Expert Services

Hands On Advanced Analytics with Apache Spark

Proiect

Introducere

În această prezentare se regăsește descrierea proiectului pentru curs, incluzând informații despre setul de date precum și sarcinile specifice pentru fiecare secțiune a cursului.

- Seturile de date al proiectului se află pe drive, în directorul dedicat Data, subdirectorul „Project”. Fiecare director este dedicat pentru un set de date, iar pentru fiecare set de date se găsesc trei sub-directoare „json”, „csv” și „parquet” care conțin aceleași date, dar fișiere din ele în formatul cu numele directorului.

Sarcinile pot fi realizate oricând după prezentarea corespunzătoare a secțiunii. Titlurile paginilor care conțin sarcinile din această anexă vor include numele secțiunii respective.

Despre Proiect

Acest proiect are ca scop principal analiza consumului de energie al unui grup de consumatori fictivi, pe parcursul unui an, a unei companii de energie, folosind tehnicile de analiză din motorul Apache Spark.

- Primul set de date reflectă atât consumul total de energie, cât și detalii specifice, dacă există, despre producția din panouri solare, consumul pentru vehicule electrice (EV), energia furnizată înapoi către rețeaua electrică, consumul și încărcarea bateriilor.

- Al doilea set de date oferă atât tariful pe an și prețurile per kWh în diferite intervale de timp, unice pentru fiecare client în parte, cât și prețul de vânzare la nivelul companiei de energie pe diferite intervale de timp.

Proiectul implică curățarea și prelucrarea datelor, completarea valorilor lipsă, iar la final calculul facturii de energie, bonus fiind compararea facturii cu alți clienți similari.

Setul de Date – Raw Time Series

Acest set de date conține informații despre consumul de date a unor clienți fictivi ai unei companii de energie. Structura datelor este una similară cu schemele folosite la momentul actual pentru astfel de date.

Nume Coloană	Descriere
contract_id	Numărul de contract al clientului.
timestamp	Data și ora la care s-a efectuat măsurarea consumului.
value	Consumul de energie din ultimele 15 minute.
value_source	Tipul măsurătorii.
annotations	Alte date despre măsurătoare sau client in format JSON.

Setul de Date – Customer Tariff

Acest set de date conține informații despre tarifele și prețurile a unor clienți fictivi, într-un interval de timp, ai unei companii de energie. Structura este una similară cu cea folosite la momentul actual pentru astfel de date.

Nume Coloană	Descriere
contract_id	Numărul de contract al clientului.
target_local_start_timestamp	Data și ora de start când tariful devine activ, inclusiv.
target_local_end_timestamp	Data și ora de final când tariful încetează să mai fie activ, exclusiv.
tariff_name	Numele planului de tarifare.
charge_type	Indică dacă prețul este pentru cumpărare (buy) sau vânzare (sell) de energie. • „buy”: Această intrare este pentru consumul de energie de la rețea. • „sell”: Această intrare este pentru vânzarea de energie către rețea.
price	Costul sau venitul per kWh, în funcție de tipul de tarifare, buy sau sell.

Curățarea Datelor

Setul de date al consumului prezintă unele probleme, precum spații suplimentare și informații lipsă sau semi-structurate. Acestea trebuie rectificate înainte de a trece mai departe:

- Curățarea coloanei „contract_id” de spații suplimentare
- Setarea coloanei „value_source” cu valoarea „missing” atunci când valoarea lipsește.
- Coloana „timestamp” are variații mici ce trebuie rectificate. Contoarele măsoară consumul odată la 15 minute exact începând cu ora 00:00, dar din cauza procesării, pot apărea variații la timpul pe care îl trimit.
  - Hint: Se găsește cea mai apropriat multiplu de 15 la minute și se scot secundele

Extragerea Informațiilor de localizare și filtrarea datelor invalide

Setul de date al consumului prezintă unele probleme, precum spații suplimentare și informații lipsă sau semi-structurate. Acestea trebuie rectificate înainte de a trece mai departe:

- Extragerea unei noi coloane „region” din „annotations”
  - Se recomandă folosirea funcțiilor de Spark de procesare JSON
- Clienții cu regiuni invalide se vor scoate din setul de date și se vor salva pe disk într-o locație separată.
- Extragerea datei din coloana „timestamp” într-o nouă coloană „utc_date”
- Calcularea datei locale pentru data și ora din „timestamp”, pe baza regiunii, într-o nouă coloană „local_timestamp”
- Extragerea datei din coloana „ local_timestamp” într-o nouă coloană „local_date”

Extragerea Informațiilor de consum

Setul de date al consumului prezintă unele probleme, precum spații suplimentare și informații lipsă sau semi-structurate. Acestea trebuie rectificate înainte de a trece mai departe:

- Extragerea din coloana „ annotations” a consumului de vehicul electric (EV), baterie (BATTERY_IN) și consumul trimis spre rețeaua electrică (GRID_SELL) în coloanele „sent_to_ev”, „sent_to_battery” și „sent_to_grid”. În cazul în care valoare lipsește, se consideră consumul 0.

- Extragerea din coloana „ annotations” a energiei primite de la panourile solare (PV) și baterie (BATTERY_OUT) în coloanele „received_from_pv” și „received_from_battery”. În cazul în care valoare lipsește, se consideră energia primită 0.

Filtrarea consumului neobișnuit

Anumite valori ale consumului sunt neobișnuit de mari și este necesară scoaterea lor:

- Setarea coloanei „value_source” în „plausability_check_failed” pentru valorile cu consum neobișnuit din setul de date.
  - Decizia valorilor mari fie se face cu o analiză vizuală a datelor (sortarea și identificarea lor vizual fie prin agregări simple fie mai complicate) sau bonus, pentru cine dorește, prin tehnici de învățare automată.

- Salvarea separată a datelor cu „value_source” având valoare „plausability_check_failed” într-o locație separată. Atenție, datele nu se scot din setul de date.

- Setarea coloanei „value” în NULL pentru datele cu „value_source” având valoare „plausability_check_failed”.

Completarea valorilor lipsă

Anumiți clienți au lipsuri prezintă câteva lipsuri în consum, unele dintre ele adăugate de noi la pasul precedent:

- Prezicerea valorii „value” atunci când ea este NULL folosind următoarea metodă:
  - Calculăm mai întâi media din ultimele 8 săptămâni a valorilor din aceeași zi a săptămânii la aceeași oră, minut și secundă. Se folosesc din ultimele 8 săptămâni doar valorile care nu lipsesc, au coloana „value_source” setată pe valoarea „measurement”.
  - Facem suma coloanelor „sent_to_ev”, „sent_to_battery” și „sent_to_grid”.
  - Facem suma coloanelor „received_from_pv” și „received_from_battery”.
  - Completăm coloane „value” cu maximum dintre aceste 3 valori.
  - Bonus, pentru cine dorește, puteți folosi și algoritmi de învățare automată, precum Linear Regression, în loc de calcularea mediei din ultimele 8 săptămâni și să comparați cele 2 metode.

- Calcularea valorii „received_from_grid” atunci când avem toate informațiile.
  - Facem suma coloanelor „sent_to_ev”, „sent_to_battery” și „sent_to_grid” și scădem valorile din coloanele „received_from_pv” și „received_from_battery”.

Asocierea cu tarifere

Curățând datele și completând valorile lipsă, putem acuma să trecem la asocierea cu tarifele consumatorilor:

- Asocierea intrărilor de consum cu cele de tarifare, pe bază numărului de contract, al timpului și al tipului de preț, cumpărare sau vânzare.
  - Prețul tarifat la cumpărare este prețul de cumpărare pentru intrare de consum înmulțit cu coloana „received_from_grid”
  - Prețul tarifat la vânzare este prețul de vânzare pentru intrare de consum înmulțit cu coloana „send_to_grid”.

Calcularea consumului și a facturii

După asociere, putem trece la calculul factorii:

- Per zi / săptămâna / lună / an, pentru fiecare client, să se calculeze:
  - Consumul de energie total (suma coloanei value) – „ kWh_total”
  - Energia extrasă din baterie și cea din PV – „kWh_from_battery”, „kWh_from_PV”
  - Energia folosită pentru EV – „kWh_for_EV”
  - Consumul de energie folosit de la rețea – „kWh_from_grid”
  - Costul consumului de energie folosit de la rețea – „price_billed”
  - Consumul de energie trimis către rețea – „kWh_to_grid”
  - Costul primit înapoi, a consumului de energie trimis către rețea – „price_cashback”
  - Costul total (diferența de cost) – „price_final”
  - Puneți codul de calculare al agregatelor pe orice perioadă de timp într-o funcție, iar pentru fiecare interval, zi / săptămâna / lună / an, rulați această funcție. Afișați pentru fiecare perioada top 10 clienți cu cel mai mare preț.