International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/IJSRCSEIT

Automation In Data Engineering Using SQL and BI Tools

Harish Goud Kola
Independent Researcher, USA

Article Info
Volume 7, Issue 5
Page Number: 329-336

Publication Issue :
September-October-2020

Article History
Accepted : 10 Oct 2020
Published : 30 Oct 2020

ABSTRACT

In this research article it will provide information regarding the data engineering that utilizes automation in using SQL and BI tools to enhance the workflow of data. The automation of ETL (extraction, transformation, and loading) processes will improve the efficiency, consistency, and quality of a data pipeline. Reduced human interference in the processing of queries in SQL and the integration with BI tools streamline data processes, increase speed, and impact real-time decision-making. This automation minimizes errors and allows for scaling when dealing with large volumes of data. The informations that are presented in this research article is gathered from books, journals, articles and online websites.

Keywords : Automation, Data Engineering, SQL, and BI Tools.

1. Introduction

In data engineering using SQL and BI tools, automation in data workflows, mainly in large-scale workflows, has changed the approaches to management. Organizations have significantly relied on data-driven insights to make its key decisions. Thus, the increasing need for automating the ETL processes is due to efficiency, accuracy, and scalability needs. Utilizing SQL queries and BI tools, data engineers can automate repetitive tasks, ensure consistency of data, and increase data processing speed overall. This integration does not only ensure less manual intervention but also helps businesses deal with enormous amounts of data with greater ease, thus enabling faster, better-informed decisions. With this ever-increasing demand for real-time analytics, automation of data engineering is important in order to maintain integrity, optimize resource usage, and encourage innovation in data management practice.

2. Literature Review

2.1 Automation in SQL through Snowflake Tasks for BI Tools and Dashboards

According to the author XIA et al.2019, it states that in this research the aim was to analyze the manner in which automation of SQL enhances the user experience on BI tools as it streamlines the access of data and improves the speed of generating reports. It demonstrates how automating repetitive SQL processes empowers business users, hence leading towards better decision-making and efficiency. The approach of this method was about the automation of ETL and report generation through the use of SQL scripts on BI tools. It provides significant outcomes such as the automation saves person-effort, enhances data quality, and accelerates reporting cycles. Increased engagement and better utilization of data-driven decisions emerge from the analysis. Probable future scope would be expanding automation into predictive analytics by integrating real-time data for deeper insight.

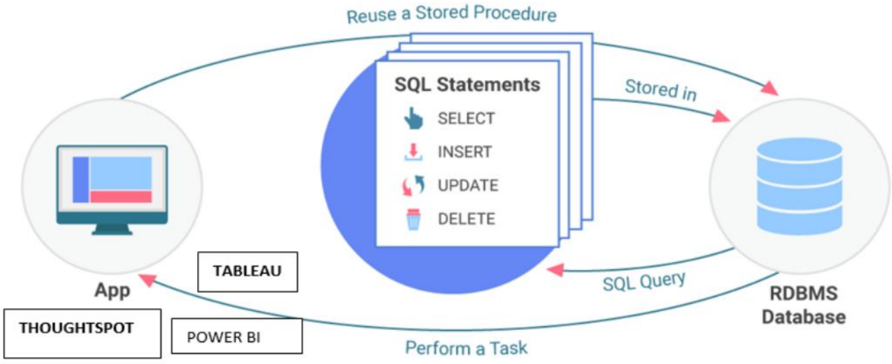

Figure 1 : Architecture SQL automation

(Source: https://www.espjeta.org)

2.2 An Analysis of Traditional ETL Tools Against the Latest Unified Platform

According to the author Sethi et al.2020, it states that in this research the aim was to compare data analytics platforms which are recently developed with the legacy of ETL tools, moving on in the areas of performance, scalability, and capabilities integration. The objective was to compare the level of technology innovation into achieving increased real-time processing, agility, and making real-time data processing cost-efficient, thus deeming modern platforms as more favorable to organizations looking to maximize the data value. The methodology has based the analysis of strengths and weaknesses of traditional systems and modern ones. The results indicated that modern platforms provide more flexibility, scalability, and ease of use. It made an important point in making decisions based on proper information while opting for data integration approaches. The scope for the future is further research into more developments relating to ELT, streaming data, and unified approaches for special business needs.

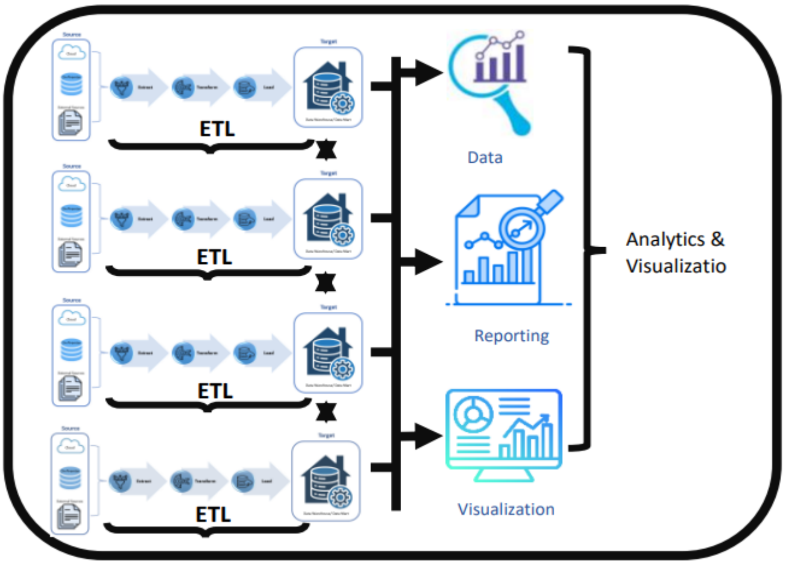

Figure 2: Data Analytics Platform

(Source: https://www.researchgate.net)

2.3 A Survey of Pipeline Tools for Data Engineering

According to the author Ashok et al.2020, it states that this research focused on a number of data pipeline tools applied in data engineering concerning the activities in wrangling and preparation for applying machine learning processes. The objective was to categorize and review these tools around the design and the intended data engineering functions, including ETL/ELT, data integration, ingestion, orchestration, and ML pipelines. The approach considered the analysis of both commercial and open-source tools, with references and case studies of practical applications. The choice of pipeline tool for application depended on specific task needs. It also found out some challenges of tool integration and data preparation for machine learning applications which can be focused in further research works. The results also reflected a space for further scope in research on advanced pipeline orchestration techniques and optimization of automation for large-scale ML projects.

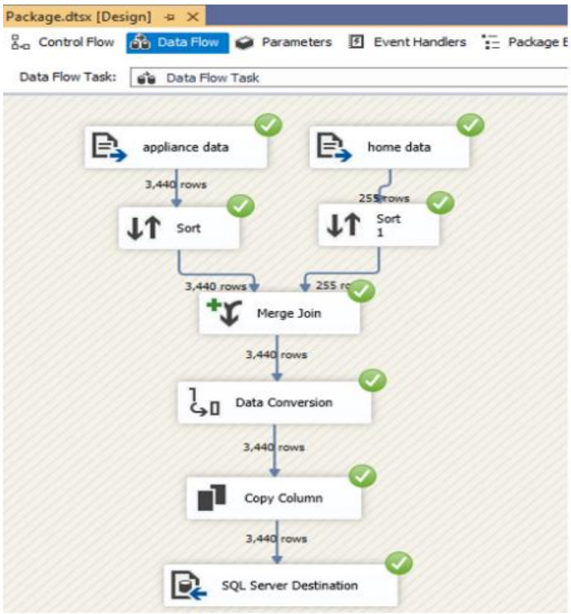

Figure 3: Microsoft SSIS Pipeline

(Source: https://arxiv.org)

3. Methods

3.1 Data Collection and Preprocessing Techniques

The techniques for Collecting and Preprocessing Data with automation in data engineering using SQL and BI tools, can be performed as a preliminary step towards structuring and cleaning raw data for efficient processing. Data collection can be defined as the process of extracting data from reliable sources such as relational databases, flat files, APIs, or even cloud storage into one central data repository (Banga et al.2020). SQL queries are highly common in any application that automates the extraction of relevant datasets; thus, the right data is retrieved in real time or in batch mode. Once the data is collected, preprocessing begins, where it involves data cleaning, transformation, and standardization, which helps in eliminating inconsistencies so that adding up can be done towards quality. SQL scripts and BI tools handle missing values, remove duplicates, and normalize data formats to automate several tasks. In most data preprocessing, aggregation and joining various sources into one single view for analysis are part of the requirements. Automation of collection and preprocessing reduces manual effort, enhances the accuracy of the data, and accelerates the process flow of the data pipeline, making it possible for more complex analysis and results.

3.2 Automation of ETL process using SQL and BI tools

It states that streamlining data workflows, so that timely and accurate data is made available for analysis purposes that is present directly at the center of automation through SQL and BI tools. In the extraction phase, SQL queries automatically bring out data from databases, APIs, or files, thereby minimizing the extent of manual efforts needed. After extraction has been done, the data must be cleaned and aggregated and reformatted according to business requirements as part of the transformation phase. Advanced transformations, such as data normalization, type conversion, and filtering, SQL scripts are applied to, but BI tools can be instrumental in visualizing how the transformation process works for easier monitoring (Kretz et al.2019). In the loading phase, the data transfer happens automatically as the data warehouses or reporting systems get loaded. These are often the scheduled SQL works or ETL tools like Talend or Apache Nifi. These tools provide the scheduling and managing of ETL workflows in a very intuitive interface. The processing can be either real-time or batch. Automation ensures consistency, reduces errors, and increases the efficiency of data pipelines, hence speeding up decision-making.

3.3 Design and Deployment of Automated Data Systems

The design and deployment phases of automated data systems actually hold an important role in ensuring data pipelines are scalable, efficient, and reliable. In the design phase, there is planning of the architecture that involves integration of various sources of data, different storage solutions, and processing systems using SQL and BI tools. Data models are structured for proper design of the data so that queries can be easily set up and analyzed (Michael et al.2020). The automation at each step would ensure that extraction, transformation, and loading can take place with less human intervention. SQL scripts are optimized for efficiency in execution. It states that during the deployment phase, the packaged automated data systems would be integrated into the production environments. This includes scheduled job setup, data pipeline configuration, and BI tool usage for system health and performance monitoring. It is probably deployed on cloud-based platforms like AWS or Azure, allowing deployment to work on scalable and flexible scalability. The totally automated systems are monitored, fine-tuned, and kept in order so that efficiency with data integrity, performance optimizations, and overall systematic working can be realized.

4. Results

4.1 Efficiency Improvements through Automation

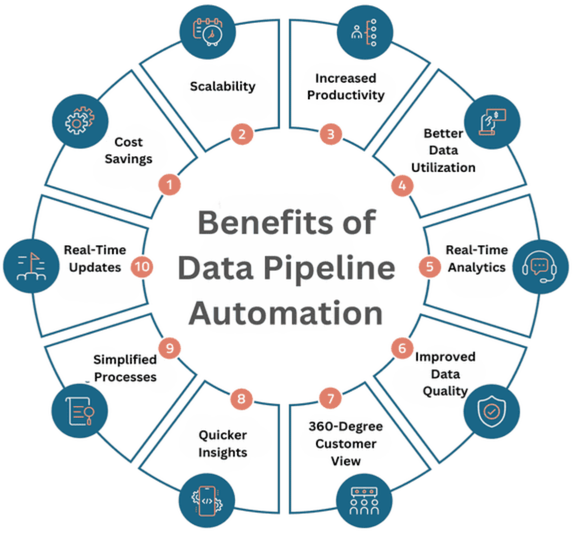

Figure 4: Efficiency Improvements through Automation

(Source: https://www.kohezion.com)

In data engineering, automation-based efficiency improvement is said to be the primary advantage of using SQL and BI tools. Automating data extraction, transformation, and loading leads to fewer manual interventions that could otherwise be time-consuming and error-prone. Data pipelines can work seamlessly and at scale by processing massive volumes of data unattended, whereas the automated SQL scripts that run at intervals. BI tools make this process even more efficient by allowing real-time data monitoring, automated reporting, and the quick identification of bottlenecks in the workflow (Berezovskyi et al.2019). Data engineers can then allocate more strategic work while automation ensures consistency throughout the pipeline. Second, automation accelerates data availability such that what could take hours or even days may be completed in minutes, ETL processes previously taking hours or days are now accomplished in minutes. As a consequence, businesses get to access the latest insights quicker, enhance the decision-making capabilities, and respond more effectively to changing conditions.

4.2 Case studies and real world applications

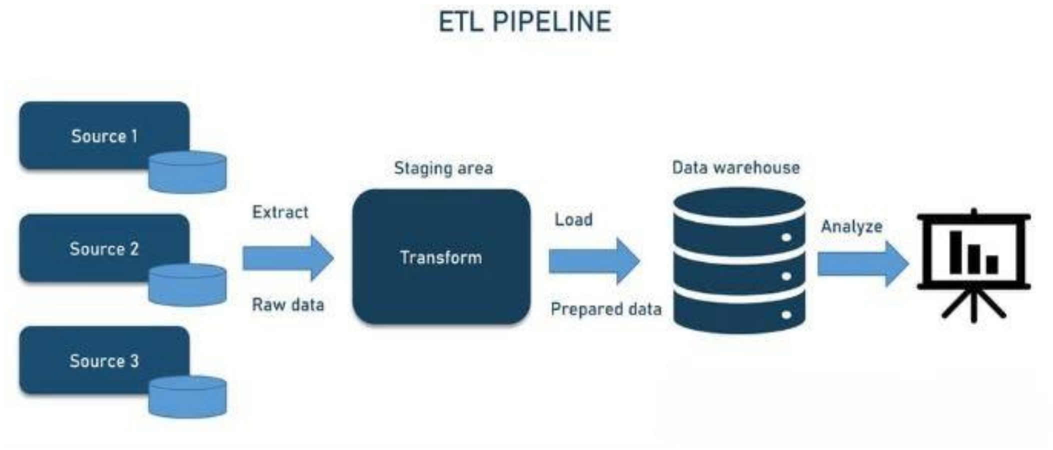

Figure 5: ETL Pipeline Processes

(Source: https://www.altexsoft.com)

Case studies and real-world applications illustrate the wide reach of automation in data engineering using SQL and BI tools across various sectors. For instance, within the retail sector, firms have implemented the use of automated pipelines of data in tracking inventory, sales, and customer behavior in real-time to produce far more accurate demand forecasting and dynamic pricing methods. The ETL automated processes help streamline patient data management for the healthcare sector by integrating information from different systems into timely and accurate reporting. It states that in a recent scenario, consider a case with a financial institution that automated its reconciliation through SQL scripts and BI dashboards (Romero et al.2020). The entire process was free of manual errors, making it possible to advance the monthly reporting cycle from several days to hours. Automation of analysis has also significantly improved predictive maintenance and production scheduling in the manufacturing industries. These case studies reflect how automation with SQL and BI tools would contribute to operational efficiency, cost savings, and better decision-making across various sectors.

4.3 Impact on Data quality and Consistency

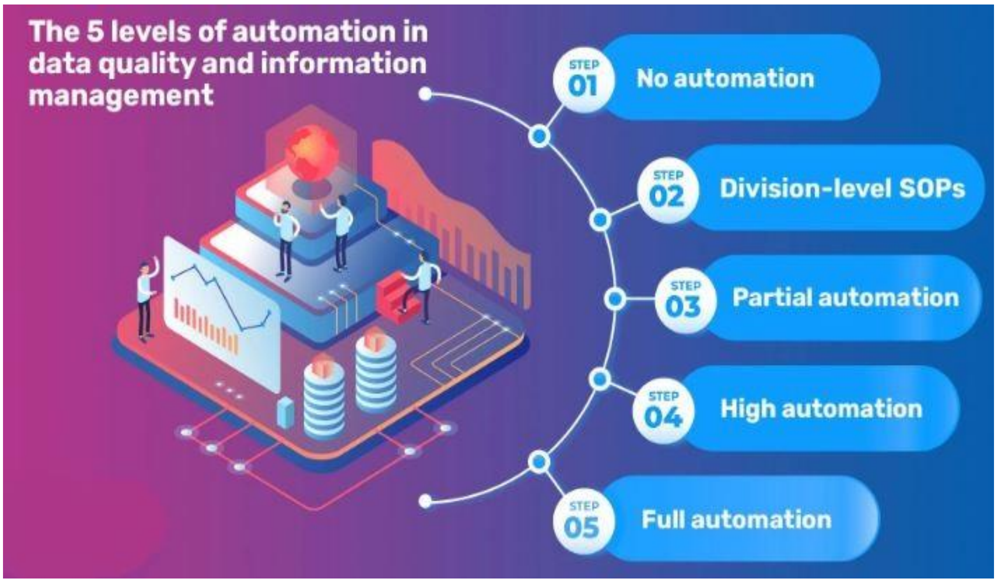
The 5 levels of automation in data quality and information management
STEP 01 No automation
STEP 02 Division-level SOPs
STEP 03 Partial automation
STEP 04 High automation
STEP 05 Full automation

Figure 6 : Impact on Data quality and Consistency

(Source: https://info.italentdigital.com)

It provides information that among other things, automation fundamentally affects data quality and consistency, especially in using SQL and BI tools with data engineering. Data inconsistencies are normally minimized by automation, which is mostly caused by human error that is likely to come across when uniform processing of datasets during data extraction, transformation, and loading processes are not maintained. SQL queries can be designed to clean and validate data automatically, or even in real-time so that anomalies or discrepancies are detected. In this process, high-quality data will, in turn, always enter the system. BI tools provide dashboards and alerts monitoring the data quality, thereby taking corrective measures automatically if problematic issues have arisen within it (Sharma et al.2020). Automated data pipelines ensure that the input lines observe the same structure in terms of formatting and transformation rules. Organizations can depend on the integrity of its data, thus leading to more accurate analysis and better reporting. As a result, decision-making improves with automation of such processes, which in the long run can effectively help maintain a reliable flow of data across systems.

5. Discussion

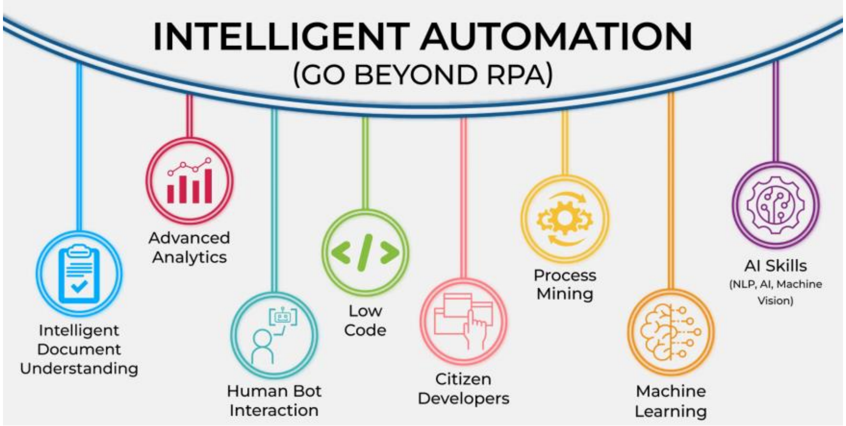

Figure 7: Intelligent Automation

(Source: https://deltalogix.blog)

In this research article it presented automation of data engineering using SQL and BI tools where it covers both the positive and negative sides about using such systems. The first positive is significant efficiency improvement because the automation process removes repetitive labor tasks on the data processing, which gains better time to get insights. Data accuracy is also improved by making sure the result from the automation reduces human error and maintains consistent data transformations and validations all through the pipeline. It also states that these aspects also bring upon challenges in some scenarios. One kind of a problem or negative issue is on designing and maintaining data pipelines, mainly when integrating more than one data source or especially with large volumes of data. It also has information such as the heavy dependence on automated processes may sometimes mask underlying problems in the data, thereby making it hard to notice in real time (Khan et al.2020). It states that despite how much difficult these are, automation remains to be one of the most important tools for modern data engineering workflows because of benefits such as scalability, instant decision-making, and enhanced quality of data. Future advancements in artificial intelligence and machine learning will look into facing all these problems and raising the efficiency of automation.

6. Future Directions

The future of automation in data engineering, using SQL and BI tools, might be to integrate with advanced technologies like AI/ML, creating a more powerful capability in data processing. AI can automate more complex tasks such as anomaly detection and predictive analytics, and real-time decision-making can be done with deeper insights from data through minimal human intervention. The cloud-based platforms and serverless architectures will enable agile and scalable data systems to handle increasing volumes of data. Data governance and security automation will also be crucial in the same areas, where many BI tools automatically support data privacy monitoring, compliance, and auditing (Rao et al.2019). As more organizations may move towards a more connected ecosystem, the ability to seamlessly integrate disparate data sources through automated workflows will become one of the important ingredients of success. It also states that adopting self-service BI tools will enable non-technical users to automate even more data reporting and analytics, providing greater leverage to automation for more people.

7. Conclusion

The automation of SQL and BI tools in the work role of data engineers results in efficiency gains in process, improvement in data quality, and decision-making processes. Organizations would reduce errors, lower the efforts, and increase speed to serve accurate real-time analytics by automating critical extraction, transformation, and loading processes. One of the areas where SQL queries seamlessly integrate with BI tools is the aspect of smooth data workflows that enhance scalability, allowing organizations to handle high volumes of data. As automation and machine learning are becoming more prevalent, data engineering will become more dependent on AI, machine learning, and cloud-based solutions in order to optimize pipelines and work toward innovation in industries.

he global rubber industry, offering potential for local economic development, from reducing its environmental impact to superior agricultural adaptation. With much work to overcome full scale commercialization challenges, the analysis shows a robust framework for future development. To maximize guayule's potential as an alternative to natural rubber production based on sustainability particular continued investment in research, technological innovation and strategic implementation will be required.

REFERENCES

[1] XIA Jr, Y.U.Q.I., 2019. Data Automation, Data Analytics and Processing System.

[2] Åstrand, A., 2020. Re-engineering a database driven software tool: Rebuilding, automating processes and data migration.

[3] Sethi, F., 2020. Automating software code deployment using continuous integration and continuous delivery pipeline for business intelligence solutions. Authorea Preprints.

[4] Ashok, H., Ayyasamy, S., Ashok, A. and Arunachalam, V., 2020, July. E-business analytics through ETL and self-service business intelligence tool. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 747-751). IEEE.

[5] Banga, D. and Khang, A., 2020. Application of Data Technologies and Tools in Business and Finance Sectors. In Data-Driven Modelling and Predictive Analytics in Business and Finance (pp. 1-17). Auerbach Publications.

[6] Kretz, A., 2019. The data engineering cookbook. Mastering the plumbing of data science.

[7] Michael, A.V. and Ahirao, P., 2020, April. Improved use of ETL tool for updation and creation of data warehouse from different RDBMS. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

[8] El-khoury, J., Berezovskyi, A. and Nyberg, M., 2019. An industrial evaluation of data access techniques for the interoperability of engineering software tools. Journal of Industrial Information Integration, 15, pp.58-68.

[9] Romero, O., Wrembel, R. and Song, I.Y., 2020. An alternative view on data processing pipelines from the DOLAP 2019 perspective. Information Systems, 92, p.101489.

[10] Sharma, S., Goyal, S.K. and Kumar, K., 2020. An Approach for Implementation of Cost Effective Automated Data Warehouse System. International Journal of Computer Information Systems and Industrial Management Applications, 12, pp.13-13.

[11] Khan, M., 2020. Distributed and scalable parsing solution for telecom network data.

[12] Rao, T.R., Mitra, P., Bhatt, R. and Goswami, A., 2019. The big data system, components, tools, and technologies: a survey. Knowledge and Information Systems, 60, pp.1165-1245.