INFS5710 IT Infra. for BA

Business Intelligence and Data Warehouses

Dr Vincent Pang

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Chapter 13

Business Intelligence and Data Warehouses

13-1 TO 13-7

Pp. 598 - 631

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

After three weeks of heavy material, this week is a bit easier. Most of the materials are on the slides! ☺

This week, we will be looking at Business Intelligence and Data Warehouse.

Database Systems Infrastructure
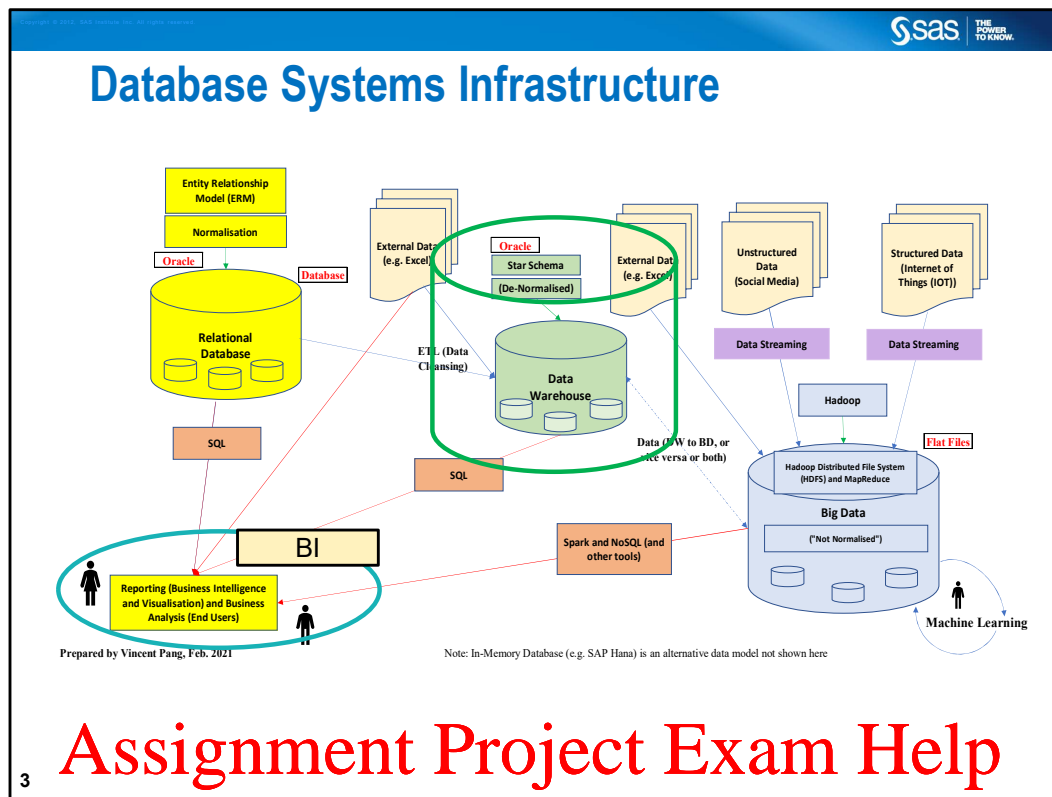
Coming back to this slide...

So far, we have looked at Normalisation and ERD, and today we will be looking at Business Intelligence (BI) and Data Warehouse (DW).

In the 1990s, Cognos and Business Objects and a few other vendors are moving in the direction of Business Intelligence (BI), Data Warehouse (DW) is the key of consolidated the data before populated to the BI applications.

The purpose of DW is the data is cleaned and I do not have to worry about cleaning the data again. Every day or every week I only have to clean and add the current data to the database.

The data were used in BI. Reports are printed or on screen for managers to review and make decisions.

Business Intelligence (BI)

- Comprehensive, cohesive, integrated set of tools and processes
  - Captures, collects, integrates, stores, and analyzes data
  - Generates and presents information to support business decision making
- Allows a business to transform:
  - Data into information
  - Information into knowledge
  - Knowledge into wisdom

4

From the previous diagram, we have seen what has to do with the users.

We have to gather data and turn the data into information, and usually on a report on paper or screen.

The decision maker will review and make a decision on the information on a report or they will come back and ask for more information.

As stated in the slide…

Business Intelligence (BI)

- Concepts, practices, tools and techniques to help business
  - Understand its core capabilities
  - Provide snapshots of the company situation
  - Identify key opportunities to create a competitive advantage
- Provides a framework for
  - Collecting and storing <u>operational data</u> and aggregating it into <u>decision support data</u>
  - Analysing <u>decision support data</u> and presenting generated information to end users to support business decisions
  - Making <u>business decision</u> which generates more data
  - Monitoring results to evaluate outcomes and predicting future outcomes with a high degree of accuracy

5

Data are collected and stored, and then the data will then be used to populate (i.e. added to) in DW.

As stated on the slide...
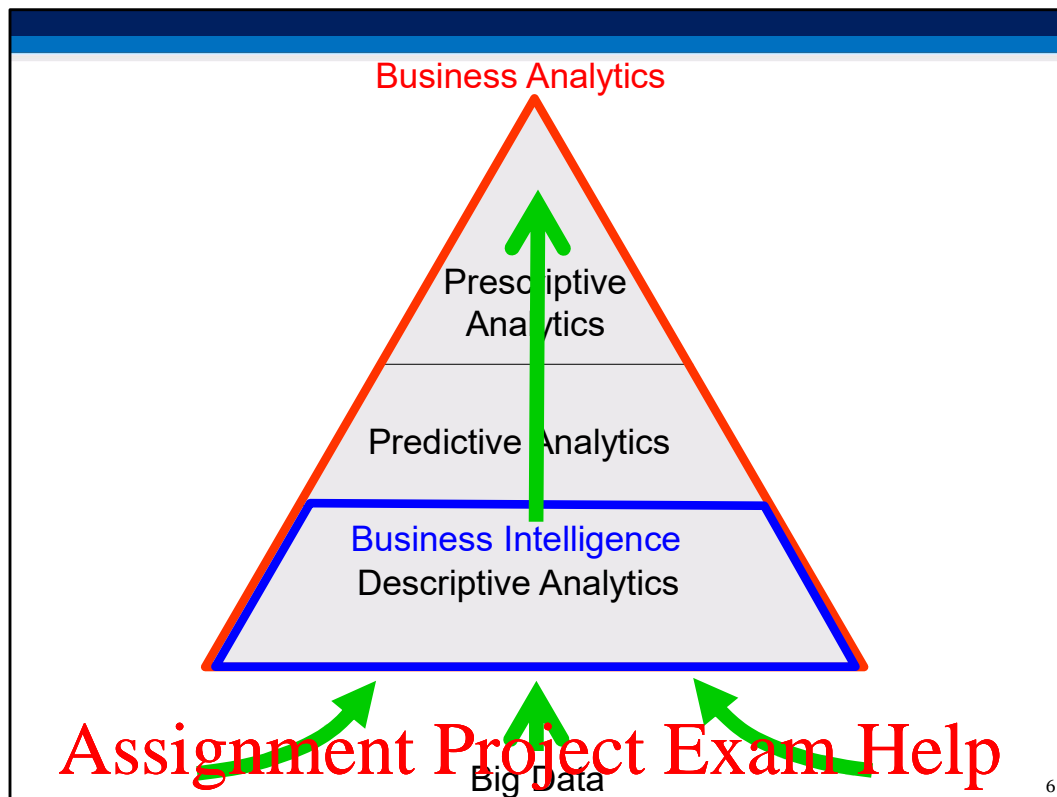
As stated in the slide.

There are about four to five analytics. There are other terms of analytics are used such as exploratory analytics or explanatory analytics.

Exploratory Analysis is the first step in the data analysis process looking what are available in the data and any interesting insight. This is the same when you look at your data too. You first might look at any interesting story and do you need to get more data to get the insights you are after.
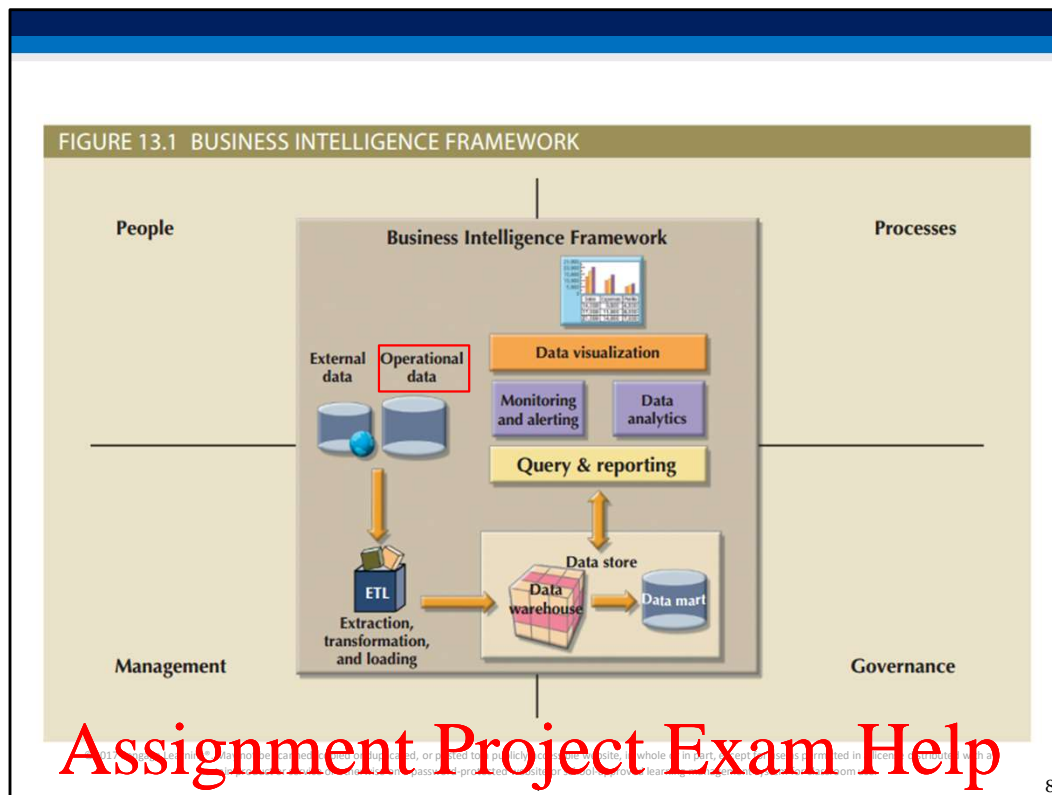
In explanatory analysis, you need data to support your insight and able to explain to others what you have done is supported by the data.

BI is usually related to Descriptive Analytics because in the 1990s and 2000s, the technologies drive what you can do!

Last Year Project: LA Bike-Sharing

FIGURE 13.1 BUSINESS INTELLIGENCE FRAMEWORK
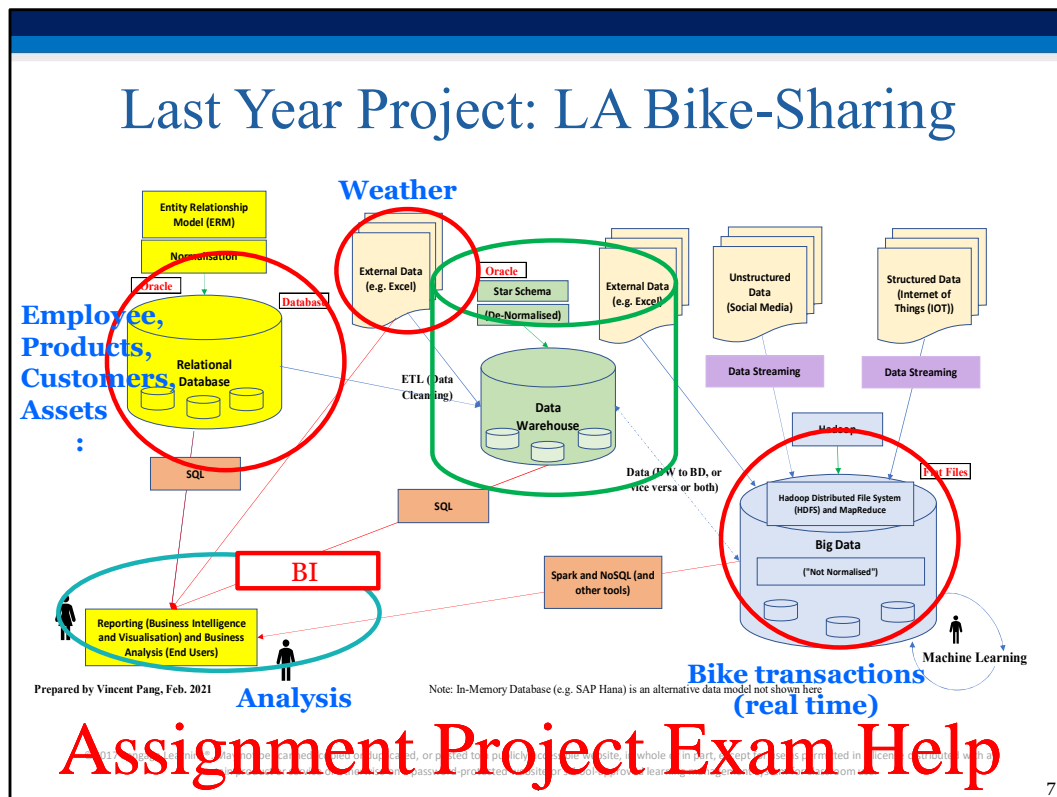
This is a typical BI framework, i.e. an IT infrastructure for BI. You have people, process, management and governance.

1. People are the technical people or users.

2. Processes are what you need to do get the report produced.

3. In this case, the data, external and operational, have to undergo ETL (Extract, Transform and Loading) which is to clean the data. Management refers to management of data.

- **Extract** is to get the data.
- **Transform** is to clean the data to make sure they are readable, blanks are filled and data are corrected to the correct format. For instance, you might want to consolidate "University of New South Wales", "U. New South Wales", "University of NSW", or even "Unsw" etc.. To become "UNSW".
- **Loading** is to load into DW, and there is usually a process of checking and evaluating the data uploaded into DW (Data store in the diagram), if they are not good or enough information, you do it again.

Data cleaning and **data cleansing** are two different terms used but they are the

same thing.

You can then run a query on the data and create reports, or do data analysis etc.

4. As for **governance**, this is to deal with how you data is to be managed. Who have access to the data, and what data is to be stored and where?

Read Table 13-2 for the BI architectural component.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

TABLE 13.3

SAMPLE OF BUSINESS INTELLIGENCE TOOLS

TOOL	DESCRIPTION	SAMPLE VENDORS
Dashboards and business activity monitoring	Dashboards use web-based technologies to present key business performance indicators or information in a single integrated view, generally using graphics that are clear, concise, and easy to understand.	Salesforce IBM/Cognos BusinessObjects Information Builders iDashboards
Portals	Portals provide a unified, single point of entry for information distribution. Portals are a web-based technology that use a web browser to integrate data from multiple sources into a single webpage. Many different types of BI functionality can be accessed through a portal.	Oracle Portal Actuate Microsoft SAP
Data analysis and reporting tools	These advanced tools are used to query multiple and diverse data sources to create integrated reports.	Microsoft Reporting Services MicroStrategy SAS WebReportStudio
Data-mining tools	These tools provide advanced statistical analysis to uncover problems and opportunities hidden within business data. Chapter 14 covers data mining in more detail.	SAP Teradata MicroStrategy MS Analytics Services

9

As stated on the slide.

TABLE 13.3

SAMPLE OF BUSINESS INTELLIGENCE TOOLS

TOOL	DESCRIPTION	SAMPLE VENDORS
Data warehouses (DW)	The data warehouse is the foundation of a BI infrastructure. Data is captured from the production system and placed in the DW on a near real-time basis. BI provides company-wide integration of data and the capability to respond to business issues in a timely manner.	Microsoft Oracle IBM/Cognos Teradata
OLAP tools	Online analytical processing provides multidimensional data analysis.	IBM/Cognos BusinessObjects Oracle Microsoft
Data visualization	These tools provide advanced visual analysis and techniques to enhance understanding and create additional insight of business data and its true	Dundas Tableau QlikView Actuate

A **dimension** is a structure that categorizes facts and measures in order to enable users to answer business questions. Commonly used dimensions are people, products, place and time.

10

As stated on the slide...

A **dimension** is a structure that categorizes facts and measures in order to enable users to answer business questions. Commonly used dimensions are people, products, place and time.

Use a rubric cube as an example, each side of the rubric cube is a dimension, so you can say you have a six dimensions. One dimension could be sales, another dimension could be states lot like NSW, Victoria, and so on, another dimension could be products, another dimension could be customers, and so on

Practices to Manage Data

- **Master data management (MDM)**: Collection of concepts, techniques, and processes for identification, definition, and management of data elements
to provide a comprehensive and consistent definition of all data within an organization
- **Governance**: Method of government for controlling business health and for consistent decision making
to ensure accountability of decision-making
- **Key performance indicators (KPI)**: Numeric or scale-based measurements that assess company's effectiveness in reaching its goals
profit, profit margin, employee turnover, graduation rate

11

As stated on the slide.

KPI is a popular measurement! To see how the business is going, or how a person is going with his work.

Practices to Manage Data

- **Data visualization**: Abstracting data to provide information in a visual format
  - Enhances the user's ability to efficiently comprehend the meaning of the data
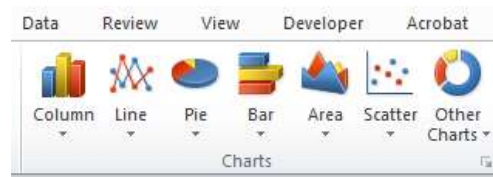 - Techniques:
    - Pie charts and bar charts
    - Line graphs
    - Scatter plots
    - Gantt charts — Project schedule
    - Heat maps — a representation of data in the form of a map or diagram in which data values are represented as colours.
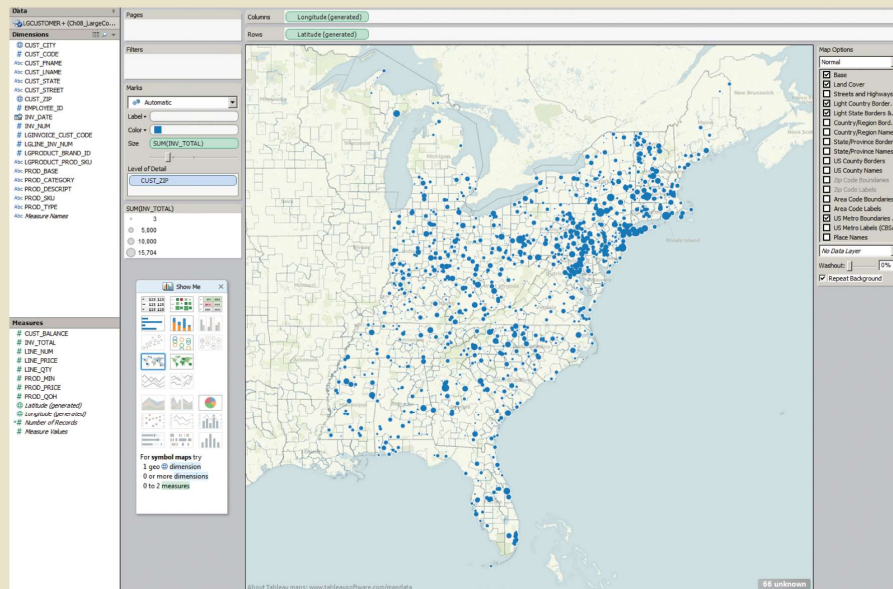
12

As stated on the slide.

In SAS EG and Studio, both have these tools. The graphs are not pretty, colours and fonts are not the best but you can see the results.

FIGURE 13.2 VISUALIZING SALES TOTAL BY zip CODE

Courtesy of Tableau
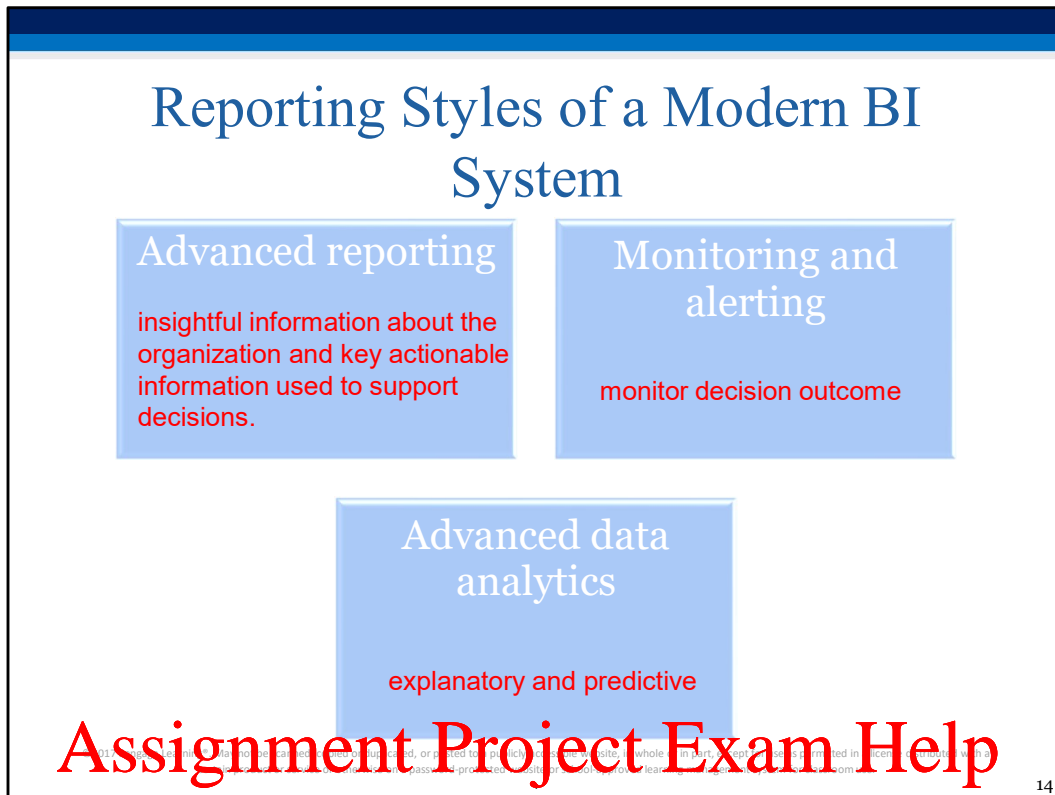
You can view the data by zip code – this can be done IN SAS/OD using SAS VA or SAS Viya.

Reporting Styles of a Modern BI System

Advanced reporting

insightful information about the organization and key actionable information used to support decisions.

Monitoring and alerting

monitor decision outcome

Advanced data analytics

explanatory and predictive

14

As stated on the slide...

Different types and purpose of different Reporting Styles of a modern BI System.

Monitoring could be to see how things are going, so you might have different colours such as red, amber, and green. Green means everything it's OK. Amber means something needs attention. Red means something needs attention.
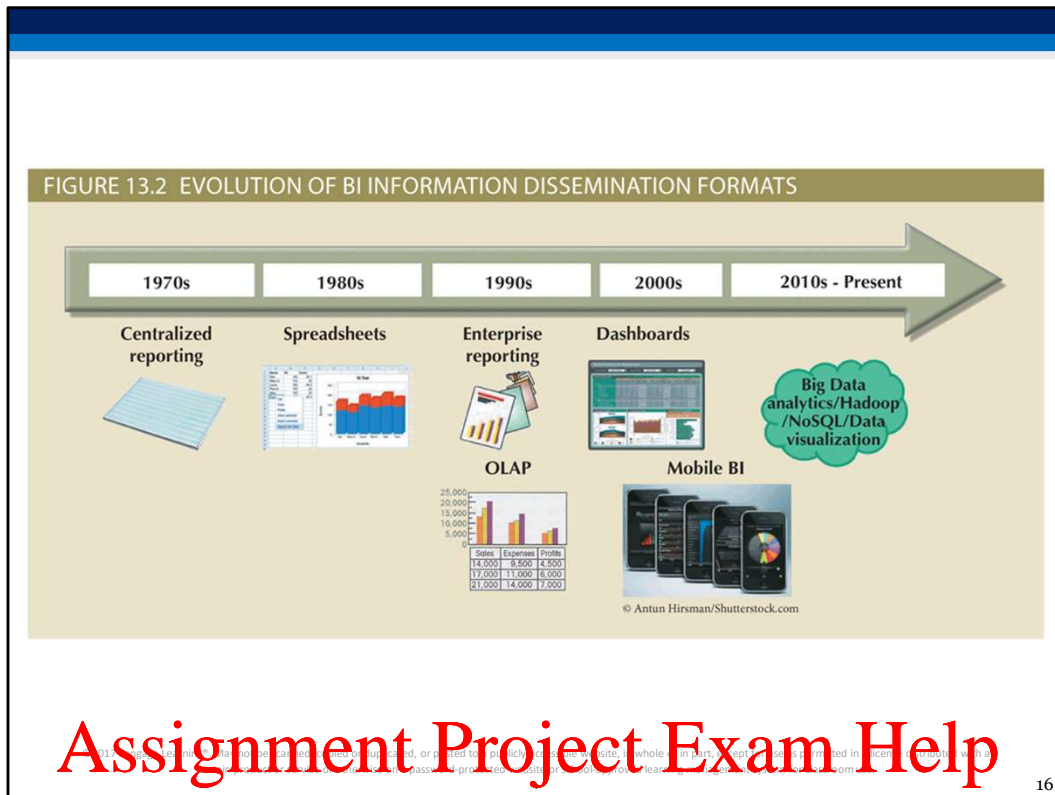
Business Intelligence Benefits

- Improved decision making

- Integrating architecture

  BI can be the integrating umbrella for a disparate mix of IT systems

- Common user interface for data reporting and analysis

- Common data repository fosters single version of company data

- Improved organizational performance
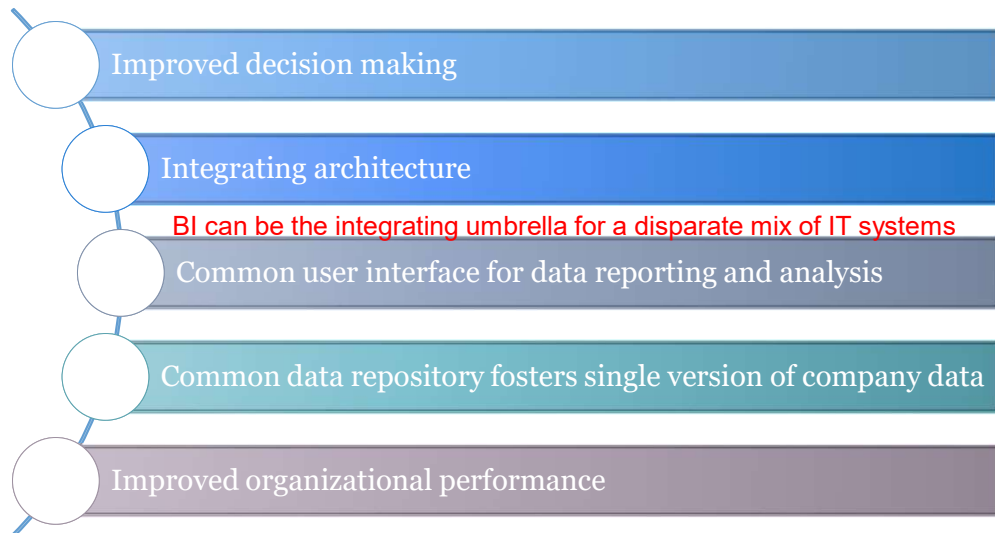
15

As stated on the slide...

Please read the textbook for BI Benefits.

FIGURE 13.2 EVOLUTION OF BI INFORMATION DISSEMINATION FORMATS

1970s — Centralized reporting
1980s — Spreadsheets
1990s — Enterprise reporting / OLAP
2000s — Dashboards / Mobile BI
2010s - Present — Big Data analytics/Hadoop /NoSQL/Data visualization

© Antun Hirsman/Shutterstock.com

As stated on the slide...

This diagram shows how we read data and report data! It changes over time due to improved in technologies and the internet. So now, instead of a company, we can do as an individual. For instance, how many steps did you do today, last week, last month? Your patterns of getting up and going to bed, and how is your wellbeing.

Another example is how much electricity produced by the solar panels of the house during the day. It might change the pattern of how a mother cooks Instead of cooking late in the evening, she might start to cook in the early afternoon because of "free solar electricity", and only heat up before dinner. She might instead do the washing at the early morning; she might do a bit late, say after kids are in school, because more electricity is generated.

Business Intelligence Evolution

TABLE 13.4

BUSINESS INTELLIGENCE EVOLUTION

SYSTEM TYPE	DATA SOURCE	DATA EXTRACTION/ INTEGRATION PROCESS	DATA STORE	END-USER QUERY TOOL	END USER PRESENTATION TOOL
Traditional mainframe-based online transaction processing (OLTP)	Operational data	None Reports read and summarized data directly from operational data	None Temporary files used for reporting purposes	Very basic Predefined reporting formats Basic sorting, totaling, and averaging	Very basic Menu-driven, predefined reports, text and numbers only
Managerial information system (MIS)	Operational data	Basic extraction and aggregation Read, filter, and summarize operational data into intermediate data store	Lightly aggregated data in RDBMS	Same as above, in addition to some ad hoc reporting using SQL	Same as above, in addition to some ad hoc columnar report definitions
First-generation departmental decision support system (DSS)	Operational data External data	Data extraction and integration process populates DSS data store Run periodically	First DSS database generation Usually RDBMS	Query tool with some analytical capabilities and predefined reports	Spreadsheet style Advanced presentation tools with plotting and graphics capabilities

ATM

A transaction is any action that reads from or writes to a database (SELECT, UPDATE, CREATE)

Relational DB

17

As stated on the slide...

A summary of BI Evolution.

TABLE 13.4

BUSINESS INTELLIGENCE EVOLUTION

SYSTEM TYPE	DATA SOURCE	DATA EXTRACTION/ INTEGRATION PROCESS	DATA STORE	END-USER QUERY TOOL	END USER PRESENTATION TOOL
First-generation BI	Operational data External data	Advanced data extraction and integration Access diverse data sources, filters, aggregations, classifications, scheduling, and conflict resolution	Data warehouse RDBMS technology Optimized for query purposes Star schema model	Same as above	Same as above, in addition to multidimensional presentation tools with drill-down capabilities
Second-generation BI Online analytical processing (OLAP)	Same as above	Same as above	Data warehouse stores data in MDBMS Cubes with multiple dimensions Multidimensional DB	Adds support for end-user-based data analytics	Same as above, but uses cubes and multidimensional matrixes; limited by cube size Dashboards Scorecards Portals
Third-generation Mobile, cloud-based, and Big Data	Same as above Includes social media and machine-generated data	Same as above Cloud-based	Same as above Cloud-based Hadoop and NoSQL databases	Advanced analytics Limited ad hoc interactions	Mobile devices: smartphones and tablets

18

As stated on the slide...

18

Decision Support Data

- **Decision support system (DSS)** is an arrangement of computerised tools used to assist managerial decision making
  - Typically has a much narrower focus and reach than a BI solution
- BI information technology has evolved from centralized reporting styles to the current, mobile BI and Big Data analytics style in the span of just a few years
  - The rate of technological change is not slowing down; technology advancements are accelerating the adoption of BI to new levels
- Effectiveness of BI depends on quality of data gathered at operational level
- **Operational data**
  - Seldom well-suited for decision support tasks
- Stored in relational database with highly normalised structures
  - Optimized to support transactions representing daily operations

As stated on the slide...

Decision Support Data

- Differ from operational data in:
  - Time span — Operational data short time frame; Decision support data longer
  - Granularity for decision support data
    - **Drill down**: <u>Decomposing</u> a data to a lower level
    - **Roll up**: <u>Aggregating</u> a data into a higher level
  - Dimensionality

    High for decision support data (e.g., data requirement for multi-periods, multi-location, multi-companies)

20

As stated on the slide...

FIGURE 13.4 TRANSFORMING OPERATIONAL DATA INTO DECISION SUPPORT DATA

You can transform operational data into decision support data by using Pivot Table in **Excel.**
Pivot Table is a cheap and free data tools to do **BI**.

- **Column A Year** becomes Time
- **Column B Region** remains as Region
- **Column C Agent** remains as Agent
- **Column D Product** remains as Product
- **Column E Value** becomes Sales

You then can run a report, say, Sales by product, region over the last two years (2014-2015).

Decision Support Database Requirements

organization of data

- Database schema

  - Must support complex, non-normalized data representations

  - Data must be aggregated and summarized

  - Queries must be able to extract multidimensional time slices

Slice-and-dice

Business Intelligence Framework

22
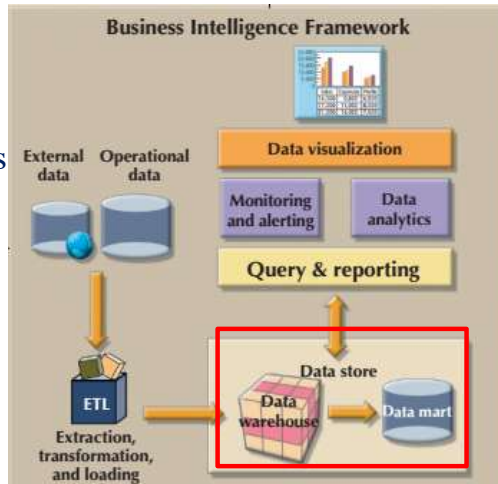
As stated on the slide...

We will talk more about data warehouse in a moment and that will fit in with this.

Decision Support Database Requirements

- Data extraction and loading (ETL)
    - Allow batch and scheduled data extraction
    - Support different data sources and check for inconsistent data or data validation rules
    - Support advanced integration, aggregation, and classification
- Database size should support
    - **Very large databases (VLDBs)**
    - Advanced storage technologies
    - Multiple-processor technologies

23

As stated on the slide...

Data Warehouse

24

FIGURE 13.5 THE ETL PROCESS
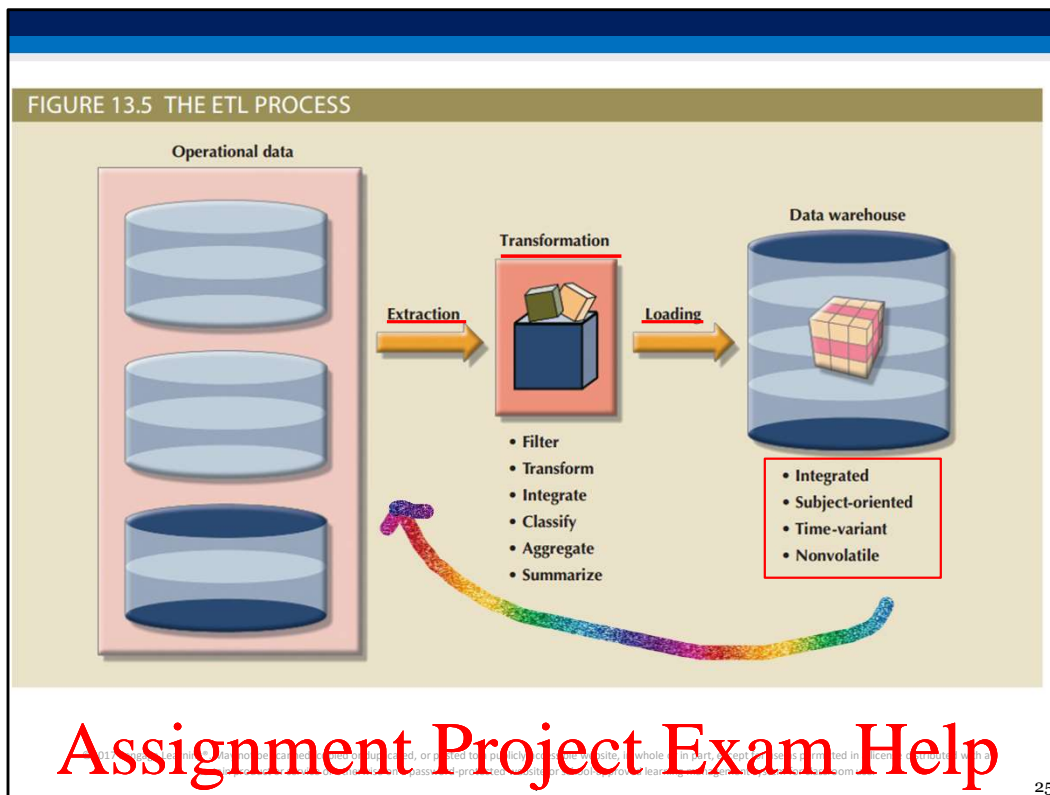
Remember, we talked about ETL a few slides back?

The step is associated with the Collection of data. You have plenty of transaction data like in a supermarket, they have several millions of transactions, probably most on toilet paper ☺ but you need to **extract** from the database to be cleaned and processed before the data is stored in the Data Warehouse (DW).

Your data might not be cleaned, so you will have to clean it. Once you cleaned the data, you might not want to do it again, so you want to save it somewhere – this is the data warehouse comes in. Thus, **transformation** is the process of doing this!

The transformation is the critical part of the whole process. As part of the cleaning process, you can transform the data in various way by ensuring all the columns are filled with data. New columns could be created, or new columns are created for aggregation or summarisation.

Once, everything is done, then you **load** the data into DW.

Next, we will talk about integrated, subject-oriented, time-variant and non-volatile in DW.

Table 13.8 - Characteristics of Data Warehouse Data and Operational Database Data

TABLE 13.8

CHARACTERISTICS OF DATA WAREHOUSE DATA AND OPERATIONAL DATABASE DATA

| CHARACTERISTIC	OPERATIONAL DATABASE DATA	DATA WAREHOUSE DATA
Integrated	Similar data can have different representations or meanings. For example, Social Security numbers may be stored as ###-##-#### or as ########, and a given condition may be labeled as T/F or 0/1 or Y/N. A sales value may be shown in thousands or in millions.	Provide a unified view of all data elements with a common definition and representation for all business units.
Subject-oriented	Data is stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, and credit amounts.	Data is stored with a subject orientation that facilitates multiple views of the data and decision making. For example, sales may be recorded by product, division, manager, or region.
Time-variant	Data is recorded as current transactions. For example, the sales data may be the sale of a product on a given date, such as $342.78 on 12-MAY-2016.	Data is recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.
Nonvolatile	Data updates are frequent and common. For example, an inventory amount changes with each sale. Therefore, the data environment is fluid.	Data cannot be changed. Data is added only *Read-only* periodically from historical systems. Once the data is properly stored, no changes are allowed. Therefore, the data environment is relatively static.

26

As stared on the slides.

The four characteristics of DW of integrated, subject-oriented, time-variant and non-volatile are on the slides.

Star Schema
organization of data

- Data-modeling technique — is the approach most widely used to develop data warehouse and dimensional data marts
- Maps multidimensional decision support data into a relational database
- Creates the near equivalent of multidimensional database schema from existing relational database
- Yields an easily implemented model for multidimensional data analysis

FIGURE 13.11 STAR SCHEMA FOR SALES
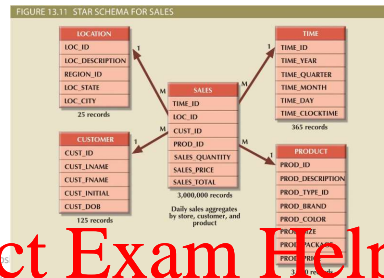
27

ERD is used in the development of a database.

To develop a data model for DW, we called it Star Schema!. It is different from ERD – the model is de-normalised.

It is called a star schema because it looks like a star! You will see later…

Don't worry, I am not going to ask you to create data model using Star Schema. ☺ If you are interested, I have written materials and they are available in the Workshop folder.

Data Marts

e.g., finance, sales, human resource

- Small, single-subject data warehouse subset
- Provide decision support to a small group of people
- Benefits over data warehouses e.g., By State – only people in that state can see the numbers.
  - Lower cost and shorter implementation time
  - Technologically advanced (a test system for a full data warehouse)
  - Inevitable people issues (minor resistance to changes)

28

Assignment Project Exam Help

https://powcoder.com

Data mart is trimmed to a small subset of DW contains data for a specific area and for a specific group of people. The size of the data mart is much smaller than data warehouse, so the access speed is much faster.

Add WeChat powcoder

For instance, **Finance data mart** for **finance department** or sales data mart for sales department.

It has benefits over larger DW – as stated on the slide…

Components of Star Schemas
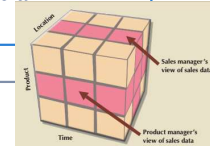
Facts — Sales data for example
- Numeric values that represent a specific business aspect

Dimensions — (Sales by) location, time, customer, products
- Qualifying characteristics that provide additional perspectives to a given fact — people, products, place and time

Attributes
- Used to search, filter, and classify facts
- **Slice and dice**: Ability to focus on slices of the data cube for more detailed analysis

Each dimension (above) has its own table with attributes.

Attribute hierarchies
- Provides a top-down data organization (e.g., region – state – city - store)

Drill-down and roll-up

29

For example for the sales star schema, you can have sales data in the Facts, and location, time, customer and product as dimensions.

For example the attributes for product could include Product_id, Product_descrtiption, Weight, Bar_code and so on.

There are attributes of the facts and dimension you have to add.

Slice and Dice

Sorry, the picture is not very clear!

Slice a cube, like cut a slide of cake. You will left with 2-dimensional data.

Dice means you can cut into smaller cubes! You still have 3-dimensional data.

To Roll-up and Drill Down

To roll-up, you can think of adding the data together as you move up, whereas

To drill-down, you want to look the data in more details.

How to create a cube?

IBM Cognos Cubes

https://www.ibm.com/support/knowledgecenter/en/SSD29G_2.0.0/com.ibm.swg.ba.cognos.tm1_prism_gs.2.0.0.doc/c_paw_modeling_cubes.html

See here on about Cubes – Cognos in the start of about cubes

Star Schema Representation

- Facts and dimensions represented by physical tables in data warehouse database
- Many-to-one (M:1) relationship between fact table and each dimension table
- Fact and dimension tables
  - Related by foreign keys
  - Subject to primary and foreign key constraints
- Primary key of a fact table
  - Is a composite primary key because the fact table is related to many dimension tables
  - Always formed by combining the foreign keys pointing to the related dimension tables *(see next slide)*
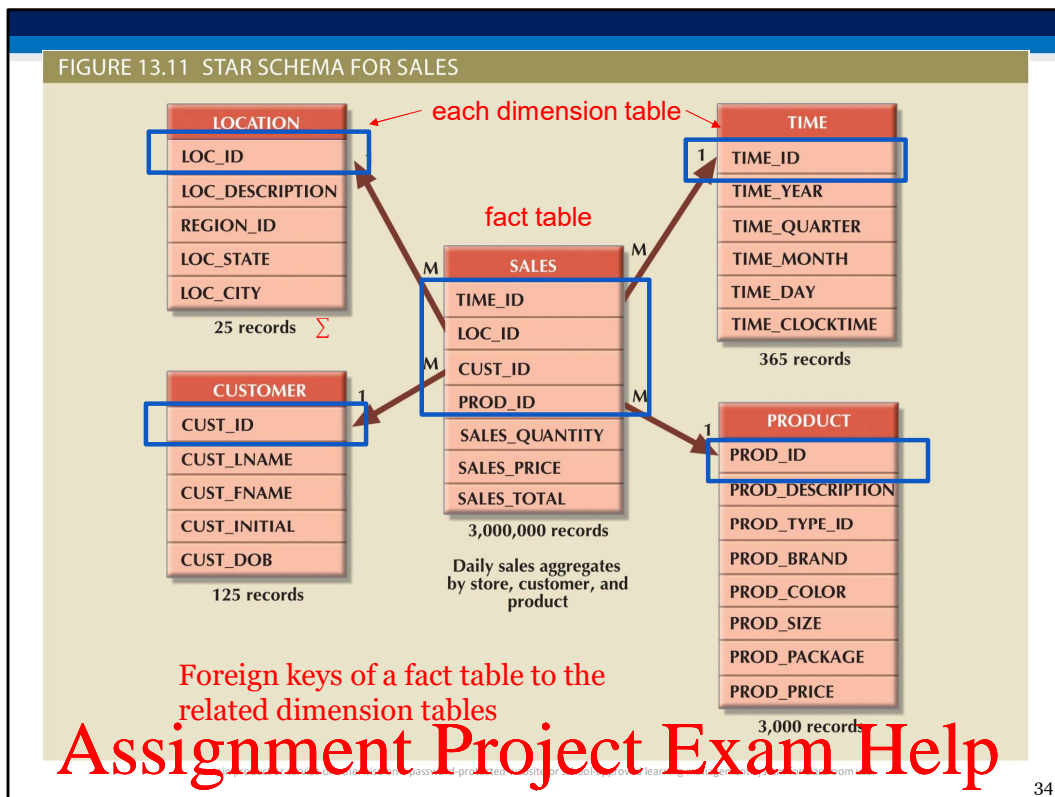
33

The key facts about Star Schema is shown on the above slide.

It is probably better to show as an example on the next few slides.

FIGURE 13.11 STAR SCHEMA FOR SALES

The fact table has four dimension tables. They are location (loc_id PK), time (time_id PK), customer (cust_id PK), and product (prod_id PK).

The fact table is sales, and the PK is made up of a composite key of four Ids, namely loc_id , time_id, cust_id and prod_id.

Look at the schema and try to work it out!

FIGURE 13.5 SIMPLE STAR SCHEMA
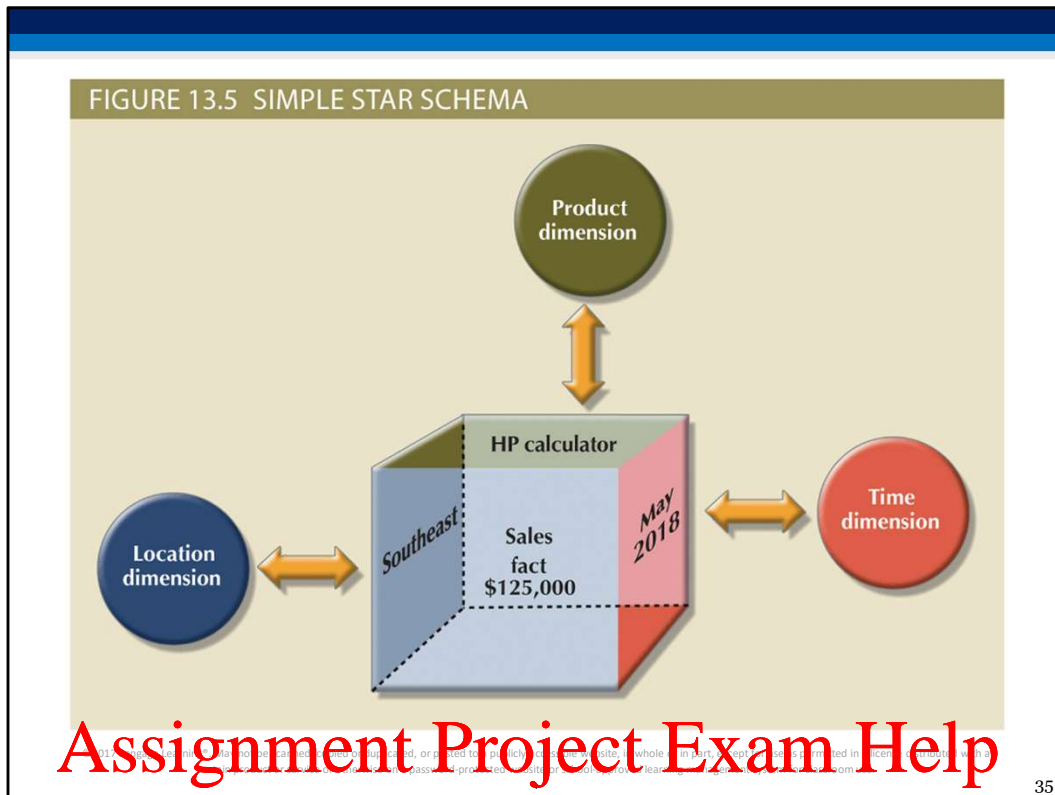
For example, from the previous slide:

Customer dimension: You (not shown here)

Product dimension: HP Calculator

Location dimension: Southeast

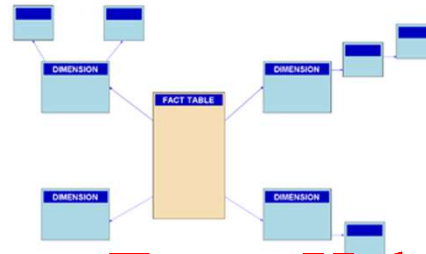Time dimension: May, 2018

Sales fact: $125,000

Performance-Improving Techniques for the Star Schema

- "Normalising" dimensional tables (might not be in 3NF)
- **Snowflake schema**: Dimension tables can have their own dimension tables
- Maintaining multiple fact tables to represent different aggregation levels
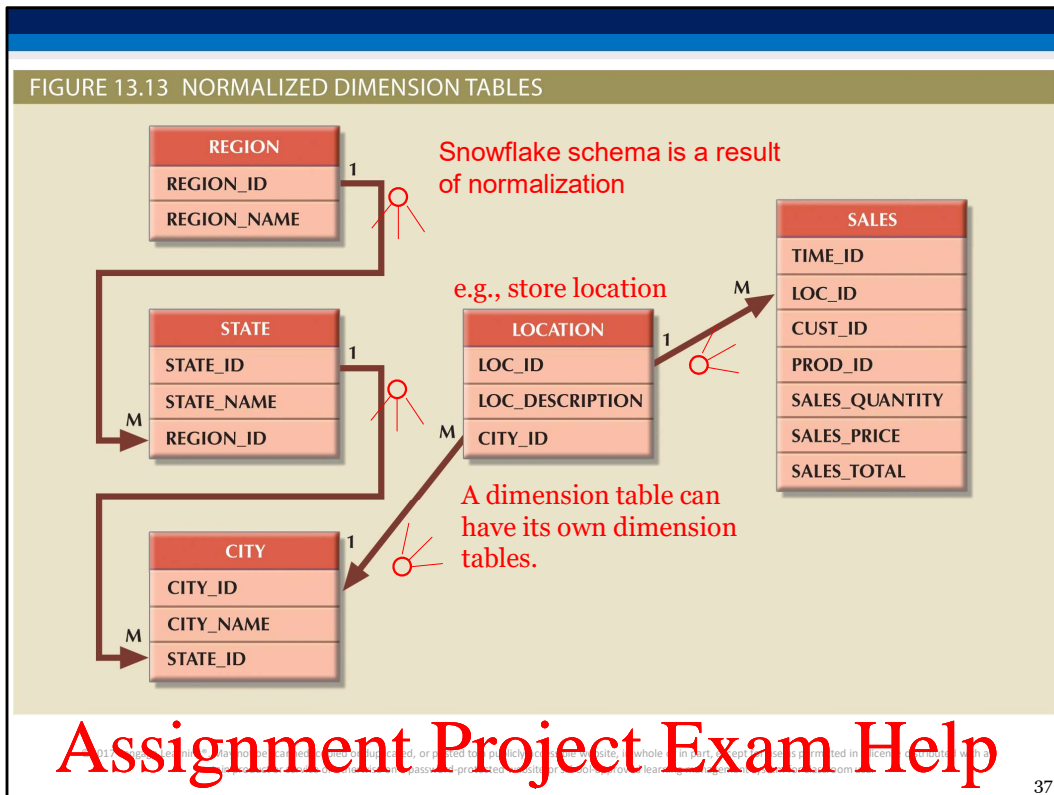- Denormalizing fact tables

Note: Star Schema is not normalised! It probably is in 2NF or 1NF! The reason it you want fast access to the data.
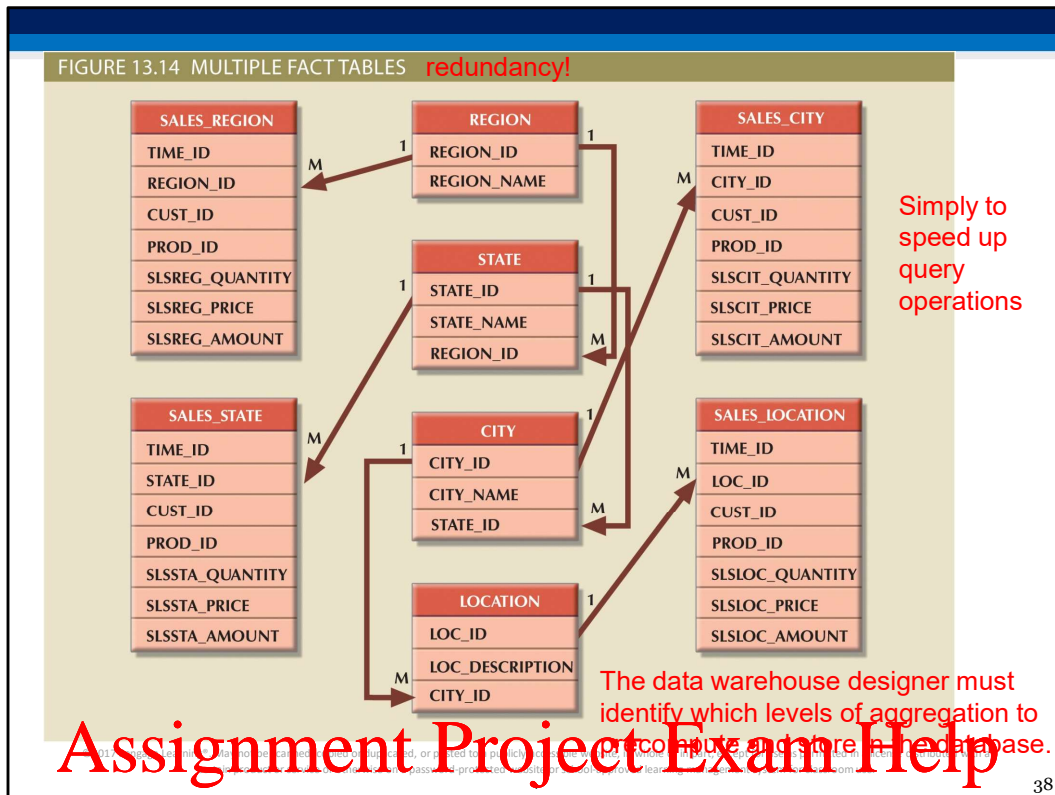
Sometimes, you want to do some normalisation in order to ensure the integrity of the data. Snowflake schema, evolved from Star Schema, allows some normalisation on one of the dimension table as shown in the snowflake.

FIGURE 13.13 NORMALIZED DIMENSION TABLES

Snowflake schema is a result of normalization

e.g., store location

A dimension table can have its own dimension tables.

For example, a snowflake schema with some normalizations.

FIGURE 13.14 MULTIPLE FACT TABLES redundancy!

Simply to speed up query operations

The data warehouse designer must identify which levels of aggregation to capture and store in the database.

You can have multiple fact tables and this will speed up accessing the data.

Performance-Improving Techniques for the Star Schema

- Partitioning and replicating tables
  - **Partitioning**: Splits tables into subsets of rows or columns and places them close to customer location

    or called fragmentation

    Distributed database
  - **Replication**: Makes copy of table and places it in a different location
  - **Periodicity**: Provides information about the time span of the data stored in the table

    data of current year, previous years, or all years stored in different tables

Why distributed database? (store data near users)
- *Internet as the platform for data access and distribution*
- *Mobile wireless revolution*

39

As stated on the slide.

We cover some of these terms next week.

39

Online Analytical Processing (OLAP)

- Advanced data analysis environment that supports decision making, business modeling, and operations research
- Characteristics:
  - Multidimensional data analysis techniques
  - Advanced database support
  - Easy-to-use end-user interfaces

As stated on the slide.

40

Multidimensional Data Analysis Techniques

- Data are processed and viewed as part of a multidimensional structure

- Augmented by the following functions:

  - Advanced data presentation functions 3D graphic, pivot table

  - Advanced data <u>aggregation, consolidation, and classification</u> functions Slice, dice, drill down, roll up

  - Advanced computational functions Market share, return, profit margin

  - Advanced data-modeling functions
    What if scenarios, predictive modeling
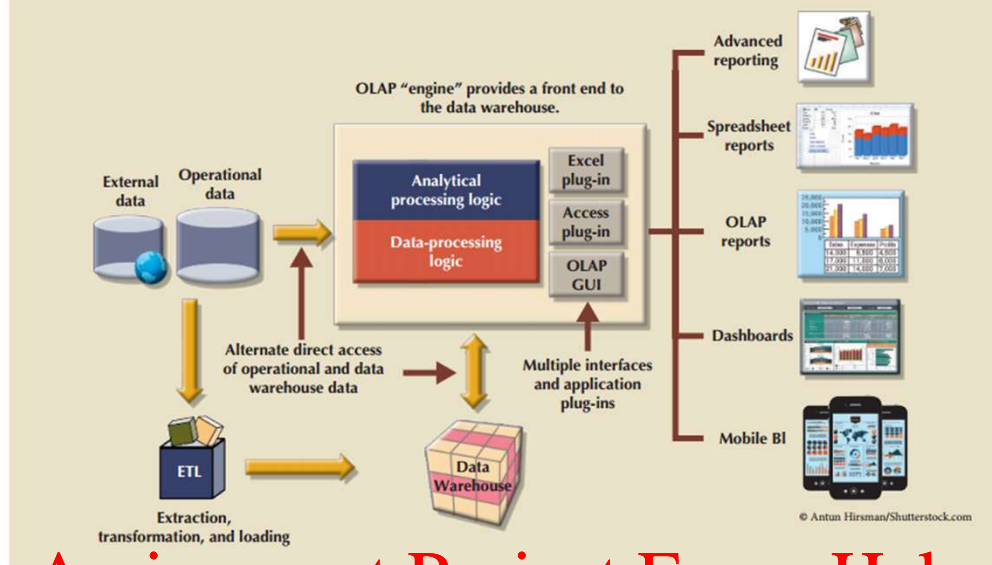
Assignment Project Exam Help

41

As stated on the slide.

https://powcoder.com

Add WeChat powcoder

Figure 13.17 - OLAP Architecture

FIGURE 13.17 OLAP ARCHITECTURE

The OLAP is virtually looks like DW except the data can come from either external data directly or from DW.

The data go directly to OLAP usually are the current or confidential data.

Once the data are gathered together, you can access directly to create reports or view on your Mobile.