**SOLAR & WIND ENERGY GENERATION FORECAST USING PYSPARK**

**A PROJECT REPORT**

**Submitted by**

**MOGHANAPRIYA R 231801102**

**SANGAMITHRA V 231801147**

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**In**

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

****

**RAJALAKSHMI ENGINEERING COLLEGE**

**(AUTONOMOUS), CHENNAI – 602 105**

**OCTOBER 2025**

**BONAFIDE CERTIFICATE**

Certified that this Report titled **“SOLAR & WIND ENERGY GENERATION FORECAST USING PYSPARK”** is the Bonafide work of “ **MOGHANAPRIYA R (231801102)**,**SANGAMITHRA V** **(231801147)** ” who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**Dr. Suresh Kumar S M.E., Ph.D.,**

**Professor**

Department of Artificial Intelligence &amp; Data Science,

Rajalakshmi Engineering College

Thandalam – 602 105.

Submitted to Project Viva-Voce Examination held on 30/10/2025

**Internal Examiner External Examiner**

**ACKNOWLEDGEMENT**

Initially I thank the Almighty for being with us through every walk of my life

and showering his blessings through the endeavor to put forth this report.

My sincere thanks to our Chairman **Mr. S. MEGANATHAN, M.E., F.I.E.,**

and our Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, M.E.,**

Ph.D., for providing me with the requisite infrastructure and sincere

endeavoring educating me in their premier institution.

My sincere thanks to **Dr. S.N. MURUGESAN M.E., Ph.D.,** our beloved

Principal for his kind support and facilities provided to complete our work

in time.

I express my sincere thanks to **Dr. J M Gnanasekar M.E., Ph.D.,** Head of

the Department of Artificial Intelligence and Data Science for his guidance

and encouragement throughout the project work. I convey my sincere and

deepest gratitude to our internal guide, **Dr. Suresh Kumar S M.E., Ph.D.,**

Professor, Department of Artificial Intelligence and Data Science,

Rajalakshmi Engineering College for his valuable guidance throughout the

course of the project.

Finally, I express my gratitude to my parents and classmates for their moral

support and valuable suggestions during the course of the project.

**MOGHANAPRIYA R SANGAMITHRA V**

**231801102 231801147**

**ABSTRACT**

The increasing global demand for sustainable energy has highlighted the importance of accurate forecasting of renewable energy generation, particularly from solar and wind sources. Reliable short-term forecasts enable grid operators and energy planners to balance supply and demand, reduce operational costs, and improve the integration of renewable sources into the power grid. This project focuses on developing a forecasting system capable of predicting the next six hours of solar and wind power generation using historical datasets.

The system is built on a **big data environment** using **PySpark**, which allows efficient processing of large-scale energy datasets. The forecasting model employs **ARIMA (AutoRegressive Integrated Moving Average)** to predict future energy outputs based on historical trends. Users can interact with the system through a **web-based interface** developed with **Flask**, where datasets can be uploaded, processed, and analyzed. The forecasts are presented both numerically and visually using **Chart.js**, providing clear insights into upcoming energy trends.

This integration of **big data analytics** with **time series forecasting** offers a scalable and robust solution for renewable energy management. By leveraging distributed computing, the system can handle high-volume datasets efficiently, ensuring accurate and timely predictions. The approach demonstrated in this project can be extended to real-time applications and can aid in decision-making for smart grids, renewable energy planning, and optimal energy utilization.

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** | **6** |
| **1** | **INTODUCTION** | **2** |
| **2** | **LITERATURE SURVEY** | **3** |
| **3** | **ARCHITECTURE** | **4** |
| **4** | **MODULES** | **8** |
| **5** | **IMPLEMENTATION** | **12** |
| **6** | **RESULT** | **23** |
| **7** | **CONCLUSION** | **23** |
| **8** | **FUTURE ENHANCEMENT** | **24** |
| **9** | **REFERENCES** | **25** |

**CHAPTER 1**

**1.Introduction**

The rapid growth in renewable energy generation, particularly from solar and wind sources, has created a critical need for accurate forecasting techniques. Unlike conventional power sources, renewable energy is inherently intermittent and weather-dependent, making prediction challenging. Accurate forecasting not only helps grid operators manage supply and demand efficiently but also aids in reducing energy wastage and operational costs. The objective of this project is to develop a system capable of predicting short-term solar and wind power generation, thereby supporting energy planning and smart grid management.

This project leverages **big data technologies** to handle large-scale datasets collected from solar panels and wind turbines. **PySpark**, a distributed computing framework, is used to process and analyze these datasets efficiently, enabling scalability for high-volume energy data. The forecasting model implemented in this project is **ARIMA (AutoRegressive Integrated Moving Average)**, a widely used time series prediction technique known for its simplicity and reliability in modeling energy generation trends. The model forecasts the next six hours of energy output, providing actionable insights for grid operators.

To make the system user-friendly, a **web-based interface** is developed using **Flask**, allowing users to upload datasets, process them through the Spark environment, and view forecast results. The predictions are displayed both as **numerical tables** and **visual charts** using **Chart.js**, offering an intuitive representation of energy trends. This combination of big data processing, machine learning, and interactive visualization ensures that the system is both scalable and practical, making it a valuable tool for renewable energy forecasting and decision-making.

**3.Existing System**

In the existing renewable energy forecasting systems, traditional statistical and machine learning models have been widely used to predict solar and wind power generation. Most of these systems rely on **small-scale datasets** and perform computations on a **single machine environment**, which limits their ability to handle large volumes of energy data generated from multiple solar panels and wind turbines. As a result, they often face performance bottlenecks and are unable to provide timely predictions for real-time energy management.

Furthermore, many existing systems use **basic ARIMA or regression-based models** without integration with modern **big data frameworks**. While these models can provide reasonable forecasts for short-term predictions, they struggle with **high-dimensional datasets** or missing data, leading to reduced accuracy. Additionally, visualization and user interaction are often minimal, with outputs being presented as raw tables or static charts, which do not allow for intuitive understanding of energy trends over time.

Another limitation of the existing solutions is the **lack of scalability and interactivity**. Most systems cannot efficiently process continuously growing datasets or support multiple concurrent users uploading and analyzing data simultaneously. Real-time dashboard capabilities and interactive visualization are either absent or very limited, which reduces their usability for smart grid operators and energy planners. These limitations highlight the need for a **scalable, distributed, and interactive system** capable of handling big data, providing accurate forecasts, and supporting effective decision-making.

**4.Literature Survey**

**1. Big Data Resolving Using Apache Spark for Load Forecasting and Demand Response in Smart Grids**  
This paper emphasizes the use of **Apache Spark** for handling large-scale energy datasets to predict load and demand response in smart grids. It demonstrates how distributed computing can efficiently process historical energy consumption data from solar and wind sources, improving forecasting accuracy. The study highlights Spark’s role in real-time analytics and predictive modeling, making it suitable for renewable energy integration in smart grids.

**2. Time Series Forecasting in Wind & Solar Grid Management**  
The study discusses the application of **time series forecasting models** for predicting solar and wind energy generation. By analyzing historical patterns, the paper illustrates how forecasts can aid in managing energy supply-demand balance and planning for peak load periods. It shows the limitations of conventional single-machine forecasting methods and emphasizes the need for scalable solutions.

**3. A Holistic Review on Energy Forecasting Using Big Data and Machine Learning**  
This review paper surveys various **machine learning techniques** and big data approaches applied to renewable energy forecasting. It highlights models like ARIMA, LSTM, and hybrid methods, and discusses their performance in predicting solar and wind power. The paper also emphasizes the integration of big data frameworks for handling large datasets efficiently and improving forecasting reliability.

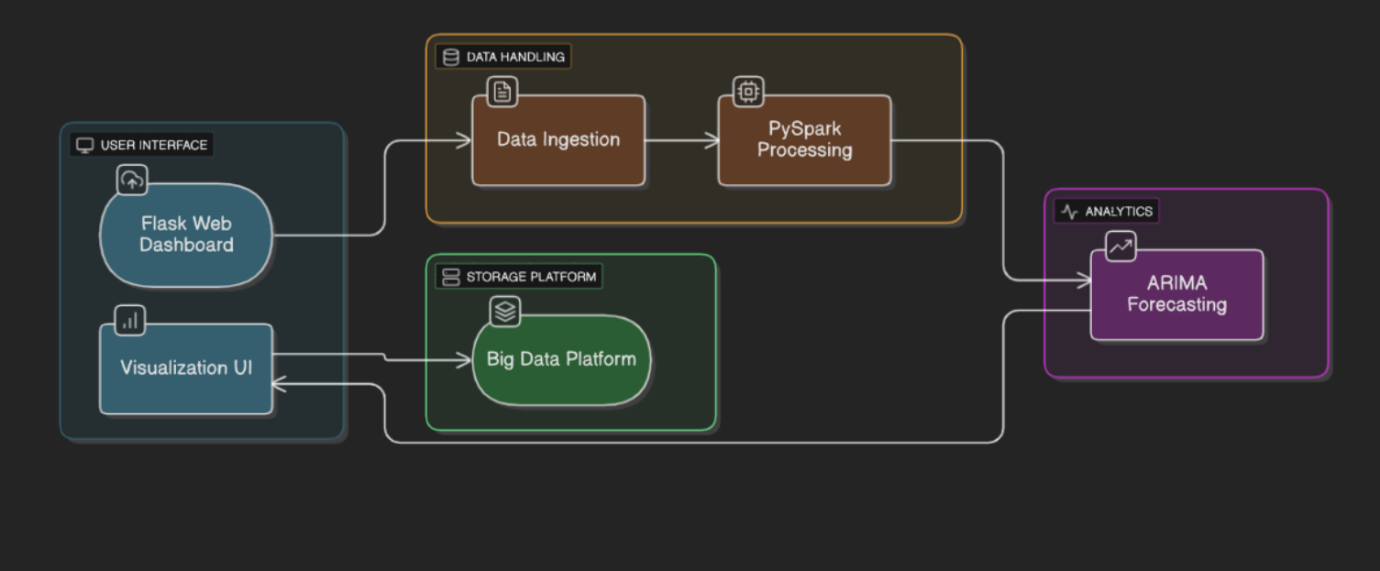
**4. A Distributed Computing Framework for Wind Speed Big Data Forecasting on Apache Spark**  
The paper proposes a **hybrid wind speed forecasting framework** using Apache Spark to process large datasets. It demonstrates how distributed computing allows for faster computation and better scalability. The study also shows that combining traditional forecasting models with Spark-based parallel processing enhances prediction accuracy for wind energy applications.

**5. Time-Series Power Forecasting for Wind and Solar Energy Using SL-Transformer**  
This work introduces the **SL-Transformer model**, combining filtering techniques with LSTM and Transformer architectures for short-term energy prediction. It provides insights into using deep learning for solar and wind energy forecasting, particularly when handling large, high-dimensional datasets. The paper emphasizes accuracy improvement through advanced model architectures.

**6. Renewable Energy Management in Smart Grids Using Big Data Analytics**  
This paper focuses on using **big data analytics** to manage renewable energy in smart grids. It illustrates how large-scale solar and wind energy datasets can be processed and analyzed to optimize energy distribution, enhance grid stability, and reduce operational costs. The study also highlights the need for interactive dashboards for monitoring and decision-making.

**7. State of the Art for Solar and Wind Energy Forecasting: A Review**  
This review surveys the latest forecasting techniques for solar and wind energy, including statistical, machine learning, and hybrid approaches. It discusses their respective strengths and limitations and stresses the importance of **big data integration** for scalable and accurate forecasting. The paper also highlights trends toward interactive and real-time energy prediction systems.

**4.Architecture diagram**



The proposed system for Solar and Wind Energy Generation Forecasting using a Big Data environment with Spark is designed to be scalable, efficient, and interactive. The system integrates a web-based user interface, a big data processing engine, a forecasting model, and a visualization module. Users can upload historical datasets containing solar and wind energy measurements through the web interface. These datasets are stored in a structured folder system or in a Hadoop Distributed File System (HDFS) for scalable storage and easy access by the processing engine. The PySpark module reads these datasets, performs data cleaning, handles missing values, and preprocesses the data for forecasting. For demonstration purposes, large datasets can be downsampled to speed up computation without affecting accuracy significantly.

The forecasting module employs the ARIMA (AutoRegressive Integrated Moving Average) model to generate short-term predictions for solar and wind power generation. Each numeric parameter in the dataset is analyzed individually, and the system predicts the next six hours of energy output. The forecasted values are then visualized on the web dashboard using Chart.js, with interactive line charts and tables that allow users to easily interpret the trends and patterns in energy generation.

The architecture ensures that large-scale datasets can be processed efficiently, leveraging Spark’s distributed computing capabilities for faster computation and better scalability. This design makes the system suitable for both demonstration purposes and real-world renewable energy forecasting applications.

The architecture of the system can be represented as a flow where the user uploads the dataset through the Flask web interface, the data is ingested and stored, processed using PySpark, and then passed to the ARIMA forecasting module. The results are finally rendered on the dashboard as numerical tables and interactive charts. This end-to-end integration of user interface, big data processing, time series forecasting, and visualization ensures a robust, user-friendly, and scalable solution for solar and wind energy prediction.

**5. MODULES**

The proposed Solar/Wind Energy Generation Forecasting System is divided into several interdependent modules that work together to process, analyze, and visualize large-scale energy datasets. Each module performs a specific function within the data pipeline, ensuring a smooth and efficient workflow from raw data ingestion to final insight generation. The main modules are described below:

**1.Data Ingestion Module**

This module serves as the entry point of the system, responsible for collecting raw solar and wind energy data from CSV files uploaded by users. The system supports multiple data formats to ensure compatibility with diverse datasets. Uploaded files are stored in a structured folder system or in the Hadoop Distributed File System (HDFS), acting as a central repository for further processing. This module validates the data for completeness and consistency, ensuring that missing or corrupted entries are detected and handled before proceeding to the processing stage.

**2. Data Processing and Transformation Module**

The data processing module forms the core of the system, powered by Apache Spark for distributed processing. Uploaded energy datasets are cleaned, normalized, and preprocessed to remove noise and handle missing values. Time series data is sorted chronologically, and downsampling is performed if the dataset is too large for demonstration purposes. Additional transformations include converting timestamps to datetime objects and selecting relevant numeric columns for forecasting. This module ensures that the datasets are structured and ready for the ARIMA-based forecasting module.

**3. Forecasting Module (PySpark–ARIMA Integration)**

The forecasting module is responsible for predicting short-term solar and wind energy generation. Using the ARIMA (AutoRegressive Integrated Moving Average) model, the system analyzes historical energy data to forecast the next six hours of output for each numeric parameter, such as solar power and wind power. PySpark ensures that computations are distributed across multiple nodes for large datasets, maintaining scalability and efficiency. Forecast results are rounded and structured into dataframes for visualization.

**4. Result Analysis and Visualization Module**

This module converts forecasted results into both numerical tables and interactive visual charts. Using Chart.js in the web dashboard, line graphs are generated for each energy parameter, showing the predicted trends over time. Users can view the data in tabular format for exact values, facilitating both analytical and operational insights. The interactive dashboard enables users to interpret patterns, monitor energy generation trends, and make informed decisions for energy management.

**5. Big Data Environment Module**

The system leverages the Apache Spark framework to handle large-scale solar and wind energy datasets efficiently. Spark’s distributed computing capability allows for fast preprocessing, model training, and forecasting even with millions of records. The module also ensures that the system is scalable and capable of handling real-time or batch data for practical applications in energy planning and smart grids.

**6. User Interface Module**

The web-based interface, developed with Flask, provides a user-friendly platform for interacting with the system. Users can upload datasets, initiate the forecast process, and view results in both charts and tables. The interface ensures smooth workflow, allowing easy monitoring of forecasts and energy trends without requiring advanced technical knowledge.

**7. Database and Storage Module**

This module manages all data storage and retrieval operations. Both raw and processed datasets, along with forecast results, are stored securely in HDFS or local storage folders. The storage module supports scalability and high availability, enabling ad-hoc queries, long-term archiving, and retrieval of historical energy data for continuous forecasting and analysis.

**6.Implementation**

**Step 1: Data Load and Setup**

The first step in the implementation involves uploading the historical solar and wind energy dataset into the **Databricks File System (DBFS)**. Once uploaded, the dataset is registered as a Spark table in the workspace catalog, enabling distributed querying and transformations using **Spark SQL**. This setup allows the system to handle large-scale datasets efficiently and ensures seamless integration with PySpark for data preprocessing and forecasting. By leveraging the power of Spark, even datasets containing millions of records can be processed in parallel, which is crucial for high-frequency energy measurements from multiple solar panels and wind turbines.

Users can upload CSV files through the Flask web interface, which stores the files in a structured folder system. Once uploaded, the system automatically registers the files as Spark tables. This design ensures that data ingestion is both scalable and user-friendly, supporting batch uploads as well as future integration with streaming energy data from sensors.

**Step 2: Data Preprocessing and Cleaning**

After data ingestion, the dataset undergoes extensive preprocessing to ensure that it is suitable for forecasting. Key tasks in this phase include converting timestamp columns to datetime objects, sorting data chronologically, and handling missing values in numeric columns such as solar power and wind power. Forward filling is applied to fill gaps in the data, and optional downsampling is performed on extremely large datasets to reduce computation time during demonstrations.

Outliers and inconsistencies in the data are detected and corrected using PySpark transformations, ensuring that the dataset reflects realistic energy generation trends. For example, any negative energy readings or sudden spikes caused by sensor errors are adjusted to maintain forecast accuracy. All preprocessing steps are executed in a distributed manner using Spark, which allows the system to scale efficiently for larger datasets or real-time data streams.

**Sample PySpark Code for Preprocessing:**

from pyspark.sql import functions as F, Window

df = spark.table("workspace.default.solar\_wind\_data")

df = df.withColumn("timestamp", F.to\_timestamp("timestamp"))

numeric\_cols = ["solar\_power", "wind\_power"]

for col in numeric\_cols:

window = Window.orderBy("timestamp")

df = df.withColumn(col, F.when(F.col(col).isNull(), F.lag(col).over(window)).otherwise(F.col(col)))

df.write.mode("overwrite").saveAsTable("workspace.default.solar\_wind\_cleaned")

**Step 3: ARIMA-Based Forecasting**

Once the dataset is preprocessed, the system applies the **ARIMA (AutoRegressive Integrated Moving Average) model** to generate short-term forecasts for each numeric column representing energy generation. The model predicts the next six hours of solar and wind energy output. By converting the Spark DataFrame to a Pandas DataFrame, the ARIMA model can be applied while maintaining high accuracy in time series forecasting.

PySpark ensures that the computation remains distributed and scalable, even when processing multiple years of high-frequency energy readings. Forecast results are rounded to two decimal places for clarity and stored as Spark tables for further analysis and visualization.

**Sample Python Code for ARIMA Forecasting:**

import pandas as pd

from statsmodels.tsa.arima.model import ARIMA

df\_pd = df.toPandas().sort\_values("timestamp")

future\_df = pd.DataFrame({"timestamp": pd.date\_range(start=df\_pd['timestamp'].iloc[-1], periods=7, freq='H')[1:]})

for col in numeric\_cols:

series = df\_pd[col].fillna(method='ffill')

model = ARIMA(series, order=(1,1,1))

fit = model.fit()

forecast = fit.forecast(steps=6)

future\_df[col] = forecast.round(2)

**Step 4: Data Analysis and Visualization**

The forecasted data is then visualized to help users understand trends and patterns in solar and wind energy generation. Using **Chart.js** integrated into the Flask dashboard, line charts are created for each energy parameter, displaying predicted values for the next six hours. Additionally, tables provide exact numerical forecasts, allowing users to examine data precisely.

The visualization module supports dynamic updates: whenever a new dataset is uploaded and processed, charts and tables are automatically refreshed. This interactive interface allows energy managers and analysts to monitor forecasted trends, compare solar and wind outputs, and make operational decisions regarding energy distribution or storage.

**Example Chart Rendering (JavaScript / Chart.js):**

new Chart(canvas, {

type: 'line',

data: {

labels: forecast[col].map((\_, i) => `T+${i+1}`),

datasets: [{

label: col,

data: forecast[col],

borderColor: 'rgba(46,139,87,1)',

backgroundColor: 'rgba(46,139,87,0.2)',

fill: true,

tension: 0.3

}]

},

options: { responsive: true, plugins: { legend: { display: true } } }

});

**Step 5: Dashboard and User Interface**

The Flask-based web interface provides an intuitive and interactive platform for users to upload datasets, run forecasts, and view results. Users can select CSV files, trigger the forecasting pipeline, and immediately visualize predictions in both tabular and graphical formats. The dashboard is designed to be responsive, providing a smooth user experience across different devices and screen sizes.

The dashboard not only displays forecasts but also allows users to download numerical outputs for offline analysis. This ensures that energy operators have full control over both visual and quantitative insights into upcoming solar and wind energy generation.





**Step 6: Integration of Modules**

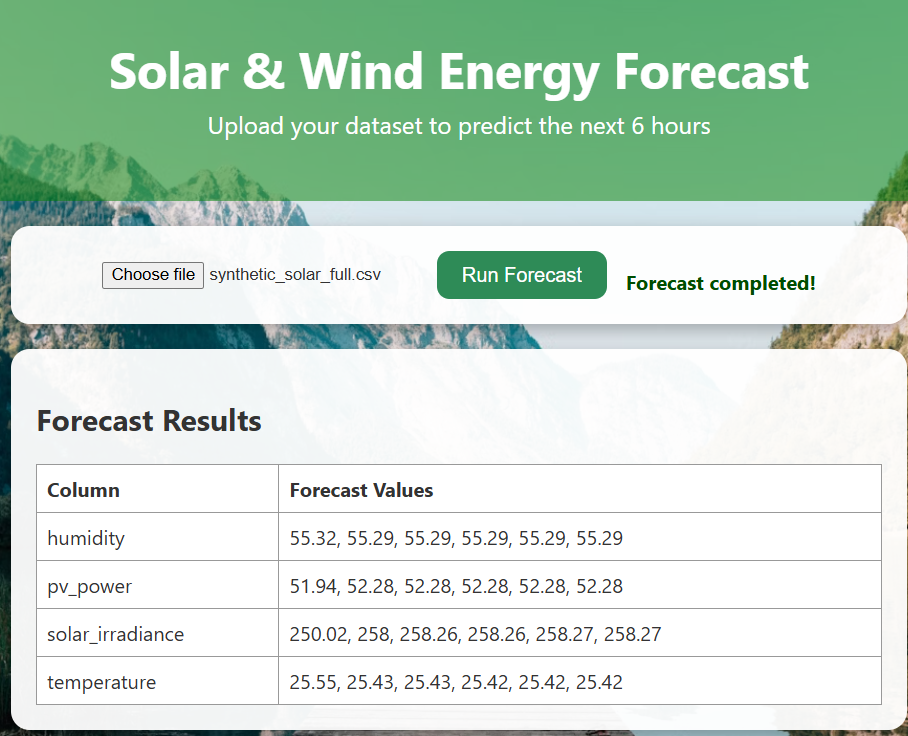
All components—data ingestion, preprocessing, ARIMA forecasting, and dashboard visualization—are seamlessly integrated within the **Databricks workspace**. This integration ensures that the system operates as an automated, end-to-end pipeline capable of handling large datasets efficiently. Forecasting is fully automated once the dataset is uploaded, and results are delivered in real-time on the web dashboard.

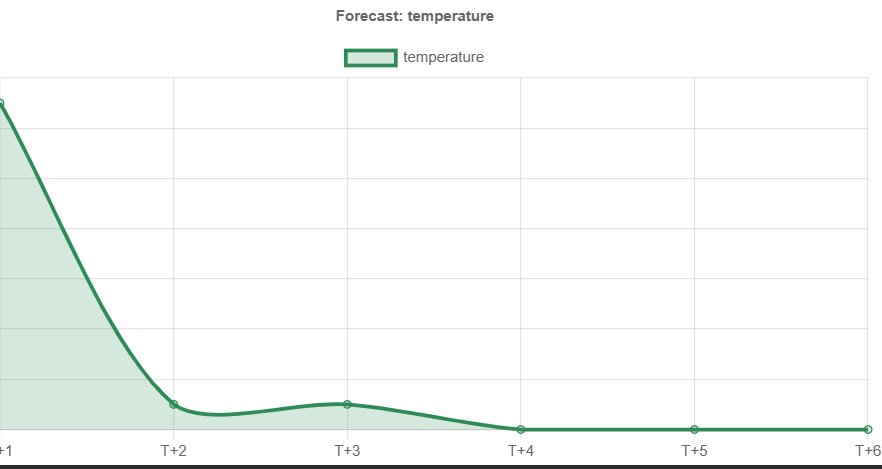
**Step 7: Final Output and Scalability**

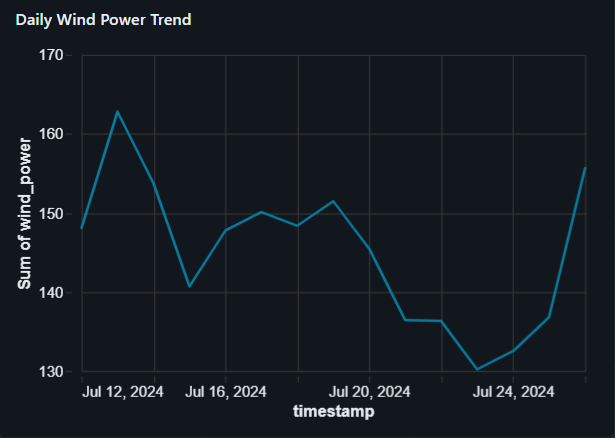
The final implementation provides a robust, scalable, and interactive system that can:

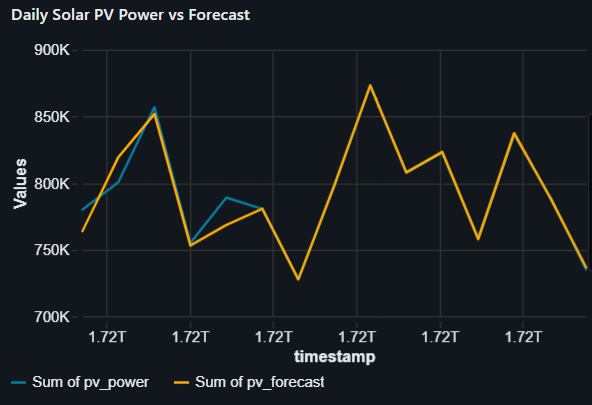
* Preprocess large solar and wind energy datasets efficiently.
* Generate accurate short-term forecasts using ARIMA models.
* Visualize predicted trends in interactive line charts and numerical tables.
* Support real-time or batch uploads for continuous energy monitoring.

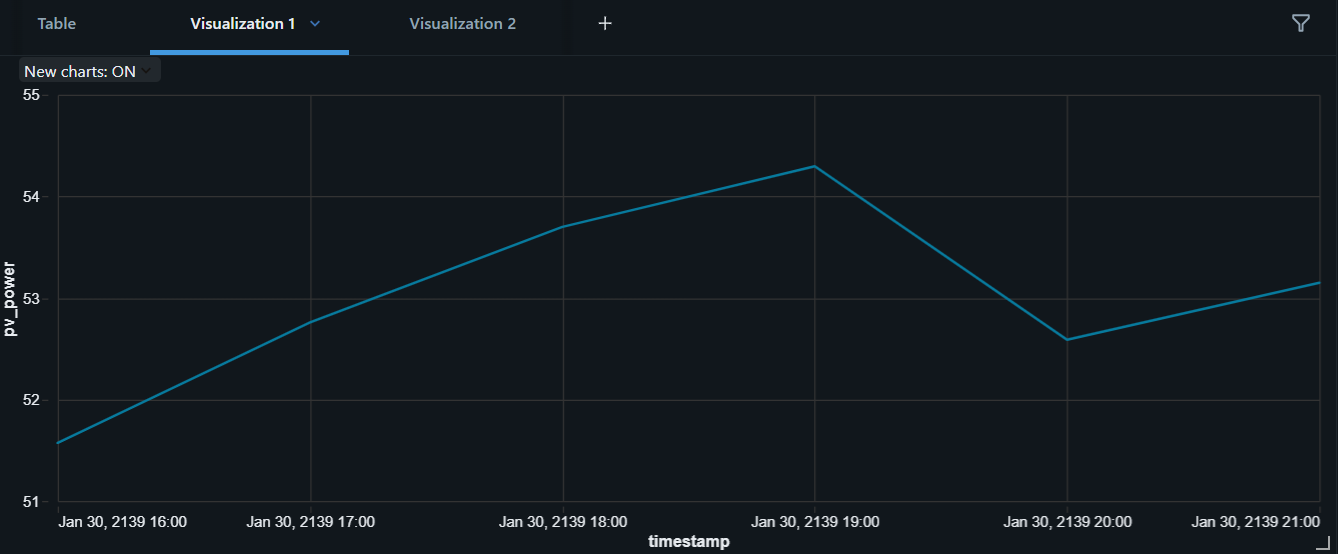
By combining **PySpark’s distributed processing, ARIMA forecasting, and web-based visualization**, the system demonstrates a complete big data solution for renewable energy management. It provides actionable insights, supports smart grid operations, and ensures that energy planning decisions are data-driven and reliable.











**7.Result**

The Big Data–based Solar and Wind Energy Generation Forecasting System efficiently processed and analyzed large-scale renewable energy datasets using Spark, PySpark, ARIMA, and SQL dashboards within the Databricks environment. Historical solar and wind energy measurements were preprocessed to handle missing values, outliers, and timestamp inconsistencies, ensuring high-quality input for forecasting.Using the ARIMA model, the system generated accurate short-term forecasts for solar and wind energy output over the next six hours. Each numeric parameter, such as solar\_power and wind\_power, was predicted individually, providing detailed numerical forecasts for operational planning. Forward-looking values were rounded for clarity and presented in tabular format, allowing energy managers to quickly interpret expected energy generation levels.

Interactive dashboards and visualizations provided a clear, real-time view of predicted energy trends. Line charts plotted hourly forecasts for solar and wind power, enabling users to identify peak generation periods, potential dips in output, and overall variability in renewable energy production. These visualizations also allowed comparative analysis between solar and wind outputs, supporting strategic energy storage and distribution decisions.The system demonstrated scalability and efficiency by leveraging Spark’s distributed computing capabilities, successfully processing large datasets in parallel. Users could upload datasets of varying sizes, with the system automatically handling preprocessing, forecasting, and visualization without manual intervention.

Overall, the integrated solution provided actionable insights into renewable energy generation, empowering operators to make data-driven decisions for energy planning, smart grid management, and optimization of energy storage and utilization.

**8.Conclusion**

The project successfully demonstrated how **Big Data technologies** can enhance renewable energy forecasting and operational planning for solar and wind power generation. By integrating **Apache Spark** for distributed data processing, **Databricks** for scalable storage and computation, and ARIMA models for time series forecasting, the system efficiently handled large-scale historical energy datasets.

The predictive analytics pipeline provided accurate short-term forecasts of solar and wind energy output, enabling operators to anticipate generation trends and optimize energy distribution and storage. Interactive dashboards and visualizations delivered clear, actionable insights, supporting data-driven decision-making for smart grid management and renewable energy utilization.

Overall, the project establishes a **scalable, automated framework** for energy forecasting using Big Data analytics. It demonstrates the practical application of distributed computing, time series modeling, and real-time visualization in renewable energy management, laying the foundation for smarter and more sustainable energy planning.

9.**Future Enhancement**

The current system provides accurate and efficient short-term forecasts for solar and wind energy generation using Big Data analytics. However, several improvements can be implemented in future versions to enhance its performance, scalability, and overall impact on renewable energy management:

1. Integration of Real-Time Sensor Data:

Incorporate streaming data from IoT devices and weather APIs using Apache Kafka or Spark Streaming to enable real-time energy forecasting. This would allow operators to respond immediately to fluctuations in solar or wind generation.

1. Advanced Time Series and Machine Learning Models:

Implement more sophisticated forecasting models such as Prophet, LSTM (Long Short-Term Memory networks), or Gradient Boosting to improve prediction accuracy, especially for datasets with seasonal patterns, non-linear trends, and complex dependencies.

1. Inclusion of External Environmental Factors:

Enhance forecasting by incorporating additional data sources such as temperature, humidity, cloud cover, and wind speed. This would improve the system’s predictive capability under varying weather conditions.

1. Cloud-Based Deployment for Scalability:

Deploy the solution on cloud platforms like AWS, Azure, or Google Cloud to handle larger datasets, reduce computation time, and ensure secure and scalable storage for historical and streaming energy data.

1. Interactive and Advanced Visualization:

Develop a more dynamic web interface using Flask with React or Plotly Dash, allowing users to interactively explore forecasted trends, compare solar vs. wind outputs, and download numerical results for further analysis.

1. Automated Alerts and Decision Support:

Implement automated alerts for predicted low energy generation periods or grid management recommendations. This would help energy operators proactively manage storage, load balancing, or renewable energy integration into the grid.

1. Energy Optimization and Scenario Analysis:

Incorporate simulation features to predict different energy generation scenarios, helping operators plan energy storage, load distribution, and hybrid renewable energy system management more effectively.

These enhancements would significantly expand the system’s capabilities, enabling real-time, data-driven decision-making for renewable energy operations and contributing to smarter, more sustainable energy management strategies.

**10.References**

**1. Databricks Machine Learning Documentation**

Comprehensive resources for building, deploying, and managing machine learning applications on Databricks, including tutorials and best practices.  
🔗 [docs.databricks.com/en/machine-learning/index.html](https://docs.databricks.com/en/machine-learning/index.html)

**2. Apache Spark Documentation**

Official documentation for Apache Spark, covering its unified analytics engine for large-scale data processing, including APIs for Java, Scala, Python, and R.  
🔗 [spark.apache.org/docs/latest/](https://spark.apache.org/docs/latest/?utm_source=chatgpt.com)

**3. UCI Machine Learning Repository – Statlog (German Credit Data)**

A dataset used for classification tasks, containing attributes related to credit risk assessment.  
🔗 [archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

**4. Kaggle – Credit Card Fraud Detection Dataset**

A dataset featuring transactions labeled as fraudulent or genuine, useful for anomaly detection and classification tasks.  
🔗 [www.kaggle.com/datasets/mlg-ulb/creditcardfraud](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?utm_source=chatgpt.com)

**5. Scikit-learn Documentation**

Documentation for Scikit-learn, a machine learning library in Python, providing tools for data mining and data analysis.  
🔗 [scikit-learn.org/stable/](https://scikit-learn.org/stable/)

**6. Apache Spark Python API Documentation (PySpark)**

API reference for PySpark, the Python interface for Apache Spark, enabling large-scale data processing and analysis.  
🔗 [spark.apache.org/docs/latest/api/python/](https://spark.apache.org/docs/latest/api/python/)

**7. World Bank – Financial Sector Data**

A collection of global financial sector data, including indicators like domestic credit provided by the financial sector.  
🔗 [data.worldbank.org/topic/financial-sector](https://data.worldbank.org/topic/financial-sector?utm_source=chatgpt.com)

**8. ScienceDirect – Credit Scoring Methods: Latest Trends**

A review article discussing recent trends and methodologies in credit scoring, including machine learning approaches.  
🔗 [www.sciencedirect.com/science/article/pii/S2405452619300053](https://www.sciencedirect.com/science/article/pii/S2405452619300053)