# OMOP on a Data Lake: Addressing the Critical Need for Scalable Solutions in Healthcare Data Management with OHDSI Tools and AWS Services

**Lance Eighme, Lisa McEwen, Simon White, Tobias Cauoette, Oliver Tucher, Anna Swigart**
**Helix, Inc. 101 S Ellsworth Ave #350, San Mateo, CA 94401, United States**

## Background

The OHDSI community has made significant strides in standardizing healthcare data through the OMOP CDM and stack of open-source tooling for data quality and analytics. These tools have facilitated large-scale observational studies, but managing and analyzing vast and complex healthcare datasets remains challenging and is only becoming more and more of a critical need with the increasing digitization of health records. Prior literature has explored the benefits of cloud-based infrastructures for big data analytics, highlighting the potential of such systems to handle large-scale data efficiently. Our project leverages AWS' scalability, performance, and governance capabilities to enhance healthcare analytics, with one project focusing on developing a comprehensive dataset of over 10 million patients and billions of records across multiple US health systems. This novel integration addresses the critical need for scalable solutions in healthcare data management, offering significant benefits to the OHDSI community.

## Methods

We implemented an architecture combining OHDSI tools with AWS services to create a scalable data lake environment. Our workflows include ingesting data across multiple health systems into Apache Iceberg tables, a table format optimized for high-performance big data handling and access[1]. We also employ AWS Data Quality Definition Language (DQDL)[2], a customizable data quality suite that enables scalable and automated quality control management for the OMOP CDM to evaluate referential integrity constraints, data types, uniqueness, as well as SQL-based checks. For our ETL processes, we have integrated AWS Glue jobs[3], which leverage Apache Spark to perform distributed data processing tasks efficiently. Implementing an access layer to the data that leverages AWS Lake Formation[4] has also allowed us to streamline data lake creation and management, ensuring robust data security and governance to the data from inside our organization as well as by health system partners. With Lake Formation, fine-grained access policies and comprehensive auditing apply uniformly to Apache Spark workflows, Athena SQL queries, and Redshift queries accessing this data. Amazon Redshift Spectrum[5] was utilized without any data duplication to query the data for analytical workflows and to connect with OHDSI tools, such as Data Quality Dashboard[6], Achilles[7], and ATLAS[8]. This infrastructure supports the integration and analysis of data from several US health systems, contributing to a combined dataset of over 10 million patients and billions of records.

## Results

Deploying OMOP on a scalable data lake architecture with AWS significantly enhanced the efficiency and scalability of healthcare data management and analytics. AWS Lake Formation and Iceberg facilitated seamless data ingestion and governance, ensuring compliance with data security standards. The use of AWS Glue DQDL automated and scaled data quality management, and with DataQualityDashboard we have been able to effectively identify data quality issues to ultimately flag for remediation. Analytical operations in Amazon Redshift exhibited substantial performance improvements, enabling faster and more complex queries. This approach also facilitated the creation of a novel dataset comprising several millions of patients across multiple US health systems, showcasing the system's scalability and its capacity to handle massive volumes of data efficiently.

## Conclusion

Deploying OMOP on a scalable data lake architecture using AWS services represents a significant advancement in healthcare analytics. This integration enhances the scalability, performance, and quality of data analysis while supporting the creation of a standardized, comprehensive dataset to efficiently manage the rapidly-growing clinical data across several US health systems. By leveraging scalable infrastructure services available within AWS, analysts and researchers can conduct timely and in-depth inquiries into these data, pushing the boundaries of observational health data sciences and fostering new insights into healthcare outcomes. This approach offers a robust and modern solution to the OHDSI community, addressing critical needs in large-scale data management and analysis.

1. Ryan Blue, Fokkema S, Vander V, et al. Apache Iceberg: A High-Performance Table Format for Huge Analytic Datasets. Apache Software Foundation. Available from: https://iceberg.apache.org/.
2. AWS Data Quality Definition Language. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/glue/latest/dg/dqdl.html.
3. AWS Glue. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/glue/
4. AWS Lake Formation. Amazon Web Services (AWS). Available from: https://aws.amazon.com/lake-formation/
5. AWS Redshift Spectrum. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum.html
6. Blacketer C, Schuemie FJ, Ryan PB, Rijnbeek P (2021). “Increasing trust in real-world evidence through evaluation of observational data quality.” Journal of the American Medical Informatics Association, 28(10), 2251-2257. Version 2.6.0.
7. DeFalco F, Ryan P, Schuemie M, Huser V, Knoll C, Londhe A, Abdul-Basser T, Molinaro A (2023). Achilles: Achilles Data Source Characterization. R package version 1.7.2.
8. OHDSI ATLAS. Observational Health Data Sciences and Informatics (OHDSI). Available from: https://atlas.ohdsi.org/.