



WSTĘP DO DATA ENGINEERINGU

IGOR RĘBAS

Spis treści

Wstęp	4
1.1 Do kogo skierowany jest niniejszy kurs- wymagania wstępne	4
1.2 O kursie	4
1.3 Czym właściwie jest Data Engineering	5
2. Google Cloud Platform – informacje ogólne	6
2.1 Czym jest GCP – wady i zalety	6
2.2 Konfiguracja projektu	7
2.3 Billing i koszty GCP	9
2.4 Zarządzanie dostępem (IAM)	10
2.5 Cloud logging i monitoring	11
3. Google BigQuery	12
3.1 Czym jest BigQuery	12
3.2 DataSets w BigQuery	13
3.3 Pisanie zapytań w BigQuery	14
3.4 Tworzenie tabel w BigQuery	15
3.4.1 Tworzenie tabel w BigQuery poprzez ustawienia zapytania	15
3.4.2 Tworzenie tabel w BigQuery z Google Sheet'a	16
3.4.3 Tworzenie tabel z plików zlokalizowanych w GCS	18
3.5 Scheduling zapytań	20
3.6 Data Transfer	22
4. SQL	24
4.1 Struktura zapytania	24
4.2 Funkcje skalarne w Standard SQL	26
4.3 Instrukcje warunkowe	29
4.4 Podzapytania i CTE	29
Zadania 1-16 (wszystkie zadania odnoszą się do datasetu otomoto)	31
4.5 Złączenia tabel	33
4.6 Funkcje okienkowe	36
4.7 Funkcja PIVOT	41
Zadania 27-29	43
4.8 Arrays	43
4.9 Struct	46
4.10 Praca z obiektami typu JSON	50

Zadania 35-37	52
4.11 Optymalizacja zapytań w BigQuery.....	53
5. Python	55
5.1 Konfiguracja środowiska pracy	55
5.2 Podstawy składni języka Python.....	57
5.2.1 Komunikacja z użytkownikiem	57
5.2.2 Zmienne i typy danych	58
5.2.3 Funkcje	60
5.2.4 Operacje na tekstach	63
5.2.5 Operacje na datach – biblioteka datetime	65
Zadania 1-11	68
5.3 Instrukcje warunkowe	69
5.3.1 Instrukcja if, elif, else	70
5.3.2 Instrukcja 'match'	71
5.4 Struktury danych	73
5.4.1 Listy.....	73
5.4.2 Krotki (tuples).....	76
5.4.3 Słowniki	78
5.4.4 Sety	81
5.4.5 Porównanie struktur danych	83
Zadania 18-23	84
5.5 Pętle	84
5.5.1 Pętla for	85
Zadania 23-31	88
5.6 Biblioteka pandas	89
5.6.1 Wprowadzenie do biblioteki Pandas.....	89
5.6.2 Podstawowe struktury danych w Pandas. Tworzenie DataFrame'a.	89
5.6.3 Dostęp i manipulacja danymi w DataFrame.....	91
5.6.4 Agregacja i grupowanie danych	94
5.6.5 Łączenie DataFrame'ów	96
5.6.6 Wykorzystanie niestandardowych funkcji w Pandas	98
5.6.6.1 Funkcje lambda	98
5.6.6.2 Wykorzystanie funkcji apply, map do przekształcania danych	99
Zadania 32-42	102
5.7 Praca z plikami typu JSON	103
Zadania 43-45	105
5.8 Praca z API – biblioteka requests	106
5.8.1 Czym jest API	106
5.8.2 Biblioteka requests.....	107
Zadania 46-48	109
6. Google Cloud Shell 6.1 Wprowadzenie do Google Cloud Shell.....	110
6.2 Konfiguracja gcloud	111

6.3 Konfiguracja środowiska wirtualnego	112
6.4 Praca z bibliotekami google cloud	113
Zadania 1-6	119
6.5 Integracja z GitHubem.....	120
6.5.1 Czym jest GitHub.....	120
5.6.2 Klonowanie repozytorium i dodawanie plików.....	121
5.6.3 Tworzenie nowej gałęzi	123
7. Cloud Functions	127
7.1 Czym są Cloud Functions	127
7.2 Tworzenie Cloud Functions triggerowanych http	128
7.3 Wywoływanie funkcji za pomocą Cloud Schedulera	132
7.3.1 Czym jest Cloud Scheduler i CRON	132
7.3.2 Tworzenie zaschedulowanego zapytania	133
7.4 Tworzenie Cloud Functions triggerowanych operacjami na GCS	134

Wstęp

1.1 Do kogo skierowany jest niniejszy kurs- wymagania wstępne

Kurs jest skierowany do wszystkich, którzy myślą o potencjalnym rozwoju w kierunku inżynierii danych i nie tylko. Jeżeli nawet nie jesteś przekonany, czy jest to właściwa droga Twojego rozwoju, to nabyta w kursie wiedza (czy to z zakresu Pythona czy SQLa) na pewno przyda Ci się w życiu zawodowym.

Minimalne wymagania przed przystąpieniem do kursu to średnio zaawansowana znajomość **SQLa** oraz podstawowa znajomość **Pythona**. Jeżeli nie posiadasz wiedzy na takim poziomie – nie ma żadnego problemu – będziesz mógł ją nadrobić w trakcie kursu, gdyż zawiera on obszernie powtórzenie materiału z zakresu **SQLa** ze szkolenia „Od zera do SQL Codera” oraz kompleksowe rozdziały dotyczące podstaw programowania w **Pythonie**.

1.2 O kursie

Kurs przygotowany jest w formie samouczka, podobnie jak poprzednie szkolenie „**Od zera do SQL Codera**”. Podzielony jest on na **trzy główne części** – pierwsza część dotyczy zagadnień związanych z **BigQuery** oraz **SQL**, druga część poświęcona jest w całości **Pythonowi**, zaś trzecia część skupia się na najczęściej wykorzystywanych **narzędziach w inżynierii danych** i wykorzystaniem uprzednio zdobytej wiedzy z Pythona i SQLa. Po pierwszej i drugiej części przewidziany jest mikro test, zaś po trzeciej części samodzielny projekt do zrobienia. W trakcie kursu do wykonania będą zadania o różnym poziomie trudności:

- * - zadania proste (31 zadań)
- ** - zadania średnie (38 zadań)
- *** - zadania trudne (26 zadań)

Na końcu kursu znajdują się rozwiązania do zadań związanych z pisaniem kodu. Podczas kursu będziemy korzystać z chmury google’a – **Google Cloud Platform (GCP)** oraz z Pythona w wersji 3.11. Sugerowane IDE do części związanej z Pythonem to **Pycharm Enterprise**. Procentowo zagadnienia na kursie przedstawiają się mniej więcej następująco:

- **SQL** – powtórzenie – 15%
- **BigQuery** i zaawansowany SQL – 15%
- **Python** – podstawy – 15%
- **Python** – biblioteki wykorzystywane w Data Engineeringu – 25%
- **Narzędzia GCP wykorzystywane w Data Engineeringu** – 30%

1.3 Czym właściwie jest Data Engineering

Data Engineer, określany również jako Data Architect bądź Data Infrastructure Specialist, to jeden z niedawno wyodrębnionych zawodów z dziedziny analityki danych. Inżynier danych to osoba, której zadaniem jest zbieranie i przetwarzanie surowych danych, ocena przydatności nowych źródeł informacji oraz projektowanie i uruchamianie nowych relacyjnych baz danych, pozwalających na przechowywanie oraz przetwarzanie informacji napływających do systemu. W praktyce oznacza to zbieranie, przeniesienie, przechowywanie, przygotowywanie do obróbki oraz wstępne analizowanie dużych zbiorów danych.

Praca Data Engineera nieodzownie związana jest z pracą w chmurze. Najpopularniejsze obecnie chmury to AWS, Azure oraz Google Cloud Platform – dlatego ważne jest, aby dobrze poznać przynajmniej jeden z wyżej wymienionych systemów oraz jego funkcjonalności. Niezbędnym elementem pracy jest też bardzo dobra znajomość języka SQL oraz przynajmniej dobra znajomość jednego z języków programowania – najpopularniejsze języki wykorzystywane w codziennej pracy to Python oraz Java.

2. Google Cloud Platform – informacje ogólne

2.1 Czym jest GCP – wady i zalety

Google Cloud Platform (GCP) to usługa chmurowa oferowana przez Google, umożliwiająca firmom i programistom korzystanie z infrastruktury chmurowej do przechowywania danych, uruchamiania aplikacji i wykonywania obliczeń. Jest to kompleksowy zestaw narzędzi i usług zaprojektowanych w celu wspierania różnorodnych zastosowań, od małych startupów po duże przedsiębiorstwa.

ZALETY GOOGLE CLOUD PLATFORM:

- **Elastyczność i skalowalność:**
GCP umożliwia elastyczne dostosowywanie zasobów do zmieniających się potrzeb, co pozwala na płynne skalowanie aplikacji w górę lub w dół w zależności od obciążenia.
- **Wysoka dostępność i niezawodność:**
Google dysponuje rozbudowaną infrastrukturą globalną, co zapewnia wysoką dostępność usług chmurowych i minimalizuje ryzyko awarii.
- **Zaawansowane narzędzia i technologie:**
GCP oferuje szeroki zakres narzędzi do analizy danych, uczenia maszynowego, sztucznej inteligencji i innych zaawansowanych technologii, co umożliwia tworzenie innowacyjnych rozwiązań.
- **Bezpieczeństwo:**
Google Cloud Platform zapewnia zaawansowane mechanizmy zabezpieczeń, takie jak szyfrowanie danych w ruchu i w spoczynku, kontrole dostępu oraz monitorowanie bezpieczeństwa.

WADY GOOGLE CLOUD PLATFORM:

- **Koszty:**
Korzystanie z usług chmurowych może generować duże koszty, zwłaszcza przy intensywnym użyciu zasobów. Niektóre funkcje, takie jak przesyłanie danych między regionami, mogą być dodatkowo płatne.
- **Skomplikowana oferta:**
Dla początkujących użytkowników może być trudne zrozumienie i wybór odpowiednich usług spośród bogatej oferty Google Cloud Platform.