ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Configuration-Driven No-Code Data Ingestion Engine with Rule-Based Processing

Naveen Koka

USA

ABSTRACT

The efficient management of batch processes is critical for modern data-driven organizations, especially within the context of complex data ingestion and processing workflows. This abstract explores the essential components and functionalities required to streamline batch job execution within such environments. Firstly, a robust batch management system is essential, providing capabilities for job scheduling, dependency management, and error handling. Such a system orchestrates the execution of batch jobs, ensuring they run in the appropriate sequence and handle dependencies effectively. Additionally, real-time event-driven triggers enable immediate batch job initiation upon file drops, ensuring timely processing of incoming data and swift response to business needs. Conversely, scheduled batch executions provide predictability and automation, allowing organizations to optimize resource utilization and maintain operational efficiency by running batches at predetermined intervals or specific times.

Furthermore, the abstract delves into the necessity of a comprehensive configuration setup to support various data formats, transformation requirements, and validation criteria. A rich configuration engine allows for seamless mapping, transformation, and validation of incoming data, enabling flexibility and adaptability to diverse data sources and requirements. It also emphasizes the importance of containerization for batch job execution, ensuring scalability, resource isolation, and efficient utilization of computing resources. Lastly, the provision of job IDs facilitates seamless invocation of batch jobs by external systems, enabling streamlined interaction and integration across different platforms. Together, these components form a cohesive framework for efficient batch job management, empowering organizations to optimize data processing workflows and derive actionable insights from their data assets.

***Corresponding author**

Naveen Koka, USA.

Received: October 07, 2023; **Accepted:** October 11, 2023; **Published:** October 21, 2023

Keywords: Data Ingestion, Rules Engine

Introduction

Processing, revolutionizing the way businesses import and manage partner data. With intuitive configuration options, users can seamlessly set up and automate data ingestion processes without the need for extensive coding expertise. This innovative solution streamlines operations, reducing the need for manual intervention and minimizing maintenance efforts. By empowering users to define rules and workflows easily, it enhances flexibility and scalability, enabling businesses to adapt swiftly to changing requirements.

Problem Statement

The E-commerce platforms and data ingestion systems play a vital role in the modern digital marketplace, serving as the backbone for businesses to import partner data, market products, and facilitate sales. However, a significant challenge lies in efficiently integrating partner data into these systems. Without robust configuration options, the process often demands extensive tweaking and maintenance, leading to inefficiencies and increased operational overhead.

The absence of comprehensive configuration capabilities can result in a cumbersome and time-consuming data import process. Businesses may find themselves grappling with manual adjustments and frequent updates to ensure accurate data ingestion. This not only hampers productivity but also introduces the risk of errors and inconsistencies, potentially undermining the integrity of the platform's offerings.

Moreover, the lack of rich configuration options can impede the scalability and adaptability of e-commerce platforms and data ingestion systems. As businesses expand their partnerships and product offerings, the need for flexibility in importing and managing diverse sets of data becomes paramount. Without adequate configuration tools in place, organizations may struggle to keep pace with evolving requirements, hindering their ability to capitalize on growth opportunities and stay competitive in the dynamic digital landscape.

Solution

A rich configuration engine is the cornerstone of efficient data management for marketers, offering unparalleled flexibility in mapping, transforming, and validating data. With such a tool at their disposal, marketers can expedite the process of bringing their products to market, as it streamlines the ingestion of partner data. By enabling seamless mapping of diverse data formats and structures, along with robust validation mechanisms, the configuration engine ensures that the imported data meets quality standards, thereby enhancing the marketer's ability to make informed decisions and drive sales. Furthermore, the flexibility

provided by the engine empowers marketers to adapt quickly to changing requirements and optimize their strategies for maximum impact.

In addition to a rich configuration engine, proper batching and event mechanisms are crucial for efficient data import, especially in managing multi-tenant data and responding promptly to events. Batch processing allows for the efficient handling of large volumes of data, optimizing resource utilization and ensuring smooth operations across multiple tenants. Meanwhile, event-driven import mechanisms enable real-time data ingestion, ensuring that marketers have access to the latest information as soon as it becomes available. By integrating these mechanisms into the data ingestion process, businesses can maintain data integrity, minimize processing delays, and capitalize on opportunities with agility and precision.

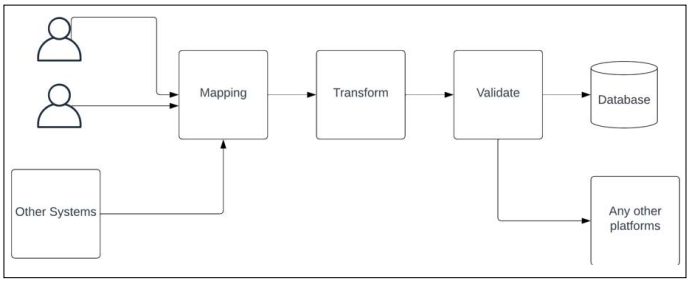

Figure 1: Data Ingestion Engine

Engine Details

To accommodate customers dropping files in various locations, a diverse array of file drop options must be available for seamless import from these sources. These options include FTP, S3, and an HTTP service, where files are received along with corresponding job IDs. This multifaceted approach ensures flexibility and accessibility, allowing customers to utilize their preferred file transfer methods while simplifying the data import process.

To effectively process files provided by marketers, the engine must possess the capability to comprehend various file formats. A broad spectrum of file support is essential to achieve this objective, extending beyond conventional formats like CSV, JSON, and XML. Diversifying file support ensures adaptability to the diverse needs of marketers and the data they provide. Upon specifying the file format, the engine should leverage available open-source libraries tailored to each format. This strategic utilization of libraries streamlines the reading process, enhancing efficiency and accuracy in data extraction and interpretation.

By embracing a comprehensive range of file formats, the engine empowers marketers with the freedom to utilize data in their preferred formats without encountering compatibility constraints. This flexibility fosters smoother collaboration and integration of diverse data sources, ultimately enhancing the marketer's ability to derive actionable insights and drive business growth. Leveraging open-source libraries specific to each file format not only ensures robust file parsing capabilities but also facilitates continuous improvement and updates through community contributions, keeping the engine aligned with evolving industry standards and customer requirements.

The engine's next task involves mapping the fields extracted from the file and transforming them into a format conducive to seamless integration into the system. While mapping CSV files typically follows a straightforward process, handling multi-level structures like JSON and XML requires additional flexibility. Empowering the user interface with elasticity to facilitate complex mappings, coupled with an internal engine capable of converting data into the desired format, significantly enhances the capabilities of the data ingestion system. This ensures compatibility with diverse data sources and optimizes the efficiency of data processing workflows, ultimately enhancing the system's overall effectiveness in managing and leveraging imported data.

By enabling the engine to dynamically adapt to various data structures and formats, including those with nested hierarchies like JSON and XML, the system enhances its versatility and usability. This adaptability not only simplifies the mapping process for users but also ensures that the imported data is accurately transformed and integrated into the system. With the capability to handle complex mappings seamlessly, the system empowers businesses to leverage a wider range of data sources and capitalize on valuable insights, thereby driving innovation and competitive advantage in today's dynamic marketplace.

The engine is now tasked with transforming incoming data according to the predefined map. In addition to this core function, the transformation engine is expected to offer supplementary capabilities such as merging multiple fields into two or vice versa, as well as providing support for mathematical operations, string manipulation, and date functions. These functionalities are considered essential prerequisites for modern data ingestion systems, ensuring not only seamless data transformation but also facilitating advanced processing and manipulation of data elements.

These capabilities serve as foundational components for efficient data management and analysis, enabling businesses to derive deeper insights and maximize the value extracted from their datasets. By incorporating such features into the transformation engine, the data ingestion system empowers users with the tools necessary to streamline data processing workflows, enhance data quality, and unlock the full potential of their data assets. Additionally, these capabilities lay the groundwork for scalability and innovation, enabling organizations to adapt to evolving data requirements and leverage emerging technologies to drive business growth and competitiveness.

The engine's next responsibility is to validate incoming data according to predefined rules. To accomplish this, the configuration must offer options to define criteria for data validation. If the data fails to meet the specified criteria, the import process is halted to prevent the ingestion of erroneous or incomplete information. This validation mechanism ensures the integrity and accuracy of the imported data, minimizing the risk of errors and inconsistencies within the system.

By enabling users to configure validation criteria, the system enhances its adaptability to diverse data sources and requirements. This empowers businesses to establish tailored validation rules that align with their specific data quality standards and regulatory compliance needs. Furthermore, the ability to halt the import process upon encountering invalid data helps maintain data integrity and reliability, ultimately contributing to more informed decision-making and improved operational efficiency within the organization.

Jobs

The Running these batches will inevitably consume time, especially considering the volume of data being processed.

Additionally, to ensure optimal performance, running the batches within a container is necessary. However, this approach may demand significant memory resources. Therefore, there arises a need for a structured template that can facilitate the creation of the container and the development of a startup script to execute the batch within the container efficiently.

The template serves as a blueprint, providing a standardized framework for configuring the container environment. It outlines the necessary specifications and dependencies required to set up the container environment effectively. By adhering to this template, developers can ensure consistency and reliability across different container instances, streamlining the deployment process and minimizing potential compatibility issues.

Furthermore, the startup script plays a crucial role in orchestrating the execution of the batch process within the container. This script automates the initialization steps, such as loading data, executing commands, and managing resources, thereby simplifying the deployment and execution of batch jobs. Through the integration of a well-designed startup script, organizations can enhance the efficiency and scalability of their batch processing workflows, ultimately optimizing resource utilization and accelerating time-to-insight.

Once the batch runs are established, individual jobs may consist of multiple batches to execute, with some running concurrently while others depend on the completion of preceding batches. Therefore, a comprehensive batch management system becomes indispensable to orchestrate and oversee the execution of these jobs effectively. This system facilitates the organization and coordination of batch jobs, ensuring they are executed in the desired sequence and according to predefined dependencies.

Central to the batch management system is its ability to handle job scheduling, prioritization, and dependency management. By defining dependencies between batches within a job, the system can intelligently sequence their execution, ensuring that dependent batches are only initiated once their prerequisite batches have successfully completed. Moreover, the system enables parallel execution of independent batches, optimizing resource utilization and minimizing overall job execution time.

Additionally, the batch management system provides monitoring and error handling capabilities to ensure job reliability and integrity. It tracks the progress of batch executions, monitors resource utilization, and detects and handles errors or failures gracefully. This proactive approach to job management enhances system robustness and reliability, enabling organizations to maintain operational efficiency and meet critical processing deadlines with confidence.

Once all configurations are set up within the batch management system, an essential requirement is the provision of a unique job ID to facilitate invocation by external systems. This job ID serves as a crucial reference point, enabling seamless interaction between the external system and the batch management infrastructure. With the job ID in hand, external systems can efficiently trigger the execution of specific batch jobs, ensuring timely processing of data and tasks according to predefined parameters.

Events

There are two distinct types of events that trigger batch job execution within the system. The first type involves immediate action upon the occurrence of a file drop event. As soon as a file is dropped into the designated location, this event serves as a signal to initiate the corresponding batch job without delay. This real-time response mechanism ensures prompt processing of incoming data, enabling timely execution of tasks and actions aligned with business requirements.

In contrast, the second type of event revolves around scheduled batch job execution. In this scenario, batch jobs are programmed to run at predetermined intervals or specific times according to a predefined schedule. This scheduling mechanism allows for the systematic and planned execution of batch processes, enabling organizations to optimize resource allocation, manage workload distribution, and maintain operational efficiency. By leveraging scheduled events, businesses can automate routine tasks, streamline data processing workflows, and ensure consistent and reliable execution of batch jobs over time.

Events play a pivotal role in driving batch job execution within data processing systems. Real-time file drop events trigger immediate batch job initiation, ensuring rapid response to incoming data and facilitating timely processing and analysis. Conversely, scheduled events provide a structured approach to batch job execution, enabling organizations to automate routine tasks and optimize resource allocation. By leveraging both types of events, organizations can achieve a balance between real-time responsiveness and systematic scheduling, thereby enhancing operational efficiency and enabling agile decision-making based on up-to-date data insights. Moreover, events serve as the catalyst for seamless interaction between the batch management system and external systems, facilitating streamlined data workflows and promoting interoperability across diverse platforms and applications.

Uses

The data ingestion system finds application across a spectrum of industries and platforms, notably within e-commerce platforms and transaction-based systems. In e-commerce, it serves as the backbone for importing partner data, managing product catalogs, and facilitating seamless

In transaction-based systems such as financial institutions or online banking platforms, the data ingestion system plays a crucial role in processing and analyzing transactional data in real-time.

E-Commerce Platforms

Importing partner data is indispensable within the e-commerce landscape, where the seamless integration of products, inventory, and pricing information is critical for maintaining competitiveness. In this dynamic market, any delay in data import can result in missed opportunities and customer dissatisfaction. Therefore, a robust data ingestion system is paramount, ensuring the timely and uninterrupted import of partner data. With such a system in place, businesses can guarantee the constant flow of essential information without experiencing downtime, thereby empowering them to stay agile and responsive to market demands.

Banking Platforms

In banking platforms, the need for efficient data ingestion is equally crucial, albeit in a different context. Here, the system is tasked with importing various types of financial data, including transaction records, account balances, and customer information. Timely and accurate data ingestion is essential for providing real-time insights into account activities, detecting fraudulent transactions,

and offering personalized financial services to customers. With a robust data ingestion system, banking platforms can ensure the seamless integration of data from diverse sources, enabling them to maintain operational efficiency, comply with regulatory requirements, and deliver superior customer experiences. Moreover, by leveraging advanced data ingestion technologies, banks can enhance data security, mitigate risks, and unlock new opportunities for innovation in financial services.

Conclusion

The robust data ingestion systems across diverse industries, including e-commerce and banking. These systems serve as the backbone for importing, processing, and managing critical data, enabling businesses to stay competitive and responsive to market demands. Whether in e-commerce platforms or banking environments, timely and accurate data ingestion is essential for driving operational efficiency, enhancing customer experiences, and facilitating informed decision-making. By leveraging advanced technologies and methodologies, organizations can optimize their data ingestion processes, ensuring seamless integration of diverse data sources and maintaining a competitive edge in today's dynamic marketplace. Ultimately, investing in robust data ingestion systems is essential for organizations looking to unlock the full potential of their data assets and drive sustainable growth in the digital age [1-5].

References

1. Alwidian Jaber, Rahman Sana, Gnaim Maram, Al-Taharwah Fatima (2020) Big Data Ingestion and Preparation Tools. Modern Applied Science 14: 12.
2. Yadav Rakhee, Kumar Yogesh, Patil Rajendra (2023) Study of Data Ingestion Tools. Bi-Lingual International Research Journal 10. 246-253.
3. Tunjic A (2019) The Automation of the Data Lake Ingestion Process from Various Sources. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) 1276-1281.
4. Kannan Sobti, Deepika Dash (2020) Top Big Data Technologies for Data Ingestion. International Research Journal of Engineering and Technology 7: 4143-4148.
5. Ji C, Shao Q, Sun J, Liu S, Pan L, et al. (2016) Device Data Ingestion for Industrial Big Data Platforms with a Case Study. Sensors 16: 279.