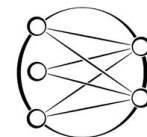




Big Data dla początkujących

Koło Naukowe Machine Learning
Politechnika Rzeszowska
mgr inż. Paweł Kuraś, EMBA



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



O czym będzie ten wykład?



Big Data - jak to się zaczęło?

Archie Query Form 

Search for:

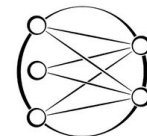
Database: ☒ Worldwide Anonymous FTP ☐ Polish Web Index

Search Type: ☒ Sub String ☐ Exact ☐ Regular Expression

Case: ☒ Insensitive ☐ Sensitive

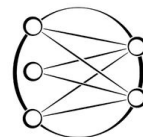
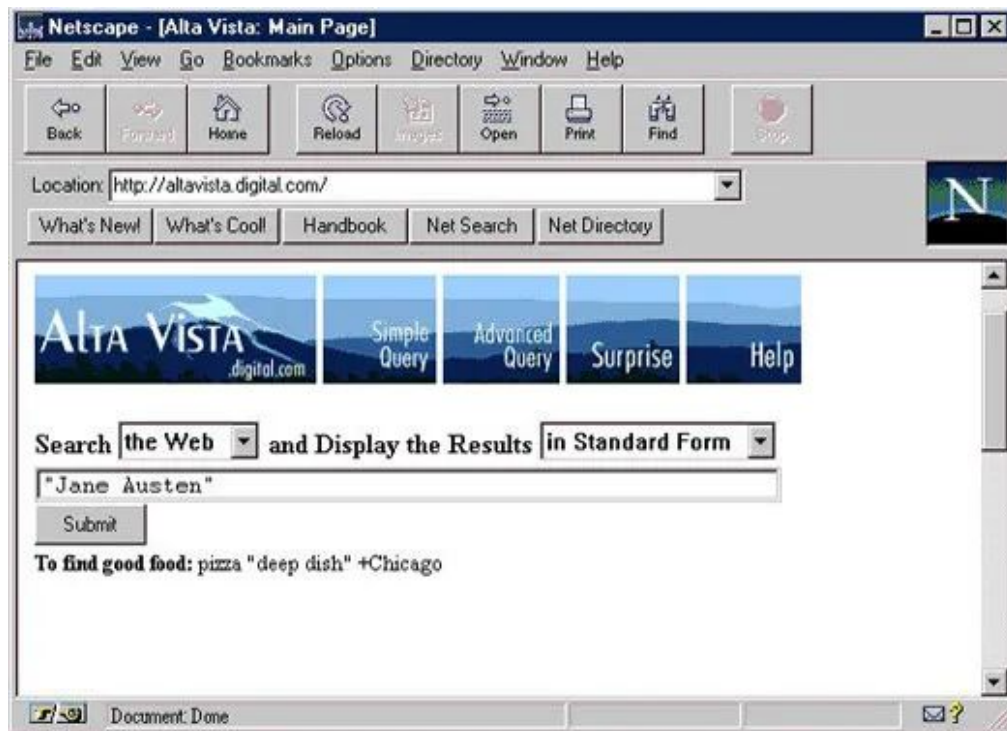
Do you want to look up strings only (no sites returned):
☒ NO ☐ YES

Output Format For Web Index Search: ☐ Keywords Only
☒ Excerpts Only
☐ Links Only





Big Data - jak to się zaczęło?





Big Data - jak to się zaczęło?



Search the web using Google!

10 results

Google Search

I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

[About Google!](#)

[Stanford Search](#) [Linux Search](#)

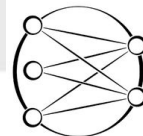
Get Google! updates monthly!

your e-mail

Subscribe

[Archive](#)

Copyright ©1997-8 Stanford University

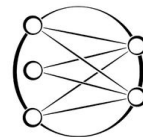


KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



Big Data - jak to się zaczęło?

- 2003 rok - Google File System (GFS)
- 2004 rok - Map Reduce
- 2005 rok - Big Table
- 2007 rok - Apache Hadoop
- 2008 rok - Apache HBase, Accumulo





Filozofia Big Data - czym różni się od myślenia “tradycyjnego”?

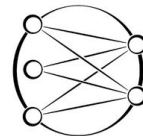


Czym nie jest Big Data?



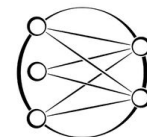
Big Data - czym nie jest?

- Big Data nie jest prostym i skutecznym rozwiązaniem złożonych problemów dostępnym “od ręki”
- Big Data nie jest narzędziem dostępnym dla każdego bez przeszkolenia
- Big Data nie jest "BI na sterydach"
- Big Data nie jest "rozwiązaniem"
- Big Data nie nadaje się do "nisko wiszących owoców".





Zasada 5V

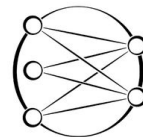




KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



Co odróżnia Big Data?

- Podejście rozproszone
- Podzespoły ulegają awarii - jest to norma, a nie anomalia
- Gromadzimy wszystkie możliwe dane, które być może się przydadzą, ale nie muszą
 - część danych może dostać nowe życie w nowych okolicznościach
 - czasem będziemy mieć mało danych
- Decentralizacja
- Długie czasy przetwarzania danych
- Duży poziom zróżnicowania danych
- Wiele iteracji!



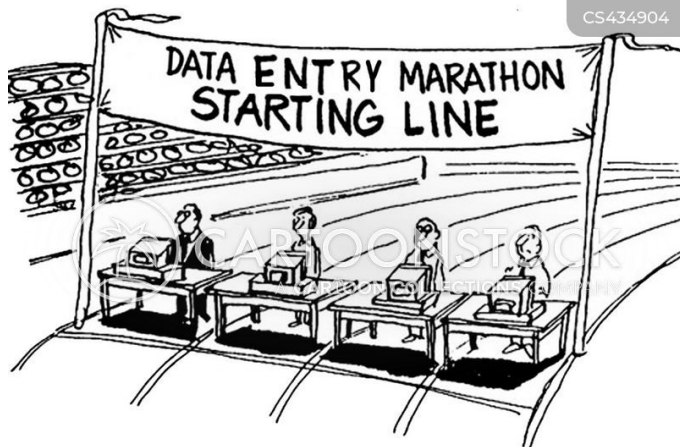


Obowiązkowe “pytania” podczas pracy nad systemem Big Data



Obowiązkowe “pytania” podczas pracy nad systemem Big Data

1. Wprowadzanie danych

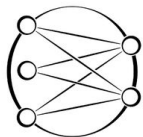


THE LAVA IS ENTERING THE SEA, AND
NEW RIFTS ARE OPENING TO THE NORTH!

GET A GIS SURVEY TEAM IN
THE AIR! WE NEED TO REVISE
OUR COASTLINE SHAPEFILES!



I WANT TO MAKE A DISASTER MOVIE
THAT JUST SHOWS SCIENTISTS RUSHING
TO UPDATE ALL THEIR DATA SETS.

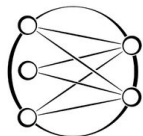
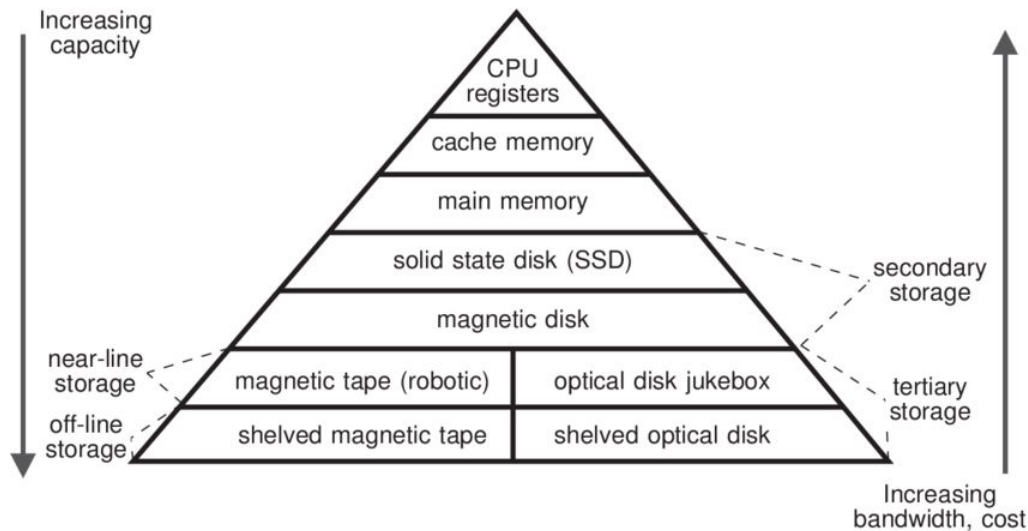


KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



Obowiązkowe “pytania” podczas pracy nad systemem Big Data

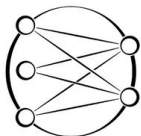
1. Wprowadzanie danych
2. Przechowywanie danych





Obowiązkowe “pytania” podczas pracy nad systemem Big Data

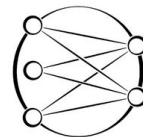
1. Wprowadzanie danych
2. Przechowywanie danych
3. Jakość danych





Obowiązkowe “pytania” podczas pracy nad systemem Big Data

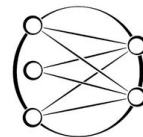
1. Wprowadzanie danych
2. Przechowywanie danych
3. Jakość danych
4. Operacje na danych





Obowiązkowe “pytania” podczas pracy nad systemem Big Data

1. Wprowadzanie danych
2. Przechowywanie danych
3. Jakość danych
4. Operacje na danych
5. Skalowalność i bezpieczeństwo danych





Budowanie strategii Big Data

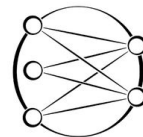


Pięć “P” procesu analizy danych



Pięć “P” analizy danych

1. Purpose (cel)
2. People (ludzie)
3. Process (proces)
4. Platforms (platformy)
5. Programmability (programowalność)



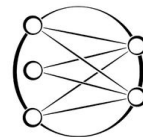


Technologie Big Data: Grupy



Omawiane grupy technologii:

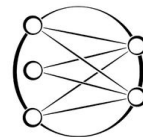
1. Storages
2. Bazy danych (nierelacyjne)
3. Full-text search
4. Przetwarzanie danych
5. Komunikacja z danymi
6. Schedulers
7. Messaging





Storages (magazyny danych?)

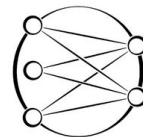
1. HDFS
2. Ozone
3. ADLS (Azure)
4. Amazon S3 (AWS)
5. Google Cloud Storage (GCP)





Bazy danych (nierelacyjne!)

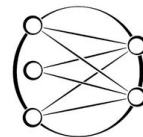
1. HBase
2. Accumulo
3. MongoDB
4. Cassandra
5. CosmosDB (Azure)
6. Dynamo DB (AWS)
7. Firestore (GCP)
8. Kudu
9. Ozone (wymienione także w storages)
10. Neo4j
11. Druid





Full-text search

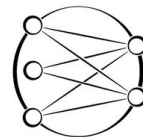
1. Lucene
2. Elasticsearch
3. Solr





Przetwarzanie danych

1. Spark
2. Spark Structured Streaming
3. Kafka Streams
4. Flink
5. Map Reduce

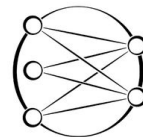




Komunikacja z danymi

1. Hive

2. Impala

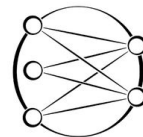


KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



Schedulers

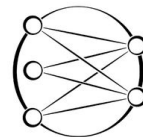
1. Oozie
2. Airflow
3. Step Functions (AWS)
4. Workflows (GCP)
5. Logic Apps (Azure)





Messaging

1. Kafka
2. RabbitMQ
3. EventHub (Azure)
4. Kinesis (AWS)
5. Pub/Sub (GCP)



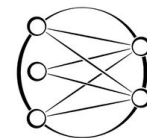


Przegląd technologii Big Data



Hadoop

Grupa: ekosystem technologii BD

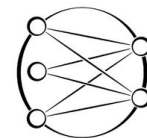


KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



HDFS

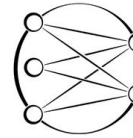
Grupa: Storage



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ



Ozone



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Storage, Bazy danych

Podstawowy opis struktury:

1. Volumes
2. Buckets
3. Keys





ADLS (gen 2)



Grupa: Storage

Cechy:

- Kompatybilność z Hadoop
- Bezpieczeństwo
- Efektywność
- Dobra optymalizacja





Amazon S3



Grupa: Storage

Budowa:

- Buckets
- Objects
- Keys





HBase i Accumulo

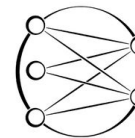


Grupa: Bazy Danych





MongoDB



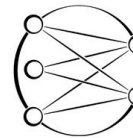
KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Bazy Danych





CosmosDB



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Bazy danych

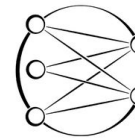
Cechy:

- API dla wielu języków
- Szybkość, wydajność
- Model danych - JSON





Dynamo DB



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Bazy danych

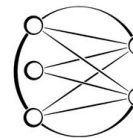
Model danych:

- Tables
- Items
- Attributes





Kudu



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Bazy danych

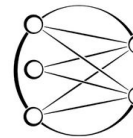
Model danych:

- Tabele
- Kolumny
- Nazwy, typy
- Klucze





Neo4j



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Bazy danych

Model danych:

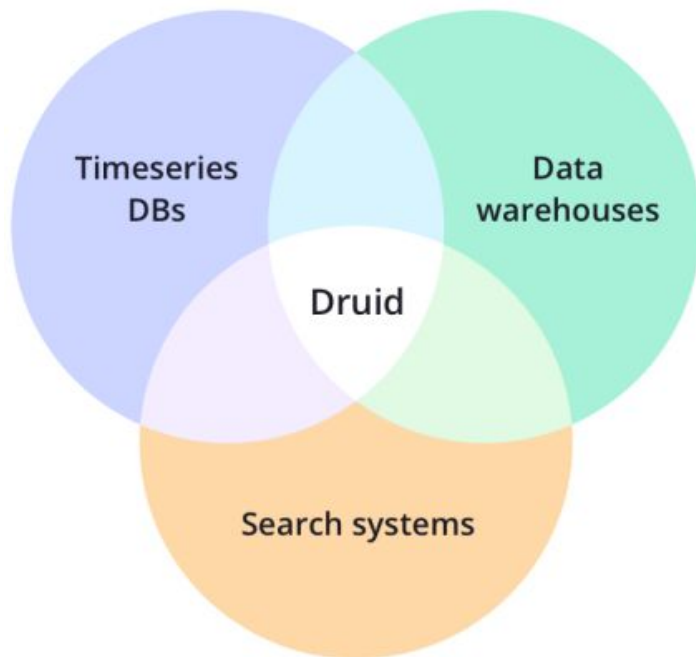
- Wierzchołki (nodes)
- Relacje (relationships)





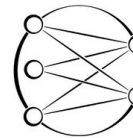
Apache Druid

Grupa: Bazy danych





Solr



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Full-text Search

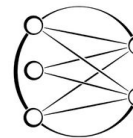
Komponenty

- Request Hangler
- Search Component
- Query Parser
- Response Writer
- Analyzer/Tokenizer
- Update Request Processor





Solr



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Full-text Search

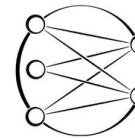
Pojęcia

- Instancja (instance)
- Core
- Shard
- Kolekcja (collection)
- Replika (replica)





Elasticsearch



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Full-text Search

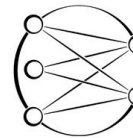
Budowa

- Node
- Cluster
- Index
- Type
- Document





Spark



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Przetwarzanie danych

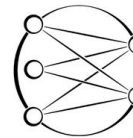
Biblioteki:

- Spark SQL
- Spark ML
- GraphX
- GraphQL
- Spark Structured Streaming
- Spark Streaming
- Connectory





Flink



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

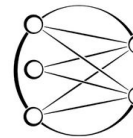
Grupa: Przetwarzanie danych

Apache Flink to framework i silnik do rozproszonego przetwarzania danych dla stanowych (ang. stateful), ograniczonych i nieograniczonych strumieni danych





Kafka Streams



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Przetwarzanie danych

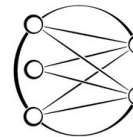
Budowa

- Stream Processor
- Table
- Aggregation Operation
- Join Operation
- Windowing
- Interactive Queries





Map Reduce



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Przetwarzanie danych

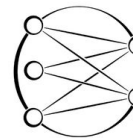
Po co dziś?

- Kompatybilność wsteczna
- Fundament innych technologii (np. Hive)





Hive



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Komunikacja z danymi

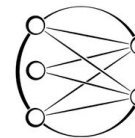
Budowa

- User Interface
- Meta Store





Impala



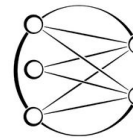
KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Komunikacja z danymi





Oozie



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Schedulery

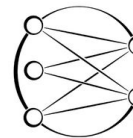
Rodzaje jobów:

- Oozie Workflow jobs
- Oozie Coordinator jobs
- Oozie Bundle





Apache Airflow



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Schedulery

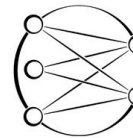
Podstawowe zasady Airflow:

- Dynamic
- Extensible
- Elegant
- Scalable





Step Functions



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Schedulery

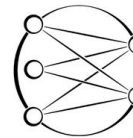
Dwa typy Workflow:

- Standard (domyślnie)
- Express





Workflows



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Schedulery

**W przypadku Workflows,
przeptywy tworzone są
“ręcznie” - przy pomocy
plików typu yaml.**





Azure Logic Apps



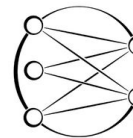
Grupa: Schedulery

Jest to usługa analogiczna jak dwie poprzednie, z tą różnicą, że znajduje się na Azure, czyli chmurze Microsoftu.





Kafka



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Messaging/Kolejkowanie

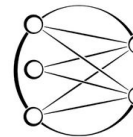
Budowa Kafki:

- Broker
- Klaster
- Topic
- Partycje
- Producer
- Consumer
- Message





RabbitMQ



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Messaging/Kolejkowanie

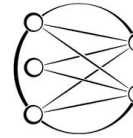
Budowa RabbitMQ:

- Broker
- Klaster
- Kolejka
- Exchange
- Binding
- Producer
- Consumer





Event Hubs



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Messaging/Kolejkowanie

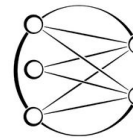
Budowa (w stosunku do Kafki):

- Cluster (Kafka) to Namespace
- Topic (Kafka) to Event Hub
- Partition, Consumer group i offset pozostają bez zmian.





Kinesis



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Messaging/Kolejkowanie

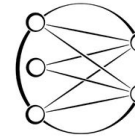
Cechy:

- Real-time
- Fully managed
- Scalable





Pub/Sub



KOŁO NAUKOWE
MACHINE LEARNING
POLITECHNIKI RZESZOWSKIEJ

Grupa: Messaging/Kolejkowanie

Budowa:

- Topic
- Subscription
- Message
- Message attribute
- Publisher
- Subscriber
- Push i Pull



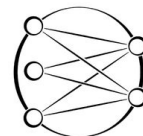


Cloud vs On-premise - jak wybrać infrastrukturę?



Cloud vs. własna architektura

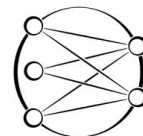
1. Własna architektura





Cloud vs. własna architektura

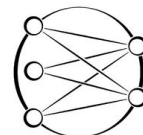
1. Własna architektura
2. Chmura





Cloud vs. własna architektura

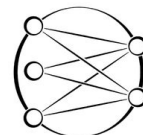
1. Własna architektura
2. Chmura
3. Chmury prywatne
 - a. Infrastructure as a Service (IaaS)
 - b. Platform as a Service (PaaS)





Cloud vs. własna architektura

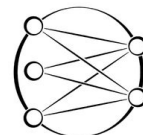
1. Własna architektura
2. Chmura
3. Chmury prywatne
 - a. Infrastructure as a Service (IaaS)
 - b. Platform as a Service (PaaS)
4. Chmury hybrydowe





Uprozczone porównanie chmury i on-premise

1. Koszt zasobów
2. Koszt obsługi
3. Jakość obsługi
4. Stabilność kosztów
5. Gdzie przechowywane są dane
6. Z jakich technologii można korzystać
7. Łatwość skalowalności
8. Dostęp do danych w kontekście prywatności



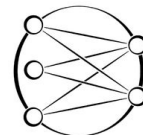


Zrozumienie chmury na tle on-premise



Prostota chmury

1. Utworzenie klastra klikając “Utwórz”
2. Określenie ilości nodów
3. Przestanie kodu na serwer
4. Określenie kilka opcji potrzebnych do uruchomienia programu.
5. Kliknięcie “Uruchom”.





“Prostota chmury”

1. Bardzo mocne uzależnienie
2. Niestabilne koszty

