Projekt końcowy “minimum”

Projekt pozwalający na otrzymanie pozytywnej oceny.

Projekt wykonany w oparciu o poniższy opis z uwzględnieniem wszystkich wskazanych punktów odpowiada minimalnym wymaganiom wobec pracy końcowej. W związku z tym jego poprawna realizacja pozwala na otrzymanie oceny dostateczneji i ukończenie studiów.

Wykonany projekt należy opisać i przedstawić w formie pracy pisemnej zgodnie z wytycznymi wskazanymi na zajęciach “Projektowanie rozwiązań Big Data”.

Technologie

- Linux
- Python
- Matplotlib
- Pyspark
- Parquet
- Cloud9
- S3
- EMR

Krótki opis

Projekt polega na utworzeniu generycznego potoku, w ramach którego dane z jednego (dowolnego) forum ze StackExchange są pobierane, przetwarzane i analizowane. Całość wykonana jest w chmurze AWS.

Kroki

Cloud9

1. pobranie danych
2. rozpakowanie danych
3. utworzenie bucketu na S3
4. przeniesienie danych do utworzonego bucketu

EMR Notebook*

1. wczytanie danych do DataFrame’u
2. oczyszczenie danych (zadbanie o odpowiednie typy i nazwy kolumn, oczyszczenie kolumn z tekstem)

3. zapis danych w formacie Parquet na S3
4. analiza danych i wizualizacja; pytania analityczne:
    a. liczba postów na przestrzeni czasu (lineplot/barplot)
    b. czas na forum (od pojawienia się użytkownika do ostatniego posta/komentarza) 10 najdłużej aktywnych użytkowników (pomijając boty) (barplot)
    c. porównanie najwyżej i najniżej ocenianych pytań (długość, tagi, liczba odpowiedzi)
    d. procent przypadków kiedy najwyżej oceniana odpowiedź to nie zaakceptowana odpowiedź
    e. rozkład ocen odpowiedzi zaakceptowanych vs pozostałych (średnia, odchylenie, minimum, maksimum)
    f. top N tagów które wygenerowały najwięcej wyświetleń
    g. liczba postów w czasie dla każdego z top N tagów (lineplot/barplot)
    h. najczęściej pojawiające się słowa w tytułach (z pominięciem stopwords)
    i. procent użytkowników którzy nigdy nic nie zapostowali
    j. średni czas od pojawienia się pytania do pojawienia się zaakceptowanej odpowiedzi

* wszystkie operacje na danych powinny odbywać się przy pomocy Sparka (DataFrame), jedynie na potrzeby samej wizualizacji można przejść do kolekcji/struktur lokalnych

Wskazówki

- należy przetworzyć wszystkie pliki z pojedynczego archiwum
- wszystkie operacje w ramach Cloud9 można wykonać przy pomocy dedykowanych narzędzi linuxowych
- wczytanie i wstępne przetworzenie danych ułatwi skorzystanie z modułu spark-xml
- EMR Notebook pozwala na załadowanie dodatkowych modułów sparkowych oraz doinstalowanie bibliotek pythonowych - patrz linki poniżej
- wstępne przetwarzanie i analizy można rozbić na dwa osobne notebooki co może ułatwić zarządzanie środowiskiem
- cały projekt można wykonać w ramach AWS Academy Learner Lab

Przydatne linki

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-studio-magics.html
- https://aws.amazon.com/blogs/big-data/install-python-libraries-on-a-running-cluster-with-emr-notebooks/
- https://github.com/databricks/spark-xml