**Alva Energies – Big Data Management Report Using Hadoop MapReduce**

What Alva Energies Does:

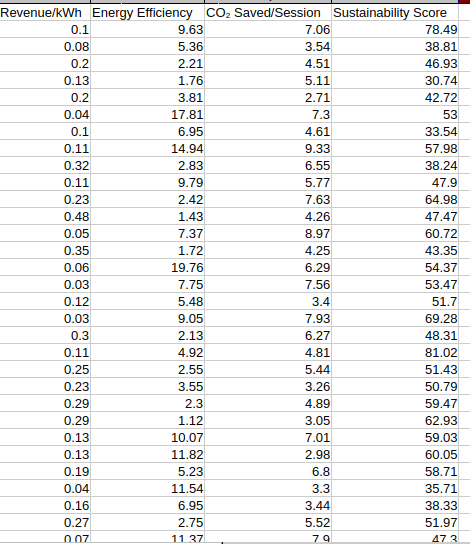
* Alva Energies provides EV charging services as a station as well as a portable module.
* The modules get it’s energy from Solar Energy and the Power Grid

Why are we analysing Energy Data?

* Analysing EV charging data helps us identify the current power needs of the business as well as the usage, and in this case, will also help us evaluate the sustainability of the business.

Why do we use Hadoop?

* MapReduce is designed to process massive datasets by distributing the workload across multiple nodes in a Hadoop cluster. This makes it ideal for big data scenarios, like energy usage or EV charging logs collected over time.
* The data used as of now is in a *CSV* format.
* The MapReduce function was made to do basic pre-processing on the data and then compute the following metrics:
  + Revenue per kWh - Revenue / Energy Delivered
  + Energy efficiency - Energy Delivered / Average Charging Time
  + CO₂ saved per session - CO₂ Saved / Charges per Day
  + Sustainability score - Weighted sum of renewable %, utilization, CO₂ saved per session



**Dataset Description**

The dataset used contains data that is extracted from the systems and is either generated by the systems or manual input from the employees:

| **Column Name** | **Description** |
| --- | --- |
| Date | Date of the charging session (daily record) |
| Charger Type | Type of charger used (e.g., Fast, Regular) |
| Charges per Day | Number of charging sessions per day |
| Energy Delivered (kWh) | Total energy delivered during sessions on that day (in kilowatt-hours) |
| Revenue per Session (₹) | Average revenue generated per charging session (in Indian Rupees) |
| Avg Time per Charge (min) | Average time spent per charging session (in minutes) |
| Charger Utilization Rate (%) | Percentage of time the charger was actively used |
| Energy Source | Source of the electricity used (e.g., Grid, Solar) |
| % Renewable Energy Used | Percentage of energy from renewable sources |
| CO₂ Saved (kg) | Estimated amount of CO₂ emissions avoided (in kilograms) |
| ICE Kilometers Avoided | Kilometers not driven by Internal Combustion Engine vehicles. |
| Fuel Avoided (liters) | Amount of fuel (in liters) not consumed due to EV usage |
| Carbon Offset Potential (tonnes) | Potential reduction in carbon emissions (in tonnes) |
| Equivalent Trees Planted | Estimated equivalent number of trees planted based on CO₂ savings |

Initial Observations:

* Some rows were **duplicates** or **incomplete**.
* The file contains special characters such as "₹" and "CO₂" which could cause **encoding issues** if not handled correctly.
* size: 391kb

Pre-processing Goals

* Removal of duplicate entries that might cause misjudgments
* Skip invalid or empty rows
* General formatting
* Derive new sustainability metrics from the initial base data

**MapReduce Process**

* **Mapper Role**:
  + Removes duplicates
  + Parses each line and extracts values
  + Calculates:
    - Revenue per kWh = Revenue / Energy Delivered
    - Energy Efficiency = Energy Delivered / Average Charging Time
    - CO₂ Saved per Session = CO₂ Saved / Charges per Day
    - Sustainability Score = Weighted sum of renewable %, utilization, CO₂ saved per session
* **Reducer Role**:
  + Since duplicates are already removed in Mapper, Reducer just passes data along
* **Output**:
  + A new cleaned dataset with 4 extra columns:
    - *Revenue per kWh*
    - *Energy Efficiency*
    - *CO₂ Saved per Session*
    - *Sustainability Score (0-100)*

**Hadoop Environment Setup**

* A dedicated project directory was created to manage the input CSV file, compiled Java classes, and output results. This helped organize the pre-processing workflow cleanly.
* The original CSV file (casestudydata.csv) was moved into an input folder created within the project directory. This made it easier to track files used in HDFS operations.
* The CSV file was uploaded to Hadoop's Distributed File System (HDFS) under a specified directory structure (e.g., /csvproject/input). This allowed the file to be distributed across the cluster for parallel processing.
* A custom MapReduce job was written in Java to clean the data, remove duplicates, and compute new sustainability-related metrics.
* After compilation and packaging into a JAR file, the job was executed by pointing it to the input directory in HDFS and specifying an output directory.
* Once the job was complete, the cleaned output was stored in HDFS. This output file was then downloaded back to the local machine. It now includes additional columns like Revenue per kWh, Energy Efficiency, CO₂ Saved, and Sustainability Score.

**Conclusion**

* This prototype demonstrated the use of Hadoop MapReduce for performing automated processing and metric derivation on the huge datasets generated by the business.
* These derived features provide **valuable insights** into the **performance, efficiency, and sustainability** of EV charging stations. They can support strategic decisions like optimizing charger usage, setting pricing models, or tracking progress toward sustainability goals.

Code Screen Shots:

