

Assignment 1

by Kishen Sukumaran

Submission date: 25-Mar-2022 06:02PM (UTC+0000)

Submission ID: 174594444

File name: 35305_Kishen_Sukumaran_Assignment_1_398298_1210149945.docx (1.01M)

Word count: 1872

Character count: 10071



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

ACCA6026: Data Warehousing and Data Mining

Name: Kishen Sukumaran

Student ID: 2121314

Assignment 1

Contents

1.0 Introduction	3
2.0 Differences between OLTP and Data Warehouse	3
3.0 Architecture of a Data Warehouse	4
4.0 Data Marts and Functions	5
5.0 Normalised OLTP versus Dimensional Data Warehouse	6
6.0 Internal and External Source systems	8
Internal Source Systems	8
External source system	8
7.0 Star Schema	9
8.0 Data Mart Schemas	10
New Item Acquisition Data Mart	10
Item Location Data Mart	11
Movement Order Data Mart	11
Order Data Mart	12
Return Order Data Mart	13
9.0 ETL Diagram	14
References	15

1.0 Introduction

A data warehouse is a centralised repository of data that collects and aggregates data from multiple different sources. These source systems which the data warehouse collects data from are known as operational systems like the transaction processing system used in Rags'n'Riches Ltd. The data is then collated into relational tables using online analytical processing (OLAP) which allows users to query information and for reporting. The benefit of this system is its ability to analyse large datasets like transaction data from different source systems and aggregates said data to be used for analytical purposes like making business decisions and tracking data trends. On the other hand, an online transaction processing (OLTP) database is a system which tracks daily transaction data. This allows for real-time execution of large quantity of transactions typically used in systems with many users over the internet. OLTP databases can manage large quantity of simple processes for insertion, deletion and updating data.

2.0 Differences between OLTP and Data Warehouse

Online transaction processing is a technique designed for managing business transactions by storing data from transaction which are added on to on a daily basis whereas, a data warehouse collects data from different sources for example, the sales department and delivery department which may use different databases and creates a central repository of data which can be used for analysis. In essence, OLTP is used for tracking business transactions and a data warehouse using online analytical processing (OLAP) is used to aggregate data to make business decisions for example on a month with high number of deliveries to customers, managers can identify the month with higher number of deliveries, then decide to either increase delivery costs or increase number of delivery vehicles to match the demand. This can only be done on a data warehouse because this system relies on query processing which allows data to be sorted to increase business intelligence because an OLTP database can only handle transaction processing. A data warehouse is optimised to collect and analyse large datasets and aggregating data from said datasets which can be done on a separate system to the OLTP database thus reducing the load on the OLTP database. In the current OLTP system when large analytical queries are being performed like creating performance reviews and generating sales reports it will impact the performance of the database creating a bottleneck when accessing the database, meaning some users may not be able to access or input data into the database while the system is performing queries.

3.0 Architecture of a Data Warehouse

A data warehouse is a system which consists of source systems, (internal databases external databases, transaction processing systems, enterprise resource planning (ERP) systems and flat or text files) transport agents which collect and move data from one source system to another, a centralised repository where metadata of the system is stored, data marts where specific data is extracted to suit the needs of different departments of the business and users. The figure below shows the basic architecture of a data warehouse (Figure 1).

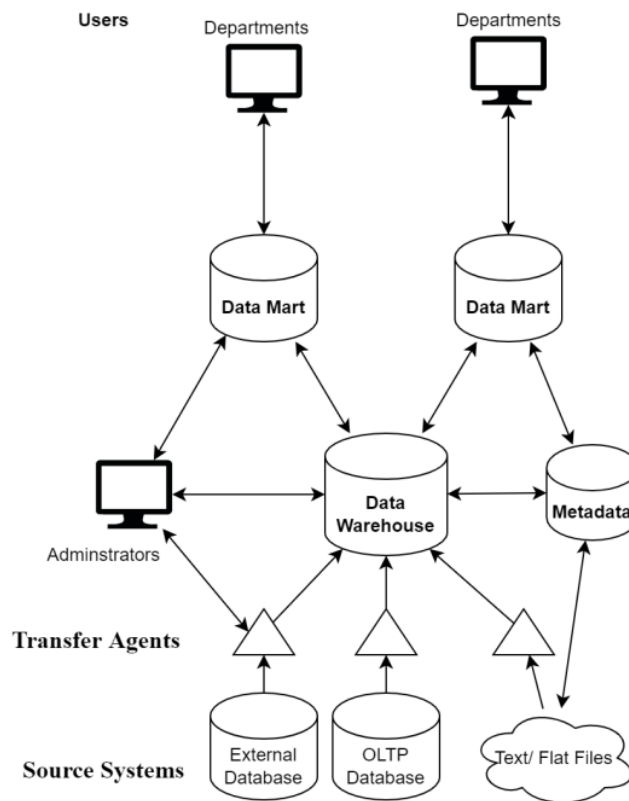


Figure 1: Data Warehouse Architecture

Internal data like the OLTP database can be integrated into the data warehouse relatively easily because the data is structured. External data from sources outside of the business can also be integrated into the data warehouse through the use of structuring functions which can give unstructured external data structure and meaning. Furthermore, metadata which is the details given to unstructured data like external data can help provide information about the external data. As such external data can act as an independent data mart in which the data

is not tied to existing data from the data warehouse. In contrast, a dependant data mart commonly accessed by business users consists of data drawn from the centralised repository of the data warehouse which is focused on a single subject like sales for a sales department.

The difference between these data marts is how they are populated, the process in which data mart is populated is known as the Extraction Transformation and Loading process (ETL). This process includes moving data from source systems, filtering data into relevant categories and then loading the data into the data mart. When data is from an internal source system is loaded into a data warehouse the ETL process can be completed easily compared to an external source system because the data from the internal database is formatted and summarised in a format which the data warehouse can aggregate. On the other hand, data from external source systems is considered unstructured data and must be filtered, formatted, and summarised because data from external source systems may not have the same attributes as the data in the data warehouse. As such, integrating the existing databases into a data warehouse should not affect the operations of the business since all databases are internal and share attributes among different departments.

4.0 Data Marts and Functions

A data mart is a form of a data warehouse that is hold data specific to a functional area which can be controlled by a department of the business. This means that the data is sent to the data mart is specific to the department. Moreover, data marts can collect data from operational sources, the central data warehouse repository, and external systems. Since data marts oriented to a single subject like orders or sales, they can be used by different departments of the business to access data relating to their department rather than from the centralised data warehouse. As data marts are subject oriented it means that the system will provide concise information on a specific subject for analysis and modelling. This is done to allow users to make decisions by only providing relevant data and excluding data that does not support making business decisions for example, if the sales manager wants to view a sales report, they can use the data mart to view the aggregated data regarding sales while excluding data relating to the invoice IDs of sales. This can be used by all departments to allow users in this case the managers of each department to view targeted data relating to their departments.

5.0 Normalised OLTP versus Dimensional Data Warehouse

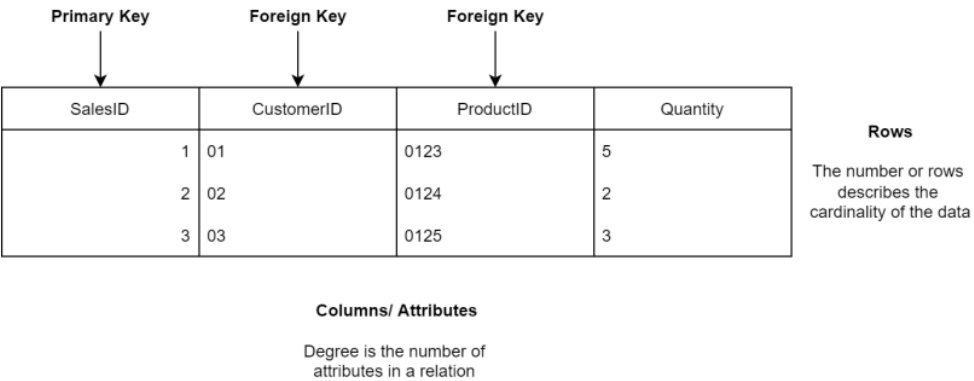


Figure 2: Relational Model

A normalised structure in a database like an OLTP system refers to the process of breaking down complex relational data structures into simpler ones based on normal forms. Normalising data can be done up to three levels, the first normal form (1NF) where each attribute has a single value. The second normal form (2NF) where each attribute that is not a primary key (a key that other attributes are dependant on) is functionally dependant on a primary key. The last is the third normal form (3NF) where attributes are already normalised to 2NF and has a primary key attribute is then also functionally dependant on another non-primary key. This process reduces large datasets into their smallest forms while maintaining data cardinality (Figure 2). As such, relational models like OLTP databases are optimised to perform addition, deletion and updating data in real time. This means that to add new orders or to update them

using the OLTP database is sufficient. However, to perform analysis and reporting a dimensional model like the technique used in a data warehouse is more suitable (Figure 3).

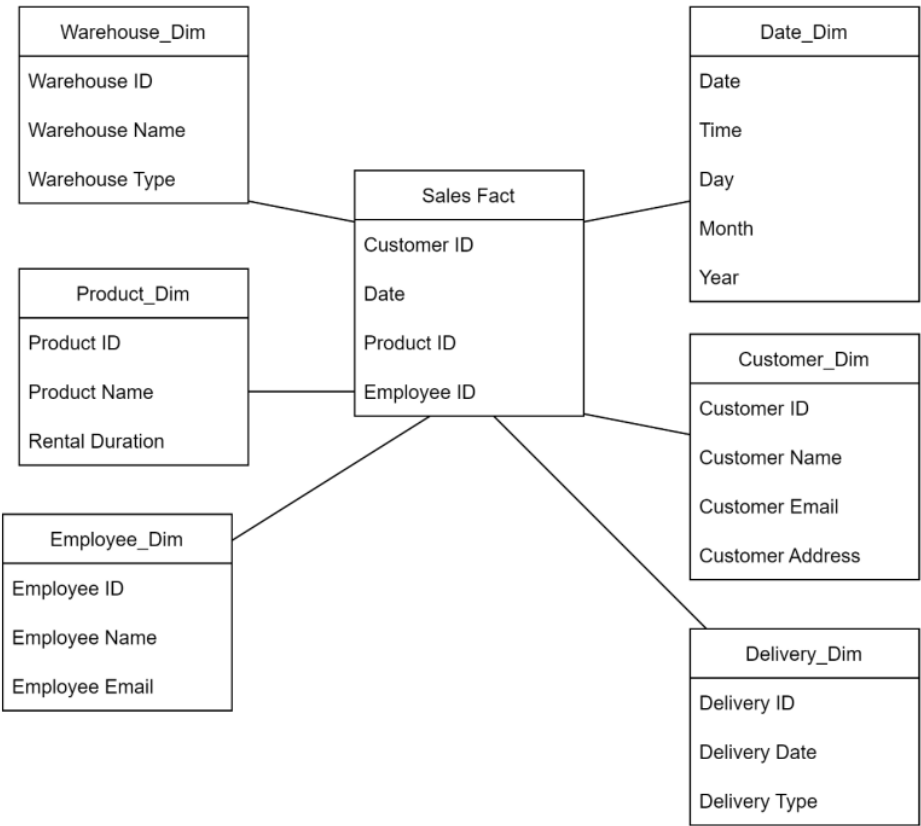


Figure 3:Dimensional Model

This is because using the dimensional model it allows users to analyse and summarise the data values. Where relational models use normalisation to reduce data redundancy, the dimensional model is aggregates to allow data to be retrieved easily and to generate reports with only relevant information as shown in the figure above. The dimensional model used in figure 3, shows a fact table which is the primary table used in this dimensional model. The Sales Fact consists of attributes from dimensions which are tables with context data relating to the fact table. This allows users to view details of a sales fact to identify the relevant data to the sale as well as information about other aspects of the sale like customer location, warehouse and products used in the sale. The benefit of system is that is allows for much higher query performance and the ability for users to generate reports and perform analysis on the specific fact.

6.0 Internal and External Source systems

Internal Source Systems

- Warehouse database
- Prop database
- Order database
- Rental item database
- Customer database
- Discount database
- Transport System database
- Movement order database
- Return order database
- Sales database
- Prop Damage

External source system

- Acquisition database

7.0 Star Schema

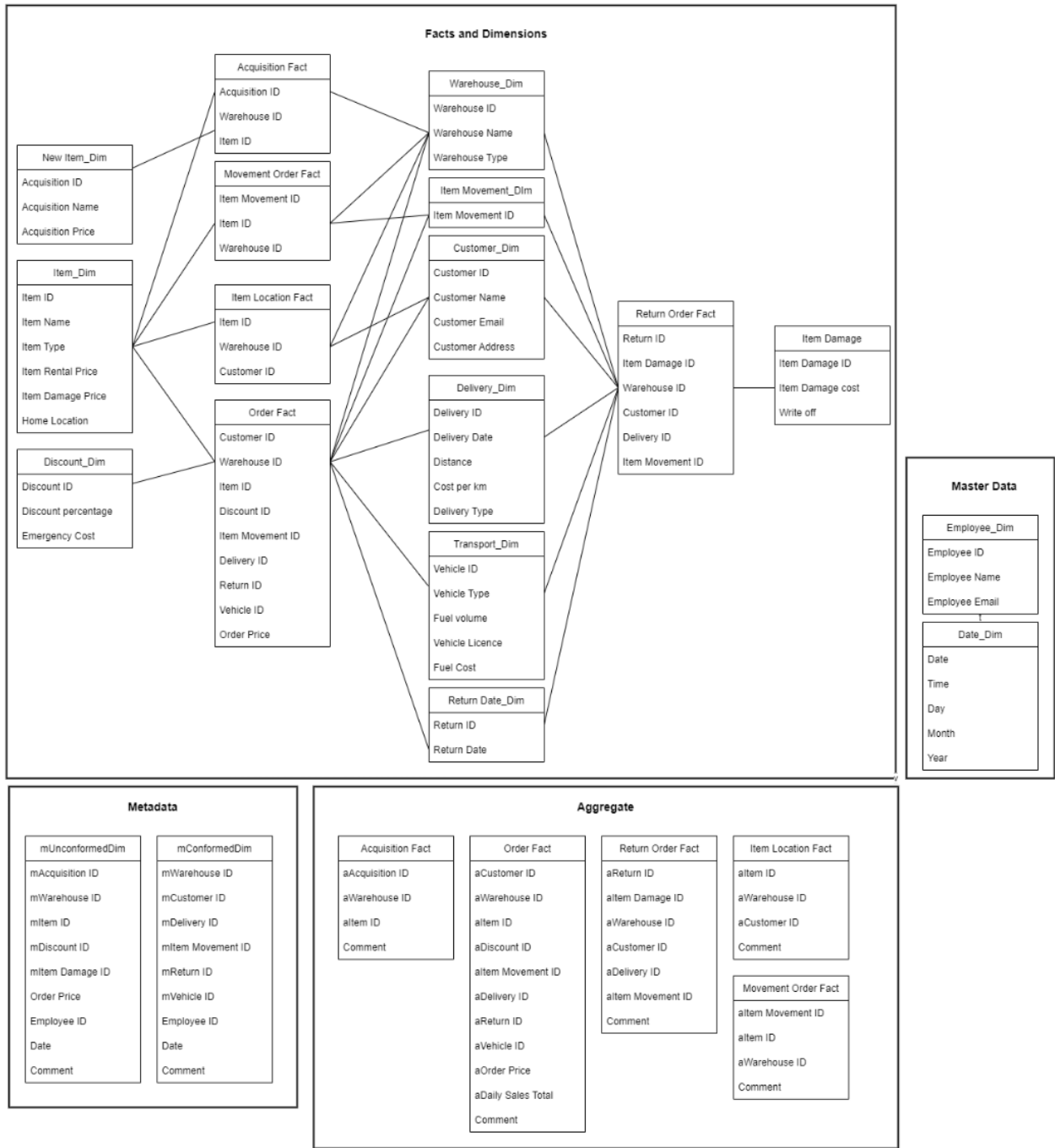


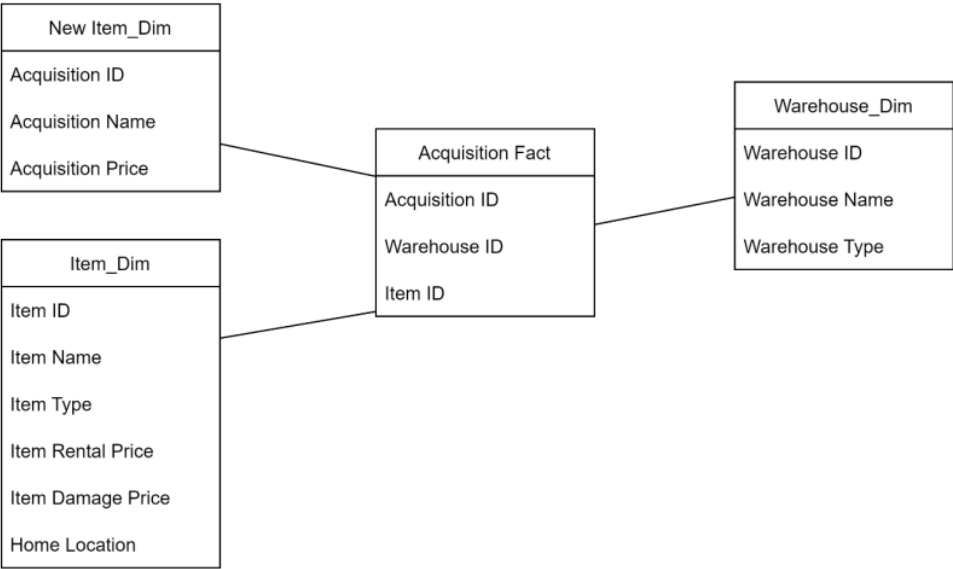
Figure 4: Star Schema

In figure 4 above, the facts used were defined through analysis of the business to identify the different departments of the business and their functions. Then the date and employee dimensions were used as the master data as they are used in every fact. Finally, the metadata for the system was selected by extracting the facts and organising them into unconfirmed and conformed attributes. The aggregates were taken from the facts and a daily sales aggregate was included as part of the orders fact to collect daily sales data.

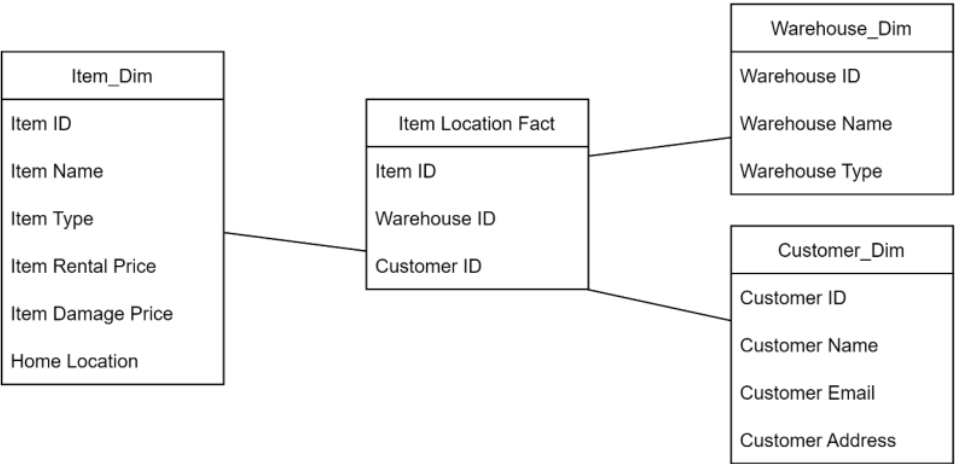
8.0 Data Mart Schemas

Data marts for this schema should be the facts and their corresponding dimensions.

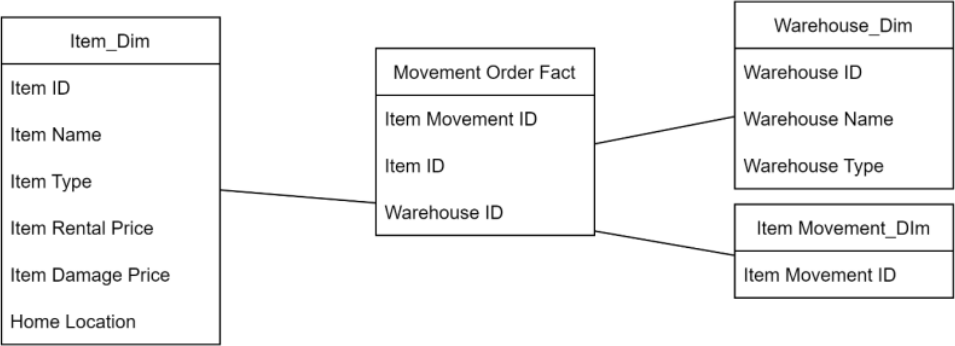
New Item Acquisition Data Mart



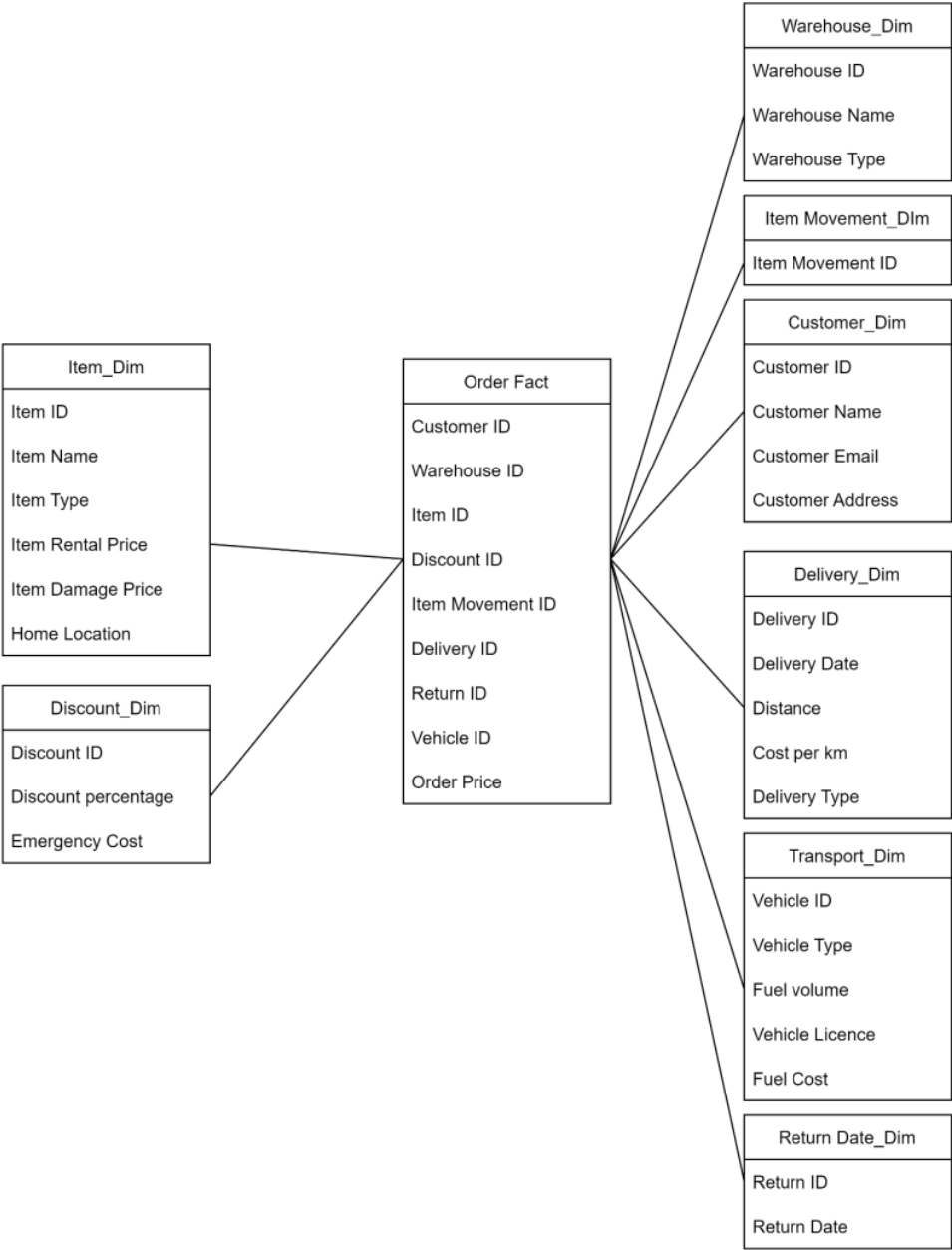
Item Location Data Mart



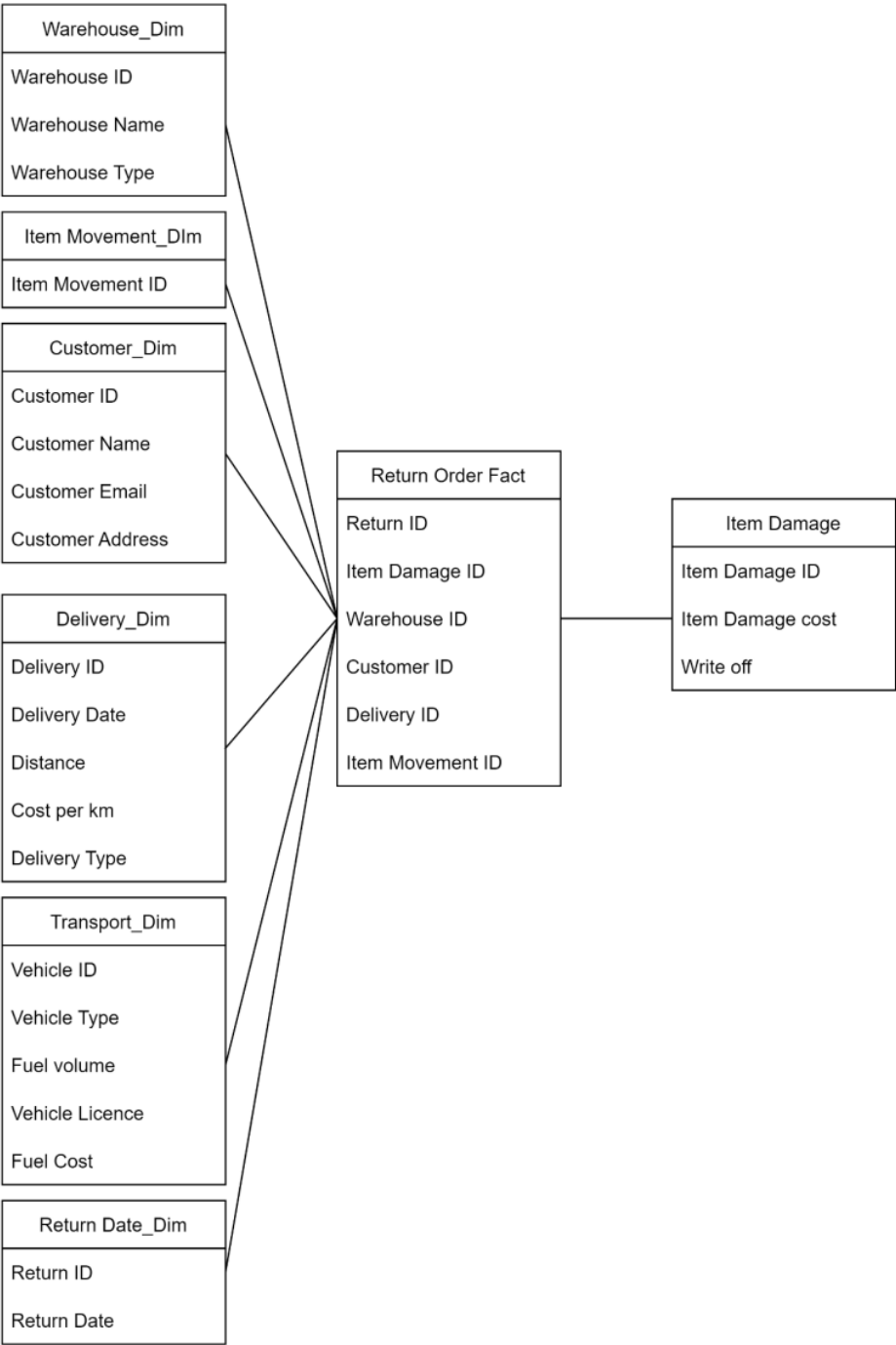
Movement Order Data Mart



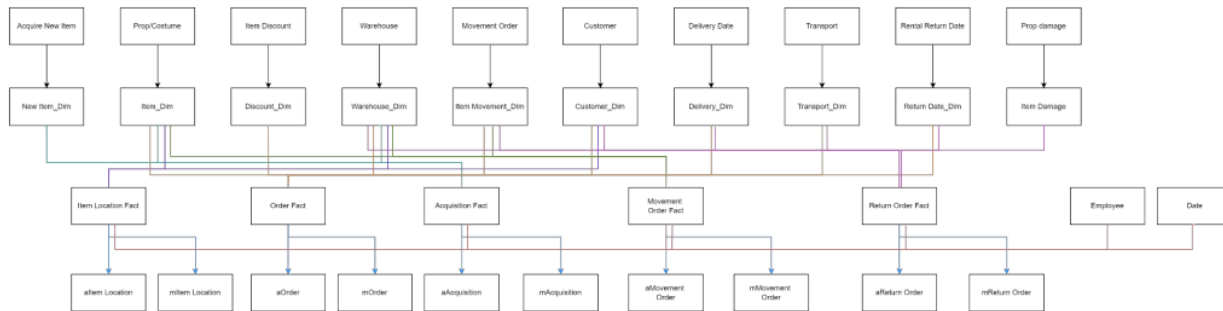
Order Data Mart



Return Order Data Mart



9.0 ETL Diagram



The data that requires transformation before it can be loaded into the data warehouse is the external database data from acquisition database. The data from this database is inputted by employees as suggestions to the managers and as such the data may be unstructured. As such the data will need to be extracted filtered and loaded by the data warehouse before it can be used to analyse the trends in the database.

References

Boussaid, O., Tanasescu, A., Bentayeb, F., & Darmont, J. (2006). Integration and dimensional modeling approaches for complex data warehousing. *Journal of Global Optimization*, 37(4), 571-591. doi:10.1007/s10898-006-9064-6

Garmany, J., Walker, J., & Clark, T. (2005). *Logical database design principles*. Boca Raton (Florida): Auerbach Publications.

Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (1998). Architecture and quality in data warehouses. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, 93-113. doi:10.1007/bfb0054221

M Kirmani, M. (2017). Dimensional modeling using star schema for Data Warehouse Creation. *Oriental Journal of Computer Science and Technology*, 10(04), 745-754. doi:10.13005/ojcs/10.04.07

Parekh, A. (2017). Introduction on Data Warehouse with OLTP and OLAP. *International Journal of Engineering and Computer Science*, 2(08). Retrieved from <http://www.ijecs.in/index.php/ijecs/article/view/1819>

Sharma, Vaibhav. (2021). An Enlightening Assessment of Data Mart Exploration in Promptly Mounting Data Warehousing Consequence. 8. 264-268. 10.17148/IARJSET.2021.8544.

Sinha, T. (2021, March 16). OLAP vs. OLTP: What's the difference? Retrieved March 25, 2022, from <https://www.ibm.com/cloud/blog/olap-vs-oltp>

Assignment 1

GRADEMARK REPORT

FINAL GRADE

50/100

GENERAL COMMENTS

Instructor

Kishen Sukumaren

1 engagement with literature
comment available

only 7 references from a range of good
sources, more references are required at
level 6 little evidence of citations within
the body of text - use IEEE rather than
Harvard in computing!!! 20

2 Knowledge and Understanding
comment available

knowledge demonstrated and
appropriate images used to reinforce the
discussion - but no references used or
credit given to image sources.... At level 6
this is considered the bare minimum 20

3. Cognitive and Intellectual skills
comment available

all areas covered but some sections could
benefit from substantially more
explanation 10

4. Practical Application Skills
comment available

design diagrams are very good and show
adequate practical implementation of
knowledge gained (some are a little difficult
to read due to level of zoom and number
of overlapping relations) 40

5. Transferable Skills for life and Professional practice
comment available

nicely laid out report but definitely more
work required in expansion and
explanations of topics within the body of
the report 10

Total: 50

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15
