Spark – Kafka

Exploration Task:

1. Understanding Kafka Basics (Kafka Architecture, Kafka Topics / Partitions and use Producer / Consumer Console to produce & consume messages) - help link
2. Spark Understanding (Spark Architecture, Basic Concepts, In Depth of the concepts being implemented) – help link

Note: Please prepare PPT according to your understanding on above

Implementation Task:

Background:

The electricity based on renewable energy sources perceived as an alternate source of energy and their penetration within the power system is rising at an extremely fast rate. Among new sources of renewable energy, wind energy has seen tremendous growth over recent years; in various countries, it is a true alternative to fossil fuels. Furthermore, wind power generation capacity varies constantly, stochastic, intermittent in nature and associated with generation of other ramp events and to understand this stochastic variation in wind power generation the right set of data plays a significant role.

Here is the attached 10m dataset for wind power generation of **Turkey** region

Task:

=> Kafka Producer

1. Read this CSV with headers using spark.
2. Publish these records into Kafka in streaming fashion.

=> Kafka Subscriber

1. Read the data from Kafka in streaming fashion using spark
2. Write this data received in delta format where schema of delta table will be
   [Hint: While reading the data from CSV, please note the formats and data types cautiously and then target to derive the following schema]

```
signal_date: DateType [column-type: Date/Time [format: yyyy-MM-dd]]
signal_ts: TimestampType [column-type: Date/Time [format: yyyy-MM-ddTHH:mm:ss]
create_date: DateType [column-type: Date/Time [current-date]]
creare_ts: TimestampType [column-type: Date/Time [current-time]]
signals: MapType<String, String> [with signal name as key and signal value as value of map, where signal columns are LV ActivePower (kW), Wind Speed (m/s), Theoretical_Power_Curve (KWh), Wind Direction (°)]
```

=> Analysis Task:

1. Read the data from delta lake using spark.
2. Calculate number of distinct `signal_ts` datapoints per day
   [Since the data is 10m it should be 6 datapoints per hour which means 144 datapoints per day]
3. Calculate Average value of all the signals per hour
4. Add a column in the dataframe of above named as `generation_indicator` and it will have value as
   a. if LV ActivePower<200KW then *Low*
   b. 200<= LV ActivePower < 600 then *Medium*
   c. 600< =LV ActivePower < 1000 then *High*
   d. 1000<=LV ActivePowe then *Exceptional*
5. Create a new dataframe with following json:

```
[
  {
    "sig_name": "LV ActivePower (kW)",
    "sig_mapping_name": "active_power_average"
  },
  {
    "sig_name": "Wind Speed (m/s)",
    "sig_mapping_name": "wind_speed_average"
  },
  {
    "sig_name": "Theoretical_Power_Curve (KWh)",
    "sig_mapping_name": "theo_power_curve_average"
  },
  {
    "sig_name": "Theoretical_Power_Curve (KWh)",
    "sig_mapping_name": "theo_power_curve_average"
  },
  {
    "sig_name": "Wind Direction (°)",
    "sig_mapping_name": "wind_direction_average"
  }
]
```

6. Change signal name in dataframe from step no 4 with mapping from step no 5 by performing broadcast join.