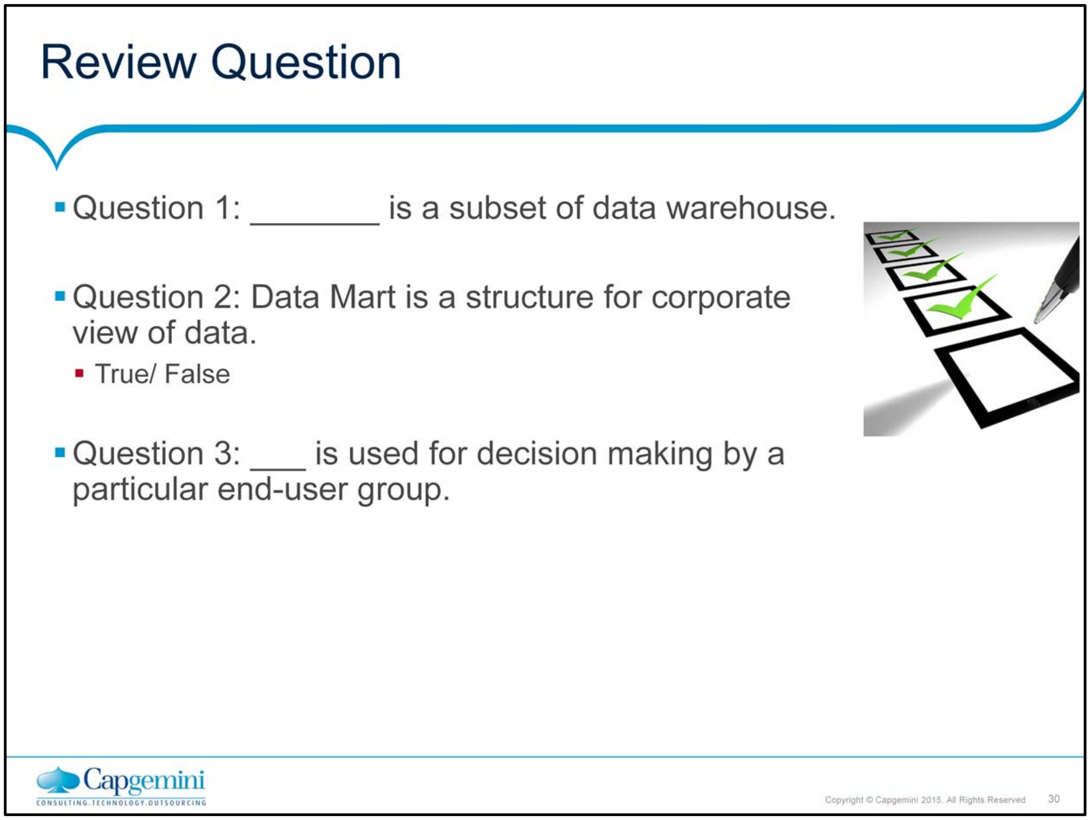


Summary

- 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

Summary
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.