1 Dimensional Design I

- Explain the dimension Design Approach
- Explain the Kimball's Enterprise Bus Matrix
- You're building a customer dimension per business requirement, your dimension must include City and Zip Code for grouping. The customer can have multiple addresses.

A)

- Explain the dimension Design Approach

Dimensional modelling is a design technique to structure data so that it is intuitive for business users (so that it refers to business concepts like: sales, customers, products, stores etc.) and delivers fast query performance.

Dimensions and facts

To do this, the world is divide into measurements (Facts) and context (Dimensions).

- Dimensions - describe business entities, usually textual data. Things we would like to know something about (customers, products, stores etc.);
- Facts - measures of properties, usually numeric data. Things that describe properties of dimensions (sales amount of a customer, sales quantity of a store etc.).

If we think from a star schema perspective, then:

- A fact records a property of multiple dimensions, for example: *a sale is related to a customer, a product, a store, a date and time.*
- A dimension may be measured by multiple facts, for example: *a customer can be related to one or multiple sales.*

Type of facts

- additive - can be added across all the dimensions, example: *sales;*
- semi-additive - can be added across most of the dimensions: *facts populated by regular snapshots like current stock, current account balance, stock level (meaning there is no sense to add yesterday's stock level to today's);* (typically time)
- fact-less - no measures, typically records and event.

Type of dimensions

- dynamic - the content of the dimension changes as the data in the source system updates, for example: *Product, Customer etc.;*
- static - the content of the dimension is populated once and forever, for example: *Date, Time.*

b)

▼ Explain the Kimball's Enterprise Bus Matrix

It is both a design tool and a project artefact. It is a simple representation of the dimensionality to be associated with your data warehouse/BI environment, a guide to the logical design phase, and a mechanism to communicate the data in the overall architecture back to the business. It lists out subject areas and dimensions of the environment we'll build-out and support.

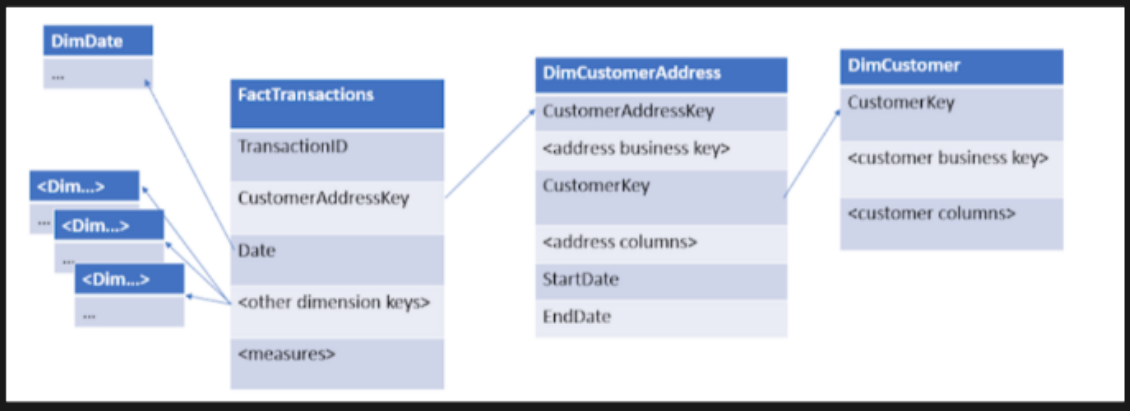
- -It shows the connection between dimensions and facts

AdventureWorks DW	DimAccount	DimCurrency	DimCustomer	DimDate	DimDepartmentGroup	DimEmployee	DimGeography	DimOrganization	DimProduct	DimProductCategory	DimProductSubcategory	DimPromotion	DimReseller	DimSalesReason	DimSalesTerritory	DimScenario
FactCallCenter				x												
FactCurrencyRate		x		x												
FactFinance	x	x		x	x			x								x
FactInternetSales		x	x	x			x		x	x	x	x			x	
FactInternetSalesReason		x	x	x			x		x	x	x	x		x		
FactProductInventory				x					x	x	x					
FactResellerSales		x		x		x	x		x	x	x	x	x		x	
FactSalesQuota				x		x									x	
FactSurveyResponse			x	x			x								x	

c)

▼ You're building a customer dimension per business requirement, your dimension must include City and Zip Code for grouping. The customer can have multiple addresses.

the DimCustomerAddress would be an SCD type 2. This will make the granularity of your new DimCustomerAddress be the "address history". I.E: I would assume that your business key for this dimension is the Customer ID, AddressID and Address type, so you should add a new row every time you get an address change; keeping tracking of the start and end dates that the address is valid. Your fact table should be related with DimCustomerAddress by the surrogate key "CustomerAddressKey '', and DimCustomerAddress should relate with DimCustomer using "CustomerKey". The Start and End dates can be used during the ETL time to populate the Fact table to determinate the right "CustomerAddressKey".



2 Dimensional Design II

▶ Explain the Dimensional Model and how you applied it into your hand-in.

▼ Explain the differences between DWH and DB.

▼ Draw a star schema from tables.

▼ Explain the Dimensional Model and how you applied it into your hand-in.

The accuracy in creating your Dimensional modelling determines the success of your data warehouse implementation. Here are the steps to create Dimension Model.

1. Identify Business Process.
2. Identify Grain (level of detail).
3. Identify Dimensions.
4. Identify Facts.
5. Build Star.

3 ETL Design I

- Practical (based on hand in): On an overall level, discuss the Kimball architecture for a Data Warehouse. Specifically, discuss the principles of design and implementation of ETL flows. Show how these principles were applied in your hand-in.
- Theoretical: Explain what a Bridge Table is. Give examples of a use case.
- Practical: Suppose we have a table in our source system that stores data on customers. The fields in that table are: First Name, Middle Name, Last Name, Title, Street Address, Street Number, City, Post Code, Region, Country. Which of these fields may be relevant to include as attributes on a customer dimension? Explain.

A)

▼ Practical (based on hand in): On an overall level, discuss the Kimball architecture for a Data Warehouse. Specifically, discuss the principles of design and implementation of ETL flows. Show how these principles were applied in your hand-in.

1. **Extraction** - the priority is to transact business efficiently. Keep extracts simple. You just copy data from source DB to the staging area. Not cleansing yet. You only decide what to keep in the DW and what not.

   Types of extraction:
   Full extraction - extract everything from source DB;
   Partial extraction - only changed or new data will be extracted since last update.
2. **Transformation** - once the data is extracted to the staging area, there are several potential transformations:
   Cleansing - correcting misspellings, resolving domain conflicts, dealing with missing elements or parsing into standard formats;
   Combining multiple sources;
   Deduplication;
   Assign warehouse keys.
3. **Load the presentation area** - that's the final step of the ETL process.
   - Load each data mart in bulk;
   - Index the newly arrived data for query performance;
   - Notify the user community.

B)

▼ Theoretical: Explain what a Bridge Table is. Give examples of a use case.

If we have a many-to-many relationship between a fact table and a dimensional table we can create bridge table between them. A good example would be: imagine having a case where a property might have multiple owners and a owner might have multiple properties

5 ETL Data Quality

- Explain about the different types of data quality ---- encountered in a data warehouse project. Show example from your hand in and what you did to address data quality issues.
- What is a Junk dimension?
- What if a thermometer was measuring 2 Celsius and then it suddenly spiked to 32, what should I do, what quality issue it violates.

A)

▼ Explain about the different types of data quality ---- encountered in a data warehouse project. Show example from your hand in and what you did to address data quality issues.

Data quality is the process of ensuring that the data is reliable and relevant.

Data quality

- accuracy - data represents what it is intending to represent in the real world, e.g. the age of a person is trusted or an address is the correct one by performing a sample measurement. Collect only from a verifiable source;
- completeness - the data is complete as a whole. Deliver all the required values that are available;
- consistency - same data from two or more source systems should not conflict with each other e.g. when a product is discontinued, there should not be any sales of the product;
- validity - data is valid with a range of values, e.g. an age of a person cannot contains a negative value and cannot be higher than 125 years;
- uniqueness - data identifies one and only one entry. Unique data means that there is only one instance of a specific value appearing in a data set, so it is free from data duplication, e.g. a Social Security number (SSN) to ensure that each person has a unique Id, therefore duplicate SSN values are not allowed within the data set;
- timeliness and availability - data is available when it is required. Timeliness is also about how frequently data is likely to change and for what reasons.

Improving the data quality

- data profiling - understanding the data. initial assessment of the current state of the data;
- data standardization - conform datasets to a common data format;
- matching or linking - identify and merge matching pieces of information together;
- data quality monitoring - data quality software in combination with machine learning can automatically detect, report and correct data variations based on predefined business rules and parameters.

B)

▼ What is a Junk dimension?

The reason for creating a junk dimension could be that there are some attributes in some dimensions that have only a few distinct values (low cardinality). The junk dimension would represent a cross `JOIN` (cartesian) product of every possible combination of the attributes from dimensions. It is also needed to add a surrogate key for every combination.

> 💡 Example: *DimMaritalStatus*, and *DimGender*, so instead of storing 2 keys, we can create a junk dimension with all the possible combinations of gender and marital status and store only one key.

C)

▼ What if a thermometer was measuring 2 Celsius and then it suddenly spiked to 32, what should I do, what quality issue it violates.

change the spiked value if it's an error with the median value

6 Slowly Changing Dimensions I

- ▶ Practical (based on hand in): Discuss the handling of change in a data warehouse context. Your discussions should include types of slowly changing dimensions as well as show examples from hand in.
- ▶ Theoretical: Explain the different types of keys we encounter in data warehousing and relate them to Slowly Changing Dimensions.
- ▶ Practical: Give an example of how you might define a test case to test Slowly Changing Dimensions Type 2.

A)

▼ Practical (based on hand in): Discuss the handling of change in a data warehouse context. Your discussions should include types of slowly changing dimensions as well as show examples from hand in.

A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse. It is considered and implemented as one of the most critical ETL tasks in tracking the history of dimension records.

1. Type 1 - whenever the data updates in the source system, we overwrite the existing record. This way we do not keep track of the history of changes.
2. Type 2 - whenever the data update in the source system, we insert a new record for the new data. Using this method, we can keep track of the history of changes because the business key from the source system will always remain the same, in contrast to the surrogate key which will keep updating each time. Querying the business key, will return all versions of the same product.
3. Type 3 - each time the data update, we create a new column, for the new value. This is most useful when we only need to remember the previous and new value for some time, before we fully transition to the new one.
4. Type 4 - this approach is used for more rapidly changing dimensions. In this case, a rapidly changing column is moved to a separate dimension table, eliminating the unnecessary volume from the main dimension, still being able to perform the required analysis.

B)

▼ Theoretical: Explain the different types of keys we encounter in data warehousing and relate them to Slowly Changing Dimensions.

- Primary key (PK) - a column or columns in a database table that uniquely identify each row in a table.
- Foreign key (FK) - a column in a relational database table whose value are drawn from the values of a primary key in another table. In a dimensional model, the components of a composite fact table key are foreign keys with respect to each of the dimension tables.
- Composite key - key in a database table made up of several columns. Same as concatenated key. The overall key in a typical fact table is a subset of the foreign keys in the fact table.
- Natural key - meaningful value that identify records, such as social security numbers that identify specific customer, SKU numbers in a product dimension. In some cases, natural keys are unique identifiers and can serve as primary keys.
- Surrogate key - (usually) integer keys, that are sequentially assigned as needed in the ETL system. In a dimension table, the surrogate key is the Primary key. A surrogate key cannot be interpreted by itself. Surrogate keys are required in many data warehouse situations to handle slowly changing dimensions.
- Business key - each dimension table consists of its business key attribute, the one that uniquely identifies a row in the table.

  In a **slowly changing dimension**, the destination table will have two **types of keys** , the surrogate **key** which will tie out to the fact table, and the business **key**, which identifies the record from the source.

C)

▼ Practical: Give an example of how you might define a test case to test Slowly Changing Dimensions Type 2.

1. Verifying the Current Data
2. Verifying the uniqueness of the key columns in the SCD
3. Verifying that historical data is preserved and new records are getting created

[Testing Type 2 SCD using ETL Validator| Datagaps]

8 Data Analytics and Visualization on Dimensional Models I

- ▶ Explain the Dimension Design approach and show how you used this in your hand-in. Explain S-O-R model. Explain with example from your hand-in.
- ▶ Explain the concept of pre-attentive attributes. Discuss the importance in relation to dashboard design.
- ▶ You are tasked with creating a visualization to show the relationship between time spending on learning paths and grades. What kind of visualization are you going to show and why?

A)

▼ Explain the Dimension Design approach and show how you used this in your hand-in. Explain S-O-R model. Explain with example from your hand-in.

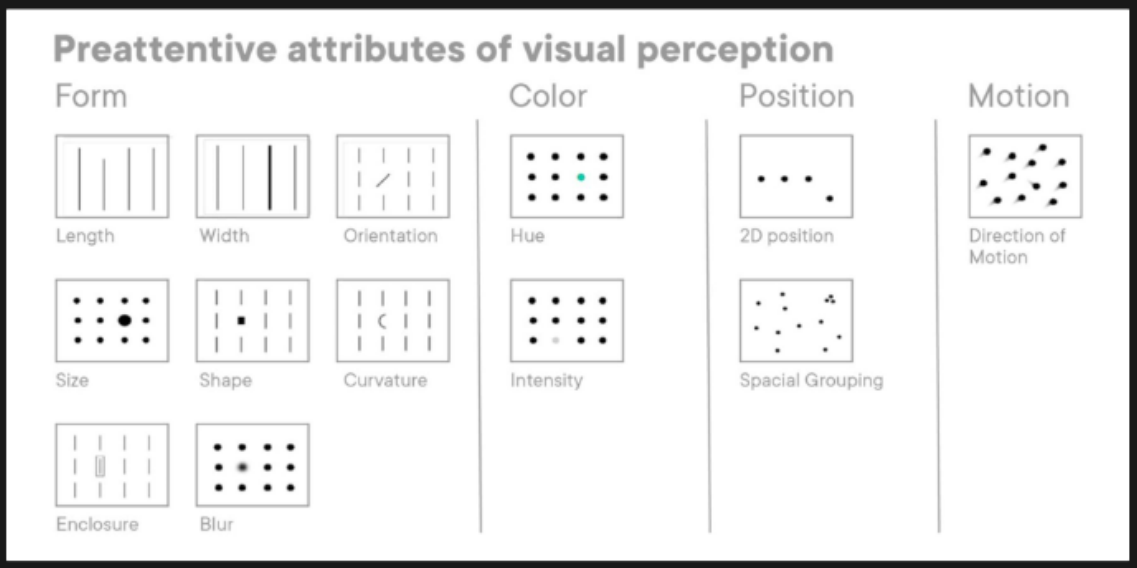
SOR stands for Stimulus Organism Response. A *Stimulus* is anything that can trigger a physical or behavioral change, the *Organism* refers to the individual, and the *Response* to the reaction.

Depending on the data you want to display, you have to use specific charts, for instance, for displaying information related to ranking, it is better to use pointer based charts or bar based ranking charts as they tend to describe data more accurately.

▼ Explain the concept of pre-attentive attributes. Discuss the importance in relation to dashboard design.

Pre-attentive attributes are visual properties that we notice without using conscious effort to do so. In is an important element in visualization, as it enables to direct the viewer's attention towards the most important information in our visuals.

Designing with pre-attentive processing in mind

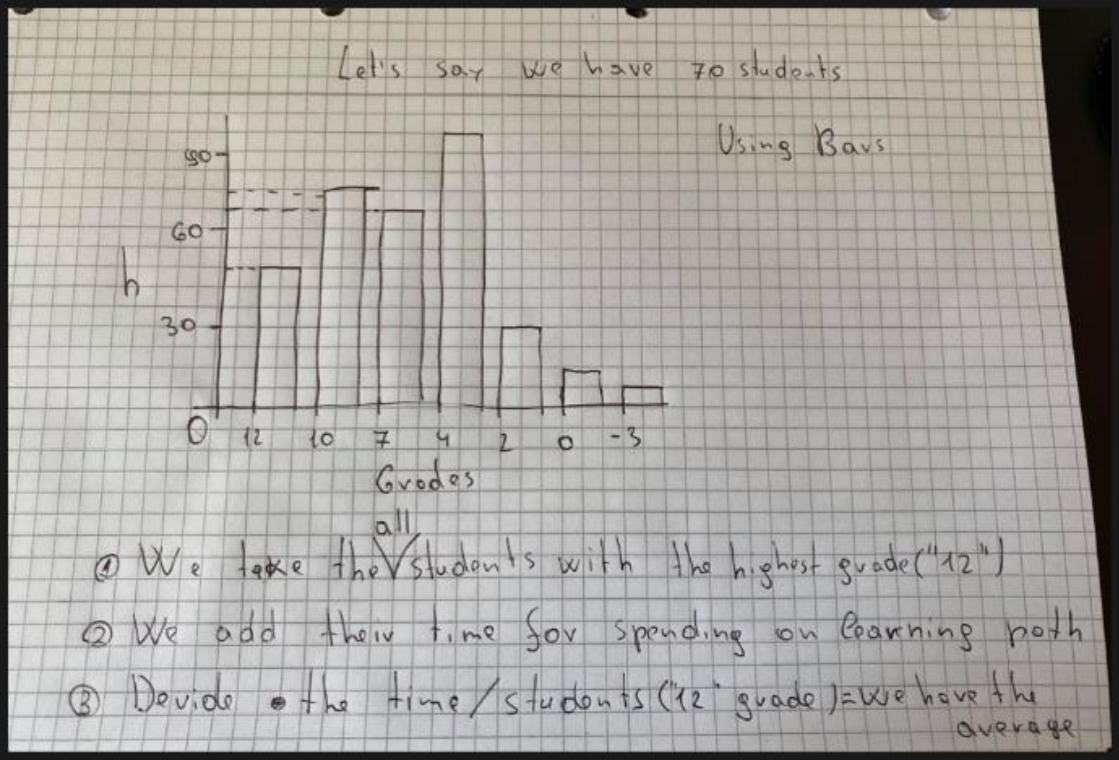


Visual things our eyes are drawn to:

- Form (length, width, size and shape);
- Color (hue, intensity);
- Position (2D, grouping);
- Motion.

C)

▼ You are tasked with creating a visualization to show the relationship between time spending on learning paths and grades. What kind of visualization are you going to show and why?



9 Data Analytics and Visualization on Dimensional Models II

- ▶ Explain the Dimensional Model
- ▶ Explain the concept of cognitive bias. Enumerate several type and provide examples
- ▶ You are tasked with creating a visualization to show daily data about temperature and precipitations.

A)

▼ Explain the Dimensional Model

The accuracy in creating your Dimensional modeling determines the success of your data warehouse implementation. Here are the steps to create Dimension Model.

1. Identify Business Process.
2. Identify Grain (level of detail).
3. Identify Dimensions.
4. Identify Facts.
5. Build Star.

B)

▼ Explain the concept of cognitive bias. Enumerate several type and provide examples

Cognitive bias is a systematic error in thinking that affects the decisions and judgment of people. Cognitive biases can be caused by a number of different things, such as heuristics (mental shortcuts), social pressures, and emotions.

Types of cognitive bias:

- overconfidence - results from someone's false sense of their skill, talent, or self-belief. The most common manifestations of overconfidence include the illusion of control, timing optimism, and the desirability effect; (The desirability effect is the belief that something will happen because you want it to.)
- herd mentality - when investors blindly copy and follow what other famous investors are doing;
- framing - when someone makes a decision because of the way information is presented to them, rather than based just on the facts. In other words, if someone sees the same facts presented in a different way, they are likely to come to a different conclusion about the information.

Cognitive Bias (corporatefinanceinstitute.com)

C)

▼ You are tasked with creating a visualization to show daily data about temperature and precipitations.