National College of Ireland

Data Warehousing and Business Intelligence Project

on

Insert Title Here

Forename Surname
1234567

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Forename Surname
Student ID:	1234567
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Insert Title Here

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	October 26, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Insert Title Here

Forename Surname
1234567

October 26, 2018

Abstract

Abstract goes here. You should provide a high-level (approx. 150 – 250 words) overview of your project, its motivation, and the core objectives / business requirements addressed.

1 Introduction

In this section you need to motivate your project (using citations). Why are you doing it? You should articulate the business requirements that your project seeks to address as motivation. These should be quantifiable and addressed via the BI Queries in Section 7.

Lorem ipsum dolor sit amet, ut veri deleniti eloquentiam sea Feng & Buyya (2016). Ea commodo aperiam complectitur pri, usu et case dolore. Kune et al. (2016) ad quidam regione percipitur, est ut possit bonorum persecuti. Quis utinam offendit eu usu, eu accumsan disputando per, id cibo reprehendunt sit Beloglazov & Buyya (2015), Gomes et al. (2015). In melius legendos corrumpit pro. Eos dico dignissim voluptatibus et, duo nisl cibo ut. Diceret periculis posidonium cum eu. Gomes et al. (2015) regione nam ex. Vix id viris phaedrum. Pri augue cetero probatus ut.

(Req-1) my first requirement

(Req-2) my second requirement

(Req-3) ...

2 Data Sources

Here you should present and formally describe your sources of data used in the project.

2.1 Source 1: Kaggle

The kaggle movies dataset downloaded this source from: `https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv` provides 20 columns of information on 5000 movies.

These are ....

However, relevant to this project are: ...

This dataset addresses the business requirements listed in Section 1 in the following ways ...

Source	Type	Brief Summary
Kaggle	Structured	Used because I was too lazy to get a difficult dataset
Data.world	Structured	Basically provides the same data as the above, but I'm hoping no one will notice
Twitter	Unstructured	I didn't know what else to do and can't think for myself.

Table 2: Summary of sources of data used in the project

2.2 Source 2: ...

...

2.3 Source 3: ...

...

3 Related Work

In this section, discuss related work on your topic of choice. Consider answering the following questions:

Q1 How have these (or similar) datasets been used before?

Q2 What is generally known about the domain within which your requirements (in Section 1) are situated?

Q3 What significant results exist in this area, and how to you expect to add to them by undertaking this project?

See Hall et al. (2018) for an example of a lit review that looks at specific challenges and approaches within a given area.

4 Data Model

In this section, provide details of your dimensions, why they are in your data model, how sources of data contribute to each of the dimensions, and present your star schema.

Do not do this:

So i have 4 dimensions, as seen in Figure 1. They are good and answer all the queries. I was lazy, so i took a screen shot of SSAS, which my lecturer cannot read.

Instead, provide a detailed discussion on the composition of each dimension, justifying its contents and any hierarchies you have designed in to facilitate drill down. It's seriously unlikely that you can map a dataset 1 to 1 with all columns to a dimension, so explain here which components of which dataset map to which dimension. For the fact table, be precise with respect to the granularity of the facts, and motivate each fact included linking it to at least one of your requirements.

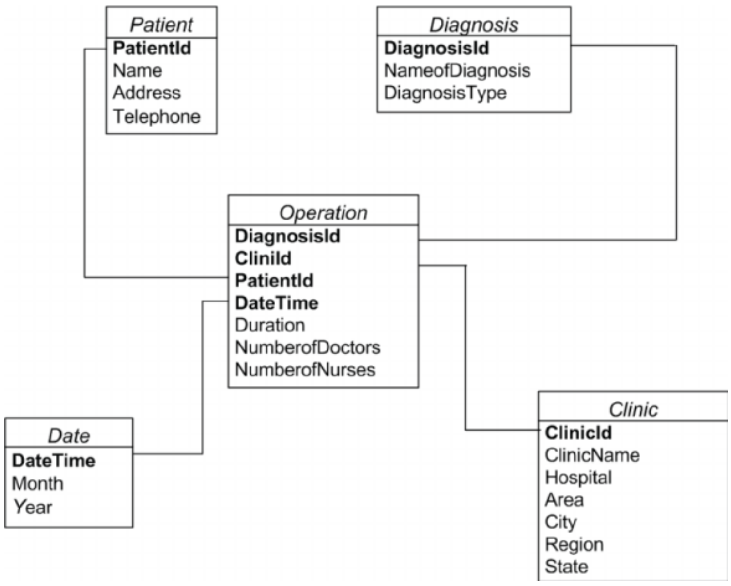

Figure 1: Some star schema i found online

5 Logical Data Map

In this section, describe your logical data map, i.e. how every row of every data source is handled such that it is a part of your star schema.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	Column	Destination	Column	Type	Transformation
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Movie Genre	DimMovie	Genre	Dimension	Primary Genre (1st one) only, all characters after first comma removed
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $

6 ETL Process

Describe the high-level strategy of the ETL process, very specific details can be highlighted in the video.

Essentially, explain the main SSIS (or similar) flow, noting specifically what key challenges of the data were, how they were overcome, the degree of automation, this section should provide additional key implementation details to the logical data map.

7 Application

Rationale and evaluation approach with respect to addressing the business requirements noted in Section 1, i.e. how have you used the case studies / BI Queries to address and demonstrate the attainment of your business requirements.

7.1 BI Query 1: Which genre has the most active engagement on Twitter

For this query, the contributing sources of data are: ...

The general findings are that ... as illustrated in Figure 2.

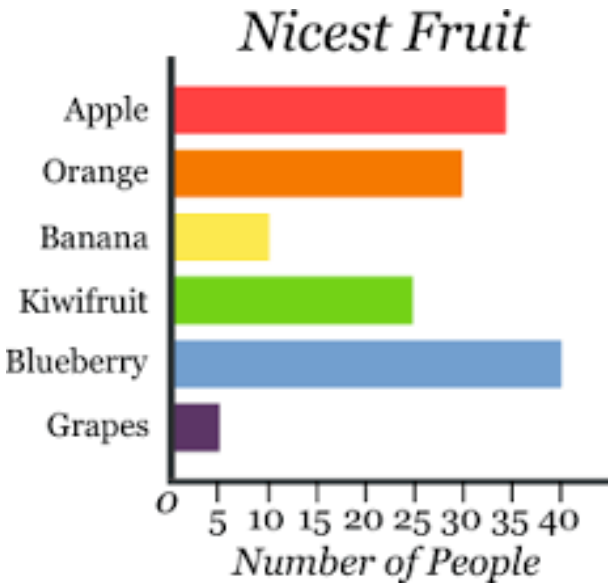

Figure 2: Results for BI Query 1

7.2 BI Query 2: ...

For this query, the contributing sources of data are: ...

The general findings are that ... as illustrated in Figure 2.

7.3 BI Query 3: ...

For this query, the contributing sources of data are: ...

The general findings are that ... as illustrated in Figure 2.

Figure 3: Results for BI Query 2

7.4 Discussion

A detailed discussion / summarisation of the findings from the 3 queries. Note that this discussion will have a lot more detail than the discussion in the following section (Conclusion). You should relate your main findings to the literature that you reviewed in Section 3, i.e. those with a similar topic to your data warehousing project (but which are not necessarily data warehousing projects), and compare and contrast your findings with theirs.

8 Conclusion and Future Work

(Partially) answer your research question and discuss the implications of your (partial) answer, talk about the efficacy of your research, and discuss its limitations.

. . .

Present **MEANINGFUL** future work. Sweeping more parameters in your simulation / model / platform is probably not meaningful. More discuss what could a follow up research project do, to better / differently approach / extend etc. your work.

References

Beloglazov, A. & Buyya, R. (2015), 'Openstack neat: a framework for dynamic and energy-efficient consolidation of virtual machines in openstack clouds', *Concurrency and Computation: Practice and Experience* **27**(5), 1310–1333.

Feng, G. & Buyya, R. (2016), 'Maximum revenue-oriented resource allocation in cloud', *IJGUC* **7**(1), 12–21.

Gomes, D. G., Calheiros, R. N. & Tolosana-Calasanz, R. (2015), 'Introduction to the special issue on cloud computing: Recent developments and challenging issues', *Computers & Electrical Engineering* **42**, 31–32.

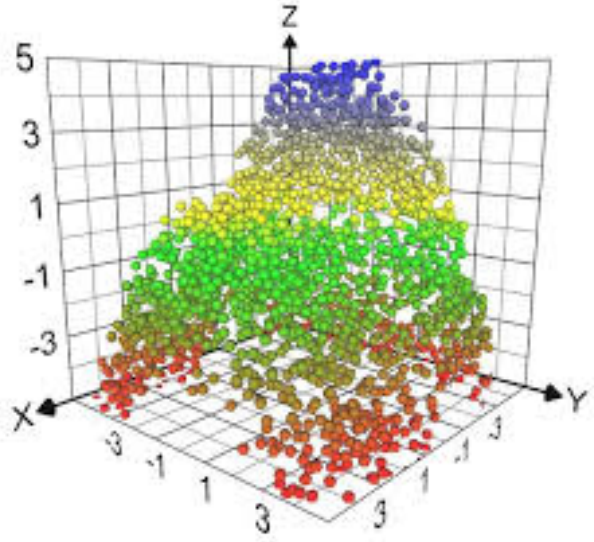

Figure 4: Results for BI Query 3

Hall, M., Mazarakis, A., Chorley, M. & Caton, S. (2018), 'Editorial of the special issue on following user pathways: Key contributions and future directions in cross-platform social media research'.

Kune, R., Konugurthi, P., Agarwal, A., Rao, C. R. & Buyya, R. (2016), 'The anatomy of big data computing', *Softw., Pract. Exper.* **46**(1), 79–105.

Appendix

R code example

```
#Calculate CE for each counterparty
Value.A <- data.frame() #MTM value of each contract within cp A
# ...
for(i in 1:length(foo)){
  if(isTrue(as.character(portfolio_data[i,1])=="A")==TRUE){
    # ...
  }
}
```