# Big Data Analytics Project Report

**Sumbitted by Fawad Ahmad and Asim Ali**

## 1. Use Case Selection

For this project, I selected a Big Data Analytics use case from the Energy & Oil & Gas domain: Oil Well Production Optimization Using Real-Time & Historical Sensor Data.

## 2. Problem Definition & Scope

Business Problem:

Oil wells frequently underperform due to hidden anomalies in pressure, flow, or temperature signals. These instabilities reduce production rates, increase downtime, and cause costly intervention delays.

**Objective:**

Build a data-driven pipeline capable of:

- Continuously analyzing multivariate sensor streams.

- Detecting undesirable well events early.

- Classifying the type of operational inefficiency.

- Supporting petroleum engineers in optimizing production and minimizing downtime.

## 3. Dataset Selection

Dataset: 3W Dataset (Well Working Window)

Source: Kaggle Public Dataset

Link: https://www.kaggle.com/datasets/afrniomelo/3w-dataset

Origin: Petrobras Offshore Oil Wells

Type: Multivariate Time-Series

Format: ~1,984 individual CSV files

**Why this dataset fits the use case:**

- Real offshore well sensor measurements.
- Designed for fault detection and undesirable events.
- Includes normal vs 8 fault categories.

**Dataset Metadata:**

- Instances: 1,984 sequences
- Features per sample: 8 sensor variables
- Labels: 9 classes (0 = normal, 1-8 = fault types)

**4. Big Data Pipeline Design**

4.1 Data Ingestion:

- Kafka for real-time IoT ingestion.
- Spark Batch for historical data loading.

4.2 Storage Layer:

- Raw Zone: HDFS/S3 storing raw CSV/JSON.
- Processed Zone: Delta Lake (Parquet).
- Feature Store: Redis/Cassandra.

4.3 Data Processing:

Batch (PySpark):

- Missing value handling

- Normalization

- Windowing

- FFT features

Real-Time (Spark/Flink):

- Sliding windows

- Online anomaly scoring

- Drift detection

**5. Machine Learning Pipeline**

Models:

- LSTM/GRU for temporal sequences.

- XGBoost for engineered features.

Hybrid:

1. LSTM → embeddings

2. XGBoost → classification

Evaluation Metrics:

- Precision, Recall, F1

- False Alarm Rate

- Detection Latency

**6. Real-Time Deployment**

- Docker + Kubernetes

- Kafka consumer service

- Grafana dashboards

- MLflow tracking

## 7. Conclusion

This project delivers a fully scalable big data + ML pipeline capable of early detection of offshore well anomalies. Using the 3W dataset, the system improves operational efficiency, reduces downtime, and enables data-driven production optimization.