

Designing and Implementing a Data Warehouse

Dimensional Data Modeling



March 28, 2023

Vincent Nguyen

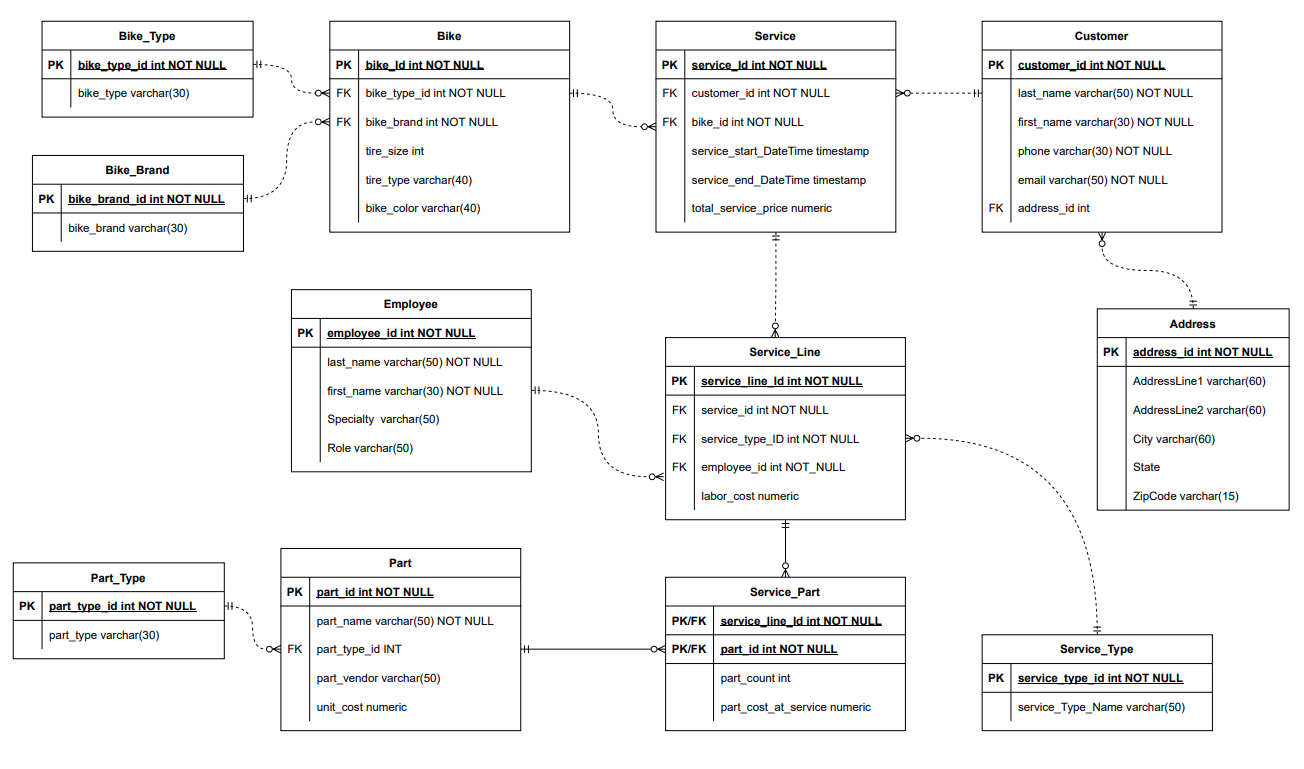
NVUI@BU.EDU

# Overview of the Assignment:

This assignment will go through steps to develop a data warehouse design.

# Part 1 – Review the business requirements

*BikeShop* is a national chain of bike repair shops which maintains a relational database to keep track of bike repairs. (example <https://www.landrys.com/> however consider them being more of a national business with hundreds of US locations)

*BikeShop* would like to introduce data warehousing and analytics to build their business. You have been hired as a data architect to create an initial Constellation data warehouse design. Below is *BikeShop’s* relational database. In addition, they would like to correlate the repair data with weather data to see if there are trends from weather which might impact repairs.

Here is a brief description of each of the tables in the OLTP system:

Customer and Address tables contain information about the customer. The address table is not normalized and may contain potential anomalies, especially if this data will need to be correlated with other external data.

For each bike, the type (i.e. mountain) and brand are tracked, along with tire size, type and bike color. Note that customers sometimes change the type of tire, and this data gets lost if a bike is brought back for service having a different tire type.

Service includes when the service started and ended, including the overall price. For each time the bike is serviced, there are multiple service lines (for example bike is brought in for tire and chain replacement which would be two service lines. An employee is tracked who performs the service. Some of the data around Employee is not normalized and contains anomalies. Each service line could have a part associated which includes prices for part as well as it’s cost at the time of service.

**External Weather data** is to come via an API such as <https://openweathermap.org> Here is a sample JSON API response to give you a sense of the data returned: <https://openweathermap.org/current#current_JSON>

{ "coord": { "lon": 10.99, "lat": 44.34 }, "weather": [ { "id": 501, "main": "Rain", "description": "moderate rain", "icon": "10d" } ], "base": "stations", "main": { "temp": 298.48, "feels\_like": 298.74, "temp\_min": 297.56, "temp\_max": 300.05, "pressure": 1015, "humidity": 64, "sea\_level": 1015, "grnd\_level": 933 }, "visibility": 10000, "wind": { "speed": 0.62, "deg": 349, "gust": 1.18 }, "rain": { "1h": 3.16 }, "clouds": { "all": 100 }, "dt": 1661870592, "sys": { "type": 2, "id": 2075663, "country": "IT", "sunrise": 1661834187, "sunset": 1661882248 }, "timezone": 7200, "id": 3163858, "name": "Zocca", "cod": 200 }

## Part 1 – Business Rules

1. ***Determine four business questions your data warehouse design will answer. Keep these questions in mind as you move on to the rest of the assignment. One of the questions needs to consider some sort of correlation with external weather data.***
   1. Which service type generates the most revenue over time?
   2. Which part are most sold during certain weather condition?
   3. Which employees generate the highest revenue for the company?
   4. What is total revenue of each product type of each month?

## Part 2 – Design a constellation schema warehouse.

1. ***Create and insert an ERD showing the constellation ERD schema below. Please see requirements in the next two sections.***

Diagram

Description automatically generated

1. ***Determine four to five (non-date/time) dimension tables.***
   1. ***SCDs need to be considered- make sure to include at least one type 2 and one type 3***

There are 7 (non-date/time) dimension tables in this constellation schema. It is described in the question C below.

* 1. ***Consider the date dimensions of different grains, consider a role-playing or bitemporal dimension, if so, explain why this is the case.***

In this context, there is one time dimension table which consider a role-playing with fact\_service table because a time dimension in this context use to represent both the start and end times of a service. There are 2 foreign keys to the dimension in the fact\_service table such as start\_DateTime\_id and end\_DateTime\_id

* 1. ***For EACH table answer the following question: What kind of SCD is the dimension table and why did you choose this type of SCD?***

|  |
| --- |
| Diagram  Description automatically generatedI choose SCD 1 – Overwrite for Weather, Bike Customer dimension table because even though the data in those table may change but the historical data does not need. For example:  Weather data is not expected to change. Bike data gets lost if a bike is brought back for service having a different tire type. And Customer data does not affect to the analytic data much. |
| A picture containing text  Description automatically generatedI choose SCD 2 – adding new record for the Employee dimension table because the specialty and role of the employee might change time to time, so it is important to track the historical changes. It would help to look back the employee career path. |
| I choose SCD 2 – adding new record for the Part dimension table because the Text  Description automatically generated with medium confidenceunit cost might change over time depend on the market. So, it is important to track the historical changes, it will help the business owner to analysis the price change curve and take an action on it to minimize their cost. |
| I choose SCD 2 – adding new record for the Service\_type dimension table Text  Description automatically generatedbecause the labor cost might change over time. Especially in the peak season or the holiday season, the labor cost would increase, so it is important to track historical changes. It will help the business owner to take an action to minimize their cost. |
| A picture containing table  Description automatically generatedI choose SCD 3- Current & Previous for Location dimension because location data is very infrequent changes, and I would like to keep it simple by tracking the current and the previous changes only. |

1. ***Determine two to three fact tables.***
   1. ***How are the measures tied to your questions?***
      1. ***Which service type generates the most revenue over time?***

To answer this business question, The measures tied to the question are total\_service\_price in the Fact\_Service table and service\_type in the Service\_Type dimension table. It is illustrated below:

Diagram

Description automatically generated

* + 1. ***Which part are most sold during certain weather condition?***

To answer this business question, The measures tied to the question are part\_count and part\_cost\_at\_service in the Fact\_Service table, these measures used to determine the parts that are most sold during weather condition when we joint Fact\_service table to other tables such as location, Fact\_weather, weather, Time, and Part tables. It is illustrated below:

Diagram

Description automatically generated

* + 1. ***Which employees generate the highest revenue for the company?***

To answer this business question, The measures tied to the question is total\_service\_price in the Fact\_Service table and join with Employee dimension table. It is illustrated below:

Diagram

Description automatically generated

* + 1. ***What is total revenue of each product type of each month?***

To answer this business question, The measures tied to the question is total\_service\_price in the Fact\_Revenue table and join with Service\_Type and Month dimension tables. It is illustrated below:

Diagram

Description automatically generated

* 1. ***Explain if your fact tables cumulative or snapshot.***

|  |  |
| --- | --- |
| A picture containing table  Description automatically generated | Fact\_Weather table and Fact\_Service table are snapshot because the measures are captured in the particular instance of the time, particular of the customer, grain is the lowest time. For example: Fact\_Service capture snapshot everytime the customer visit to get the service. |
| Graphical user interface, application, table, Excel  Description automatically generated | Fact\_Revenue table is cumulative because the data are at the higher grain (at the month). The measures are aggregate up to the month. (Happen in the period of time) |

* 1. ***Which attributes might be degenerative in your fact tables?***

There are no attributes which might be degenerative in the fact tables in the BikeShop’s constellation schema.

* 1. ***Which attributes in the OLTP schema will transform to measures?***

|  |  |
| --- | --- |
| Attribute from OLPT | Fact and Measure |
| Service\_Line.labor\_cost | Fact\_Service.labor\_cost |
| Service\_Part.part\_count | Fact\_Service.part\_count |

* 1. ***What measures can be derived/calculated that should be included?***

|  |  |  |
| --- | --- | --- |
| Attribute from OLPT | Fact and Measure | How is it transformed |
| Service\_Line.labor\_cost | Fact\_Service.labor\_cost averaging | Averaging the cost. |
| Service\_Part.part\_count | Fact\_Service.part\_count sum | Sum the part count. |

Use the **Ask the Teaching Team Discussion Forum** if you have any questions regarding the how to approach this assignment.

Save your assignment as ***lastnameFirstname\_assign2\_0.docx*** and submit it in the *Assignments* section of the course.

For help uploading files please refer to the *Technical Support* page in the syllabus.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Criterion | A | B | C | D | F | Letter Grade |
| Technical mastery (50%) | Evidence of excellent mastery throughout | Evidence of good mastery throughout | Evidence of basic mastery throughout or good mastery intermittently | Minimal mastery evidenced | Virtually no mastery evidenced |  |
| Depth and thoroughness of coverage (25%) | Excellent depth and coverage of significant topics and issues | Good depth and coverage of significant topics and issues | Basic depth and coverage of significant topics and issues | Minimal depth and coverage of significant topics and issues | Virtually no depth and coverage of significant topics and issues |  |
| Clarity in presentation (25%) | Ideas and designs are exceptionally clear and organized throughout | Ideas and designs are clear and organized throughout | Ideas and designs are somewhat clear and organized throughout | Ideas and designs are mostly obscure and disorganized | Ideas and designs are entirely obscure and disorganized |  |
|  |  |  |  |  | Assignment Grade: |  |