234073	234102	234106
Numer indeksu	*Numer indeksu*	*Numer indeksu*
Aleksandra Kowalczyk	Zbigniew Nowacki	Karol Podlewski
Imię i nazwisko	*Imię i nazwisko*	*Imię i nazwisko*

Kierunek Informatyka Stosowana
Stopień II
Specjalizacja Data Science
Semestr 2

Data oddania 21 października 2020

Przetwarzanie i analiza dużych zbiorów danych

Zadanie 1

Spis treści

1 Cel zadania

Celem zadania było porównanie czasów analizy danych dotyczących zgłoszeń do władz Nowego Jorku za pośrednictwem numeru 311. Należało znaleźć najczęściej zgłaszane skargi (ogólnie, a także dla każdej dzielnicy) oraz urzędy, do których najczęściej zgłaszano te skargi - w przedstawionym rozwiązaniu za każdym razem szukaliśmy 10 najczęściej występujących wartości.

Zadanie należało wykonać za pomocą skryptu napisanego w języku python - wyszukując odpowiednie treści bezpośrednio w pliku csv oraz przez zapytania do bazy danych. Następnie należało przeprowadzić działania mające na celu redukcję czasu wykonywania kwerend.

2 Opis implementacji

W celu realizacji zadania wykorzystano bazy danych MS SQL oraz MySQL. Przy skryptach niezbędne okazały się biblioteki mysql, pandas, pyodbc oraz timeit.

Z pliku .csv wczytywano jedynie kolumny, których treść była związana z zadaniem: *Agency Name*, *Complaint Type* oraz *Borough*. W przypadku bazy danych MS SQL oznaczało to stworzenie pomocniczego pliku .csv z wybranymi kolumnami – jego tworzenie zostało zawarte w czasie wczytywania danych.

W celu optymalizacji czasu zapytań dla każdej z baz zostało zastosowane indeksowanie na wszystkich kolumnach.

3 Uzyskane wyniki

Pierwszym krokiem było zmierzenie czasu wczytywania danych. Działania mające na celu optymalizację czasu zapytań nie dotyczyły bezpośrednio wczytywania danych, dlatego nie porównywano czasów wczytywania dla bazy danych przed oraz po dodaniu indeksów. Wyniki przedstawiono w tabeli 1.

Tabela 1: Porównanie czasów wczytywania bazy danych

	Próba 1	Próba 2	Próba 3	Próba 4	Próba 5	**Średnia**
Skrypt	129,67	128,44	130,00	127,78	128,94	**128,97**
MS SQL	498,44	351,10	433,34	362,29	359,55	**400,94**
MySQL	2625,13	2664,90	2515,30	2870,82	2811,83	**2697,60**

Następnie zmierzono czasy indeksowania dla baz danych. Wyniki zostały przedstawione w tabeli 2.

Kolejnym krokiem było sprawdzenie jak dużo czasu zajmowały konkretne zapytania. Pierwsze zapytanie poszukiwało najczęściej zgłaszanych skarg. Były to odpowiednio: *Noise - Residential*, *Heat/Hot Water*, *Illegal Parking*, *Street Condition*, *Blocked Driveway*, *Street Light Condition*, *Heating*, *Plumbing*, *Water System* oraz *Noise - Street/Sidewalk*. W tabeli 3 przedstawiono

Tabela 2: Porównanie czasów indeksowania tabel

	Próba 1	Próba 2	Próba 3	Próba 4	Próba 5	**Średnia**
MS SQL + Indeks	809,48	708,06	742,05	710,20	738,95	**741,75**
MySQL + Indeks	1093,88	1117,06	1131,49	1068,48	1072,74	**1077,55**

czasy zapytań.

Tabela 3: Porównanie czasów zapytania poszukującego najczęściej zgłaszane skargi

	Próba 1	Próba 2	Próba 3	Próba 4	Próba 5	**Średnia**
Skrypt	3,45	3,32	3,37	3,51	3,44	**3,42**
MS SQL	38,94	40,76	40,79	39,08	40,55	**40,02**
MS SQL + Indeks	21,52	19,70	20,66	20,07	20,80	**20,55**
MySQL	75,89	77,84	76,17	75,50	77,78	**76,64**
MySQL + Indeks	17,78	18,04	19,70	19,72	16,77	**18,40**

Drugie zapytanie dotyczyło najczęściej zgłaszanych skarg dla każdej dzielnicy Nowego Jorku. Uzyskano następujące wyniki:

- Bronx - *Noise - Residential, Heat/Hot Water, Street Light Condition, Heating, PLUMBING, Blocked Driveway, Noise - Street/Sidewalk, Unsanitary Condition, Water System, Illegal Parking.*
- Brooklyn - *Noise - Residential, Heat/Hot Water, Illegal Parking, Blocked Driveway, Street Condition, Street Light Condition, PLUMBING, General Construction/Plumbing, Heating, Water System.*
- Manhattan - *Noise - Residential, Heat/Hot Water, Noise - Street/Sidewalk, Noise, Street Condition, Illegal Parking, Noise - Commercial, Heating, Street Light Condition, Water System.*
- Queens - *Noise - Residential, Blocked Driveway, Illegal Parking, Street Condition, Street Light Condition, Water System, Heat/Hot Water, Damaged Tree, Sewer, General Construction/Plumbing.*
- Staten Island - *Street Condition, Street Light Condition, Noise - Residential, Illegal Parking, Water System, Missed Collection (All Materials), Sewer, Damaged Tree, Dirty Conditions, Blocked Driveway.*
- Brak przypisanej dzielnicy - *Heating, General Construction, PLUMBING, Benefit Card Replacement, Paint - Plaster, Nonconst, DOF Parking - Payment Issue, HPD Literature Request, Electric, DCA / DOH New License Application Request.*

W tabeli 4 przedstawiono czasy zapytań.

Tabela 4: Porównanie czasów zapytania poszukującego najczęściej zgłaszaną skargę dla każdej dzielnicy

	Próba 1	Próba 2	Próba 3	Próba 4	Próba 5	**Średnia**
Skrypt	104,49	90,25	101,8	95,5	102,22	**98,85**
MS SQL	38,89	40,72	40,24	38,87	40,18	**39,78**
MS SQL + Indeks	21,51	22,49	21,20	21,09	21,25	**21,51**
MySQL	86,15	84,42	84,87	86,73	83,98	**85,23**
MySQL + Indeks	49,47	47,65	49,35	48,05	47,18	**48,34**

Ostatnie zapytanie miało na celu znaleźć urzędy do których najczęściej zgłaszano skargi, były to odpowiednio: *New York City Police Department*, *Department of Housing Preservation and Development*, *Department of Transportation*, *Department of Environmental Protection*, *Department of Buildings*, *Department of Parks and Recreation*, *Department of Health and Mental Hygiene*, *Department of Sanitation*, *Taxi and Limousine Commission* oraz *Department of Consumer Affairs*. W tabeli 5 widoczne są uzyskane rezultaty pomiarów czasy wykonania kwerend.

Tabela 5: Porównanie czasów zapytania poszukującego urzędu, do którego najczęściej zgłaszano skargi

	Próba 1	Próba 2	Próba 3	Próba 4	Próba 5	**Średnia**
Skrypt	3,13	3,16	3,09	3,12	3,10	**3,12**
MS SQL	42,61	39,92	38,84	38,49	36,57	**39,29**
MS SQL + Indeks	21,61	21,88	22,28	22,73	22,01	**22,10**
MySQL	76,63	78,82	79,23	79,49	79,40	**78,71**
MySQL + Indeks	23,66	25,65	23,16	24,32	23,97	**24,15**

W 6 podsumowano uzyskane wyniki poprzez zsumowanie średnich czasów uzyskanych przez każdy wariant na wszystkich etapach.

Tabela 6: Porównanie i zsumowanie czasów średnich dla każdego z wariantów

	Wczytywanie	Indeks.	Zap. 1	Zap. 2	Zap. 3	**Suma**
Skrypt	128,97	-	3,42	98,85	3,12	**234,36**
MS SQL	400,94	-	40,02	39,78	39,29	**520,03**
MS SQL + Indeks	400,94	741,75	20,55	21,52	22,10	**1208,15**
MySQL	2697,6	-	76,64	85,23	78,71	**2938,18**
MySQL + Indeks	2697,6	1077,55	18,40	48,34	24,15	**3866,04**

4 Analiza

W przypadku gdy porównamy czasy wczytywania, pythonowy skrypt okaże się najszybszy, natomiast wykorzystanie bazy danych MySQL zajmie najwięcej czasu. W dodatku skrypt w języku Python był najszybszy dla 2 z 3 zapytań - nie poradził sobie tak dobrze jak bazy danych, kiedy do zapytania trzeba było wprowadzić dodatkowe kryteria - było to spowodowane tworzeniem dodatkowych obiektów.

Baza danych MS SQL była szybsza w każdym sprawdzanym aspekcie od MySQL: wczytywania, indeksowania oraz zapytań.

Indeksowanie znacznie przyśpiesza proces analizy danych, należy jednak mieć na względzie to, że także zajmuje sporo czasu, który nie zostanie efektywnie spożytkowany w przypadku ciągle aktualizowanej bazy danych.