data symphony

Creating Business Value, Driven by Data Intelligence

CASE STUDY

Building a scalable and affordable Open-Source powered data and financial modeling platform

CASE STUDY

Case Study

Building a scalable and affordable Open-Source powered Data and Financial Modeling Platform

Discover an Open-Source powered platform, combining scalable data storage, flexible ETL processing, and automated financial modelling for seamless performance and cost efficiency.

CASE STUDY

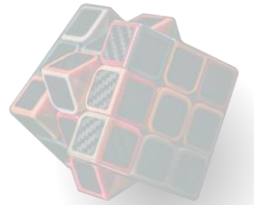
Overview

A fintech customer sought to establish a robust and agile platform that could seamlessly handle its essential data storage and financial modeling requirements. This platform was envisioned to be scalable, decentralized, and flexible, and high availability/uptime to support uninterrupted business operations. The primary objective was to craft a solution that not only provided state-of-the-art performance and rapid processing capabilities but also remained cost-effective, aligning with the financial realities.

Challenge

Data Storage and Management: The organization faced the critical challenge of developing a comprehensive data storage system capable of efficiently accommodating both unstructured and structured data. A robust data storage architecture was essential, designed to handle vast quantities of data seamlessly, whether deployed on-premises or in the cloud. Furthermore, the platform needed to incorporate structured database capabilities with cost-effective solutions for storing schema information about the data.

ETL (Extract, Transform, Load): The platform faced the critical challenge of efficiently processing both small and large datasets, necessitating a robust ETL framework capable of extracting, transforming, and loading structured and unstructured data seamlessly. This framework needed to support complex workflows that could dynamically scale based on fluctuating data volumes, ensuring optimal performance regardless of the dataset size.



Challenge

CASE STUDY

Financial Modeling: To support critical financial projections and strategic decision-making, the platform required an advanced system that could distribute complex financial calculations across multiple machines. This distributed computing capability was essential for achieving high efficiency and rapid processing of large-scale financial models, particularly under time-sensitive conditions. The challenge lay in ensuring that this architecture could handle the intricacies of financial data without introducing delays or errors, thereby maintaining the integrity and timeliness of insights that drive business success.

Automation: Automation emerged as a crucial element for enhancing operational efficiency. The challenge was to implement comprehensive automation, including data processing and financial modeling, while providing a user-friendly interface for developers to manage tasks in real-time. This required not only robust automation tools but also a seamless integration of these tools into existing workflows to minimize disruption and maximize productivity. The complexity of orchestrating these automated processes added another layer of difficulty, necessitating a solution that could adapt to evolving operational demands.

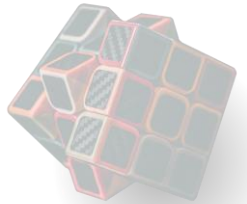
Data Interaction: The platform faced the pressing need for advanced OLAP (Online Analytical Processing) cube capabilities that would allow users to interact with processed data through familiar tools like Power BI and Excel pivot tables. The challenge was to ensure that these capabilities could handle vast amounts of data efficiently while providing actionable insights quickly. The integration of such functionalities demanded a robust architecture capable of delivering high performance under heavy analytical workloads, all while ensuring user accessibility and ease of use.

Solution

To effectively address the challenges and requirements outlined, a cost-effective, self-hosted solution was strategically implemented, leveraging industry-leading technologies designed to deliver exceptional scalability, flexibility, and performance while minimizing operational costs. The core components of this solution included Apache Hadoop, Apache Spark, PostgreSQL, and Apache Kylin, each selected for their unique capabilities to drive data-driven decision-making.

Data Storage - For scalable and flexible data storage, a Hadoop-based data lake was deployed to efficiently manage both structured and unstructured data:

- Hadoop was chosen for its high capacity and flexibility in handling vast amounts of unstructured data. This platform supports deployment options that include either file server-based systems or a Hive data warehouse for structured data management, ensuring adaptability to various data types.
- For structured data storage, PostgreSQL was integrated as a low-cost, open-source relational database system. PostgreSQL offers versatile deployment options—whether on-premise or in the cloud—while providing robust schema storage capabilities that facilitate efficient management and querying of structured datasets with minimal infrastructure costs.



Solution

CASE STUDY

ETL and Financial Modeling - The ETL (Extract, Transform, Load) processes and financial modeling functionalities were powered by Apache Spark, recognized as the industry standard for both small and large-scale ETL jobs:

- Apache Spark was selected for its unparalleled ability to handle extensive ETL tasks efficiently. By utilizing distributed computing, Spark enabled rapid data transformations and processing, adeptly managing complex workflows and large datasets with ease.
- Spark's SQL-based querying interface allowed users to access and analyze data intuitively, while its advanced capabilities supported machine learning operations for predictive analytics and deeper insights. This flexibility ensures that organizations can adapt their financial modeling efforts to changing business needs without compromising performance.
- For financial modeling specifically, Spark's distributed architecture facilitated the division of intricate financial calculations across multiple nodes. This ensured rapid, scalable, and accurate results while significantly reducing processing times and allowing the platform to handle high data loads without bottlenecks.
- Apache Spark was incorporated not only limited to ETL, but also for advanced data processing through its integrated machine learning library, MLlib. This feature allows organizations to perform predictive analytics and build machine learning models at scale, unlocking deeper insights from their data and enhancing decision-making processes.

Solution

Automation - To streamline operations and enhance efficiency in data processing and financial modeling workflows, automation was realized through a combination of Apache Airflow and Apache NiFi:

- Apache Airflow enabled robust task automation, scheduling, and orchestration of Python scripts. This powerful tool allowed for real-time monitoring and management of workflows, automating repetitive tasks such as data ingestion, model runs, and other critical steps in the pipeline. By reducing manual intervention, Airflow improved overall efficiency and reliability.
- Apache NiFi provided a user-friendly graphical interface for managing data flows from source to destination. This efficient tool allowed developers to visually design and automate the movement of data throughout the platform, ensuring seamless integration across various components.

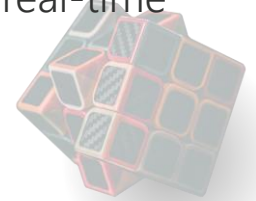
Interacting with Data and Results - For seamless interaction with processed data and financial results, Apache Kylin was deployed to provide OLAP (Online Analytical Processing) cube capabilities:

- Apache Kylin offered a cloud-compatible OLAP engine designed for efficient storage and rapid query processing. The platform compressed large datasets to optimize storage requirements while enabling fast real-time querying of information.
- Through its integration with Power BI, Excel pivot tables and cube formulae, Apache Kylin empowered users to interact effortlessly with extensive financial datasets. This functionality provided near-instant access to insights, facilitating dynamic reporting and analysis that significantly enhanced decision-making speed and effectiveness.

Outcomes

The solution delivered significant business value, successfully addressing the startup's requirements for scalability, flexibility, and cost efficiency. Key outcomes included:

- High availability: The platform achieved 99.9% uptime, ensuring continuous access to data and the uninterrupted processing of financial models.
- Scalable data storage: The Hadoop-based data lake and PostgreSQL implementation provided a robust foundation for storing both structured and unstructured data, offering low-cost, flexible storage options that could scale with business growth.
- Efficient ETL and financial modeling: Apache Spark enabled high-speed, distributed data processing and financial modeling, reducing the time required for data transformations and allowing the startup to handle complex financial computations with ease.
- Automated workflows: The integration of Apache Airflow and NiFi streamlined workflow automation, significantly reducing manual efforts and improving the efficiency of data processing pipelines and financial models.
- Real-time data interaction: With Apache Kylin, the startup gained the ability to interact with financial data in real-time through familiar tools like PowerBI and Excel, facilitating faster, more informed decision-making.



data symphony

Creating Business Value, Driven by Data Intelligence

GET IN TOUCH

ask@datasymphony.com

www.datasymphony.com