

Accelerating Subsurface workflows thru Automated Data and ML Pipelines

Author information – First name, middle initial, last name. (left justified, italics, 10pt font)

Affiliation – SLB, USA

Summary (All headings should be Arial 12pt bold, DELETE SECTIONS THAT ARE NOT USED)

Understanding the subsurface is key to deliver reliable well construction and efficient production operations. Embracing digitalization in subsurface helps to improve accuracy, reduce risks, and accelerate cycle time. Crew change, availability of high-precision data sets and increased pressures on margins imply that human interpreters need to be supplemented by automation and machine learning (ML) driven insights through digital techniques.

ML solutions can accelerate interpretation to modeling to reservoir engineering workflows by optimizing first break picking in seismic processing, improving fault detection, stratigraphic interpretation in seismic and reconstructing logs and detect outliers in wellbore. Scalable machine learning requires reliable data products, but operators are fraught with data wrangling challenges across sources without lineage or context. Domain users cannot collaborate well with data scientists further impeding ML models from moving from innovation to production.

New wellbore and seismic data can be aggregated across vendors, data stores and contextualized to reliable data products by automated DataOps pipelines. Domain experts can understand these data products and collaboratively work with data scientists on an intuitive ML workbench democratizing the ML craft and providing first principal guard-rails. An MLOps pipeline can manage model versions and continuously deliver qualified ML models into elastic compute-clusters for reliable result prediction on new datasets.

Such a digital system can account for ML models drift recalibration and regional localization ensuring the solution remains operational and reliable over time. Reliable data products through DataOps pipelines feeding contextualized information to ML models deployed and operationalized using MLOps in the cloud, result in efficient and intelligent solutions that optimize subsurface processing, interpretation, and modeling workflows.

Theory / Method / Workflow

Moving from idea and concept to a proof-of-value to a production deployment that delivers value at scale requires structured approach to digital delivery. Many projects die at innovation stage due to the lack of a strong software development, delivery, and operationalization approach.

Subsurface influences key well construction and production operations. Comprehensive understanding of subsurface geological and reservoir characteristics is critical for successful well construction and production operations. These operational decisions are hampered by lack of connectivity and reliable subsurface data delivery.

Smart data-driven algorithms can greatly improve Subsurface accuracy, risk reduction, and cycle times. Digitalizing seismic processing, interpretation, and reservoir engineering leads to improved accuracy, risk reduction, and faster cycle times from seismic data to simulation.

Data-driven insights require ML algorithms which in turn need reliable data products. Many data scientists have trouble with data wrangling and use curated Excel/CSV exports or staged datasets for ML prototypes. These are good for demos, but not for making ML models operational.

Data-driven organizations use inclusive and agile data governance with distributed, federated data management – treating Data as a "product". DataOps is a method that combines the practices of agile software engineering, DevOps, and data engineering to improve the quality, speed, and collaboration around data products.

Automated and event-driven automation ensures data product flows are continuous and not impeded by manual interventions. Technology foundations key for this automation includes both infrastructure versioning and automation (infrastructure as code) and data pipeline versioning and automation.

Data scientists and ML engineers need easy data access, intuitive data analysis, and ability for faster experimentation to create ML models that produce data-driven insights. DataOps techniques offer continuous integration and deployments with federated governance that deliver high-quality and accessible data products. Data scientists consume these data products initially through visual analysis and identify the data features and explore possible algorithm choices. They then build machine learning models with the best-fit algorithm(s), necessary features, and optimal hyper-parameters. As an iterative process, they constantly adjust the hyperparameters and validate the model by comparing the predictions with test datasets.

ML models must be treated like any other code and should therefore be 'source controlled' along with metrics/results for each model alternative. Reliable ML driven insights in production setting requires a governed and automated rollout strategy, a Build/Delivery pipeline like DevOps. MLOps is a methodology that combines the practices of machine learning (ML), DevOps, and data engineering to help organizations deploy and maintain ML models in production. MLOps increases automation and improves quality of production ML models while federated governance helps to meet business and regulatory requirements. It applies to the entire lifecycle of ML models, from data preparation to model training and evaluation to deployment and monitoring.

The deployed model should be versioned and managed so that it can be easily rolled back in case of errors. Discerning model performance issues vs data quality/drift issues is a challenge and robust DataOps procedures including automated data transformations, data curation, and data quarantine are necessary to supplement MLOps and handle partner or local datasets.

MLOps and DataOps are complementary methodologies that can be used together to improve the efficiency and effectiveness of machine learning and data analytics. By combining the best practices of both methodologies, organizations can build and deploy ML models that are more accurate, reliable, and secure.

Results, Observations, Conclusions

The paper presents an integrated digital solution that includes:

- a cloud-based cognitive environment which abstracts the infrastructure, software and services deployment, security, operational monitoring.
- a DataOps pipeline built on open standards (OSDU® Standard, described later) allows data to be ingested from several sources seamlessly, aggregating data from several vendor, corporate/project stores, unstructured content like well reports, etc. and delivers the foundation for data products.
 - These raw data entities are then contextualized and curated using rules and ML techniques to derive trusted data products. These trusted data products can be easily discovered and consumed in both conventional physics-driven and newer ML-driven workflows.
- a simple and intuitive ML workbench to democratize machine learning and bring both domain experts and data scientists together to collaborate on building and publishing ML models on the platform.
 - MLOps techniques allow model versioning and continuous delivery to production use. Elastic compute clusters run validated models to score new datasets on-demand and generate business insights.
- continuous monitoring of infrastructure, model performance, model drifts, data drifts ensuring reliability and system health in production.
 - Account for contingencies such as model recalibration and localization for different asset types/geographies, manually entered data from the field or partner datasets not connected online.
- security of the operational system with all access logged and auditable by CloudOps procedures (not covered in this paper).

The following subsurface workflows are a small sliver of those that have benefited from leveraging the MLOps and DataOps pipelines in our Digital solution accelerating the processes and bringing more efficiency and uncertainty reduction to these workflows:

- Optimized seismic processing through ML-driven first break picking.
- Improved fault detection and stratigraphic interpretation
- Reconstruction of well logs and detection of outliers

The paper will briefly explain the challenges and the results observed in operationalizing ML-driven solutions for these subsurface workflows.

Acknowledgements

- Vikas Jain, SLB for helping with a significant portion of the content of this paper.
- SLB AI team for the references and projects that illustrate the benefits of the approach.

References

- Definitive guide to industrial DataOps, Cognite, 2022, <https://www.cognite.com/en/resources/definitive-guide-to-industrial-dataops>
- Data Mesh: Delivering Data-Driven Value at Scale 1st Edition, (Zhamak Dehghani, ThoughtWorks), 2019 - <https://www.thoughtworks.com/en-us/insights/books/data-mesh>
- What is MLOps?, Lynn Heidmann, ScalingAI, Dataiku, 2022 - <https://blog.dataiku.com/what-is-mlops-why-does-it-matter>

- The Open Group Launches the Open Subsurface Data Universe™ Forum, 2019, <https://www.opengroup.org/open-group-launches-open-subsurface-data-universe-forum>
- Deep Learning Ensemble for Seismic First-Break Event Picking, 83rd EAGE Annual Conference & Exhibition, (Zhao, T., P. Bilsby, S. Manikani, G. Busanello, M. Benzaoui, and A. Abubakar), 2022
- Interpreting seismic faults with machine learning techniques, U.S. Patent 62/853,681, (Li, C., and A. Abubakar), 2019
- Unsupervised well log reconstruction and outlier detection (X. Chen, H. Maniar, and A. Abubakar), US Provisional Patent Application, 62/897,088, Sep 6, 2019)