Dimensional Modelling

Week 5

Introduction

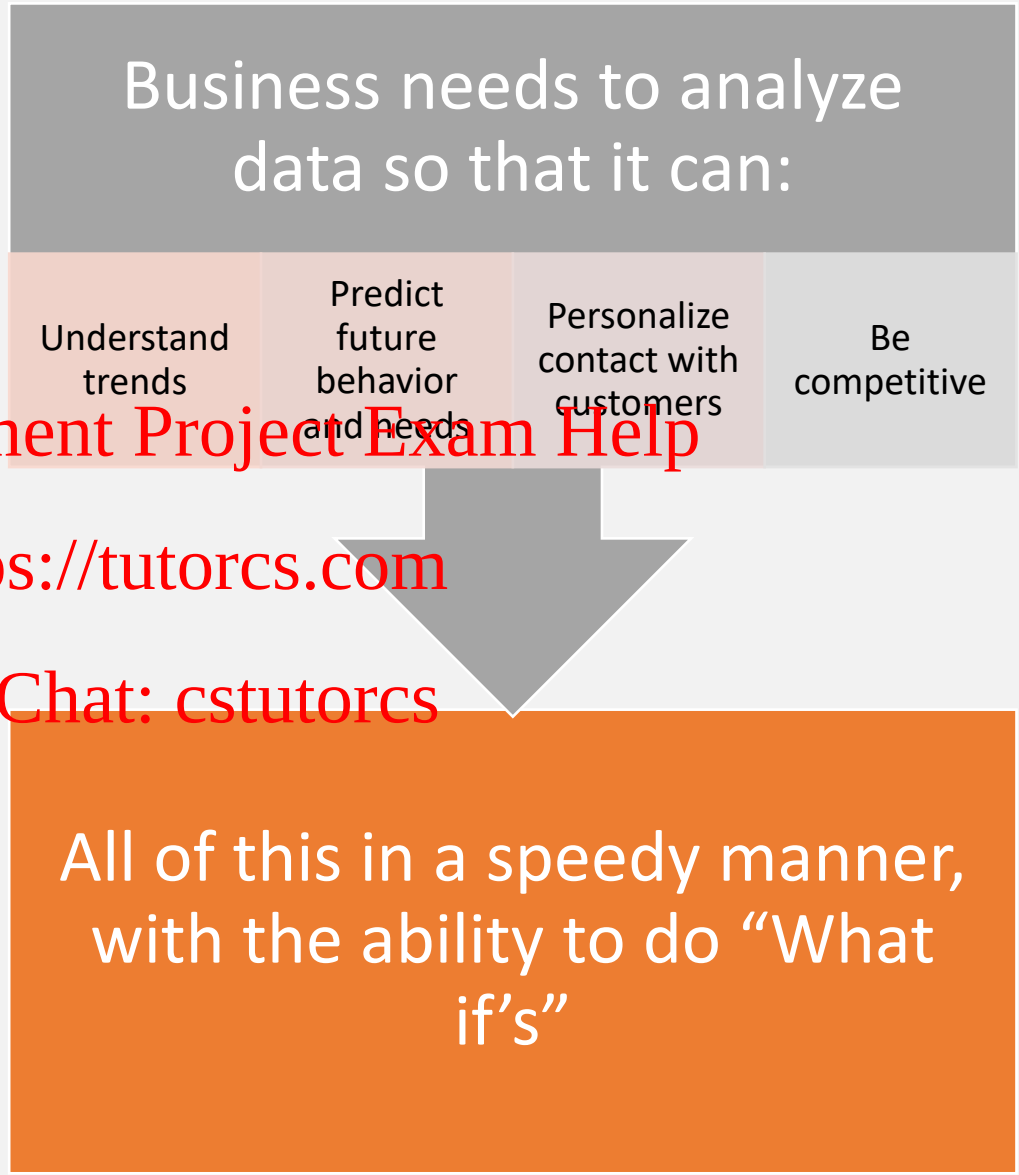
What is Dimensional Modelling?

- Data Dimensional Modelling (DDM) is a modelling/structuring technique
- It uses **Dimensions and Facts** to store the data in a Data Warehouse efficiently
- It optimises the database for faster retrieval of the data.
- Dimensional Models have a specific structure and organise the data to generate reports that improve performance

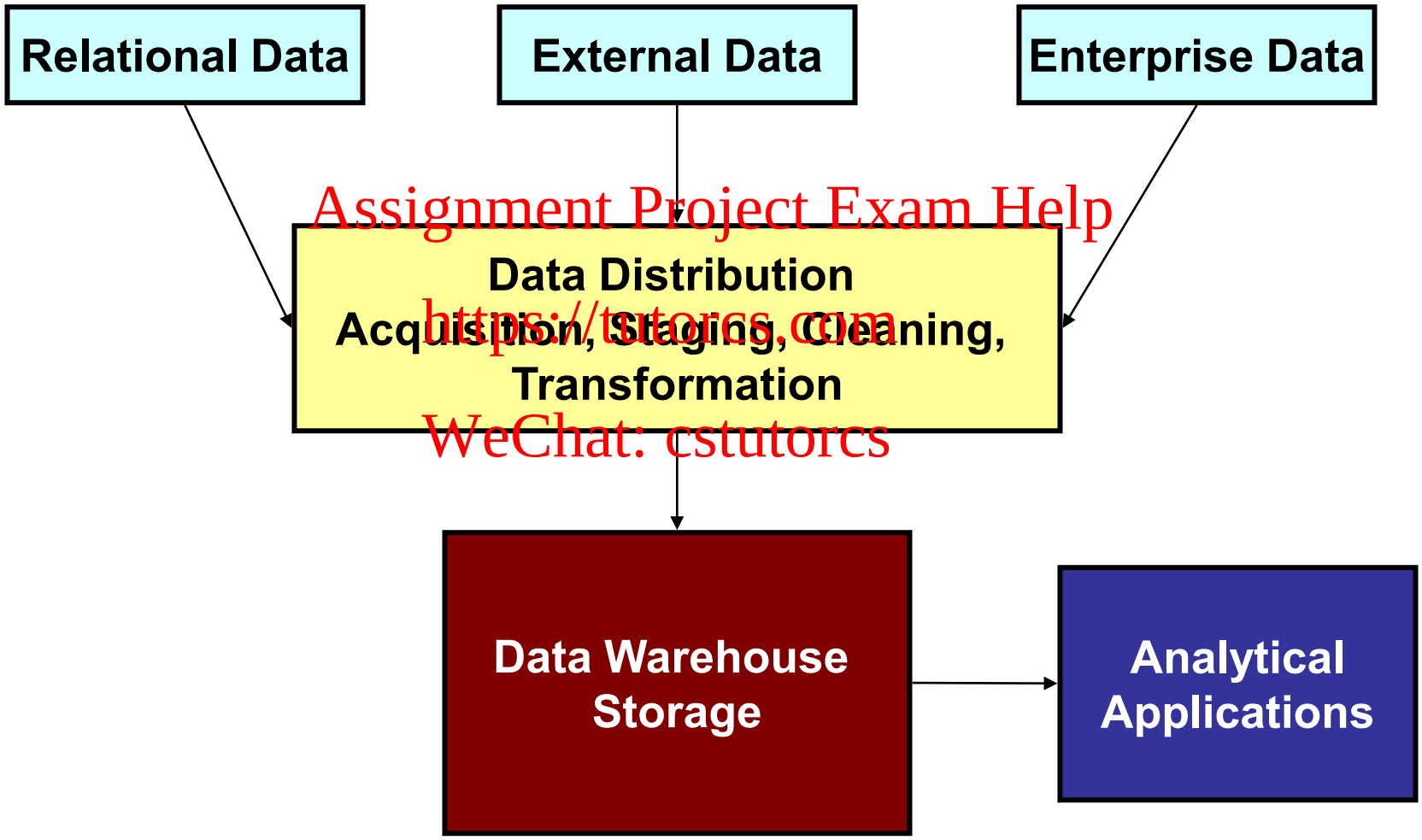
What is Dimensional Modelling?

- Dimensional Models have a specific structure and organise the data to generate reports that improve performance

Why Dimensional Data Warehouses ?



Dimensional Data Warehouse Architecture

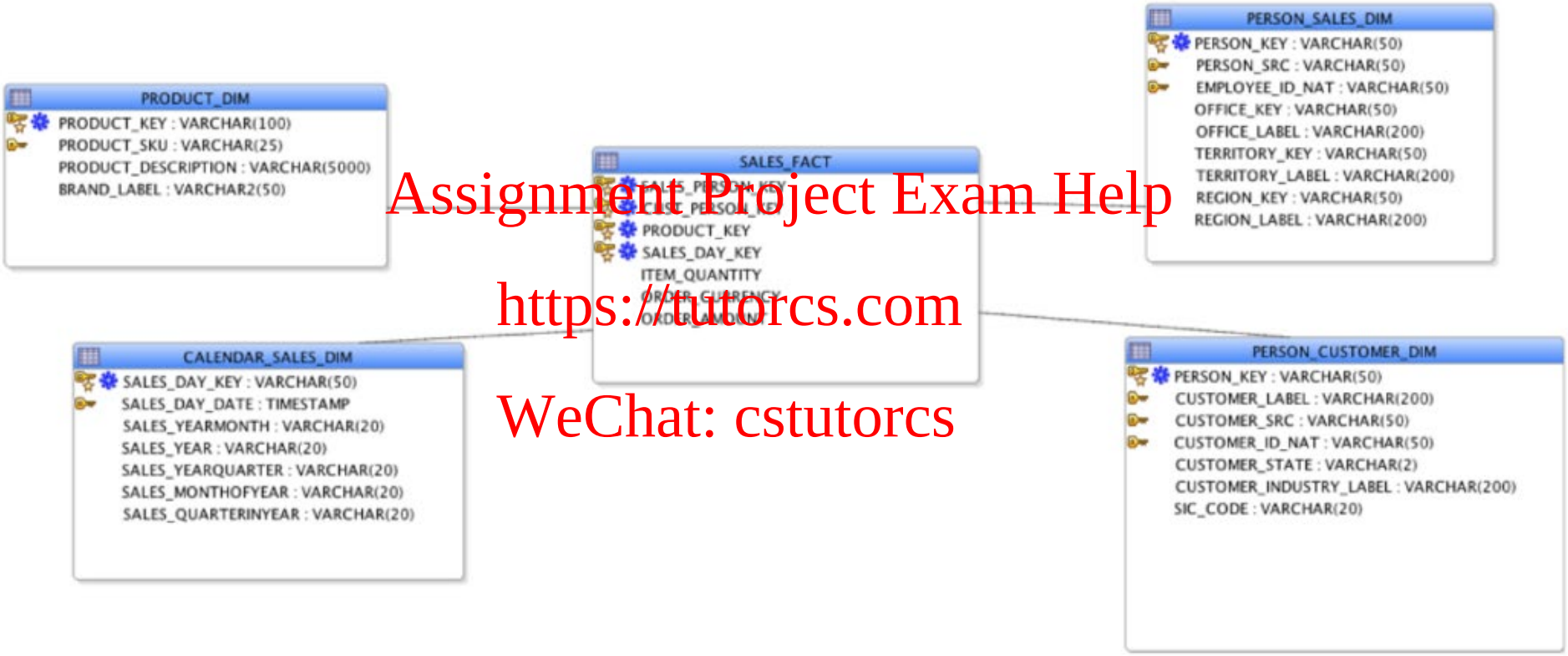


Dimension and Fact

- A fact:
  - A quantitative piece of information - such as a sale or a download.
  - Facts are stored in fact tables, and have a foreign key relationship with a number of dimension tables.

- Dimensions:
  - Are companions to facts, and describe the objects in a fact table.
  - Facts can be linked to multiple dimensions.

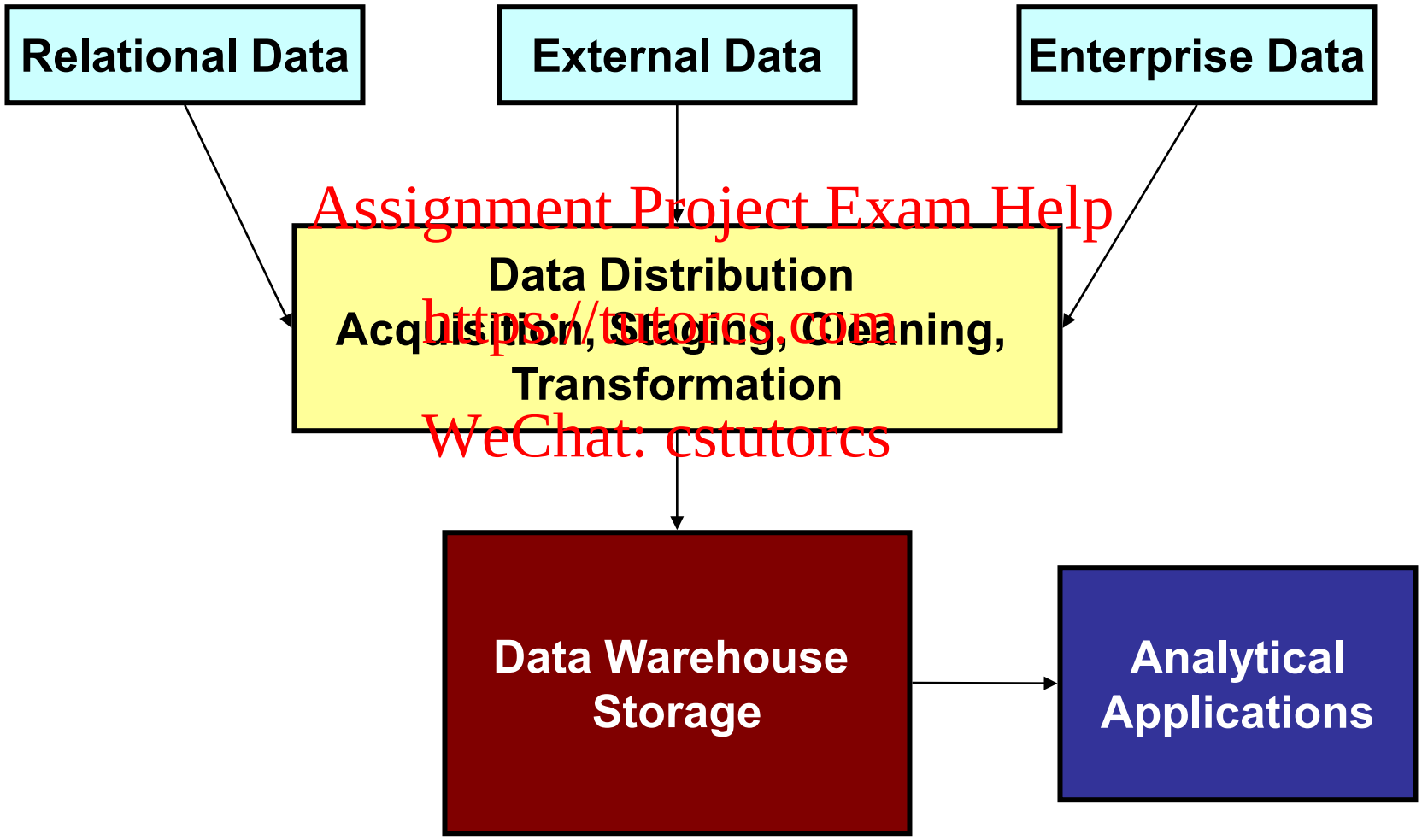
Dimension and Fact



What is a Data Warehouse?

- A database (and it can be built using a relational DB like Oracle)
- It is not a live / day-by-day database
- Day-by day transactions go to another database(s), usually fully relational databases fully normalized

Dimensional Data Warehouse Architecture



What is a Data Warehouse?

- A DW:
  - is updated at specific points in time
  - is mainly read-only (analytics)
  - is optimized for (read) performances
  - is a collection (integration) of different sources
- In our diagram on the previous slide
  - The “yellow” box (= the staging area) is permanent and it is where data are clean and integrated

Dimension Modelling

- Part of the Business Dimensional Lifecycle methodology developed by Ralph Kimball
- Includes a set of methods, techniques and concepts for use in data warehouse design.
- Focuses on identifying the key business processes within a business and modelling and implementing these first before adding additional business processes, a bottom-up approach.

DM Objectives

- Make information easily accessible
- Present information consistently
- Adaptable and receptive to change
- Present information in a timely way
- Protect information assets
- Serve as an authoritative and trustworthy foundation for improved decision making (single source of truth in data engineering language)
- Keep key stakeholders (VIPs) happy

Data Warehouse - Definition

- Subject-Oriented:
  - Data is organised around the major subjects of the enterprise (e.g. customers, sales, products) rather than application areas (e.g. invoicing, stock control etc.)
- Integrated:
  - Data from different sources are combined.
    - These sources may be inconsistent and formatted differently.
  - The DW establishes a consistent combined data source.

Data Warehouse - Definition

- Time Variant:
  - Data is only accurate at a particular point of time or over some time interval.
  - Time variances is shown in the extended time that data is held.
  - The implicit or explicit association of time with all the data and the fact that the data represents a series of snapshots.
- Non-Volatile:
  - Data is not updated in real-time but refreshed from operational data at regular intervals.
  - New data is already added as a supplemental to the database rather than as a replacement.
  - The database is constantly absorbs new data, incrementally integrating it with the previous data.

Data Warehouse - Requirements

- Accessibility:
  - Understandable- legible, meaningfully labelled
  - Intuitive and obvious to the business user – not just developers
  - Requires well-designed tools that are simple and easy to use in accessing data
  - Tractable – minimal wait time on data operations
- Consistency
  - Credible data – data must be clean and quality assured
  - Cross Business Process Compatible – a customer is always a customers, otherwise it should be labelled differently
  - Common definitions should be available for end users
  - Consistent information is high quality information that is accounted for and complete

Data Warehouse - Requirements

- Adaptive and Resilient
  - Tolerant to business changes (which are inevitable)
  - Warehouse changes should be graceful and should not invalidate existing data or applications
  - New case or busines cases should not disrupt existing applications
  - If changes to descriptive data cannot be avoided, appropriate measures must be in place to account for these changes
- Security
  - A warehouse contain business critical, sensitive, confidential and valuable information that may be harmful in the wrong hands
  - Requirements include:
    - Access control
    - Data distribution
    - Encryption
    - Redundancy
    - Etc

Data Warehouse - Requirements

- Improved decision making
  - Need the right data, visualization and analytical tools
  - There is only one true output of a DW – the decision made after viewing the evidence from the DW
  - Evidence should support decisions that deliver business impact and value
  - Decision support systems

Data Warehouse - Requirements

- Acceptability
  - User acceptance = success
  - Senior management must also buy in and support the increased use of this approach and technology
  - Requires that users trust the data
  - Tools must be intuitive

What is a Dimensional Model?

What is Dimensional Modeling?

- A logical design technique that seeks to present the data in a standard, intuitive framework that allows for high-performance access.
  - Can be implemented using a relational or a DBMS

What is Dimensional Modeling?

- Every dimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables.

What is Dimensional Modeling?

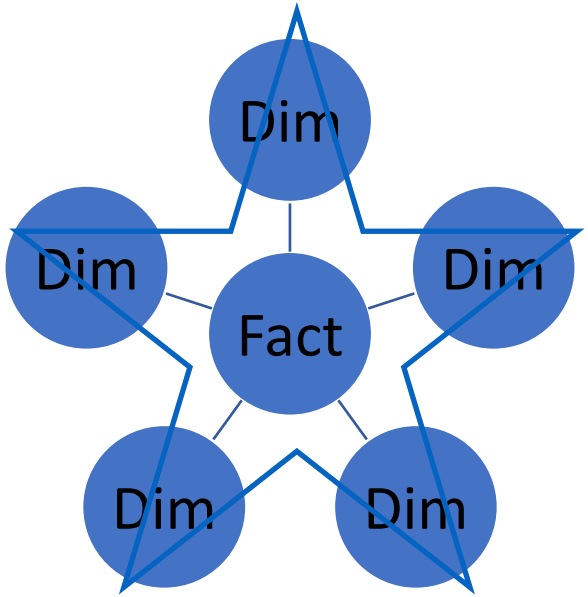
- Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table.

What is Dimensional Modeling?

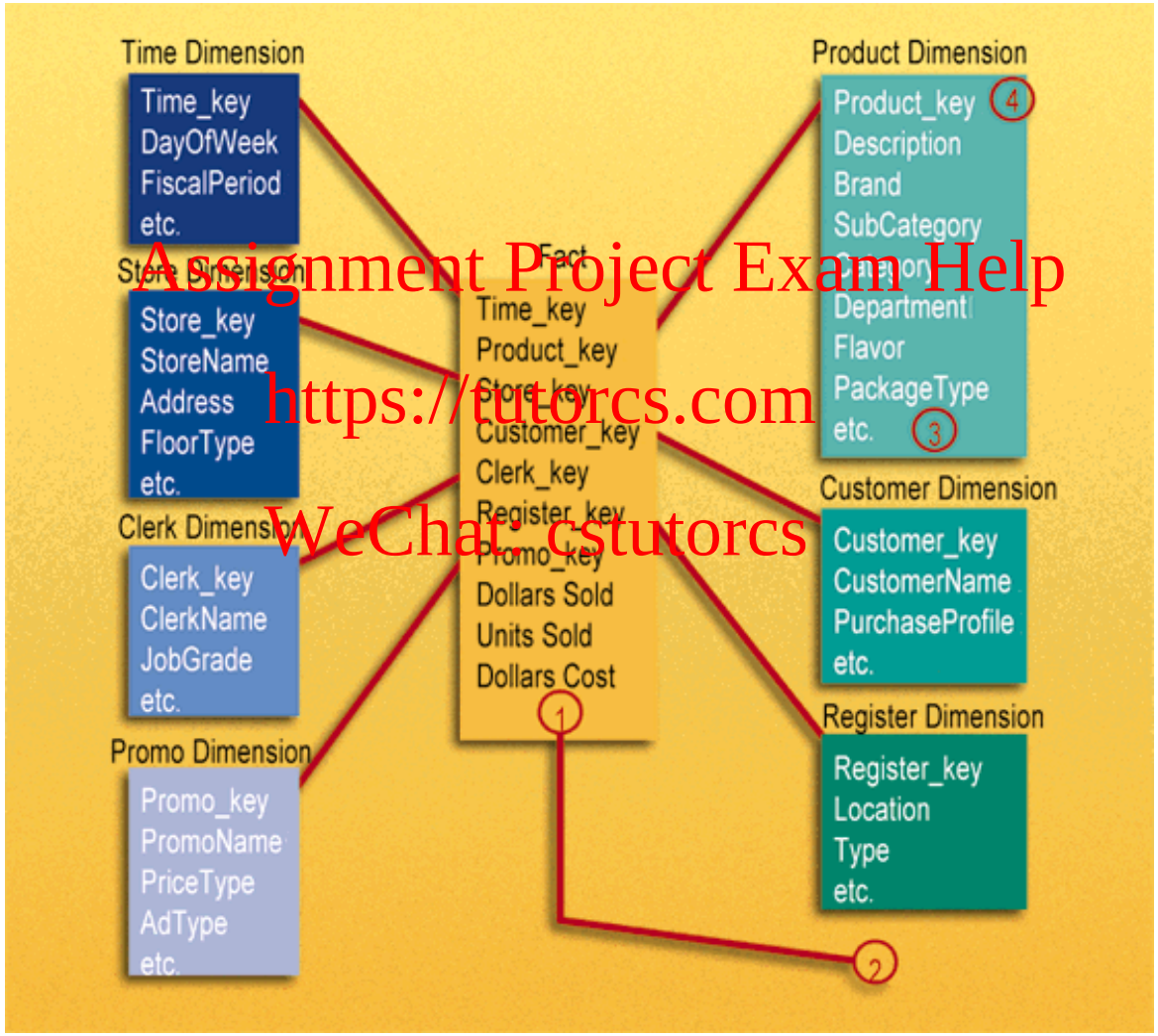
- This characteristic "star-like" structure is often called a star join.
  - The term star join dates back to the earliest days of relational databases.

Star Schema

- Singe data (fact) table surrounded by multiple descriptive (dimension) tables



Dimensional Model Example



Dimensional Schema

- Fact Tables
  - Contain related **measures**
    - Also called Measures.
    - Store quantifiable business data (such as sales, expenses, and inventory). ".
  - Usually, the largest tables
  - Usually appended to
  - Can contain detail or summary data
  - Measures are usually additive

Dimensional Schema

- Dimension Tables
  - Contain descriptors
  - Utilize business terminology
  - Textual and discrete data
  - Attributes through which the table measures are analyzed

Dimensional Schema – Fact Table

- A fact table contains information about things that an organization wants to measure.
- A fact table's key is made up from the keys of two or more parents.
- A fact always 'resolves' a many-to-many relationship between the parent, or dimension tables.
- The most useful fact tables also contain one or more numerical measures, or facts, that occur for the combination of keys that define each record.
- Example: the facts are Dollars Sold, Units Sold, and Dollars Cost.

Dimensional Schema – Fact Table

- The most useful facts in a fact table are numeric and additive.
- Additivity is crucial because data warehouse applications almost never retrieve a single fact table record;
  - Rather, they fetch back hundreds, thousands, or even millions of these records at a time, and often the most useful thing to do with so many records is to add them up.

Dimensional Schema – Dimension Table

- Dimension tables contain information about how an organization wants to analyze facts:
  - "Show me sales revenue (fact) for last week (time) for blue cups (product) in the western region (geography)
- Dimension tables most often contain descriptive textual information 'Blue cups', 'Western Region'
- Dimension attributes are used as the source of most of the interesting 'constraints' in data warehouse queries., and they are virtually always the source of the row headers in the SQL answer set

Dimensions vs Facts

Dimensions

- The time independent, textual and descriptive attributes by which users describe objects.
- Combining all the attributes including hierarchies, rollups and sub-references into a single dimension is denormalization.
- Often the “by” word in a query or report
- Not time dependent

Facts

- Business Measurements
- Most Facts are Numeric
- Additive, Semi-Additive, Non-Additive
- Built from the lowest level of detail (grain)
- Very Efficient
- Time dependent