# Overview of the Assignment:

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

# Part 1 – Review the business requirements

*CityTour* is a national chain of metropolitan tours (yes they have double decker tour busses!). City Tours specializes in themes, for example haunted tours, tours in the river and the harbor (you may want to review Boston Duck boat tours, and Boston trolly tours for ideas)

*CityTour* 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. On the next page is *CityTour’s* relational OLTP data model. In addition, they would like to correlate the tour data with weather data to see if there are trends from weather which might impact tours.

***CityTour* OLTP Data Model on next page**

**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](https://openweathermap.org/current" \l "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 }

*CityTour* **OLTP Data Model**

**A diagram of a computer

Description automatically generated with medium confidence**

## 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.

**Tour Performance: Which tours have the highest and lowest bookings? This can help identify popular tours and those that may need improvement.**

**Customer Analysis: What are the demographics of customers who book tours? Understanding the customer base can help in tailoring marketing strategies.**

**Revenue Analysis: What is the revenue generated from each tour? This can provide insights into the most profitable tours.**

**Weather Impact: How does weather impact tour bookings? By correlating tour data with weather data, we can understand if certain weather conditions affect tour bookings.**

## Part 2 – Design a constellation schema warehouse

1. Create and insert an ERD showing the **constellation** ERD schema below. Requirements are as follows:
2. Determine four to five (non-date/time) SCDs - make sure to include at least one type 2 and one type 3
3. Determine two to three date dimensions of different grains, consider a role-playing or bitemporal date dimensions. You will use table in question 2b to outline the facts and explain your design choice for the temporal dimensions
4. Determine two to three fact tables, your design should include at least one Snapshot (transactional) fact, and at least one cumulative fact.

**Slowly Changing Dimensions (SCDs):**

**Customer (Type 2**): Attributes can include CustomerID, Name, Address, Phone, Email, etc. Changes in customer information like address or phone number can be tracked using Type 2 SCD.

**Product (Type 3)**: Attributes can include ProductID, Name, Category, Price, Old\_Price, etc. Changes in product price can be tracked using Type 3 SCD.

**Role-Playing Tables:**

**Date dimension** can play multiple roles in our fact tables. For example, in the Sales fact table, we can have OrderDateKey, DeliveryDateKey etc., each representing a different role of the Date dimension.

**Bitemporal Tables:**

**SalesHistory:** This can be a bitemporal table that keeps track of changes in the Sales fact table over time. It can have similar structure as the Sales table but with additional columns to track the validity period (from-to dates) for each fact record and the system time when the record was inserted or updated in the table.

**Fact Tables:**

**Sales** (Snapshot Fact): This table can store transactional data like OrderID, CustomerID, ProductID, OrderDateKey, Quantity, TotalAmount etc.

**Inventory** (Cumulative Fact): This table can store cumulative data like ProductID, DateKey (for the period), OpeningStock, StockIn, StockOut and ClosingStock.

1. Questions on Dimensions
   1. For EACH non-date/time SCD use the following table explain your SCD design

|  |  |  |
| --- | --- | --- |
| **Table Name and Screenshot** | **SCD Type** | **Design explanation** |
| Customer | Type 2 | This table stores information about customers. Each row represents a customer with details like name, address, phone number, and email. The EffectiveDate and EndDate columns are used to track the period during which a particular version of the customer record is valid. This allows us to keep a history of changes for each customer (for example, if a customer’s address changes). The CurrentFlag column is used to quickly identify the current version of each customer record. |
| Product | Type 3 | This table stores information about the products (in this case, tours). Each row represents a product with details like name, category, and price. The Old\_Price column is used to keep track of the previous price whenever the current Price changes. This allows us to see how the price of a product has changed over time. |
|  |  |  |

* 1. Outline the fact tables which contain role-playing or bitemporal design

|  |  |  |
| --- | --- | --- |
| **Table Name and Screenshot** | **Temporal Type** | **Design explanation** |
| DateDimension | Role-Playing | This table stores information about dates. Each row represents a unique date with details like day, month, quarter, and year. This allows us to analyze data at different time grains. The DateKey column is a unique identifier (primary key) for each date. This table can play multiple roles in our fact tables. For example, in the Sales fact table, we can have OrderDateKey, DeliveryDateKey etc., each representing a different role of the Date dimension. |
| SalesHistory | Bitemporal | This table stores information about dates. Each row represents a unique date with details like day, month, quarter, and year. This allows us to analyze data at different time grains. The DateKey column is a unique identifier (primary key) for each date. This table can play multiple roles in our fact tables. For example, in the Sales fact table, we can have OrderDateKey, DeliveryDateKey etc., each representing a different role of the Date dimension. |
|  |  |  |

1. Questions on Fact tables
   1. **Question** For each of your business questions in part 1, how are the measures tied to your questions? Highlight the dimensions, facts and measures involved.

|  |  |
| --- | --- |
| **Question 1:** | |
| **Screenshot of tables answering question** | **How measures are tied to your question** |
|  | **Measure: Number of bookings. This is calculated as the count of OrderID in the Sales table. It gives us the total number of bookings for each tour.** |
| **Question 2:** | |
| **Screenshot of tables answering question** | **How measures are tied to your question** |
|  | Measure: Number of bookings. This is calculated as the count of OrderID in the Sales table. It gives us the total number of bookings made by each customer. |
| **Question 3:** | |
| **Screenshot of tables answering question** | **How measures are tied to your question** |
|  | Measure: Revenue. This is calculated as the sum of TotalAmount in the Sales table. It gives us the total revenue generated from each tour. |
| **Question 4:** | |
| **Screenshot of tables answering question** | **How measures are tied to your question** |
|  | Measure: Number of bookings. This is calculated as the count of OrderID in the Sales table. It gives us the total number of bookings for each date, which can then be correlated with weather data to analyze the impact of weather on tour bookings. |

* 1. **Question:** Outline the fact tables and explain why they are cumulative or snapshot

|  |  |  |
| --- | --- | --- |
| **Table Name and Screenshot** | **Fact Type** | **Design explanation** |
| Sales | Snapshot Fact | This table stores transactional data about sales. Each row represents a unique sale with details like order ID, customer ID, product ID, order date key, quantity, and total amount. The OrderID is a unique identifier (primary key) for each sale. CustomerID, ProductID, and OrderDateKey are foreign keys referencing the Customer, Product, and DateDimension tables respectively. This table is a snapshot fact table, meaning it captures the state of facts at a particular point in time. |
| Inventory | Cumulative Fact | This table stores cumulative data about inventory. Each row represents the state of inventory for a product at the end of a period with details like product ID, date key, opening stock, stock in, stock out, and closing stock. The ProductID and DateKey together form a composite primary key for each record. OpeningStock, StockIn, and StockOut are used to track inventory movements, and ClosingStock is a computed column that gives the closing stock for each product at the end of each period. This table is a cumulative fact table, meaning it accumulates facts over time. |
|  |  |  |

* 1. **Question:** Which attributes in the OLTP schema will transform to measures and what measures can be derived/calculated that should be included?

|  |  |  |
| --- | --- | --- |
| **Table and Attribute name from OLTP** | **Fact and Measure name** | **How is it transformed** |
| Sales.TotalAmount | Sales.Revenue | The TotalAmount attribute from each sales record in the OLTP schema is summed up to calculate the Revenue measure in the data warehouse. |
| Sales.Quantity | Sales.QuantitySold | The Quantity attribute from each sales record in the OLTP schema is summed up to calculate the QuantitySold measure in the data warehouse. |
| Inventory.StockIn, Inventory.StockOut | Inventory.StockMovement | The StockIn and StockOut attributes from each inventory record in the OLTP schema are summed up to calculate the StockMovement measure in the data warehouse |

## Part 3 – Assumptions/Appendix (optional)

You may need to make some assumptions in order to work on this assignment. Clearly state any assumptions you needed to make here, along with your reasoning why the assumption was appropriate (one or two sentences should be sufficient in most cases). Do not remove any functionality, or trivialize, any of the assignment requirements. Feel free to augment.

I made the following assumptions:

Data Availability: I assumed that all the necessary data is available and accurately recorded in the OLTP system. This is important because the quality and completeness of the data in the OLTP system directly affect the accuracy of the data warehouse.

Business Processes: I assumed that the business processes of CityTour are stable and well-defined. This is because the design of the data warehouse is based on the business processes.

Data Relationships: I assumed that the relationships between different data entities (like Customers, Products, Sales, etc.) are correctly defined in the OLTP system. This is crucial for correctly transforming the OLTP schema into a data warehouse schema.

Weather Data: For correlating tour data with weather data, I assumed that CityTour has access to reliable weather data that can be integrated into the data warehouse.

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: |  |