

Unfinished Report

by Mark Vaughan

Submission date: 18-Mar-2022 01:57PM (UTC+0000)

Submission ID: 173862827

File name: 19487_Mark_Vaughan_Unfinished_Report_398298_2041779824.docx (667.2K)

Word count: 6938

Character count: 36053

a) **Network Operating Systems**
Assessment
Paper Number Four

Mark L. E. Vaughan

**Undergraduate, University of Wales Trinity St David Swansea Wales*
1909160@student.uwtsd.ac.uk

1. Contents

Report Heading Area	Error! Bookmark not defined.
Task 1 – CEO Report	3
1. Benefits of Data Warehousing	3
2. Explanation of the Architecture	4
3. The Star schema	5
a. Source systems	5
4. Explain ETL	6
b. Extract	6
c. Transform	7
d. Load	8
e. Metadata	9
5. Datamart's	9
Data marts for managing financial and ops directors	9
6. IT director – explain differences between relational normalised and data Warehouse systems structure	10
-----	...Error! Bookmark not defined.
Task 2 – The Rags 2 Riches Data Warehouse build	12
7. Data sources	12
f. Known operational sources	12
g. Non-operational sources (presumed)	12
h. Hire dept/Clients	12
i. Warehouses	12
j. Transport	12
k. Finance – income – outgoings	12
l.	12
8. Star schema	12
m. Dims	12
9. Data marts	12

n.	Mart identification	14
o.	Data requirements	14
10.	ETL PLAN	14
p.	Table data sources	14
q.	Extraction	14
r.	Transformation requirements	14
s.	Post Processing and Staging	15
11.	Bibliography Error! Bookmark not defined.	
12.	Appendices	16
13.	Star Schema Tables	17
14.	Data Marts	20
t.		22

*Abstract text***2. Task 1 – CEO Report**

The use of relational databases in modern business is essentially a given. Their effectiveness in storing and processing business data in an efficient space optimised storage structure makes them an ideal solution. Note the use of the words storing and processing are used. However, there is an important word missing. This word is Reporting. Without proper reporting, a business will cripple itself. Without proper reports, effective decisions cannot be decided, operations within the business become inefficient as current and factually correct data is not on hand and potentially profitable avenues of expansion are left undiscovered. This is where data warehouses come to the fore.

b) Benefits of Data Warehousing

Having read the last paragraph, the reader will be considering the question: What is the benefit of having a data warehouse? The answer to that question can be broken down into four key areas. These are: -

- Improved reporting
- Data archive/backup
- Compartmentalised security
- Data Mining Capability

The prime characteristic of a data warehouse is its capability for reporting. Online Transactional Processing systems (OLTP) are structured to store transactions with minimal duplication and fast record retrieval using relational links across multiple tables. This set up is good for processing records but are slow

where the retrieval of multiple related records is concerned.

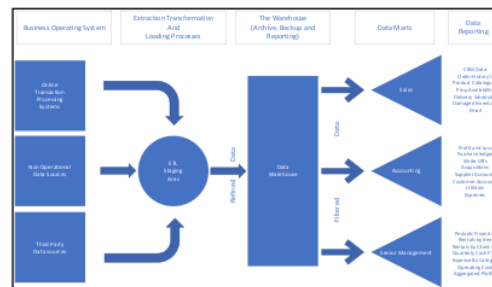


Figure 1

The structure of Data warehouses is different. The design is optimised for the collation and filtering of data from OLTP sources and more importantly, non-operational data sources. Unlike relational systems, the use of record duplication, table collation and reduced indexing allows for faster and more versatile reporting functionality that can be tailored to provide customised reports for all levels of the enterprise.

This is where the second area is concerned. The warehouse is in essence an archive of all the data that the enterprise uses or has used. This not only includes the operational data like orders and invoices. it can also include paper documents, photographs, supplier price lists and research data. It is also a record of the past activities of the enterprise forming an audit trail of past transactions and a base upon which the restoration of damaged data stores can be actualised. In this guise, the warehouse is a form of backup which can be conveniently accessed for fast restoration but, it should not be used as a substitute for other protection systems. The wisdom of multigenerational backups stored in offsite protected facilities should not only be remembered but should be leveraged alongside the warehouse to maximise data security up to and beyond ISO2700 specification.

In the area of security, the warehouse has another useful feature in the form of compartmentalisation. Collecting all the enterprises data and storing it all in one place

may appear to be a big security issue. As mentioned previously, the use of a proper backup plan and offsite backup up sites does mitigate the risk. There is one aspect that is not covered. This is concerning the breach of data by users and intruders. There are many forms of data breach, some are malicious and are caused by hackers and disgruntled employees, some are deliberate which are instigated by opportunists, the insubordinate and the careless. To allow warehouse access directly would be unwise so, the warehouse uses a feature called data marts. This feature allows data to be compartmentalised to restrict access to authorised personnel and user groups. The data is generally read only in nature and cannot be altered by design. The risk here is in confidentiality and access rights. A full explanation of data marts will be showcased in section x.

As data warehouses are primarily designed for reporting, a very useful and important investigation method can be leveraged. The method is called Data Mining. With the expert manipulation of the warehouses tables and the use of well-designed queries, information about different aspects or company information, performance and demographics can be compared and combined to discover new information hidden within the stream of data created by the varied and many facets of corporate activities. The discovery of such information can advise higher management of new opportunities, potential improvements and warn of the dangerous pitfalls that can affect any growing company or enterprise. Use of this system can be a force multiplier for growth and stability for SME's and mega corps alike.

3. Explanation of the Architecture

Data Warehouses comprise to two main systems. The first is the staging area which is used for the ETL process that will be discussed in section 4. The second area is the warehouse proper that is comprised of four forms of tables. It is this area which will be described next.

The first two forms of tables make up the control mechanism for the warehouse. They

control the guiding parameters (Master Tables) and the history of the warehouse (Metadata Tables) without which the integrity and robustness of the information system will fail.

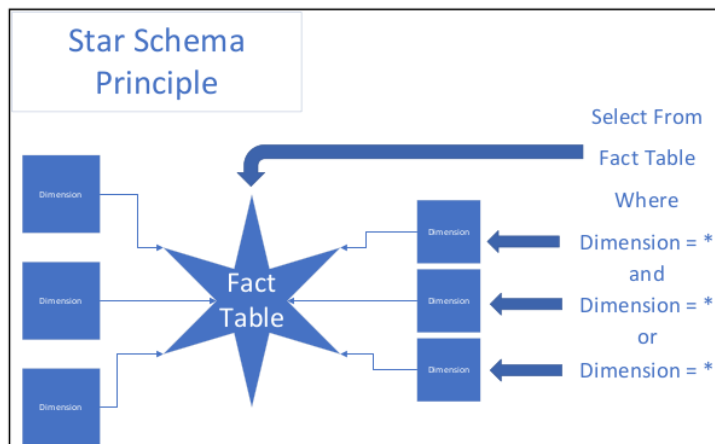
The main set of tables used to control the warehouse are master tables. These tables are a form of fact table that hold key unifying data fields that enable unambiguous identification of key data from varied sources. The advantage is that these fields allow the connection between different databases within the fact table based on the same data definition. In the ETL process within the warehouse, these master fields are defined in data dictionary's which define the data type and constraints for these fields and state the reason for the fields use and how it relates to other data values within the operating data set. In use, the master tables hold the static data fields from the source data and are usually the primary or foreign keys in relational databases. As the data rarely changes between updates, these master fields add to the stability of the warehouse and are essentially the skeleton of the data framework on which the warehouse is built.

The metadata tables are the heart of the key to the warehouse's integrity. The records the results of and the changes to the warehouse tables. This audit trail can provide the managers and staff of the warehouse with key diagnostic data to correct flaws in its design, provide advanced warning of potential failures and most importantly provide guidance in the changes to the warehouses operation when the data stream changes over time as it invariably does. It can also be used to perform an autopsy on the data system in the event of a major crash. With the information gathered, the causal code can be corrected and with a mix of historical data and snapshots, a failed system can be restored to full functionality.

The metadata can also be used to catalogue changes in the source systems for the warehouse. As databases are updated and new software is introduced into the operating structure of the business, these changes are catalogued and annotated by the metadata tables. Typically, the tables record the changes on a record-by-record basis and list details such

as the data's source table, primary key, time of extraction, the responsible operator the data field and its value/entry and its new value/entry to name a few. This attention to detail is the key reason for the robustness of the warehouse system and guarantees the integrity of the data held within.

The second set of tables that will now be discussed form the main data storage segment of the warehouse. These tables are arranged into what is known as a star schema.



co

In data science terminology, a schema is a plan or description on how data is stored and associated by links (also known as relations). In a star schema, a central table known as the fact table is surrounded by peripheral tables known as dimension tables; hence the name star schema as the layout of the schema looks like a star. The star schema works by placing all the available data for a specific aspect of the enterprise's operations into the Fact table. There may be as many as twenty to thirty fact tables in a typical warehouse and these can relate to operational areas such as sales, production values and finance. Fact tables can and do often hold multiple copies of data as the same data is used in different areas of the enterprise for instance, the sales department generate sales order documents with associated order numbers. These order numbers are also used by production and warehousing departments to monitor the production and delivery process associated with these orders.

Using the relational schema of the transactional systems in the warehouse would cause problems in retrieval and aggregation times for the warehouse systems. Instead, the star schema permits the duplication of data that speeds up the operational times for the warehouse as the amount of time spent searching indexes across multiple tables in different databases is vastly reduced or eliminated entirely.

With fact tables, there is essentially a single table database which forms part of the structure of the warehouse with other fact tables. However, stored data is useless if it cannot be retrieved and processed into reports or business intelligence dashboards for the systems users. This is where the dimension tables con into play. Simply put, dimension tables hold information that can be used by users

to create queries that can match entries in the fact tables and use the results to create aggregated reports that summarise groups of figures or produce lists of facts for use in managerial or operational reports. For special reports or exploratory data mining, queries can be created with special combinations of dimensional table entries to form a database of experimental data. This technique is known as ad-hoc reporting.

1. The Star schema

Source systems

A data warehouse is useless without data. So where does the data come from. Initially, the data came from the various files, plans and documents of an enterprise. As digital integration progressed, the data was converted into digital files which made the entry into the warehouse easier and widened the range and type of data that is stored. Today, there are three main domains from which the data is acquired. These are:

- Online Transaction Processing Systems (OLTP)
- Offline and Archival Data.
- Third party computer systems and documentation

As the name implies OLTP systems use the live data from processing systems as a data source. These records are processed on very short timescale schedules ranging from hourly to daily updates. Warehouse data from these systems is often used for production reporting, web updating and status enquiries.

The other main source of data for warehouses is the offline and archival records. This data holds information about previous production data, sales records, customer records, accounting records and other forms of general data of data that needs to be held for legal compliance. The bulk of this data is mainly uploaded to the warehouse on the initial upload. The data in this category does have several problems regarding formatting and definition. Old data was often recorded inconsistently leaving gaps in the record or has data entries that are varied, misspelt or of the wrong data type. This form of error is dealt with in the transform phase of the ETL process.

Third Party data is the domain that covers data from other computer systems, documentation such as price lists, product catalogues and written correspondence. The range and type of this data can be variable and needs careful handling. With documentary data, the usual approach is to use an indexed table with links to a filesystem that holds digitised versions of the documents. Electronic data is handled differently. In this case, the company needs to set up a data dictionary for the incoming data from the data link and process the data in real time if it time permits, buffer the data stream in a file buffer for later processing as a batch procedure.

2. Explain ETL

Having been mentioned several times previously, the term ETL has been used in relation to data warehouse processing. In this context, the term ETL refers to Extract, Transform and Load that is used to process the

data from source systems, correct basic errors and insert the unified data set into the tables in the warehouse. The process starts with the extract process which will be discussed next.

c) Extract

To use an analogy; ETL can be thought of as taking a set of coloured plastic bricks from several containers, then throw them into a small box, sort the contents by colour and then place the sorted bricks into a larger divided box with a single colour per division. The extract process is the act of throwing the bricks into the small box. However, this is not as simple as it seems. As the data sources are part of an operating system, the method and timing of the extract needs to be performed in such a manner to prevent disruption to the processing cycle of the production system. With SME's the timing is often based on a local working day. In this scenario, the bulk of production is performed within the nine to five working day which naturally lends itself to extracting the days information during the slack period in the evening and early morning. Larger enterprises generally operate on an international scale where different time zones become a factor and the natural lulls in processing tend to not appear or a continuous processing cycle is used to aid in global systems synchronisation. The extraction method used here is to operate concurrently with the production process cycle and extract the relevant information in real time into a buffered extraction system which introduces a degree of tolerance to the extraction stage and minimise extraction backlog interruptions or extraction flow dropouts.

With archival and historic data, the problem is different. Here, the problem is not with velocity but with volume. The quantity of data to be processed is often very large and a parallelised normalised processing system is generally the norm for dealing with this problem. In this context, normalisation refers to the aligning of data to conform with a common definition that is defined in the warehouses data dictionaries. Initially, this class of data is loaded as a single initial loading procedure so processing timing is not a factor. The only limitation being the deadline for the initial load. Once the initial

load has been performed, subsequent processing is performed on a periodic basis usually as monthly or quarterly updates.

With third party sources, the solution to external sources is essentially a mix of the other processing methods. Where relatively static data is being extracted i.e., price lists or specification sheets, the data extraction is performed on an as needed basis. Here the focus is on the voracity and value of the data. Online third-party data connections require similar handling to transactional processing systems. In this use case, it should be remembered that the source system is outside of the warehouse manager's control. Data from this source is liable to alteration, change of entry parameters, change of format, or change of perceived definition. This complicates the extract and transform components of the ETL process and greater human interaction will be required to regularise the extracted data and adapt the extraction process in line with the changes.

In all cases, the extraction process must organise the collected data into a structure that is conducive to the operating principles of the ETL process. The key to the structuring of the data is the use of data dictionaries. These dictionary's have several functions. Firstly, they hold the definition of the information. That is the meaning in the business context and its relation to other data fields in the data set. Secondly, the dictionary holds details of the source system storage location, naming conventions and data types. This specifies where the source data can be found, its field name and data type as well as any constraints applied to that data. Thirdly, the dictionary defines how the data needs to be formatted, constrained, and applied in preparation for uploading to the warehouse. The purpose here is to determine how the data needs to be transformed, checked for both consistency and conformity and into which fact tables the information needs to be applied to.

Implementations of the storage structures vary and are often dependent on the unique characteristics of the warehouse requirements. The most common approach is to store the records of the extracted data into tables that

map multiple field names onto a single field entry and append unmappable fields to the end of the table. The structure will also have fields for metadata such as source name, date/time of extraction and a global key that uniquely identifies the data item extracted. This data will be used for fill metadata tables later in the load process.

d) Transform

Transform is the second stage of the ETL process. This is where the information gained from the extract process is transformed from its raw state and converted into a form that is compatible for inclusion within the design of the warehouse. This stage of the ETL process deals with the variety within the extracted data sets. The main cause of this variety is the fact that the data is entered by human beings. Some errors such as misspellings, substitutions and mis-keying's can be mitigated by checking mechanisms within the entry procedures in properly written software. However, humans also write software and this diligence to detail is often missing or falls short of the ideal. Therefor it can be a given that any data exported to the data prep area will have errors and omissions. The first step in transforming the data is to clean the data to remove these errors. The first cleaning step is to deal with the variation in data types. As part of the extraction process, data with the same definition, as defined by the dictionary's, are stored in the same column in the storage table. The data in the column may be numeric, string or Boolean in type and must be converted to the type as specified in the warehouse's design. In modern systems, this cleaning process revolves around the use of workflows that systemises operations to the data in a step wise fashion. An example of which is the conversion of currency data. Here, accounting formats can add prefixes to numeric data and store the result as string data. If the warehouse requires the data to be numeric to two decimal places, the conversion process starts by removing the prefixes in the first stage. The second stage then converts the string type into a numeric type ensuring that the two decimal point constraint is applied. These steps from a workflow for this data type and field and is stored in the warehouse transform system for

use with later extraction data. In this way, an automated procedure is created which is the ideal solution for this stage of data processing remembering that the volumes of data can be very large and unnecessary interruptions have a negative effect on performance.

Workflows are also used to convert entries in the record to an entry that is valid within the confines of the warehouse data design. This deals with the problem of varying character cases, misspellings, and alternate variations. The result being a table that has a more uniform data set. At this point in the process, the problem of dealing with null and zero value entries needs to be addressed. This can be an area where confusion about the two value types exist. Data entry operators are generally not cognisant of the difference between the two value types. Depending on the experience of the operator, a zero value can be a numeric zero, the word 'zero' or a blank field entry. However, the computer system could interpret a blank field as a null entry (means nothing entered) or as a zero-character string which is considered as an entry. It is a desirable feature that the difference be made obvious in the data set. This is achieved by creating two defined entry variables, one for null entry and one for a zero-value entry. Workflows can be created to convert null entries in the record to a string such as "No Entry". Likewise, a zero value would be converted to a numeric value which would be a conversion rule for the warehouse transform process. With this form of methodology, the rules created within workflows should cover the bulk of transformations in the data prep area. However, there will still be exceptions. The process to handle these exceptions will vary with the implementation but will still fall into identification, remediation, and inclusion phases. The identification phase marks the meta data entries as requiring human intervention. This flags the record as requiring a human decision to remedy the event. The remediation phase details on how the event is handled. This can take the form of adding an extra rule to a workflow to handle future occurrences of the event or it could simply require the entry to be edited or deleted by a warehouse operative. Where the event is the result of a restructuring

of source data, the procedures within the ETL process would need to be altered by the development team to reflect the restructuring of the source.

In a mature warehouse system, this type of intervention should be minimal. With new warehouses, this would be a common occurrence as the list of variant entry types is still growing and suitable remediation has not been implemented. It also means that the inclusion of remedied records might not have been decided on. To elaborate, corrected data can either be used to replace the erroneous record in the extraction structure or it could be stored for later inclusion of the extracted data in the next round of the ETL procedure. The choice of inclusion would depend on the timing requirement of the ETL procedure. Where the ETL process is real-time based or has very tight deadlines then the latter option would be preferable as the 'faulty' record is omitted from the initial upload and the corrected record is reintroduced on the next upload. This option would minimise upload delays but could introduce synchronisation artifacts where real time reporting is used. The use of the other method would result in the reverse result where there may be delays in upload timing and no synchronisation artefacts. This method would be more desirable where batch processing is used, and time availability is not critical.

e) Load

The load is the third stage of the ETL process. As discussed, the structure of the warehouse is based around fact tables and dimensions. During the load process the tables are extended by appending the extracted data to the relevant tables. Depending on the implementation, the loading process can either be serial in nature where a table is updated one at a time or may be parallel where all tables are updated simultaneously.

In most cases, the parallel load process is used. This is primarily because of the volume of data that needs to be loaded. This is achieved by the creation of temporary staging tables that are created using the star schema's layout and then being populated with data from the extraction table. These staging tables are then used by

append queries to add the new data into the existing fat tables in the warehouse. As the process is parallel in action, the queries are run in independent subprocesses in the server with one table per subprocess.

Whether a single table or multiple tables are being processed, the actions performed are logged as metadata in the metadata tables of the warehouse on a record-by-record basis. This forms an audit trail of the state and changes of the table and allows failures to be flagged in the event of a load error. This would facilitate a reload operation from the point of failure saving time and potential corruption from inaccurate overwriting of the current data.

f) Metadata

Metadata is by description information about data and the majority of data warehousing operations rely on metadata at some level or other. The bulk of the metadata is stored in dedicated tables in the database structure however some of the metadata is attached to fact tables as separate fields that cover details such as creation date and date of last update. These fields can also be crossed indexed with upload tables as an integrity check. Coming back to the metadata tables, these tables record the operational metrics of the warehouse. As mentioned, upload records record the operation of the upload process, the source records involved, time taken, fields changed with initial states and errors in loading, associated error codes and the surrogate record number of the affected record(s). Other tables will record changes to the warehouse star schema changes, user access and associated data like terminal number, times and dates and records accessed. Security processes are also recorded by these tables with backup process data being stored to aid recovery from failed backup/restore operations and to facilitate partial and full archival recordings. One factor to bare in mind is that the production of metadata by volume is far greater than the sum of the factual data entered into the warehouse data system. Therefore when planning the resources for a warehouse, the metadata storage and expansion capabilities need to be upgradable and flexible to facilitate changes when storage technology evolves and offers better and new options.

3. Datamart's

With most of the enterprises data being held on one system, security is a prime concern. Security covers many aspects where the warehouse is concerned and include data integrity, physical security, and access control. It is in relation to access control that data marts come to the fore.

Data marts are an extension of the warehouse that provide secure access to selected segments of data for authorised users. The data is read only in nature and can only be updated via the originating source data system the security principle being that data flows only one way through the warehouse. Data marts work by creating datasets from one or more fact tables using SQL view queries and making the filtered data available to applications and presentation software that is used by the information workers in that department. Data marts can be created for sales and PR, accounting, and managerial departments for the generation of reports. For senior management, data marts are useful for aggregated reporting such as yearly or quarterly sales figures and regional profitability reports as well as being used with data mining techniques to discover trends and profitable leads in marketing. For the latter, data analysts can modify existing queries by the simple addition of additional dimension tables. This is an easy feature to implement and does not require the adaptation of existing fact tables and related schema and can be undone just as easily.

4. Data marts for managing financial and ops directors

Dedicated reporting structure for specific departments

Within the previous section, it was mentioned that data marts used filtered queries to present limited and role specific access to staff. This feature is the basis for an efficient and effective business tool. This is not a tool or a piece of hardware, it is a data structure geared to providing numerous departmentalised reports for low to mid-level management. The development and use of a dedicated reporting system improves decision making and provides prompt and timely notification of potential

problems. As the reporting scheme develops the warehouse managers can fine tune the underlying data provision for the mart to optimise the marts functionality.

Data interface for 3rd party software's (CRM)

In departments where data marts are used the output can be configured to interface with third party software. A typical example of this is the use of Customer Relationship Management software where existing client data, purchase history and communications records can be presented by the mart for use by the software for use by call centre staff with little chance of being changed by untrained staff or being accessed by unauthorised terminals.

Provision for ad-hoc querying without altering active data systems

The provision of a dedicated reporting system is useful for users at the operational level however, upper management require a 'ten-thousand-foot view' with their data which the reporting system can provide. They will also need the ability to run 'what if?' information searches. Again, here is where the data mart shines, data analysts can create ad-hoc queries based on current data across several systems without interrupting the operation of active control systems across the company.

Ability to display historical and non-operational data

Not all departments within the business will require operational data. They will require data of various types. This can include photographs, old price lists and historic documents. The data mart comes to the fore again as such documents can be made available. Departments such as Public Relations, Advertising and sales can obtain the full range of documents to produce new literature, adverts, and sales promotions. Without the warehouse, access to these documents would be spread across the company and availability between departments would be limited.

4. IT director – explain differences between relational normalised and data Warehouse systems structure

Relational vs archival systems

Databases are like horses. Some horses are built for speed and others are made for hauling and their structure is a result of their abilities. In the business world, there are operational processing systems, archival storage networks and dedicated monitoring and reporting systems. Their function determines the form of the underlying data structures and operational methodology. In the Rags to Riches scenario, the two forms of data systems are operational which prioritises individual data record retrieval and update and dimensional warehouse system that prioritises mass record retrieval and aggregated reporting functions.

The relational database system was devised by Dr Edward Codd[1] in 1973 and was realised by IBM in 1974 as the System R data system and was the grandfather of all modern relational systems. As a child of the 70's relational data systems developed with the optimisation of storage space and efficiency of cost. To this end Dr Codd's rules of normal form were devised to allow large data sets to operate at optimal speed within limited storage capacities. Relational database systems optimise storage by removing duplicated field entries in records and storing the duplicated field into a single record in a dedicated indexed table. Where a primary data record has several duplicated fields, the dedicated tables are linked to the primary via a keyed index field. Searches based on queries that use the relation link scheme produce data sets that search the dedicated tables for matching keys and combine the results into the original data records. Searches for multiple data records would involve searching these tables multiple times for each returned record. The time to retrieve data with this system exponential multiplies with the number of parameters times the number of records. Therefore, finding one data record is fast but finding data records with three parameters over 10,000 records would be $(3 \times 3) \times 10,000$ accesses to the system. Very Slow!

Dimensional warehouse systems are different from relational systems. Based on the older flat bed or single table database structures, the warehouse system is fast on record retrieval but inefficient on storage management using far more space than an optimised relational system. This feature makes this form of data structure highly efficient when generating reports. Queries that have multiple search parameters only need to search one table one record at a time. To produce a result data set, only one set of records needs to be scanned once making reporting very quick. Warehouses double up on this form of storage by using a star schema as the main storage structure for the warehouse.

Star schemas work by creating a central fact table that holds a combination of measures (data capable of performing as source data for a function) from the original source table and linked foreign tables, primary key fields (from relational tables) and static data. Linked to these fact tables are dimensional tables. These are tables that hold a special variant of static data (data that rarely changes) which act as query parameters for a linked field. Dimension tables are often derived from lookup table data. The beauty of star schemas is that data from different sources can be converted into fact tables using a standardised set of conversion rules which makes them linkable and with the correct set of dimension tables can create functional reports quickly for all levels of users. The data can come from many sources and of different types. With specialised groups of reports, data can be presented as a data mart that can provide a read only data source that would be secure, adaptable, and depending of the ETL frequency almost instantaneous.

Retrieval complexity timings

Fact tables run faster because there is only one table to search that does not require access multiple table sources and results in what is known as lower retrieval complexity value. This value being a simple calculation that uses the time to retrieve an item of data divided by the number of machine operations to obtain the data. As shown in previous paragraphs the

difference in timing is decidedly marked. The retrieval time per record for the relational system example would be 8 as the number of accesses per related table is multiplied by the number of permutations of the search keys minus 1. The timing for a fact table is 1 as the retrieved record can be tested three times without further access to other tables leaving the fact table as the only searchable structure.

5. Task 2 – The Rags 2 Riches Data Warehouse build

The following section of the document will provide an incomplete but workable design specification for this theatrical company.

5. Data sources

In preparation for the design, several assumptions have been made. Firstly, it is assumed that the database systems are available company wide on a protected network. Secondly, the source databases are built using the same schema and lastly, the data entries will follow the same kind of format. With buildings in different locations, procedures and types of data entry tend to vary with the locality and is dependant on the skill level of the system operators. It is presumed that all operators will act equally and in the same way.

On studying the provided description, the following data sources have been identified.

- g) Known operational sources
 - Orders Database (includes client data, rental history data and package details). Possibly may include a package listing which itemises the stock holding/substitutions information.
 - Transfer Database (ties in with the transport database)
 - Transport Database (Includes delivery data, vehicle data and the local acquisitions and damaged stock lists)
- h) Non-operational sources (presumed)
 - Archived Legal documents
 - Price lists
 - Paper documents – Orders, invoices, transfer, and delivery notes
 - Workshop repair records and costings
 - Write off list
 - Daybooks, accounting ledgers and related spreadsheets (Legal Requirement)
 - HR Documentation, Contract of employment etc.
 - Marketing documents, historic company documents and photographs.
 - Stockholder reports, board meeting minutes, quarterly and yearly reports.

Third Party data (presumed)

- Web based email
- Written communications(letters)
- Acquired historical research data
- Specialist supplier listings

Based on the above operational source list a generalised relational schema was devised and is shown in figure 3. This is not a complete or thorough description but does show the general connections between the tables and how they interact between the databases used within the company.

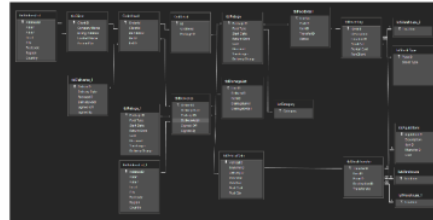


Figure 2

- i) Hire dept/Clients
- j) Warehouses
- k) Transport
- l) Finance – income – outgoings
- m)

6. Star schema

Facts

Orders

Inventory

Package Hire

n) Dims

Date

Location

Client

Stock Cat

7. Data marts

Company managers (top level)

Division Managers

Sales Dept.



Figure 3 Sale Relational Schema

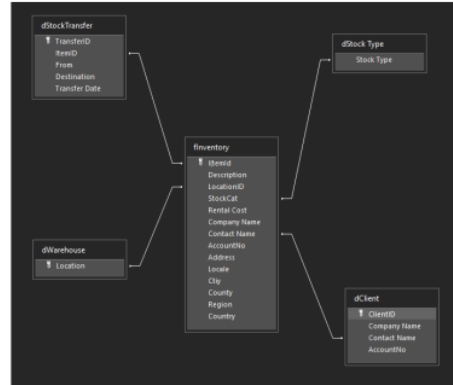


Figure 6 Inventory Star Schema

Transport Dept

Deliveries

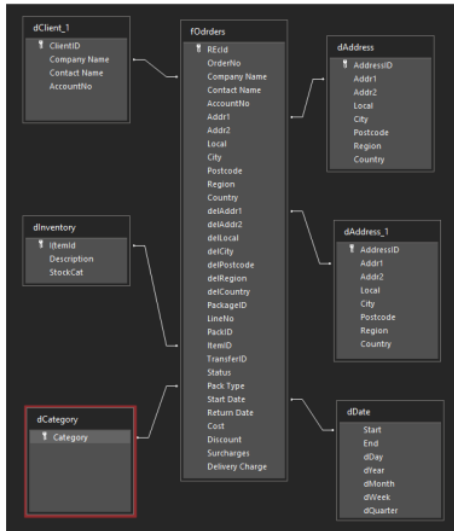


Figure 4 Sales Star Schemas

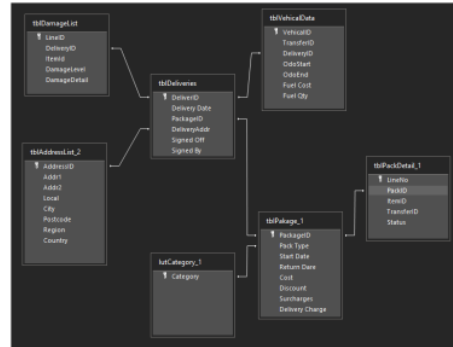


Figure 7 Deliveries Relational Schema

Inventory/Stock Dept.

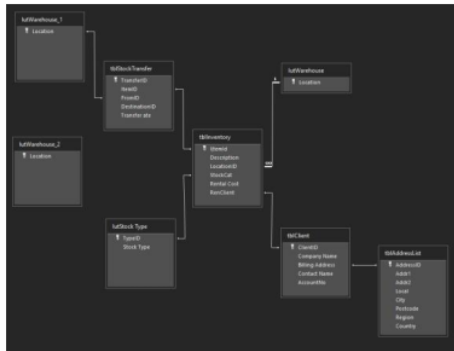


Figure 5 Inventory Relational Schema

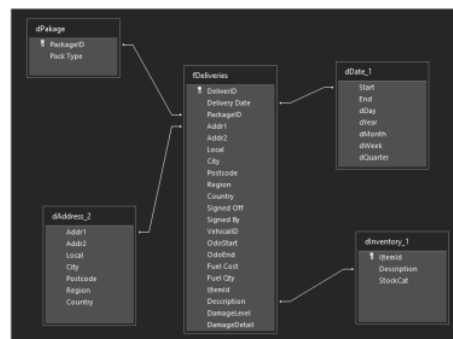


Figure 8 Deliveries Star Schema

Transfers

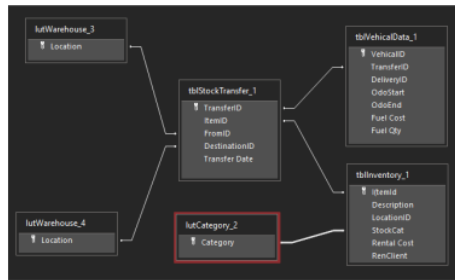


Figure 9 Transfers Relational Schema

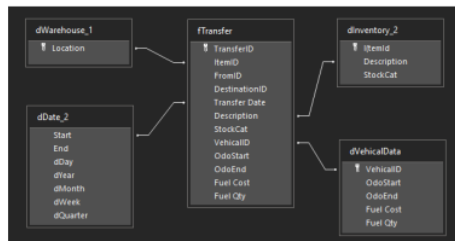


Figure 10 Transfers Star Schema

Vehicle Data

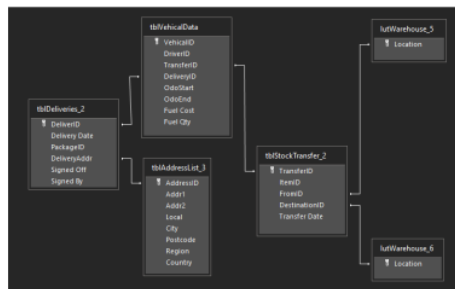


Figure 11 Vehicle Relational Schema

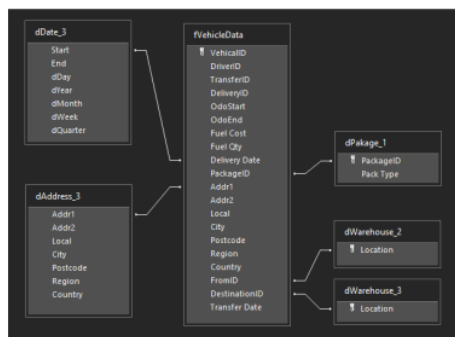


Figure 12 Vehicle Star Schema

Accounting.

Invoicing and payments

Acquisitions

Workshop

o) Mart identification

p) Data requirements

8. ETL PLAN

This is an outline ETL plan detailing the general process of operation for Rags to Riches system. Table details have been shown in section 11.

q) Table data sources

Step 1 create temporary storage tables in staging area. Tables to hold data for master tables, primary fact tables and dimension tables. delayed records from previous ETL process to be pre-loaded into the tables.

Step 2 retrieve cut off data from warehouse metadata tables Data retrieved to include last record processed, date/time and last surrogate key value.

r) Extraction

Step 3 execute extraction queries using cut off parameters. Extraction queries are of filtered append type with temporary storage tables as data sinks for the queries. Surrogate keys to be created for each record as a calculated field. Error events to be logged to extraction metadata file for later correction.

s) Transformation requirements

Step 4 Execute first stage automated flow processes of extracted data files.

Step 5. Check error files on extract and flow processes and manually fix or mark for delayed loading. Delayed records to be logged to metadata system for detailed correction and loading on the next ETL cycle.

Step 6. Execute secondary workflow processes to update dimension tables.

Step 7. Process master tables for inclusion into master fact tables. This will also convert look up table referrals into short text data type fields.

t) Post Processing and Staging

Step 8. Execute prestaging queries to create temporary fact and dimension tables for each table in the main warehouse system. Process will use make table queries to load transformed data ready for the load process. Errors and processing details to be logged into metadata subsystems.

Step 9. Execute parallelised Staging process. Threaded append queries will bulk load the contents of the temporary staging tables to the end of the active warehouse tables, dimension tables first followed by fact tables. metadata tables to be updated with upload data as records are processed.

6. Bibliography

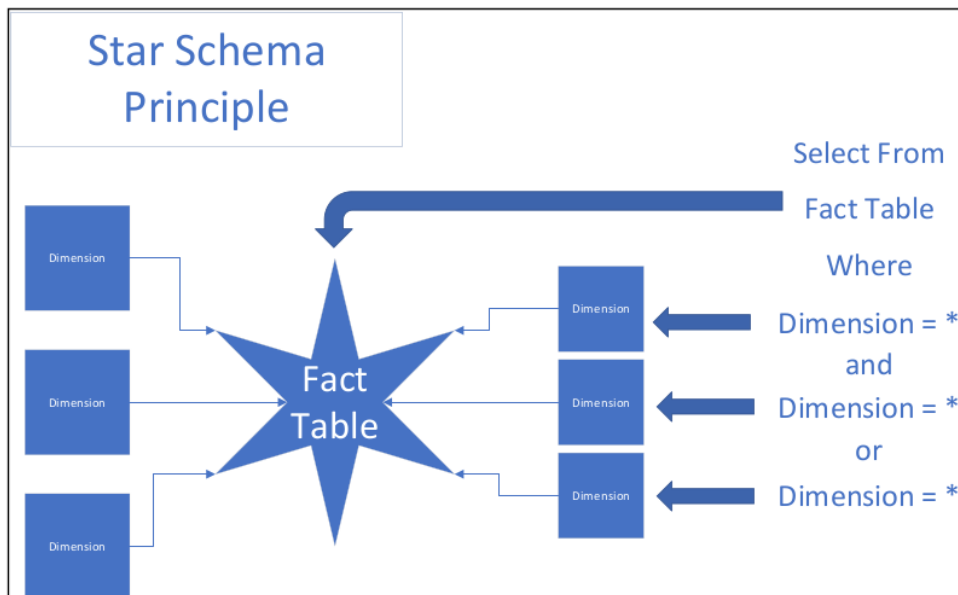
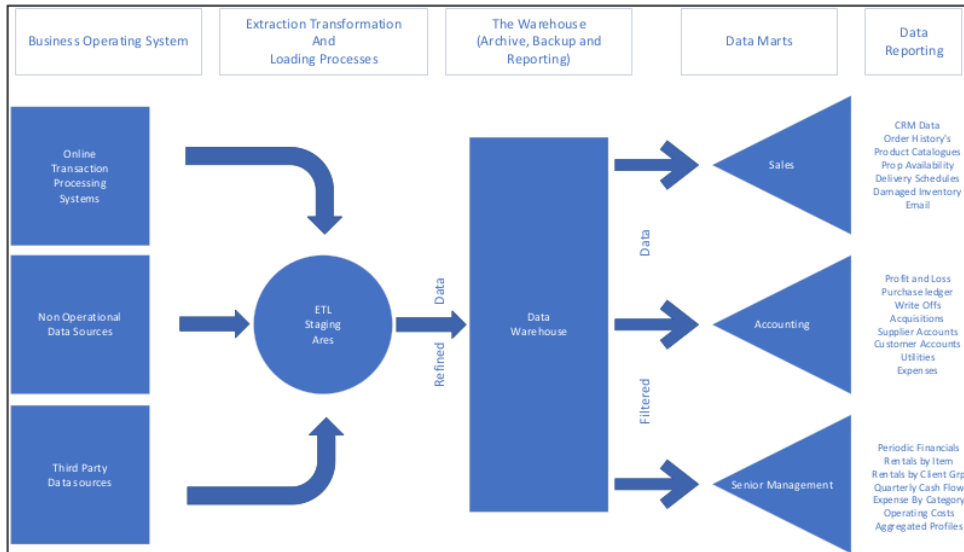
- [4] P. Ponniah, Data warehousing fundamentals for IT Professionals, Second edition, 2011.
- [5] C. Tupper, Data Architecture: From Zen to Reality (Google eBook), 2011.
- [6] C. Tupper, Data Architecture, 2011.
- [7] T. Connolly and C. Begg, Database Systems: A Practical Approach to Design, Implementation and Management (International Computer Science Series), vol. 29, 2010.
- [8] R. Kimball, L. Reeves, M. Ross and W. Thornthwaite, The Data Warehouse Lifecycle Toolkit, 2008.
- [9] R. Wrembel and C. Koncilia, Data warehouses and OLAP: Concepts, architectures and solutions, 2006.
- [10] L. C. Rivero, J. H. Doom and V. E. Ferraggine, Encyclopedia of database technologies and applications, 2005.
- [13] P. B. a. E. F. Codd, "Information Retrieval a Relational Model of Data for Large Shared Data Banks", 1973.

Figure 1	3
Figure 3	12
Figure 4 Sale Relational Schema	13
Figure 5 Sales Star Schems	13
Figure 6 Inventory Relational Schema	13
Figure 7 Inventory Star Schema	13
Figure 10 Deliveries Relational Schema	13
Figure 11 Deliveries Star Schema	13
Figure 12 Transfers Relational Schema	14
Figure 13 Transfers Star Schema	14
Figure 8 Vehicle Relational Schema	14
Figure 9 Vehicle Star Schema	14

Appendices

Appendices

Full sized Images.



9. Star Schema Tables

Fact Tables

Appendices

Field Name	Data Type	Description (Optional)
SurrogateKey	Short Text	ETL process Generated Key used for identifying Data Records . Primarily used for metadata table indexing/
Itemid	AutoNumber	
Description	Short Text	
LocationID	Short Text	
StockCat	Short Text	Stock Category
RentalCost	Currency	Cost of Rental by day
Company Name	Short Text	
Contact Name	Short Text	
AccountNo	Short Text	
Address	Short Text	
Locale	Short Text	
City	Short Text	
Country	Short Text	
Region	Short Text	
Country	Short Text	

Orders		
Field Name	Data Type	Description (Optional)
SurrogateKey	Short Text	ETL process Generated Key used for identifying Data Records. Primarily used for metadata table indexing/
OrderNo	Number	
Company Name	Short Text	
Contact Name	Short Text	
AccountNo	Short Text	
Addr1	Short Text	
Addr2	Short Text	
Local	Short Text	
City	Short Text	
Postcode	Short Text	
Region	Short Text	
Country	Short Text	
delAddr1	Short Text	
delAddr2	Short Text	
delLocal	Short Text	
delCity	Short Text	
delPostcode	Short Text	
delRegion	Short Text	
delCountry	Short Text	
PackageID	Short Text	
LineNo	Short Text	
PackID	Number	
ItemID	Number	
TransferID	Number	
Status	Short Text	
Pack Type	Short Text	
Start Date	Date/Time	
Return Date	Date/Time	
Cost	Currency	
Discount	Currency	
Surcharges	Currency	
Delivery Charge	Currency	

Field Name	Data Type	Description (Optional)
WarehouseKey	Short Text	ETL process Generated Key used for identifying Data Records. Primarily used for metadata table indexing/
TransferID	AutoNumber	
ItemID	Number	
FromID	Short Text	
DestinationID	Short Text	
Transfer Date	Date/Time	
Description	Short Text	
StockCat	Number	Stock Category
VehicleID	Short Text	
OdoStart	Number	
OdoEnd	Number	
Fuel Cost	Currency	
Fuel Qty	Number	

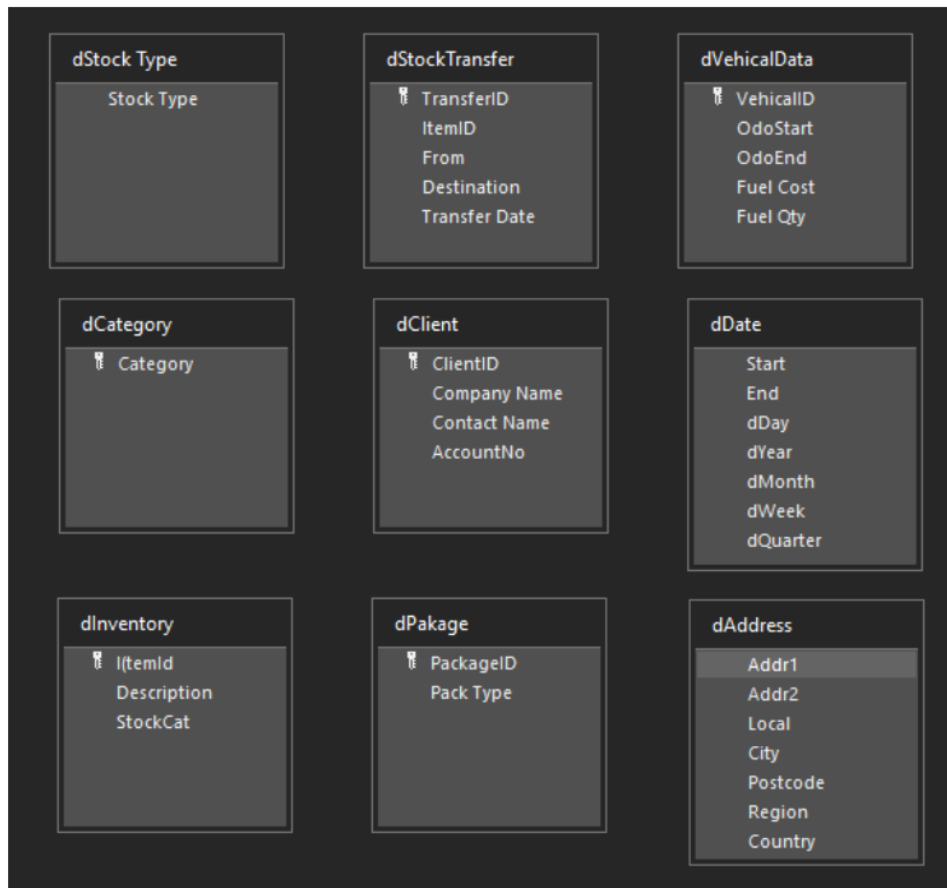
Appendices

Field Name		Data Type	Description (Optional)
VehicleID	AutoNumber		
DriverID	Short Text		
TransferID	Short Text		
DeliveryID	Short Text		
OrderStart	Number		
OrderEnd	Number		
Fuel Cost	Currency		
Fuel Qty	Number		
Delivery Date	Date/Time		
PackageID	Short Text		
Addr1	Short Text		
Addr2	Short Text		Deliver Address
Local	Short Text		
City	Short Text		
Postcode	Short Text		
Region	Short Text		
Country	Short Text		
FromID	Short Text		transfer data
DestinationID	Short Text		
Transfer Date	Date/Time		

Database		Description (Optional)
Field Name	Data Type	
SurrogateKey	Short Text	ETL process Generated Key used for identifying Data Records. Primarily used for metadata table indexing/
DeliveryID	AutoNumber	
Delivery Date	Date/Time	
PackageID	Number	
Addr1	Short Text	
Addr2	Short Text	
Local	Short Text	
City	Short Text	
Postcode	Short Text	
Region	Short Text	
Country	Short Text	
Signed Off	Yes/No	
Signed By	Short Text	
VehicleID	Short Text	
Odolstart	Number	
OdolEnd	Number	
Fuel Cost	Currency	
Fuel Qty	Number	
Itemid	Short Text	
Description	Short Text	
DamageLevel	Number	
DamageDetail	Long Text	

Dimension Tables

Appendices



10. Data Marts

Datamart: Company managers (top level)	
Report type	Report Format
<i>Price lists</i>	
<i>Orders</i>	
<i>Stock availability</i>	
<i>Client accounts</i>	
<i>Prop dictionary</i>	
<i>Internal Stock Transfer Order</i>	

Datamart: Division Managers	
Report type	Report Format

Appendices

Rental Reports	
Expense reports	

Datamart: Sales Dept.	
Report type	Report Format
<i>Price lists</i> <i>Orders</i> <i>Stock availability</i> <i>Client accounts</i> <i>#Prop dictionary</i> <i>Internal Stock Transfer Order</i>	

Datamart: Inventory/Stock Dept.	
Report type	Report Format
<i>Inventory tracking</i> <i>Delivery record</i> Processing & status records Despatch record <i>Goods in record – stock & acquisitions</i> Inventory Holding	 Agregate Total.

Datamart: Transport Dept	
Report type	Report Format
<i>Inventory tracking</i> <i>Delivery record</i> Despatch record <i>Goods in record – stock & acquisitions</i> <i>Delivery sheets – includes acceptance document</i> <i>Fuel logs</i> <i>Vehicle logs</i> <i>Collection's sheet</i> <i>Collection reports</i>	

Datamart: Accounting.

Appendices

Report type	Report Format
Monthly income/expense Profitability <i>Collection reports</i> Inventory • Acquisitions • Write offs • Damage • Asset disposal Employee list	

Datamart: Invoicing and payments	
Report type	Report Format
Monthly income/expense Acquisitions Write offs Profit and Loss Residuals Reporting Asset disposal	

Datamart: Acquisitions	
Report type	Report Format
New Stock Acquired	
Acquisition Expense	
u)	

Datamart: Workshop	
Report type	Report Format
Damaged Stock List	
Write Off Stock List	
Work in Progress	
Supplier Source	

Unfinished Report

GRADEMARK REPORT

FINAL GRADE

69/100

GENERAL COMMENTS

Instructor

Mark Vaughn

1 engagement with literature
comment available

8 references - mainly books - try to include
a variety of source types

for some reason your bibliogoraphy starts
on 4 20

reference sources listed in bibliograpy,
but little direct citations evident in the
report body

2 Knowledge and Understanding
comment available

Good description of all elements in task1

very detailed discussion of ETL 20

Managerial explanation well explained in
neutral terms

3. Cognitive and Intellectual skills
comment available

Managerial explanation well explained in
neutral terms

good discription of how the theory can be 10
translated into practice throughout - some
sections not discussed

4. Practical Application Skills
comment available

some sections in task two have headings
which would apper to have been thought
of but lack further discussions

40

Diagrams and designs are generally very
good but some are diffiult to read even at
high zoom

5. Transferable Skills for life and Professional pract
comment available

Technical terminology broken down into 10
non expert language - I particularly
enjoyed the database-horses comparison

Nicely laid out columnar report

Incorrect title

Some numbering and indenting
inconsistent

Try to be more succinct in your writing

no clear introduction

Total: 69

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
