PROJECT TEMPLATE: FOR DATA SCIENCE AND DATA ENGINEERING

Introduction of the Task:

Based on the literature, the most used AI algorithm is ML algorithms in Microgrid System. There are also different techniques like MPC, FCS-MPC control techniques that have been found form literature. According to the literature, ANN was the most used method in MG problems.

Microgrids were first introduced in 2001 by Bob Lasseter. Microgrids should, in theory, be continuously linked to the utility grid, allowing any surplus energy from the microgrid to be sent to the primary grid and any energy shortfall in the microgrid to be met by the utility grid. The significant components of MG included DERs, power converters, energy storage, loads, master controller, intelligent switches, protective devices, communication, control, and automation systems. In the context of MG systems, three control architectures have been developed: the centralized, the decentralized and the distributed topology. A centralized architecture consists of a single controller which manages and communicates with all the other components. The main feature of the decentralized architecture is that the control system is composed of several individual controllers. AI-powered tools can help determine the production from DER, such as photovoltaic (PV) systems. The use of AI-powered tools can help to check the output data and find valuable insights to reduce operational costs and ensure stability.

Objective of the Task (Data Science/Data Engineering):

We have to conduct a study to evaluate the potential of machine learning (ML) tools using data from the TwInSolar Consortium Work Package: WP4 – A smart microgrid in a tropical island by the University of La Reunion. Several datasets were used to gain a deep understanding, including load profile data for different campuses such as IUT_load, ESIROI_SEASOI, and CROUS_load data.

The IUT_load data combines the data from ***Dpt. 1_2***, ***3_4,*** and ***Enerpos building***. Some buildings are already equipped with PV, including ***ESIROI_PV, ENERPOS_PV,*** and ***Dept 1_2_PV***

data. An ML approach must be employed to analyze their MG consolidated data to determine the potential load patterns and behavior of PV productions. The ***load (demand/requirement)*** and ***PV data*** should be utilized for forecasting based on historical ***data (only based on the data you have in Datetime and KW (production from PV),*** and PV production analysis based on ***meteo data***, which was collected from the weather station. It means you need to perform the analysis for the area where you have PV (Photovoltaik) already installed. You should make two analysis ***(one without meteo/weather data and one with weather data).***

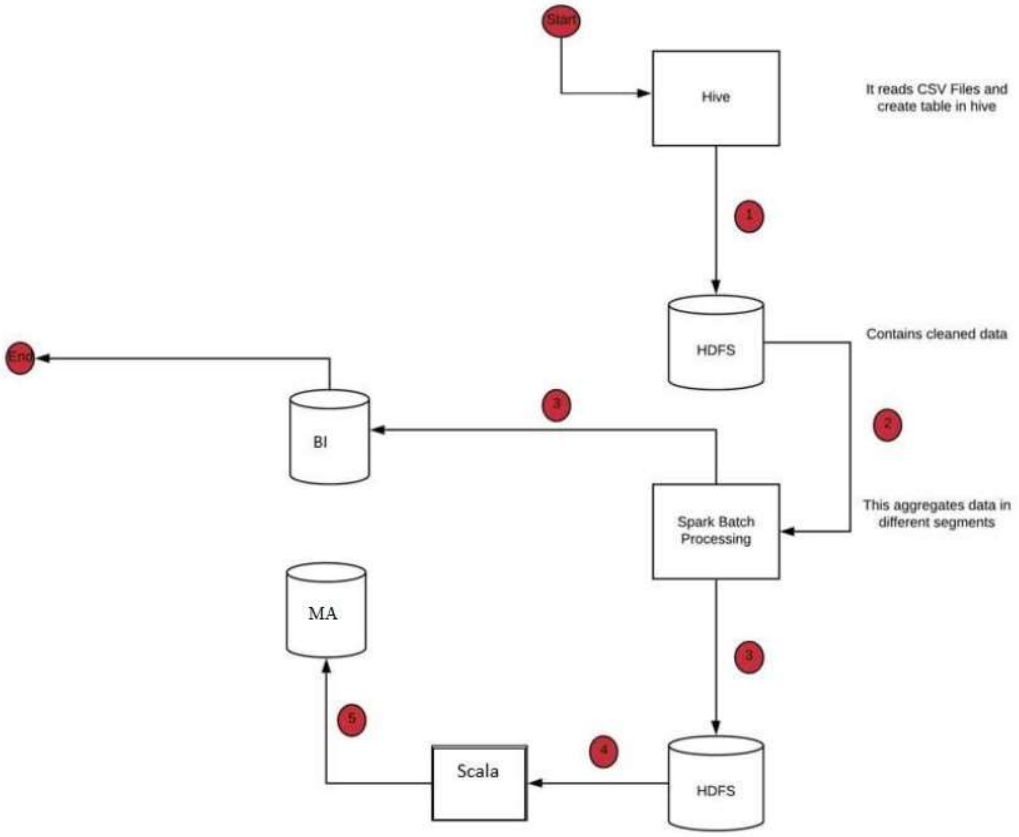
The use of ML models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) has been considered for forecasting approaches. GRU and LSTM are types of recurrent neural networks (RNN) that are preferred due to their better mechanism in handling long-term dependencies. The CROUS_load, ESIROI_SEASOI_Load, IUT_load, ESIROI_IUT2_HVAC, ESIROI_PV, ENERPOS_PV, and Dpt_1_2_PV data were trained for forecasting. (Note: Load data is mandatory for your task. Load data means the requirements of the buldings. ***Few buildings has PV installed already and that you should consider). You should not use GRU and RNN. You must use some model which is mostly used in prediction of weather solar forecasting, pattern of weather analysis, pattern of solar (PV) production analysis.***

- Make the evaluation matrix for your predictions.
- Use different models and add the data together for second task (meteo + datasets).
- Tell us why PV production is higher and lower and when.
- Explain you model accuracy and when most PV was produced.
- Also find our which weather parameters has highest influence on PV production.
- You can use the Load data for these 3 campuses where PV is installed and tell us the gap of energy (Load (KW) – PV production (KW).

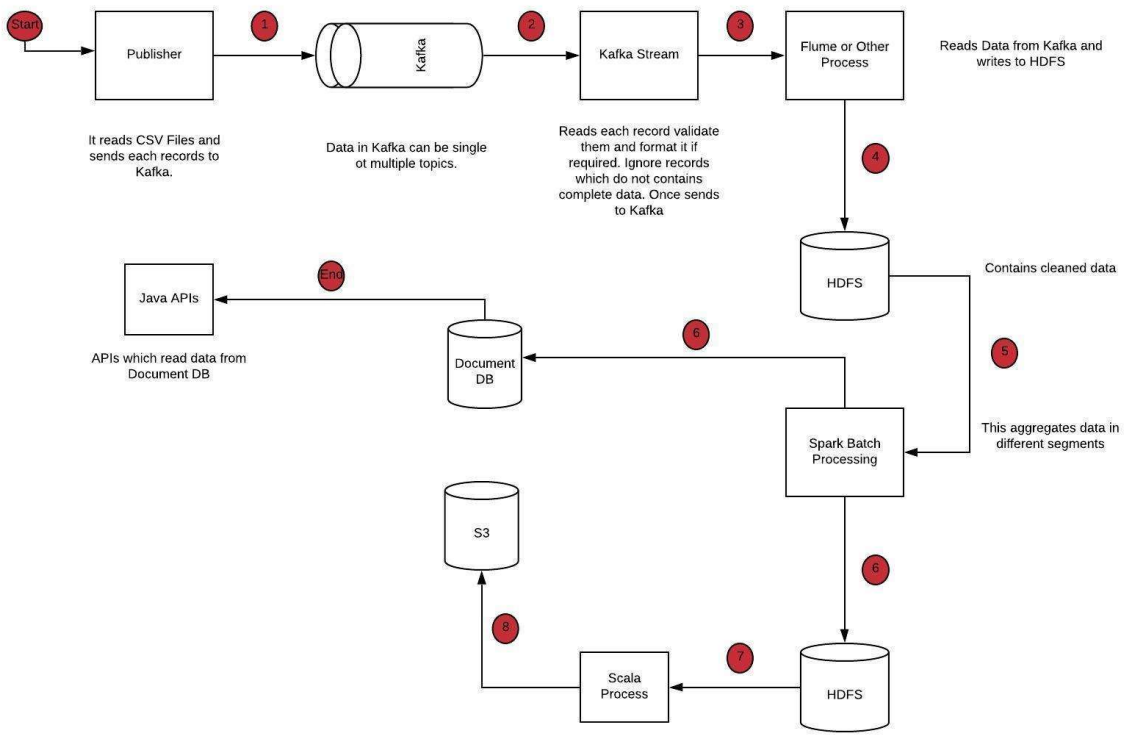
Important for Data Engineering Students:

- The data is same for your task but you have to perform the following analysis:
- You need to take 3 files for PV production (see above the Campuse where PV is installed)
- Make three tables based on the column has given (HIVE/SQL
- Make another table for the weather data (meteo data)
- Merge the tables together (Dept 1_2 PV + meteo), (ESIROI_PV+meteo), (ENERPOS_PV+meteo).
- Make the following visualization and analysis (You use use Scala/Python/PySpark)
- Use the appropriate services from Azure and make the Visualization. You can use also Databricks for the analysis.
- For streaming you can use Stream Job Analytics and for Visu BI.

- Calculate maximum, lowest production of PV, When PV production was lowest (in which month), Make the weather data (GHI mainly) visualization, make the Visualization of Gap of energy production.



If you want to explore more please use the following structure:



- If you use this structure please use Kafka for streaming, flume is not required, try to read the data directly from Kafka and write to HDFS.
- From HDFS use the data and do process of it for insights like (X axis Date time and Y axis must be production).
- Make the Weather insights analysis too **(see point 7)**
- Tell us also the Gap of the energy PV production and requirement which is Load (KW).
- Scala is not required but you can save the output of your analysis in any storage (Google storage/Blob etc.)

What you have to do:

1. Setup Spark 2 shell
2. Use Intellij to validate or modify source code
3. Click “mvn clean install” to build jar file
4. Create Kafka Producer providing Kafka configurations
5. Read CSV/Txt files to send data to Kafka
6. Run Kafka Producer based on the dataset you produced (it can be 4 as I mentioned above) with different file and topics

If you use Scala :

- Create sample Maven Scala Project
- Add necessary spark dependencies
- Create Schema of CSV/txt files
- Create Spark Session (if you use Scala)

  a) Add storage details

  b) Add all variables to your environment as they have sensitive data
- Read CSV file and convert into dataset

With/Without Scala:

7. Calculate maximum, lowest production of PV, When PV production was lowest (in which month), Make the weather data (GHI mainly) visualization, make the Visualization of Gap of energy production.
8. Save the output in storage.

Note: Please do not share this template with any external (outside from SRH)