Energy Consumption Analytics Project for GreenPower Utilities

Introduction

GreenPower Utilities is seeking proposals for a six-week data engineering capstone in the energy sector. Energy companies collect ever-larger datasets from smart grids and IoT devices, presenting opportunities to improve efficiency. In fact, the big data market in energy is projected to grow from about $9.6 billion in 2024 to $16.2 billion by 2029. In this project, teams will act as data engineers at an energy provider, building pipelines to analyze energy usage, optimize production, and support sustainability.

Project Scope and Objectives

Participants will tackle realistic energy data challenges:

- **Data Collection:** Utilize real-world public datasets such as electricity usage records (e.g., UCI's Household Power Consumption or Kaggle's SCADA power data), renewable generation data (e.g., wind/solar output time series), and weather data (e.g., NOAA or open weather APIs). Integrate multiple sources (e.g., consumption, generation, weather) to form a comprehensive dataset.

- **Data Cleaning:** Address sensor anomalies (spikes or dropouts in meter readings) and missing data points. Convert units to standard measures (kilowatt-hours, degrees Celsius, etc.). Handle time-series irregularities (e.g., align timestamps across sources to a common interval).

- **Ingestion & Storage:** Develop an ingestion pipeline (using Python scripts or streaming tools) to load time-series data into a database optimized for time-series (e.g., TimescaleDB or cloud time-series storage). Apply optimization such as data partitioning (by date or region) to enhance query performance.

- **Transformation and Modeling:** Create derived features (e.g., hourly/daily energy consumption aggregates, peak-to-average ratios). Build forecasting models for energy demand or renewable supply, and clustering models to identify usage patterns among consumers.

- **Optimization & Insights:** Analyze consumption patterns to suggest optimizations (e.g., identify periods for demand response, plan maintenance). Use anomaly detection to flag outages or faults. Provide actionable

recommendations to reduce costs and emissions.

By engaging with public energy datasets, teams will experience real analytics tasks in the energy industry, such as forecasting demand and planning production around weather variability.

Phases and Weekly Deliverables

The project is structured into weekly phases:

- **Phase 1 (Weeks 1–2): Data Acquisition and Preprocessing**
    - *Week 1:* Identify and retrieve relevant data sources: smart meter logs, grid sensor data, and weather records. Understand metadata (sensor locations, units). **Deliverable:** Data inventory document and metadata summary.
    - *Week 2:* Clean and preprocess time-series data: detect and remove outliers (e.g., erroneous sensor spikes), impute missing timestamps, and align data to consistent time intervals. **Deliverable:** Cleaned datasets and a data preprocessing report.
- **Phase 2 (Weeks 3–4): Pipeline & Feature Engineering**
    - *Week 3:* Implement automated pipelines to load incoming data into storage (e.g., nightly batch scripts or streaming). Set up the database schema (e.g., time-series database) with efficient partitioning. **Deliverable:** Working ingestion pipeline and database schema overview.
    - *Week 4:* Engineer analytical features: aggregate consumption by hour/day, compute peak load statistics, and correlate energy use with weather. Index or partition the data to speed up queries. **Deliverable:** Transformed feature tables and documentation of the transformations.
- **Phase 3 (Weeks 5–6): Forecasting, Visualization, & Reporting**
    - *Week 5:* Apply analytics: build demand forecasting models (e.g., ARIMA, LSTM) for short-term energy usage. Perform anomaly detection (e.g., identify sudden drops indicating outages). Validate model predictions.

Deliverable: Modeling code and evaluation summary.

- *Week 6:* Visualize results in dashboards (e.g., consumption trends, forecast vs. actual). Prepare final recommendations for operational changes (e.g., load shifting strategies). **Deliverable:** Interactive visualization dashboard and final presentation.

Regular mentor check-ins will cover model validation and domain accuracy. Teams should collaborate, dividing tasks like pipeline engineering, data analysis, and dashboard creation.

Use of Generative AI

Teams should incorporate AI tools to augment their workflow:

- **ChatGPT (OpenAI):** Use ChatGPT to generate sample code for data normalization (e.g., resampling time-series data) and to explain statistical analysis methods. For example, ChatGPT can suggest ways to handle seasonality in consumption data.

- **GitHub Copilot:** Assist in writing code for ETL processes and data transformations, particularly loops over time intervals or writing SQL for time-series aggregation.

- **Gemini (Google):** Leverage Gemini's capabilities for domain knowledge (e.g., ask about best practices in energy forecasting) or for code suggestions using Google's libraries and tools.

- **Claude (Anthropic):** Apply Claude to summarize technical reports or to ensure clarity when writing up methodology (e.g., explaining anomaly detection results).

- **LangChain:** Use LangChain to build multi-step pipelines involving LLMs, such as generating a narrative summary of key findings from model outputs, or chaining prompts to iteratively refine forecasts.

Document AI usage, noting cases where tools like Copilot significantly sped up code development or where ChatGPT clarified a concept.

Expected Outcomes

By project completion, teams will:

- Successfully ingest, store, and manage large-scale time-series energy data.
- Clean and integrate multiple sources (grid sensors, weather data) to enable comprehensive analysis.
- Produce demand forecasts or identify consumption patterns that can guide utility decisions (e.g., optimize generation schedule).
- Create visualizations (peak usage charts, anomaly dashboards) to communicate findings to engineers and managers.
- Reflect on AI's role in the workflow, demonstrating how tools like ChatGPT or LangChain improved efficiency.

This project builds real-world data engineering skills relevant to smart grids and renewable integration.

Deliverables

- A repository (GitHub) with all code (data pipeline scripts, analysis notebooks), thoroughly documented.
- Cleaned energy datasets with clear descriptions of each feature.
- Phase reports documenting data preprocessing steps and intermediate findings.
- Forecast or anomaly detection model outputs with evaluation metrics.
- Dashboards or charts (using tools like Grafana, Tableau, or Python visualization libraries) showing energy usage patterns and predictions.
- Final presentation summarizing how data-driven insights could improve utility operations.

Proposal Submission Requirements

Submit a PDF proposal detailing:

- **Project Goals:** Objectives such as improving forecast accuracy or reducing peak load.
- **Data Plan:** List intended datasets (smart meter data, weather records) with sources and access method.
- **Technical Solution:** Tools (e.g., databases, Python, cloud services) and pipeline architecture.
- **Team Roles and Timeline:** Week-by-week plan (1–6) and team assignments.
- **Relevant Experience:** Background in data analysis, time-series modeling, or energy systems (if any).

Proposals should rely only on public data. Highlight any novel visualization or modeling ideas.

Evaluation Criteria

Submissions and projects will be evaluated on:

1. **Implementation Quality:** Robustness of data pipelines and storage solutions for time-series data.
2. **Analytical Insight:** Accuracy of forecasting models and value of identified insights (e.g., cost savings).
3. **AI Tool Use:** Effective leveraging of AI (e.g., Copilot for code, ChatGPT for methodology guidance).
4. **Innovation:** Creative feature engineering or optimization (e.g., novel way to detect demand spikes).
5. **Communication:** Clarity of documentation and visualizations, tailored to utility stakeholders.

Submission Deadline

The Project is due by **July 10, 2025 (Tentative)** via email.