

# PROJECT SEMESTER REPORT

## **“Enhancing Drilling Operations with Data Analytics and Machine Learning in the Oil and Gas Industry”**

by

**Vasudhara Mahajan**

**Roll No. : 102103032**

Under the Guidance of

**Industrial Mentor - Dr. Rajarshi Roy Chowdhury,  
Manager, Advanced Analytics, IRODA Helmerich &  
Payne, India.**

**Faculty Mentor- Dr. Rahul Nijhawan, Assistant Professor - I**



Submitted to the

**Computer Science & Engineering  
Department Thapar Institute of Engineering &  
Technology, Patiala**

In Partial Fulfilment of the Requirements for the Degree  
of Bachelor of Engineering in Computer Engineering

at

Thapar Institute of Engineering & Technology, Patiala

**June 2025**

# **Title of the Project**

**“Enhancing Drilling Operations with Data Analytics and Machine Learning in the Oil and Gas Industry”**

by Vasudhara Mahajan

Place of work: Helmerich and Payne

Submitted to the Computer Science & Engineering Department, Thapar Institute of Engineering & Technology

June 2025

In Partial Fulfilment of the Requirements for the Degree of Bachelor of Engineering in Computer Engineering

## Abstract:

This project focuses on enhancing data-driven decision-making in the oil and gas drilling industry through the analysis of operational data and the development of automation-friendly tools. Conducted in collaboration with Helmerich & Payne (H&P), the project addresses key challenges in identifying slide criticality, understanding regional setpoint usage, and creating resilient data logging systems.

The work began with exploratory data analysis (EDA) on well logs and survey data to understand drilling behaviors during slide and rotate operations. Clustering techniques, particularly Spectral Clustering, were used to classify operations into critical and non-critical intervals based on anomalies in drilling parameters. Simultaneously, distribution analysis of drilling setpoints was performed across counties, operators, and basins to identify operational trends and optimization opportunities.

To support real-time monitoring and transparency, a Docker-based in-message logging application was developed. The system was built to capture key events and retain messages even during container failures, network disruptions, and manual interrupts. Custom Python classes and API utilities were also created to fetch and process survey data efficiently.

The combined outcomes of this project contribute toward safer, more efficient, and more automated drilling practices. It demonstrates the potential for machine learning and robust engineering to support operational decisions in real-time environments.

Author

Vasudhara Mahajan  
(Student's name)

Certified by



(Industrial Mentor: Rajarshi Roy Chowdhury )

Certified by



( Faculty Mentor: Rahul Nijhawan )

**CERTIFICATE (PROJECT SEMESTER TRAINING) FROM THE COMPANY OR THE ORGANIZATION**

Date: 28/05/2025

Hi Vasudhara Mahajan,

I hope you're doing well.

This letter is intended to formally acknowledge the successful completion of your internship at the Advanced Analytics team of Helmerich & Payne, India office. As your industrial mentor during the period, I can certify that you have made meaningful contributions across several key areas, and I want to commend you on the quality and consistency of your work.

Your efforts have contributed positively to our understanding of Flex-app setpoints trends, well plan classification strategies and building containerized applications with exception handling. Your initiative through the various project has been impressive. Thank you for your dedication, professionalism, and curiosity.

Please feel free to reach out if you need a reference or support in the future. I wish you all the best in your continued studies and future endeavours.



Best regards,  
Rajarshi Roy Chowdhury  
Manager, Advanced Analytics, IRODA  
Helmerich & Payne, India.  
Mob. +91-8409827271  
Email. Rajarshi.chowdhury@hpinc.com

# TABLE OF CONTENT

1.	Company Profile	1-2
	1.1 Overview	1
	1.2 Core services and work of the company	2
	1.3 Products delivered	2
2.	Introduction	3-4
	2.1 Project Overview	3
	2.2 Technology Stack used	3-4
	2.3 Scope and Utility	4
3.	Background	5-6
	3.1 Industry Context and need	5
	3.2 Motivation behind the project	5
	3.3 Prior research and references	6
	3.4 Proposed layout and workflow	6
4.	Objectives	7
5.	Methodology	8-14
	5.1 Exploratory data analysis and clustering evaluation	8-9
	5.2 Setpoint Distribution	9-10
	5.3 Slide criticality project	11-12
	5.4 Development of Resilient Message Logging Application	13-14
6.	Observations and findings	15-18
	6.1 Setpoint distribution and analysis	15
	6.2 EDA and clustering results	16-17
	6.3 Slide Criticality	17
	6.4 Docker based logging application	18
7.	Limitation	19-22
	7.1 Clustering and analysis limitations	19
	7.2 Slide Criticality limitations	20
	7.3 Setpoint and Distribution limitations	20-21
	7.4 Docker based logging application limitations	21

8.	Conclusion	22-23
	8.1 Conclusion	22
	8.2 Future Work	22-23
9.	Bibliography	24
10.	Peer Review	25-26
11.	Certificate of the course	27

## **LIST OF FIGURES**

Fig 5.1	Analysis of well sections	9
Fig 5.2	Analysis of setpoints	10
Fig 5.3	Slide Criticality project	12
Fig 5.4	Dockerized applications	14
Fig 6.1	Violin plots for distribution of setpoint values	15
Fig 6.2	Box plots	16
Fig 6.3	Pairplots	16
Fig 6.4	Trajectories	17
Fig 6.5	Docker Desktop screenshots	18

# 1. Company Profile



## 1.1 Overview

Helmerich & Payne, Inc. (H&P) is a globally renowned petroleum contract drilling company that has been at the forefront of oil and gas drilling innovation for over a century. Founded in 1920 and headquartered in **Tulsa, Oklahoma**, the company has grown from a regional drilling contractor into one of the most technically advanced and respected names in the energy sector.

H&P specializes in both land and offshore drilling services, with a particular emphasis on **automated, data-enhanced onshore drilling operations**. The company is known for its **FlexRig® technology**, a fleet of purpose-built rigs designed for **unconventional horizontal drilling**. Since their introduction in 1998, FlexRigs have transformed how shale reservoirs are developed, providing unmatched mobility, safety, and speed. These rigs feature programmable logic controllers, sensor integration, and digital communication systems that allow real-time control and optimization of the drilling process.

As of recent years, H&P has maintained over **20% of the U.S. land drilling market share** and more than **40% of the super-spec land rig share**, affirming its dominant position in North America's most productive basins, including the **Permian Basin, Bakken, and Eagle Ford**.

Beyond North America, the company has expanded its operational footprint globally, serving clients in **Argentina, Colombia, Bahrain, Australia, the United Arab Emirates, and Saudi Arabia**. These international ventures demonstrate H&P's adaptability across various geological, regulatory, and operational environments.

The company's vision is deeply aligned with **digital transformation in energy**, focusing on real-time data acquisition, machine learning models for predictive maintenance, and the development of autonomous drilling capabilities—all essential for reducing human error, increasing efficiency.



## 1.2 Core services and work of the company

Helmerich & Payne (H&P) is a leading provider of drilling services and technology solutions for the oil and gas industry. The company specializes in operating both conventional and unconventional drilling rigs, offering a wide range of services that include directional drilling, wellbore management, and survey optimization. H&P is also at the forefront of automation and innovation, developing advanced technologies to enhance drilling efficiency, precision, and safety. Their operations span globally, with active drilling projects in countries such as the United States, Argentina, Australia, Bahrain, Colombia, the United Arab Emirates, and Saudi Arabia. Through its integrated drilling and technology services, H&P delivers end-to-end solutions for production needs.

## 1.3 Product delivered

As part of the internship, several projects that aligned with H&P's cutting-edge products and services are:

- **AutoSlide® Technology:** The intern worked on slide criticality analysis, directly supporting AutoSlide®'s goal to automate directional drilling by identifying critical intervals where intervention is needed. This involved analyzing drilling modes and clustering slide vs. rotate patterns.
- **Survey Management & Setpoint Distribution Tools:** The student contributed to analyzing the distribution of key drilling parameters (e.g., RPM, WOB, flow rate) across various counties, operators, and intervals. These insights aid in regional performance evaluation and optimization.
- **Digital Roadmap® Logging System:** A fault-tolerant, containerized in-message logging system was developed by the student using Docker. It successfully recorded runtime events and managed interruptions (e.g., docker kill, network failures), supporting real-time data analysis and operational transparency.

## 2. Introduction

### 2.1 Project Overview

This project is designed to enhance the safety, efficiency, and intelligence of oil and gas drilling operations by transforming raw drilling data into actionable insights. Drilling is a high-risk, high-cost activity where real-time decisions significantly impact operational success. As drilling environments become more complex—with deeper wells, more demanding formations, and tighter regulations—there is an urgent need for intelligent systems that assist in monitoring, diagnosing, and guiding drilling activity.

The core objective of this project is to analyze historical and near real-time operational data—including rig logs, survey data, slide sheets, and equipment settings—to uncover critical patterns and anomalies. These insights empower drilling engineers to detect deviations early, predict failures, and recommend timely corrective actions. By classifying operational behaviors into typical vs. atypical patterns, the system improves the decision-making process during critical operations such as sliding or directional changes.

The analysis particularly focuses on slide criticality—a phase where well trajectory adjustments are made. Accurate detection of abnormal behavior in these intervals can significantly reduce rework, non-productive time (NPT), and tool damage. Through intelligent clustering, deviation analysis, and logging, this project supports a data-driven decision support system that brings greater clarity and safety to well operations.

### 2.2 Technology Stack Used

- **Languages & Libraries:**
  - Python (Pandas, NumPy, Matplotlib, Seaborn, Asyncio)
  - Scikit-learn, Scikit-optimize, Plotly, JSON
- **Machine Learning & Analysis:**
  - PCA for dimensionality reduction
  - Clustering algorithms (KMeans, GMM, Spectral Clustering)
  - GridSearchCV for hyperparameter optimization

- **Infrastructure:**
  - Docker (containerization and orchestration)
  - Docker Compose for multi-service deployment
  - Environment variables for secure configuration
- **Data Sources & Tools:**
  - Databricks SQL for API-based data ingestion
  - Directional survey data, slide sheets, and operational logs
- **Deployment & Logging:**
  - Real-time logging with Docker volume persistence
  - Signal handling for graceful shutdown and log retention
  - Recovery from network and container-level interruptions

## 2.3 Scope and Utility

- The system detects abnormal slide behavior during drilling, enabling early warnings and proactive adjustments to improve accuracy and efficiency.
- Analyzes setpoint usage across counties and operators, revealing patterns that can guide the development of standard practices and highlight operational inefficiencies.
- Docker-based architecture ensures continuous and resilient message logging, even in the face of unexpected shutdowns or failures.
- Converts complex telemetry and drilling data into digestible visual insights and logs to support real-time decision-making for engineers.
- Designed to allow for future improvements such as UI dashboards, predictive analytics, and cloud-based deployment for enterprise scalability.

## **3. Background**

### **3.1 Industry Context and Need**

Modern oil and gas drilling operations are data-intensive, involving complex workflows and advanced machinery deployed in diverse and often remote environments. These operations generate terabytes of data through downhole sensors, surface equipment, and control systems. Despite this wealth of information, decision-making during drilling often depends on manual interpretation and heuristics derived from experience. This reliance introduces inconsistencies, delays in identifying operational issues, and increases the risk of unplanned non-productive time (NPT).

As the industry shifts toward digital transformation, there is a clear need for systems that can analyze data at scale, surface hidden insights, and assist engineers in real-time. The integration of data science, machine learning, and automation in drilling operations presents a transformative opportunity—improving efficiency, reducing costs, and enhancing wellbore placement accuracy.

### **3.2 Motivation Behind the Project**

The core motivation for this project is to bridge the gap between the massive volume of raw drilling data and actionable operational insights. While vast data sources such as rig logs, slide sheets, survey reports, and setpoint configurations exist, they are often underutilized. These files are frequently stored without adequate post-analysis, or are manually reviewed long after drilling decisions are made.

By leveraging data-driven methodologies such as clustering, deviation detection, and real-time logging, this project aims to shift the paradigm from reactive to proactive drilling operations. Identifying early signs of abnormal slide behavior or inefficiencies in setpoint usage can reduce tool wear, enhance trajectory adherence, and improve the safety of directional drilling.

### 3.3 Prior Research and References

Prior academic and industrial research has demonstrated the effectiveness of data analytics in high-frequency industrial operations:

- [1] Time-series analysis and machine learning for real-time condition monitoring in mechanical systems have shown promise in reducing downtime and optimizing performance.
- [2] Data-driven decision support systems in well planning have enabled better parameter tuning and have highlighted the importance of continuous feedback loops in high-risk operations.
- Building on these foundations, this project integrates proven techniques—such as PCA for dimensionality reduction and Spectral Clustering for classification—into the drilling domain to classify critical intervals and provide explainable insights to field engineers.

### 3.4 Proposed Layout and Workflow

To achieve the objectives of data-driven insight generation and operational support, the following phased approach is adopted:

- **Data Collection and Preprocessing** :Collect well trajectory data, operational logs, setpoint values, and slide sheets from sources like Databricks SQL. Standardize formats, filter relevant columns, and clean anomalies to ensure quality inputs for analysis.
- **Exploratory Data Analysis(EDA)**:Visualize operational trends over time, plot trajectory paths, and identify typical vs. atypical behaviors during slide and rotate phases. Compare key parameters (ROP, WOB, RPM, Flow Rate) across various wells, operators, and basins.
- **Behavior Classification**Use unsupervised machine learning algorithms (e.g., Spectral Clustering, KMeans) combined with PCA to classify operational states. Separate critical from non-critical intervals based on deviation from planned well paths and parameter thresholds.

## 4. Objectives

The main objectives of this project are:

- To perform exploratory data analysis (EDA) on well sections and evaluate various clustering algorithms to identify the most effective method for detecting patterns and anomalies in drilling operations.
- To analyze the distribution of well data to identify distinct counts of wells based on various setpoints used in drilling processes across different counties, operators, and basins, providing insights into regional and operational variations.
- To assess and compare planned versus actual wellbore paths by processing and analyzing directional survey data. This comparison helps in identifying deviations, interpolating missing data points, and evaluating critical slides, all aimed at improving drilling accuracy and efficiency in slide Criticality project.
- To design and develop a fault-tolerant, Dockerized Python application that performs real-time data extraction from Databricks and logs operational messages with resilience. The system connects to a Databricks SQL endpoint at regular intervals, and stores results while capturing runtime logs. It is engineered for portability, automated execution, error resilience, and data persistence, even under failure conditions such as container kills, network outages, or manual interrupts.

## 5. Methodology

To achieve the defined objectives, the following structured methodology will be followed:

### 5.1 Exploratory Data Analysis and Clustering Evaluation

- Extract data from drilling logs and well reports focused on specific well sections.
- Clean and preprocess the data to handle missing values and standardize formats.
- Conduct Exploratory Data Analysis (EDA) to visualize trends, operational behaviors, and anomalies.
- Use PCA for reducing the dimensionality of the dataset for visualization purposes.
- Use GridSearchCV from scikit to find the best hyperparameters for the clustering algorithms.
- Apply and compare clustering algorithms (e.g., KMeans, GMM, Spectral Clustering) using performance metrics like silhouette score .
- Select the most suitable clustering method based on effectiveness and interpretability.

Here below is the block diagram describing the procedure about what was done in this section.

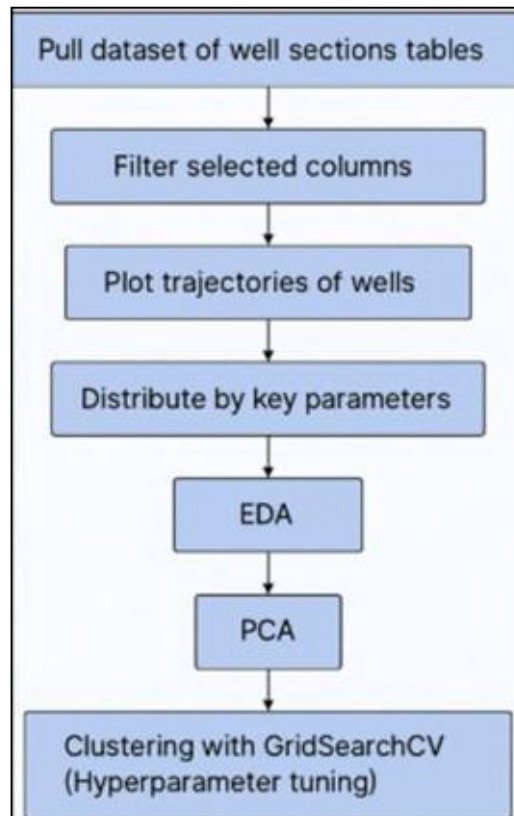


Fig 5.1 Analysis of well sections

## 5.2 Setpoint Distribution

- Load the well data into a Spark DataFrame.
- Ensured the data was clean and ready for analysis by filtering and selecting relevant columns.
- Used PySpark's filter and col functions to achieve this.
- For each setpoint column, calculated the distinct count of wells where the column value is 1.
- Used the select and distinct methods to get the distinct counts.
- Calculated the ratio of each distinct count to the total number of distinct wells
- Converted these ratios to percentages for better interpretability.



- Created a bar chart using Plotly to visualize the distinct counts and their corresponding ratios.
- Joined the tables to combine relevant information.
- Display the plot in the notebook for analysis and interpretation.
- Created the violin plots to check the density of the datapoints and figure the value of setpoints used the most

Here below is the block diagram describing the procedure about what was done in this section.

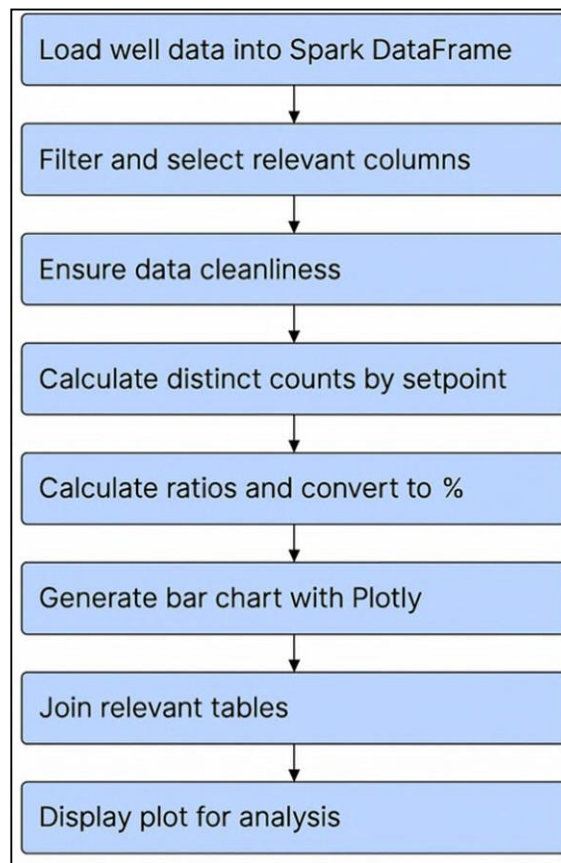


Fig 5.2 Analysis of setpoints

### 5.3 Slide Criticality Project

The project is structured across multiple Python modules, with clear separation of concerns:

#### **API Integration (utils.py)**

- APIUtils Class:

Handles PI-related helper functions including:

- Authentication
- URL generation
- File path construction
- Request handling (sync/async)

- APIClient Class:

- Provides high-level API interaction methods:

- Fetches account details
- Performs GET requests
- Handles asynchronous requests
- Saves responses for future processing

#### **Interpolation Logic (interpolation.py)**

- MinCurvature Class:

- Performs wellbore trajectory interpolation using the minimum Curvature Method.
- Functions include:
  - Data validation and initialization
  - Angular conversion (Inclination ↔ Theta, Theta & Phi ↔ INC & AZI)
  - Interpolating survey data to finer depth steps
  - Applying the min curvature algorithm for more accurate path representation

#### **Survey & Slide Sheet Processing (slide\_criticality.py)**

- Custom functions handle:
  - Cleaning and filtering survey data
  - Retrieving survey data at specific measured depths (MD)
  - Computing deviation distance between planned and actual surveys
  - Merging this data with slide sheets for final analysis
  - Saving criticality results in CSV format

Here below is the block diagram describing the procedure about what was done in this section.

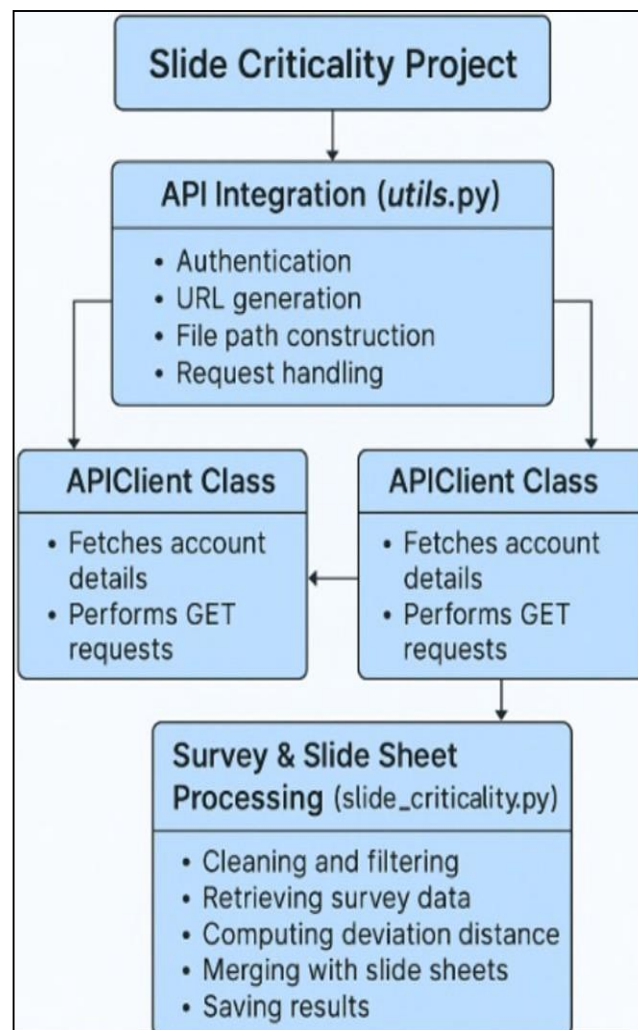


Fig 5.3 Slide criticality project

## 5.4 Development of Resilient Message Logging Application

- Developed a primary Python script to connect securely to Databricks via SQL connector using access tokens.
- Extract key columns (RigNumber, UTCTime, BottomDepth) and filter rows (BottomDepth > 5000).
- Save filtered data and log messages to CSV and log files respectively.
- Implemented a message logging mechanism that continuously logs system and data-related messages in real-time.
- Preserves log state across container restarts or system shutdowns.
- Captures critical runtime events such as errors, disconnections, and shutdown signals.
- Created a periodic data fetch loop using `time.sleep()` every 5 minutes for a 30-minute cycle.
- Integrated signal handling (SIGINT, SIGTERM) to gracefully shut down and retain state.
- Simulated multiple failure scenarios docker kill, unexpected container exits, Internet/network outages, Invalid credential use, Manual termination (Ctrl+C)
- Verified continued operation and data preservation post-recovery.
- Built a Docker image using a custom Dockerfile with Python dependencies.
- Built docker compose-yml .
- Written environment variables as well to make the connection between databricks and VS code.

Here below is the block diagram describing the procedure about what was done in this section.

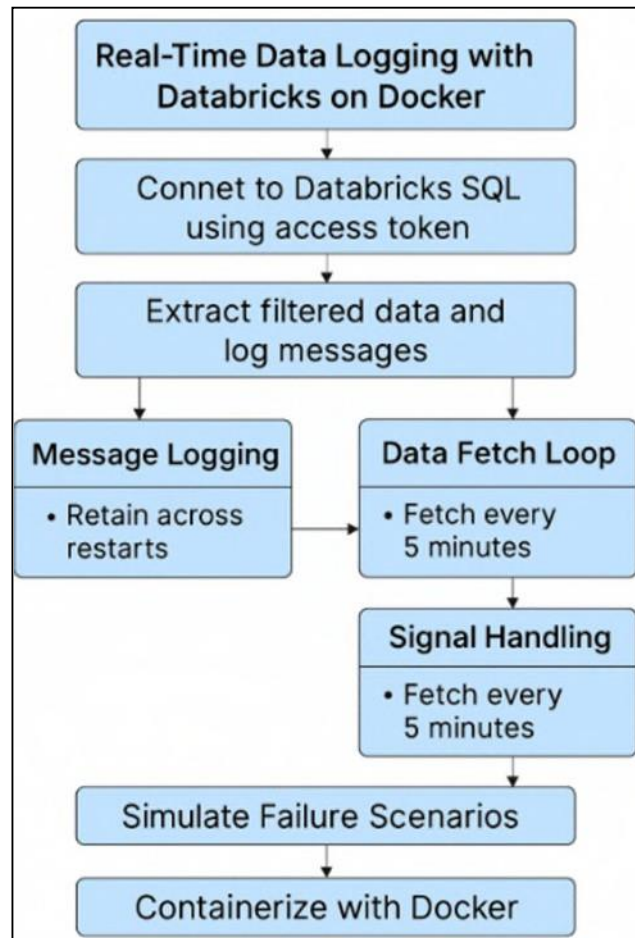


Fig 5.4 Dockerized applications

## 6. Observations and Findings

This section outlines the key discoveries and observations identified throughout the duration of the project.

### 6.1 Setpoint Distribution Analysis

- We identified the count of wells categorized by operators, rigs, and different drilling intervals.
- An in-depth analysis was conducted to study the distribution of setpoint values (e.g., RPM, WOB, and flow rate) across various basins, counties, operators, and well intervals.
- Clear patterns emerged showing that certain operators and regions adhere to more standardized practices, while others display a wide range of operational configurations.

Plotted the violin graphs , histograms and pie charts that gives the idea of the distribution of the setpoints over regions, county and operators for Rig Ids.

Here given below are the snapshots of those.

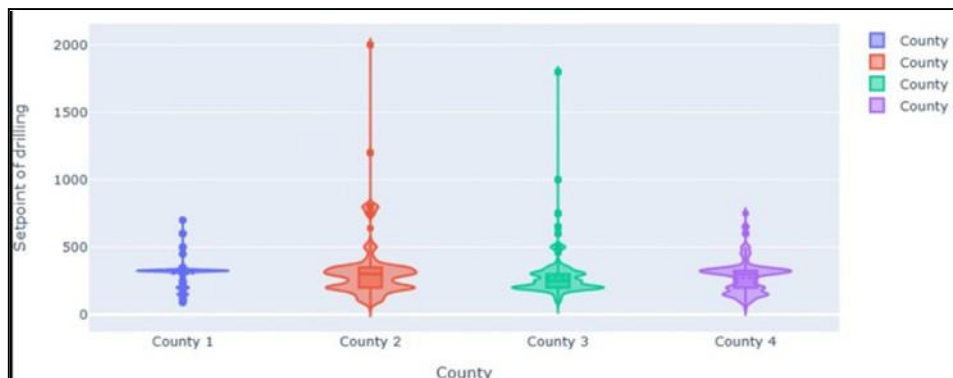


Fig 6.1 Violin plots for distribution of setpoint values

## 6.2 EDA and Clustering Results

- Exploratory Data Analysis (EDA) revealed that drilling operations exhibit distinct behavioral trends depending on operational modes and setpoint configurations.
- Among the clustering algorithms tested, Spectral Clustering offered the most effective separation of operational states, especially in distinguishing critical from non-critical intervals.

For visualization of the best clustering algorithms the pairplots were plotted and the box plots were created and also the trajectories of well sections, here is the snapshot of those.

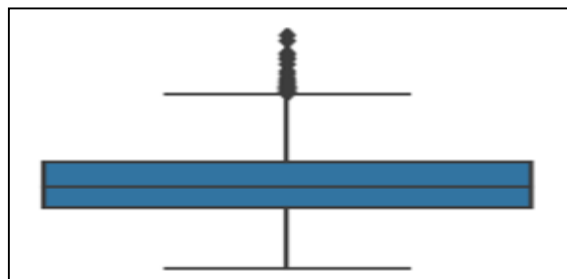


Fig 6.2 Box plots

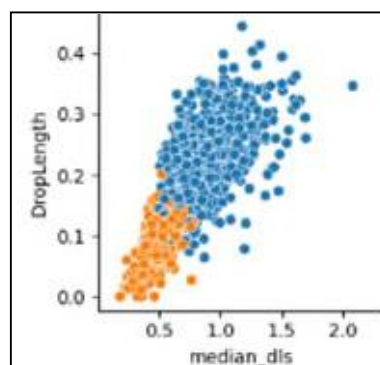


Fig 6.3 Pairplots

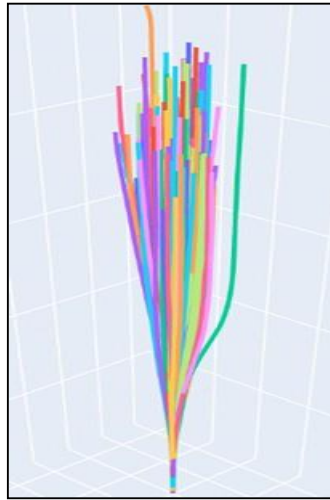


Fig 6.4 Trajectories

### 6.3 Slide Criticality

- Interpolated data provides a more continuous and accurate representation of the well path.
- Deviation distances between planned and actual trajectories are critical in understanding drilling inefficiencies.
- The framework supports both planned and actual data comparisons at specific depths, improving traceability.
- Async API request handling improves performance for large-scale well datasets.
- Min curvature-based interpolation significantly enhances the fidelity of subsurface path estimations.



## 6.4 Docker-Based Logging Application

- The logging system reliably extracted and stored messages at 5-minute intervals even in adverse scenarios such as container termination, network outages, and manual interruptions.
- Docker's volume management and restart policies ensured that logs and CSV outputs were preserved and restored after system restarts.
- Real-time message logging efficiently recorded runtime events, aiding in diagnostics and transparency.
- The application's modular design allows for quick customization, proving its adaptability for real-time data logging in drilling environments.

The docker desktop had containers running along with the images built, the wsl terminal was used to run docker commands in the vs code for running the python applications. Here is one of the screenshot of the docker container created.

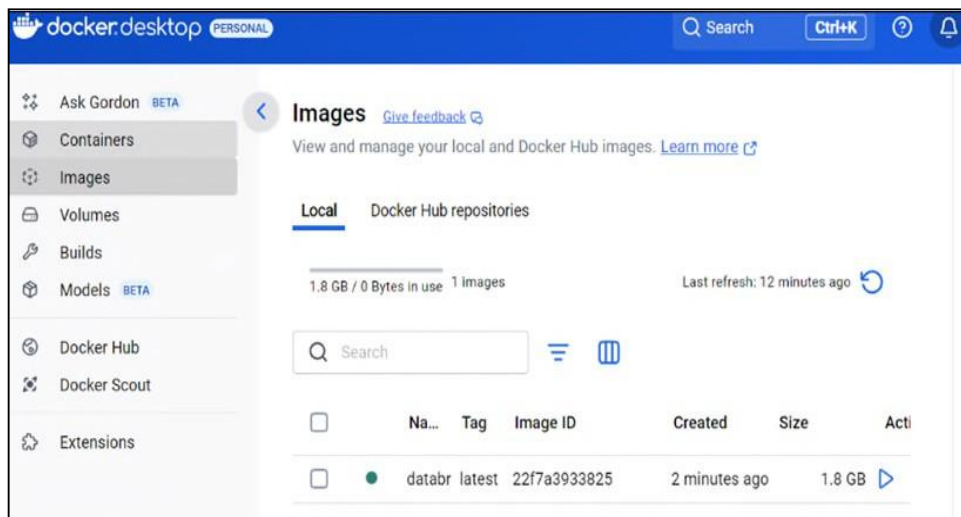


Fig 6.5 Docker Desktop screenshot

## 7. Limitations

This section outlines the primary limitations encountered during the project. While these do not undermine the project's validity, they identify areas that could benefit from further development.

### 7.1 Clustering and Analysis Limitations

- The clustering pipeline is built using third-party libraries like scikit-learn, scikit-optimize, and internal helper modules. As these libraries evolve, certain functions or dependencies may be deprecated, requiring codebase maintenance to ensure long-term stability.
- The current implementation uses pandas and numpy for data transformation and clustering, which is efficient for moderate datasets but not suitable for large-scale data. Processing millions of well sections or high-frequency drilling telemetry would require a shift to distributed frameworks such as Apache Spark or Dask.
- Hyperparameter optimization using GridSearchCV is computationally intensive, especially for algorithms like KMeans, GMM, and Spectral Clustering. As dataset size increases, so does the training time, which could become a bottleneck without GPU acceleration or parallel processing.
- PCA (Principal Component Analysis) was employed for dimensionality reduction to improve clustering efficiency and visualization. However, PCA is limited to capturing linear relationships and may not effectively capture complex, non-linear trends inherent in high-dimensional drilling datasets.
- The effectiveness of clustering was measured using silhouette and Davies-Bouldin scores. While these provide a basic evaluation, they may not fully capture the operational relevance of the clusters without domain-specific interpretation or expert validation.

## 7.2 Slide Criticality Limitations

- The accuracy of the slide criticality analysis heavily depends on the completeness and resolution of the directional survey data. Sparse or inconsistent data can significantly affect the quality of interpolation and deviation measurements.
- The Minimum Curvature Method, while widely used, assumes smooth curvature changes and may not capture abrupt directional shifts or real-world mechanical irregularities in well paths. This can lead to under- or over-estimation of deviation in critical intervals.
- The analysis was performed in a post-processing format. Without integration into a real-time drilling system, the flagged critical slides are not immediately actionable during live operations.
- The thresholds used to define a “critical” deviation were manually chosen based on visual inspection and not dynamically learned or tuned using historical outcomes or expert feedback. This reduces the adaptability of the solution across different well types or basins.
- The slide criticality analysis does not account for surrounding formation data or geospatial constraints, which could be relevant in determining the operational impact of deviations.
- Currently, the model evaluates one well at a time. There is no comparative framework to benchmark slide performance across similar wells or detect systemic issues across a basin or operator level.
- Some stages of the workflow (e.g., slide sheet matching, CSV generation) require manual intervention or script execution, limiting its deployment in a fully automated decision-support environment.

## 7.3 Setpoint and Distribution limitations

- The analysis was conducted within Databricks notebooks, which, while convenient for exploratory work, may face performance bottlenecks during high- volume computation or visualization rendering.

- Dependency and environment control in notebook-based environments can be fragile. Lack of continuous integration/continuous deployment (CI/CD) support limits the reproducibility and version tracking of the analysis pipeline.
- Visualizations were developed using libraries like Plotly within the notebook interface. These are difficult to scale or automate for reporting across large operational datasets without integrating dashboarding tools or external servers.
- Limited debugging tools in notebook environments compared to full-featured IDEs (e.g., PyCharm, VS Code) made troubleshooting certain API or data transformation bugs more time-consuming.
- As the analysis grows in complexity or data volume, transitioning to a dedicated data engineering platform may be required to avoid computation timeouts or memory allocation issues.

## **7.4 Docker-Based Logging System**

- The Dockerized logging application is currently time-bound, relying on `time.sleep()` to trigger data fetch intervals. This static schedule limits flexibility, particularly in adapting to dynamic runtime conditions or custom scheduling needs.
- The system lacks real-time alerting, meaning failures (e.g., lost connections, API errors, memory issues) are only evident via log files. There is no active monitoring or notification system to inform users of runtime anomalies, which could delay response times.
- Retry and recovery mechanisms for transient failures such as API timeouts or network blips are minimal. Enhancing error handling with exponential backoff strategies or retry loops could improve resilience.
- The architecture is designed for single-container deployment and does not currently support multi-threaded or distributed data ingestion. Scaling to multiple data sources or clusters would require significant reengineering, including message queuing and load balancing.
- Environment variable-based authentication (e.g., access tokens for Databricks) must be managed carefully to avoid security breaches. Without proper secret management tools (like AWS Secrets Manager or HashiCorp Vault), sensitive credentials may be exposed during version control or deployment.

## 8. Conclusions and Future Work

This project explored the use of data-driven techniques to improve operational awareness and decision-making in oil and gas drilling. Through clustering, setpoint analysis, slide criticality detection, and real-time logging, the work demonstrated how complex drilling data can be transformed into meaningful insights.

### 8.1 Conclusions:

This project demonstrated the value of combining data science techniques with robust software design to improve operational intelligence in drilling workflows. The multi-faceted approach—ranging from exploratory data analysis and clustering to deviation tracking and fault-tolerant system logging—provided a comprehensive toolkit for analyzing and enhancing drilling efficiency.

- **Exploratory Data Analysis (EDA)** uncovered clear behavioral trends in well operations based on drilling mode (slide vs rotate) and setpoint configurations like RPM, WOB, and flow rate. These visual and statistical insights enabled a better understanding of operational performance under varying geological and technical conditions.
- **Clustering algorithms**, particularly Spectral Clustering, were effective in segmenting well data into critical and non-critical intervals. This allowed for deeper interpretability of drilling patterns and better identification of potentially hazardous or sub-optimal operations.
- **Setpoint distribution analysis** across counties, basins, and operators revealed notable variations in operational strategies, equipment usage, and efficiency. These insights can help standardize best practices and identify areas needing corrective action.
- **Slide Criticality Detection**, through the comparison of planned and actual well paths using interpolated survey data, provided an automated mechanism to detect significant deviations. These deviations are vital for course correction and accurate well placement.
- **The Dockerized real-time logging system** proved resilient to various system failures—including container shutdowns, network drops, and user interrupts—

preserving critical runtime data and enhancing operational transparency.

## 8.2 Future Work:

While the project successfully achieved its core objectives, several areas have been identified for further improvement and expansion:

- **Real-Time Alerting:** Integrate an automated alert system to notify engineers when critical deviations or slide intervals are detected, or when data logging encounters failures. This will help ensure timely interventions during live operations.
- **Advanced Dimensionality Reduction:** Employ non-linear techniques such as t-SNE or UMAP to better visualize and model high-dimensional drilling data, which may reveal patterns missed by PCA.
- **Interactive UI for Logging System:** Extend the Docker-based logging tool with a lightweight web interface to allow users to monitor logs, configure settings, and visualize trends in real time.
- **Scalable Infrastructure:** Transition data processing from in-memory tools (e.g., Pandas) to distributed systems like Apache Spark or Dask to handle large-scale datasets more efficiently, especially during live operations.
- **Feedback-Driven Clustering Evaluation:** Collaborate with domain experts to validate and refine clustering results, ensuring the flagged intervals are operationally significant and aligned with real-world drilling conditions.
- **Long-Term Data Pipeline Integration:** Integrate the developed modules into a larger pipeline that continuously ingests, processes, and flags data as it arrives from multiple wells, supporting predictive maintenance and optimization across the fleet.

## 9. Bibliography/References


- [1] J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Boca Raton, FL, USA: CRC Press, 1998.
- [2] R. O. Flanagan, H. Onur, and M. A. Davies, “Decision support for drilling optimization using machine learning and real-time data,” in *Proc. SPE/IADC Drilling Conf.*, The Hague, Netherlands, 2019, doi: 10.2118/194285-MS.
- [3] P. W. M. Courteney, “Directional Drilling Surveying: Calculation Methods,” *Oilfield Review*, vol. 5, no. 1, pp. 45–55, 1993.
- [4] G. B. Andrews and D. S. Carter, “The Minimum Curvature Method,” *Journal of Petroleum Technology*, vol. 27, no. 5, pp. 555–561, May 1975.
- [5] M. P. Brown, “Directional Survey Calculations for Oil Wells Using Minimum Curvature Method,” *Petroleum Engineering Handbook*, SPE, 2006.
- [6] M. Lutz, *Learning Python*, 5th ed., O’Reilly Media, 2013.
- [7] T. Rehurek and R. Sojka, “Gensim – Python Framework for Vector Space Modelling,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50, May 2010.
- [8] Python Software Foundation, “Python 3 Documentation,” [Online]. Available: <https://docs.python.org/3/>. Accessed: May 15, 2025.
- [9] Pandas Development Team, “pandas: powerful Python data analysis toolkit,” [Online]. Available: <https://pandas.pydata.org/>. Accessed: May 15, 2025.
- [10] A. Sambasivan, “Working with Asynchronous APIs in Python,” *Real Python*, [Online]. Available: <https://realpython.com/async-io-python/>. Accessed: May 15, 2025.
- [11] T. Oliphant, *A Guide to NumPy*, Trelgol Publishing, 2006.
- [12] Plotly Technologies Inc., “Plotly: Collaborative Data Science,” [Online]. Available: <https://plotly.com/python/>. Accessed: May 15, 2025.

## 10. Peer review

### Annexure A. Evaluation Form for Peer Review

<b>Name of the student: (to be reviewed)</b>	Vasudhara Mahajan	<b>Roll no. of the student:</b>	102103032
Title of the project:	<i>“Enhancing Drilling Operations with Data Analytics and Machine Learning in the Oil and Gas Industry”</i>		
Name of the company:	Helmerich and Payne		
Project report (Tick the appropriate)	Excellent ✓	Good	Average
Project poster (Tick the appropriate)	Excellent ✓	Good	Average
Project video (Tick the appropriate)	Excellent ✓	Good	Average
Rate the work done	0 – 10 points	(Provide rating here) →	10
Give marks to the student on the basis of the overall performance	0 -5 marks	(Provide marks here) →	5
<p>Abstract of the project (max. 100 words):</p> <p>This project focuses on enhancing data-driven decision-making in the oil and gas drilling industry through the analysis of operational data and the development of automation-friendly tools. Conducted in collaboration with Helmerich &amp; Payne (H&amp;P), the project addresses key challenges in identifying slide criticality, understanding regional setpoint usage, and creating resilient data logging systems.</p>			



<p>Mention three strengths of the work done:</p> <ul style="list-style-type: none"> <li>• <b>Real-Time Reliability:</b> Developed a robust, Docker-based logging system capable of functioning under failure conditions like network loss or container crashes, ensuring uninterrupted data capture.</li> <li>• <b>Insightful Data-Driven Analysis:</b> Applied advanced clustering and visualization techniques to uncover critical operational patterns in drilling data, aiding in early anomaly detection and strategic decision-making.</li> <li>• <b>Practical Industry Alignment:</b> All components—from data extraction to logging—were designed with scalability, automation, and real-world operational use in mind, directly supporting H&amp;P’s digital transformation goals.</li> </ul>			
<p>Provide some useful recommendations (It may be some improvements, some suggestions to further raise the quality of the project):</p> <ul style="list-style-type: none"> <li>• <b>Integrate Real-Time Alerting Mechanisms:</b> Enhance the logging system with a real-time alert feature (e.g., email or dashboard notification) to inform operators of anomalies or failures as they occur.</li> <li>• <b>Develop a Visual Monitoring Dashboard:</b> Create an interactive dashboard using tools like <b>Plotly Dash</b>, <b>Streamlit</b>, or <b>Power BI</b> to visualize drilling performance, setpoint trends, and log summaries for faster decision-making.</li> <li>• <b>Enhance Model Evaluation with Domain Expert Feedback:</b> Involve drilling engineers in validating clustering outcomes to ensure the results are operationally meaningful and accurately reflect on-field behavior.</li> </ul>			
Name of the evaluator student:	Swati Jain	Roll no. of the evaluator student:	102103 735
Signature of the Evaluator student:			

## 11. Certificate of the course

