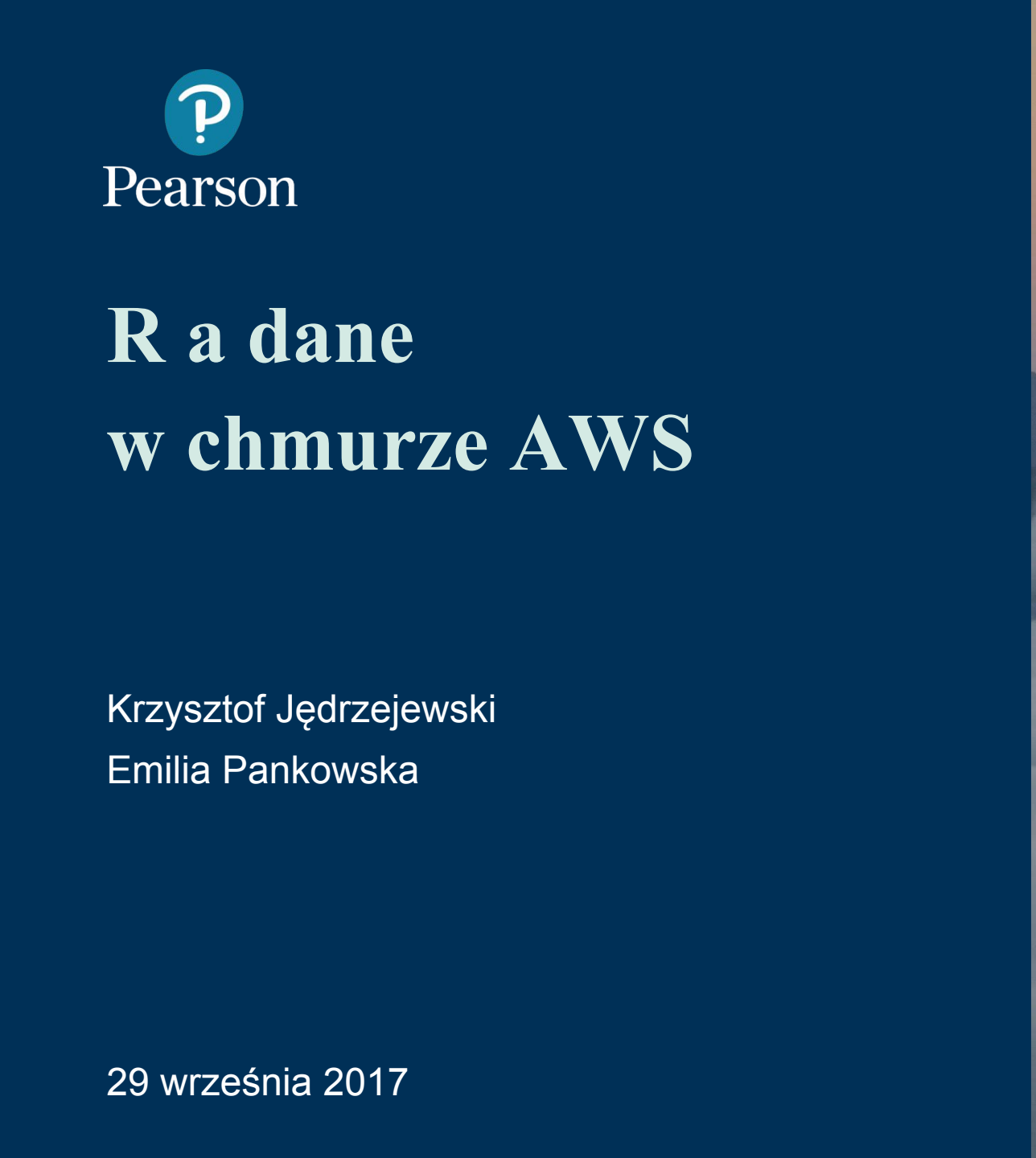


Pearson

R a dane w chmurze AWS

Krzysztof Jędrzejewski
Emilia Pankowska

29 września 2017

Kim jesteśmy?

Kim jesteśmy?

Gdzie jesteśmy?

Czym jest AWS?

Czym jest AWS?

- *Amazon Web Services*
- Zestaw różnego rodzaju usług chmurowych:
  - usługi magazynowe - S3
  - usługi bazodanowe, np. Redshift, Athena
  - usługi obliczeniowe i hostingowe, np. EC2, Lightsail
  - usługi “AI”, np. rozpoznawanie obrazów

https://aws.amazon.com

http://www.imfdb.org/wiki/File:AWS_1800.jpg

Czym jest Redshift i Athena?

Czym jest Redshift?

- Bazująca na Postgresie kolumnowa, rozproszona baza danych
- Składnia języka SQL zbliżona do Postgresa 8
  - są jednak pewne różnice
- Działa sprawnie na danych o rozmiarze rzędu petabajtów
- Płatny za czas pracy węzłów obliczeniowych

http://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html

Czym jest Athena?

- SQL-owy interfejs do danych przechowywanych w S3
- Zbudowana w oparciu o Presto DB
- Obsługuje bogatą składnię SQL
- Płatna za rozmiar danych odczytanych
- Może być używana w połączeniu z Redshiftem

https://aws.amazon.com/athena/
https://aws.amazon.com/redshift/spectrum/

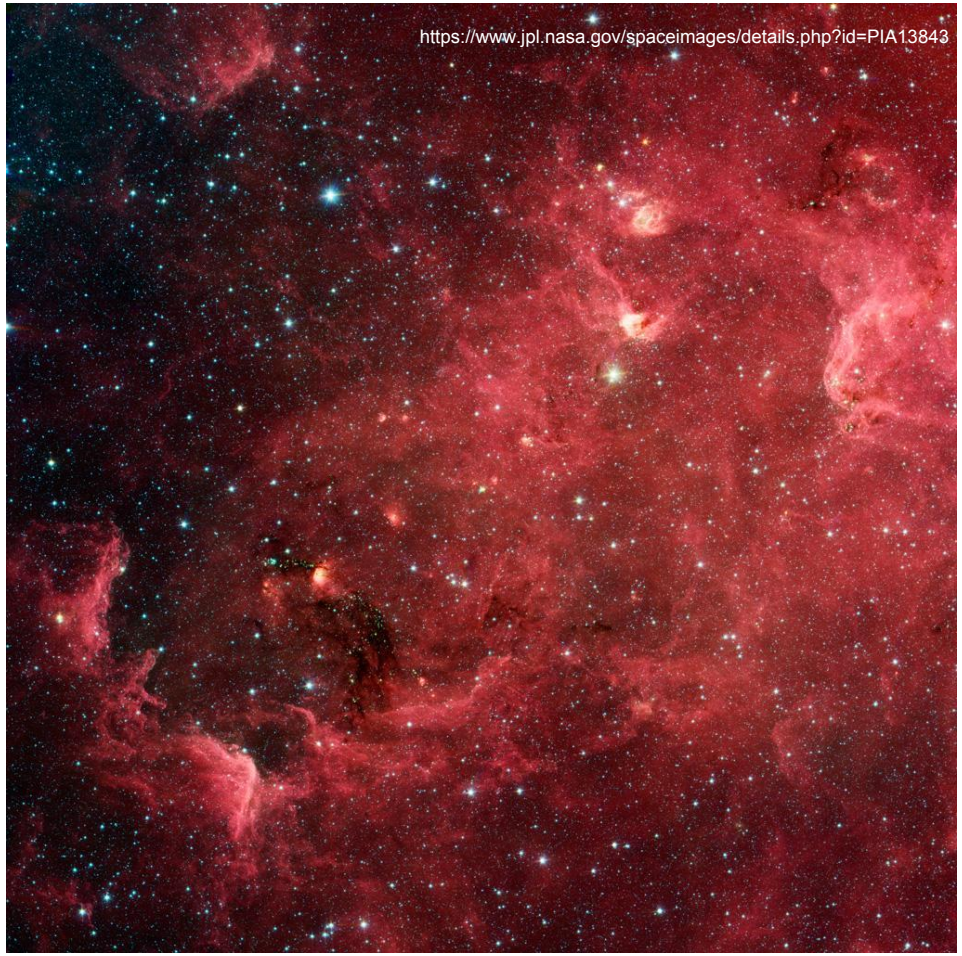
Dlaczego korzystamy z Redshifta?

Dlaczego Redshift?

Bo pozwala pracować bardzo sprawnie z dużymi danymi

- Operacje na jednej kolumnie w tabeli z ~1.1 miliarda wierszy
  - Obliczenie średniej: **~1s**
  - Obliczenie mediany: **~3m**
- Znalezienie liczby unikalnych kombinacji wartości w dwóch kolumnach pośród ~1.1 miliarda wierszy: **~40s**

```
SELECT count(*)
FROM (
  SELECT DISTINCT a, b FROM tabela
) a;
```



Dlaczego Athena?

- Działa szybciej niż Redshift, np.
- Znalezienie liczby unikalnych kombinacji wartości w dwóch kolumnach pośród ~16.6 miliarda wierszy: **10-20s**

Image by Pearson Scott Foresman

Redshift i R

Jak korzystać z Redshifta w połączeniu z R

Redshift daje możliwość sprawnego manipulowania dużymi zbiorami danych. Jednak analizowanie tych danych wymaga pobrania ich na lokalny komputer.

Dane z Redshifta można pobrać bezpośrednio, tak jak z każdej innej bazy danych, jest to jednak bardzo czasochłonne w przypadku dużych wolumenów. Oprócz tego, do czasu zakończenia pobierania danych, część zasobów klastra jest zablokowana.

Czy istnieje lepszy sposób?

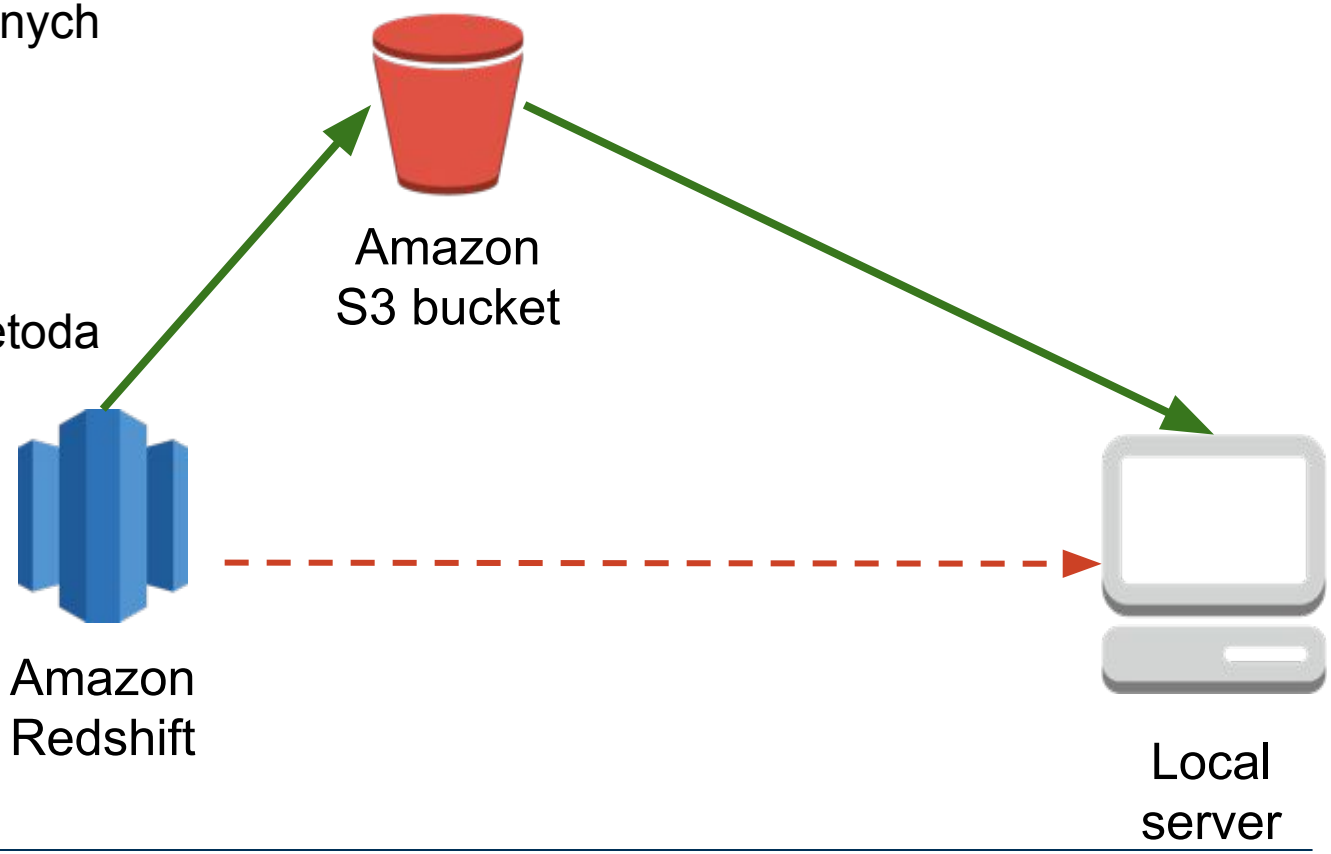
Jak pobierać dane?

Jak pobrać dane z Redshifta?

1. Bezpośrednio
   - Dobra metoda dla małych zbiorów danych

     paczki: RPostgreSQL oraz DBI
2. Korzystając z pośrednictwa S3
   - Zdecydowanie bardziej efektywna metoda dla dużych zbiorów danych

     paczki: aws.S3



Schemat pracy z danymi

1. Przygotowanie danych na serwerze Redshift
2. Załadowanie finalnej tabeli z danymi na serwer S3

   Służy do tego komenda UNLOAD wywoływana z poziomu Redshifta
3. Ściągnięcie danych z serwera S3 na serwer R

   Dane można przekopiować z serwera S3 bezpośrednio na lokalny serwer korzystając na przykład z konsoli amazonowej.
4. Obróbka danych w R

Wszystkie te kroki można wykonać z poziomu R.

Pobieranie danych z poziomu R

1. Załadowanie finalnej tabeli z danymi na serwer S3

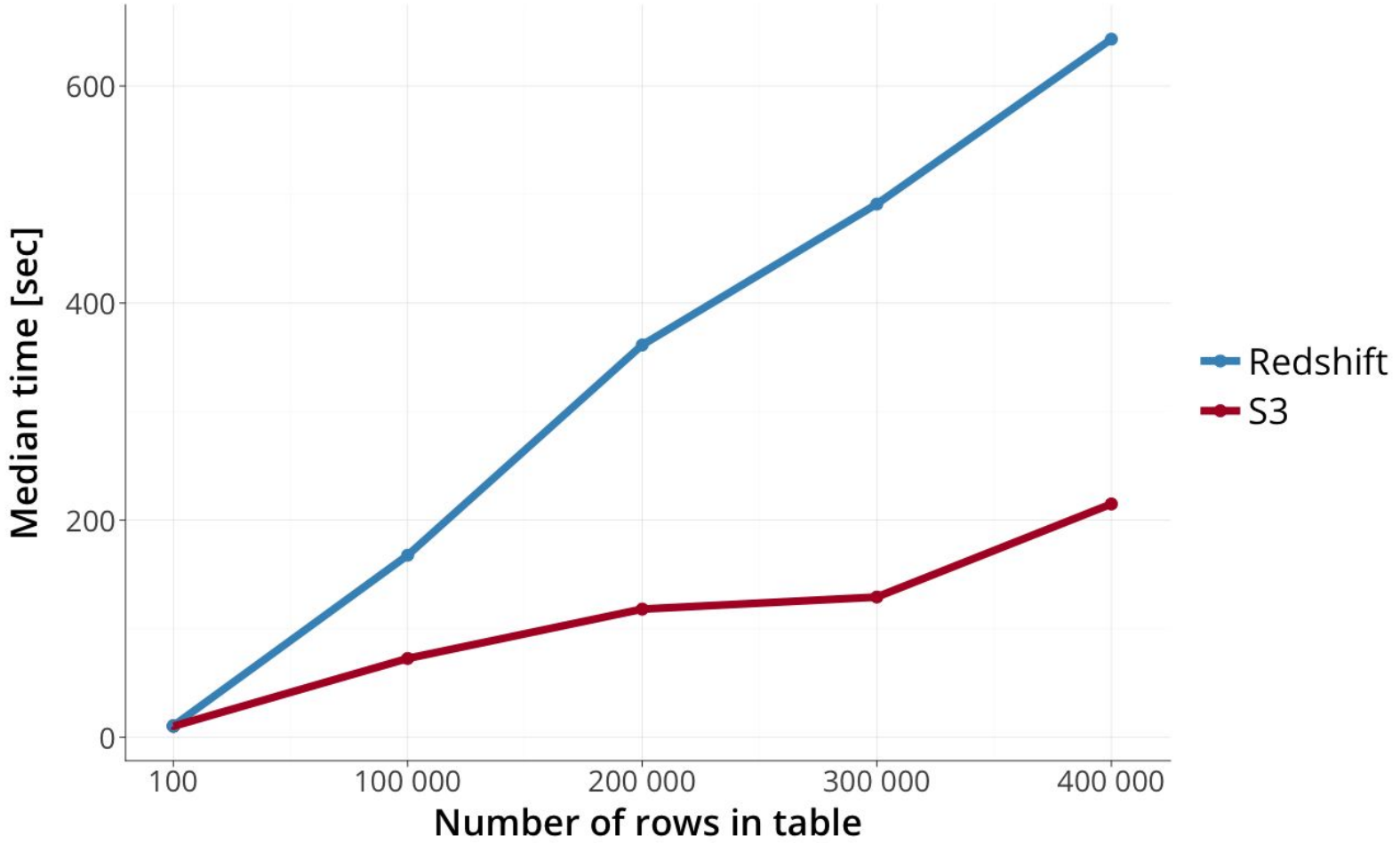
   Korzystając z paczek DBI oraz RPostgreSQL można wywołać dowolne zapytanie na serwerze bazodanowym - w tym również UNLOAD.

2. Ściągnięcie danych z serwera S3 na serwer R

   Paczka aws.S3 umożliwia połączenie oraz pobranie pliku z S3 z poziomu R

   - Pobranie wszystkich plików z serwera S3
   - Połączenie danych w jeden obiekt (data.frame, data.table)
   - Dodanie nazw kolumn oraz zmiana typów danych

Porównanie czasów pobierania danych



cloudyr

aws.s3:

https://github.com/cloudyr/aws.s3

Thomas J. Leeper (2017). aws.s3: AWS S3 Client Package. R package version 0.3.3.

cloudyr:

https://github.com/cloudyr

ALWAYS LEARNING