International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

A Data Mesh-Driven Data Lake Architecture for Oil Field Data Consolidation

Gaurav Kumar Sinha

Amazon Inc.
Email: *kr.gaurav1314[at]gmail.com*

Abstract: ***The proliferation of distributed data sources and specialized applications across different oil field operations has led to data silos and integration challenges. This hamper discovering insights across disconnected datasets. To consolidate the vast data landscape, traditional centralized approaches prove inadequate. I propose a decentralized data lake architecture aligned with data mesh principles for oil data consolidation. Self-service data infrastructure is provisioned through reusable data products that encapsulate domain-specific datasets with relevant policies and tooling. This encapsulation aids discoverability and access. Standardized interfaces between the data products using a domain ontology and catalog facilitate consolidation at scale while distributing data governance. My reference architecture deploys on AWS leveraging analytics services like Redshift, Athena, and SageMaker. The mesh of domain data products provides scalable data consolidation and governance while accelerating analytics velocity through easier findability, accessibility, and understandability of datasets. I demonstrate applicability across key use cases like production optimization, drilling plan analysis, and equipment failure prediction for an oil field. This decentralized governance approach unlocks value from distributed data while retaining organizational flexibility and autonomy.***

Keywords: Data mesh, data lake, data consolidation, data governance, decentralized architecture, domain-oriented design, self-service analytics, data discoverability, data accessibility, data understandability, oil and gas industry, exploration and production data, seismic data, well logs, sensor data, data interoperability, data silos, knowledge graphs, ontologies, reference architecture, AWS cloud, Amazon Redshift, Amazon Athena, Amazon SageMaker

1. Introduction

The oil and gas industry generate vast amounts of complex data across upstream, midstream, and downstream operations. Seismic surveys, well logs, reservoir models, equipment sensors, and more produce petabytes of data distributed across functional silos [1]. However, insights often remain trapped within data and application boundaries of exploration, drilling, production, and distribution functions [2]. This hampers cross-domain analytics to optimize field planning, predict failures, estimate reserves, and track end-to-end operations. Prior consolidation approaches using rigid centralized data warehouses prove inadequate for the scale, diversity, and changing dynamics of oil field data [3].

This calls for flexible consolidated data platforms without creating tightly coupled dependencies. I propose a decentralized Data Mesh architecture aligned to domains that facilitates scalable consolidation of distributed datasets [4]. Encapsulated domain data products with self-service tooling aid findability, accessibility, and understandability while easing governance [5]. I demonstrate data mesh realization on AWS for oil data consolidation. Standard data formats and interfaces aided by AWS analytics services like Redshift, Athena, and SageMaker to accelerate domain-specific analytics across key oil field use cases [6].

2. Problem Statement

The oil and gas industry today grapple with massive, distributed and heterogeneous data sources across upstream, midstream and downstream operations. Data is locked away in operational siloes across seismic, well logging, reservoir simulation, production telemetry, equipment maintenance and more Domain-specific formats, semantics, access restrictions and technology variances exacerbate these data siloes [7]. Though huge volumes of data are generated and managed, getting a consolidated 360-degree view remains challenging [8]. This severely impedes cross-domain analytics across the E&P lifecycle. Questions around optimal well placements by correlating seismic data with production telemetry, analyzing drilling equipment failure patterns across fields, or tracking end-to-end production operations require unified data analysis. Such analysis is either not possible or requires enormous manual efforts today. Rigid centralized data warehouses are also poorly suited to the evolving nature of oil field data landscape and user needs.

The above limitations highlight the need for a flexible, scalable consolidation of oil field data sources providing easy discoverability, accessibility and understandability of distributed data to accelerate analytics [9]. This serves as the key problem motivator that i attempt to address in this paper through a decentralized data mesh architecture.

3. Solution

To address the complex data challenges for discovery, access and analytics in oil and gas, I propose a decentralized data mesh architecture aligned to domain areas.

a) Reference Architecture
- Self-serving domain data products with standard interfaces
- Separated data management responsibilities per domain
- Unified global catalog providing consistency

b) Domain Data Products
- Encapsulate datasets for specific domains
- Focused data governance policies and access control
- Facilitate discoverability and understanding

c) Data Delivery Pipelines

- Pipelines for ETL from data sources to domain products
- Handle variety of formats - structured (relational), semi-structured (geospatial, logs)
- Mapping and transformation leveraging ontology

d) Unified Catalog

- Global catalog referencing published domain data products
- Standard taxonomy
- Enable interoperability across domains

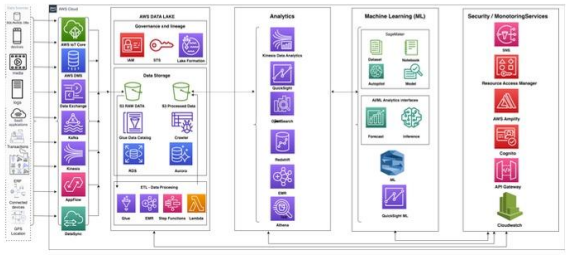
e) Reference Implementation

- Realization on AWS leveraging S3, Athena, Glue, SageMaker, Data Zone
- Usage demonstration for oil analytics spanning domains

This empowers decentralized teams to build self-operating domain data products while enabling organization-wide interoperability, consolidation and governance [10].

4. Architecture Overview

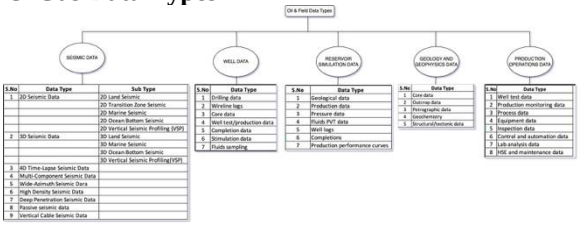
Producer Domain - Entities or teams that create, source, and provide data, ensuring quality and accessibility for downstream consumption and analysis.



Consumer Domain - End users or systems utilizing data products, focusing on analysis, reporting, and insights derived from multiple data sources.

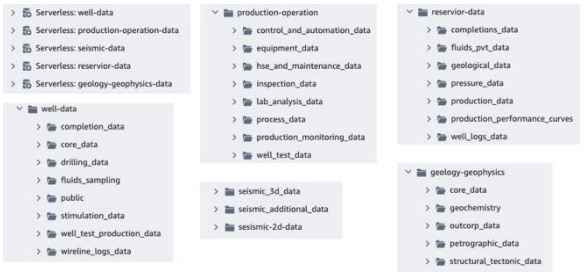
Central Governance - Oversight body that enforces policies, standards, and quality across all domains, maintaining coherence and compliance within the data ecosystem.

Oil & Gas Data Types

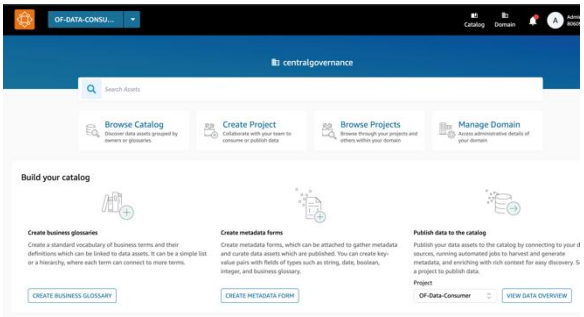


Source Data in S3

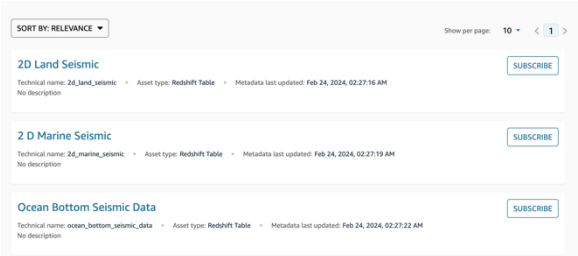
Ingested Data in Redshift



Central Governance



Central Catalog



Uses

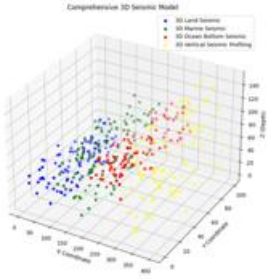
A key motivation for a consolidated data environment is to enable cross-domain analytical use cases that span data siloes. Below are tested analytical use cases

5. Exploration Phase

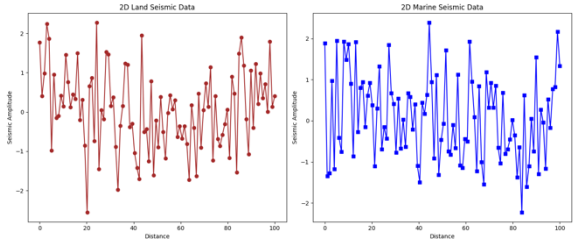
5.1 Seismic Data Analysis

To construct a comprehensive 3D seismic model of the subsurface, integrate data from 3D land, marine, ocean bottom, and vertical seismic profiling. This integrated

approach enhances the identification of potential hydrocarbon reservoirs.

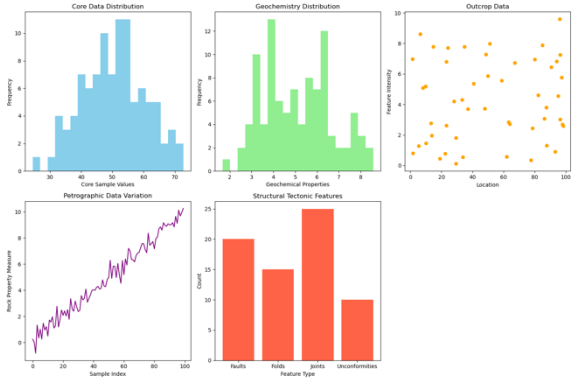


Merge 2D land and marine seismic data to facilitate early-stage exploration efforts.



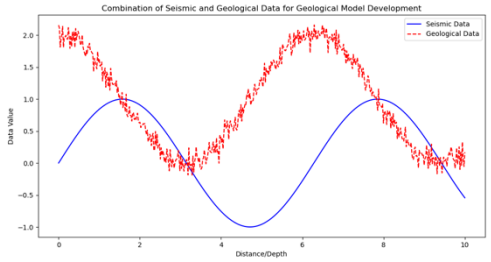
5.2 Geology and Geophysics

Analyzing core_data, geochemistry, outcrop_data, petrographic_data, and structural_tectonic_data provides insights into subsurface geology, pivotal for evaluating reservoir size and quality.



- geological model

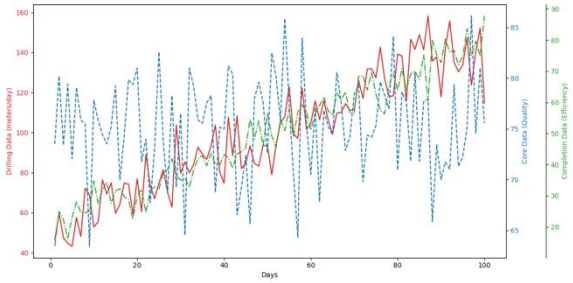
Integrate seismic and geological data for a comprehensive understanding of subsurface structures and properties.



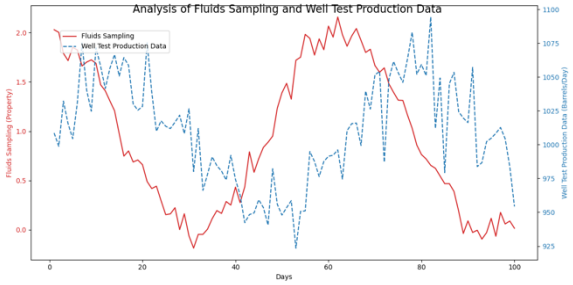
5.3 Drilling Phase

5.3.1 Well-Data Analysis

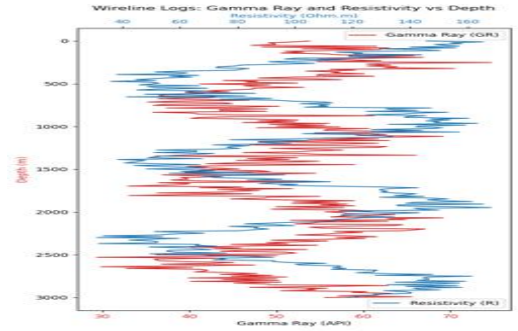
a) The integration of completion_data, core_data and drilling_data aids in optimizing well drilling and completion methodologies.



b) Fluid sampling and well test production data offer valuable insights into the composition of fluids and their performance during production operations.



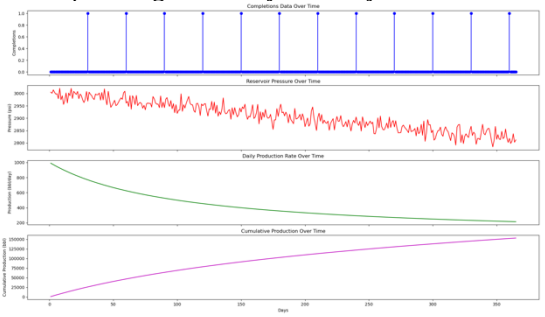
c) Wireline logs data facilitates real-time decision-making during drilling, enabling adjustments to techniques and enhancing the accuracy of the geological model.



5.3 Production Phase

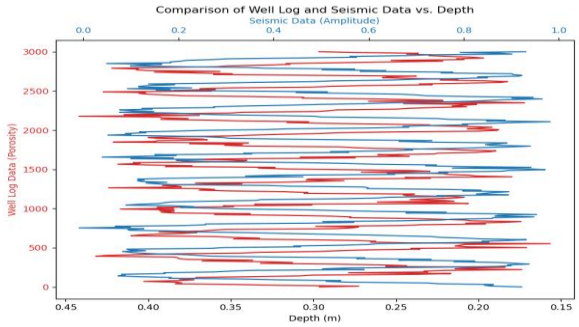
5.3.1 Reservoir and Production Data

a) The integration of completions_data, pressure_data, production_data, and production_performance_curves enables continuous monitoring and optimization of production, facilitates future performance forecasting, and assists in planning secondary recovery methods.



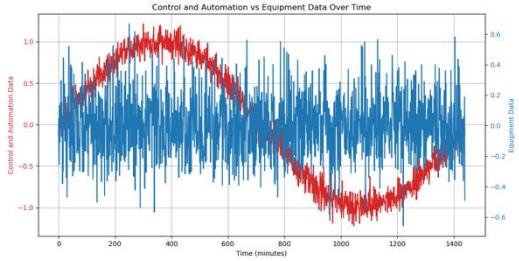
b) Analysis of well logs data from production wells, coupled with seismic data, enhances the refinement of the reservoir model and deepens understanding of the reservoir's

performance.

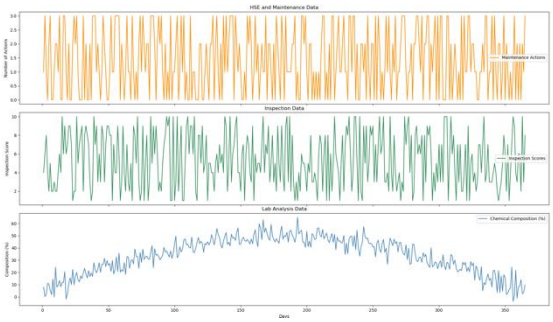


5.3.2 Production Operation

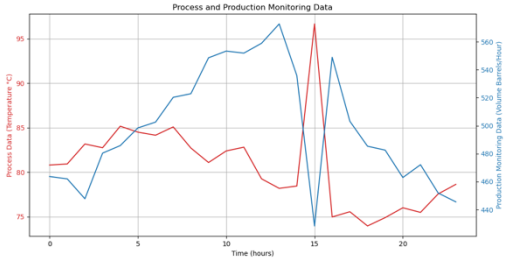
a) Utilizing both control and automation data alongside equipment data aids in optimizing production operations, ensuring the integrity, and enhancing the efficiency of the production system.



b) HSE and maintenance data, inspection data, and lab analysis data play pivotal roles in upholding safety standards, ensuring equipment reliability, and maintaining environmental compliance.



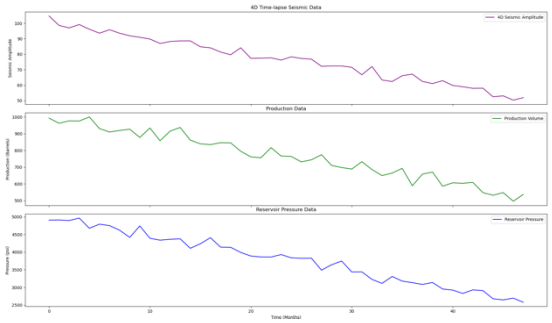
c) Process data and production monitoring data facilitate real-time monitoring of production, enabling the detection of anomalies, prediction of equipment failures, and identification of opportunities for process optimization.



5.3.4 Integrated Data Analysis

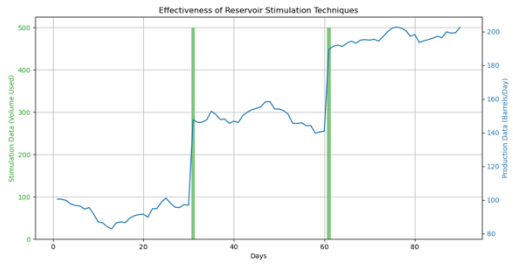
a) 4D Seismic Monitoring

By leveraging 4D time-lapse seismic data, you can monitor reservoir changes over time attributed to production activities. Integration with production data and pressure data enhances comprehension of reservoir dynamics and enables effective reservoir management.



b) Enhanced Recovery Technique

Data from stimulation operations is crucial for planning and assessing the effectiveness of reservoir stimulation techniques like hydraulic fracturing or acidizing.



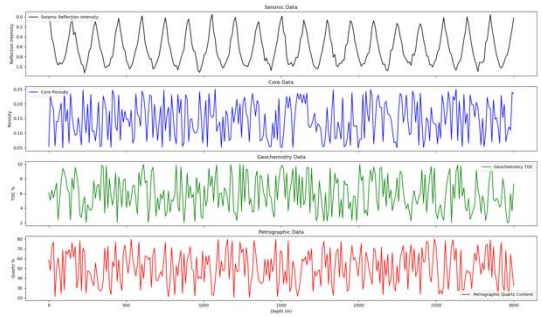
c) Health, Safety, and Environment (HSE)

Integrating HSE, maintenance, and inspection data ensures safe operations, minimal environmental impact, optimized extraction, reduced downtime, improved drilling, and enhanced modeling.

5.4.5 Resource Estimation and Field Development Planning

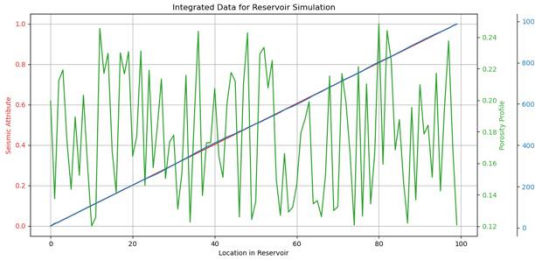
a) Reservoir Characterization

Combining seismic with geological and petrophysical data improves reservoir characterization, crucial for optimizing well placement and predicting production rates accurately.



b) Reservoir Simulation

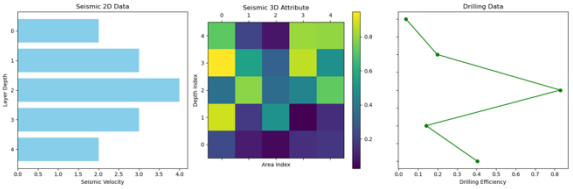
By combining seismic 3D data, reservoir data, and well data, detailed reservoir simulation predicts reservoir behavior over time, guiding field development decisions.



5.4.6 Operation Efficiency and Cost Reduction

a) Drilling Optimization

Combining drilling data with both seismic 2D and 3D data aids in pinpointing efficient drilling paths, cutting time, cost, and environmental impact.

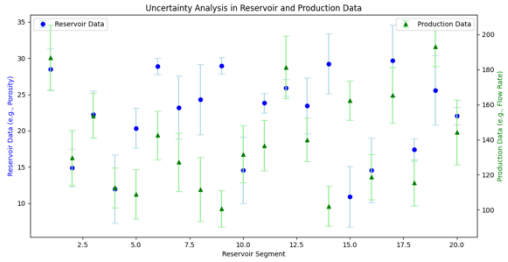


Integrating equipment data with production monitoring enables predictive maintenance planning, preempting failures, minimizing downtime, and cutting operational costs.

5.4.7 Risk Management and Uncertainty Reduction

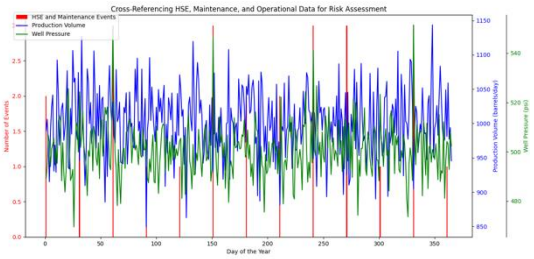
a) Uncertainty Analysis

- Combining diverse datasets aids in conducting uncertainty analysis. For example, examining the variance in reservoir and production data can pinpoint areas requiring additional data for improved certainty.



b) HSE Risk Assessment

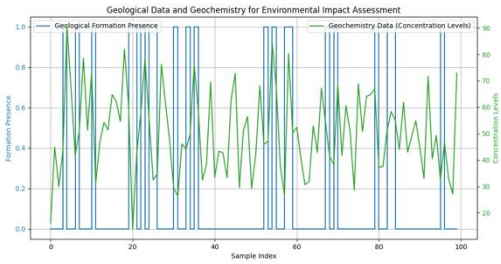
Cross-referencing HSE and maintenance data with operational data such as production operation and well data enables thorough risk assessment, ensuring effective safety protocol implementation.



5.4.8 Sustainability and Environmental Stewardship

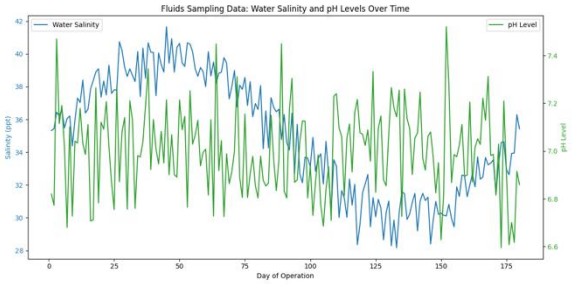
a) Environmental Impact

Leveraging geological data and geochemistry aids in assessing the environmental impact of drilling and production, promoting sustainable practices.



b) Water Management

Examining fluids sampling data supports improved management of water produced during oil and gas extraction, crucial for minimizing environmental impact.



Impact

Adopting the proposed decentralized architecture for unified access across distributed oil data sources provides multifaceted impact on end users and the analytical ecosystem:

a) Agility in Analytics

- Self-service access to curated domain data products accelerates analytics velocity
- Enables iterating over hypotheses faster through readily available consolidated data

b) Discoverability and Accessibility

- Well-documented data products with meaningful interfaces improve discoverability
- Guided user experience for domain data facilitates access to authorized consumers

c) Understandability
- Common vocabulary and ontological alignment enhance interpretability
- Contextual metadata aids in correctly interpreting data elements

d) Innovation Catalyst
- Availability of integrated data products fuels ideation
- Fosters new cross-domain analytical opportunities previously hindered by data silos
- Opens up ability to leverage modern AI/ML algorithms

e) Collaboration Culture
- Virtualizing access to high-quality domain data products incentivizes sharing
- Brings together multi-disciplinary stakeholders across silos

Scope

The paper covers:

a) Rationale and context for decentralized data management tailored to the oil industry based on data mesh principles
b) A high-level logical architecture demonstrating viability using access and governance oriented domain data products with standard interfaces
c) Implementation reference specs using AWS cloud data and analytics services like S3, Redshift, Athena etc.
d) Applicability evaluation through analytics use cases around production optimization, drilling planning and predictive maintenance

With this paper I aims to provide a comprehensive architectural blueprint and implementation guidelines for organizations to adopt decentralized mesh-oriented structures for managing and consolidating oil field data at scale.

Aspects outside the scope:

a) Detailed documentation of all ontologies, taxonomies and metadata standards
b) Physical deployment architectures for full production grade implementation
c) Quantitative performance benchmarks
d) Detailed total cost analysis

While these extensions may be valuable in future studies, the primary focus is to introduce the paradigm, justify its relevance, and offer an adoptable architectural template to technical and business stakeholders. The paper aims to balance conceptual robustness, demonstrable credibility, and a pragmatic focus on adoption, addressing the decentralization of complex information landscapes in the oil and gas industry.

Extended Use Cases

Here are 10 high-level use cases from different industries that could benefit from a data mesh architecture for enabling better data consolidation and analytics:

a) Oil and Gas - Production optimization by correlating well data across fields
b) Mining - Predictive maintenance planning for mining equipment by connecting sensor data with maintenance logs
c) Power and Utilities - Demand forecasting by combining smart meter readings, customer data, and appliance telemetry
d) Pharmaceuticals - Clinical trial analytics by linking patient diagnosis records with treatment effectiveness data
e) Banking - Fraud pattern detection across payment systems by consolidating transaction, network and cybersecurity logs
f) Telecom - Churn prediction by analyzing customer usage data, network performance, and customer ticket logs
g) Transportation - Route optimization in logistics by integrating vehicle location data with traffic, weather and inventory data feeds
h) Government - Social welfare program outcomes analysis by connecting domains like healthcare, housing, and occupation
i) Insurance - Risk modeling for underwriting using historical claim statistics, policy types, demographics
j) Retail - Targeted real-time promotions by combining customer 360 data with product catalogs, inventory and sales transactions

The use cases highlight the common challenges around connecting distributed, siloed data landscapes. A data mesh inspired, decentralized consolidation approach helps unlock unified analytics.

2. Conclusion

The proposed decentralized data lake architecture, grounded in data mesh principles, presents a transformative approach for the oil and gas industry. It effectively tackles the challenges of data silos, promoting enhanced data discoverability, accessibility, and interoperability. By fostering a culture of shared responsibility and collaboration, the architecture encourages innovation and facilitates more efficient decision-making processes. Looking ahead, the scalability and flexibility offered by this model hold the promise of reshaping how data is managed across industries, urging a broader adoption and continuous exploration of the data mesh concept. This evolution towards a more open, collaborative, and efficient data ecosystem is pivotal for driving future advancements and sustaining competitive advantage in an increasingly data-driven world**.**

References

[1] Enterprise Data Strategy: a decentralized data mesh approach. (2022, October 25). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/1004167
[2] Machado, I., Costa, C., & Santos, M. Y. (2022). Data Mesh: Concepts and Principles of a Paradigm Shift in data architectures. Procedia Computer Science, 196, 263–271. https://doi.org/10.1016/j.procs.2021.12.013
[3] Vlasiuk, Y., & Onyshchenko, V. (2023). Data Mesh as distributed data platform for large enterprise companies. In Lecture notes on data engineering and communications technologies (pp. 183–192). https://doi.org/10.1007/978-3-031-36118-0_17
[4] M, A. A., Aseel, A., Roy, R., & Sunil, P. (2023). Predictive big data analytics for drilling downhole

problems: A review. Energy Reports, 9, 5863–5876. https://doi.org/10.1016/j.egyr.2023.05.028
[5] Big data integration architectural concepts for oil and gas industry. (2016, October 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7991832
[6] OPSDS: a semantic data integration and service system based on domain ontology. (2016, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7866141
[7] Towards goal-driven access to process warehouse: Integrating goals with process warehouse for business process analysis. (2011, May 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/6006861
[8] Govern Analytics – Amazon DataZone – AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/datazone/
[9] Big data integration architectural concepts for oil and gas industry. (2016, October 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7991832
[10] Challenges of data integration and interoperability in big data. (2014, October 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7004486
[11] Yuan, J., & Li, H. (2023). Research on the standardization model of data semantics in the knowledge graph construction of Oil&Gas industry. Computer Standards & Interfaces, 84, 103705. https://doi.org/10.1016/j.csi.2022.103705
[12] Digital transformation in oil and gas industry: Developing an OSDU Third-Party Application. (2021, October 27). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9659636
[13] High-level feasibility analysis of an electronic distributed control architecture for oil & gas turbomachinery. (2016, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7555520
[14] Padi, R. K., Douglas, S., & Murphy, F. (2023). Techno-economic potentials of integrating decentralized biomethane production systems into existing natural gas grids. Energy, 283, 128542. https://doi.org/10.1016/j.energy.2023.128542
[15] Tang, X., Feng, Z., Xiao, Y., Wang, M., Ye, T., Zhou, Y., Meng, J., Zhang, B., & Zhang, D. (2023). Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development. Geoscience Frontiers, 14(5), 101426. https://doi.org/10.1016/j.gsf.2022.101426
[16] Survey on Interface Usability Evaluation for Oil and Gas Critical Control Systems. (2021, December 16). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9672449
[17] Sarrakh, R., Renukappa, S., & Suresh, S. (2022). Evaluation of challenges for sustainable transformation of Qatar oil and gas industry: A graph theoretic and matrix approach. Energy Policy, 162, 112766. https://doi.org/10.1016/j.enpol.2021.112766
[18] Al-Rbeawi, S. (2023). A review of modern approaches of digitalization in oil and gas industry. Upstream Oil and Gas Technology, 11, 100098. https://doi.org/10.1016/j.upstre.2023.100098
[19] Aragão, R. R., & El-Diraby, T. E. (2019). Using network analytics to capture knowledge: Three cases in collaborative energy-oriented planning for oil and gas facilities. Journal of Cleaner Production, 209, 1429–1444. https://doi.org/10.1016/j.jclepro.2018.10.346
[20] Desai, J. N., Pandian, S., & Vij, R. K. (2021). Big data analytics in upstream oil and gas industries for sustainable exploration and development: A review. Environmental Technology and Innovation, 21, 101186. https://doi.org/10.1016/j.eti.2020.101186
[21] Enhanced Oil Recovery, second edition. (n.d.). PennWell Books. https://www.pennwellbooks.com/enhanced-oil-recovery-second-edition-green-willhite-book-9781613994948/
[22] Ikelle, L. T. (2010). Introduction to multishooting: challenges and rewards. In Handbook of geophysical exploration (pp. 1–53). https://doi.org/10.1016/s0950-1401(10)03907-8
[23] Ppdm. (2023, August 8). Data Dilemma: Unraveling the challenges and downsides of data in oil and gas. JPT. https://jpt.spe.org/data-dilemma-unraveling-the-challenges-and-downsides-of-data-in-oil-and-gas
[24] Chemistry for enhancing the production of oil and gas. (n.d.-b). PennWell Books. https://www.pennwellbooks.com/chemistry-for-enhancing-the-production-of-oil-and-gas-frenier-ziauddin-book-9781613993170/
[25] Mohammadpoor, M., & Torabi, F. (2020). Big Data analytics in oil and gas industry: An emerging trend. Petroleum, 6(4), 321–328. https://doi.org/10.1016/j.petlm.2018.11.001