

1004334_Assignment_1

by Adam Moore

Submission date: 25-Mar-2022 01:59PM (UTC+0000)

Submission ID: 174561776

File name: 12919_Adam_Moore_1004334_Assignment_1_398298_1348875654.docx (342.88K)

Word count: 3669

Character count: 19159

LEVEL 6 ASSESSMENT SPECIFICATION

Student name:	Adam David Moore	SN:	1004334
Programme:	Data and Information Systems BSc		
Module:	Data Warehousing and Data Mining		
Module code:	ACCA6026	Contribution to Overall Module Assessment (%):	50%.
Lecturer:	Dr Nik Whitehead	Internal Verifier:	Bev Holland
Assignment Title:	Assignment 2: Rags'n'Riches	Word count (or equivalent):	2853

Table of Contents

Task 1 3

 Part 1..... 3

 Part 2..... 4

 Part 3..... 6

 Part 4..... 7

Task 2 9

 Part 1..... 9

 Part 2..... 10

 Part 3..... 12

 Part 4..... 14

References 16

Task 1

Imagine that you are a data warehousing consultant who has been asked to provide a report to the chief executive and board of directors of the above company on the following technical issues. The CEO and the directors are not computing specialists, but have been advised that a data warehouse might be suitable for their organisation.

Part 1

Discuss the difference between an online transaction processing (OLTP) database and a data warehouse. The company already has a working OLTP system and the CEO needs to learn the benefits that investing in a data warehouse system will bring to their company. (10 marks)

Both OLTP databases and data warehouses are a type of database. While an OLTP database is essential for maintaining operations within your business, and captures all business data like sales/orders, customer data, stock/products and so on, a data warehouse is purpose-built for information reporting and analytical applications [1].

The OLTP database is very much a live/active database in that, it will likely be changed to update prices, add or remove product types, capture sales/order and customer data and so on. A data warehouse will be populated via ETL processes directly from the OLTP database; this has the advantage that, anything removed from the OLTP for operational reasons (certain products removed, for example) does not need to be removed from the data warehouse. This results in intact data archives and enables accurate, reliable and complete information reporting. An OLTP system conventionally has a high frequency of small/fast online transactions (reading records, inserting records, updating records and deleting records) [1] [2].

Raw data from source systems utilised by the business can be stored in a data warehouse via extract, transform and load (ETL) processes. This data, stored in the data warehouse, is then optimised for analytics and reporting as it is structured and clean, perhaps unlike its raw format in the source systems [3]. Users within the business, such as analysts, information reporters, data scientists, data engineers and more can then easily leverage this structured, clean data that has been optimised for analytics to automate information reporting and provide business insights through analytics such as machine learning or predictive analytics.

As the warehouse data is clean and structured, it can be considered a 'single source of truth' for business data [4]. As users do not have to perform stringent ETL processes themselves, data will be far more aligned which minimises issues/errors in automation, differences in metrics and measures and other generic and common business information issues that arise because of unstructured ETL processes.

Part 2

Explain, with the aid of appropriate diagrams, the architecture of a data warehouse and its source systems. Show how the different internal and external sources of data can be used to populate a data warehouse. The CEO has heard the term 'ETL' but does not understand it, so they will need to be convinced that this will not affect the data-to-day operations of the company. (20 marks)

The figure drawn up below demonstrates the process flow for data in the proposed data warehouse. Data from various sources can be loaded into the data warehouse and conform to the dimensional database design, if they go through an appropriate and aligned extract, transform and load (ETL) process. This ETL process is designed to ensure the integrity of the data is upheld and the resultant outputs uploaded to the warehouse ensure that the needs of the business are met [3].

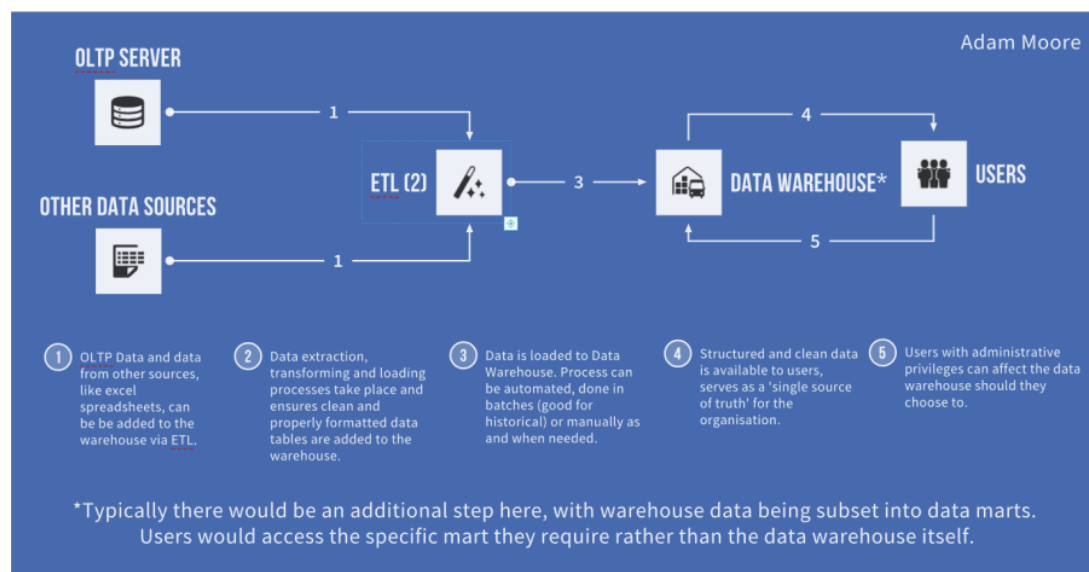


Figure 1. Demonstration of data flow to data warehouse and end-users from source systems.

By going through this ETL process, all data is cleaned and aligned into such a format that it can be combined or otherwise function effectively in a dimensional database.

The ETL process handles structured or unstructured data from any number of sources and extracts it into a 'staging' area. The data is then transformed, which can involve 'cleaning', re-structuring, validating, de-duplicating and many more functions that transforms the source data in some way. This ensures alignment and improves the quality of the data [5]. This transformed data is the loaded into the data warehouse where it becomes useable. Loading can be handled according to business needs. For example, batch loading can take place out of office hours to ensure there is no adverse effect or added strain on the network and/or systems.

Developing an effective, reliable, transparent and robust ETL process requires engagement from many stakeholders. Domain experts need to guide ETL development to ensure the resulting series of

processes are effective, sustainable and result in accurate information being conveyed to the business.

Dependant on how decision makers in the business currently receive information reports, it is highly likely that the information being conveyed to them is undergoing some form of extraction and transformation. Those in the organisation that provide business intelligence/information reporting will likely be taking data from sources (Extracting), applying some transformations to make it useable to produce business intelligence/information reports.

The major difference with a structured and integrated ETL process is that it is transparent and automated. Through transparent discussion and effective communication, domain experts build the rules and processes that govern this ETL process, and the result is a single source of truth [4] (Data Warehouse) that has the same transformations applied to it [6]. There is less burden on analysts or those reporting to clean and transform any data they use, which is a positive factor in terms of productivity. This should be a benefit to the organisation, in that it gives decision makers more confidence in the validity and robustness of information conveyed to them.

Part 3

The managing director, finance director, and the operations director have their own specific needs in terms of seeing relevant data to allow them to make well-founded management decisions. Explain how data marts may be used to provide targeted information for these managers and their departments. (10 marks)

Data marts are effectively a subset of a data warehouse; these are efficient in a business environment because each data mart handles transactions from its users, rather than having a singular data source which is used by all [7]. The reduced traffic for each server has positive performance implications for the business and can improve load times and ensure a stable warehouse environment.

Data marts are also responsible way to also ensure, in combination with an effective authentication and permissions system, that confidential information or unnecessary data is not visible to or used by any users that do not need it. This minimisation improves performance, efficiency, and helps to prevent any GDPR breaches.

There is also no downside in terms of the ETL process and getting the information needed to make effective business decisions; whatever is required to make those decisions will be present in the data mart. In terms of user experience, a reduced number of tables – all of which are highly relevant, lends itself to effective analytics.

A data mart is a curated environment for a specific community/department of individuals [8]. This aids developers who may otherwise be forced to wade through large amounts of irrelevant tables and redundant data, development time increases, and other negative implications can occur such as mis-identifying tables and so on. Creating data marts is like taking a subset of the data warehouse which is specific and relevant to the needs of the use case; for example, in this instance, a data mart would be created for the managing director, finance director and operations director, which contains curated data that is highly relevant and needed to meet their information/analytics needs and support effective decision making.

Part 4

The IT director knows a lot about OLTP systems but little about data warehousing. Explain to them the difference between the 'normalised' structure of an OLTP system and the 'dimensional' structure of a data warehouse. Use a suitable diagram to explain how facts and dimensions can be used to develop a suitable star schema for their system.

N.b. As this is a management report you do not need to develop the entire schema at this stage, but should instead use examples related to the company's data to illustrate the concepts involved. (10 marks)

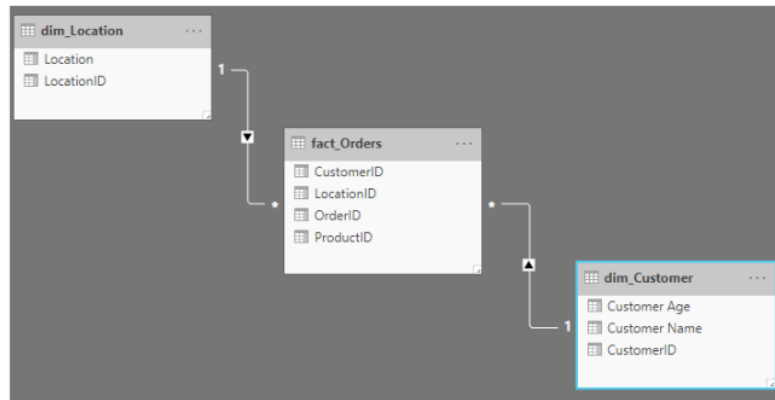
In the normalised structure of an OLTP system, data is generally stored based on the rules of database normalisation. Tables are categorised or banded into groups that are relevant to a particular area, such as finance or products. These tables are then linked by joins, creating a relational database. Querying languages such as SQL are used within database management systems (i.e. Microsoft SQL Server) are leveraged to join tables following the normalised rules and relational database design via keys or IDs [9]. The result is many tables that contain relatively high amounts of data.

This type of database is easier to implement, as there is no need to design and carefully think through a schema for implementation with defined linkages. It is also, however, considered more cumbersome and difficult to use. Users that did not take part in building the relational database may lack the expertise and understanding needed to properly interpret and leverage the data in an optimal way, unless the naming conventions and the data is very clear and easy to interpret.

Dimensional approaches split tables into two classes: fact tables, or dimension tables. Fact tables are usually transactional data with a high frequency, these may be customer orders or invoices issued for example. Dimension tables are contextual tables that are linked to fact tables to provide more information. The relationships are explicit and so less expertise is required to understand these concepts in practice than with normalised databases [10]. Fact tables will generally contain several ID or Key columns, such as a product ID or a customer ID, instead of information about the product or customer. This minimisation improves efficiency and helps to optimise use of the database. The equivalent dimension table can be linked and leveraged to provide context around that ID – such as the customer's name or date of birth, with regards to the customer ID (if such fields are held in the database). This is considered optimal for analytical operations as retrieval is quicker and more efficient, but also the context and relationships are better defined and clearer, which supports effective and proper use of the data.

The figure below (non-exhaustive, limited fields for demonstrative purposes) demonstrates a very simple data model comprising of one fact table (fact_Orders), and two dimension tables (dim_Location and dim_Customer). Relationships are formed via the use of IDs (or keys). In this case, the contextual dimension table 'dim_Location' would contain all contextual information related to each location; it is joined to the 'fact_Orders' table via the LocationID key. Similarly, the 'dim_Customer' dimension table is joined to the 'fact_Orders' table by the CustomerID key.

All contextual data relating to both the location and the customer could be added to the orders table via these relationships but enables minimisation and easier management of tables.



A star scheme can further be developed by expanding on this concept and introducing more dimension tables such as a product table, and other relevant tables associated with the business.

Task 2

The CEO and the board have been convinced that a data warehouse is the way forward for their company and you have been employed as a contractor to support the company in its development of the new data warehouse solution. You should now produce a report that develops and describes a suitable star schema for the company's new data warehouse. This should include the following elements:

Part 1

Identification of internal and external source systems (5 marks)

1. There is an orders database. Staff manually enter orders onto this database as and when orders come in; this is done before the staff member works out how to fulfil the order.
2. There is an internal stock management system, which details the dates any given item of stock is available/unavailable and the whereabouts of the stock.
3. There is a delivery database which is used to raise delivery orders. Client agreement of order fulfilment is also stored on this database.
4. There is a transport system that is used to raise and manage orders for collection and movement of stock between warehouses.
5. 'Company database system', staff record inventory checks and potential damage on a tablet and upload it to the company database system. Metadata (and possibly normal data) is stored regarding the staff recording the inventory checks.
6. Invoice system to raise invoices relating to orders, in the relevant format (based on conditions such as total order value) and discounts applied as appropriate.
7. Fuel card data/distance covered by each vehicle is recorded. This could be its own data source, as it may be mileage logs or fuel receipts, or it could be recorded or calculated from existing systems (measures could be used to calculate distance based on journeys made from delivery/transport systems or to wrangle the fuel cost data).

There appears to be at least **five** but potentially as many as **eight** source systems.

If delivery and transport are separate, plus fuel card data is an individual source and mileage is from a different source, there *could* be up to eight sources; however, there are at least **five**.

- Orders Database
- Stock Management System/Database
- Delivery/Transport Database
- Company System/Database
- Invoice System/Database

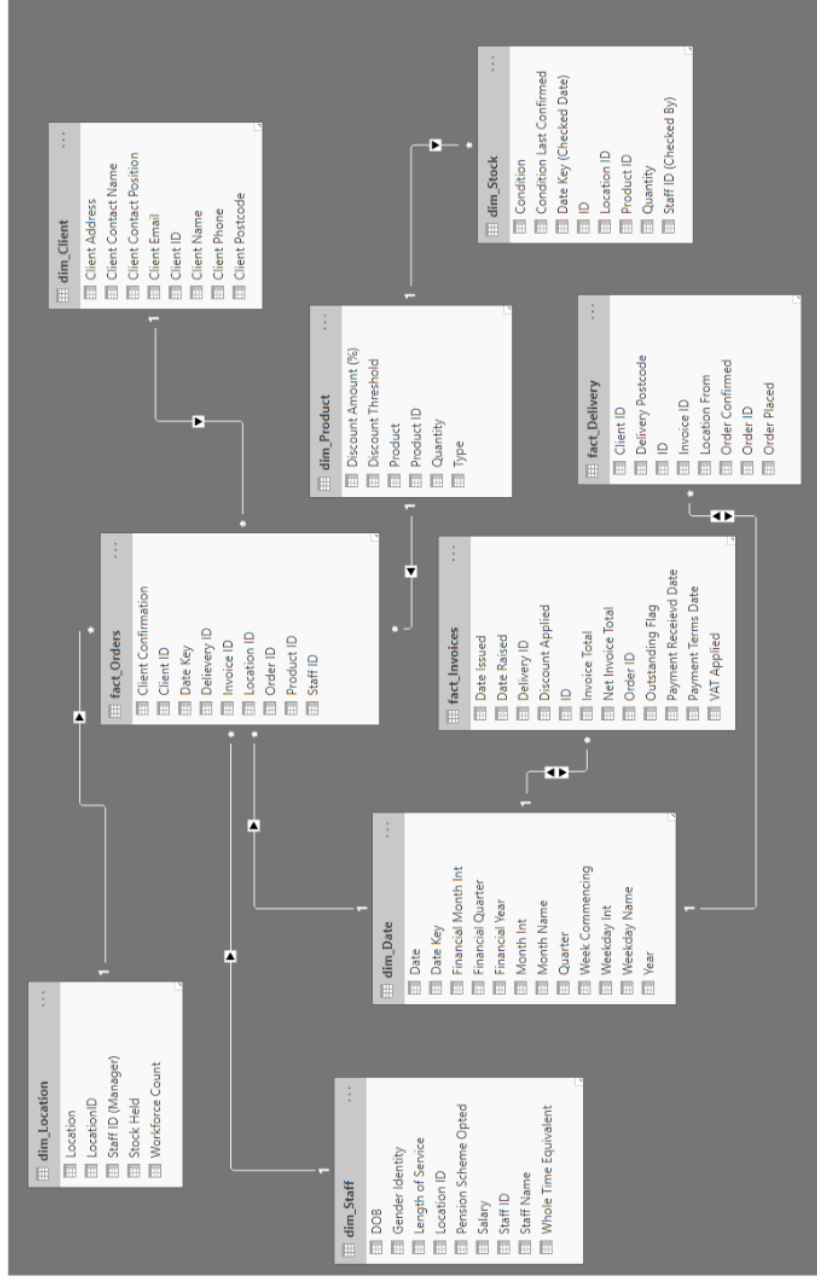
It is not absolutely clear (and further clarification would be requested from the company), but it appears as though there are two separate systems for delivery/transport.

The delivery system handles orders to clients and logs relevant dates. It also records clients' confirmation that all stock has been received in good condition. The delivery system interfaces with the orders system. The transport system is an internal system that is used by Rags'n'Riches staff to log condition of stock as it is transported to Rags'n'Riches locations; be that between locations or returning to locations following rental.

Part 2

An appropriate star schema (which may include multiple stars) for the system which includes and clearly shows tables for master data, metadata and aggregate data. (25 marks)

Rather than rely on multiple stars and joins; a multi-fact star schema is a better way to integrate multiple fact tables.



A multi-fact star schema provides a data model that allows full integration of tables. Metadata shows who processed each order, which staff member and at which location, as well as condition checks by staff members for each unit of stock. Aggregate measures can be used to provide aggregate data to relevant stakeholders. Some dimension tables are master data tables. Date Keys (integer dates) are used instead of linkage on date data types; this is more efficient in database and reporting as there is no need for the conversion to take place within the system.

Each individual item has a unit/stock ID and all information relating to each individual item is recorded in dim_Stock, each type of product has a product ID which aggregates the number of these products that are available (aggregates dim_Stock data). Stock is held at locations, so each unit (regardless of product type) is recorded against its current location to ensure effective stock management. This will help in raising delivery orders for moving stock between locations to fulfil orders. Users can check at an aggregate level, the levels of stock and where it is.

Similarly for orders and deliveries, these are highly transactional processes. Orders should be raised as a result of external clients requesting services, whereas deliveries can be both internal and external. Units are often moved between locations to meet the needs of client's orders.

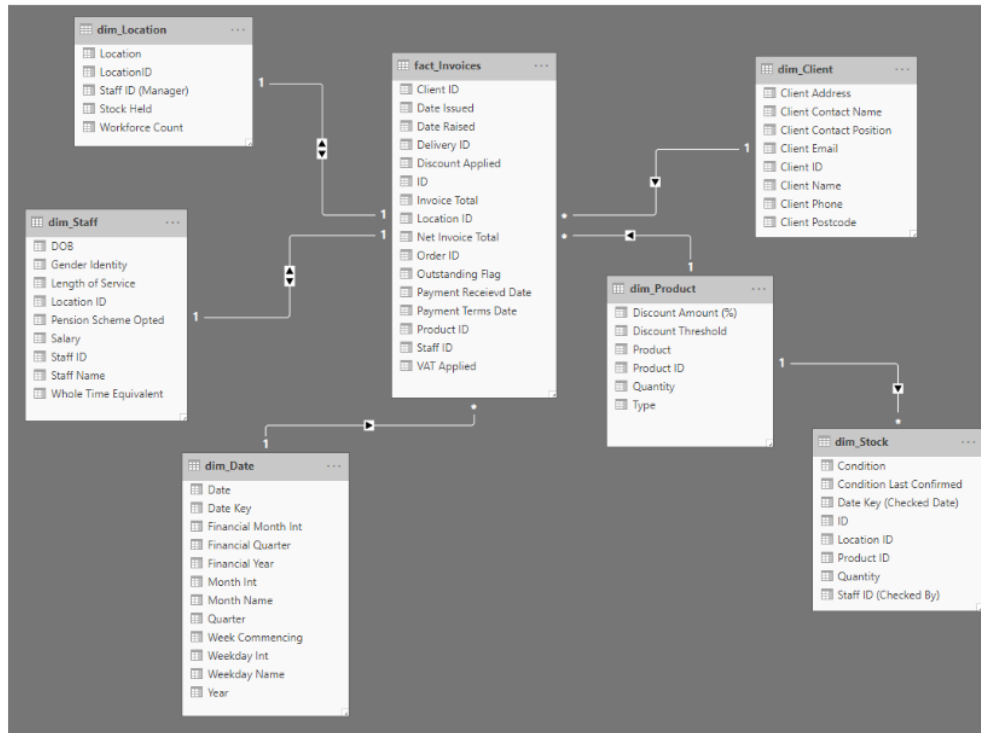
Date dimensions with an array of 'normal' and 'financial' means of presenting data on a time-series have been provided to align with both calendar and financial years. Financial quarters and years are offset due to the financial calendar running from April to March rather than January to December. This will be useful for tax purposes and should align with the needs of the financial department and operational department.

Part 3

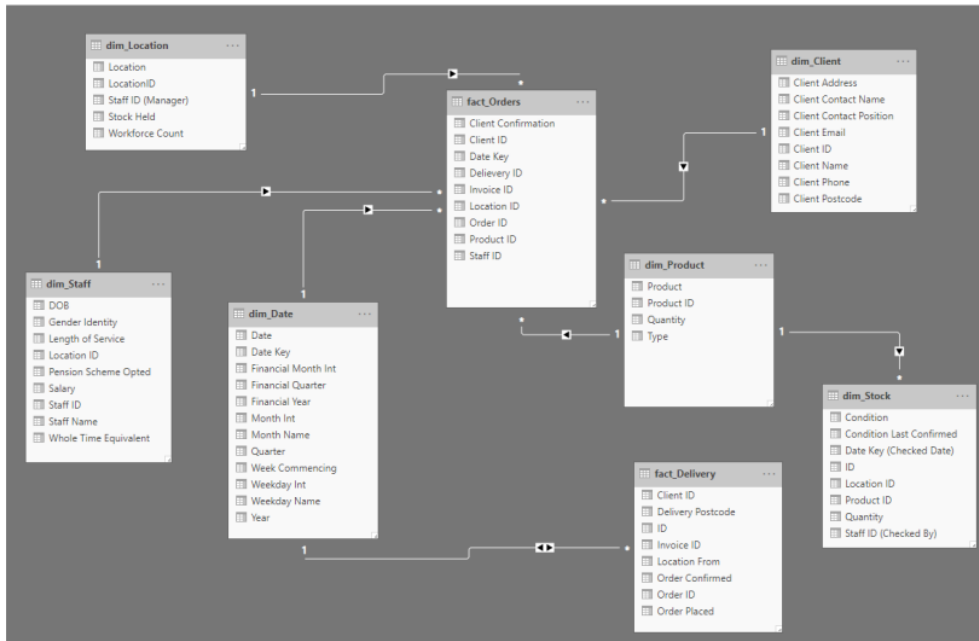
Identification of appropriate data marts, and the schema for data that would be extracted to each data mart (10 marks)

Three data marts will be needed to optimise operations.

The following schema will serve as the data mart for the finance department:



The following schema will serve as the data mart for operations:



An additional data mart with a focus on aggregates will be used for the managing directors' team and information needs also.

Fundamentally, to ensure an optimal process, redundant data for the purposes of each department should not be transported into the data marts. Meta data should only be included where it is relevant to the functional roles of staff utilising data. Reporting staff/operational staff should only see what they need to perform their roles, but it should also be in a well-structured format for their use.

Part 4

A simplified ETL plan that describes where the data for each table will be sourced and what, if any, transformations need to be applied to that data before it can be loaded into the data warehouse.
(10 marks) All design decisions should be explained and justified

dim_Location

Will be sourced from the company system/database. This set of locations will be maintained within this system; as part of the ETL process, it is likely that an indexing system will need to be used in the transform section and IDs assigned to all locations; preferably in chronological order in which they were acquired by the business. Stock Held is an aggregate measure that needs to be a calculated field based on the dim_Stock table. Workforce Count is an aggregate measure that needs to be a calculated field based on the dim_Staff table.

dim_Staff

Will be sourced from the company system/database. This set of locations will be maintained within this system; as part of the ETL process, it is likely that an indexing system will need to be used in the transform section and IDs assigned to all staff members; preferably in chronological order of appointment.

dim_Date

This is a generic and logical table that can be generated as a one-off and uploaded to the warehouse. Having dates until 2040 and from the earliest record (Company established 1923).

Additional future dates may need to be added as a batch upload in future, but it is likely the company will undergo changes to its databases in the next 15 years.

dim_Product

This dimension table will be population via the stock management system, it can be considered an aggregate table based on the stock table, whereby each type of product (differentiated by ID) is held in a distinct table, along with aggregate measures such as sum quantity.

dim_Stock

This dimension table will be populated via the stock management system. Each individual item will have a stock ID / unit ID that will allow differentiation between individual items.

dim_Client

This dimension table will be populated via the orders system, each client will have to register details onto said system (possibly via an internal staff member who is taking the order).

The client dimension table will be a distinct table of clients and an ID will be issued in chronological order of their engagement with rags'n'riches.

fact_Orders

This fact table will be populated via the orders system and is essentially a complete set of orders placed with rags'n'riches. The order number should be useable as an order ID.

fact_Invoices

Invoice fact table will be populated via the invoice system/database. The Invoice number should be useable as an invoice ID.

Care will need to be taken with currency transforms, many systems default to USD as a currency, so care will need to be taken to ensure this is GBP and functions correctly in this case as rags'n'riches is a UK trading company and trades in GBP.

fact_Delivery

This will be populated via the delivery/transport database. Both internal movements of stock between locations and deliveries to clients will be recorded and geospatial data like postcode will be included to estimate delivery costs ahead of time or for information reporting purposes.

References

- [1] R. S. e. al, "Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries," *nternational Journal on Computer Science and Engineering*, vol. 2, no. 9, pp. 2865-2873, 2010.
- [2] R. R. a. J. Gehrke, Database management systems, New York: McGraw-Hill, 2003.
- [3] S. S. e. al, "Nuts and Bolts of ETL in Data Warehouse," *Emerging Trends in Expert Applications and Security*, pp. 1-9, 2019.
- [4] G. a. D. H. Shan, "Research on Business Intelligence Application in Electric Grid Industry Based on Data Mining," in *2014 7th International Conference on Intelligent Computation Technology and Automation*, 2014.
- [5] R. a. P. K. M., "A comparative review of data warehousing ETL tools with new trends and industry insight," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, 2017.
- [6] S. A. e. al, "Optimizing analytic data flows for multiple execution engines," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012.
- [7] P. Pipe, "The data mart: A new approach to data warehousing," *International Review of Law, Computers & Technology*, vol. 11, no. 2, pp. 251-262, 1997.
- [8] C. Ghezzi, "Designing data marts for data warehouses," *CM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 10, no. 4, pp. 452-483, 2001.
- [9] S. C. a. U. Dayal, "An overview of data warehousing and OLAP technology," *ACM Sigmod record*, vol. 26, no. 1, pp. 65-74, 1997.
- [10] R. K. a. M. Ross, The data warehouse toolkit: the complete guide to dimensional modeling, John Wiley & Sons, 2011.

FINAL GRADE

48/100

GENERAL COMMENTS

Instructor

Adam David Moore

1 engagement with literature
comment available

good range of academic sources used and
cited throughout task one, task two would
have benefited with some citations to
justify your choices 20

2 Knowledge and Understanding
comment available

knowledge demonstated in all areas,
however, discussion is rather brief and in
many cases is not related back to the case
study 20

3. Cognitive and Intellectual skills
comment available

Source sysetems are correctly identified
but are not really discussed in any great
detail, no discussion on user groups or
who would want what from the system 10

4. Practical Application Skills
comment available

design is flawed as you have snowflake
schemas in operation (while these are
legitimate they will slow sown the
peoduction of any query results). Also,
your Delivery fact only seems to have one
dimention? 40

5. Transferable Skills for life and Professional pract
comment available

Avoid putting elements of the brief in your
final submission, this does not indicate
understanding, increases your word count
unnecessarily and gives a high similarity
rating on turn-it-in (if you must use
elements from the brief insert them as a
captioned image rather than text) 10

neutral language used troughout -
technical terms are broken down

structure needs work, no clear intro or
conclusion

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