VisiLake

Self-Service Data Lake Exploration & Analysis on AWS

Daniil Aleshechkin, Ben Hajdukiewicz, Adam Metz

Project Background

K + S Potash Canada (KSPC), a leader in Saskatchewan's potash industry, faces a significant challenge: an overwhelming amount of unorganized and difficult to access data. Without proper sorting, processing, or exploration, this data remains largely untapped, hindering operational efficiency. Consequently, valuable insights that could drive improvements and innovation within the organization are often overlooked.

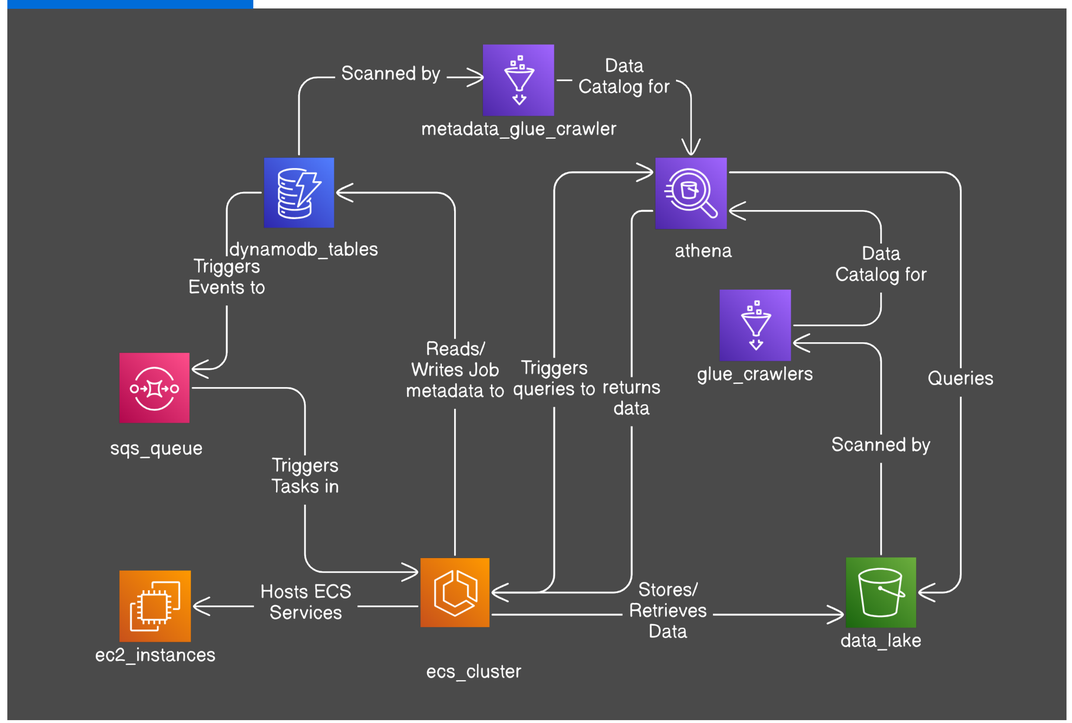
What is a Data Lake?

A data lake is a storage repository that holds vast amounts of raw data in its native format until required for analysis or processing. Unlike traditional databases, which require data to be structured beforehand, a data lake stores data in its original form, allowing for flexibility and scalability.

Our Solution

- We've developed a web application specifically designed to meet the needs of KSPC, offering a no-code application for the users
- This tool facilitates streamlined access to data from a data lake, enabling users to effortlessly retrieve the information they need
- Users can conduct analyses directly within the application, using various data analysis methods
- Automatic generation of visuals and metrics provides quick insights into the data
- By simplifying the process and removing technical barriers, we ensure efficient and easy data utilization, supporting the company's goals of operational excellence

Architecture



Process Overview

Data Request
User submits a customizable request for data retrieval and processing

Data Processing
Data is fetched from the data lake, and processed accordingly

View Results
Users can easily view and interact with the fetched and processed data via an automatically generated report

Power BI Integration
Users can also seamlessly pull the data into Power BI

Future Work

- Broaden and enhance the variety of available data analysis methods
- Provide users the option to fine tune and tailor each data analysis method
- Integration with KSPC's data lake and authentication system

Tech Stack

Web Development

- React.js for frontend development
- tRPC for API communication

Infrastructure

- AWS for cloud services
- Localstack for local emulation of AWS

Deployment & Containerization

- Nix for dependency management and environment reproducibility
- Docker for containerization

Data Processing

- Rust for data analysis method processing

Acknowledgements

- Dr. Timothy Maciag
- Dr. Karim Naqvi
- K+S Potash Canada

University of Regina | FACULTY OF ENGINEERING & APPLIED SCIENCE

VisiLake GitHub