# Terms of Reference

Business Intelligence Systems and Data Mining

Project Title: Evolution of Data Loading techniques: From Batch ETL to Stream Processing

Grigorios Moumtzis

Supervisor: Ioannis Vourgidis

# Background to the Project

A data warehouse can be defined as the central repository of integrated data that is modeled in such as way that the business intelligence applications built on top of it can assist the management of the business to improve decision making. Moving the data between the operational systems and the data warehouse is always a complicated task as the design and implementation of ETL processes requires high skilled IT specialists. In addition, the ETL requirements have changed as we have entered the Data Analytics era. A modern ETL approach requires near real-time data integration to the so-called real-time data warehouse. The management of almost all businesses, including start-ups and ecommerce companies, demand now reports or dashboards that are refreshed with a very high frequency – something that in the past was a requirement only for very specific industries such as telecommunications, fraud detection in banks or stock market data.

For very large-scale enterprises, the implementation of modern ETL and data integration techniques is something crucial – but these companies have the capacity and the know-how to develop in-house customized solutions that can do the work. On the other hand, small and medium size businesses are still struggling to provide meaningful business intelligence. Although the Business Intelligence concept is not new, the high-level management of most business don’t consider Business intelligence/Data Warehousing as a strategic; The consequence is that the business intelligence teams of small companies do not have the capacity to meet the reporting/data analysis requirements set by other departments. Most of the times, the main difficulty is the design and the implementation of the ETL processes. Many companies do not even have a dedicated BI Team. As a result, the data integration is slow and not scalable – it’s not rare to see whole dashboards that use as a source the operational system itself and to pull the data by using long, complex and not readable SQL queries.

This project is tackling the aforementioned data integration problem. It focuses on a specific part of the BI Lifecycle which is the design and implementation of scalable, efficient and near real-time ETL solutions that can be easily implemented without the need of self-coding or the purchase of expensive tools.

# Deliverables

* Dissertation in digital form that will include:

A literature review/fact finding section

A software implementation and the evaluation of it

A critical review of the project

* Hard copy if needed

# Academic Objectives

* Acquire a solid theoretical understanding about all different methods that are used to import data to data warehouses with an emphasis on modern ETL technologies.
* Improve my technical skills by designing and implementing a near real-time ETL process.
* Working systematically by researching this specific topic will improve my overall knowledge and soft skills as I will be able to critically evaluate research papers in my area of expertise and get a holistic overview of the topic. This will be a good foundation that will help me to further research similar topics outside of the scope of this project but also will help me professionally to work in a more systematic way.

# Background Research Objectives

* The main research objective is to learn more about ETL methods that can provide real-time or near real-time data loadings to data warehouses.
* Another research objective is to research all existing ETL methods that are used for data imports.
* To provide a solid foundation for the need of ETL, I will also research the concept of data warehousing and why is it important for a business
* Another research objective is to find out the challenges the BI specialists are facing nowadays and link them to the topic I am dealing with.
* I will conduct also a research about the specifications of commercial or open-source ETL tools as well as databases to understand how feasible is for a small organization to have a modern ETL solution.

# Research Questions

* What is the status of research in the field of data warehousing and ETL data pipelines?
* What are the new technologies in terms of ETL and data integration?
* Is design and implementation of ETL the most difficult challenge of a BI Team?
* Do small business need a data warehouse?
* Do small business need real-time ETL?
* Is it feasible for a small business to invest to modern ETL technology?
* Is it a way for a small business to build a business intelligence system that is efficient?
* In terms of real-time ETL, is any method that I easy to implement without purchasing expensive software?
* How are event streaming platforms linked to real-time ETL?
* Can an event streaming platform be used for integrating data to data warehouse?
* The project will suggest a simple framework for real-time ETL. Is this system feasible, efficient and easy to implement? Does the use of this technology add any value to the business? Where should be used instead of old-fashioned ETL methods or instead of simple SQL Queries?

# Product Objectives

Α Consultancy Project that will include

* A high-level overview of a Business Intelligence System and its components
* A more detailed review about data modelling of a data warehouse and how ETL is linked to that.
* A description and critical evaluation of ETL techniques in terms of updating a data warehouse, with an emphasis on real-time ETL techniques.
* Two recommended systems for real-time ETL that is scalable and easy to implement by small businesses.
* A comparison of these systems

# Risk Assessment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk identification** | **Probability** | **Impact** | **Assessment** | **Risk monitoring, mitigation and management** |
| The deadline of the project is not met | medium | high | high | A very detailed project plan including a Gant Chart will be the basis to work systematically and split the project in small chunks. This will ensure that the project will finish on time as all milestones should be accomplished on time. |
| The folder that contains the project files is deleted / lost because of hardware malfunction or an accidental event. | low | high | medium | All related files will be pushed to a Git Repo |
| Risk of not founding enough data for the implementation part of the project | low | medium | low | If no reliable source is found, a self-developed script will produce it in short time intervals |