­­­­­­



**School of Science and Technology**

**BIS3214**

**Data Warehousing and Business Intelligence**

**Autumn/Winter term**

**2014/2015**

|  |
| --- |
| **Date:** 25rd January 2015 |
| **Supervisor:** Dr. Syed Murad |
| **Campus:** Hendon |
| **Group Members:** |

|  |  |  |  |
| --- | --- | --- | --- |
| Cristiano Cardoso Maia  M00322036 | Jade Van Nelson  M00405270 | Kateryna Vyshnyak  M00330646 | Veeraj Bhura  M0041555 |

Table of Contents

[1. Introduction 3](#_Toc409615585)

[1.1 Middlesex University 3](#_Toc409615586)

[1.2 Types of Management 3](#_Toc409615587)

[1.2 Objectives, Values and Mission 4](#_Toc409615588)

[1.3 Marketing Department 4](#_Toc409615589)

[2. Analysis of Operational Systems and Data Sources 5](#_Toc409615590)

[2.1 Entity Relationship Diagram (ERD) description 5](#_Toc409615591)

[2.2 Advantages and disadvantages of using a Data Warehouse 9](#_Toc409615592)

[2.3 Table of Comparison 9](#_Toc409615593)

[3. Data Mart Design 10](#_Toc409615594)

[3.1 Subject areas identification 10](#_Toc409615595)

[Event Analysis 10](#_Toc409615596)

[Expenditures 10](#_Toc409615597)

[3.2 Start Schema Diagrams 11](#_Toc409615598)

[3.3 Granularity 12](#_Toc409615599)

[3.4 Snowflake Schema 12](#_Toc409615600)

[4. ETL process 13](#_Toc409615601)

[Data Extraction, Transformation and Loading (ETL) 13](#_Toc409615602)

[Extraction Methods in DWs 13](#_Toc409615603)

[5. OLAP 14](#_Toc409615604)

[Different OLAP Techniques 15](#_Toc409615605)

[MOLAP 15](#_Toc409615606)

[ROLAP 15](#_Toc409615607)

[HOLAP 15](#_Toc409615608)

[Definitions 15](#_Toc409615609)

[Slicing 15](#_Toc409615610)

[Dicing 16](#_Toc409615611)

[Drill Down/Up 16](#_Toc409615612)

[Implementation 16](#_Toc409615613)

[6. Conclusion 20](#_Toc409615614)

[7. References 20](#_Toc409615615)

# 1. Introduction

## 1.1 Middlesex University

A University in England with roots back to 1878 who has since build up a reputation in London and international as a providing a high level of higher education from its wealth of experience and strategic operational capabilities. Through strengthening our efficiency by employing highly qualified academics, attracting the best ambitious students and strengthening our research reputation we can capitalise internationally on becoming an excellent international tertiary education establishment.



Figure 1 - Reference http://www.bbc.co.uk/bitesize/higher/business\_management/business\_enterprise/decision\_making\_business/revision/1/

## 1.2 Types of Management

The Government Business Reference Model shown above illustrates the three levels of control that the university runs with. All major decisions that has to be made as far as what the university needs to achieve long-term and how they need to achieve it are made at the Strategic level (Senior Management). On the same level you will find job titles such as dean, chancellor, vice chancellor and many other high-ranking job titles as they have the power and knowledge that it takes to make complex decisions.

All the decisions made by the Senior Management team will be passed to the Tactical level of management (Middle Management). This action is important to make sure the whole University is aligned with the same objectives and aims. Middle management staff (such as the head of department) will be responsible to develop a plan to implement the direction given to the managers in the operational bracket.

Finally, the Operational level (junior management) is responsible to implement the overall goal. This level is where professors, librarians and other university staff that deals directly with the students, will take the instructions from the above two management teams. With the training and experience provided by Middlesex they will be able to implement the strategies to turn the strategic plans in reality. They will also be responsible do report their progress directly to the Middle management team.

## 1.2 Objectives, Values and Mission

Our Objectives are to attract students that are determined to succeed and support these students with the correct support and insure student satisfaction by developing policies to maximise their performance. We strive to enhance graduate level employability in an effort to improve our university rankings. We have strategies and policy to organise our schools to enhance excellence.

“Our vision is to become a leading University of choice, recognised internationally for excellence in all that we do.”

## 1.3 Marketing Department

Middlesex University Marketing department is based at the Middlesex University London campus. The department specialises and focuses on each of the five schools across three different campuses universally. The department receive funding for our advertising projects from the relevant departments. The marketing department managers will detail projects and allocate tasks accordingly. This will be inline with equity and diversity strategic strategy ([Equality and Diversity at Middlesex University](https://www.mdx.ac.uk/__data/assets/pdf_file/0009/58707/Equality-and-Diversity-Strategy-2013.pdf) - LINK), University regulations ([University Regulations for the Academic Year 2004/2005](http://www.web.mdx.ac.uk/regulations/archive/regs0405/STS.HTM) - LINK) and decision-making and discussions held at middle management meetings. There after the relevant departments would handle the assignments and would also manage the suppliers and supplies to the project.

Middlesex University will not consider all staff to be participating in one project each staff member will be assigned to a specific project and task in that project. The projects are based on the development of current marketing tools available. E-commerce and online marketing trends are generating popularity and there for more funding is spend on online marketing, thus to insure that the marketing department achieves its aims and goals. These projects are the same projects from the departments of each school and the content is made available in the means of online marketing and various other outlets depending on the projects.

The list of local and international suppliers are available either locally, nationally or internationally. The supplier catalogues and the supplier contracts are saved in the contracts database. A specific enquiry regarding contracts and agreements with the relevant suppliers will be approved by the marketing department and will be inline with the university's framework, mission or scope, the suppliers can provide anything that will physically assist the projects needs. Permanent MDX staff will not be considered for these projects as they are the “clients”. Overall the objectives of the marketing department is to attract students that work hard and willing to go the extra mile to reach success in their studies.

The marketing department is important for the whole University as it influences the direct correlation between the success rate and the rates of new students attending the University. The marketing department is one of the most important business departments of most any company but especially so in this case as the University renders a services namely education.

# 2. Analysis of Operational Systems and Data Sources

## 2.1 Entity Relationship Diagram (ERD) description

In introduction to the following report we will be covering various topics, thus starting with the link between the advantages and disadvantages of operational databases to data warehouses. We will also cover the how the marketing department has developed its own operational database in correlations with an ERD.

The different projects allow the University to meet its objectives by planning events and managing these events. These projects have cost and they are directly related to the project thus a budget is set by each department as to how much would be spend.

The values are met by taking into consideration the long-term effects these projects will have on staff members, attendees and the different attractions held at these events that could generate positive publicity in line with the values of the university.

The mission is to attract the highest amount of high calibre students and by having different projects like open days, digital and social media events, career events and more traditional events at university like research projects the marketing fulfil its mission.

The ERD that we have displayed below shows all the different entities and their attributes. This will allow us to show relationships across all the entities and compose a data mart that will display relevant information so that Middlesex University at the strategic level can make informed choices in how the budget should be used. The ERD will also show the university a cornucopia of other information like what suppliers are used at certain events and what the turnout of the event was. The ERD is vital to the success and longevity of the database and the data mart as if any relationships are missing or any entities have been missed this could result in the error in running queries and retrieving accurate results.

At every event each member of staff plays a crucial role like events management, relations officer from marketing would be there to insure that everything goes according to plan. The member of staff is picked for each event by a few criteria; these criteria are their expertise as to the role that they are required to fill.

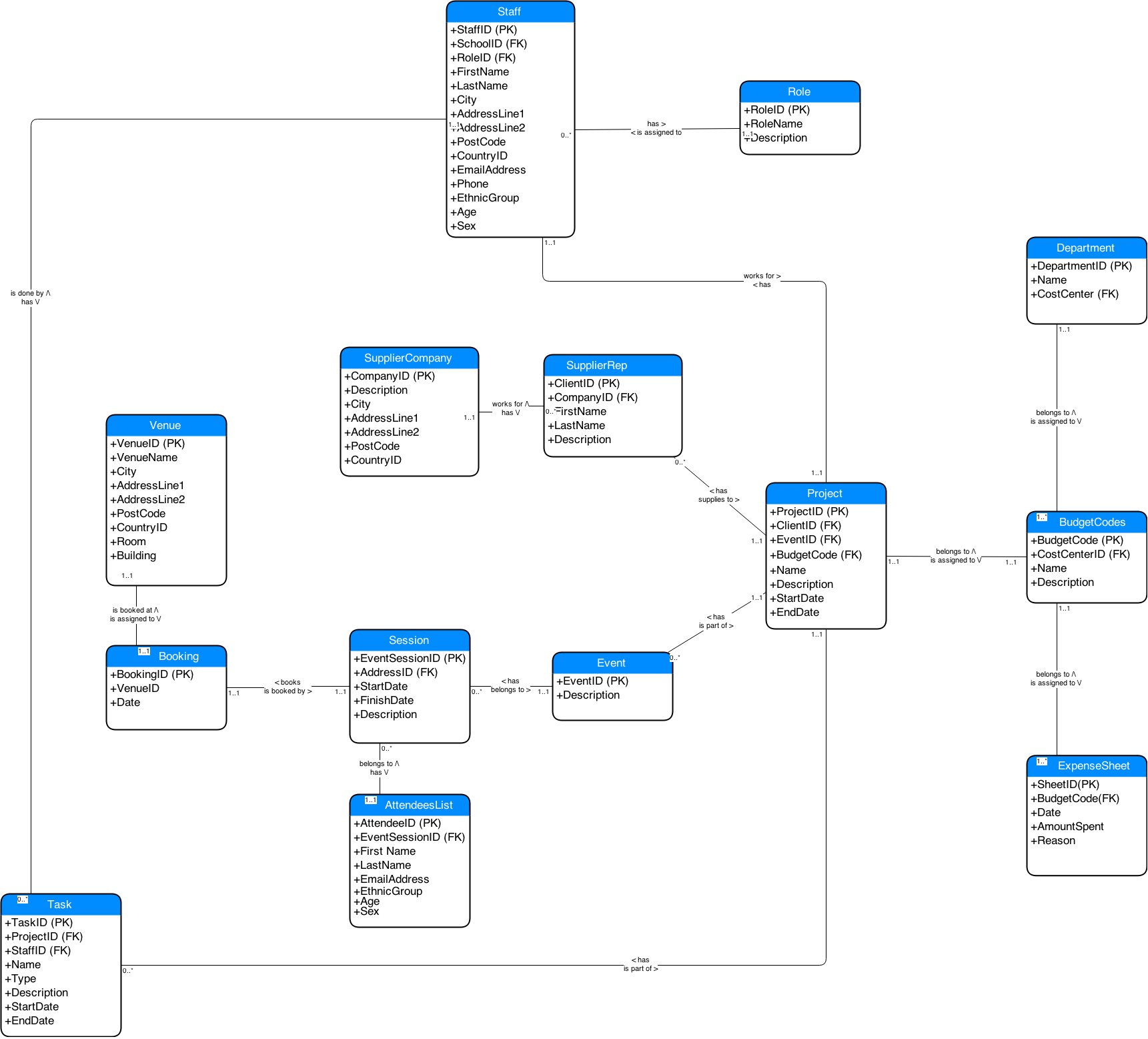
After events and during the course of the year all staff members that are part of the Middlesex University marketing department or are part of an event that the marketing department has held need to be paid for their work.

To keep a track of all the task and how they were accomplished in the past will help us to understand perhaps how to overcome current difficulties but also will be beneficial to justify our expenditure. How we have managed the task indirectly relates to our budget and therefore we could justify requesting more budget for projects as we have specific amount of task that was not completed with other projects and therefore we perhaps under performed as a marketing department

Different venues can be held in different countries thus the university fulfils its mission of attracting international attention as well as focusing on the international students who represent a large amount of the university target audience.

Each project would occur at a venue and to organise the events would have different sessions booked in by booking therefore being able to ascertain the correct amount of space allocated. This takes care of the logistical matters like seating plan, amounts expected to be catered for and falls under general event management.

1. **Role – Staff**
   1. *Cardinality*: 0..\* – 1..1
      1. Each member of staff in the department has a specific role to play.
      2. Each Role has a name and a staff member is added to that Role.
2. **Staff – Tasks – Projects**
   1. *Cardinality*: 1..1 – 0..\* and 0..\* – 1..1
      1. A project can have 0 (in case the tasks hasn't be assigned yet) or many tasks but a specific task can only be related to one project. An extra table called Task History is responsible to keep track of all the tasks from any project even if it has been archived. Along with the keeping track of the staff that was part of the task.
3. **Venue – Booking**
   1. *Cardinality*: 1..1 – 1..1
      1. The marketing department has a list addresses of venues names and locations. These venues are booked by bookings and each session is part of an event.
4. **Booking – Sessions**
   1. *Cardinality*: 1..1 – 1..1
      1. The sessions is booked by the bookings.
      2. A booking has only one session.
5. **Attendee list – Sessions**
   1. *Cardinality*: 0..\* – 1..1
      1. For every session there is an attendee list so that the number of attendees can be recorded for after analysis.
      2. One attendee list belongs to only one session.
6. **Events – Projects**
   1. *Cardinality*: 0..\* – 1..1
      1. An event can have 0 (if the event is online) or many to many projects. Events are optional for each Project but one event can only be related to one Project at time.
7. **Supplier Rep – Supplier Company**
   1. *Cardinality*: 1..1 – 0..\*
      1. Not all the supplier companies have a supplier representative, but all the supplier representatives belongs to one, and only one supplier company.
8. **Supplier Rep – Project**
   1. *Cardinality*: 0..\* – 1..1
      1. Not all projects will need to have external suppliers while other suppliers will have many of them. A supplier representative will be supplying to only one project at time.
9. **BudgetCodes – Department**
   1. *Cardinality*: 1..1 – 1..\*
      1. Each department will create different budget codes as each one of them will be assigned to a specific project. One budget code is directly linked with one specific department.
10. **BudgetCodes- Projects**
    1. *Cardinality*: 1..1 – 1..1
       1. All the projects will have a budget code but each project can only receive budget from one department.
11. **Expense sheets - Budget Codes**
    1. *Cardinality*: 1..\* – 1..1
       1. The project budget is controlled by the analysing the Expense Sheets. Each Budget Code will have many Expense Sheets associated with it but one sheet can only be associated with one budget code.



## 2.2 Advantages and disadvantages of using a Data Warehouse

Database is structured data gathering and data processing model. There are two main forms of databases, operational databases and analytical databases. This chapter will define each one of them and state their advantages and disadvantages.

The Operational Databases are company’s core transaction processing (OLTP) scheme that collect, modify and sustain data on a daily basis. The same type of database is also responsible to process and manage dynamic flow of real-time data. It allows access to the archived data and permits data modification of this data such as deletion, adding or updating.

Disadvantages of Operational Databases (OD) that they designed for data entry purposes only. Online quires and analytics must be simple as operational database has “relational” not “dimensional”. Therefore, OD is unable to quickly consolidate, cleanse and integrate data from multiple, disparate databases that run on different technical platforms in different geographical locations. Thus result in poor business intelligence results in poor strategic and tactical decision making

Operational database is designed to process and manage dynamic flow of real-time data. It allows access to the archived data and permits data modification of this data such as deletion, adding or updating. The significant difference between operational and warehouse systems are that operational system designed to assist with transaction process whereas data warehousing oriented for online analytical process (OLAP) (Exforsys). Therefore usage of data is optimized for different purposes, see Table 1 below.

## 2.3 Table of Comparison

|  |  |
| --- | --- |
| **Operational Database** | **Data Warehouse** |
| Provides process-oriented or process-driven analysis only around defined processes of the business and its tasks. | Provides subject-oriented analysis with main focus on the business area where company requires strategic decisions. The collected information is concise about particular issue and collected from more than one operational sources. The data analysis produce information about short-range strategies its assessment and evaluations. |
| Dealing with current data and this data updated on the regular basis. | Dealing with historical data and this data rarely changed. Data warehouse stores information as non-volatile and read only format. This data is loaded on the regular basis and growing constantly. |
| System optimized for fast data uploading but small content at the time. | System optimized for fast data retrieval and large content at the time. |
| Data is specific to application performance; therefore it is non-integrated or partially integrated causing data redundancy. | Controlling data redundancy problems within DBSM integration layer. |
| Doesn’t require expert level of computing skills to navigate system. | Requires advance computing knowledge to navigate system. |
| Supporting high-volume transaction performance with minimum rear reporting. | Supporting high-volume analytical transactions with maximum reporting that promptly used for making strategic decisions. |

Table Operational system VS data warehouse (Rensselaer).

Data Warehouse is computerised system for storing information. This information helps organisation to analyse historical patterns and make important business decisions. The advantage of having large repository of information that it helps to solve number of problems, increase company profitability while and reduce cost to access this historical data within external sources. Furthermore, the data is consistent, relevant, structured and combined from various locations into one centralised location. This data centralisation helps to find multiple solutions than where data analysed separately. Data Warehouse allows storing already retrieved data within operational level that improves turnaround time of data reporting and analysis (Power).

The advantage of having data warehouse in marketing is that the data is static and provides a "single version" of the truth about enterprise activities. To understand better its customers – students, future prospects on the marketplace and event’s opportunities to attract more students – customers, researchers and funds.

However there are a number of disadvantages that need to be addressed. Firstly, data warehouse is very expensive to maintain, as data itself must be normalised, loaded and extracted. Moreover, company have to train their users, otherwise it could lead to the security problems while conducting any queries via online access. Secondly, there is always a chance that new transaction system may not be compatible with the currently used systems. There are also a numerous ways of storing information in the data warehouse and applying one set of rules might not be beneficial if in the future company decides to change the way it conducts business (Exforsys).A Data Warehouse is also very slow.

# 3. Data Mart Design

## 3.1 Subject areas identification

### Event Analysis

One of the objectives and keeping in line with the University strategy is to increase the amount of students who would actually attend the University. This can be directly measured by to the success rate of the events held at University. Keeping track of the amount of events and the most successful event types at each location over time will allow us valuable intelligence has to how to plan for future marketing events.

### Expenditures

The projects themselves will be held at a location over a specific time period. This fact table allows user to compare the financial side of different projects and filter it by time and/or location. The user can also compare the total project budget per school or type of events for examples.

This fact table allows the user to compare the financial side of different projects and filter it by time and/or location.

**Relevant business questions that can be answered with this data mart:**

\* What percentage of total expenses does a project incur?

\* Are expenses rising? By how much?

\* What are typical expenses for a project of a certain type?

\* Are there any expense deviations among projects within same School?

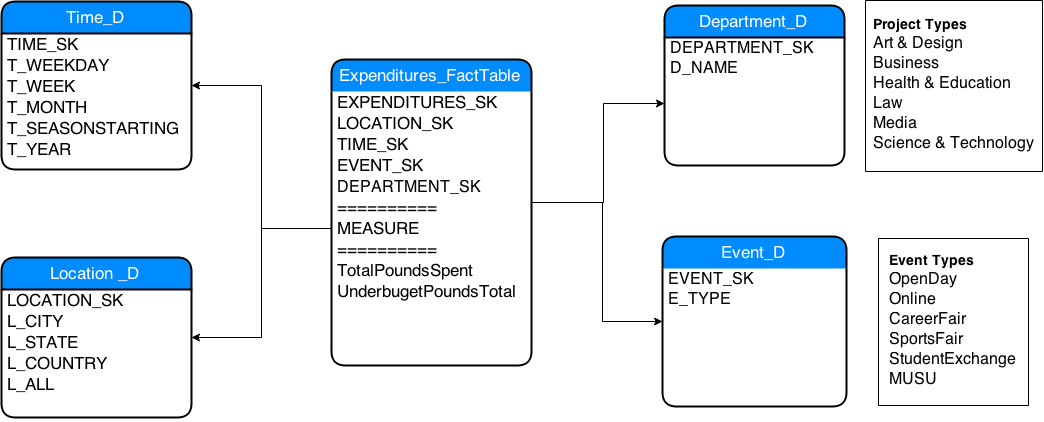
\* Are certain consultants abusing expenses?

\* What type of project incurs the highest expenses?

\* For which categories of expenses do we need to control better or plan?

## 3.2 Start Schema Diagrams

MUST UPDATE THIS DIAGRAMS WITH LATEST VERSION – ADD CAPTION



## 3.3 Granularity

The information stored and composed in a fact table can distinguish the granularity of the table by the extend of the amount of subdivide information added. The less there are the fewer pieces of distinct information means that there will be less granularity. An example would be the Location\_D with its 5 fields Venue, City, State, Country and All.

The granularity is important as it allows for the types of queries that can be queried. If there are few granularities then it means that there are less fields of information and if the granularity is high then it means that there are a lot of fields and thus we can query in a more intelligent way. Not always it’s necessary to have all the possible combinations of granularity for example on the Time dimension we are not interested in knowing how much money was spend per hour.

A fact table is the summary of the information gathered from all the other dimensions. The fact table can be filtered by location, time and any other relevant attributes from any dimension table.

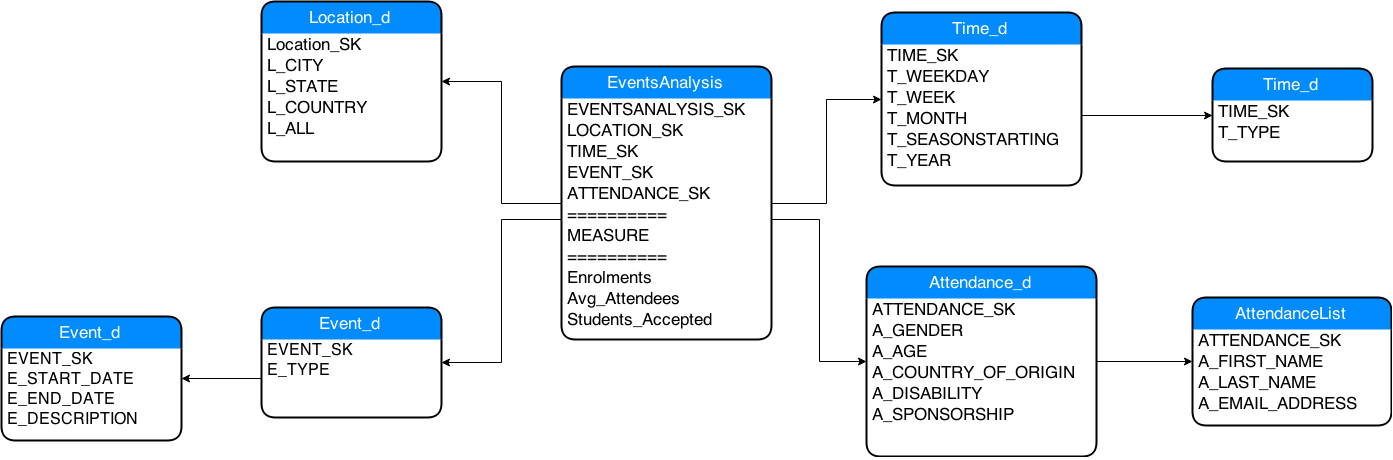
The numeric values in fact tables are divided into three main categories:

* Additive: These measures can be summed across any dimension associated with that fact table.
* Semi Additive: These can sum over some dimensions but not all and the budget balance amounts are semi addictive as there are not over time.
* Non – Additive: ??

## 3.4 Snowflake Schema

A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.

There are differences between a snowflake scheme and a start schema. A star schema is good for a data mart where as a snowflake schema is good for a data warehouse. A star schema only consist out of single dimension tables whereas snowflake schema would consist of more dimensions as shown below. A Top down approach is taken with start schema where as a bottom up is taken with snowflake schema.



# 4. ETL process

## Data Extraction, Transformation and Loading (ETL)

According to G. Goos et all., (2003) Data Warehouse (DWs) purpose is to process and facilitate decision making processes. Therefore, the Extraction-Transformation-Loading (ELT) functionality is important to facilitate any query within complex computer systems. ETL processes are liable to extract data from operational data sources, transformation of these data, including normalization, and loading clean data back to DWs. That process is crucial component of DWs data flow as incorrect or misinterpreted data will lead to the inaccurate business decision, therefore data quality must be observed at its early stage of loading. The design of ETL processes structured in 6 tasks:

1. **SOURCE EXTRACTION**

Includes extraction of data from many operational systems. To obtain desired dataset more data might be extracted, the size of it depends on actual system and business requirements.

1. **TRANSFORM THE SOURCES**

Includes processes of filtering to required dataset. The filtering tasks will include calculating values originated values, converting between different data formats and codes, sequence numbers automatic generation, elimination of duplicates and eliminating unwanted information. The most complex and time-consuming transformations are Multistage Data Transformation and Pipelined Data Transformation.

1. **JOIN THE SOURCES**

Includes merging different operational sources for unique data loading.

1. **SELECT TARGET TO LOAD**

Includes selection process of the target that must be loaded into the system.

1. **MAP SOURCE ATTRIBUTES TO TARGET ATTRIBUTES**

Includes mapping process of extracted attributes to the corresponding target elements.

1. **DATA LOAD**

Incudes upload of transformed data into DWs.

The vital stage of ELT process is data cleaning with identifying and eliminating errors and data discrepancies in order to increase data quality. According to Goos G, Hartmanis (2003) “Data quality problems are very significant: it has been estimated that poor quality customer data cost U.S. businesses $611 billion a year in postage, printing, and staff overhead.” Therefore the conceptual modeling of ETL processes is beneficial for running and maintaining Data Warehouse.

## Extraction Methods in DWs

The extraction method depends on the source system, warehouse environment and business needs, and this influences the data source, the transportation process and the time need it to update Warehouse. Generally speaking the data for extraction might be badly documented and therefore extraction will be acquired several times. The source system cannot be modified to accommodate needs of the data warehouse extraction processes; therefore two types of extraction identified such as Logical and Physical.

The **Logical extraction** method used when there is no possibility to add additional logic to the source systems to improve its performance. There are two types of logical extraction:

* **Full Extraction** exports currently available data on the source system. An example for a full extraction may be an export file of a distinct table or a remote SQL statement scanning the complete source table.
* **Incremental Extraction** exports data that has been changed since specific time event. This information can be recorded by source system reflecting to the date stamps or by change- capture mechanism. Instead of exporting entire tables it captures and exports only areas with latest implementation. This method is particularly useful on big chunks of data.

The **Physical Extraction** method depends what logical extraction method was applied, also on the source system effectiveness and limitations. There are two mechanisms of physical extraction online from the source system or from offline structure.

* **Online extraction** exports data from the source system or from the intermediate system that stores data in log format. The extracted information must be determined whether it is extracted from the original entity or prepared entity.
* **Offline extraction** exports data outside of original data source system. This data available in redo logs, archive logs or transportable table-spaces or was created by previous extraction routine. There are several structures to be aware of:
  + **Flat files** (contains generic format of outlined data, requires additional information for further managing),
  + **Dump** files (Oracle distinct files, outlined data might be not present),
  + **Redo and archive logs** (information stored in additional dump file),
  + **Transportable table-spaces** (used to extract and transport large volumes of data between Oracle databases), (Oracle, 2014).

# 5. OLAP

To analyse Middlesex University’s data mart and the information held within the team has chosen to use Online Analytical Processing (OLAP) technology, this technology allows the university to run queries on the data that would normally take far too long to run manually and be too complex. In addition, the university will also be able to gather information that can be utilized in strategic planning, answering various questions and business problems.

OLAP is also referred as “Data Cube” and allows access to older and current data so that users can execute analytical queries. OLAP cubes normally have 3 visual dimensions, but in theory there isn’t any official limit of dimensions. OLAP’s approach main focus is acquiring, consolidating, and condense the enormous amount of data records that are produces. There are many different types of operations that will be covered in the next few paragraphs that allow the technology to analyse the data, and then organizes the data in user-friendly graphical representations so that it can be easily understood.

## Different OLAP Techniques

### MOLAP

MOLAP or multi-dimensional OLAP is one of two techniques that are most widely used to analyse multidimensional cubes that house data. In comparison to the other technique used MOLAP differs slightly in the sense that in some application not all MOLAP requires pre-computed and allocation of information into the cube that it is managing. With this being said this leads to one of the advantages of MOLAP, which is that it allows the data to be queried faster because the data is already optimized. MOLAP however is a very expensive technology; this is one of its disadvantages.

### ROLAP

ROLAP or Relational OLAP is significantly different from MOLAP. In the sense that where MOLAP requires pre-computed and stored information to analyse the data cubes, ROLAP does not. ROLAP is generally used for large data sets as its performance it much slower than MOLAP. ROLAP access data in relational databases and then uses SQL queries to calculate, this is also only performed when the user requires the information.

### HOLAP

Hybrid OLAP (HOLAP) is a combination of the above three OLAP, ROLAP, and MOLAP. HOLAP stores the relational database in the ROLAP format, whilst keeping the aggregated data in a multidimensional table in the MOLAP format. This then allows a faster processing time for queries and on the other hand a fast response time for querying detailed data in ROLAP at a fast response rate. The advantage of HOLAP is smaller cubes for faster response times.

## Definitions

### Slicing

Slicing allows a user to take a multidimensional view and turn it into a 2d view that can show data along one slice or side of a layer. This allows the users to go through specific cubes of data along that slice for detailed analysis.

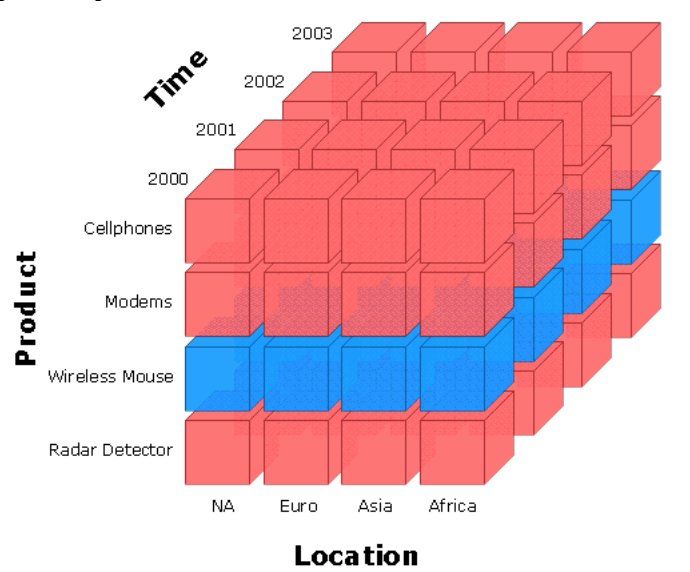
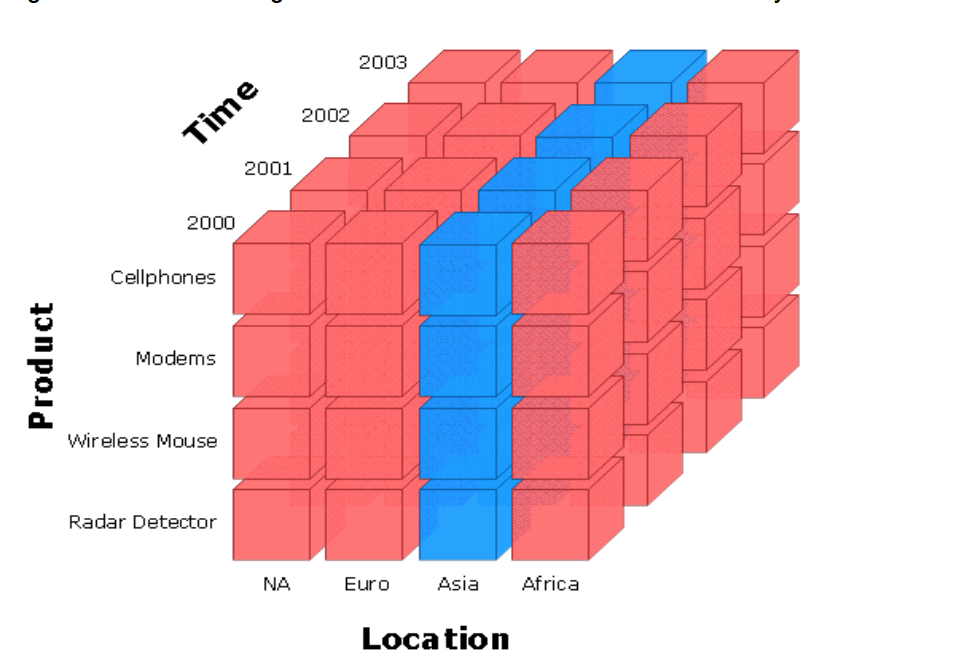


Figure - This picture shows a vertical Slice

Figure - This picture shows a horizontal slice

### Dicing

Dicing is very similar to slicing where is allows the user to examine a row of data, but dicing differs in the sense where slicing allows you to look down a slice of data. Dicing allows the user to look at much more detail and look at every cube or cubes in any row or column.

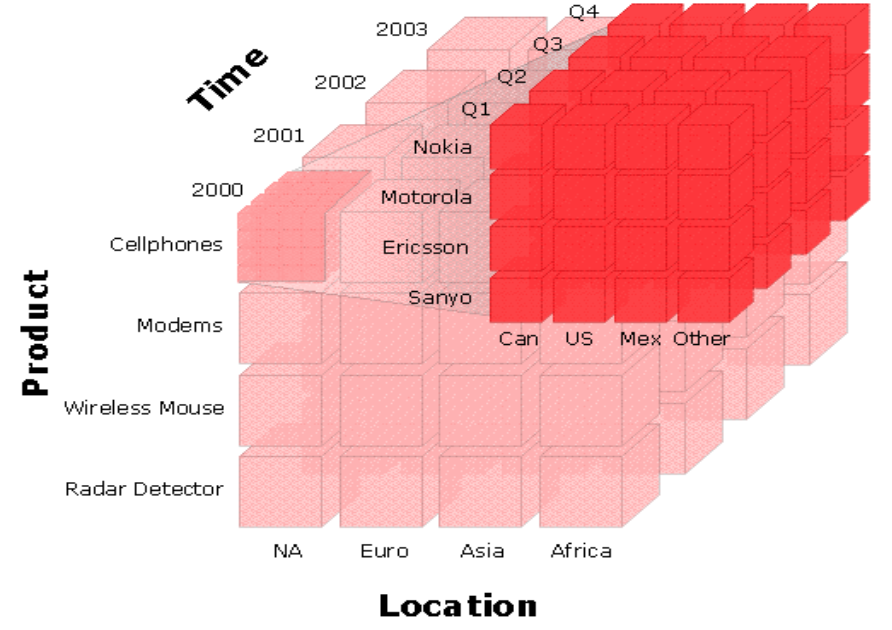


Figure - The above diagram shows dicing

### Drill Down/Up

To drill up/down is an analytic technique where the user will drill up to find data that is most summarized and drill down for data most detailed.

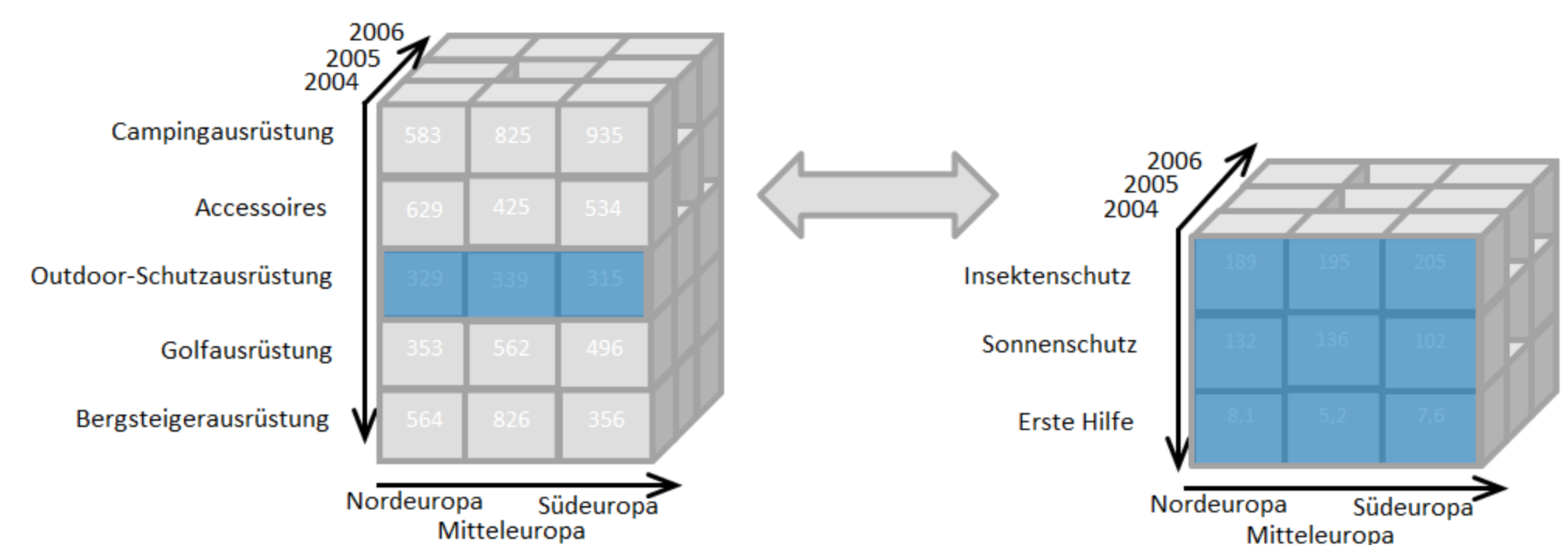


Figure - The above diagram show-drilling up/down respectively

## Implementation

**Questions structure:**

* English question
* OLAP Type
* Query
* Summary

**Question 4**

Students assigned to a group, should discuss and analyse the case and draw a star schema based on this. They also should be able to demonstrate their understanding the dimensional modelling and converting them into physical model. At this stage they don’t require to implement the tables in Oracle but they should be prepared with adequate understanding how that can be done. Week 6/7 case is based on this activity.

Discuss and analyse the case/scripts below:

Dimensional data modeling starts with a *fact table*. This is where we record what happened, e.g., someone bought a Diet Coke in East Fishkill. What you want in the fact table are facts about the sale, ideally ones that are numeric, continuously valued, and additive. The last two properties are important because typical fact tables grow to a billion rows or more. People will be much happier looking at sums or averages than detail. An important decision to make is the granularity of the fact table. If Walmart doesn't care about whether or not a Diet Coke was sold at 10:31 AM or 10:33 AM, recording each sale individually in the fact table is too granular. CPU time, disk bandwidth, and disk space will be needlessly consumed. Let's aggregate all the sales of any particular product in one store on a per-day basis. So we will only have one row in the fact table recording that 200 cans of Diet Coke were sold in East Fishkill on November 30, even if those 200 cans were sold at 113 different times to 113 different customers.

create table sales\_fact (

sales\_date date not null,

product\_id integer,

store\_id integer,

unit\_sales integer,

dollar\_sales number

);

we can pull together this table with a query JOINing the sales, products, and product\_prices (to fill the dollar\_sales column) tables. This JOIN will group by product\_id, store\_id, and the truncated date\_time\_of\_sale. Constructing this query will require a professional programmer but keep in mind that this work only need be done once. The marketing experts who will be using the data warehouse will be querying from the sales\_fact table.

In building just this one table, we've already made life easier for marketing. Suppose they want total dollar sales by product. In the OLTP data model this would have required tangling with the product\_prices table and its different prices for the same product on different days. With the sales fact table, the query is simple:

select product\_id, sum(dollar\_sales)

from sales\_fact

group by product\_id

We have a *fact table*. In a dimensional data warehouse there will always be just one of these. All of the other tables will define the *dimensions*. Each dimension contains extra information about the facts, usually in a human-readable text string that can go directly into a report. For example, let us define the time dimension:

create table time\_dimension (

time\_key integer primary key,

-- just to make it a little easier to work with; this is

-- midnight (TRUNC) of the date in question

oracle\_date date not null,

day\_of\_week varchar(9) not null, -- 'Monday', 'Tuesday'...

day\_number\_in\_month integer not null, -- 1 to 31

day\_number\_overall integer not null, -- days from the epoch (first day is 1)

week\_number\_in\_year integer not null, -- 1 to 52

week\_number\_overall integer not null, -- weeks start on Sunday

month integer not null, -- 1 to 12

month\_number\_overall integer not null,

quarter integer not null, -- 1 to 4

fiscal\_period varchar(10),

holiday\_flag char(1) default 'f' check (holiday\_flag in ('t', 'f')),

weekday\_flag char(1) default 'f' check (weekday\_flag in ('t', 'f')),

season varchar(50),

event varchar(50)

);

Why is it useful to define a time dimension? If we keep the date of the sales fact as an Oracle date column, it is still just about as painless as ever to ask for holiday versus non-holiday sales. We need to know about the existence of the holiday\_map table and how to use it. Suppose we redefine the fact table as follows:

create table sales\_fact (

**time\_key integer not null references time\_dimension**,

product\_id integer,

store\_id integer,

unit\_sales integer,

dollar\_sales number

);

Instead of storing an Oracle date in the fact table, we're keeping an integer key pointing to an entry in the time dimension. The time dimension stores, for each day, the following information:

* whether or not the day was a holiday
* into which fiscal period this day fell
* whether or not the day was part of the "Christmas season" or not

If we want a report of sales by season, the query is straightforward:

select td.season, sum(f.dollar\_sales)

from sales\_fact f, time\_dimension td

where f.time\_key = td.time\_key

group by td.season

If we want to get a report of sales by fiscal quarter or sales by day of week, the SQL is structurally identical to the above. If we want to get a report of sales by manufacturer, however, we realize that we need another dimension: *product*. Instead of storing the product\_id that references the OLTP products table, much better to use a synthetic product key that references a product dimension where data from the OLTP products, product\_categories, and manufacturers tables are aggregated.

Since we are Walmart, a multi-store chain, we will want a *stores* dimension. This table will aggregate information from the stores and cities tables in the OLTP system. Here is how we would define the stores dimension in an Oracle table:

create table stores\_dimension (

stores\_key integer primary key,

name varchar(100),

city varchar(100),

county varchar(100),

state varchar(100),

zip\_code varchar(100),

date\_opened date,

date\_remodeled date,

-- 'small', 'medium', 'large', or 'super'

store\_size varchar(100),

...

);

This new dimension gives us the opportunity to compare sales for large versus small stores, for new and old ones, and for stores in different regions. We can aggregate sales by geographical region, starting at the state level and drilling down to county, city, or ZIP code. Here is how we'd query for sales by city:

select sd.city, sum(f.dollar\_sales)

from sales\_fact f, stores\_dimension sd

where f.stores\_key = sd.stores\_key

group by sd.city

Dimensions can be combined. To report sales by city on a quarter-by-quarter basis, we would use the following query:

select sd.city, **td.fiscal\_period**, sum(f.dollar\_sales)

from sales\_fact f, stores\_dimension sd, **time\_dimension td**

where f.stores\_key = sd.stores\_key

**and f.time\_key = td.time\_key**

group by sd.stores\_key, **td.fiscal\_period**

# 6. Conclusion

What challenges do we have to keep it up and running?

Performance? Amount of data stored. Flexibility to adapt/include/remove different columns.

Add Van’s notes from lecture

# 7. References

Exforsys, Advantages and Disadvantages to Using a Data Warehouse | IT Training and Consulting – Exforsys. Available at: http://www.exforsys.com/tutorials/data-warehousing/advantages-and-disadvantages-to-using-a-data-warehouse.html [Accessed November 14, 2014].

Power, D., Ask Dan! about DSS - What are advantages and disadvantages of data warehouses? Available at: http://dssresources.com/faq/index.php?action=artikel&id=180 [Accessed November 14, 2014].

Rensselaer, Operational Systems vs. Data Warehousing. Available at: http://www.rpi.edu/datawarehouse/dw-about-cmp.html [Accessed November 25, 2014].

Goos G, Hartmanis, J. and van L.J., 2003. Conceptual Modeling-ER 2003. *Springer-Verlag Berlin Heidelberg*. Available at: http://download.springer.com/static/pdf/54/bok%3A978-3-540-39648-2.pdf?auth66=1421251779\_b8d74ae212dcb392186662a364036b7b&ext=.pdf [Accessed January 14, 2015].

Oracle, 2014. Database Data Warehousing Guide. *Oracle Database Online Documentation*. Available at: http://docs.oracle.com/cd/B19306\_01/server.102/b14223/extract.htm [Accessed January 19, 2015].

http://www.hypertextbookshop.com/dataminingbook/public\_version/contents/chapters/chapter003/section004/blue/page004.html