

Data Warehousing and Data Mining- 1809781

by Jessica Marfell

Submission date: 24-Mar-2022 05:34PM (UTC+0000)

Submission ID: 174440170

File name: 12281_Jessica_Marfell_Data_Warehousing_and_Data_Mining-_1809781_398298_488976604.docx (2.33M)

Word count: 3665

Character count: 20544

ACCA6026

Data Warehousing and Data Mining

**Author: Jessica Marfell
Student Number: 1809781
Date: 23/03/2022**

TABLE OF CONTENTS

1	INTRODUCTION.....	3
2	TASK 1.....	4
2.1	Part 1.....	4
2.2	Part 2.....	5
2.3	Part 3.....	10
2.4	Part 4.....	13
3	TASK 2.....	14
3.1	Introduction.....	14
3.2	Internal and external source systems	14
3.3	Rags'n'Riches Star Schema	15
		16
3.4	Data Marts.....	17
3.5	ETL Plan.....	21
3.6	Conclusion.....	27
4	REFERENCES.....	28
5	APPENDIX.....	29

1 Introduction

The following report answers task 1 and task 2 of the data warehousing and data mining assignment. The Rags'n'Riches case study has been reviewed and is referenced throughout. The detail of the case study is documented within the assignment specification [1] . Please refer to the appendix to view larger diagrams.

2 Task 1

2.1 Part 1

An Online Transaction Processing (OLTP) database is used prudently to manage transaction information. The benefits of using an OLTP databases is, it is can handle many transaction requests Concurrently which reduces the need for paper processes. An OLTP database is easily assessable as it allows the user to perform basic operations such as read, write, and delete. These types of databases are reliable as they can be backed up and have atomicity controls. These controls ensure a transaction is completed successfully at all stages. If one area fails, then the transaction will subsequently fail. Rags'n'Riches currently uses OLTP databases to manage their order transactions.

A data warehouse is a central repository that stores data from different operational systems. It is a data management system that enables complex queries to be performed which allows for data analysis and reporting to be completed. The use of a data warehouse improves the efficiency of accessing different data sets and makes it easier for decision makers as they have access to meaningful data.

A data warehouse is subject oriented which means, it can provide data on a particular area of interest and can analyse changes to data over a period. It is used as a foundation for business intelligence software systems as they can extract the information from the warehouse to display reports and tends to the end user.

As described on SAS Insights, data analysis can lead to cost reduction, faster and more improved decision making and new products and services being available. A data warehouse enables companies to have access to information that can identify new opportunities and lead to smarter business decisions [2]. Whereas an OLTP database can store transactional information and produce basic data analysis.

The difference between an OLTP database and a data warehouse is, an OLTP database holds data from one database. In the Rags'n'Riches model they hold a variety of OLTP databases. These databases cover applications such as order information, transport, and delivery. They are independent from one another which allows for real time transactional processing and is built for speed. However, a data warehouse accommodates the storage of all application databases. This optimises the decision-making opportunities within the company as complex queries can be carried out on the data warehouse as it stores a multitude of company information. The data warehouse links the information together and stores it in a central space. This is how more complex and meaningful data can be pulled as information will be linked together [3].

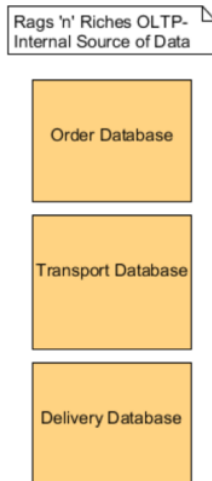
Quick and basic SQL queries can be undertaken on the companies OLTP databases, however as the data isn't normalised which means there is no data duplication these queries can become complex and tables may be required to join together to extract meaningful information as the information may need to be linked together. The result of this is, in depth analysis of the companies' data is unavailable by using OLTP databases. Joining of tables can also be more time consuming depending on the query being run.

A data warehouse holds information from separate source systems which allows analysis of data to be undertaken. It is designed to handle a large amount of data which can be systematically queried which is easier to query as there is fewer joins in the tables required. This is due to the central store of information.

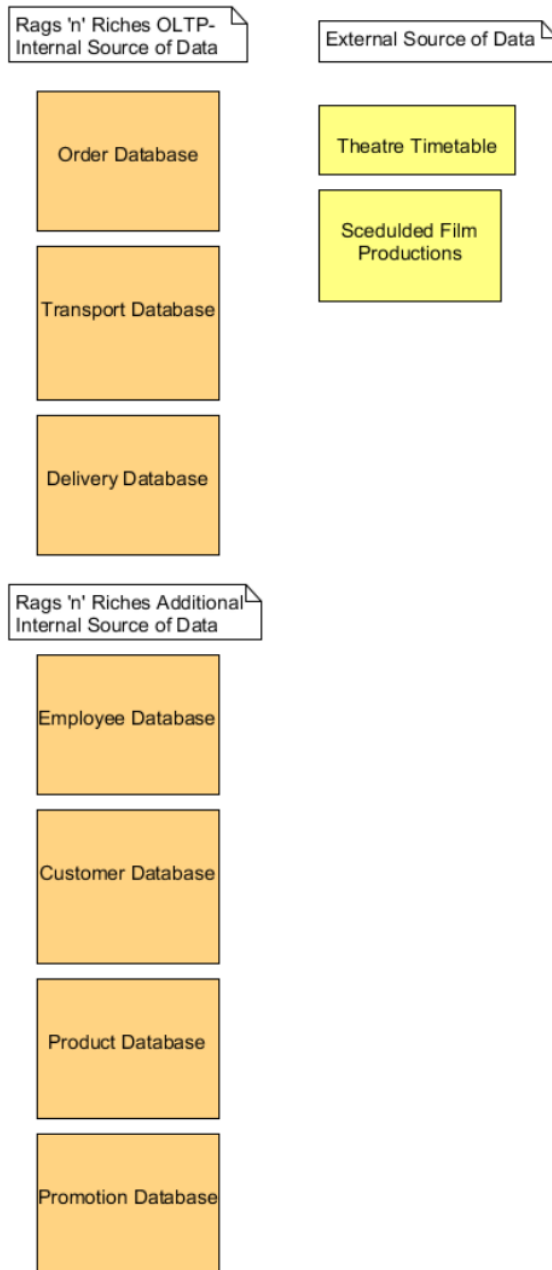
For Rags'n'Riches their OLTP databases are currently utilised to store and handle transactions. These types of databases are not suitable for in-depth data analysis. If they were to utilise a data warehouse the transaction data can be queried to perform a wide range of analysis which will ultimately improve company processes as it leads to informed business intelligence [3].

2.2 Part 2

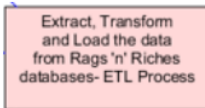
A data warehouse is a type of data management tool that stores information to perform advanced queries on the data. Its data is gathered from a range of sources which are transactional applications, in the Rags'n'Riches case study. The company has an order, transport, and delivery database within their architecture. These are also known as the companies' source systems as it stores the company information. These are separate transactional databases (Online Transaction Processing OLTP databases).



The OLTP databases displayed above are the companies existing internal source systems. This means it's the company data that is stored in the internal transactional databases. It is recommended that additional internal source systems are utilised. The company would benefit from databases that store Employee, Stock, Products and Promotional information as this would be extracted into the data warehouse. Rags'n'Riches could also utilise an external source of data that links in with the data warehouse. An example of an external source of data for the company may be a list of performances that are due to take place in certain theatres or film productions due to take place. An external timetable could provide the company with marketing scope and prepare for potential orders during certain times of the year. This information can be extracted from an external source and loaded into their data warehouse for analysis. The type of source systems may be databases, transaction processing systems or spreadsheets [4].



For Rags'n'Riches Ltd, a data warehouse will be beneficial as data can be extracted out of the company source systems for example their orders, transport and delivery transactional databases to transform the information into the same format to be loaded into the data warehouse. This process is known as ETL – Extract, Transform and Load.



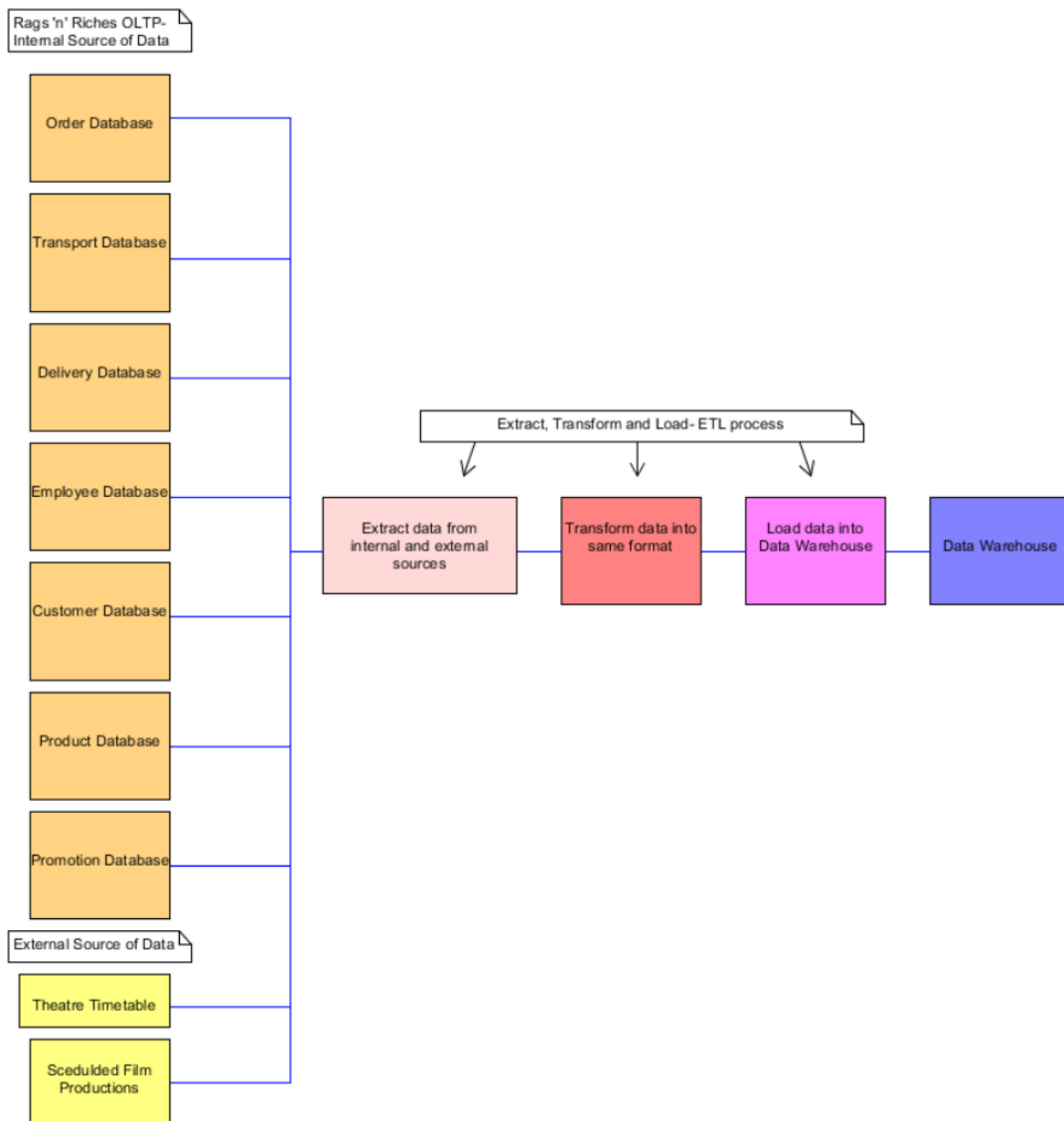
To allow for data analysis to be carried out on the data held in the databases, it needs to be extracted, transformed, and loaded (ETL) into a data warehouse. A copy of the data is extracted from the source systems through an automated procedure. An ETL process will extract the relevant data and convert it into an approved format to load into the company's data warehouse. There are some core requirements which need to be agreed with Rags'n'Riches before the data can be extracted.

- How will the business handle messy data and prioritise data to be cleaned?
- What are the legal reporting requirements and security constraints as customer information is captured?
- How often does the information need to be updated from its source for data latency and where is it coming from?
- What is the archiving process for extracted data?
- What data integration is needed for a full view of the company information?
- What available licences are there and what skill set for querying databases do Rags'n'Riches have?

Kimball further supports the ETL process and identifies 34 sub processes which can be split into four main area's [5]-

Data Extraction	Cleansing & Conforming Data	Delivering Data	Managing ETL
- Data is analysed to best describe its content through Data Profiling. - Identify the dimension values that need to be added to change the data capture. - Extract the data from the source system.	- Clean the data to ensure it provides an accurate picture of the company and create an error event schema to record every error thrown. - Audit the facts to record and problems created by the fact table. - Deduplication of records and conform the data from each source.	- Ensure the delivery of data is structured and loaded into the target dimensional models.	- Ensure the ETL system is reliable and available with the correct supporting tools. - Develop an ETL plan that maps out the data sources to target tables.

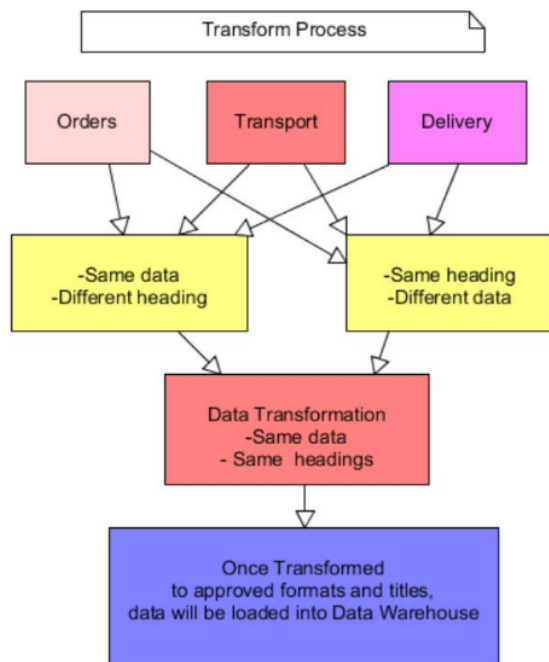
For the example above the data will be extracted from the internal and external source systems. This data will be transformed to ensure it can be loaded into the data warehouse in an approved format. This will allow the information to be accessible and queried within the company's data warehouse.



The extracting of data will not affect the day to day operations of the company as the data taken from the databases are a snapshot. The company has an option to stage the extraction of data to ensure the performance of their systems aren't impacted. As the company grows the scope of data will also increase. This will also improve quality control as it will be easier to validate the extracted information to make sure it is not corrupt before it goes into the data warehouse.

To ensure the data is transformed correctly and data validation operations can be set up. For example, if the extracted data displays the same data but with different titles it can be cleansed and streamlined to be displayed under one title. This will provide the company with meaningful data that covers all their source systems. Once the data has integrated in a suitable way for Rags'n'Riches, it can be loaded into the data warehouse [6].

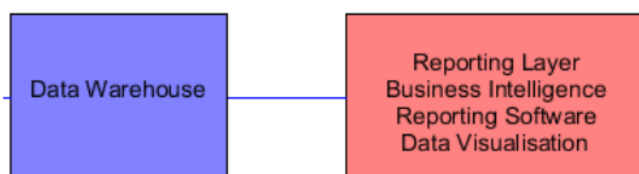
An example of data validation mapping to be considered for Rags'n'Riches-



When loading the information into the warehouse, a process can be set up to be completed at times when the company is quiet, for example, at night. A check can be done to ensure the data extracted and transformed meets the user requirements. To enhance this process further, Rags'n'Riches could check their business intelligence reports to ensure it is displayed correctly.

The ETL process documented above along with the supporting processes identified by Kimball and an introduction of an ETL plan, mitigates the risk of it affecting the day to day operations of the company. As the process is incremental and validated throughout the stages are controlled and can be rectified should errors occur.

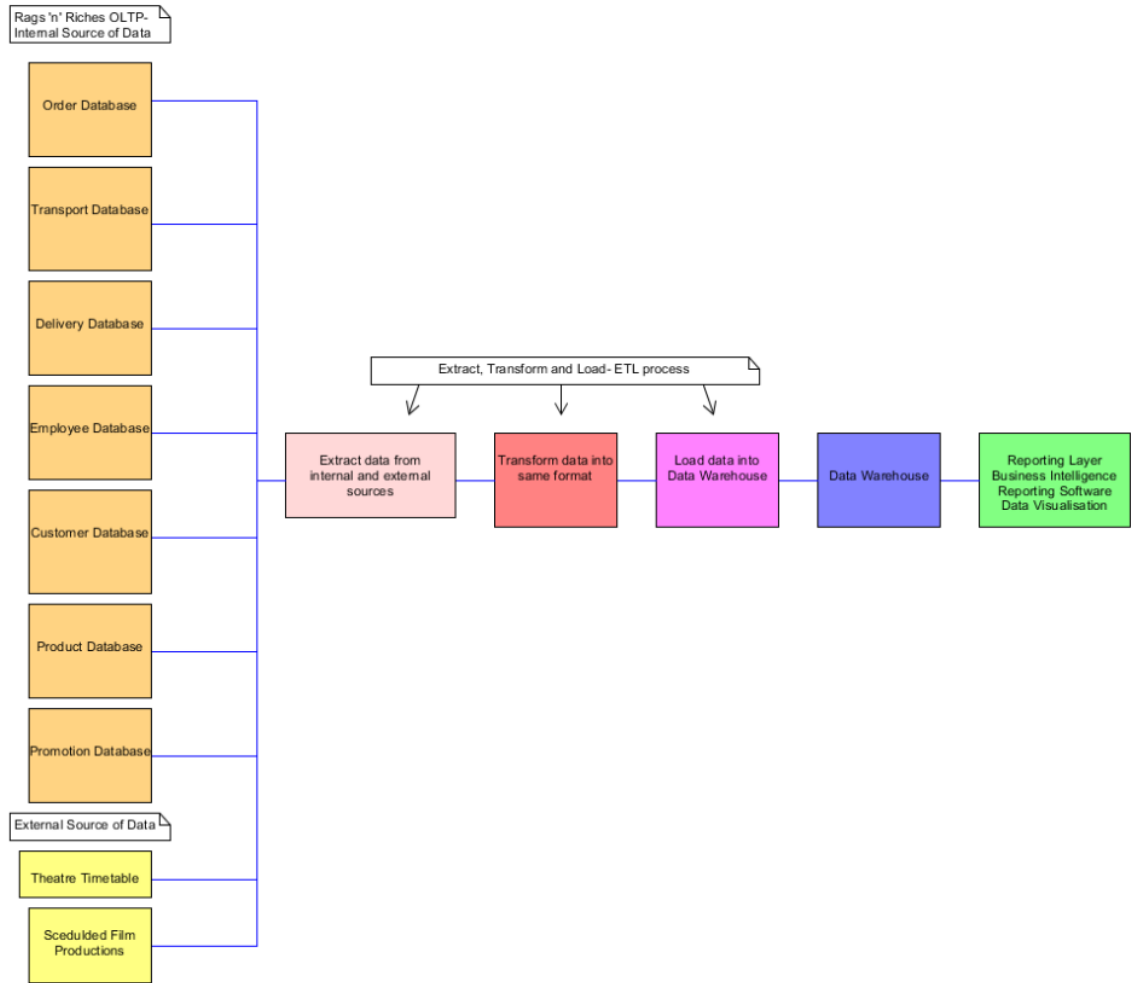
Once the data is loaded into the data warehouse the company can view this information through an operational business intelligence application such as Microsoft Power BI to present company information.



Business intelligence combines the business analytical information with data visualisation tools to display meaningful information to end users [7]. Business intelligence and data visualisation tools display the queried data to display in report and graph formats. This allows non-technical users to engage and read the information that the company has captured from their source systems. This leads to enhancements in the company processes and decision making. To ensure the data extracted is suitable a requirement gathering process will need to be undertaken with users who will be viewing the

data and making company decisions based on this information. For example, the Finance Director may require information and reports on the financial state of the business whereas the Operations Director will be more concerned with the transaction processes. To ensure all stakeholders are informed correctly, user requirements needs to be obtained.

A high-level example of how a data warehouse would be fit into the Rags'n'Riches architecture is shown below-

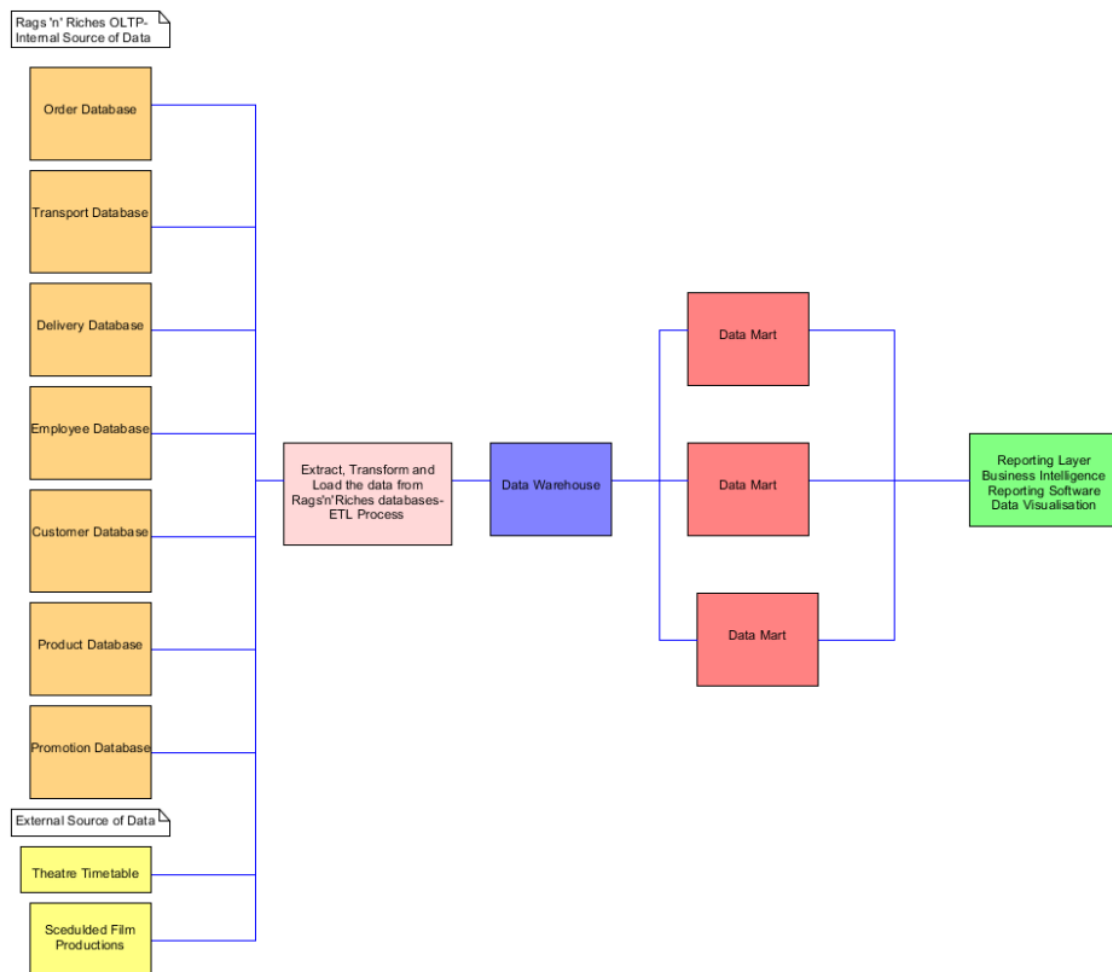


2.3 Part 3

Evidence provided above displays the advantages for Rags'n'Riches to implement a data warehouse. Data warehouses and data marts are repositories of data. The difference between the both, is a data warehouse is a central store of all data and a data mart can fulfil specific user requirements by extracting necessary information for those stakeholders to review. This extraction can be done through the company's data warehouse.

In this example, separate data marts would be applicable for the Managing Director, Finance Director and Operations Director as it will display a smaller set of data than the data warehouse. To create specific data marts for Rags'n'Riches they can use the existing data warehouse to extract the information they are concerned about.

An example is provided below for how the Rags'n'Riches architecture with a data warehouse and data marts can be utilised –

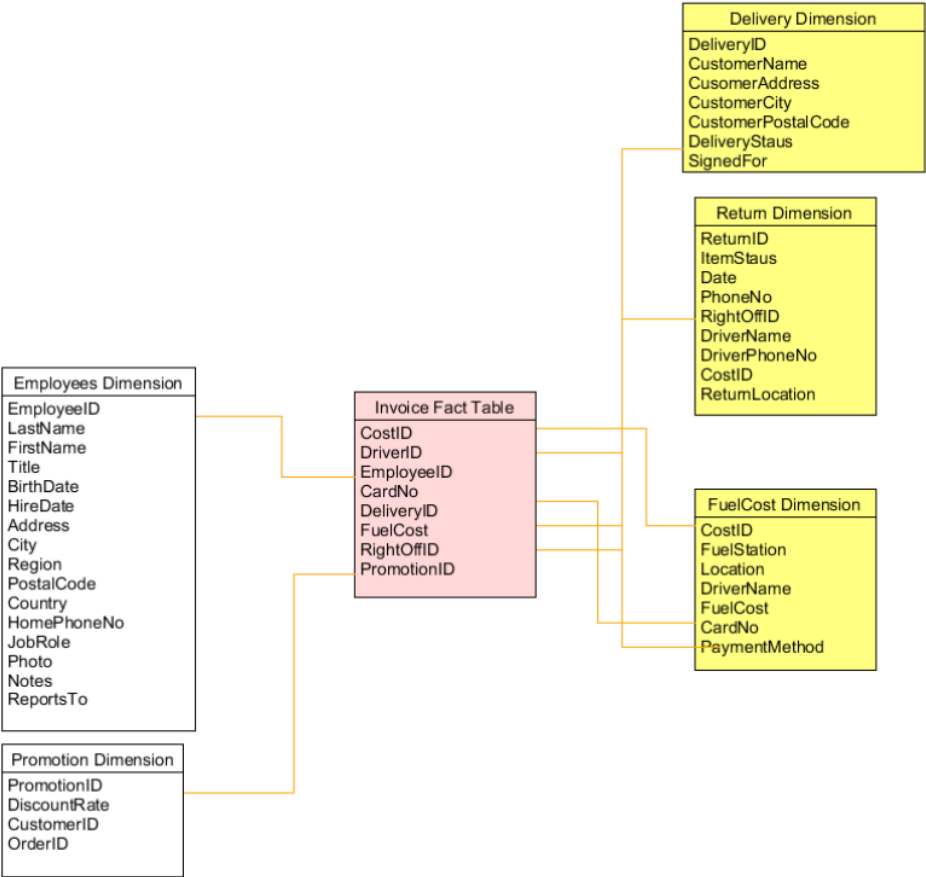


One of the advantages of using data marts is efficient access to data. They allow access to a sub section of data which saves time and enhances business intelligence as experts within the subject area can access the relevant information. If they are used alongside a data warehouse, which is the proposed solution, they can have responsibility for department maintenance of information. The Managing Director, Finance Director and Operations Director can simply access the relevant data to them, control the data being gathered and extract and analyse it against their department's key performance indicators (KPIs) to subsequently make informed decisions from the targeted data they have accessed.

Although the example above highlights the use of data marts with the company's data warehouse, it should be noted that data marts can utilise star schemas. A Star Schema separates process and

business data into facts. These facts can include data such as, Cost ID, Driver ID, Quantity, etc. The star scheme will consist of one or more fact tables that relate to several dimensional tables which reduces the need for joining tables. This architecture is in the company's data warehouse.

An example is provided below for a Rags'n'Riches star schema that demonstrates a star schema for an Invoice fact table-



Data Marts would be beneficial for the company directors as they can utilise the information extracted from their source systems and manage the display of data for their given department. The data is available for them to view but tailors it to suit their needs. Using data marts will improve the company's business intelligence and save time looking through unrelated data.

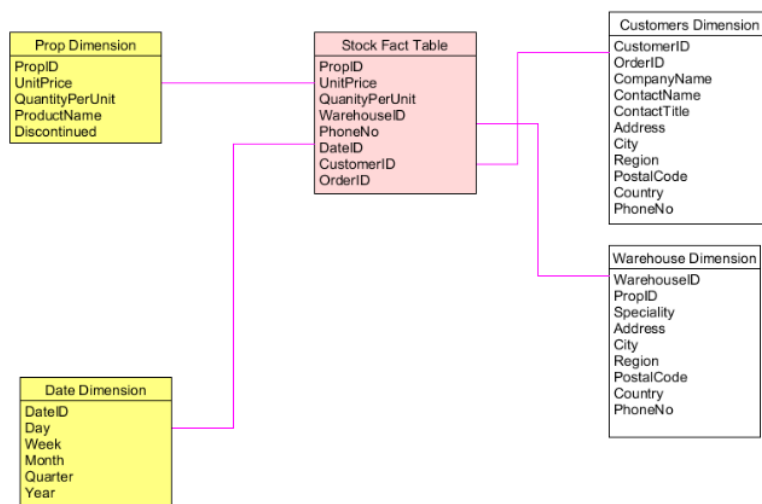
2.4 Part 4

A normalised structure of an OLTP system is processing the data into a series of normal forms. This means creating tables and establishing relationships between them to eliminate redundant information. In the 1970's Edgar F. Codd [8], highlighted the process whereby the data models should be rationalised. He set out a paper which documented a standard method of retrieving and storing computer information. Part of this process was normalising the data to the third normal form. A normalised structure includes the first (1NF), second (2NF) and third (3NF) normal form. The 1NF ensures each data field holds a single value and the 2NF eliminates repeating groups. The aim of this is to ensure the table relations are independent [9]. The 3NF uses normalising principles to reduce the duplication of data, avoid anomalies and ensure integrity within the database to simplify the data management.

A dimensional structure of a data warehouse is optimised for data Warehousing tools. Facts are used as measurements and metrics which provide context to surrounding business processes [10]. Its aim is to retrieve information faster than normalised structures. The concept was developed by Ralph Kimball where the structure consists of facts and dimension tables. It's designed to read, analyse, and summarise business data with the ability to run queries faster due to the reduced joining of tables which is displayed in the normalised structures.

A facts table can be utilised for Rags'n'Riches as it stores quantitative information for analysis. They work alongside dimension tables as the fact table will display information that needs to be analysed whereas the dimension tables will store the information that feeds into the fact table. Foreign keys are used to join the information together. This is also referred to as a star schema. It is where one large fact table is used to store transactional information which is linked to dimensional tables. Star schemes are also used to develop data marts and data warehouses as the fact tables references a number of dimensional tables [11]. It's a more effective way of handling the data and producing quick queries rather than normalising the data and joining the tables.

An example is provided below on how this can work for Rags'n'Riches-



3 Task 2

3.1 Introduction

The following information is suited to the CEO of Rags'n'Riches and uses the information provided in the case study to produce a suitable star schema for the company's new data warehouse.

3.2 Internal and external source systems

The company already have some internal source systems available. These are their transactional databases-

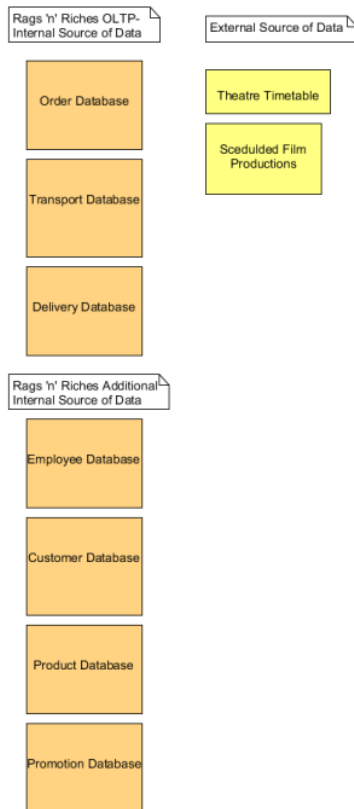
- Orders database
- Transport database
- Delivery database

To improve the architecture of the business and access to data information, it is proposed that that Rags'n'Riches introduce the following internal source systems to capture their company information-

- Employee database
- Customer database
- Product database
- Promotional database

To further enhance the company's business intelligence, external data sources can be utilised within the data warehouse. An appropriate example for Rags'n'Riches is a theatre timetable and scheduled film productions within geographical areas. These external data sources would enhance the company information to identify trends with busier periods if film productions and theatre shows are being promoted within certain areas. Using the geographical information, the company can analyse their specialist stocked areas and prepare for potential request. This could reduce their internal transporting costs as these could be moved with other transport requests.

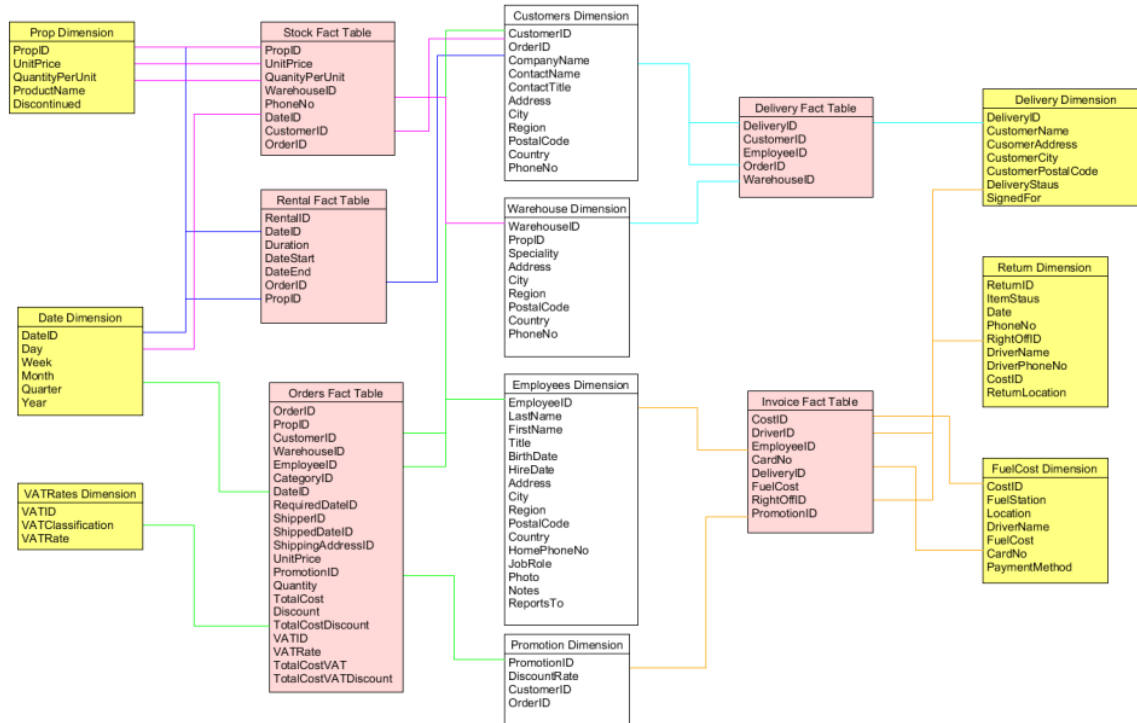
The scope of the company's internal and external source systems is displayed below:



3.3 Rags'n'Riches Star Schema

Rags'n'Riches require a number of star schemas to identify their facts and dimensions. It will separate the numerical fact data that is related to the descriptive dimensional data. The fact table is the centre of the star schema which allows machine learning tools to quickly analyse and read the data. This will lead to quicker and more efficient queries on the company's data [12].

An example is provided below which displays five star schemas and the metadata, aggregate and master data (Appendix 1).



Metadata

Stock Table Metadata	Rental Table Metadata	Orders Table Metadata	Delivery Table Metadata	Invoice Table Metadata
Metadata PropID	Metadata RentalID	Metadata OrderID	Metadata DeliveryID	Metadata CostID
TableModifiedID	TableModifiedID	TableModifiedID	TableModifiedID	TableModifiedID
RowModified	RowModified	RowModified	RowModified	RowModified
AttributeModified	AttributeModified	AttributeModified	AttributeModified	AttributeModified
OldValue	OldValue	OldValue	OldValue	OldValue
NewValue	NewValue	NewValue	NewValue	NewValue
DataSourceID	DataSourceID	DataSourceID	DataSourceID	DataSourceID
UserID	UserID	UserID	UserID	UserID
Timestamp	Timestamp	Timestamp	Timestamp	Timestamp
Comment	Comment	Comment	Comment	Comment

Aggregate Data (AD)

Stock Table AD	Rental Table AD	Orders Table AD	Delivery Table AD	Invoice Table AD
AD PropID	AD RentalID	AD OrderID	AD DeliveryID	AD CostID
AggregateDateID	AggregateDateID	AggregateDateID	AggregateDateID	AggregateDateID
DailyTotal	DailyRentalsTotal	DailyTotal	DailyDeliveryTotal	DailyTotal
WeekToDate	DailyReturnsTotal	WeekToDate	WeekDeliveryTotal	WeekToDate
MonthToDate	DailyFinesTotal	MonthToDate	DeliveryMethod	MonthToDate
YearToDate	RentalsToDate	YearToDate	Timestamp	YearToDate
Timestamp	ReturnsToDate	Timestamp	Comment	Timestamp
Comment	FinesToDate	Comment		Comment
	Timestamp			
	Comment			

Master Data

Users	Data Sources
UserID	SourceID
UserName	SourceName
UserRole	SourceLocation

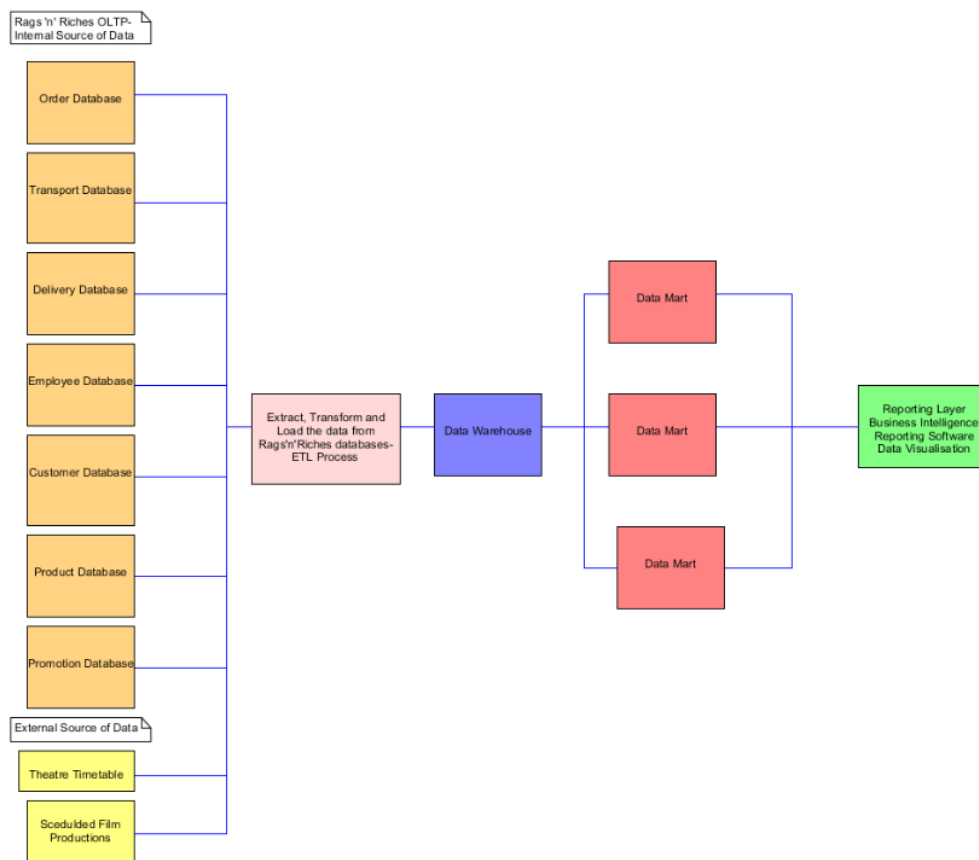
The metadata provides information about the data. It provides additional information about the data to display the changes of values or updates. For example- in the stock fact table it stores the data which is the timestamp in the metadata stock table [13].

Aggregate data contains information that has been compiled together to form a summary. For example the rental aggregate table displays the totals for daily rentals as well as other totals such as the total fines for a specific time range [14].

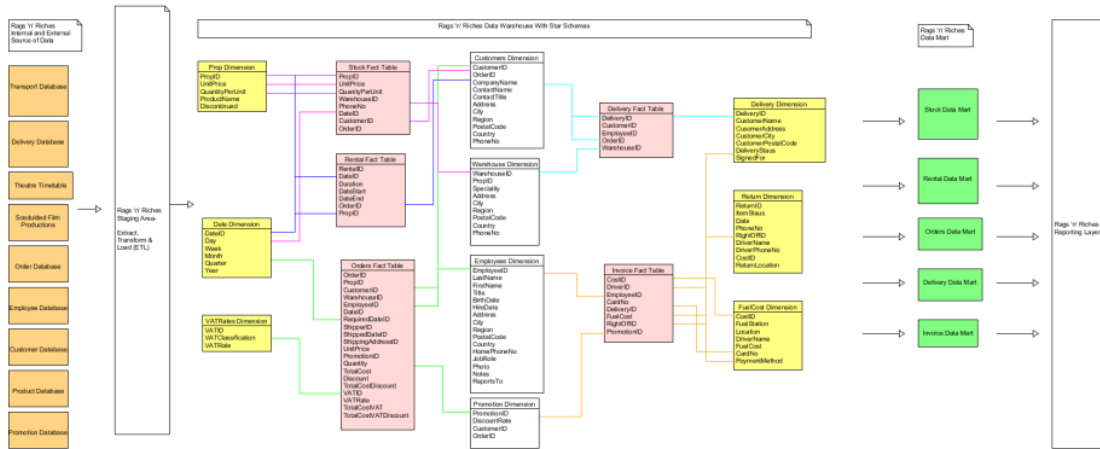
Master data tables are utilised to display the company's main business transactions and objectives. They contain the most relevant and valuable information that is shared across the organisation.

3.4 Data Marts

In the above example for the Rags'n'Riches star schemes, they will require five data marts. These data marts will be for Stock, Rental, Orders, Delivery, and Invoices. These data marts will allow a subset of company data to be accessed faster. A data mart is considered a version of a data warehouse which is designed for specific departments within the organisation [15]. They work by extracting the information from the source systems into the data warehouse and extracting the relevant information for the designed data mart. This will enhance the reporting layer and access to relevant information for the company more efficiently.

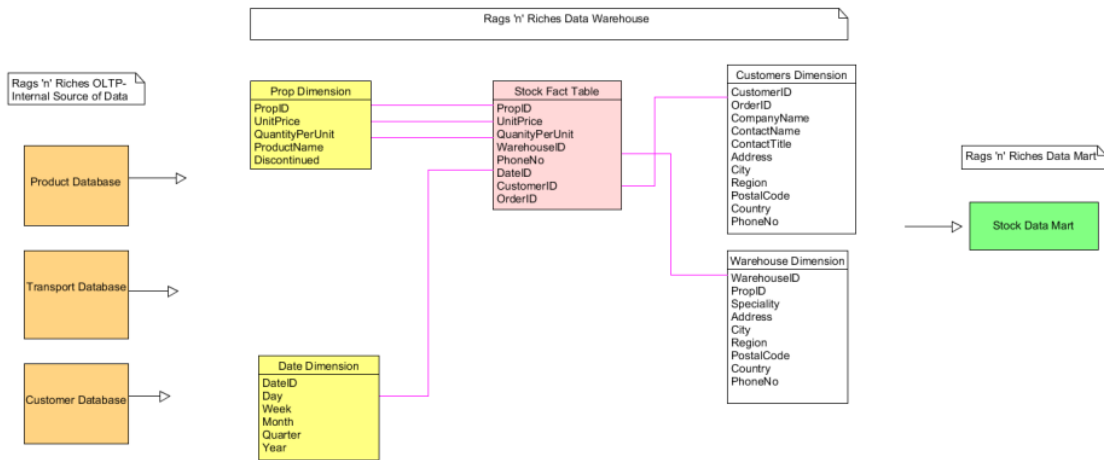


Below is an overall view of how the five data marts will be configured within the organisation (Appendix 2)-

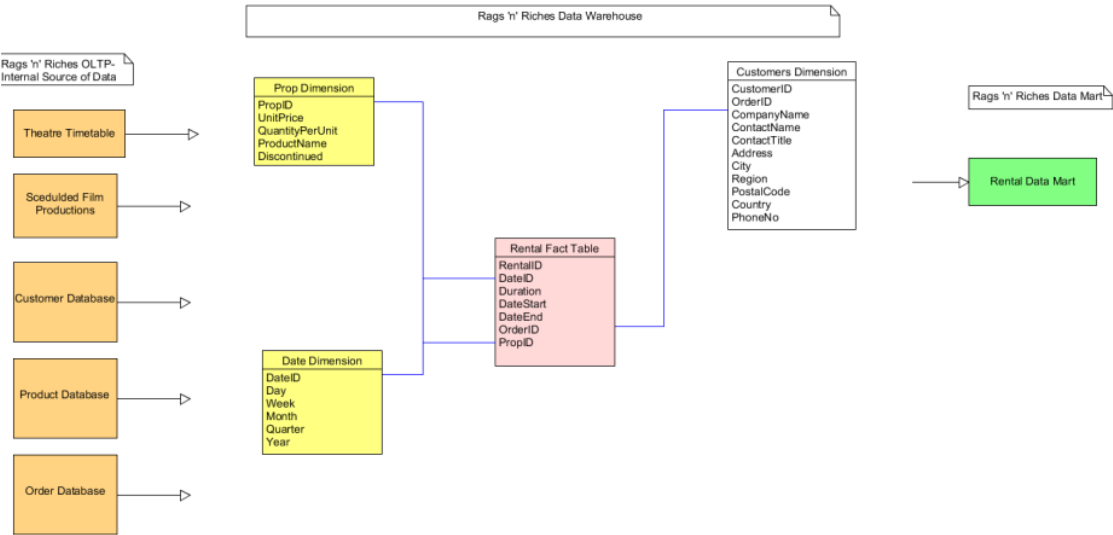


The below diagrams illustrate how the data marts will be configured using the below data sources and star schemas:

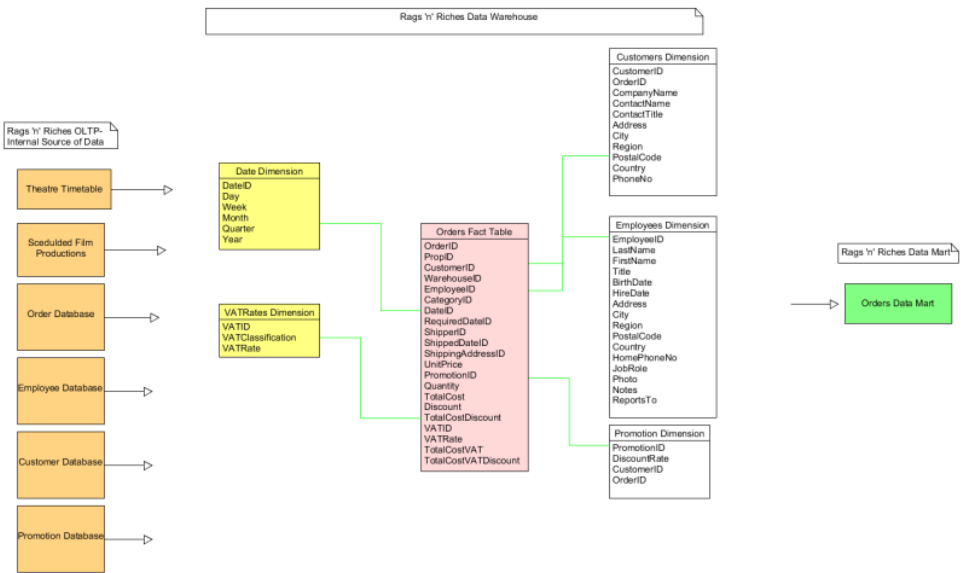
Stock data mart (Appendix 3)-



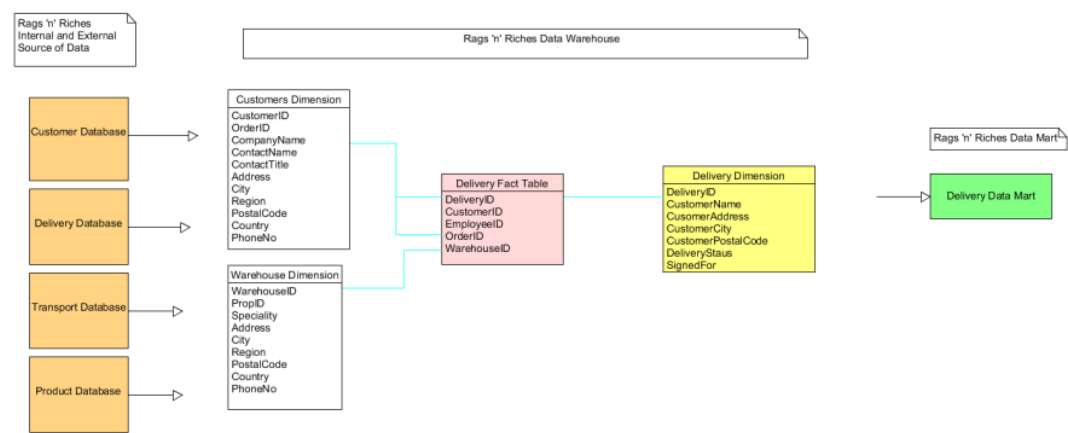
Rental data mart (Appendix 4)-



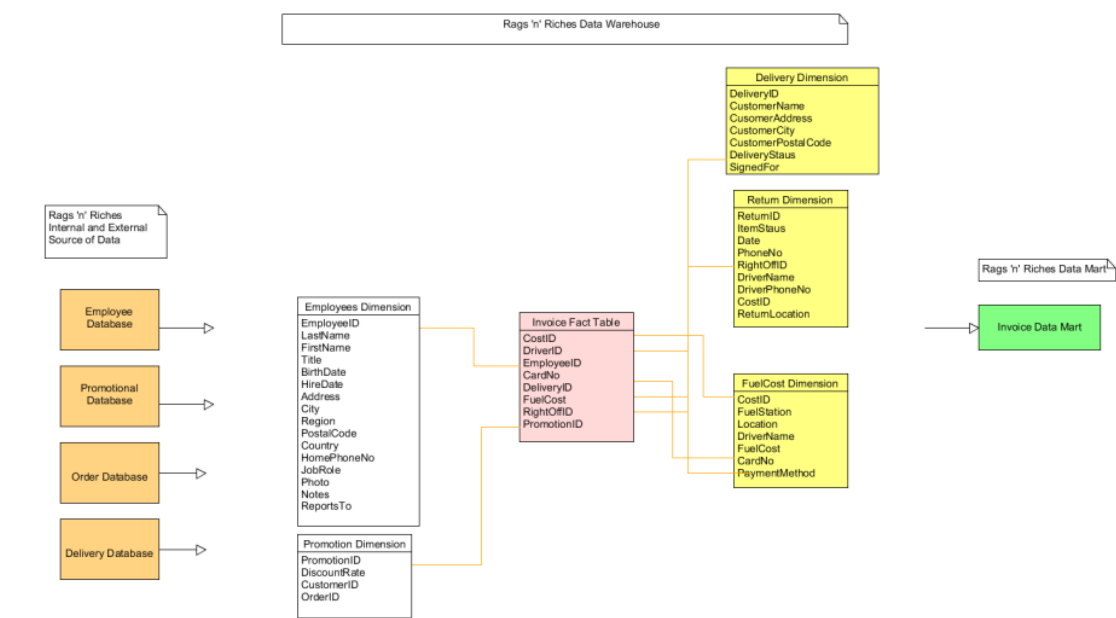
Orders data mart (Appendix 5)-



Delivery data mart (Appendix 6)-



Invoice data mart (Appendix 7)-



3.5 ETL Plan

Extracting, transforming, and loading data is a staging process to ensure the data is clean and loaded into the data warehouse successfully. This process is also referred to as ETL.

There are some core requirements when implementing an ETL plan. The business needs will need to be met. Dirty data is recommended to be handled by conforming the data into agreed dimensions from different source systems and clean the data to set out what is going to be transformed in the ETL process. As the company deals with personal and sensitive data, compliance with GDPR will need to be undertaken. This means Rags'n'Riches will have to archive their source data and document their ETL process to ensure data is allocated correctly within the data warehouse [16].

Below is a simplified version of an ETL plan suitable for Rags'n'Riches, it details where the data is sourced from and any transformations that need to be completed before it's loaded into the data warehouse.

DATABASE TABLES	DATA WAREHOUSE TABLE
DB.Order; DB.Transport; DB.Customer	Stock_Fact
DB.Order; DB.Product; DB.Customer	Rental_Fact
DB.Order; DB.Employee; DB.Cusomter	Orders_Fact
DB.Transport; DB.Delivery; DB.Order; DB.Customer; DB.Employee	Delivery_Fact
DB.Transport; DB.Delivery; DB.Order	Invoice_Fact
DB.Order; DB.Product	Product_Dimension
Not Applicable – This table needs to be generated from scratch	Date_Dimension
Not Applicable – This table needs to be generated from scratch	VATRates_Dimension
DB.Customer	Customer_Dimension
DB.Employee	Employee_Dimension
DB.ThreatreTimetable; DB.SheduldedFilmProduction; DB.Order	Warehouse_Dimension
DB.Delivery; DB.Employee	Delivery_Dimension
DB.Transport	Return_Dimension
DB.Delivery	FuelCost_Dimension

Stock_Fact

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
ProductID	DB.Order	Link Product_Dimension to	Shouldn't happen	Refer to staff
UnitPrice	DB.Order	Link Product_Dimension to	Shouldn't happen	Refer to staff
QuantityPerUnit	DB.Order	Link Product_Dimension to	Set "Unknown" value	Refer to staff
WarehouseID	DB.Transport	Link Warehouse_Dimension to	Set "Unknown" value	Refer to staff
PhoneNo	DB.Order	Link Customer_Dimension to	Set "Unknown" value	Refer to staff
DateID	DB.Order	Link Date_Dimension to	Set "Unknown" value	Refer to staff
CustomerID	DB.Customer	Link Customer_Dimension to	Set "Unknown" value	Refer to staff
OrderID	DB.Order	None	Shouldn't happen	Refer to staff

Rental_Fact

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
RentalID	DB.Order	Link Product_Dimension to	Shouldn't happen	Refer to staff
DateID	DB.Order	Link Date_Dimension to	Set "Unknown" value	Refer to staff
Duration	DB.Order	Link Date_Dimension to	Set "Unknown" value	Refer to staff
DateStart	DB.Order	Link Date_Dimension to	Set "Unknown" value	Refer to staff
DateEnd	DB.Order	= DateStart+Duration	Calculated	Calculated
OrderID	DB.Order	None	Shouldn't happen	Refer to staff
ProductID	DB.Order	Link Product_Dimension to	Shouldn't happen	Refer to staff

Order_Fact

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
OrderD	DB.Order	Link to Customer_Dimension	Shouldn't happen	Refer to staff
ProductID	DB.Order	Link to Product_Dimension	Shouldn't happen	Refer to staff
CustomerID	DB.Customer	Link to Customer_Dimension	Set to "Unknown" value	Refer to staff
WarehouseID	DB.Transport	Link to Warehouse_Dimension	Set to "Unknown" value	Refer to staff
EmployeeID	DB.Employee	Link to Employee_Dimension	Shouldn't happen	Refer to staff
DateID	DB.Order	Link to Date_Dimension	Set to "Unknown" value	Refer to staff
RequiredDateID	DB.Orders	Link to Date_Dimension	Set to OrderDateID +7 if OrderDate is known, otherwise set to "Unknown" value	Set to OrderDateID +7 if OrderDate is known, otherwise set to "Unknown" value
ShipperID	DB.Orders	Link to Shippers_Dimension	Set to "Unknown" value	Set to "Unknown" value
ShippedDateID	DB.Orders	Link to Date_Dimension	Set to "Unknown" value	Set to "Unknown" value
ShippingAddressID	DB.Orders	Link to ShippingAddress_Dimension	Set to "Unknown" value	Set to "Unknown" value
UnitPrice	DB.Order	Link to Product_Dimension	Shouldn't happen	Refer to staff
PromotionID	DB.Order	Link to Promotion_Dimension	Shouldn't happen	Refer to staff
Quantity	DB.Order	Link to Customer_Dimension	Set to 0	Set to 0
TotalCost	<i>Calculated</i>	= UnitPrice*Quantity	<i>Calculated</i>	<i>Calculated</i>
Discount	DB.Order	Link to Promotion_Dimension	Set to 0	Set to 0
TotalCostDiscount	<i>Calculated</i>	=TotalCost*(1-Discount)	<i>Calculated</i>	<i>Calculated</i>
VATID	<i>Generated</i>	Link to VATRates_Dimension	Set to "Standard" value	Set to "Standard" value
VATRate	<i>Generated</i>	=VATRates_Dimension.VAT Rate	Set to 0.2	Refer to staff
TotalCostVAT	<i>Calculated</i>	=TotalCost*(1+VATRate)	<i>Calculated</i>	<i>Calculated</i>
TotalCostVATDiscount	<i>Calculated</i>	=TotalCostDiscount*(1+VAT Rate)	<i>Calculated</i>	<i>Calculated</i>

Delivery_Fact

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
DeliveryID	DB.Order	Link Delivery_Dimension to	Shouldn't happen	Refer to staff
CustomerID	DB.Customer	Link Customer_Dimension to	Set "Unknown" value to	Refer to staff
EmployeeID	DB.Employee	Link Employee_Dimension to	Shouldn't happen	Refer to staff
OrderID	DB.Order	None	Shouldn't happen	Refer to staff
WarehouseID	DB.Transport	Link Warehouse_Dimension to	Set "Unknown" value to	Refer to staff

Invoice_Fact

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
CostID	DB.Transport	Link FuelCost_Dimension to	Shouldn't happen	Refer to staff
DriverID	DB.Delivery	Link Deliver_Dimension to	Shouldn't happen	Refer to staff
EmployeeID	DB.Employee	Link Employee_Dimension to	Shouldn't happen	Refer to staff
CardNo	DB.Transport	Link FuelCost_Dimension to	Set "Unknown" value to	Refer to staff
DeliveryID	DB.Order	Link Delivery_Dimension to	Shouldn't happen	Refer to staff
FuelCost	DB.Transport	Link FuelCost_Dimension to	Set "Unknown" value to	Refer to staff
RightOffID	DB.Order	Link Return_Dimension to	Set to "N/A" value	Refer to staff
PromotionID	DB.Order	Link Promotion_Dimension to	Shouldn't happen	Refer to staff

Product_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
ProductID	DB.Products	None – Primary key	Shouldn't happen	Refer to backup
UnitPrice	DB. Products	None	Refer to staff	Refer to staff
QuantityPerUnit	DB.Products	None	Doesn't matter	Refer to staff
ProductName	DB.Products	None	Refer to staff	Refer to staff
Discontinued	DB. Products	None	Set to FALSE	Refer to staff

Date_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
DateID	<i>Generated</i>	None – Primary key	Shouldn't happen	<i>Generated</i>
Day	<i>Generated</i>	None	<i>Generated</i>	<i>Generated</i>
Week	<i>Generated</i>	None	<i>Generated</i>	<i>Generated</i>
Month	<i>Generated</i>	None	<i>Generated</i>	<i>Generated</i>
Quarter	<i>Generated</i>	None	<i>Generated</i>	<i>Generated</i>
Year	<i>Generated</i>	None	<i>Generated</i>	<i>Generated</i>

VATRates_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
VATID	<i>Generated</i>	None – Primary key	<i>Generated</i>	Refer to staff
VATClassification	<i>Generated</i>	None	<i>Generated</i>	Refer to staff
VATRate	<i>Generated</i>	None	<i>Generated</i>	Refer to staff

Customer_Dimension

TARGET ATTRIBUTE	SOURCE(S)	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
CustomerID	DB.Customer	None – Primary key	Shouldn't happen	Refer to staff
OrderID	DB.Customer			
CompanyName	DB.Customer	None	Refer to staff	Refer to staff
ContactName	DB.Customer	None	Refer to staff	Refer to staff
ContactTitle	DB.Customer	None	Refer to staff	Refer to staff
Address	DB.Customer	None	Refer to staff	Refer to staff
City	DB.Customer	None	Refer to staff	Refer to staff
Region	DB.Customer	None	Refer to staff	Refer to staff
PostalCode	DB.Customer	None	Refer to staff	Refer to staff
Country	DB.Customer	None	Refer to staff	Refer to staff
PhoneNo	DB.Customer	None	Refer to staff	Refer to staff

Promotion_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
PromotionID	DB.Promotion	None – Primary key	Shouldn't happen	Refer to staff
DiscountRate	DB.Promotion	None	Refer to staff	Refer to staff
CustomerID	DB.Promotion	None	Refer to staff	Refer to staff
OrderID	DB.Promotion	None	Refer to staff	Refer to staff

Warehouse_Dimension

TARGET ATTRIBUTE	SOURCE(S)	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
WarehouseID	DB.Order	None – Primary key	Shouldn't happen	Refer to staff
ProductID	DB.Customer	None	Refer to staff	Refer to staff
Speciality	DB.Customer	None	Refer to staff	Refer to staff
Address	DB.Customer	None	Refer to staff	Refer to staff
City	DB.Customer	None	Refer to staff	Refer to staff
Region	DB.Customer	None	Refer to staff	Refer to staff
PostalCode	DB.Customer	None	Refer to staff	Refer to staff
Country	DB.Customer	None	Refer to staff	Refer to staff
PhoneNo	DB.Customer	None	Refer to staff	Refer to staff

Employee_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
EmployeeID	DB.Employee	None – Primary key	Shouldn't happen	Refer to staff
LastName	DB.Employee	None	Refer to staff	Refer to staff
FirstName	DB.Employee	None	Refer to staff	Refer to staff
Title	DB.Employee	None	Refer to staff	Refer to staff
BirthDate	DB.Employee	None	Refer to staff	Refer to staff
HireDate	DB.Employee	None	Refer to staff	Refer to staff
Address	DB.Employee	None	Refer to staff	Refer to staff
City	DB.Employee	None	Refer to staff	Refer to staff
Region	DB.Employee	None	Refer to staff	Refer to staff
PostalCode	DB.Employee	None	Refer to staff	Refer to staff
Country	DB.Employee	None	Refer to staff	Refer to staff
HomePhoneNo	DB.Employee	None	Refer to staff	Refer to staff
JobRole	DB.Employee	None	Refer to staff	Refer to staff
Photo	DB.Employee	None	Refer to staff	Refer to staff
Notes	DB.Employee	None	Refer to staff	Refer to staff
ReportsTo	DB.Employee	None	Refer to staff	Refer to staff

Delivery_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
DeliveryID	DB.Delivery	None – Primary key	Shouldn't happen	Refer to staff
CustomerName	DB.Delivery	None	Refer to staff	Refer to staff
CustomerAddress	DB.Delivery	None	Refer to staff	Refer to staff
CustomerCity	DB.Delivery	None	Refer to staff	Refer to staff
CustomerPostalCode	DB.Delivery	None	Refer to staff	Refer to staff
DeliveryStatus	DB.Delivery	None	Refer to staff	Refer to staff
SignedFor	DB.Delivery	None	Refer to staff	Refer to staff

Return_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
ReturnID	DB.Transport	None – Primary key	Shouldn't happen	Refer to staff
ItemStatus	DB.Transport	None	Refer to staff	Refer to staff
Date	DB.Transport	None	Refer to staff	Refer to staff
PhoneNo	DB.Transport	None	Refer to staff	Refer to staff
RightOffID	DB.Transport	None	Set to FALSE	Refer to staff
DriverName	DB.Transport	None	Refer to staff	Refer to staff
DriverPhoneNo	DB.Transport	None	Refer to staff	Refer to staff
CostID	DB.Transport	None	Refer to staff	Refer to staff
ReturnLocation	DB.Transport	None	Refer to staff	Refer to staff

FuelCost_Dimension

TARGET ATTRIBUTE	SOURCE	PROCESSING REQUIRED	MISSING VALUE?	INCORRECT VALUE?
CostID	DB.Transport	None – Primary key	Shouldn't happen	Refer to staff
FuelStation	DB.Transport	None	Refer to staff	Refer to staff
Location	DB.Transport	None	Refer to staff	Refer to staff
DriverName	DB.Transport	None	Refer to staff	Refer to staff
FuelCost	DB.Transport	None	Refer to staff	Refer to staff
CardNo	DB.Transport	None	Refer to staff	Refer to staff
PaymentMethod	DB.Transport	None	Refer to staff	Refer to staff

3.6 Conclusion

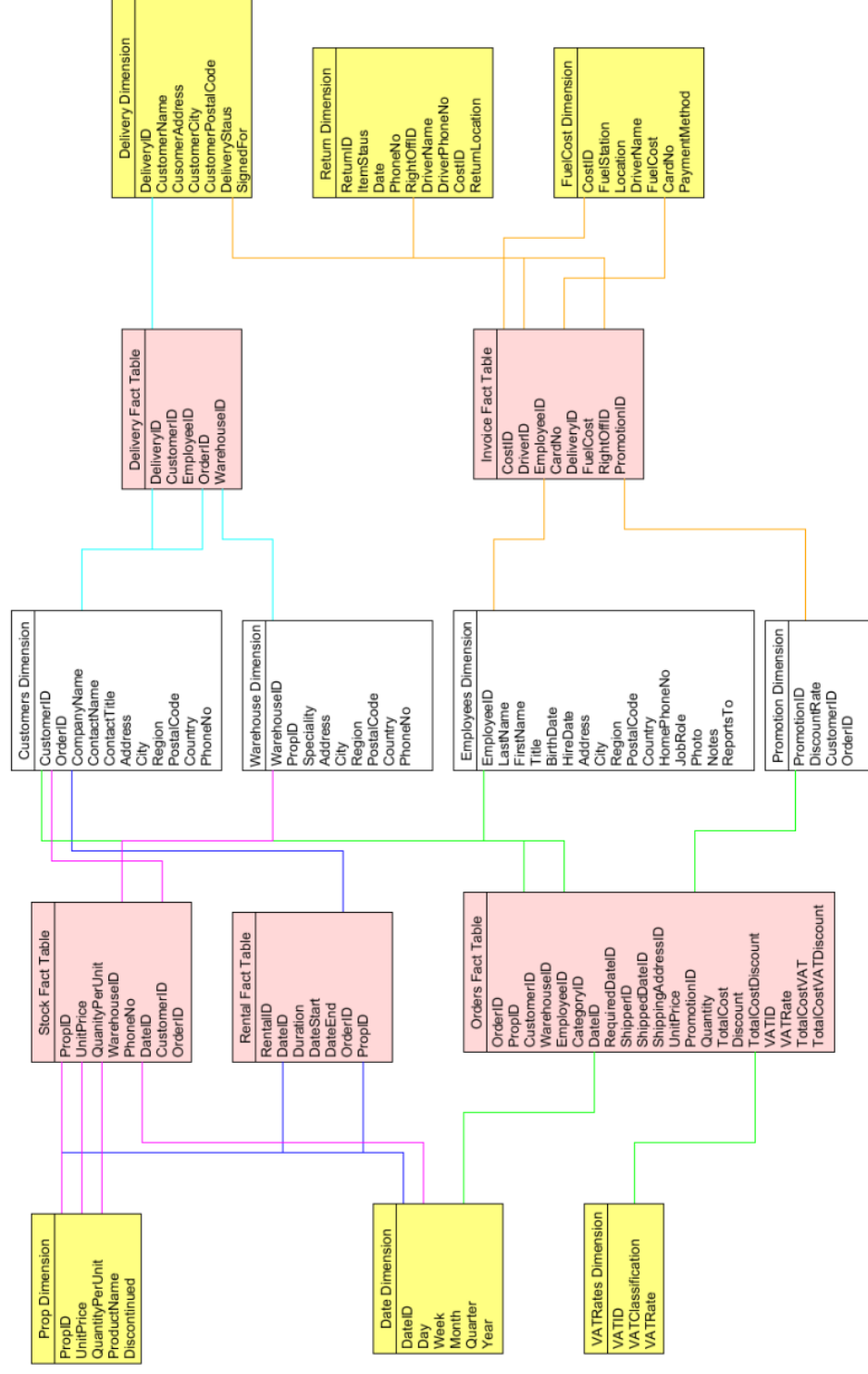
In conclusion Rags'n'Riches should implement the architecture detailed in the report. Improvements to their source systems will gather more raw data to be extracted, transformed, and loaded into the data warehouse by following the ETL plan. The processed data will be contained within the warehouse and extracted through individual data marts which are suitable to the company's requirements. This will allow more streamlined queries and faster access to meaningful data. This data can be loaded into a reporting tool for non-technical and decision makers to access to ultimately make informed decisions on business processes.

4 References

- [1] UWTSD, "Data Warehousing and Data Mining Assignment," UWTSD, [Online]. Available: https://moodle.uwtsd.ac.uk/pluginfile.php/565756/mod_resource/content/0/ACCA6026_sem1_CW1_2021-22_DataWarehousingDataMining.pdf.
- [2] SAS, "Big Data Analytics," SAS, [Online]. Available: https://www.sas.com/en_gb/insights/analytics/big-data-analytics.html. [Accessed 20 03 2022].
- [3] D. Cardon, "Database vs Data Warehouse: A Comparative Review," Health Catalyst, 03 09 2018. [Online]. Available: <https://www.healthcatalyst.com/insights/database-vs-data-warehouse-a-comparative-review/>. [Accessed 10 03 2022].
- [4] Aunalytics, "Understanding Analytics Part 1: Top Internal Sources of Big Data," Aunalytics, [Online]. Available: <https://www.aunalytics.com/understanding-analytics-part-1-top-internal-sources-of-big-data/>. [Accessed 21 03 2022].
- [5] K. Group, "ETL Architecture's 34 Subsystems," Kimball Group, [Online]. Available: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/etl-architecture-34-subsystems/>. [Accessed 18 03 2022].
- [6] Matillion, "The Importance of ETL Tools in Data Warehousing," Matillion, 27 03 2020. [Online]. Available: <https://www.matillion.com/resources/blog/the-importance-of-etl-tools-in-data-warehousing>. [Accessed 20 03 2022].
- [7] Tableau, "What is business intelligence? Your guide to BI and why it matters," Tableau, [Online]. Available: <https://www.tableau.com/en-gb/learn/articles/business-intelligence>. [Accessed 23 03 2022].
- [8] W. L.Hosch, "Edgar Frank Codd," Britannica, [Online]. Available: <https://www.britannica.com/biography/Edgar-Frank-Codd>. [Accessed 20 03 2022].
- [9] R. Becker, "Third Normal Form," Techopedia, 14 08 2020. [Online]. Available: <https://www.techopedia.com/definition/22561/third-normal-form-3nf>. [Accessed 16 03 2022].
- [10] D. Taylor, "Dimensional Modeling," Guru99, 19 02 2022. [Online]. Available: <https://www.guru99.com/dimensional-model-data-warehouse.html>. [Accessed 23 03 2022].
- [11] Wikipedia, "Star schema," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Star_schema. [Accessed 21 03 2022].
- [12] M. Smallcombe, "Snowflake Schemas vs Star Schemas: What Are They and How Are They Different," Integrate.io, 12 02 2022. [Online]. Available: <https://www.integrate.io/blog/snowflake-schemas-vs-star-schemas-what-are-they-and-how-are-they-different/>. [Accessed 21 03 2022].
- [13] G. Kranz, "Metadata," WhatIs.com, [Online]. Available: <https://whatIs.techtarget.com/definition/metadata>. [Accessed 21 03 2022].
- [14] IBM, "Data aggregation," IBM, 05 03 2021. [Online]. Available: <https://www.ibm.com/docs/en/tnpm/1.4.2?topic=data-aggregation>. [Accessed 18 03 2022].
- [15] "What is Data Mart?," CA03CA3405Notes, [Online]. Available: http://www.nitjsr.ac.in/course_assignment/CA03CA3405Notes%20on%20Data%20Mart.pdf. [Accessed 21 03 2022].
- [16] A. R.Earls, "6 data storage compliance strategies for the enterprise," TechTarget, 19 05 2021. [Online]. Available: <https://www.techtarget.com/searchstorage/tip/6-data-storage-compliance-strategies-for-the-enterprise>. [Accessed 21 03 2022].

5 Appendix

Appendix 1-



Metadata

Stock Table Metadata									
Metadata PropID	TableModifiedID	RowModified	AttributeModified	OldValue	NewValue	DataSourceID	UserID	Timestamp	Comment

Rental Table Metadata									
Metadata RentalID	TableModifiedID	RowModified	AttributeModified	OldValue	NewValue	DataSourceID	UserID	Timestamp	Comment

Orders Table Metadata									
Metadata OrderID	TableModifiedID	RowModified	AttributeModified	OldValue	NewValue	DataSourceID	UserID	Timestamp	Comment

Delivery Table Metadata									
Metadata DeliveryID	TableModifiedID	RowModified	AttributeModified	OldValue	NewValue	DataSourceID	UserID	Timestamp	Comment

Invoice Table Metadata									
Metadata CostID	TableModifiedID	RowModified	AttributeModified	OldValue	NewValue	DataSourceID	UserID	Timestamp	Comment

Aggregate Data (AD)

Stock Table AD									
AD PropID	AggregateDateID	DailyTotal	WeekToDate	MonthToDate	YearToDate	Timestamp	Comment		

Rental Table AD									
AD RentalID	AggregateDateID	DailyRentalsTotal	DailyReturnsTotal	DailyFinesTotal	RentalsToDate	ReturnsToDate	FinesToDate	Timestamp	Comment

Orders Table AD									
AD OrderID	AggregateDateID	DailyTotal	WeekToDate	MonthToDate	YearToDate	Timestamp	Comment		

Delivery Table AD									
AD DeliveryID	AggregateDateID	DailyDeliveryTotal	WeekDeliveryTotal	DeliveryMethod	Timestamp	Comment			

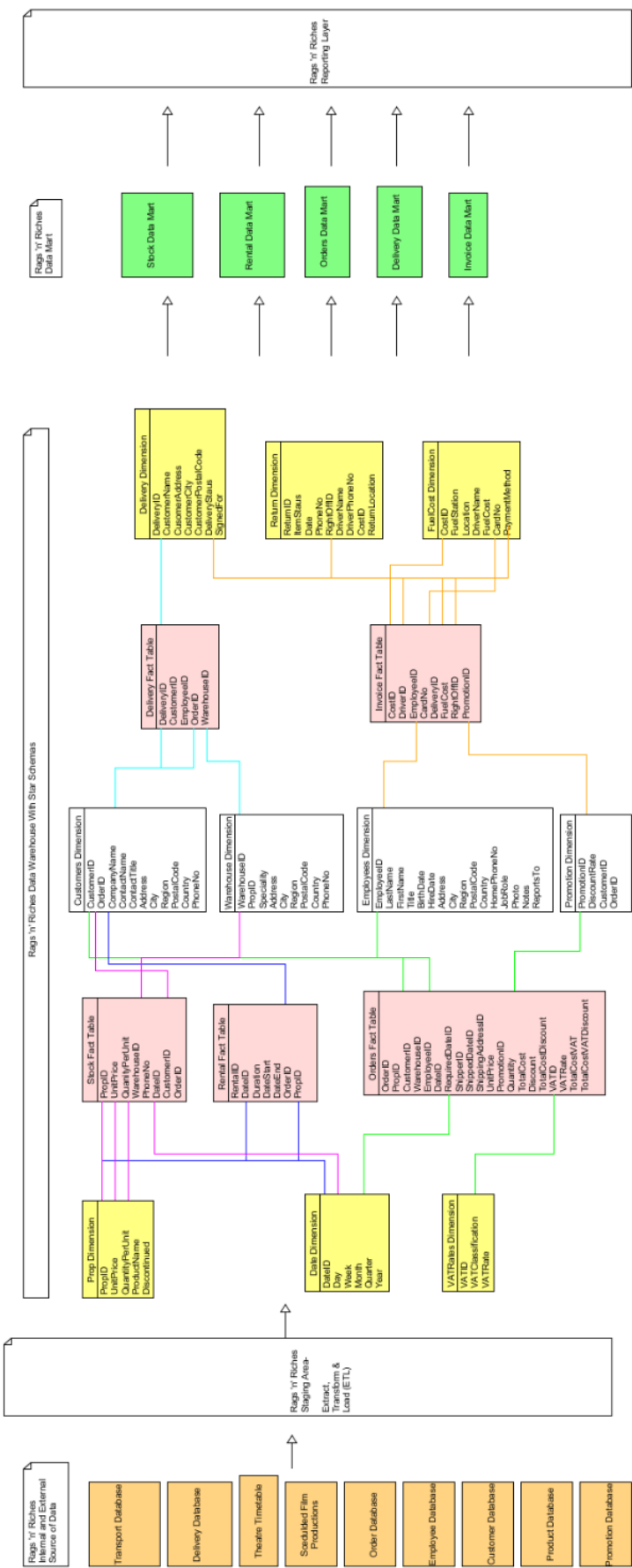
Invoice Table AD									
AD CostID	AggregateDateID	DailyTotal	WeekToDate	MonthToDate	YearToDate	Timestamp	Comment		

Master Data

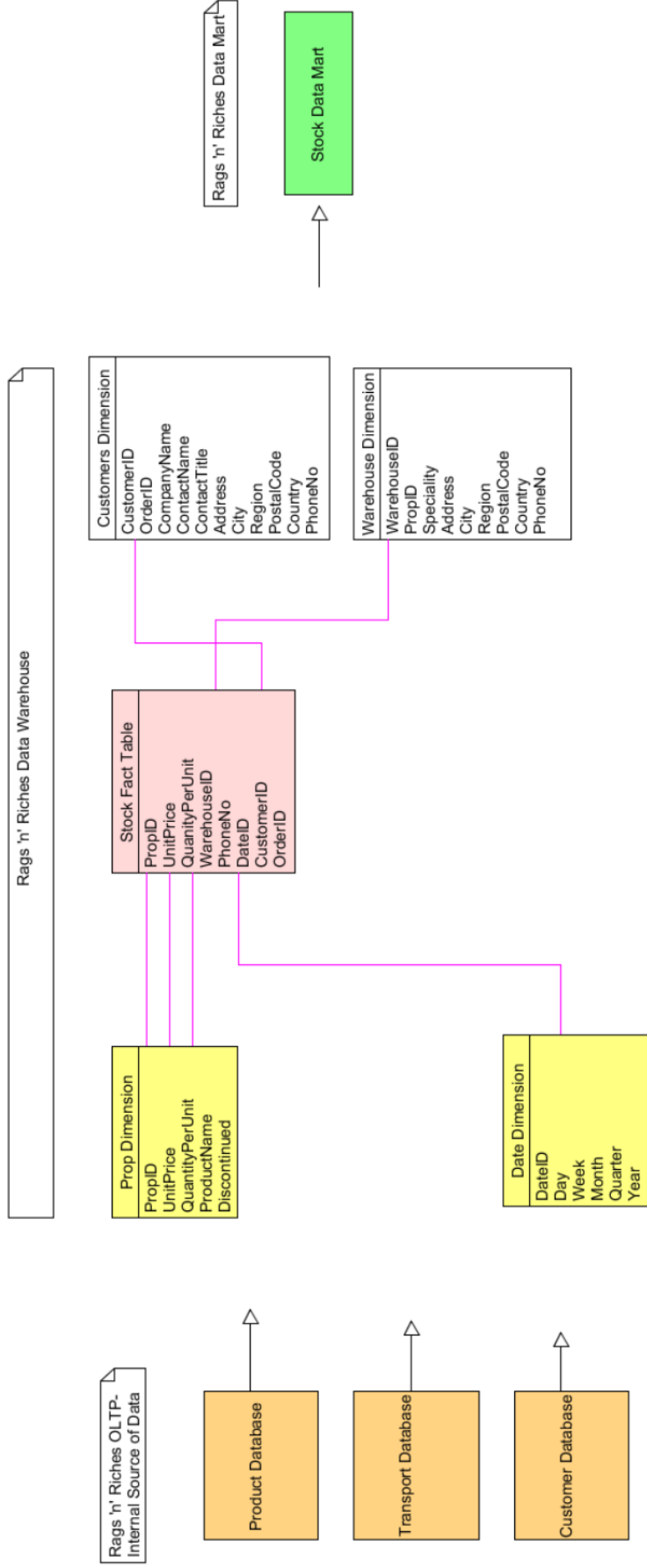
Users	
UserID	UserName
UserRole	

Data Sources		
SourceID	SourceName	
SourceLocation		

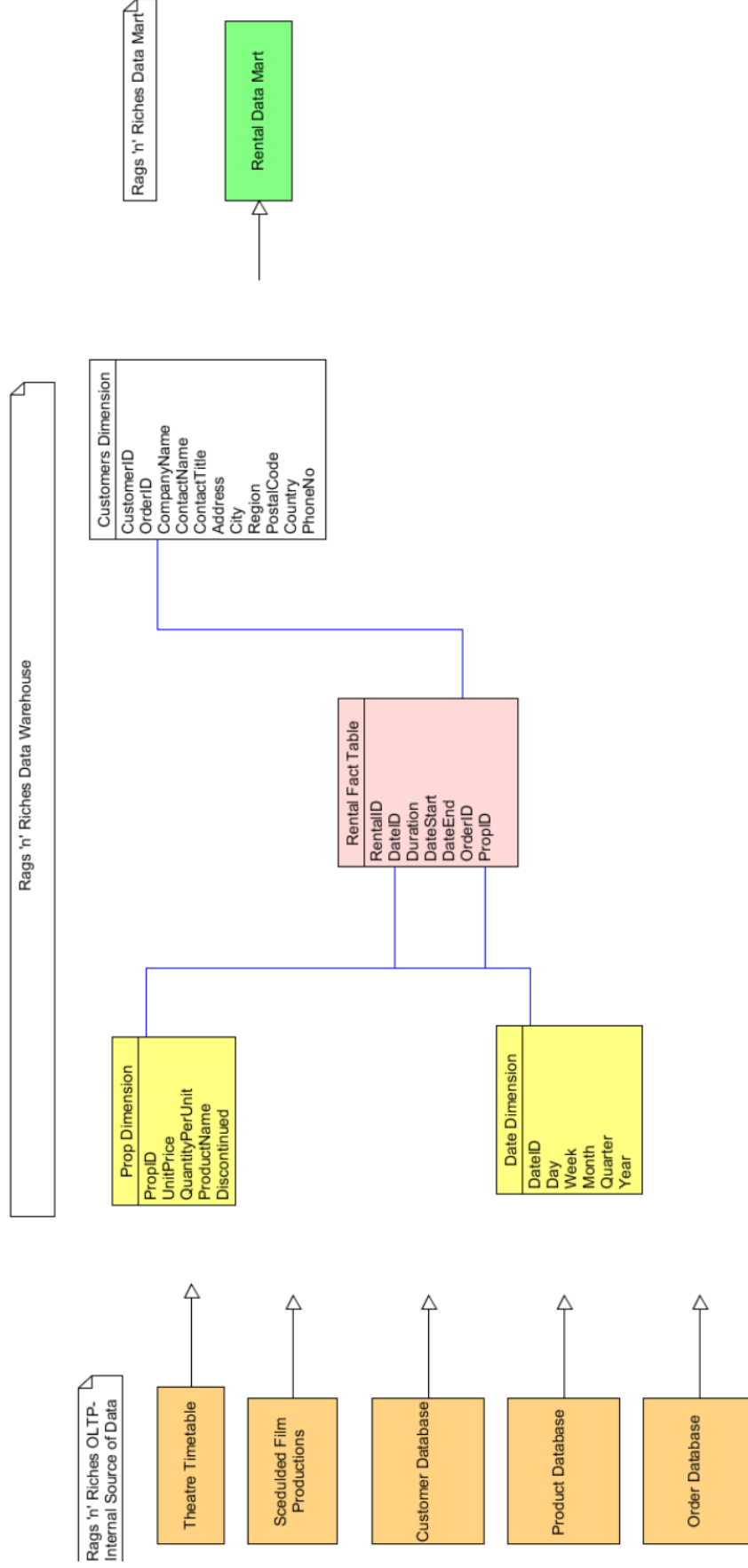
Appendix 2-

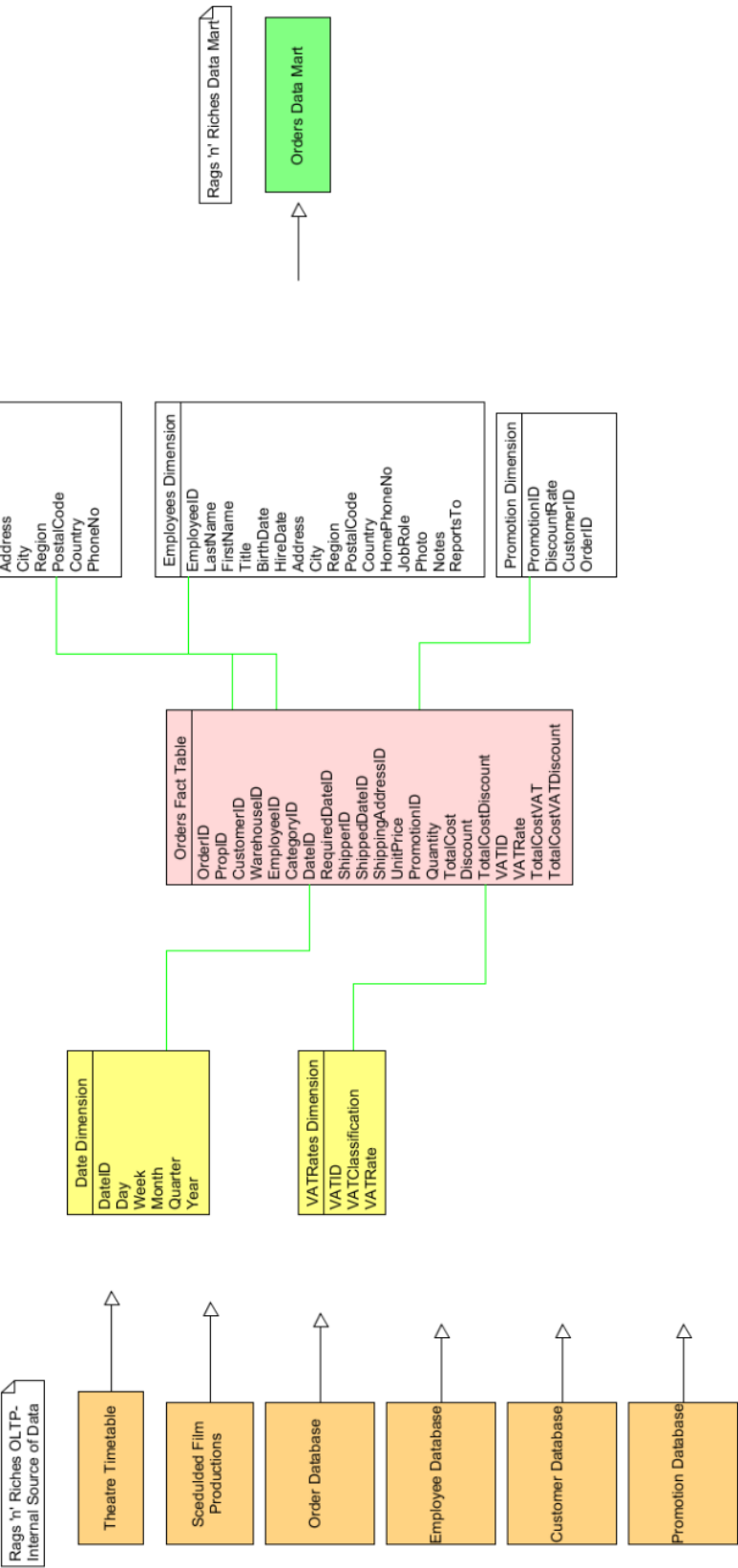


Appendix 3-



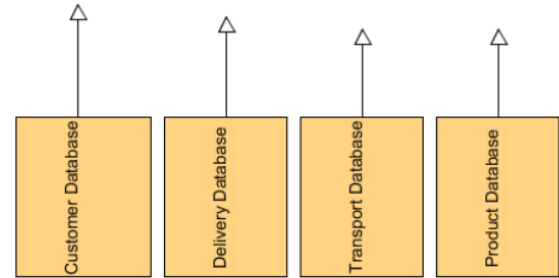
Appendix 4-





Appendix 6-

Rags 'n' Riches
Internal and External
Source of Data



Customers Dimension									
CustomerID	OrderID	CompanyName	ContactName	ContactTitle	Address	City	Region	PostalCode	Country
PhoneNo									

Warehouse Dimension									
WarehouseID	PropID	Speciality	Address	City	Region	PostalCode	Country	PhoneNo	

Delivery Fact Table				
DeliveryID	CustomerID	EmployeeID	OrderID	WarehouseID

Delivery Dimension						
DeliveryID	CustomerName	CustomerAddress	CustomerCity	CustomerPostalCode	DeliveryStatus	SignedFor

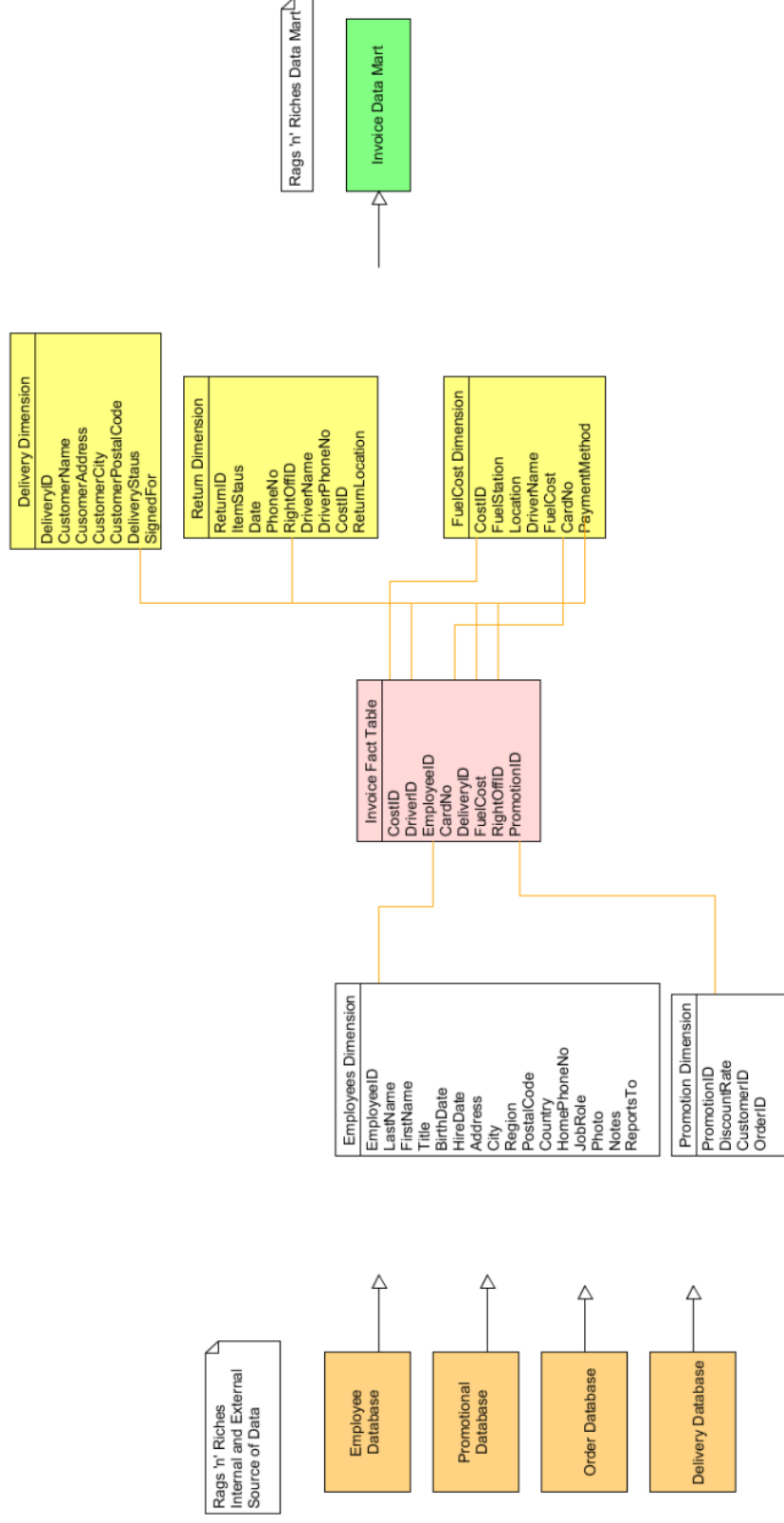
Rags 'n' Riches Data Mart

Delivery Data Mart

Rags 'n' Riches Data Warehouse

Appendix 7-

Rags 'n' Riches Data Warehouse



Data Warehousing and Data Mining- 1809781

GRADEMARK REPORT

FINAL GRADE

70/100

GENERAL COMMENTS

Instructor

Jessica Marfell

1 engagement with literature
comment available

16 references sourced but try to use a wider range of reference types (don't be so reliant on web sources and blogs, make greater use of books and peer-reviewed journals) 20

references cited throughout the body of the report- more citations would have been beneficial (new fact new citation)

2 Knowledge and Understanding
comment available

Sound understanding shown throughout

Very Good Descriptions of the ETL process 20

3. Cognitive and Intellectual skills
comment available

Good identification of the source systems and user groups and discussion of who will use different elements of the data. 10

4. Practical Application Skills
comment available

Very Good designs of starschemas and analysis of meta data - some diagrams are a little difficult to understand due to the way the relations overlap 40

reasonable internal and external datasources assumed

5. Transferable Skills for life and Professional practice
comment available

Report is well laid out and structured 10

minor grammatical and spelling errors

bullet points within table on page 7 are
unnecessary- consider a separate row
Total:

PAGE 1
PAGE 2
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
PAGE 16
PAGE 17
PAGE 18
PAGE 19
PAGE 20
PAGE 21
PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36