SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V10I11P102

Volume 10 Issue 11, 7-12, November 2023
© 2023 Seventh Sense Research Group®

Original Article

Implementation of a Data Lakehouse for Efficient Recovery and Processing of Massive Data

N'GUESSAN Behou Gérard[1], ASSIE Brou Ida[2], WAMBA Samuel Fosso[3], ACHIEPO Odilon Yapo Melaine[4]

[1]*Virtual University of Côte d'Ivoire, Research and Digital Expertise Unit (UREN), Abidjan, Ivory Coast.*
[2]*UFR Mathématiques et Informatique (UFRMI), Félix Houphouet-Boigny University, Ivory Coast, Abidjan, Ivory Coast.*
[3]*Toulouse Business School, France.*
[4]*Virtual University of Côte d'Ivoire, Research and Digital Expertise Unit (UREN), Abidjan, Ivory Coast.*

[2]*Corresponding Author : assie.ida@ufhb.edu.ci*

Received: 07 September 2023 Revised: 16 October 2023 Accepted: 04 November 2023 Published: 18 November 2023

Abstract - *In the field of Big Data, a recurring problem is that of rapid data processing with a view to their recovery and exploitation, in particular on dashboards. This problem leads to a shortening of decision-making and an inefficiency in using analytical solutions because of excessive latency times. The objective of this article is to set up a Big Data architecture capable of accelerating queries and processing on massive data in order to offer very good performance, in particular for real-time or near-real-time applications, regardless of the amount of data available and the rate at which new data is produced. This architecture is based on object storage, data virtualization and Data Lakehouse technologies. More specifically, it is based on the MinIO and Dremio technologies, which allow optimization mechanisms useful for achieving the defined objectives, particularly the reflection mechanism of Dremio. The combination of these technologies has made it possible to develop dashboards with very low latency with global COVID-19 data.*

Keywords - *Big data, Object storage, Data virtualization, Data lakehouse, Technology.*

1. Introduction

Digitalization is a socio-technical process that allows to benefit from digitized elements to develop new organizational procedures, business models, or commercial offers [1]. It has taken off in all fields, including health. This process has led to generating a large amount of data [2]. These volumes of data represent major technological challenges in terms of storage, processing, and exploitation [3] [4]. It was in this context that, during the Coronavirus pandemic, each country began collecting data on the pandemic. Many statistics have been produced and analyses conducted as part of epidemiological pandemic management.

However, when all these data are centralized, their volume makes their storage, processing, and analysis complex. Indeed, storage formats are not uniform and automated analytical solutions are either paper-based or in the form of web applications that cannot present all global data simultaneously on a dashboard. The data are thus presented, in the best case, by country, which does not allow overall comparisons of the situations of several countries. This draft article was initiated to provide a viable technical solution.

The goal is to provide a solution that not only stores all global COVID-19 data in a single coherent system but also both the available data and the data that continues to be collected. But also make it possible to use all the global COVID-19 data to develop automated dashboards. To do so, the article is subdivided into three (3) parts. The first presents the different reference data storage architectures for developing data-based solutions or services. The second part presents the proposed architecture and the various components of it. Finally, the third part presents the architecture's implementation and the results obtained.

2. State of the Art

Developing data-centric (data-centric) applications has given rise to different data storage architectures. These architectures have been the subject of much research and are presented in three forms: Data Warehouse, Data Lake, and Data Lakehouse. This includes work by NAMBIAR et al., SCHOLLY et al., TALHA et al., and LOPEZ et al., which provides a detailed overview of the roles of data warehouses and data lakes in modern enterprise data management [3, 4-8]. SCHOLLY et al. proposed a generic MEDAL-built model and criteria for assessing the metadata system. This model allows you to manage metadata from a data lake through a list of features. It is a model of the graph-based metadata system [8]. Michael et al. do a comparative study of storage systems and emphasize Lakehouse data [9]. ZHAO et al. developed a system using the platform's middleware. This platform has several services for parallel processing and multi-cloud capability. Multi-cloud speeds up the data processing pipeline and article categorization process using machine learning on a hybrid cloud [10]. Ericka et al., Maltais et al., and Renard et al. presented the various technological innovations to combat COVID-19 [11-13]. Cepeda et al. developed a forward-looking scoreboard. This table is based on tools from business intelligence (BI), Docker and Dremio to facilitate decision-making. It also improves the customer service of the Guayaquil Business Unit branches [14].

However, the digitalization used mainly in the health field generates a large amount of data. However, the configuration of the architecture of the literature struggles to manage this mass of data produced every day. Some platforms are obliged to present only a few data to facilitate their loading. Also, these data are available in different forms (quantitative, qualitative, and textual). This mechanism causes a complexity of storage, processing, and analysis. There are latencies in the loading of data in automated reports and dashboards. Hence, the interest of proposing an architecture capable of facilitating the presentation and loading of Covid-19 data.

3. Architecture

The proposed architecture follows the operating mechanism in Figure 1 below. This architecture enables the storage, processing, and exploitation of global COVID-19 data. This is a Data Lakehouse architecture based on object storage in figure 1.

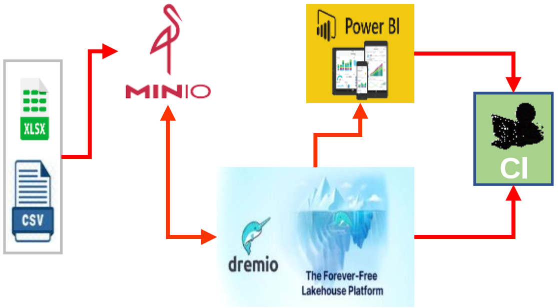


Fig. 1 Proposed architecture

This Lakehouse data architecture, as shown in the operating process in Figure xx, not only stores all the global COVID-19 data but also facilitates the processing and exploitation of this data. The stored data are managed as autonomous and discrete units called objects. Indeed, instead of using a table format, a data virtualization solution (Dremio) has been proposed to make directly usable the structured data directly available in the object storage system (MinIO) in a very efficient way. This architecture contains three main sections:

- Object Storage System: In this system, data is stored and managed as stand-alone units called objects. Each unit has a unique identifier or key that allows it to be found wherever it is stored in a distributed system. This data stored as objects guarantees data availability, search capability and enhanced data security as it protects data from accidental deletion or corruption. Here, MinIO, being open source, is also faster, with a read and write speed ranging from 171 GB per second to 183 GB per second on a standard computer. Its performance features, combined with its compatibility with the Amazon S3 file system, have made it the standard for Artificial Intelligence, Machine Learning and Data Science applications.

- Data Virtualization: This section allows you to access data without transferring it. The virtualization platform Dremio the platform offers an interface with many storage technologies traditionally used in Big Data, including Elasticsearch, Hive, HDFS and Amazone S3. The fact that Dremio supports Amazone S3 makes it compatible with MinIO. The Amazone S3 connector is therefore used to connect Dremio to MinIO to virtualize the data stored in MinIO and their use by Dremio customers. Dremio also provides a web interface for building virtual datasets that can be manipulated with SQL queries. It can generate queries itself from a visual manipulation of the data. In addition to this, Dremio contains query optimization mechanisms called reflections. They are used to speed up requests by using one or more criteria on the data to be used. When optimizing queries, data processing jobs are generated and orchestrated automatically by the platform. It is very important to understand that Dremio does not store data. It not only provides optimal access to the data but also allows processing of the data as would be done with any ETL. The performance of Dremio can be increased by deploying it in cluster mode. A Dremio cluster can contain hundreds or even thousands of nodes. Data Visualization: Data visualization is a very important approach to data analysis. In this work, the choice is made on the Open-Source Power BI Desktop solution. It allows you to create professional-quality, personalized and interactive data visualizations. It has the advantage of providing a sufficiently simple interface for end users to create their own automated reports and dashboards. Power BI natively provides a connection interface to Dremio, making it an ideal tool for this article. Visualizations developed with Power BI Desktop can be easily industrialized in two ways:

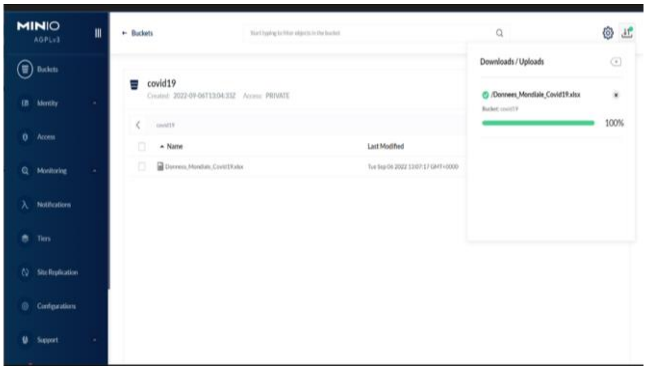
- Or rent a publishing space on an online Power BI (Cloud) server to publish its automated reports and dashboards.

- Or buy a Power BI server and deploy it yourself (on-premise) to publish its automated reports and dashboards.

- Any automated reports and dashboards published on a Power BI server (cloud or on-premise) are not only accessible via a web browser but also using the Power BI mobile client available for Android and iOS phones.

4. Implementation

As part of the implementation, we used an online dataset with a capacity of 36.1MB and containing 214,478 observations. In the context of Big Data, although these data are not large enough, they already pose the problem of a relatively large loading time on the dashboards when several variables are used simultaneously on the dashboards. So, we are dealing with the velocity problem and not with the volume of data, even though the two issues are still related. In practice, Big Data clusters can be sized according to the results obtained in a single location with 36 MB of data. This data was obtained from the Covid-19 global data collection platform (https://ourworldindata.org/). This dataset comprises 63, including the continent, the date of registration and the number of deaths of the day. To facilitate the storage of datasets, a Buckets account is created in the MinIO web interface where Covid-19 data is deposited see Figure 2 below. Buckets are used to organize objects. A Space account is created in the Dremio web interface and connected to MinIO via the Amazone S3 connector. This Space account is used to create virtual Datasets. Dremio offers the possibility to do all kinds of data processing.



covid19
Created: 2022-09-06T13:04:33Z Access: PRIVATE

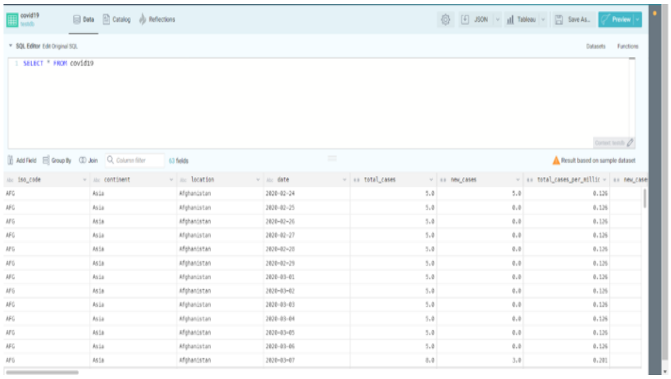
covid19

Name	Last Modified
Donnees_Mondiale_Covid19.xlsx	Tue Sep 06 2022 13:07:17 GMT+0000

/Donnees_Mondiale_Covid19.xlsx
Bucket: covid19
100%

Fig. 2 Global COVID-19 data in MinIO

Exploiting the global COVID-19 data stored in MinIO requires an optimized virtual version available on Dremio. A connection is thus established between Power BI and Dremio. This connection makes the Virtual Datasets data available in Power BI, as shown in Figure 3.

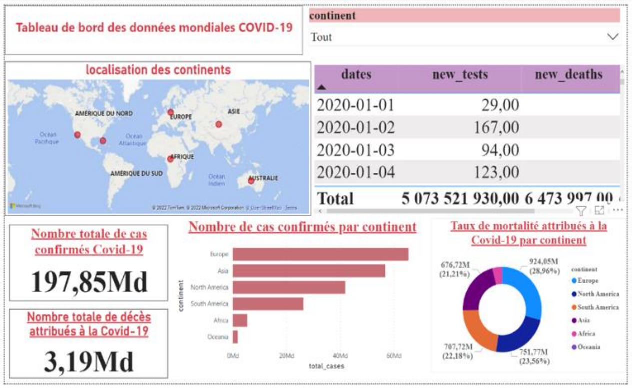


```
SELECT * FROM covid19
```

iso_code	continent	location	date	total_cases	new_cases	total_cases_per_millic	new_case
AFG	Asia	Afghanistan	2020-02-24	5.0	5.0	0.126	
AFG	Asia	Afghanistan	2020-02-25	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-02-26	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-02-27	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-02-28	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-02-29	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-01	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-02	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-03	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-04	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-05	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-06	5.0	0.0	0.126	
AFG	Asia	Afghanistan	2020-03-07	8.0	3.0	0.201	

Fig. 3 Creating reflections in Dremio

This virtual dataset (COVID-19) has created reflections to make it possible for Power BI to use it efficiently and optimally. The data available in Power BI allows you to build two dashboards. The first concerns global data with filters by continent, as shown in the following Figure 4:



dates	new_tests	new_deaths
2020-01-01	29,00	
2020-01-02	167,00	
2020-01-03	94,00	
2020-01-04	123,00	
Total	5 073 521 930,00	6 473 997 00

Fig. 4 Global COVID-19 data dashboards by continent

This dashboard contains a map showing the areas most affected by the pandemic on each continent. Then, there is a table that gives, on a daily basis, the number of new cases and the number of new COVID-19 deaths on each continent. Two boxes show the total number of cases detected worldwide and the total number of deaths due to COVID-19 since the start of the pandemic. Also, on this dashboard, a strip chart shows the overall number of confirmed COVID-19 cases per continent, and a pie chart shows the overall mortality rate attributed to COVID-19 on each continent.

Finally, the dashboard contains a filter on the continents to obtain all the visualization elements presented for one or more selected continents. The response time of this dashboard can be measured with the Power BI Performance Analyzer feature.

A second automated dashboard concerns global data with country filters, as shown in Figure 5 below.

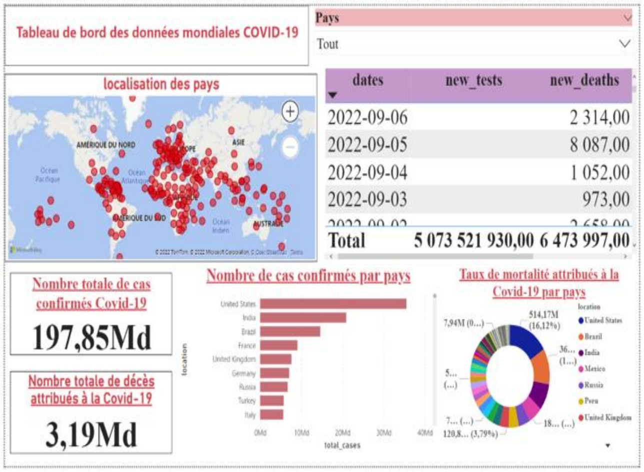


Fig. 5 Global COVID-19 data dashboards by country

Similar to the continental presentation table, the country presentation table contains a map showing the areas most affected by the pandemic in each country. Then, for each day, a table shows the number of new cases and the number of new COVID-19 deaths in each country. Also included in this chart is a bar chart showing the overall number of confirmed COVID-19 cases by country and a pie chart showing the overall mortality rate attributed to COVID-19 in each country.

Finally, the dashboard contains a country filter to obtain all the visualization elements presented for one or more selected countries. Like the first dashboard, the response time of this dashboard can be measured thanks to the Power BI Performance Analyzer feature.

5. Evaluation

Data records in Dremio are created, not copies of data, but simple virtual Datasets. These can be compared to views of relational DBMS.

The Dremio Virtual Datasets can be exploited as if they were real data. But, the volume of data makes their exploitation problematic. Dremio has optimization mechanisms called reflections, which are used to accelerate data loading. These should be defined on the variables to be used in analytical solutions as shown in Figure 6 below:

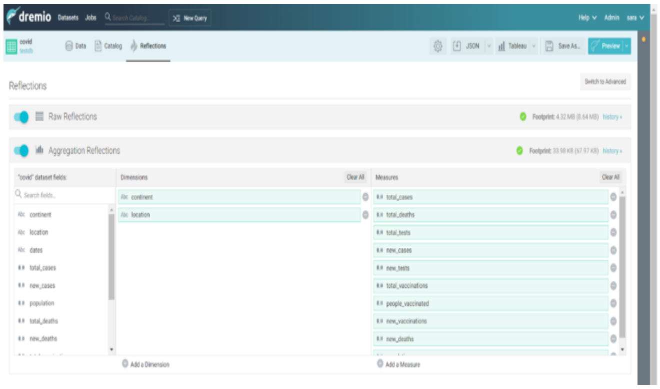


Fig. 6 Creating reflections in Dremio

After starting the recording, we can see the data query time and the display of the results for each visual with and without the Dremio reflections. Table 1 below presents a comparison of the duration of the response times for each visual before and after the activation of Dremio reflections in the dashboards with continent filter.

After analyzing the dashboard, we see a reduction in response time or a display of each visual in the dashboard with the activation of Dremio reflections. It averages a reduction rate of over 90% compared to the dash without the Dremio reflections. Table 2 below shows a comparison of the duration of response times for each visual with and without Dremio reflections in country-filtered dashboards.

After analysing the dashboard, we notice a reduction in the response time or a display of each visual in the dashboard carried out after the data processing. It has an average reduction rate of over 91% compared to the Dashboard without the activation of reflections in Dremio processing.

Table 1. Evaluation of loading times (in ms) on the dashboard with continent filtering

	Without Dremio Reflections	With Dremio Reflections	Difference	Percentage (gain)
Geographic map	32548	794	31754	97.56%
Painting	32024	704	31320	97.8%
Bar chart	32282	693	31589	97.85%
Circular diagram	31798	644	31154	97.97%
Text zone	101	26	75	74.26%
Filtered	101	27	74	73.27%
Box showing the total number of cases	29194	417	28777	98.57%
Box showing the total number of deaths	29308	459	28849	98.43%

Table 2. Evaluation of loading times (in ms) on the Dashboard with country filtering

	Without Dremio Reflections	With Dremio Reflections	Difference	Percentage (gain)
Geographic map	33196	391	32805	98.82%
Painting	No display	522	522	100%
Bar chart	No display	475	475	100%
Circular diagram	19894	527	19367	97.35%
Text zone	107	36	71	66.36%
Filtered	110	37	73	66.36%
Box showing the total number of cases	29194	411	28783	98.59%
Box showing the total number of deaths	29308	432	28876	98.53%

6. Conclusion

In this work, we proposed and implemented a Data Lakehouse architecture. It is a Big Data architecture combining object storage via MinIO and data virtualization via Dremio to set up very efficient Lakehouse Data. Global COVID-19 data was used as a data source.

Data was first loaded into MinIO. Then, using MinIO as a data source, virtual datasets were created in Dremio. These virtual datasets developed automated dashboards of global COVID-19 data using Power BI.

Tests conducted with Power BI have shown that the proposed architecture makes it possible to directly exploit global COVID-19 data from any continent and any country. Also, this architecture gives a single-user version a response time of a few milliseconds, which is almost real-time. The proposed technologies, namely MinIO and Dremio, being natively distributed systems; the use of a MinIO cluster coupled to a Dremio cluster is a device that could revolutionize the universe of data-oriented applications.

Data Engineers can improve this work by automating data loading in MinIO.

References

[1] Aymen Elhali, Imane El Yamlahi, and Amine Nabil Bouayad, “The COVID-19 Crisis is a Boost to Digital Transformation in Morocco,” *French Review of Economics and Management*, vol. 4, no. 2, 2023.

[2] Elizabeth Couzineau-Zegwaard, “The Impact of Digitalization on the Supply Chain Business Ecosystem,” *The Journal of Management Sciences*, vol. 301302, no. 1, pp. 85-97, 2020.

[3] Mohamed Talha, *Big Data between Quality & Data Security*, Doctoral Thesis, Cadi Ayyad University (Marrakesh, Morocco), 2022.

[4] Lotfi Benazzou, and Nabaouia Bennia, Covid-19 and Management Control, Controlling, Accounting and Auditing Journal, vol. 5, no. 3, 2021.

[5] Athira Nambiar, and Divyansh Mundra, “An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management,” *Big Data and Cognitive Computing*, vol. 6, no. 4, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Etienne Scholly, “*From Metadata Modeling to the Conception of a Data Lake: Application to Public Housing*,” Doctoral Thesis, University of Lyons, 2022. [Google Scholar] [Publisher Link]

[7] Lopez Chavez, and Marina Adriana, “*Proposal of a Research Data Management Platform and Its Adoption by Researchers in Environment and Computer Science in the Context of Forestry Research in Quebec*,” Doctoral Thesis, Tele-university, 2022.

[8] Etienne Scholly et al., Metadata Systems in Data Lakes: Modeling and Functionality.

[9] Michael Armbrust et al., “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics,” *CIDR Proceedings*, 2021. [Google Scholar] [Publisher Link]

[10] Jie Zhao, Maria A. Rodriguez, and Rajkumar Buyya, “High-Performance Mining of COVID-19 Open Research Datasets for Text Classification and Insights in Cloud Computing Environments,” *IEEE/ACM 13th International Conference on Utility and Cloud Computing, IEEE*, pp. 302-309, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ericka Y. Bracamonte-Arámburo, and Guillermo Foladori, “Adverse Effects of COVID-19 Protective Masks: Conflictive Cases,” *Research and Science of the Autonomous University of Aguascalientes*, no. 85, 2022.

[12] Stephanie Maltese, “Descriptive Study of Canadian Humanitarian Agility and Resilience in the Time of COVID-19,” *Canadian Journal of Development Studies*, vol. 43, no. 4, pp. 468-486, 2022. [Publisher Link]

[13] Damien Renard, Collaboration on Social Innovation Platforms: The Case of “Solidarity Covid-19 Francophonie”, Communication, Technologies and Development, no. 10, 2021.

[14] José Espinoza Cepeda, Alejandra Colina Vargas, and Marcos Espinoza Mina, Improving Customer Service through the Application of Business Intelligence Tools.