# Terms of Reference

Business Intelligence Systems and Data Mining

Project Title: Evolution of Data Integration in Data Warehousing: 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 most 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 do not 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 very specific part of the business intelligence lifecycle which is the ETL process with an emphasis on real-time ETL processes. The goal is to provide suggestions for ETL strategies that are simple to implement and have real or near real-time capabilities. The suggestions will be based on a) the literature review b) the implementation and evaluation of two different ETL architectures. The evaluation will include several criteria including performance.

# Deliverables

* Dissertation in digital form that will include:

A literature review/fact finding section

An evaluation of two ETL tools based on a set of quantitative and qualitative criteria

A critical review of the project (literature review and implementation)

* Hard copy if needed

# Academic Objectives

* Acquire a solid theoretical understanding about all different methods that are used to extract, transform and load data to data warehouses with an emphasis on modern ETL technologies.
* Improvement of technical skills by designing and implementing a near real-time ETL process that includes the use of a modern ETL tool that belongs to the big data technologies.
* Working systematically by researching this specific topic will improve the overall knowledge and soft skills of the author as he will be able to critically evaluate research papers in his area of expertise and get a holistic overview of the topic. This will provide a good foundation for further research of similar topics outside of the scope of this project. This will also help professionally to work in a more systematic way.

# Background Research Objectives

* The main research objective is to learn more about all the ETL methods (data integration) specifically for data warehousing. Emphasis will be given on real-time ETL techniques.
* Another research objective is to learn more about big data technologies and see how these are related to ETL (specifically for data warehousing)
* To provide a solid foundation for the need of ETL, the author will also research the concept of business intelligence and data warehousing and explain why it is important for a business.

# Research Questions

* How ETL processes have been evolved since the beginning of the Business Intelligence era?
* How do event streaming platforms work and how can they provide real-time ETL?
* What is the adding value of the modern technology compared to the traditional one in terms of performance, implementation and cost?

# Product Objectives

* A high-level overview of a Business Intelligence System and its components
* A more detailed review about data warehousing 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.
* Evaluation of two ETL tools that use different technologies
* Suggestions and recommendations for ETL strategies

# 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 |