Leeds Beckett University
Faculty of Arts, Environment & Technology

MSc Business Intelligence

Academic Year 2015-2016

Joanne Kennedy C3369865

Data Warehouse Models and Approaches: Designing PlaceU Data Warehouse System

Date of Submission: 13th March 2016

Contents

Article I. Identification of a case study and Data Warehouse Requirements

Section 1.01 Background and Strategy

PlaceU is an IT and media specialist recruitment agency with offices based in Leeds, Manchester and Liverpool and the managing director of the company has been focusing on new innovative ways to enhance business value, generate more revenue and gain a competitive advantage.

The MD knows that attracting new accounts can be a timely and costly process so generating a high lifetime value and building strong relationships with current accounts is the key to the sustained growth and success of the business. (Maynard, 2010)

The MD also believes that generating more revenue will come from generating more productivity and motivation in his offices, which is why recognition of staff performance is a key part of his strategy.

PlaceU are currently using traditional relational database methods to store and manage their data, however the managing director of the company would like to implement a data warehouse which will integrate operational data from his various offices into a single and consistent architecture. (Perkins, 2003)

As it stands the company currently have a large amount of unstructured data which is not being used to the best of its ability. The implementation of a data warehouse and the ability to mine and analyse this data, will be a useful tool to access quick and accurate information, and produce reports on business patterns and trends. (Intel IT, 2013)

Section 1.02 Objectives and Requirements

In order to accomplish the business goals the MD has a number of objectives and requirements that the data warehouse must be able to fulfil.

The first focus is on staff reward schemes; employees have a major impact on business performance and therefore by bringing in a staff reward scheme this will hopefully influence the behaviour of the employees for the better. (ACCA, 2013)

The main objectives are to:

- Investigate the performance of each location, which will be accomplished by determining how many placements consultants have secured and how many of these placements have turned into permanent positions, as well as looking at the staff turnover and loyalty over the past year.
- Investigate the financial performance by looking at the amount of revenue generated by each consultant over the last year in comparison to the previous year.

The MD wants to initiate the following schemes.

- Yearly reward scheme for staff loyalty – 5+ years.
- Yearly reward scheme for consultants that secure a certain number of placements.
- Yearly reward scheme for consultants that secure permanent placements.
- Yearly reward scheme for consultants that generate the most revenue for the business.

The MD also wants the ability to locate the accounts for each office that provide contractors with the maximum salary, in order to form stronger relationships with these accounts and hopefully generate more business with them in the future.

Section 1.03 Data Warehouse Reports

A way to achieve the MD's plans would be to produce a number of reports on the data once it has been transformed and loaded into the data warehouse, a list of potential reports can be seen below.

- No of placements secured by consultants in the Leeds office in comparison to the Manchester office during 2015.

- No of placements secured by consultants in the Leeds office in comparison to the Liverpool office during 2015.
- No of placements secured by consultants in the Manchester office in comparison to the Liverpool office during 2015.
- No of permanent placements secured by consultants in the Leeds office in comparison to the Manchester office during 2015.
- No of permanent placements secured by consultants in the Leeds office in comparison to the Liverpool office during 2015.
- No of permanent placements secured by consultants in the Manchester office in comparison to the Liverpool office during 2015.
- Amount of revenue generated by consultants during 2015 compared to 2014.
- No of accounts that paid the maximum salary in the Leeds office in comparison to the Manchester office during 2015.
- No of accounts that paid the maximum salary in the Leeds office in comparison to the Liverpool office during 2015.
- No of account that paid the maximum salary in the Manchester office in comparison to the Liverpool office during 2015.
- No of consultants that have worked for the company for over 5 years in the Leeds office in comparison to the Manchester office.
- No of consultants that have worked for the company for over 5 years in the Leeds office in comparison to the Liverpool office.
- No of consultants that have worked for the company for over 5 years in the Manchester office in comparison to the Liverpool office.

Section 1.04 Data Sources

The data sources chosen for this case study can be found in figure 1 below.

No	Data Source	Description	Format
1	lds_placeU	Current relational database for Leeds branch	SQL code
2	mch_placeU	Current relational database for Manchester branch	SQL code
3	lvp_placeU	Current relational database for Liverpool branch	SQL code
4	PlaceU_placements	Data on placements	Excel Spreadsheet
5	PlaceU_locations	Data on branch locations	Excel Spreadsheet

Figure 1: Data Sources

Article II. Data Warehouse Architecture and Methodology

Data warehouses are the foundation for supplying insightful business solutions and provide a great way to gain competitive advantage. When planning a data warehouse it is important to keep in mind how the data warehouse will grow in size and complexity as the company expands, and this is why the data mart approach is widely recommended as it focuses on one operational system at a time. (Chenoweth, Corral and Demirkan, 2006)

There are three key types of data warehouses; enterprise data warehouses (EDW), operational data stores (ODS), and data marts. EDW is the area of the business where the data is centrally stored, accessed and integrated. (Teradata, 2007) Although this is a common solution for decision support, it is a large-scale data warehouse and therefore would not be appropriate for this case study at the moment. (Turban, Sharda, Delen, 2011)

ODS are a quick short-term solution to problem solving and decision making, however they only store recent information, and although they would currently be able to fulfil the MD's requirements, this would not be a beneficial solution for PlaceU in the long-term, especially when the business expands and new objectives and requirements need to be fulfilled. (Turban, Sharda, Delen, 2011)

The final type of data warehouse is data marts, which are small subsections of a data warehouse that focus on a particular department within an organisation. (Turban, Sharda, Delen, 2011)

There are a number of considerations associated with designing a data warehouse, for example the type of data model to produce; it is important that the data model is fully flexible and responsive in order for it to address current, as well as future business needs. (Teradata, 2007)

With this in mind, there are two common approaches to designing a data warehouse, Ralph Kimball's bottom-up approach uses data marts as a starting point, and gradually merges them together in order to form a fully fledged data warehouse (EDW). (Standen, 2008) Whereas Bill Inmon follows a top down approach which

focuses on designing a normalised logical model of the enterprise data warehouse followed by physical dimensional models for each data mart. (George, 2016)

	Inmon	Kimball
Building Data warehouse	Time Consuming	Takes lesser time
Maintenance	Easy	Difficult, often redundant and subject to revisions
Cost	High initial cost; Subsequent project development costs will be much lower	Low initial cost; Each subsequent phase will cost almost the same
Time	Longer start-up time	Shorter time for initial set-up
Skill Requirement	Specialist team	Generalist team
Data Integration requirements	Enterprise-wide	Individual business areas

Figure 2: Inmon vs. Kimball (George, 2016)

After reviewing both approaches, Kimball's approach seems like the most appropriate solution for PlaceU. A data mart would be an adequate method in order to provide the MD with the information he currently requires. This approach is not only less time consuming, less costly and requires less specialist skills but it also allows for future progression and expansion into an EDW. (George, 2016)

It is important to consider data ethics at this stage to determine if any ethical issues may arise in terms of how the information is being used and by whom. Generally speaking people often think ethics is only an issue to consider when working on a large-scale project with big data, however ethical concerns can also arise from small data collections.

In the future PlaceU may want to gather external sources or even buy them in, for example competitor's data to determine how and where they stand in the current market, the top skills wanted by contractors and various other demographics, however it is important to know the difference between information that is acceptable for analysis and information which poses a risk to ethics.

In order to manage ethical concerns a number of things can be done, for example defining who has access to the information; which in PlaceU's case the data will not be made public and only high level management will have access to this data. Other

ways to manage ethical concerns is to involve consultants, contractors and accounts in the process of defining the ethical standards documentation, locking-down the test environments and ensuring to stick to the laws and legislations concerning the security and privacy of data and information i.e. The Data Protection Act. (Linstedt, 2004)

It is also important to consider the scope of data; how often the data will be updated and reloaded into the data warehouse, in PlaceU's case this is relatively straightforward as the MD is only wanting to look at figures on a yearly basis in order to reward his consultants, check the overall performance and know which accounts each location should be aiming to work with, so the data will be updated and reloaded at the end of every financial year.

Article III. Data Warehouse Design

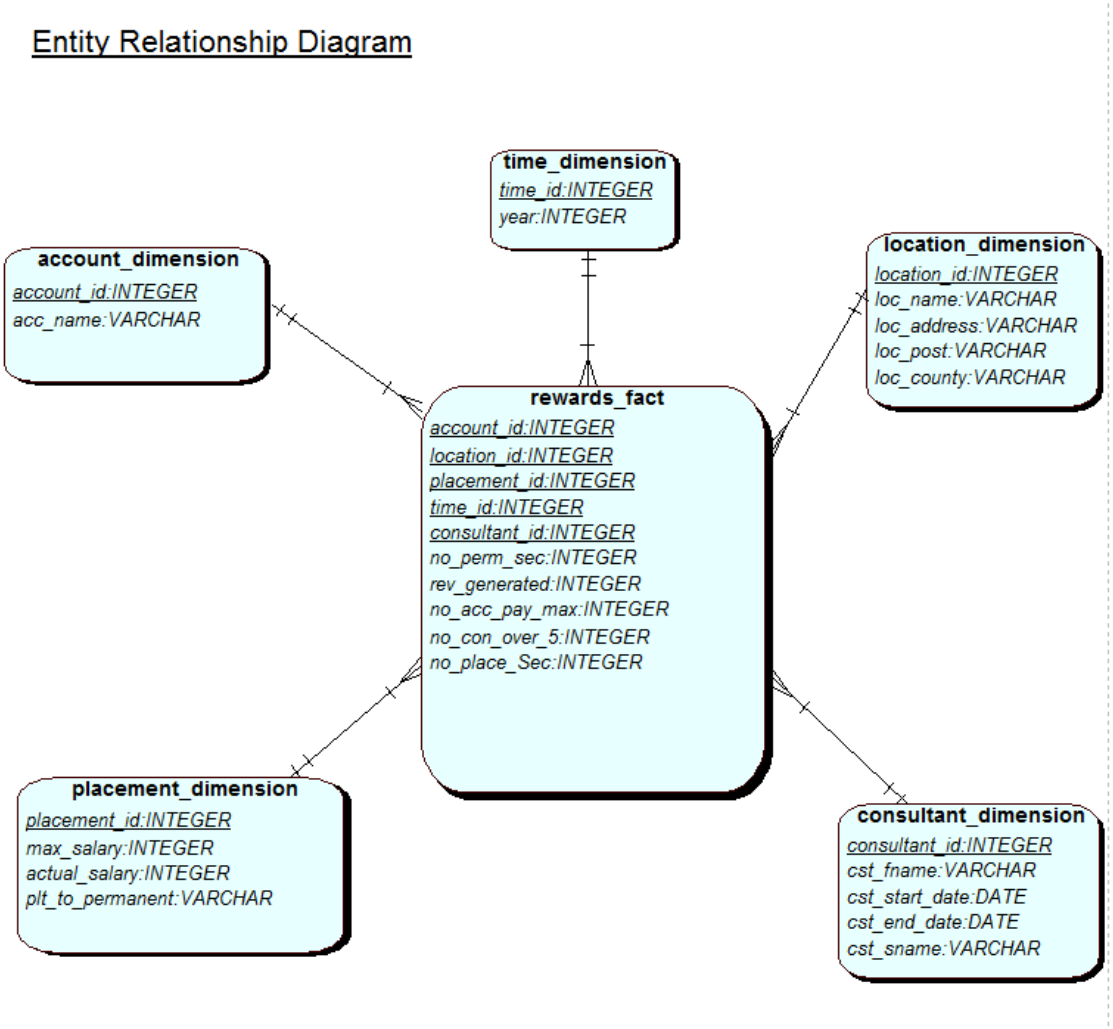

Figure 3: Data Warehouse Design

Above is the final design for the data warehouse which includes a fact tables with the dimension keys and measures which will be useful for the managing director in order to fulfil his requirements and objectives.

When designing the fact table it was important to determine the granularity of the data, or in other words the lowest level of information required for the MD to accomplish his goals. Looking back to the MD's requirements it was clear that this process was mainly to integrate reward schemes within his company, in order to find a way to motivate staff. With this in mind, that was why the decision to include a placement dimension, a consultant dimension, a location dimension and a time dimension was made. It is also notable to remember that the MD also wanted to locate the accounts that were paying the maximum salary to contractors, in order to strengthen the bond and relationship with these accounts; this is why the account dimension was also added to the star schema. (1keydata, 2016)

It is important to point out that the Rewards_Fact table has five foreign keys coming from each of the dimension tables, a surrogate key could be used for this project however the decision has been made to create a composite key made up of each of its foreign keys.

In terms of the measures selected for the Rewards_Fact table, a no_perm_sec and a no_place_sec were chosen to count the no of placements and permanent placements secured by consultants. The rev_generated is there to total up the amount of money being brought into the business via the consultants which can be worked out using the contractor's actual salary, as PlaceU get 15% of this. The no_con_over_5 was chosen to count the no of contractors who have worked for the company for over 5 years and the no_acc_pay_max was chosen to count the number of accounts that may contractors the maximum salary, as this would mean a higher income for PlaceU.

An issue to consider would be data quality due to the data coming from multiple sources, as poor data quality can result in poor analytics and reporting, leading to

poor business decisions. A brief overview of the data provided by PlaceU's existing systems do not show any major outstanding inconsistencies, duplicates or missing data, however to ensure this is the case, the data will go through the ETL process which stands for extract, transform and load. (Wentzlaff, 2014)

The ETL process encounters a cleaning stage after the initial extraction of the data. This is probably one of the most important steps in the process to ensure the quality of the data is good enough to be transformed and loaded into the data warehouse. The cleaning process is there to ensure a number of things for example, all identifiers are unique, all null values are standardised as a not available value, all phone numbers and postcodes are standardised into their correct format, all address fields are validated into a consistent naming format and then validated against each other to ensure they have been entered correctly. (Data Integration, 2016)

In a recent case, errors within the police's data led to people being wrongly mixed up in a paedophile investigation which resulted in a false arrest due to human error. (BBC, 2015) Although the outcome of wrong data within PlaceU will not have as much of an impact, it is important that data is uploaded cumulatively with the correct running total, and it is important to remember that removing someone from the upload sheet will not remove them from the data warehouse. (NHS, 2016)

A data dictionary has been prepared into order to allow business users to fully understand the data that is stored inside the data warehouse, it is intended to be accessible for not only reading, but updating as new sources are found and fresh useful information is gathered. It is important to note that the data dictionary must be well managed and kept up-to-date in order to provide rich and reliable information. The data dictionary for PlaceU can be found below in figure 4. (Simms, 2012)

Table Name: Placement_Dimension

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P	placement_id		Number			lds_placeU.sql,
	max_salary	Maximum Salary	Number		Required	mch_placeU.sql,
	actual_salary	Actual Salary	Number		Required	lv_placeU.sql,
	plt_to_permanent	Placement turned into permanent position	Text	3	Required	PlaceU_placements.xlsx

Table Name: Account_Dimension

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P	account_id		Number			lds_placeU.sql,
	acc_name	Account Name	Text	30	Required	mch_placeU.sql,
						lv_placeU.sql,
						PlaceU_placements.xlsx

Table Name: Consultant_Dimension

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P	consultant_id		Number			
	cst_fname	Consultant First Name	Text	15	Required	lds_placeU.sql,
	cst_sname	Consultant Surname	Text	15	Required	mch_placeU.sql,
	cst_start_date	Consultant Start Date	Date		Required, YYYY-MM-DD	lv_placeU.sql,
	cst_end_date	Consultant End Date	Date		YYYY-MM-DD	PlaceU_placements.xlsx

Table Name: Location_Dimension

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P	location_id					
	loc_name	Location Name	Text	30	Required	
	loc_address	Location Address	Text	30	Full Format (Street, Avenue)	
	loc_post	Location Postcode	Text	8		
	loc_county	Location County	Text	30		PlaceU_locations.xlsx

Table Name: Time_Dimension

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P	time_id					lds_placeU.sql,
	year	Year	Number	4	Required	mch_placeU.sql,
						lv_placeU.sql,
						PlaceU_placements.xlsx

Table Name: Rewards_Fact

Key	Field Name	Caption	Data Type	Field Size	Notes	Source
P/F	account_id		Number			
P/F	placement_id		Number			
P/F	consultant_id		Number			
P/F	location_id		Number			
P/F	time_id		Number			
	no_perm_sec	No of permanent placements secured	Number			lds_placeU.sql,
	no_place_sec	No of placements secured	Number			mch_placeU.sql,
	rev_generated	Amount of revenue generated	Number			lv_placeU.sql,
	no_acc_pay_max	No of account that pay maximum salary	Number			PlaceU_placements.xlsx,
	no_con_over_5	No of consultants worked over 5 years	Number			PlaceU_locations.xlsx

Figure 4: Data Dictionary

Metadata is simply data about the data, the metadata for this study will show the managing director what tables, attributed and keys the data warehouse contains, along with where each data set came from, similar to the data dictionary above. The difference between the metadata and the data dictionary is the metadata will also tell the managing director what transformations were applied with the cleaning of the data as well as how often the data gets reloaded into the data warehouse and how this data has changed over time. (Leeds Beckett University, 2016a)

To ensure data integrity it is important to define the column names and data types along with the declarative constraints, as you can see in the data dictionary the primary and foreign key attributes have been declared along with some additional notes on the declarative constraints for example in the Placement_Dimension table the maximum and actual salary say ‘required’ in the notes section, this is a data

constraint meaning that this field cannot be a null value. (Leeds Beckett University, 2016b)

In the PlaceU_locations.xlsx data file the location_id field is missing, along with the loc_post, as the address and postcode are joint together. During the cleansing process the address field will need to be split to allow a post code field along with the addition of a new field for the id as seen in figure 5 below.

LOC_NAME	LOC_ADDRESS	COUNTY
LEEDS	232 Roundhay Road, LS8 4HT	West Yorkshire
MANCHESTER	13 Rowsley Street, M11 3FF	Greater Manchester
LIVERPOOL	22 Bold Street, L1 4HR	Merseyside

LOCATION_ID	LOC_NAME	LOC_ADDRESS	LOC_POST	LOC_COUNTY
1	LEEDS	232 Roundhay Road	LS8 4HT	West Yorkshire
2	MANCHESTER	13 Rowsley Street	M11 3FF	Greater Manchester
3	LIVERPOOL	22 Bold Street	L1 4HR	Merseyside

Figure 5: PlaceU Locations Cleansing

A further examination of the PlaceU_placements.xlsx data file shows that there are a number of missing values under the plt_to_permanent field. As you can see in the data dictionary this is a required field in order to fulfil the MD's goals and objectives and therefore it cannot have any null values, if for instance a value has not been recorded in this column then it is important to standardise null values as a not available value during the cleaning process in order to ensure the data warehouse does not contain any missing data values.

PLACEMENT_ID	PLT_SHORT_DESC	PLT_REQUIRED_START_DATE	PLT_ESTIMATED_END_DATE	PTL_ACTUAL_START_DATE	PLT_ACTUAL_END_DATE	PLT_RENEWAL_NO	PLT_TO_PERMANENT
127	DBA at Asda	01/06/2013	01/09/2013	01/06/2013	01/09/2013	3	N
150	systems analyst	01/09/2013	01/12/2013	01/09/2013	01/12/2013	4	
165	data analyst	01/06/2011	01/09/2011	01/06/2011	01/09/2011	3	N
182	developer	01/03/2010	01/06/2010	01/03/2010	01/06/2010	2	N
193	developer	01/06/2013	01/09/2013	01/06/2013	01/09/2013	3	N
195	developer	01/01/2014	01/03/2014	01/01/2014	01/03/2014	1	N
204	BI Analyst	01/03/2010	01/06/2010	01/03/2010	01/06/2010	2	N
223	BI Analyst	01/06/2015	01/09/2015	01/06/2015	01/09/2015	3	
121	DBA at Asda	01/06/2011	01/09/2011	01/06/2011	01/09/2011	3	N
123	DBA at Asda	01/06/2012	01/09/2012	01/06/2012	01/09/2012	3	N
131	DBA at Asda	01/06/2014	01/09/2014	01/06/2014	01/09/2014	3	N
139	systems analyst	01/06/2010	01/09/2010	01/06/2010	01/09/2010	3	N
151	systems analyst	01/01/2014	01/03/2014	01/01/2014	01/03/2014	1	N
155	systems analyst	01/01/2015	01/03/2015	01/01/2015	01/03/2015	1	N
164	data analyst	01/03/2011	01/06/2011	01/03/2011	01/06/2011	2	
173	data analyst	01/01/2014	01/03/2014	01/01/2014	01/03/2014	1	N
175	data analyst	01/06/2014	01/09/2014	01/06/2014	01/09/2014	3	N
183	developer	01/06/2010	01/09/2010	01/06/2010	01/09/2010	3	N
187	developer	01/06/2011	01/09/2011	01/06/2011	01/09/2011	3	N

Figure 6: PlaceU Placements Cleansing

Another check which will need to be done to the PlaceU_placements.xlsx, as well as the lds_placeU.sql, mch_placeU.sql and lvp_placeU.sql is a check for outliers within the salary field. An outlier is a result that lies far away from all the other results in a random sample, although they can be valuable to a business, they are usually the result of badly recorded data. In the PlaceU case any outliers within the salary field will be eliminated as one wrongly inputted salary could alter the reports in such a significant way, that the businesses decisions made from these reports could be potentially hazardous. (Nist Samatech, 2013)

Article IV. Bibliography

ACCA (2013) **Reward schemes for employees and management**. [Online] Available from: <http://www.accaglobal.com/content/dam/acca/global/PDF-students/2012s/sa_jan13_p5_reward_a.pdf> [Accessed 26th February 2016].

BBC (2015) **Data errors implicated innocent people – watchdog**. [Online] Available from: <http://www.bbc.co.uk/news/uk-33556221> [Accessed 4th March 2016].

Chenoweth, T, Corral, K and Demirkan, H (2006) **Seven Key Interventions for Data Warehouse Success**. [Online] Available from: <http://dss.gusconstan.com/DSS/documents/p114-chenoweth.pdf> [Accessed 26th February 2016].

Data Integration (2016) **ETL (Extract-Transform-Load).** [Online] Available from: <http://www.dataintegration.info/etl> [Accessed 2nd March 2016].

George, S (2016) **Inmon vs. Kimball: Which approach is suitable for your data warehouse?** [Online] Available from: <http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse> [Accessed 26th February 2016].

Intel IT (2013) **Using a Multiple Data Warehouse Strategy to Improve BI Analytics**. [Online] Available from: <http://www.intel.co.uk/content/dam/www/public/us/en/documents/white-papers/using-a-multiple-data-warehouse-strategy-to-improve-bi-analytics.pdf> [Accessed 12th February 2016].

Leeds Beckett University (2016a) **Data Have Data – The Metadata**. [Online] Available from: <https://my.leedsbeckett.ac.uk/bbcswebdav/pid-1137662-dt-content-rid-3381636_1/xid-3381636_1> [Accessed 2nd March 2016].

Leeds Beckett University (2016b) **Data Quality & Integrity; and Data Maintenance**. [Online] Available from: <https://my.leedsbeckett.ac.uk/bbcswebdav/pid-1137662-dt-content-rid-3684855_1/xid-3684855_1> [Accessed 2nd March 2016].

Linstedt, D (2004) **Data Warehousing Ethical Concerns: Security, Access and Control**. [Online] Available from: <http://tdan.com/data-warehousing-ethical-concerns-security-access-and-control/5186> [Accessed 2nd March 2016].

Maynard, W (2010) **The Four Strategies to Generate More Revenue for Your Company**. [Online] Available from: <http://www.kinesisinc.com/the-four-strategies-to-generate-more-revenue-for-your-company/> [Accessed 12th February 2016].

NHS (2016) **Top Tips to Avoid Recruitment Upload Discrepancies: A Guide for Recruitment Data Contacts**. [Online] Available from: <https://www.crn.nihr.ac.uk/wp-content/uploads/Funders%20academics/Portfolio%20userguides/8.%20Top%20Tips%20RDCs.pdf> [Accessed 4th March 2016].

Nist Samatech (2013) **what are outliers in the data?** [Online] Available from: <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm > [Accessed 4th March 2016].

Perkins, A (2003) **A Strategic Approach to Data Warehouse Engineering**. [Online] Available from: <http://www.visible.com/Company/whitepapers/dw.pdf> [Accessed 12th February 2016].

Simms, B (2012) **Data Dictionary Requirements**. [Online] Available from: <http://community.miamioh.edu/institutionalanalytics/sites/edu.institutionalanalytics/files/Data%20Dictionary%20Requirements.pdf> [Accessed 2nd March 2016].

Standen, J (2008) **Data Warehouse vs. Data Mart**. [Online] Available from: <http://www.datamartist.com/data-warehouse-vs-data-mart> [Accessed 26th February 2016].

Teradata (2007) **Data Model Overview: Modeling for the Enterprise while Serving the Individual**.

Turban, E, Sharda, E.R and Delen, D (2011) **Decision Support and Business Intelligence Systems**. 9th ed. Prentice Hall.

Wentzlaff, A (2014) **7 challenges to consider when building a data warehouse**. [Online] Available from: <http://www.onapproach.com/7-challenges-consider-building-data-warehouse/> [Accessed 2nd March 2016].

1keydata (2016) **Fact Table Granularity**. [Online] Available from: <http://www.1keydata.com/datawarehousing/fact-table-granularity.html > [Accessed 2nd March 2016].