UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

Proceedings of the 13th International Conference on Applied Internet and Information Technologies

AIIT 2023

13 October, 2023, Bitola, Republic of North Macedonia

AIIT
International Conference

UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

13TH INTERNATIONAL CONFERENCE ON APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2023
PROCEEDINGS

Bitola, 2023

Proceedings publisher and organizer of the conference:
University “St. Kliment Ohridski” - Bitola, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD
Faculty of Information and Communication Technologies - Bitola
University “St. Kliment Ohridski” - Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Eleonora Brtka, PhD

Conference Chairmans:
Kostandina Veljanovska, University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia - chair
Eleonora Brtka, University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Andrijana Bocevska, PhD

Kostandina Veljanovska, PhD

Cover design:

Andrijana Bocevska, PhD

e-Proceedings
ISBN 978-608-5003-03-7

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2023 (13 ; 2023 ; Bitola, Republic of North Macedonia)
Proceedings / 13th International conference on applied internet and information technologies AIIT 2023, 13 October, 2023, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Eleonora Brtka]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2023. - 411 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-03-7

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 62559493

C O N T E N T

Automation and Monitoring on Integration ETL Processes while Distributing Data

Aneta Trajkovska, Tome Dimovski, Ramona Markoska, Zoran Kotevski

University "St. Kliment Ohridski" – Bitola, Faculty of Information and Communication Technologies, Studentska bb, Bitola, Republic of Macedonia

aneta.trajkovska@uklo.edu.mk, tome.dimovski@uklo.edu.mk, ramona.markoska@uklo.edu.mk, zoran.kotevski@uklo.edu.mk

Abstract:
This paper describes an approach for the automation of processes, leveraging the capabilities of various technologies and diligent data monitoring. The continuous advancement of technology presents new opportunities for development and offers a superior and more precise alternative for decision-making across various domains, not only IT areas. From an analytical standpoint, such a system is characterized by its ability to operate in a manner where multiple computers or computing devices on a network share various components responsible for distinct operations. By dividing the workload, these devices coordinate their capabilities to efficiently accomplish the intended task, resulting in enhanced performance compared to a single device handling it alone. This is precisely what SnapLogic, as an iPaaS platform, excels at. It simplifies management through APIs, which are utilized in its ETL (Extract, Transform, and Load) processes, thereby expanding communication possibilities with other systems. By using this integration platform, businesses can eliminate slow and error-prone manual methods while achieving a high level of automation for their processes. One notable feature of this platform is its process orchestration capability. It empowers users to monitor and manage data pipelines originating from various platforms and systems, allowing for seamless application integrations and API calls, all executed efficiently within a single system.

Keywords:
Integrations, IPaaS, ETL (Extract, Transform and Load) processes, Snaplogic, pipelines, API, AWS IaC, AWS CloudFormation, ELK (Elasticsearch, Logstash and Kibana)

1. Introduction

With the exponential progress of technologies and the growing need to distribute a variety of data formats (while they are easily accessible and presented with good visualization), there is also the need to use processes that that enhance efficiency while providing a high level of automation for various operations. The ever-growing volume of data within organizations necessitates finding ways to ensure that their transactions are executed in a timely manner, enabling them to be processed and transferred to the appropriate destination promptly. The systems themselves have the capability to coordinate the execution of operations and tasks in a timely manner, thereby enhancing efficiency and significantly reducing the time required to complete tasks. This is precisely one of the objectives of integration iPaaS (Integration Platform as a Service) systems, ETL(Extract, Transform and Load) processes, and distributed systems.

iPaaS is a set of automated tools that integrate software applications that are deployed in different environments [1]. Typically, an iPaaS platform provides pre-built connectors, business rules, maps and transformations that facilitate the development of applications and orchestrate integration flows [2].

The utilization of ETL processes facilitates a straightforward and standardized approach to meet the integration requirements, enabling large organizations and their businesses to process data effectively and achieve the desired outcomes. ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system [3].

In this paper it was used automation on integration, infrastructure in AWS, and monitoring on ETL processes. The remaining of this paper is organized as follows. Section 2 presents the challenges of utilization of the integration processes in the organizations. In section 3 are elaborated: ETL processes and their advantages, Snaplogic as an integration platform that is running ETL processes ELK components and their functions, AWS CloudFormation service as an Infrastructure as a Code (IaC). Section 4 describes the proposed solution with utilization of the mentioned services and technologies, while section 5 concludes the paper with a summary of the research.

2. The challenge of using integration processes in the organization

The utilization of integration processes within organizations poses certain challenges that need to be addressed effectively. In the past, companies integrated their business processes through custom programming, enterprise middleware or enterprise application integration (EAI) implementations, such as service-oriented architecture (SOA). These integration solutions worked but were expensive and time-consuming to create [4].

iPaaS offers a unified and streamlined solution for data exchange and integration across all applications within an organization, regardless of whether they are hosted on-premises or in the cloud. This enables seamless and consistent data integration processes throughout the entire organization (Figure 1).

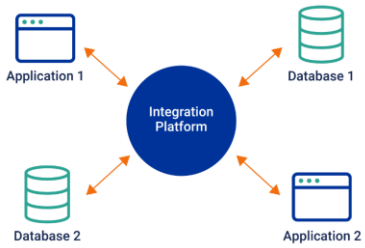

Figure 1: How integration platforms drive the flow of data between applications and databases

An organization can implement iPaaS more rapidly and cost-effectively compared to traditional technologies, offering comprehensive integration across various use cases such as cloud-to-cloud, cloud-to-on-premises, and on-premises-to-on-premises scenarios. iPaaS plays a vital role in driving digital transformation by providing several notable advantages [5], [6],[7],[8]:

- Faster deployment and cost efficiency,
- Business and IT agility,
- Flexibility for growth and scalability,
- Enhanced Connectivity,
- Simplified Management and Monitoring.

3. Automation of ETL integration processes using Snaplogic platform, AWS infrastructure and ELK monitoring

Increasing operational efficiency through streamlined data integration brings a lot of advantages. Exploring the integration of ETL processes with the Snaplogic platform, AWS infrastructure, and ELK monitoring yields a transformative approach to data management and optimization.

3.1. ETL (Extract, Transform and Load)

Modern information technology relies extensively on ETL processes to achieve enhanced data processing and secure data distribution. ETL stands for Extract, Transform, and Load, and it refers to the procedure of transferring data from its source to a designated data warehouse [9], [10].

During the extraction phase, data is gathered and extracted from various sources such as databases, APIs, or files. The data is then subjected to transformation, which involves cleaning, filtering, and structuring it to meet the requirements of the target system or data model. Lastly, the transformed data is loaded into the data warehouse, where it can be efficiently stored, organized, and analyzed (Figure 2).

Figure 2: Phases of ETL processes

The utilization of ETL processes plays a crucial role in modern information technology, enabling organizations to effectively manage and leverage their data assets for improved decision-making, analytics, and business insights [11],[17].

ETL and ELT (Extract, Load and Transform) are both viable solutions for data movement and transformation but are best suited to different business use cases [12].

3.2. Snaplogic – integration platform that runs automated ETL processes

Today, organizations are driven to achieve enhanced performance in terms of speed and quality. SnapLogic is an integration platform (iPaaS-integration-platform–as-a-service) that is continuously coming into use when solving any business solutions. It is renowned as one of the fastest and most efficient platforms and services available in the market. SnapLogic's Intelligent Integration Platform (IIP) automates all phases of IT integration projects—design, development, deployment, and maintenance—whether in cloud or hybrid environments. It makes the data integration process faster and easier.

The SnapLogic iPaaS includes 700+ pre-built connectors, called Snaps and pre-built end-to-end integrations, called templates, for common business processes. Pipeline templates enable every user to quickly create the integrations for new, improved workflows and outcomes [15]. The power of this platform is of really great importance to many organizations, as it has huge possibilities that it offers, and at the same time, easy management, and an understandable interface for users, which makes it irreplaceable and gives the impression of an integration platform that will be constantly upgraded. It is a simple platform that consists of many powerful features [18], such as:

- Unification (One integration platform connects more than 700 snaps),
- Productivity (Self Service Integration for everyone),
- Modern (Flexible Deployment for Any – sized Enterprise),
- Artificial intelligence (for speed and quality),
- Interface – easy access through a browser to the platform itself.

The products it offers are a simple way (with little or no code) to automate applications and data integrations. Performs cloud data mobilization using visualization and ETL processes to further test them at the application level.

3.3. Snaplogic Architecture

Snaplogic's architecture is designed to address the complexities of modern data integration, providing a scalable and flexible platform to handle data from on-premises systems, cloud applications, APIs, and IoT devices. At its core, Snaplogic employs a unique visual design approach, empowering users with little to no coding experience to create complex data pipelines effortlessly. The architecture of Snaplogic revolves around three fundamental components [19]:

- Snaps (Snaps are pre-built connectors that facilitate the seamless integration of various data sources and targets),
- Snaplex (The Snaplex is a distributed execution grid that executes the data pipelines created using Snaps. It operates both in the cloud and on-premises, providing the flexibility to deploy integration workloads in the most suitable environment.),
- Snaplogic Manager (The Snaplogic Manager serves as the control center for designing, monitoring, and managing data integration processes.).

3.4. ELK (Elasticsearch, Logstash and Kibana)

ELK, an acronym for Elasticsearch, Logstash, and Kibana, is a comprehensive data processing and visualization platform designed to help organizations efficiently collect, store, analyze, and visualize vast amounts of data in real-time [20]. Elasticsearch's underlying data structure, based on JSON documents, enables lightning-fast searches, aggregations, and complex queries, making it an indispensable component for performing data analysis at scale [21]. Logstash serves as the data processing engine in the ELK Stack. It provides a flexible and extensible framework for collecting, parsing, and enriching data from various sources before forwarding it to Elasticsearch [22]. Kibana is a powerful data visualization tool, allows users to create interactive and customizable dashboards, reports, and visualizations, enabling them to explore data insights effortlessly [23].

3.5. AWS CloudFormation

AWS CloudFormation revolutionizes the way infrastructure is managed by treating it as code. Instead of manually configuring resources through the AWS Management Console or using AWS Command Line Interface (CLI), users can codify their infrastructure in JSON or YAML templates. This paradigm shift enables version control, peer review, and automated deployments, ensuring that infrastructure changes are consistent, documented, and auditable [24]. By leveraging declarative templates, resource drift detection, stack management, and integrations with various AWS services, CloudFormation streamlines the process of building scalable and consistent infrastructures, thereby enabling businesses to focus on innovation and growth while maintaining optimal control over their cloud environments.

4. Proposed solution

Within this paper, I will present a solution that offers a significant contribution to how the automation itself introduced into the process can be a key tool that enables simpler maintenance and optimization of resources. By harnessing the power of the SnapLogic platform and the businesses approach of utilization of the ETL processes, fortified by the scalability of AWS infrastructure, and enhanced by ELK monitoring (system logs, application logs, system metrics), we have engineered a comprehensive system that revolutionizes data workflows.

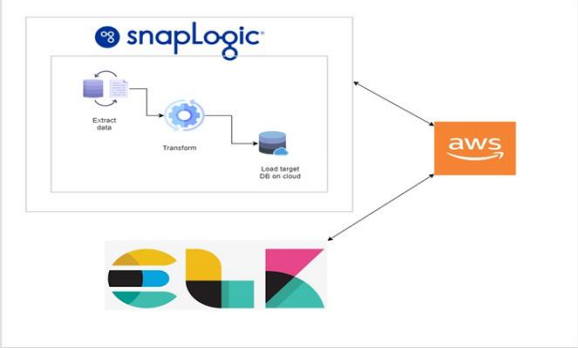

Figure 3: High overview of the proposed automated solution and used technologies

To present the overall setting of the solution in relation to the subject that is dealt with in this paper I have set up four stages of workflow: code preparation (as the first phase – staging phase), code development (the second phase of the automated ETL processes and the design of the infrastructure itself), code review (the third phase is to the review of the code itself within the prepared scripts and the verification of accuracy) and the code promotion (is the final stage of the prepared automated processes where the environments are active and the execution of the ETL processes runs smoothly).

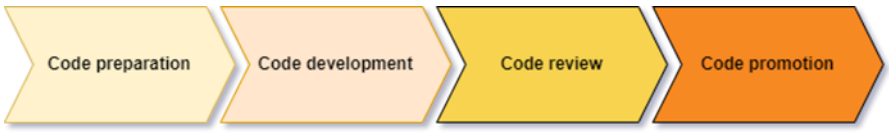

Figure 4: Workflow stages

4.1. Infrastructure of the solution in AWS

Based of the automation in the infrastructure I have used the AWS service CloudFormation, which is a service of Amazon Web Services (AWS) that enables the automatic creation and management of the infrastructure in the cloud by coding the infrastructure as code.

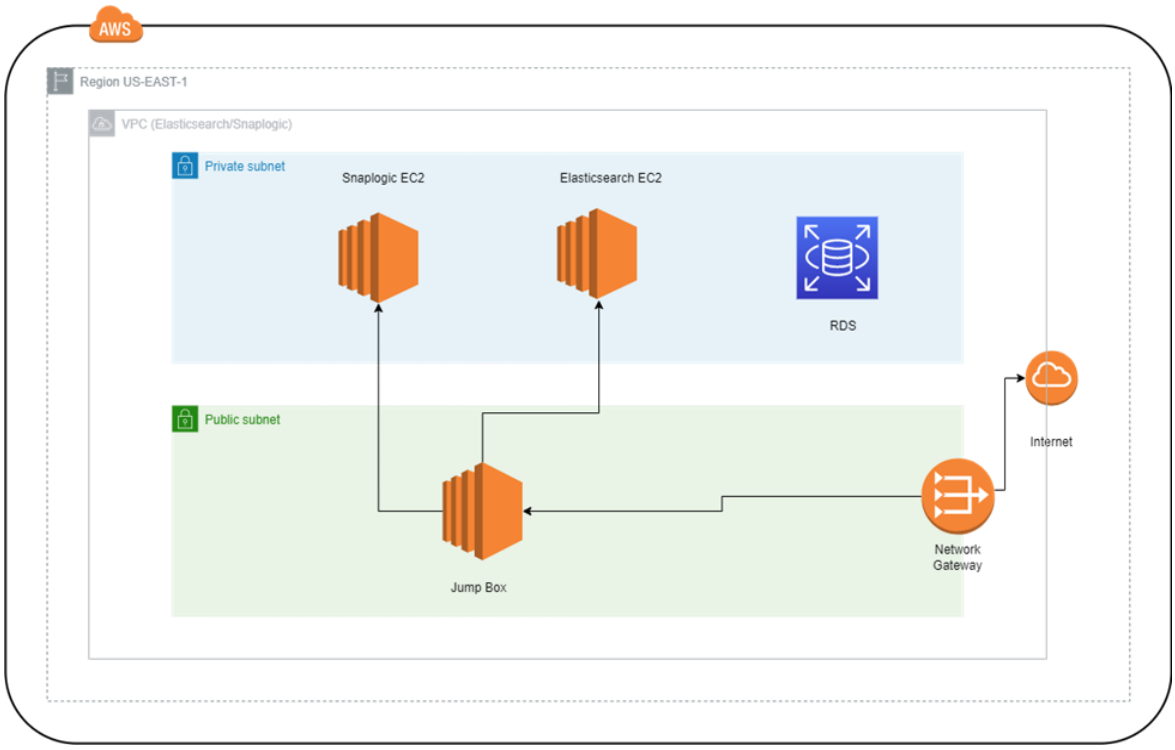

Figure 5: Infrastructure of the solution on AWS

The solution's architecture is outlined as follows: within the cloud environment, an initial work region is designated, leading to the creation of an isolated virtual private cloud within the broader AWS public cloud. This virtual private cloud is subdivided into two distinct subnets - one public and one private. A unique aspect is observed in the placement of two EC2 instances within the private network, and an additional two within the public network.

The essential contribution in selecting this configuration is in the communication between these instances. A Network Gateway instance is established to facilitate access to the Jump Box instance,

which resides alongside other components within the same public network. By utilizing the public IP address, a connection is established with the Jump Box instance. Subsequently, the connection proceeds from the Jump Box EC2 instance to access Snaplogic and Elasticsearch, both housed within the private network. This connection is facilitated via the employment of the private IP address.

Moreover, the proximity within the same subnet allows seamless access to the created RDS database, underscoring the cohesive integration of components within this comprehensive solution.

4.2. Automated ETL integration processes

The Snaplogic integration platform is very powerful and can achieve the processing of a large amount of data in just a few seconds, if the processes configured in it are created optimally.

The following scenario is composed of four processes, one of which is the main process, and the rest are sub-pipelines that are adequately called depending on where the data flow is. At the beginning, the process log information from the start of the execution. Then it pulls data from the API to further map and transform them, to finally record certain data in the MYSQL database.

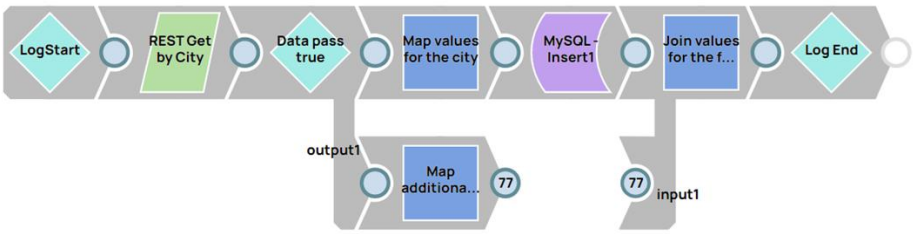

Figure 6: An integration process that aims to log into the database, read data from the API, transform it and write it to the database

The second scenario of the integration ETL process provides information about the state and activity of the snaplex itself. It works in a way that makes a call to a public API, where this time I use Snaplogic as the generator of the necessary data, making a call:

GET https://{pod_path}/api/1/rest/public/snaplex?{query_parameters}

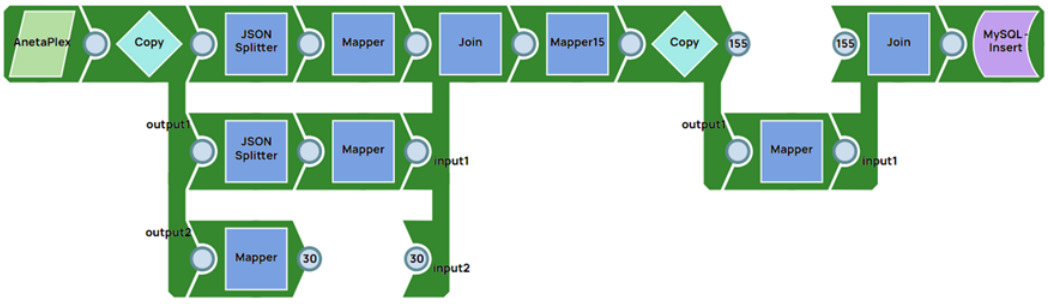

Figure 7: Successfully completed integration process that writes statistics data into the database from the information on the Snaplex

4.3. Monitoring of the data with utilization of ELK

ELK can be used to read metrics and monitoring data from a variety of sources, such as system logs, application logs, system metrics, and others. The monitoring itself is of great importance not only for the users of the data but also for the development of the business itself within the organizations.

Monitoring integration ETL processes is important to ensure that they are working properly, and that data is being transferred and transformed in a way that is appropriate for business needs. Monitoring can help detect process problems and warn them before serious consequences occur. This can be done by constantly monitoring the performance of the ETL process, as well as by reviewing the logs and occasionally testing the data.

Monitoring plays a pivotal role in elevating business efficiency. If process problems can be detected and resolved, it can lead to faster and better decision-making, as well as increase the quality of the data used to run the business.

Figure 8: Visualization of the temperature, pressure and air humidity changes for the city of Bitola in Kibana, based on the data in the table

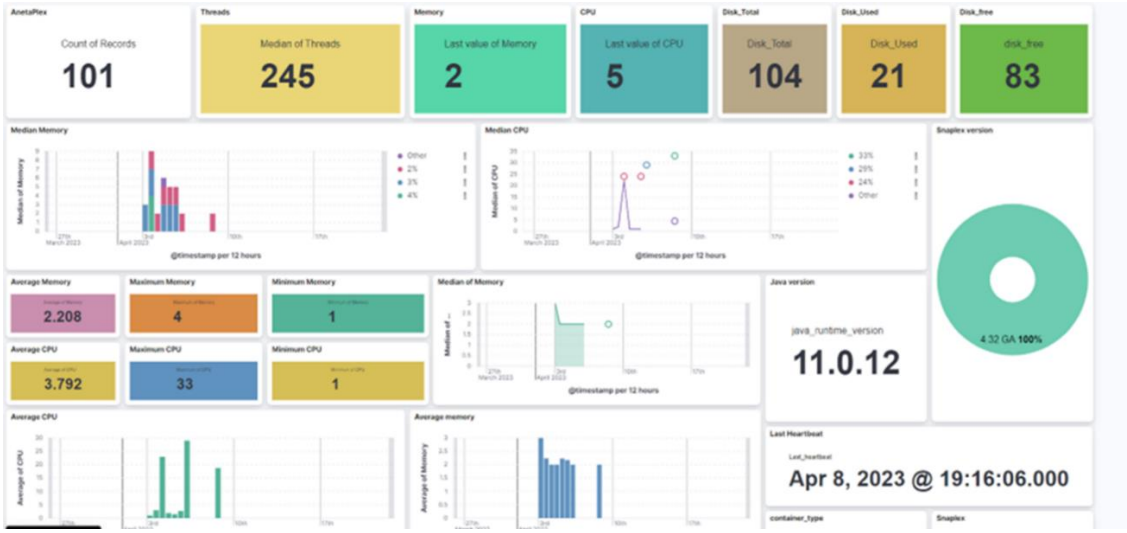

Figure 9: Dashboard that is showing data metrics from both tables SnaplexHealthTableInfoStats and SnaplexHealth

5. Conclusions

Automation control in process management is one of the most important areas of science and technological development. The automation and adaptability of integrated systems themselves

represent opportunities for better process control. The primary objective of this project was to enable the automation of the ETL process and implement monitoring for data flow through the Elastic Stack.

In this paper, were explored the numerous advantages of using automation in ETL integration processes within the AWS infrastructure. Through extensive research and practical implementation, we have observed significant benefits from utilizing this approach.

The project has successfully achieved its goals, allowing users to enhance database performance, improve infrastructure creation efficiency, and reduce the time required for manual data processing by implementing this solution.

The results of the research were successful, achieving the automation of the ETL process and providing real-time data flow monitoring through Elasticsearch and Kibana. The automated ETL process significantly improved data management efficiency, while the data flow monitoring empowered users with a comprehensive, real-time view of their data.

This project serves as compelling evidence that leveraging automated integration for ETL processes, along with AWS CloudFormation infrastructure, and ELK stack for monitoring, can yield improved data management, increased accuracy, and enhanced process efficiency.

References:

[1] Fred Churchville and Stephen J. Bigelow " What is iPaaS? Guide to integration platform as a service" Cloud infrastructure design and management, June (2021)
[2] Fred Churchville and Stephen J. Bigelow " Guide to integration platform as a service" Cloud infrastructure design and management, June (2021)
[3] IBM Cloud Education, IBM Cloud Education, " ETL (Extract, Transform, Load)", 14 December 2021
[4] IBM Cloud Education, " The enterprise integration challenge", 2022
[5] SnapLogic, iPaaS - Integration Platform as a Service Explained, https://www.snaplogic.com/blog/what-is-ipaas-integration-platform-as-a-service , 28 October 2023
[6] Asad Zulfahri, "16 Benefits of iPaaS That Show How It Can Transform Your Business", https://blog.hubspot.com/website/ipaas-benefits , 2021
[7] Stephen J. Bigelow, " Review these top enterprise iPaaS benefits and challenges", https://www.techtarget.com/searchapparchitecture/tip, May 2021
[8] IBM Cloud Education, " Benefits of iPaaS", 2022
[9] Snaplogic, The Modern ETL Guide, 16 May 2022
[10] IBM, ETL (Extract, Transform, Load), https://www.ibm.com/topics/etl
[11] Shanika Wickramasinghe, " Data 101: Extract, Transform & Load (ETL) Explained" January 26, 2023, https://www.splunk.com/en_us/blog/learn/etl-extract-transform-load.html
[12] Snaplogic, ETL vs. ELT: What's the Difference Between These 2 Processes, https://www.snaplogic.com/blog/etl-vs-elt-whats-the-difference , 13 September 2022
[13] Snaplogic documentation https://docs-snaplogic.atlassian.net/wiki/spaces/SD/overview
[14] Snaplogic, Data Pipeline, https://www.snaplogic.com/glossary/data-pipeline
[15] Snaplogic, " Intelligent Integration Platform", https://www.snaplogic.com/products/intelligent-integration-platform
[16] Martin Flower, ISBN 0321127420, Patterns of Enterprise application architecture, 11th October 2006
[17] Snaplogic, "ETL Data Pipelines", https://www.snaplogic.com/blog/all-about-etl-data-pipelines , May 2023
[18] Snaplogic, Enterprise Integration Architecture, Cloud Service Integration (snaplogic.com)
[19] Snaplogic, https://www.snaplogic.com/blog/cloud-integration-platform
[20] Elastic Stack, https://www.elastic.co/what-is/elk-stack
[21] Elastic Stack, https://www.elastic.co/what-is/elasticsearch
[22] Elastic Docs, https://www.elastic.co/guide/en/logstash/current/introduction.html
[23] Elastic Stack, https://www.elastic.co/what-is/kibana
[24] AWS Docs, https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/Welcome.html