

Anton Hall 1809131 Data Warehousing - ACCA6026

by Anton Hall

Submission date: 25-Mar-2022 07:36AM (UTC+0000)

Submission ID: 174499111

File name: 12316_Anton_Hall_Anton_Hall_1809131_Data_Warehousing_-_ACCA6026_398298_1721582199.docx (510.45K)

Word count: 3045

Character count: 16960

University of Wales, Trinity Saint David

Data Warehousing and Data Mining - ACCA6026

Anton Hall

Contents

1.Task 1 – Data Warehouse Report	1
1.1 OLTP vs Data Warehouse	1
1.2 Data Warehouse Architecture	1
1.3 Data Marts	3
1.4 Normalised Data vs Dimensional Data.....	3
2. Task 2 – Data Warehouse Solution	4
2.1 Internal and External Sources	4
2.2 Appropriate Star Schema	5
2.3 Data Marts	6
2.4 ETL Plan	9
Conclusion.....	12
References	13
Bibliography	13

1.Task 1 – Data Warehouse Report

1.1 OLTP vs Data Warehouse

Online Transactional Processing (OLTP) is a type of data processing system, it sits alongside another type of data processing system called online analytical processing (OLAP). Both types of systems are used in businesses to process large amounts of data. However, the use case of each type of system can be quite specific to the end goal of the user or business, OLTP is normally used for operational systems and OLAP is normally used to gain business insights.

OLTP systems use a relational structure to process many transactions by lots of people, normally over the internet. They are designed to be efficient in executing large numbers of simple transactions.

“OLTP applications typically access a small number of tables (preferably one, but usually only two or three) during a given transaction, a series of steps that either access data from a database or put data into it.” [1]

Some examples of OLTP include Banking, Booking Systems and Online Transactions. OLTP databases are normalised. In the case of Rags’n’Riches, OLTP databases will be used to store data for specific parts of the business.

OLAP on the other hand is what is used to describe a data warehouse structure, predominately used for business insights and analytics. Data warehouses are used to bring data together from the entire business data landscape to enable the analysis of data in a one place. A data warehouse gives the ability for all data for be stored in a single location where it can be queried.

“Data warehousing is the coordinated, architected, and periodic copying of data from various sources, both inside and outside the enterprise, into an environment optimized for analytical and informational processing.” [1]

A benefit of incorporating a data warehouse into a business is for the increased business intelligence. It helps end users make more informed decisions quicker by giving them more visibility of data across the business in a single place.

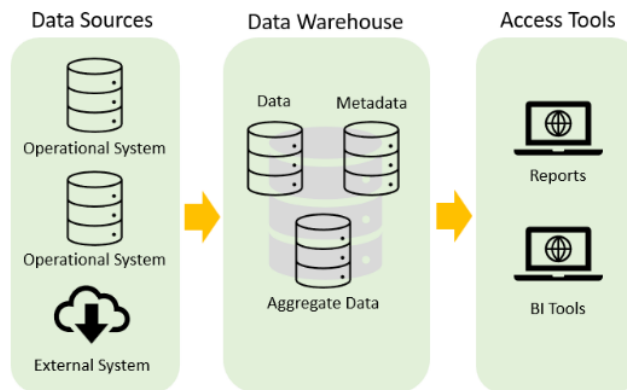
1.2 Data Warehouse Architecture

As mentioned above a data warehouse is type of data processing system that is used to support analytics across an entire business data ecosystem. The data in a data warehouse is normally extracted from a wide range of sources, both from within and outside of the business. Consolidating large amounts of data within a data warehouse allows a business to retrieve valuable insights from their data to aid them in decision making [2]. To achieve valuable insights a Datawarehouse needs careful consideration when being developed.

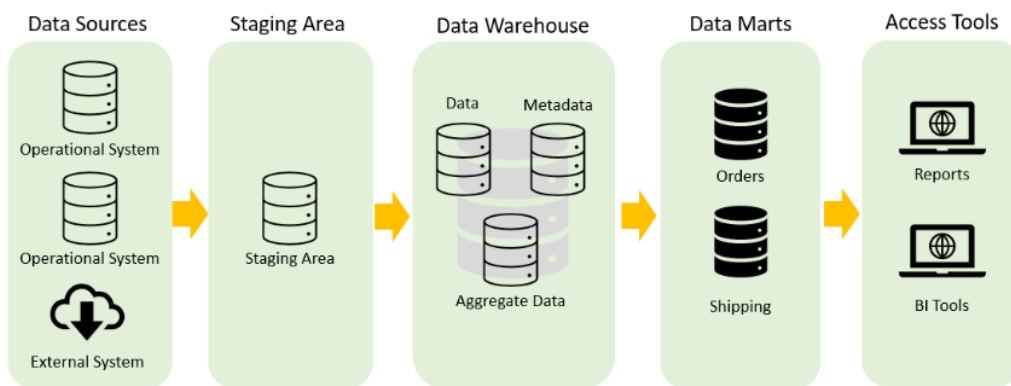
A conventional data warehouse is made up of four main components:

1. A central database – The foundation of the data warehouse, traditionally this would be a relational database on premise or in the cloud.
2. ETL (Extract Transform and Load) tools – Data is extracted from source systems, transformed into a new structure, and loaded to the central database
3. Metadata – Data about the data, contains information around the source, values and describes any alterations made to the data.
4. Access tools – Tools that allow users to interact with the data in the data warehouse. [3]

The figure below shows the basic architecture of a data warehouse:



There are two additional components that would be recommended for the Rags'n'Riches data warehouse. The first would be a staging area; this is where the data from the sources would be loaded before it is transformed and loaded into the data warehouse, this aids with the ETL process. The second would be the addition of data marts, these will be explained in the next section. The figure below illustrates the system architecture recommended:



As well as the design of the components of the data warehouse careful planning and development is needed for the ETL process. ETL is the process in which data is extracted from source systems, transformed into meaningful data, and loaded into the data warehouse.

The first step of the process is to extract the data from source systems such as business systems, APIs, and transactional databases. There are several ways and frequency in which this data can be extracted. In the case of Rag'n'Riches, the recommendation would be to extract the data outside of business hours on a daily occurrence.

The second step of the process is transforming the raw data that has been extracted from the different sources into a specific schema format. This is where the data is cleaned to meet business needs. The transformation takes place between the staging area and the data warehouse.

Finally, once the data has been transformed it is loaded into the data warehouse.

ETL is usually executed with a data pipeline. Data pipelines are developed to perform consistent moving of data from source systems into the data warehouse. [4]

1.3 Data Marts

A data mart can be defined as,

“A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.” [5]

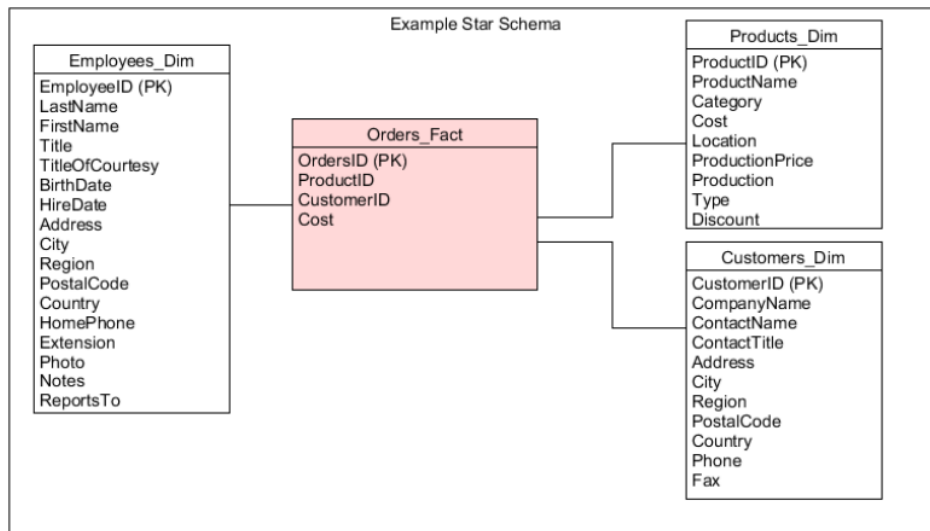
In the case of Rags’n’Riches the suggestion would be to include at least two data marts within the data warehouse architecture. The managing director would likely need a summary overview of all departments; therefore it wouldn’t be efficient to create a data mart to meet their business needs as they are not specific to a particular section of the business. The managing directors review of the data would be less frequent than that of the finance director and operations director so again the ability to gain quick insights is negated. The intention would be to create a data mart specific to the needs of the finance department. In Rags’n’Riches this DataMart would be focussed on the quick access to any of the financial data such as invoices. Another data mart would be designed dedicated to the operational team, this would be focussed on any data relevant to the daily running of the business such as orders and stock movements.

It would be suggested that dependant data marts be created, these are portioned sections within a data warehouse. This top-down approach extracts data from the main storage location of all business data into a defined data mart for analysis.

1.4 Normalised Data vs Dimensional Data

OLTP databases are normalised in structure, and normalisation refers to how data is organised within a database. Normalised databases store their data in tables, and each table is normally focused on one entity, this eliminates redundancy and inconsistent dependency. There are three levels of normalisation – first, second and third normal form. [6]

Data warehouses are designed using dimensional data modelling techniques, in a way that will aid end users in enquiring the data held within it. The dimensional data modelling techniques splits data up into two types of tables – Dimension tables and Fact tables. Dimension tables are used to store descriptive attributes whereas fact tables contain measures and key relationships. A star schema is normally made up of at least one fact table that contains foreign keys to link it to many dimension tables. In the case of Rags’n’Riches the concept of a star schema will be explained with the example of analysing orders. The measure would be the number of orders, or when they were placed so this would lead us to create an orders fact table. However, this is not the only information we would want to analyse, this is where dimension tables are used to add descriptive information to each line in the fact table. For example, we may want to see which customer placed the order, which employee handled the order or more information about the product that was ordered. The example star schema below illustrates this concept, showing the orders fact table in red with the supporting dimension tables in white:



2. Task 2 – Data Warehouse Solution

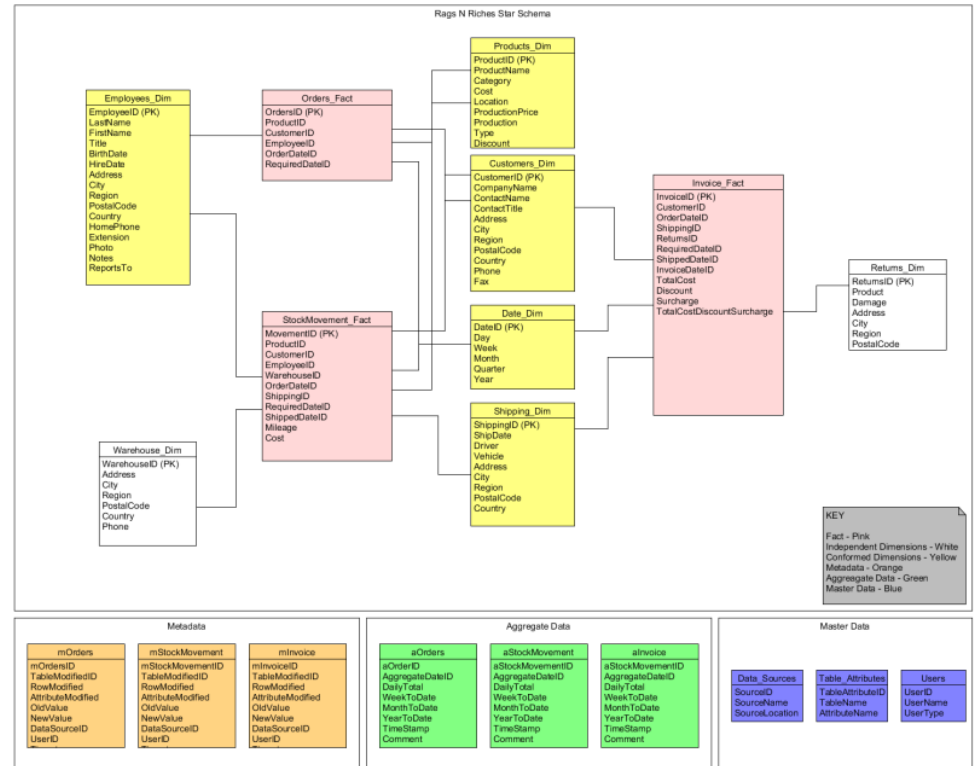
2.1 Internal and External Sources

There are several key data sources mentioned in the Rags'n'Riches case study. All of which should be included in the new data warehouse design. From an initial investigation the following data sources have been identified:

Source	Description	Internal/External
Employees	Employee information. Including job role, time with company etc.	Internal
Company Locations	Company Locations information. Including warehouses and offices	Internal
Products	All product information. Including Category, costs, productions, discounts	Internal
Orders	All order details.	Internal
Customers	Customer information. Including names and addresses etc.	Internal
Shipping	Shipping information. Including driver and vehicle information to track mileage	Internal
Returns	Return information to track damage checks via wifi/4G/5G link	External

2.2 Appropriate Star Schema

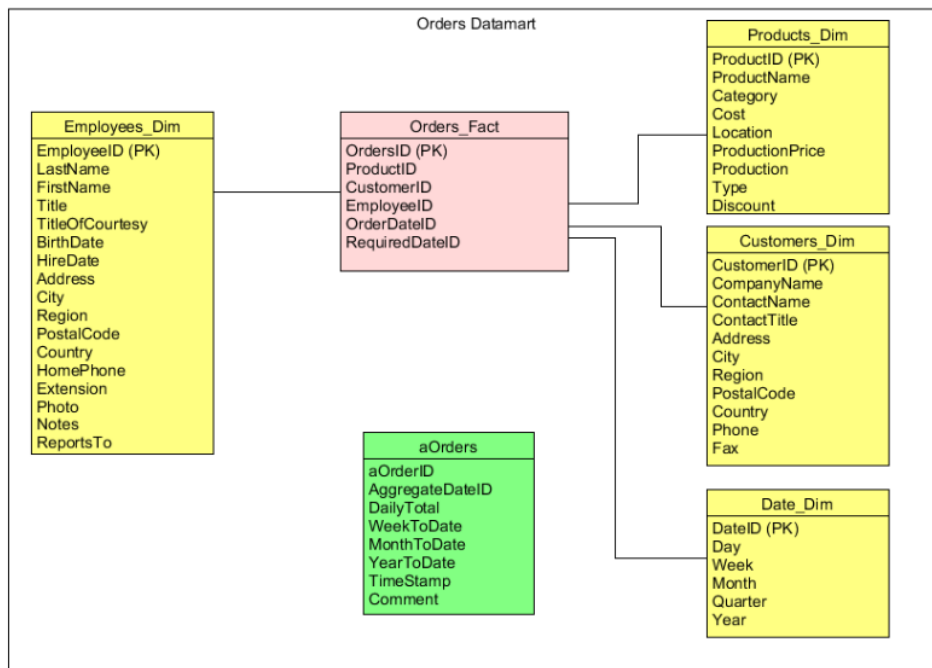
The figure below illustrates an appropriate star schema that has been developed for the Rags'n'Riches Datawarehouse. It consists of multiple stars with three fact tables, Orders, Stock Movement, and Invoices. It also consists of several conformed and independent dimension tables. It has been designed in this way to make it efficient to access the different information needed across the business. The schema also includes non-dimensional metadata, master data, and aggregate data. These tables can be structured as a set of normalised tables.



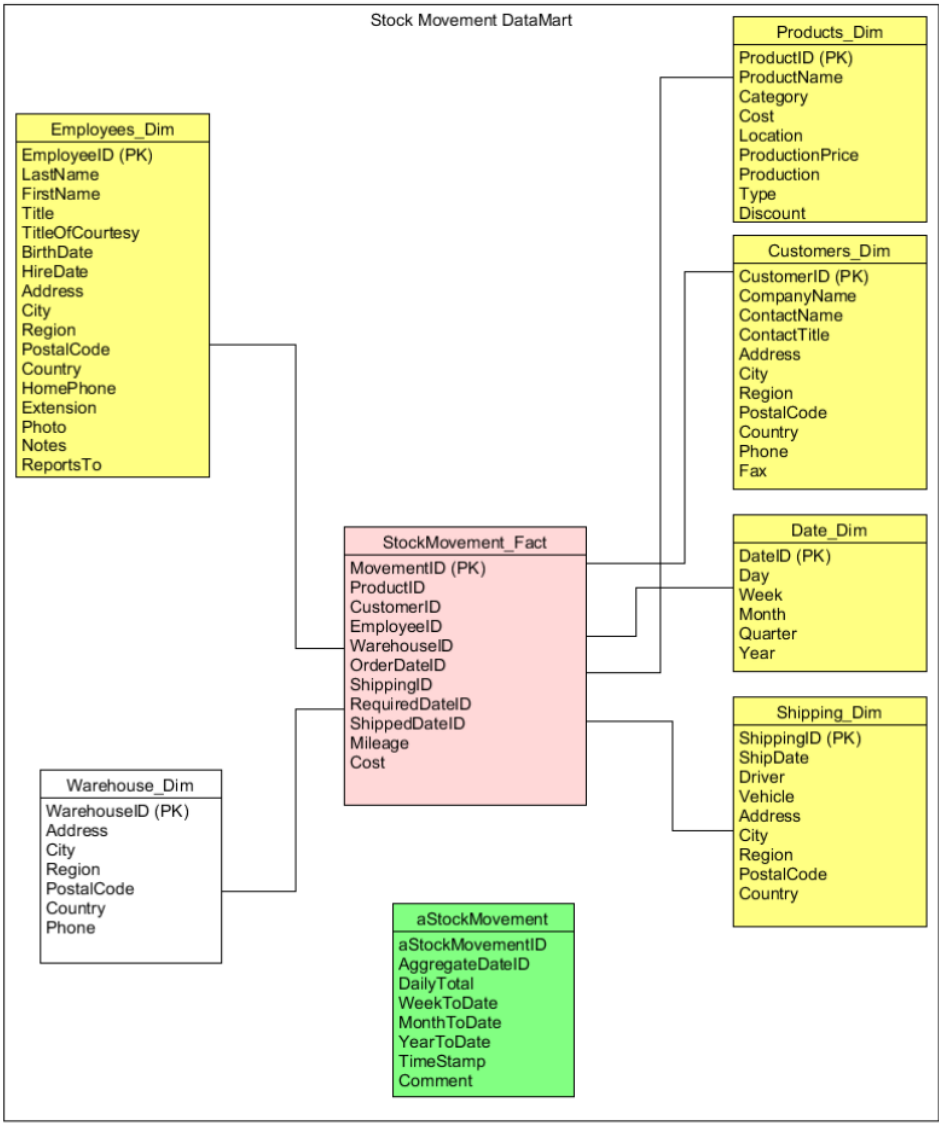
2.3 Data Marts

As well as creating a data warehouse, for efficient business line specific access to data, DataMart's should be implemented. In this case the use of dependant logical data marts would be encouraged. The design of the star schema in section 2.2 was focussed on the providing the relevant information in an efficient manner and therefore the fact tables have been created to provide the information needed to each team within the company. Three data marts will be created based around the fact tables in the star schema above.

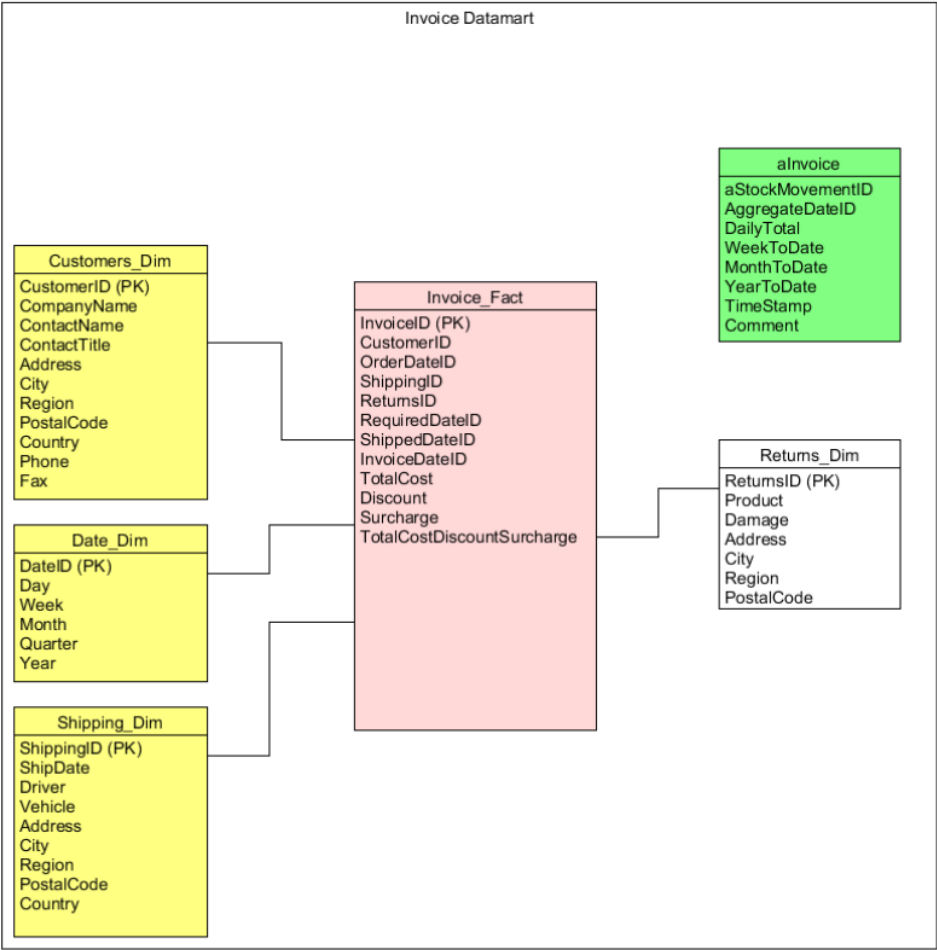
The first data mart would be based around the orders fact table to provide information to the team responsible for managing orders within the business. The schema is illustrated in the figure below:



The second data mart would be for the operational side of the business, based around the stock movement fact table. The schema is illustrated in the figure below:



The third data mart would be for the finance section of the business, based around the invoice's fact table. The schema is illustrated in the figure below:



2.4 ETL Plan

This section will describe a simplified ETL plan to show where the data for each table in the data warehouse will be sourced. The schema for the data warehouse is shown in section 2.2. The three fact tables are the tables that are likely to have more issue during the ETL process as they contain data from several different tables.

Data source to data warehouse table mapping:

Data Warehouse Table	Source Table
Employee_Dim	DB.Employees
Warehouse_Dim	DB.Locations
Products_Dim	DB.Products
Customers_Dim	DB.Customers
Date_Dim	Needs to be created
Shipping_Dim	DB.Shipping
Returns_Dim	DB>Returns – External source
Orders_Fact	DB.Employees; DB.Products; DB.Customers; DB.Orders
StockMovement_Fact	DB.Employees; DB.Products; DB.Customers; DB.Shipping; DB.Locations
Invoice_Fact	DB.Employees; DB.Customers; DB.Shipping; DB>Returns; DB.Orders

Orders Fact table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
OrdersID (PK)	DB.Orders	None – Primary Key	Shouldn't be missing	Refer to staff
ProductID	DB.Products	Link to Product_Dim	Set to 'unknown'	Refer to staff
CustomerID	DB.Customers	Link to Customer_Dim	Set to 'unknown'	Refer to staff
EmployeeID	DB.Employee	Link to Employee_Dim	Set to 'unknown'	Refer to staff
OrderDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff
RequiredDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff

Stock Movement Fact table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
MovementID	DB.Shipping	None – Primary Key	Shouldn't be missing	Refer to staff
ProductID	DB.Products	Link to Product_Dim	Shouldn't be missing	Refer to staff
CustomerID	DB.Customers	Link to Customer_Dim	Set to 'unknown'	Refer to staff
EmployeeID	DB.Employee	Link to Employee_Dim	Set to 'unknown'	Refer to staff

OrderDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff
RequiredDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff
ShippingDateID	DB.Shipping	Link to Date_Dim	Set to 'unknown'	Refer to staff
Mileage	DB.Shipping	None	Set to 0	Refer to staff
Cost	DB.Shipping	None	Set to 0	Refer to staff

Invoice Fact table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
InvoiceID	DB.Orders	None – Primary Key	Shouldn't be missing	Refer to staff
CustomerID	DB.Customers	Link to Customer_Dim	Set to 'unknown'	Refer to staff
EmployeeID	DB.Employee	Link to Employee_Dim	Set to 'unknown'	Refer to staff
OrderDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff
RequiredDateID	DB.Orders	Link to Date_Dim	Set to 'unknown'	Refer to staff
ShippingDateID	DB.Shipping	Link to Date_Dim	Set to 'unknown'	Refer to staff
InvoiceDateID	Generated	If TotalCost > 10000 Then OrderDate + 7 Days Else OrderDate	Set to 'unknown'	Refer to staff
TotalCost	DB.Orders	Sum Cost Group By Order ID	Set to 0	Refer to staff
Discount	DB.Orders		Set to 0	Refer to staff
Surcharge	DB>Returns	Link to Returns_Dim	Set to 0	Refer to staff
TotalCostDiscountSurcharge	Calculated	Total Cost – Discount + Surcharge	Set to 0	Refer to staff

Date Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
DateID	Generated	None – Primary key	Shouldn't be missing	Generated
Day	Generated	None	Generated	Generated
Week	Generated	None	Generated	Generated
Month	Generated	None	Generated	Generated
Quarter	Generated	None	Generated	Generated
Year	Generated	None	Generated	Generated

Employee Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
EmployeeID	DB.Employees	None – Primary key	Shouldn't be missing	Refer to staff
LastName	DB.Employees	None	Refer to staff	Refer to staff
FirstName	DB.Employees	None	Refer to staff	Refer to staff
Title	DB.Employees	None	Refer to staff	Refer to staff
BirthDate	DB.Employees	None	Refer to staff	Refer to staff
HireDate	DB.Employees	None	Refer to staff	Refer to staff
Address	DB.Employees	None	Refer to staff	Refer to staff
City	DB.Employees	None	Refer to staff	Refer to staff
Region	DB.Employees	None	Refer to staff	Refer to staff
PostalCode	DB.Employees	None	Refer to staff	Refer to staff
Country	DB.Employees	None	Refer to staff	Refer to staff
HomePhone	DB.Employees	None	Refer to staff	Refer to staff
Extension	DB.Employees	None	Refer to staff	Refer to staff
Photo	DB.Employees	None	Refer to staff	Refer to staff
Notes	DB.Employees	None	Refer to staff	Refer to staff
ReportsTo	DB.Employees	None	Refer to staff	Refer to staff

Warehouse Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
WarehouseID	DB.Locations	None – Primary key	Shouldn't be missing	Generated
Address	DB.Locations	None	Refer to staff	Refer to staff
City	DB.Locations	None	Refer to staff	Refer to staff
Region	DB.Locations	None	Refer to staff	Refer to staff
PostalCode	DB.Locations	None	Refer to staff	Refer to staff
Country	DB.Locations	None	Refer to staff	Refer to staff
Phone	DB.Locations	None	Refer to staff	Refer to staff

Shipping Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
ShippingID	DB.Shipping	None – Primary key	Shouldn't be missing	Generated
ShipDate	DB.Shipping	None	Refer to staff	Refer to staff
Driver	DB.Shipping	None	Refer to staff	Refer to staff
Vehicle	DB.Shipping	None	Refer to staff	Refer to staff
Address	DB.Shipping	None	Refer to staff	Refer to staff
City	DB.Shipping	None	Refer to staff	Refer to staff
Region	DB.Shipping	None	Refer to staff	Refer to staff
PostalCode	DB.Shipping	None	Refer to staff	Refer to staff
Country	DB.Shipping	None	Refer to staff	Refer to staff

Customers Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
CustomerID	DB.Cutomer	None – Primary key	Shouldn't be missing	Generated
CompanyName	DB.Cutomer	None	Refer to staff	Refer to staff
ContactName	DB.Cutomer	None	Refer to staff	Refer to staff
ContactTitle	DB.Cutomer	None	Refer to staff	Refer to staff
Address	DB.Cutomer	None	Refer to staff	Refer to staff
City	DB.Cutomer	None	Refer to staff	Refer to staff
Region	DB.Cutomer	None	Refer to staff	Refer to staff
PostalCode	DB.Cutomer	None	Refer to staff	Refer to staff
Country	DB.Cutomer	None	Refer to staff	Refer to staff
Phone	DB.Cutomer	None	Refer to staff	Refer to staff
Fax	DB.Cutomer	None	Refer to staff	Refer to staff

Products Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
ProductID	DB.Product	None – Primary key	Shouldn't be missing	Generated
ProductName	DB.Product	None	Refer to staff	Refer to staff
Category	DB.Product	None	Refer to staff	Refer to staff
Cost	DB.Product	None	Refer to staff	Refer to staff
Location	DB.Product	None	Refer to staff	Refer to staff
ProductionPrice	DB.Product	None	Refer to staff	Refer to staff
Production	DB.Product	None	Refer to staff	Refer to staff
Type	DB.Product	None	Refer to staff	Refer to staff
Discount	DB.Product	None	Refer to staff	Refer to staff

Returns Dimension table:

Target Attribute	Source	Processing	Missing Value	Incorrect Value
ReturnsID	DB.Product	None – Primary key	Shouldn't be missing	Generated
Product	DB.Product	None	Refer to staff	Refer to staff
Damage	DB.Product	None	Refer to staff	Refer to staff
Address	DB.Product	None	Refer to staff	Refer to staff
City	DB.Product	None	Refer to staff	Refer to staff
Region	DB.Product	None	Refer to staff	Refer to staff
PostalCode	DB.Product	None	Refer to staff	Refer to staff

Conclusion

This report has firstly provided an overview of why Rags'n'Riches as a business will benefit from the implementation of a data warehouse. The difference from their current OLTP system and the potential benefits of a data warehouse have been explained as well as providing information on the architecture and techniques adopted when designing a data warehouse. Secondly a potential design

has been developed from the information provided about the business. This suggests potential data sources, a star schema, data marts specific to the business needs and an ETL plan.

References

- [1] T. Hammergren and A. Simon, "Data warehousing for dummies," p. 366, 2009.
- [2] "What Is a Data Warehouse | Oracle United Kingdom." [Online]. Available: <https://www.oracle.com/uk/database/what-is-a-data-warehouse/>.
- [3] "What Is a Data Warehouse? | Definition, Components, Architecture | SAP Insights." [Online]. Available: <https://www.sap.com/insights/what-is-a-data-warehouse.html>.
- [4] "What is ETL? - Databricks." [Online]. Available: <https://databricks.com/glossary/extract-transform-load>.
- [5] "What is a Data Mart | IBM." [Online]. Available: <https://www.ibm.com/cloud/learn/data-mart>.
- [6] "Database normalization description - Office | Microsoft Docs." [Online]. Available: <https://docs.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description>.

Bibliography

- [1] "Data Warehouse Architecture: Traditional vs. Cloud | Panoply." [Online]. Available: <https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/>.
- [2] "Dimensional Modeling Techniques - Kimball Group." [Online]. Available: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/>.
- [3] R. Kimball and M. Ross, "The Kimball Group Reader," *Kimball Gr. Read.*, Dec. 2015.
- [4] R. Kimball, M. Ross, and W. Thornthwaite, "Data Warehouse Lifecycle Toolkit (2nd Edition)," 2011.

FINAL GRADE

66/100

GENERAL COMMENTS

Instructor

Anton Hall
1 engagement with literature comment available
Limited sources with only occasional in body citations
Make a citation to reenforce each new fact and use a range of reference types, websources are okay but back these up with more academic journals 20
Heavy use of direct quotes - try to paraphrase and referenc where ever possible
2 Knowledge and Understanding comment available
sound knowledge displayed and mapped to datawarehousing concepts demonstrated and related back to the brief, however, some elements could have benefitted from direct discussion rather than simply using large direct quotes 20
3. Cognitive and Intellectual skills comment available
little discussion on the source systems - the design seems to work under the assumption that the organisation works with one internal RDMS for transactional processing (which would be the ideal scenario), but little consideration is given to other internal systems and external sources 10
4. Practical Application Skills comment available
well laid out star schema design and etl process 40
5. Transferable Skills for life and Professional pract comment available
Technical terminology boken down into non expert language 10
Large amount of similarity generated due

to direct quotes and tabululated data- no
action taken on this owing to the nature of
the similarity

No Clear introduction

caption every image and include in table
of figures

Conclusion brief

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
