Wieloskalowa analiza danych

z forum internetowego przy użyciu zasobów infrastruktury chmurowej AWS oraz technologii Spark

Michał Kamiński

Cel pracy

- Szybkie i powtarzalne dostarczenie infrastruktury do wykonywania analiz na klastrze obliczeniowym `AWS EMR`
- Demonstracja użycia ww. infrastruktury przy użyciu technologii `Spark`

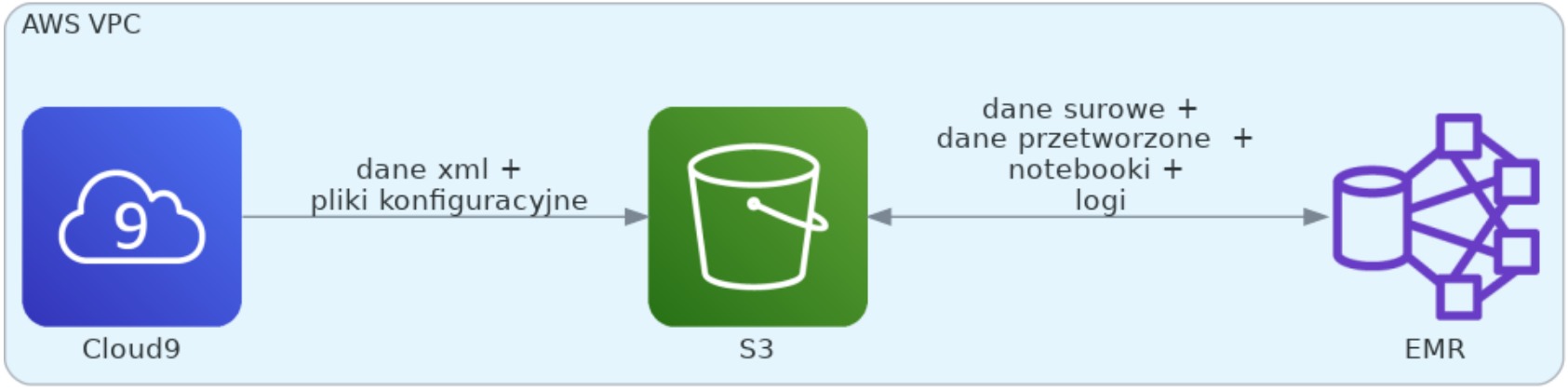
Infrastruktura i dane

Dane

- Forum internetowe *Beer, Wine and Spirits* [1]
- 6 zbiorów danych w formacie `xml`, połączonych kluczami
- Całkowity rozmiar ~20MB
- Dane z ~8,5 lat, 3700 postów

[1] Źródło - https://stackexchange.com/

Schemat infrastruktury



- `Cloud9` jako środowisko konfiguracyjne
- 5 koszyków danych S3 (dane xml, dane wstępne, repozytorium notebooków, repozytorium logów, pliki konfiguracyjne)
- Klaster `EMR` (1 MASTER, 2 CORE)

Wstępne przetwarzanie (pySpark)

1. Wczytanie danych xml - `spark-xml_2.12:0.15.0`
2. Utworzenie schematów danych
3. Czyszczenie danych tekstowych (tagi html, znaki specjalne) [2]
4. Zapis w formacie `parquet`

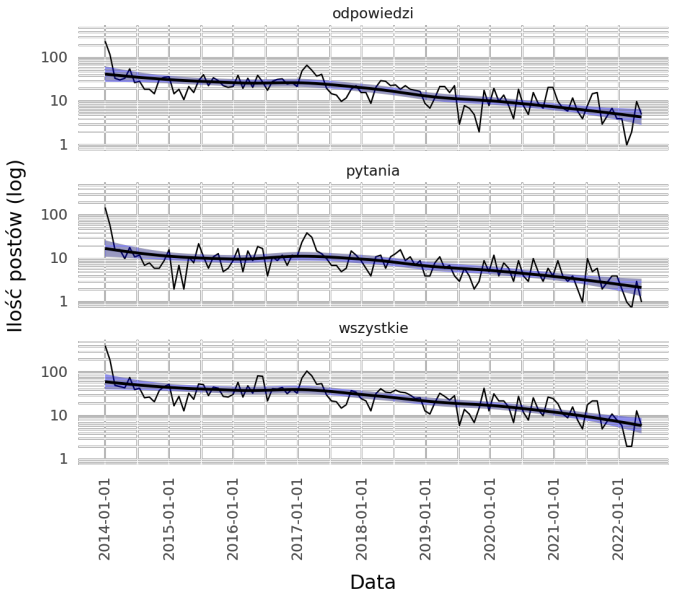
Adnotacja

50% redukcja rozmiaru danych (`xml` vs `parquet`)

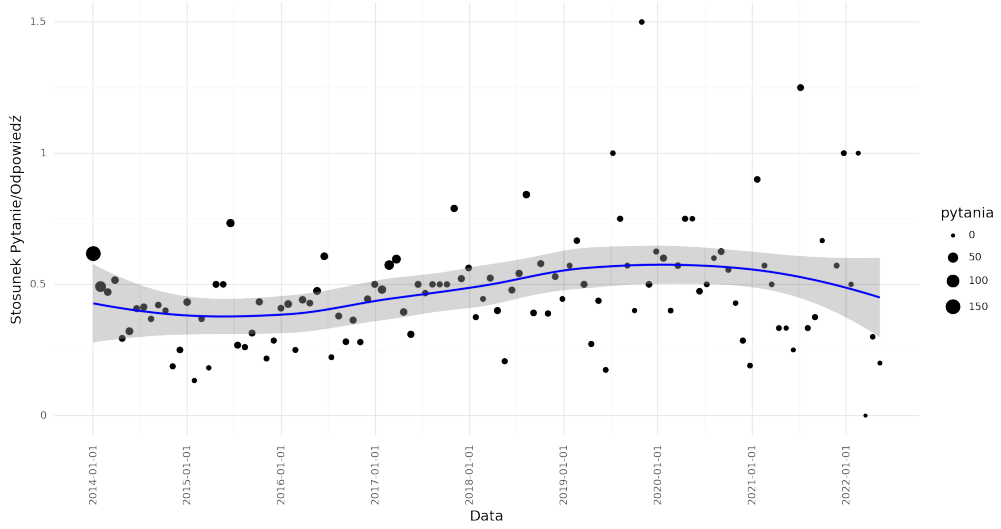
[2] np. <p>, \n, \t, \s

Analiza (pySpark)

Aktywność na forum



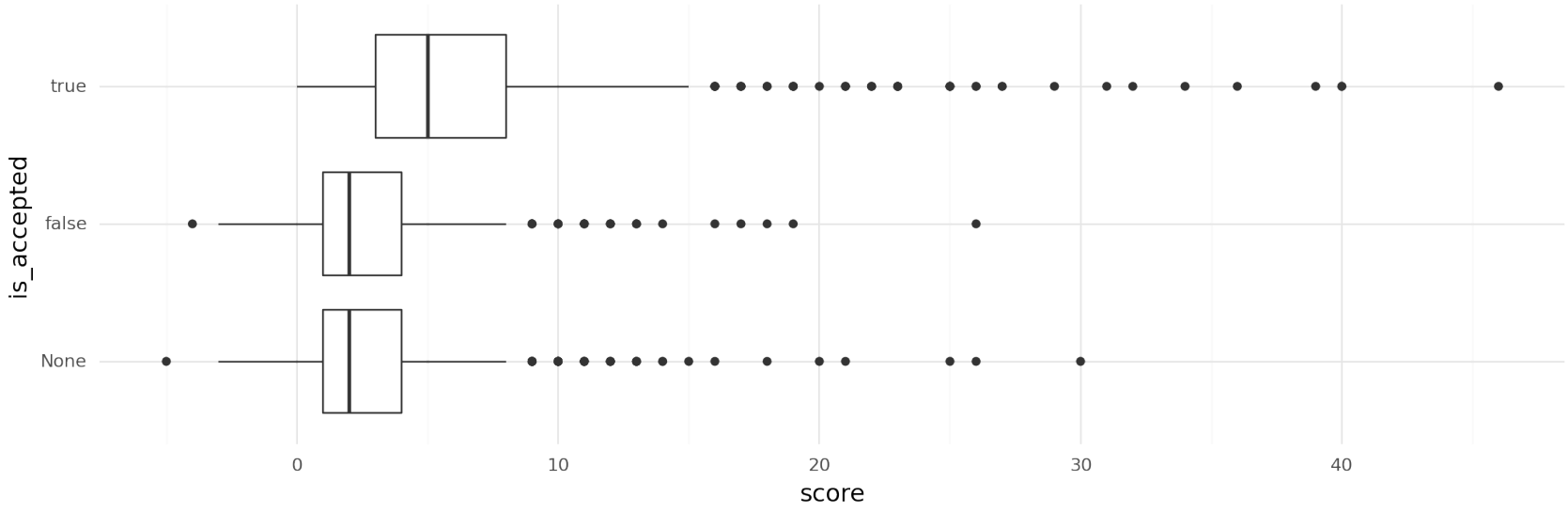
Rys. 1: Liczba postów w czasie



Rys. 2: Stosunek pytań do odpowiedzi w czasie

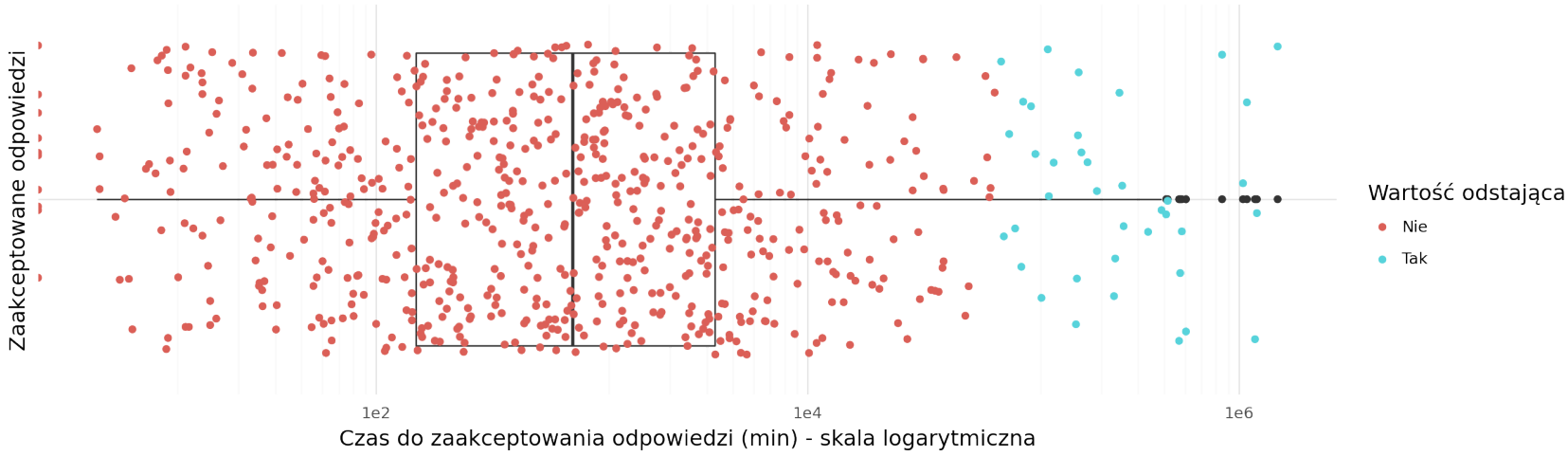
- średnio co drugie pytanie otrzymywało odpowiedź (0.47 +/- 0.22)

Statystyki postów



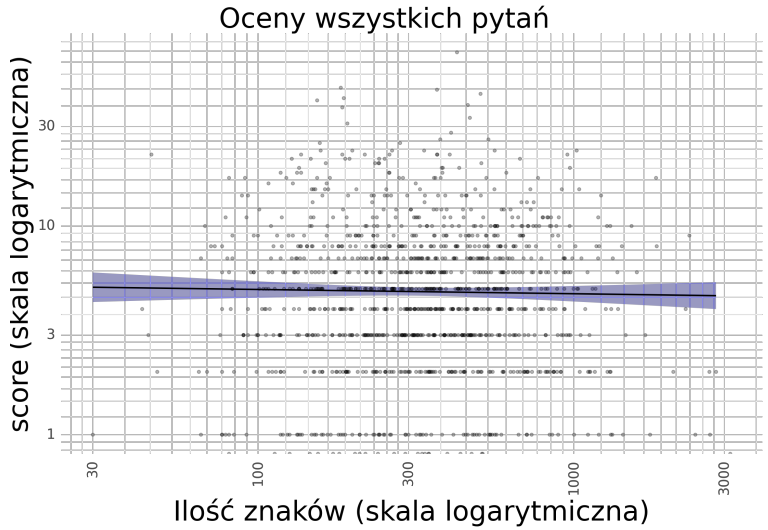
Rys. 3: Rozkład ocen

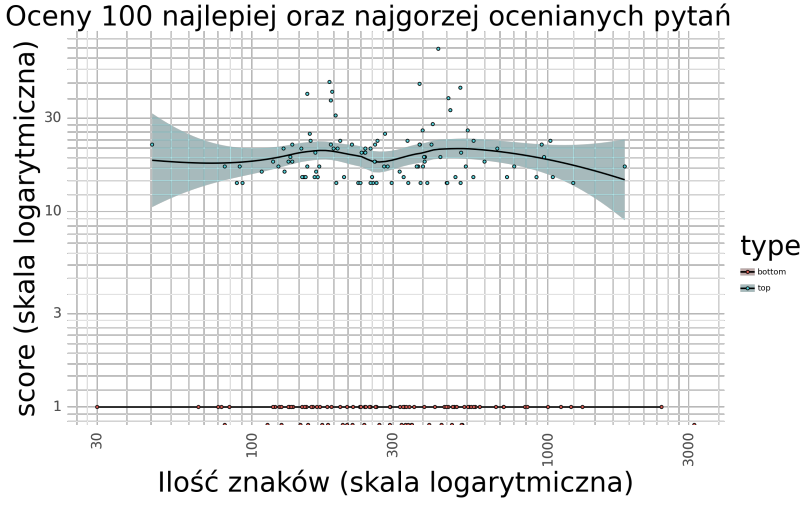
is_accepted	średnia	std	min	max	mediana
True	6.40	5.92	0	46	5
False	2.58	2.74	-4	26	2
None	2.76	3.18	-5	30	2



Rys. 4: Czas do akceptacji odpowiedzi

- Po odrzuceniu wartości odstających średni czas do akceptacji wyniósł ~66,5h (mediana 10,5h)





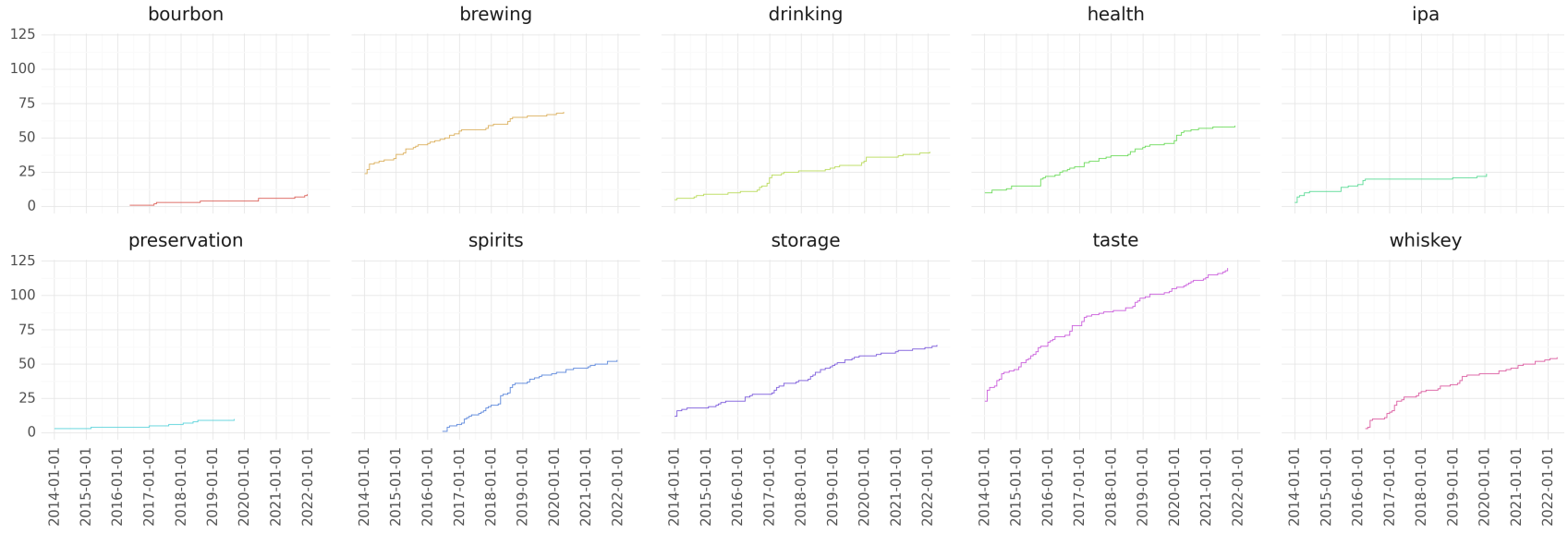
Rys. 5: Wartość score w porównaniu do długości pytania

- najlepsze pytania (top) - średnia ocen 20.6 oraz 4 odpowiedzi
- najgorsze pytania (bottom) - średnia ocen 0.4 oraz 1 odpowiedź

Analiza tematów postów - tagi

tag	liczba wyświetleń
taste	1 330 670
health	1 286 001
preservation	682 216
storage	542 860
whiskey	464 756
bourbon	330 268
brewing	307 892
ipa	291 935
spirits	255 328
drinking	225 924

Kumulatywna suma postów z danym tagiem w czasie



Analiza tematów postów - tytuły

etap	przykład
(1) tytuł	What is a citra hop, and how does it differ from other hops?
(2) token	[what, is, citra, hop, and, how, does, it, differ, from, other, hops]
(3) – stop words	[citra, hop, differ, hops]
(4) stemming	[citra, hop, differ, hop]

stem	ilość wystąpień
beer	476
wine	147
drink	104
alcohol	88
differ	72
bottl	68

Podsumowanie

- Podstawowa infrastruktura - możliwe łatwe i szybkie skalowanie, optymalizacja sposobu przechowywania danych
- Otrzymane wyniki mogą posłużyć do dalszej analizy przyczyn spadającej aktywności na forum