Declaration of Academic Integrity (Student) Form

Reference Procedure: AQAE022 Academic Integrity Policy

STUDENT NAME	Ujjwal Rastogi
STUDENT NUMBER	L00196925
PROGRAMME	Msc in Computing in Big Data Analytics & Artificial Intelligence
MODULE CODE	LY_IDAAI_M_ANAL_IT902_Y5_2025
ASSIGNMENT TITLE	Big Data Analytics Platform for Energy Forecasting in Modern Power Systems
SUBMISSION DATE	19th January 2026

Student Declaration

1. I have accurately identified and included the sources of all facts, ideas, opinions, and viewpoints from others in the assignment references. All direct quotations, paraphrasing, and discussions of ideas from books, journal articles, internet sources, course materials, or any other sources used are properly acknowledged and cited in the assignment references.
2. I have not used unauthorised artificial intelligence tools or aids.
3. I understand and am compliant with ATU's policy and procedures regarding Academic Integrity and I am aware of the consequences of any violations.
4. I have followed the referencing guidelines recommended in the assignment instructions and / or programme documentation.
5. By signing this form or by submitting material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other programme of study.
6. By signing this form or by submitting material for assessment online I confirm that I have read and understood AQAE022 Academic Integrity Policy

Student Signature	**Date**
Ujjwal Rastogi	19th January 2026

1. Introduction

Modern power systems generate massive volumes of data due to the widespread deployment of smart meters, sensors, and digital monitoring tools. Big Data technologies have therefore become essential in enabling efficient storage, large-scale processing, and advanced analytics in the energy sector.

One critical application of Big Data in power systems is energy forecasting, particularly price forecasting. Accurate price predictions support operational planning, risk mitigation, and decision-making for utilities. However, electricity prices are highly volatile and influenced by multiple factors, including demand, generation mix, fuel prices, and grid congestion.

This report summarizes the academic case study titled AI–Big Data Analytics Platform for Energy Forecasting in Modern Power Systems [1]. The paper proposes an open-source, on-premises Big Data architecture designed to support scalable data analytics and machine learning for electricity price forecasting. The focus of this report is on the Big Data architecture, the technologies used for storage and processing, and the role of analytics within the platform.

2. Overview of the Case Study

The selected case study presents the design and implementation of a Big Data Analytics Platform tailored for modern power systems. The primary objective of the platform is to enable the automated execution of intelligent forecasting models using large volumes of historical and operational data.

The platform is built using open-source technologies and deployed on-premises to reduce costs and maintain control over sensitive data. A data lake architecture is adopted to handle structured, semi-structured, and unstructured data originating from multiple sources, including electricity market data, weather variables, and fuel prices.

To evaluate the platform, the authors compare several forecasting models. The best-performing models are integrated into the platform and executed automatically to generate forecasts at regular intervals. The results are then visualized through dashboards to support operational decision-making.

3. Big Data Architecture Description

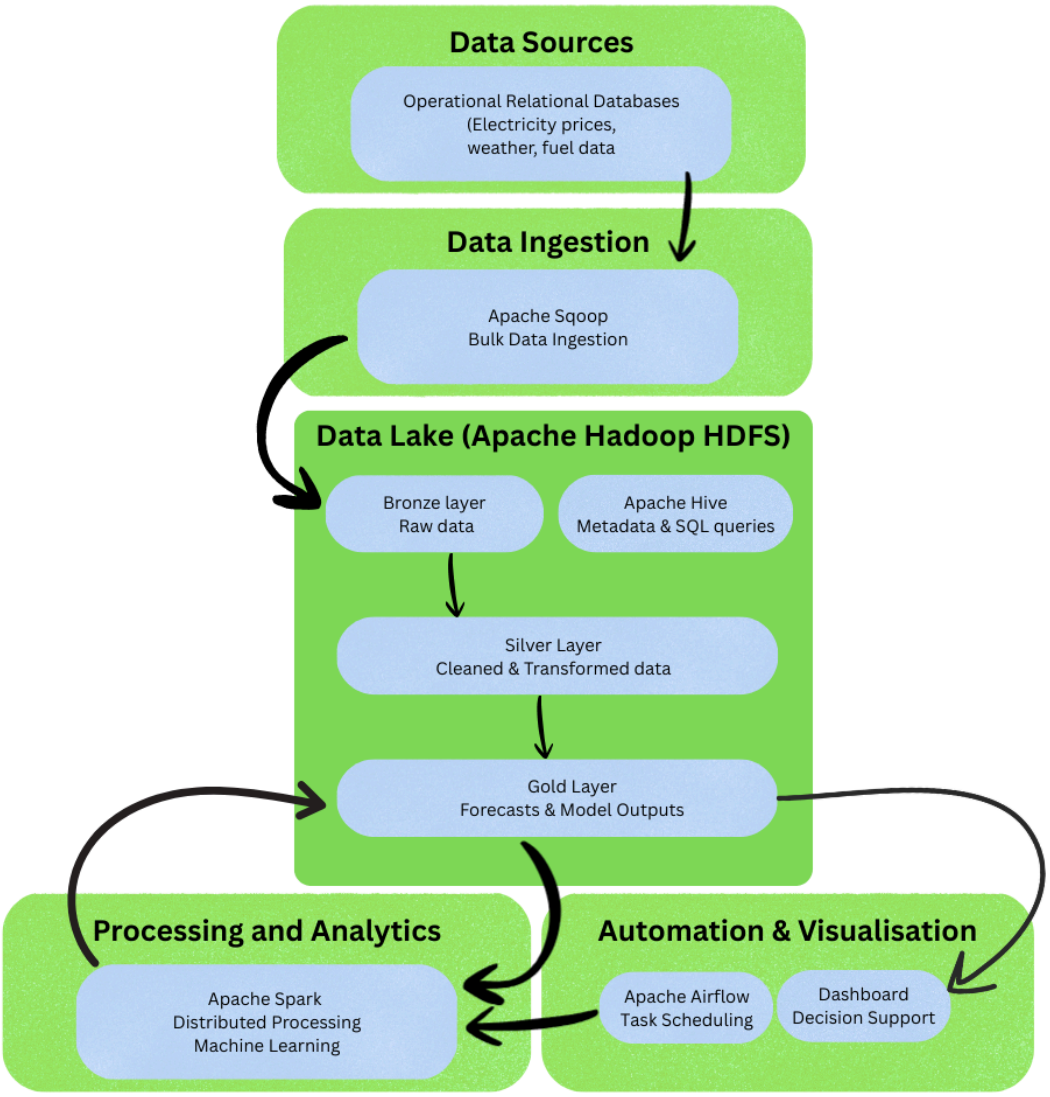

***Figure 1**. Big Data Analytics architecture for energy forecasting based on a data lake and Apache Spark.*

Figure 1 illustrates the layered Big Data Analytics architecture for energy forecasting based on a data lake and Apache Spark. The platform follows a layered data lake design that enables scalable data ingestion, storage, distributed processing, and automated analytics using open-source technologies.

3.1 Data Sources and Ingestion

The platform ingests data from multiple operational sources relevant to power system operation and electricity markets. These sources include relational databases containing historical electricity prices by node, market components such as energy, congestion, and losses, as well as external influencing variables such as weather conditions and fuel prices.

Apache Sqoop is used as the primary data ingestion tool to transfer large volumes of structured data from relational databases into the Big Data platform. Sqoop supports parallel data ingestion, allowing efficient bulk transfers without impacting the performance of operational systems. This ensures that historical and operational datasets are consistently integrated into the analytics environment.

3.2 Data Lake Storage Architecture

At the core of the platform is a data lake implemented using Apache Hadoop and the Hadoop Distributed File System (HDFS) [2]. The data lake is logically organized into three layers.

The **bronze layer** stores raw data exactly as it is ingested from source systems. This layer serves as a reliable historical record and allows data to be reprocessed if required.

The **silver layer** contains cleaned and transformed data that has undergone preprocessing steps such as data validation, normalization, and handling of missing values. Data in this layer is structured to support efficient analysis and model development.

The **gold layer** stores analytics-ready outputs, including trained machine learning models and electricity price forecasting results. This layer is optimized for consumption by analytics applications and visualization tools.

Apache Hive is used on top of HDFS to manage metadata and enable SQL-based querying of structured datasets within the data lake. Hive simplifies data access for analysts and supports integration with downstream processing and visualization components.

3.3 Processing and Analytics Layer

Apache Spark serves as the main distributed processing engine within the platform [3]. Spark is used to perform large-scale data transformations, feature engineering, and execution of machine learning models for energy forecasting. Its in-memory computation model enables faster processing compared to traditional disk-based approaches, particularly for iterative analytics workloads.

The platform supports parallel execution of forecasting models across multiple nodes, allowing region-specific or node-level models to be trained and executed simultaneously. This capability is essential for large power systems where localized forecasts are required.

3.4 Automation and Orchestration

To ensure consistent and repeatable analytics workflows, Apache Airflow is integrated as the orchestration and scheduling component. Airflow manages the execution of data ingestion, preprocessing, model training, and forecasting tasks according to predefined schedules.

By automating these workflows, the platform enables regular generation of electricity price forecasts at intervals such as 24, 72, or 120 hours without manual intervention. This design supports near real-time operational analytics and improves reliability of the forecasting process.

3.5 Visualization and Decision Support

Forecasting results generated by the analytics pipeline are stored in the gold layer of the data lake and made available to visualization tools. Dashboards present comparisons between actual and predicted electricity prices, enabling stakeholders to assess model performance and support operational and strategic decision-making.

This integration ensures that insights from Big Data processing are accessible and actionable.

4. Data Analytics and Machine Learning

The platform supports multiple types of analytics, ranging from descriptive and diagnostic analytics to predictive analytics. In the case study, predictive analytics is emphasized through electricity price forecasting.

The authors evaluate statistical, machine learning, and deep learning models for electricity price forecasting. The paper emphasizes operationalization of forecasting models within the Big Data platform rather than algorithmic novelty.

The best-performing models are selected based on error metrics and deployed into the automated pipeline, demonstrating how advanced analytics can be embedded into real-world power system operations.

5. System Scalability and Design Choices

A key design choice in the proposed platform is the use of open-source technologies deployed on-premises. This approach significantly reduces licensing costs and allows the system to be tailored to existing organizational infrastructure. Horizontal scalability is achieved by adding nodes to the Hadoop and Spark clusters as data volumes or computational demands increase.

The data lake architecture provides flexibility to incorporate new data sources and analytics use cases in the future, such as load forecasting, renewable generation forecasting, and asset health monitoring. This makes the platform a long-term foundation for data-driven energy system management.

6. Conclusion

This case study demonstrates how a well-designed Big Data architecture can support advanced energy forecasting in modern power systems. By combining a data lake architecture with distributed processing and machine learning capabilities, the proposed platform enables scalable, automated, and cost-effective analytics.

The separation of storage, processing, and orchestration layers makes the architecture easy to extend. The integration of machine learning models into an automated pipeline illustrates how Big Data technologies can move beyond experimentation and deliver operational value.

7. References

[1] *AI–Big Data Analytics Platform for Energy Forecasting in Modern Power Systems*, academic case study.

[2] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2015.

[3] M. Zaharia *et al*., "Apache Spark: A Unified Engine for Big Data Processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.