Submission code: K94Z6LVB9D

See this and other abstracts on our website: https://home.sbgf.org.br/Pages/resumos.php

Massively Parallel Distributed Computing and Agentic Framework for Subsurface Data Management and Analysis

Vikram Sahukar (Petrabytes Corp), Sashi Gunturu (Petrabytes Corp), Vineet Patel (Petrabytes Corp), Atharva Gadad (Petrabytes Corp), Juliana Fernandes (IesBrazil Technology & Innovation), Rodrigo Eiras (IesBrazil Technology & Innovation), Débora Barretto (IesBrazil Technology & Innovation)

Massively Parallel Distributed Computing and Agentic Framework for Subsurface Data Management and Analysis

Abstract

The energy sector confronts a significant challenge in managing and extracting value from ever-expanding multidimensional datasets, with seismic and sensor data reaching petabyte scales. Legacy software applications lack the scalability and interoperability to handle this volume and variety, leading to disjointed workflows and the underutilization of valuable data. This work introduces a solution that leverages a massively parallel distributed computing environment on AWS, combined with Generative AI (GenAI), to enable scalable and intelligent energy data flow orchestration. The proposed architecture utilizes AWS EMR running on EKS to process petabyte-scale data and employs GenAI for automated metadata mapping, enrichment, and validation against the OSDU® data standards. Similar architecture also works for other cloud services providers, such as Microsoft. This framework is implemented using the PIEScale solution by Petrabytes, incorporating a template-based approach to parse unstructured data and a comprehensive business rules engine to ensure data quality and compliance. This architecture enables operators to migrate their data to OSDU® efficiently. It is designed to be compatible with various data formats that are widely used across the industry. Users can choose to use one of the many pre-existing pipelines that come standard with the architecture or create their own for proprietary/unstructured data formats. It has been developed with scalability in mind, allowing for computational scaling to handle large volumes of data. This architecture has been tested successfully on multiple public datasets, including the Volve dataset, Gulfacs datasets, different public datasets and simulated test data for about 100,000 wells and 100 TB Synthetic Seismic data. This work presents a case study that was undertaken with a major North American operator demonstrating the solution's efficacy, successfully ingesting 2.9 TB of seismic data and data from 309 wells in under three hours—a process that previously took several weeks. This transformative approach unlocks the value of both archived and real-time data, creating a seamless, collaborative subsurface workflow experience and empowering more holistic and timely decision-making throughout the asset life cycle. With this implementation, energy customers can gain value in the form of Innovation, global collaboration, data quality and standardization, Gen AI transformations and Agentic workflows.