Report - Prolongation of Masters Project - MAS-ICT 2016

Simon Byrne

Identifying Value in Big Data Sources for Autodesk Localization Services

Extraction, integration and analysis of various data sources including those emanating from Big data sources for the Autodesk Localization Services department.

Master of Advanced Studies - Information and Communication Technologies

Table of Contents

Abstract 1

Introduction 2

Context of initial project 2

Problem to be addressed 3

Project Management 3

Methodology 3

Calendar 3

Meetings 4

Analysis of technologies and justification of choices 4

Analysis of source data technologies – Hadoop 4

Analysis of choice of technology SQL vs. NoSQL 4

Justification of choice of SQL for the project. 5

Analysis of choice of technology - Microsoft BI Suite 5

Justification of choice of Microsoft BI Suite 5

Architecture 6

Analysis/discussion of results obtained by statistical analysis 6

Recommendations for future development of the project. 6

Conclusion 6

Bibliography 6

Appendixes 7

Appendix A: Procédure de remédiation du travail de master du MAS ICT-4 de M. Simon Byrne 7

Appendix B: Prolongation project calendar 8

Appendix C: Notes from meetings with conseiller 8

Appendix D: Data sources used in original projects data mart. 8

Appendix E: Description of Autodesk’s Localization Services department 9

# Abstract

Executive Summary be written last should say

* The reason for the prolongation project ?
* The work done ?
* The conclusions.?

“The first sentence of an abstract should clearly introduce the topic of the paper so that readers can relate it to other work they are familiar with. However, an analysis of abstracts across a range of fields show that few follow this advice, nor do they take the opportunity to summarize previous work in their second sentence. A central issue is the lack of structure in standard advice on abstract writing, so most authors don’t realize the third sentence should point out the deficiencies of this existing research. To solve this problem, we describe a technique that structures the entire abstract around a set of six sentences, each of which has a specific role, so that by the end of the first four sentences you have introduced the idea fully. This structure then allows you to use the fifth sentence to elaborate a little on the research, explain how it works, and talk about the various ways that you have applied it, for example to teach generations of new graduate students how to write clearly. This technique is helpful because it clarifies your thinking and leads to a final sentence that summarizes why your research matters.”

1. Topic – analysis of technologies uses in previous project with justification of choices.

2. Previous work while delivering functionally did not have in-depth report on tech. Previous studies of technologies don’t look at the specifics of a small departments data mart..

3. Other studies – specifics of an enterprise department in integrating small and large data

4. To address this problem we look at the technology choices made in our data warehouse project, analyse and justify their selections and recommend future developments

5.. Research looks at history , adv , disadv of technologies and compares with nosql solution . – delivers conclustions and recommends future development of the project.

# Introduction

## Context of initial project

Brief introduction – Context and motivation of the original project

The general purpose of the original project was to provide a proof of concept to show that the Localization Services department of Autodesk\* (Appendix E – Description of Autodesk Localization Services department) could harness big data coming from cloud product usage and social media, integrate this data with other internal and external data sources and then analyze it in order to identify strategic opportunities for localizing products. (ref. SimonByrne Mas-ict.pdf - “How though to gather this data – world data, social media data, cloud usage data, and integrate with other internal data sources and then analyze it to identify business opportunities. To provide a proof of concept of the ability to and usefulness of bringing these data sources together and provide it to the Localization Services department of Autodesk is the objective of this project.”).

To this end an SQL based approach was selected for the creation of a data mart to bring together and analyse the data elements in a star schema database structure. An ETL process was implemented to create a data mart using Microsoft Visual Studio SSIS and analysis was subsequently carried out using Microsoft Visual Studio SSAS and Microsoft Excel. The various data elements were taken from different internal and external sources (Appendix D – Data sources used in original project data mart), most notably from the companies internal big data solution - a Hadoop database made accessible to Hive queries through a Hue web interface.

In terms of the analysis of the data once extracted, transformed and loaded, the approach was described in the original report as follows;

“Given the large amount of information available externally and internally on users and those interested in Autodesk products, could we identify trends which would help us to increase revenue. If we took a particular look at language use and geographic locations, an angle which is often ignored by US-centric companies but of immediate interest to a localization department such as ours, could we identify useful correlations and trends?“

The analysis was then structured around the following business questions;

1. Are there potential markets for which no localized version exists?

2. Are there other correlations or tendencies in the behavior and preferences of users of Autodesk cloud products and Autodesk’s social media followers?

The conclusions of the analysis provided the following insights in response.

1. Are there potential markets for which no localized version exists?

* When EPI is taken into consideration. the countries with low EPI seem the most obvious candidates for new localized versions, namely Turkish and Arabic where it seems there is a high level of uptake of cloud services despite no translated versions and lower English profiency.
* For countries with High EPI, Netherlands, the Scandinavian countries it may be there is no need foR translated versions as potentially customers are competent enough to use English versions when available. However this might be an avenue to investigate, could customer experience, satisfaction be enhanced by providing translated versions into these languages?
* This Is borne out when looking at language information when bookings and billings information is added into the mix.. the Scandinavian, Benelux with high EPI appear prominently again, as do low EPI Turkish and Arabic. There are some interesting new languages identified also in Bahasa(Indonesia) and Thai.

2. Are there other correlations or tendencies in the behavior and preferences of users of Autodesk cloud products and Autodesk’s social media followers?

* Tweet count at a location has a moderate to high positive correlation to that locations –Education downloads , Autodesk Homepage visits, cloud product usage (a360, oxygen) and also to revenue bookings and billings. The correlation to cloud usage is particularly marked 0.89 for a360 , 0.76 for oxygen.
* Tweeting has a positive correlation to a countries Global Competitive index statistics
* Education users and downloads have a positive correlation with cloud usage and billings and bookings.
* Higher customer satisfaction with translation quality is positively correlated with akp usage. Quality of translation counts!

The general conclusions of the proof of concept data mart project were as follows;

ref. SimonByrne Mas-ict.pdf - “Ultimately our quest to bring data sources together from internal and external sources and harness big data sources such as data from cloud usage statistics and social media(twitter) has been a successful one. Through application of an Extract Transform and Load process which involved data cleaning, data transformation and data reduction and then loading of our integrated data into a star schema database structure optimized for data mining , we have gone on to create a data cube which allowed us to easily query and analyze the integrated data.

This analysis enabled us to answer the initial business questions posed and threw up also some other unexpected correlations.

In my opinion this Is a rich area to build on. Now that the proof of concept model is available to the department I would propose to integrate other facts from the data sources which were already used here to look for further correlations and to also to look further for other data sources. in particular further social media data which may be available within the organization. the initial exploration of Twitter data in particular demonstrated strong correlations with other measures of business success such as cloud service usage and revenue. “

## Problem to be addressed

While the technical elements delivered with the original project (i.e. the data warehouse and analysis implementation) was accepted by the Examining committee (CEX) - “Aucun développement logiciel (code) nouveau n'est demandé”, the report which accompanied it was evaluated as ‘lightly insufficient’ and a remediation was recommended.. The committee highlighted several areas of improvement that should be addressed in a new report to be delivered in order to fulfill this remediation (Annexe A: Procédure de remédiation du travail de master du MAS ICT-4 de M. Simon Byrne)

In brief the recommendations of the committee were;

• General recommendations – Follow directives delivered for the final report. Adopt a more scientific approach. Present references. Validate results. Be critical and creative.

• Form of report – 20 pages maximum. Avoid too linear a structure. Use references, listed in a bibliography. Avoid repetition. Explain acronyms and concepts. Pay closer attention to detail of presentation in report to avoid spelling mistakes and punctuation errors.

• Content of report – Structure well arguments. Justify importance of contents. Provide list of recommendations for advancing with results of the project. Provide a schema of the general architecture of the solution. Provide justification of technology choices. Review statistical analysis done. Improve graphical representations.

The problem to be addressed and the purpose of this new project will be to provide a report to address the various points

To achieve this the report will focus on providing 5 main areas;

1. An in-depth analysis of the technologies evaluated for the original project
2. An in-depth justification of the technologies chosen for the original project.
3. A detailed description and presentation of the architecture chosen and implemented in the original project.
4. A review of the statistical analysis carried out in the original project with the correction of any errors and an improved presentation of statistical conclusions.
5. Recommendations for the future development of the project.

# Project Management

Section describing the planning and management of the report creation project

## Methodology

As no software implementation is required for the prolongation project, no particular project methodology will be applied. A basic high level ‘waterfall’ schedule outlined in the following ‘Calendar’ section calls out the scope, major milestones and deliveries of the project.

## Calendar

The report creation project calendar can be referred to in Appendix B - Prolongation Project calendar.

Principal milestones are.

* 01/06/2016 - Beginning of preparation of cahier des charges
* 12/08/2016 - Finalization of cahier des charges
* 29/08/2016 - Beginning of TM
* 28/10/2016 - Delivery of TM
* 21/11/2016 - Defense of TM

The total scope for the project is 75 hours work - comprising;

* Preparation of cahier des charges = 7 hours
* Preparation of TM = 60 hours
* Delivery of TM = 3 hours
* Preparation of defense = 4 hours
* Defense = 1 hours

## Meetings

Meetings with conseiller professeur Cedric Baudet to discuss the detail and progress of the project will be twice weekly to the following schedule

* Mercredi, 14 septembre  8h30
* Vendredi, 30 septembre 8h30
* Lundi, 10 octobre 8h30
* Vendredi, 21 octobre 8h30

Notes from these meetings are included in Appendix C – Notes from meeting with conseiller

# Analysis of technologies and justification of choices

Analysis of source data technologies – Hadoop

The data used in the project for Autodesk Cloud usage was extracted form a Hadoop database using Hive Queries. An important element of the report will be to explain these technologies and why they are used as Big Data Solutions.

##### Recap on core concepts - Big Data and Cloud computing

Before beginning our study of Hadoop technology, we should quickly recap on the core concepts of Big data and cloud computing, for it is these concepts as we will demonstrate that have driven the emergence of Haoop as a leading technology.

##### Big data

In their report “The rise of “big data” on cloud computing: Review and open research issues “ Ibrahim Abaker Targio Hashem, IbrarYaqoob , NorBadrulAnuar , Salimah Mokhtar , Abdullah Gani , Samee Ullah Khan. The authors comment that *“The continuous increase in the volume and detail of data captured by organizations, such as the rise of social media, Internet of Things (IoT), and multimedia, has produced an overwhelming flow of data in either structured or unstruc- tured format. Data creation is occurring at a record rate [1], referred to herein as big data, and has emerged as a widely recognized trend.”* They then go on to define Big data as follows. *“Big data are character- ized by three aspects: (a) data are numerous, (b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly.”* This is somewhat similar to the famous 4 Vs of defining data. Veracity, Volume, Velocity, Varying which were called on as definitions in the original report, however one simpler definition from Gartner is perhaps even more pertinent to our study of underlying big data technologies. Gartner defines Big data as *“high-volume, high-velocity, and/or high-variety information assets that* ***require new forms of processing*** *to enable enhanced decision making, insight discovery and process optimization”* (The importance of Big Data: A Definition. Douglas L) https://www.gartner.com/doc/2057415/importance-big-data-definition .The key here being ‘require new forms of processing’, the traditional relational database is not a good solution for big data – it is too slow in performance, and not easily scalable.. and so we must look to other solutions.

##### Big data and Cloud computing

*“Cloud computing is a powerful technology to perform massive-scale and complex computing .It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through cloud computing has been observed.”*

As discussed in the original report Autodesk is in full transformation from a desktop software company to a cloud software company “We’re transforming all our products from desktop to cloud products” Carl Bass Autodesk CEO Mar 19 2014. Cloud solutions bring many advantages to users and to the organization. The one which is relevant to us here is the vast amount of user data collected. This is the ‘big data’ element of our project . Autodesk uses Hadoop to pump upwards of 20 million hits from its cloud product sites into Hadoop. <https://www.datanami.com/2015/01/06/hadoop-behind-autodesks-cloud-ambitions/> and then aggregate this data and make it available to analyst to search for value.

“Relationship between cloud computing and big data Cloud computing and big data are conjoined. Big data provides users the ability to use commodity computing to process distributed queries across multiple data sets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms”

##### *The rise of “big data” on cloudcomputing:Review and open researchissues IbrahimAbakerTargioHashem a,n, IbrarYaqoob , NorBadrulAnuar , Salimah Mokhtar , AbdullahGani , SameeUllahKhan*

##### Overview

Apache define Hadoop as follows; “Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.”

##### History

Hadoop was created by Doug Cutting at Apache,. While working on a program to index web sites called Nutch , he was experiencing scalablity problems… .he found a google paper describing a hypothetical google file system which seemd to solv the scalability problems and set about creating and open source version of it in Java .. sorce (www.datanami.com/2015/04/15/from-spiders-to-elephants-the -history-of-hadoop . He called it the nutch distributed file system which becae the hadoop distributed file system . they also creaed then mapreduce which did the job processing work on the file system . Cutting was hired by Yahoo as they adopted hadoop to solve scalabot issues in their own serach engine… ..tiwiter, facebook .linked in soon started using the system . … and hadoopusage has grouwn from there.

In around 2004 with the objective of……..

Creatd by doug couting , had its origins in apache Nutch . an open source web serach engine. When nutch hit scaliblity issues… created a n open source version of Google distribute dFiel system (NDFS Nutch Distributed File System) .

2004 Google published the paper that introduced MapReduce to the world. Early in 2005 the nutch developers had a working mapreduce inmplementioan in Nuthc.

2006 – these elements were moved out of Lucene web search project and called Haddop … made a top level pojrect at Apache. (being used by faceobok, new yourk times.

Source: [https://books.google.ch/books?hl=en&lr=&id=drbI aro20oC&oi=fnd&pg=PR5&dq=hadoop&ots=t Dhucl1h3&sig=Gb2IH3qRfR20QGhe2xwB32WgLGU#v=onepage&q=hadoop&f=false](https://books.google.ch/books?hl=en&lr=&id=drbI_aro20oC&oi=fnd&pg=PR5&dq=hadoop&ots=t_Dhucl1h3&sig=Gb2IH3qRfR20QGhe2xwB32WgLGU#v=onepage&q=hadoop&f=false) Hadoop: The definitive guide, the third editions, by Tom White . Copyright 2011 Tom White, 978-1-449-3115 on Google scholar

##### Architecture

Description of Hadoop.

Taken from original report – here we need to go more indepth

*“. Hadoop is an open source apache project, combining a mapreduce processing engine with a distributed file system (HDFS) for storage.*

*In HDFS files are replicated across data nodes so that loss of one node will not lead to failure.*

*MapReduce is a java program which performs 2 basic functions which are specifically targeted at handling large volumes of data: - 1. Map and 2. Reduce*

*Map- Mapreduce  takes a large input  file and splits it into smaller files,  these files are then sent in parallel to nodes which run a function called a mapper. The mapper will convert the file into a structured format of key and value pairs and then output these into a new set of files. These new files then get sorted and collated by the key in the key value pairs.  All data for any given key is kept in one file (it is never split across different files). These files then become the input for the reducer function.*

*Reduce  - The reducer nodes take these new input files and run some aggregation over the data so that the reduce output consists of one key value par for each unique key … that data then goes into the output file for each of the reduce nodes and there are then merged together to produce the output for the map reduce job.”*

##### Advantages

Advantages of Hadoop

Apache Hadoop, NoSQL and NewSQL Solutions of Big Data.

Rakesh Kumar1, Bhanu Bhushan Parashar2, Sakshi Gupta3, Yougeshwary Sharma4, Neha Gupta5

1Red Hat Certified Engineer

2Assistant Professor & JECRC University, Jaipur, India

3Software Engineer & Alcatel-Lucent, Gurgaon, India

4Associate Software Engineer & Accenture, Pune, India

Benefits of HDFS:

· Portability

· Big data Capable

· Cost-effective

· Built-In Redundancy and Failover

Why MapReduce is Popular:

· Automatic parallelization and distribution

· Fault-tolerance

· Clean abstraction for developers

· Programs are written in Java language

##### Disadvantages

Disadvantages of Hadoop

##### Common uses

Where Hadoop is commonly used

## Analysis of choice of technology SQL vs. NoSQL

Analysis and comparison of the two technologies and justification of the choice of SQL

#### SQL

##### Overview

A brief overview of SQL

##### History

Origins and history of SQL

##### Architecture

Description of SQL

##### Advantages

Advantages of SQL

##### Disadvantages

Disadvantages of SQL

##### Common uses

Where SQL is commonly used.

#### NoSQL

##### Overview

A brief overview of NoSQL

Apache Hadoop, NoSQL and NewSQL Solutions of Big Data.

Rakesh Kumar1, Bhanu Bhushan Parashar2, Sakshi Gupta3, Yougeshwary Sharma4, Neha Gupta5

1Red Hat Certified Engineer

2Assistant Professor & JECRC University, Jaipur, India

3Software Engineer & Alcatel-Lucent, Gurgaon, India

4Associate Software Engineer & Accenture, Pune, India

Characteristics of NoSQL

· It support simple and flexible non-relational data models

· It stores large volume of data and having more flexible structure

· Use NoSQL without any inconsistency, in distributed environment so provide high availability

· No discontinuation of any work, if any faults or failures exist in any machine

· It allows data to store which is not require fixed table schemas

· It does not support ACID transactions as provided by RDBMS

· Ability to scale horizontally leading to high performance over many commodity servers

B. Why NoSQL

· NoSQL provides horizontal scalability better than vertical

· NoSQL support hardware getting cheaper and processing power increasing

· NoSQL support less operational complexity as against RDBMS solutions

· NoSQL provides, in most of the solutions you get automatic sharding etc. as default

D. Benefits of NoSQL

· Elastic scaling: organisations are able to scale out as well as take benefits of new nodes according

to their data storage needs

· No need for data to fit a schema: Both types of data (structured and unstructured) can be stored

as there is no fixed data model, so organisations access to much larger quantities of data

· Ability to cope with hardware failure: NoSQL database was designed with redundancy in mind

· Quick and easy development:

**A comprehensive comparison of SQL and MongoDB databases**

**Rajat Aghi, Sumeet Mehta, Rahul Chauhan, Siddhant Chaudhary and Navdeep Bohra**

Department of Computer Science, Maharaja Surajmal Institute of Technology, Janakpuri, N.delhi 110058, India

*The relational model*

Introduced in 1970s, the relational model offers a very mathematically-adapt way of structuring, keeping, and using the data. It expands the earlier designs of flat model, network model, et cetera by introducing means of relations. Relations bring the benefits of group-keeping the data as constrained collections whereby data-tables, containing the information in a structured way (e.g. a Person's name and address), relates all the input by assigning values to attributes (e.g. a Person's ID number).

Thanks to decades of research and development, database systems that implement the relational model work extremely efficiently and reliably. Combined with the long experience of programmers and database administrators working with these tools, using relational database applications has become the choice of mission-critical applications which cannot afford loss of any information. Despite their strict nature of forming and handling data, relational databases can become extremely flexible and offer a lot, granted with a little bit of effort

.

D

*The model-less (NoSQL) approach*

The NoSQL way of structuring the data consists of getting rid of the constraints imposed by the relational model, hence liberating the means of keeping, querying, and using information. NoSQL databases, by using an unstructured (or structured-on-the-go) kind of approach, aim to eliminate the limitations of strict relations, and offer many different types of ways to keep and work with the data for specific use cases efficiently (e.g. full-text document storage). By eradicating the strictly structured data keeping style defined within the relational model, these DB systems work by offering a much more freely shaped way of working with information, thus providing a great deal of flexibility and ease -- despite the fact that they come with their own problems, some serious considering the important and indispensable nature of data

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF CHEMICAL ENGINEERING

**ADVANCED PROCESS SIMULATION**

**SQL vs. NoSQL**

Author: Cansu Birgen

Supervisors:

Prof. Heinz Preisig

John Morud

**December 8, 2014**

Non-relational databases are named as NoSQL (Not Only SQL) which provides a mechanism for storage and retrieval of data which is modeled in a way different than in a relational database. NoSQL emphasizes the movement coming up with alternatives for RDBMSs/SQL where these are a bad fit rather than being being completely against them. Therefore, NoSQL is notmeant be a total replacement for relational approach. However, there is no still prescriptive definition of NoSQL. It consists of aggregation of common characteristics (Näsholm, 2012).

The term NoSQL dates back to 1998 when it was used for a particular RDBMS that did not support SQL. It was not until 2009 that it was used with approximately the same meaning that it has today.

Evolution of NoSQL databases was initiated by the need of a data storage model which enables the users work with large volumes of data with database running on clusters, since relational databases are not designed to run efficiently on clusters (Fowler et. Pramod, 2012). Common characteristics of NoSQL databases are shown below.

 \_Not using the relational model

 \_Running well on clusters

 \_Open-source

 \_Built for 21st century web estates

 \_Schemaless

After general introduction of NoSQL, a more detailed background will be provided in the following subsections.

##### History

Origins and history of NoSQL

##### Architecture

Description of NoSQL

##### Advantages

Advantages of NoSQL

Motives for Deployment of NoSQL

Motives for Deployment of NoSQL

For a long time, RDBMSs have been the solution in many applications of database management. However, some other applications have faced changing requirements, and RDBMSs were not fully capable of meeting the requirements (Strauch, 2011). This situation created a demand for a new approach in database managementthus triggered the interest in NoSQL. Some of the important features which have expanded the motives for deployment of NoSQL are listed below.

 \_For some applications, it is necessary to process huge amount of data. NoSQL handles it better compared to manyRDBMSs.

 \_RDBMSs were originally built to scale vertically referring to adding more hardware onto existing machines. However, vertical scaling has some limitations as it gives sublinear effects, sometimes requires high-risk operational efforts and there is an upper bound on how

4

 \_As open source projects, several NoSQL databases can be used for free. Many RDBMS alternatives are in comparison rather expensive. Also, because of less rigid data models and no explicitly defined schema, application development is often faster using NoSQL

##### Disadvantages

Disadvantages of NoSQL

##### Common uses

Where NoSQL is commonly used

### Justification of choice of SQL for the project.

In view of the advantages and disadvantages outlined for each technology previously and the common uses for each, justify the choice of an SQL solution for our data-mining project. E.g. time to learn and get up to speed. Static data in data warehouse, structured data in star schema.

## Analysis of choice of technology - Microsoft BI Suite

Description of tools and justification of the choice

##### Overview of Microsoft BI Suite

A brief overview of the technology

##### Advantages of Microsoft BI Suite

Advantages of SSIS/SSAS

##### Disadvantages of Microsoft BI Suite

Disadvantages of SSIS/SSAS

### Justification of choice of Microsoft BI Suite

In view of the advantages and disadvantages and possible alternatives outlined for Microsoft Visual Studio BI suite, justify its choice for our data mining project. E.g. Accepted in-house solution, industry standard, time to learn and get up to speed.

## Architecture

Description of Schema of Architecture - Describe and explain full schema of the Architecture. From manual extraction and collection of source data, through input into SQL ‘raw’ database, through transform to star schema database, through data cube creation and display in MS Excel. Include details of pre-treatment of data and requests to transform data.

# Analysis/discussion of results obtained by statistical analysis

The comments from the examination committee regarding the statistical analysis delivered in the original report were ‘step back and look at the statistical analysis, figures should be understandable, have a legend and be made explicit in the text (including graphs: axes units, legends, explanatory text).

To address these comments I will review the section ‘Analysis – Cube in Excel’ in the original project report and recreate it in the new report, modified according to the comments from the Examination committee.

#### Planned Analysis

The main objective of the ‘Analysis of results-statistical analysis’ section in the new report will be to detect and correct errors in the statistical analysis carried out in the original project and to present more clearly and exactly the results. The objective is not to carry out any new analysis or include any new data in the study. The addition of new data elements or new areas of analysis was not requested in the conclusions of the procedure de remediation. Furthermore there would be a serious limitation to any attempt to do this, as the author is no longer an employee of Autodesk he neither has access to the data sources nor to the environment where the original project was created.

#### Limitations of data in original project – data simulation

The new report will include a section outlining which data in the original project were authentic data and which data were simulated. This section will also describe the methods which were used to simulate data.

Data limitations

The most important data limitation for this project was the need to keep data confidential. As a result for the cloud applications I have taken samples (for a360 one week per month, for oxygen six months of the year... for akp one day per month) and then generated random data in the database for all other dates. I have also randomized and rearranged financial and customer satisfaction data in the raw files. This will undoubtedly have an impact on analysis but was necessary to proceed with the project.

# Recommendations for future development of the project.

A section will be added to the report evaluating the utility of the project in its current state and its benefit to the department; what problems it solves? what opportunities it opens? This section will also explore how the work form this project can be reused, improved and extended in order to bring further benefit to the department.

# Conclusion

# Bibliography

GALA. (2016). GALA – Globalization and Localization Association. Retrieved from GALA – Globalization and Localization Association: https://www.gala-global.org/language-industry/intro-language-industry/what-localization

# Appendixes

## Appendix A: Procédure de remédiation du travail de master du MAS ICT-4 de M. Simon Byrne

Procédure de remédiation du travail de master du MAS ICT-4 de M. Simon Byrne

Introduction

Suite à l'évaluation légèrement insuffisante du travail de master du MAS-ICT 4 de M. Byrne, la commission d'examen estime qu'une remédiation est recommandée. Celle-ci a été acceptée par l'étudiant. Le présent document a pour but de fixer le cadre de cette remédiation.

Cadre général

Un quart de travail normal sera demandé à l'étudiant (75 heures de travail) avec les phases principales décrites dans l'échéancier annexé.

Aucun développement logiciel (code) nouveau n'est demandé.

La CEX propose un passage en revue critique des points forts et faibles du travail, des objectifs de la remédiation et de la planification de celle-ci au sein d'une phase initiale de préparation, aboutissant à un cahier des charges validé par le conseiller, puis par la CEX (avec commentaires). Ensuite, au délai de rendu du rapport, le rapport comprenant les éléments figurant dans le cahier des charges sera fourni par l'étudiant. Enfin, une défense finale permettra d'évaluer l'ensemble du travail.

En cas de réussite de la remédiation, la note 4 sera attribuée.

La CEX fournit ci-après des pistes d'améliorations du travail.

Recommandations générales

Relisez soigneusement les directives du master et les directives concernant le rapport final.

Appliquez une approche scientifique : posez le problème, énoncez les critères de validation, présentez des arguments référencés (externes et internes), validez les résultats obtenus. Soyez critique et créatif.

Points du rapport à améliorer

Forme

• 20 pages maximum, le reste doit être fourni sous forme d'annexes référencées, notamment pour assurer un niveau et un public constant pour le rapport

• structuration trop linéaire : des renvois internes manquent, des références externes doivent être centralisées dans une bibliographie, numérotées et référencées par numéro dans le texte

• évitez les redites (faites des renvois internes ; si nécessaire, regroupez les parties traitant du même sujet)

• expliquez les acronymes et concepts, par exemple par une référence

• une relecture du rapport est obligatoire : veillez à l'orthographe, à la ponctuation, au fil rouge (ordre logique du discours), aux transitions explicites entre sections/chapitres. • mise en page soignée et relativement compacte

Contenu

• concentrez-vous à l’essentiel, structurez bien l’argumentation, justifiez l’importance ou la pertinence du contenu (pas d’affirmations gratuites), soyez précis, évitez les contradictions internes

• valorisez les résultats en fournissant une liste de recommandation claire sur la suite de la vie du projet (en quoi la solution est exploitable complètement, partiellement ; quels aspects sont réglés, lesquels doivent être améliorés, etc), expliquez comment l'on peut obtenir des résultats intéressants ; la solution est-elle réutilisable, extensible, générique ?

• un schéma général d'architecture de l'application (importation de données, prétraitement, stockage, requêtes, …), expliqué et référencé, est indispensable pour une bonne vue d'ensemble, cela permettrait aussi de donner quelques compléments comme p.ex. détailler quelles parties ont été automatisées ou non

• les choix doivent être systématiquement justifiés (par exemple dans un chapitre séparé), notamment :

◦ justifier l'utilisation de la suite BI de Microsoft

◦ expliquer pourquoi ne pas utiliser du NoSQL (avantages et inconvénients de cette solution, références p.ex. articles)

• prenez du recul sur l'analyse statistique effectuée

• les figures doivent être compréhensibles, avoir une légende et être explicitée dans le texte (y compris les graphes : unités des axes, légendes, textes explicatifs)

Points de la défense à améliorer

• ne lisez pas les transparents mot à mot : complétez les quelques points de chaque transparent par des explications orales

• évitez les transparents surchargés ou illisibles

• après le contexte général, vendez surtout les résultats de votre projet, pourquoi il est utile (cf directives de la défense)

• respectez le temps imparti qui est court : allez à l'essentiel, conservez éventuellement 1-2

transparents d'explications techniques en cas de questions

• l'architecture doit être décrite clairement

## Appendix B: Prolongation project calendar



## Appendix C: Notes from meetings with conseiller

* Mercredi, 14 septembre  8h30
* Vendredi, 30 septembre 8h30
* Lundi, 10 octobre 8h30
* Vendredi, 21 octobre 8h30

## Appendix D: Data sources used in original projects data mart.

|  |  |  |
| --- | --- | --- |
| Name | Origin | Description |
| Autodesk A360 | Hadoop | A360 is Autodesk online project collaboration software; it provides design, engineering and project teams to work together on a centralized platform, to view, search and share data and use other cloud collaboration tools |
| Autodesk Identity(Oxygen) | Hadoop | Identity is a core service that manages sign-on and permissions for Autodesk cloud products. The Identity dataset currently tracks users and API activity and can be partitioned by student, calling product, and country. This dataset is used to count monthly active users, frequency, and attrition rates. |
| Autodesk Knowledge Network | Hadoop | The Autodesk Knowledge Network program includes several communication channels for providing customers help on the web through the external Autodesk Knowledge Network web site, in product contextual help accessible from both cloud and desktop product and form the Autodesk community forums at fourms.autodesk.com. |
| BBIS Report | Excel file | The BBIS report is an excel pivot report which is published internally of Autodesk software, bookings and billings in terms of seats and also in terms of $ segmented by Site and Parent Industry , Company size , product, country and Fiscal year, quarter and month. |
| Customer Satisfaction data | Excel file | The customer satisfaction data is an excel sheet which is provided by the Global Customer Support and Operations department. It contains the results of a Global (13 languages), annual survey involving 50,000 commercial customers to get a pulse on customer loyalty to Autodesk. |
| Education community data. | Excel file | Statistics on the number of official education users and downloads of products per country are maintained by the Autodesk Education community team and |
| Autodesk Homepage visits | Excel file | A log of how many visits are recorded to each of the Autodesk principal regional websites is also shared via an internal dashboard and downloadable in excel format. |
| Autodesk tweets | Excel file | From [**http://www.tweetarchivist.com/**](http://www.tweetarchivist.com/) |
| English Proficiency Index | Excel file | http://www.ef.edu/epi/  The world’s largest index for English speaking skills. |
| WEF data | Excel file | The World Economic Forum posts annual their Global Competitiveness Report which compiles various different economic statistics on the countries of the world |

## Appendix E: Description of Autodesk’s Localization Services department

Autodesk is a world leader in the provision of software and services for engineering, design and entertainment. The ‘Localization Services department is responsible for producing localized versions of Autodesk software. That is to say, the translation of the user interfaces, documentation and in certain cases the addition of specific content for a geographic region or market. (GALA, 2016) define localization as “Localization (also referred to as "l10n") is the process of adapting a product or content to a specific locale or market. Translation is only one of several elements of the localization process. In addition to translation, the localization process may also include:

* Adapting graphics to target markets
* Modifying content to suit the tastes and consumption habits of other markets
* Adapting design and layout to properly display translated text
* Converting to local requirements (such as currencies and units of measure)
* Using proper local formats for dates, addresses, and phone numbers
* Addressing local regulations and legal requirements