

Hadoop i Spark

Mariusz Rafało

mrafalo@sgh.waw.pl

<http://mariuszrafalo.pl>

ORGANIZACJA ZAJĘĆ

Ramowy plan zajęć

1. Wprowadzenie do ekosystemu Apache Hadoop
2. Techniki i technologie przetwarzania danych
3. Wprowadzenie do platformy Databricks
4. Wybrane technologie ekosystemu Big Data
 - a) Składowanie danych
 - b) Zarządzanie klastrem
 - c) Bezpieczeństwo klastra
5. Formaty plików w ekosystemie Apache Hadoop
6. Przetwarzanie danych w czasie rzeczywistym
7. Wprowadzenie do technologii Apache Kafka

Projekt zaliczeniowy

Literatura

1. Spark. Zaawansowana analiza danych, S. Ryza, U. Laserson, S. Owen, J. Wills, Helion, 2015
2. Big Data: A Revolution That Will Transform How We Live, Work, and Think, V. Mayer-Schönberger, K. Cukier, Eamon Dolan/Mariner Books, 2014
3. Mastering Apache Spark, M. Frampton, Packt Publishing Ltd., 2017
4. Big Data Analytics with R and Hadoop, V. Prajapati, Packt Publishing Ltd.

Materiały

1. Slajdy i materiały pdf

2. Kody źródłowe Databricks

3. Inne materiały

Logistyka

BIG DATA

Definicja Big Data

- Big Data definiowane jest jako składowanie zbiorów danych o tak dużej złożoności i ilości danych, że jest to niemożliwe przy zastosowaniu podejścia tradycyjnego (np. opartego na hurtowni danych)
- Zagadnienie obejmuje identyfikację, pobieranie, składowanie, przeszukiwanie, współdzielenie, analizę i wizualizację danych

wikipedia.org

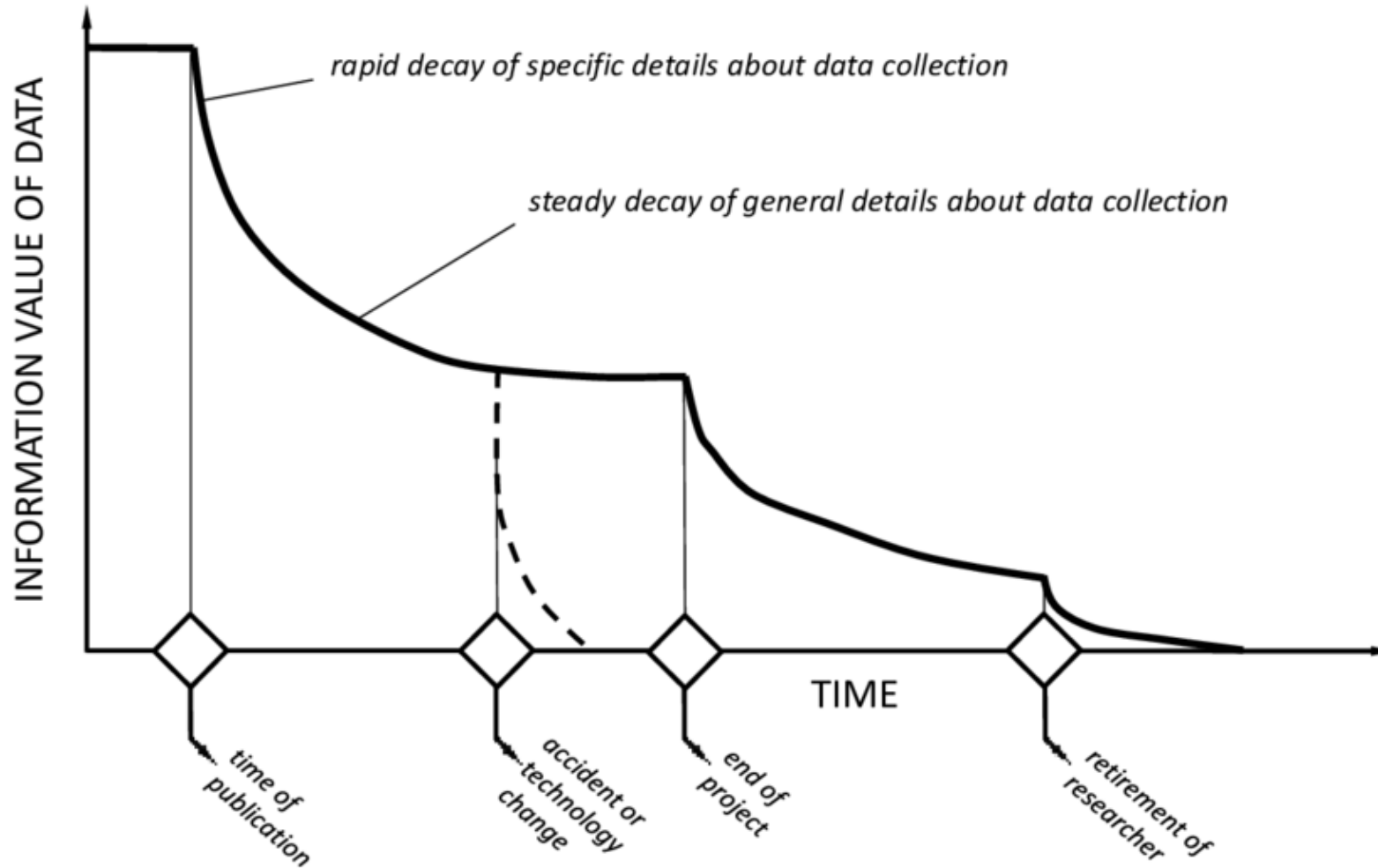
Paradygmat Big Data

Element	Paradygmat klasyczny	Paradygmat Big Data
Ilość danych	Kolejne przyrosty danych cyklicznie ładowane do hurtowni danych	Analizowanie danych w czasie rzeczywistym, zapisywanie wyłącznie informacji kluczowych
Szybkość danych	Cykliczne pobieranie wyłącznie istotnych danych. Wysoki (względnie) poziom latencji (opóźnienia) danych	Nasłuch strumienia danych, w momencie pojawienia się określonych sytuacji, natychmiastowe podjęcie działania
Różnorodność danych	Umieszczanie danych w bazie danych o określonej strukturze	Strukturyzowanie danych, które pozwalają na określenie kontekstu danych o nieustrukturyzowanej postaci

Big Data: 3V's

- Wolumen danych (***Volume***)
- Zróżnicowanie danych (***Variety***)
- Szybkość zmian danych (***Velocity***)

Big Data – kolejne **V: Value** (wartość)



https://www.researchgate.net/figure/Information-value-of-data-over-time_fig1_329062811

Big Data – kolejne **V: Veracity** (wiarygodność)

- Big data, ze względu na rodzaj danych oraz ich skalę, obarczony jest szeregiem problemów:
 - Błędy danych
 - Przekłamania
 - Szum informacyjny
 - Anomalie w danych
- W takich uwarunkowaniach istotne jest zarządzanie wiarygodnością danych dla ich użytkowników

Dziękuję za uwagę