Potok Big Data

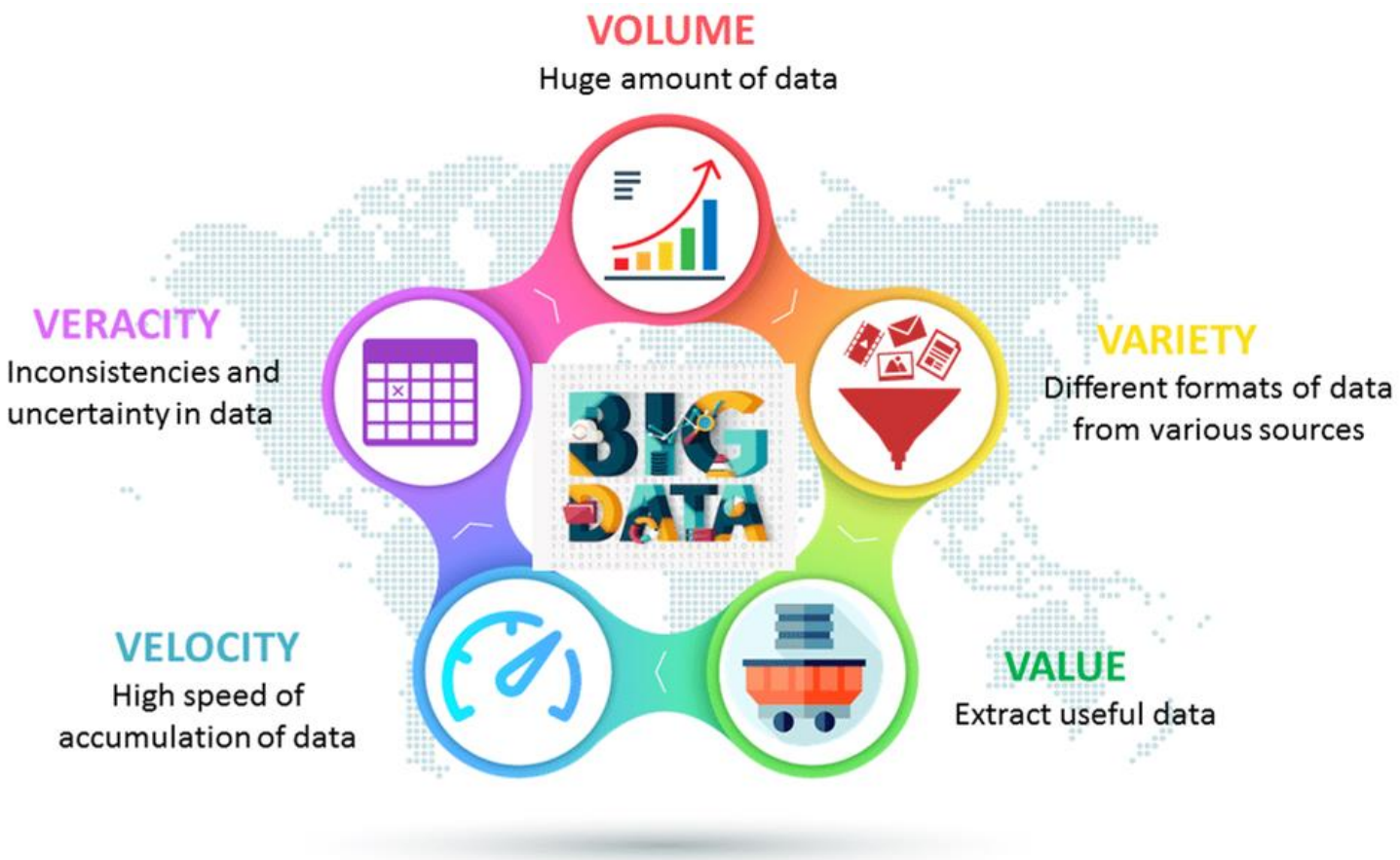
iBigWorld:
Innovations for Big Data in a Real World

Prof. dr. Dragan Stojanovic, **UNI**
Prof. dr. Natalija Stojanovic, **UNI**

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data 5V



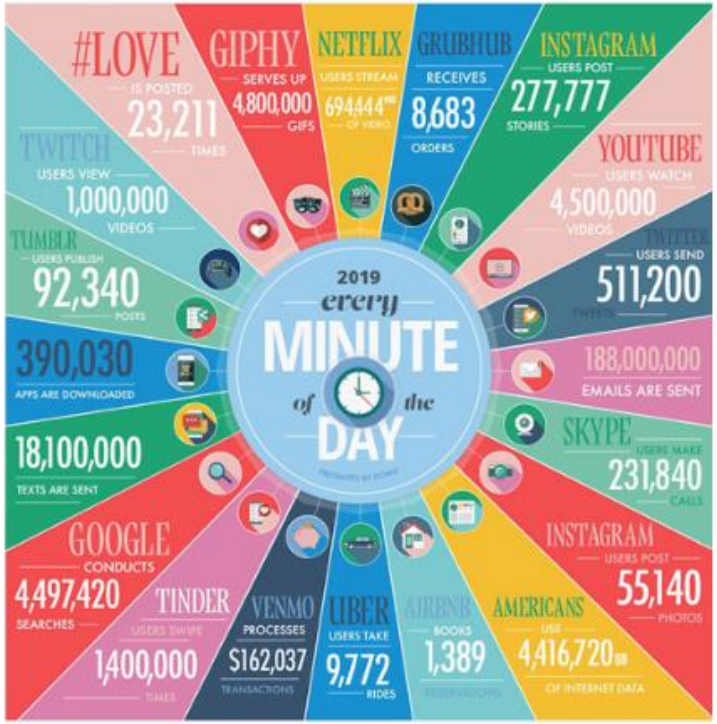
https://statanalytica.com/blog/4-vs-of-of-big-data/

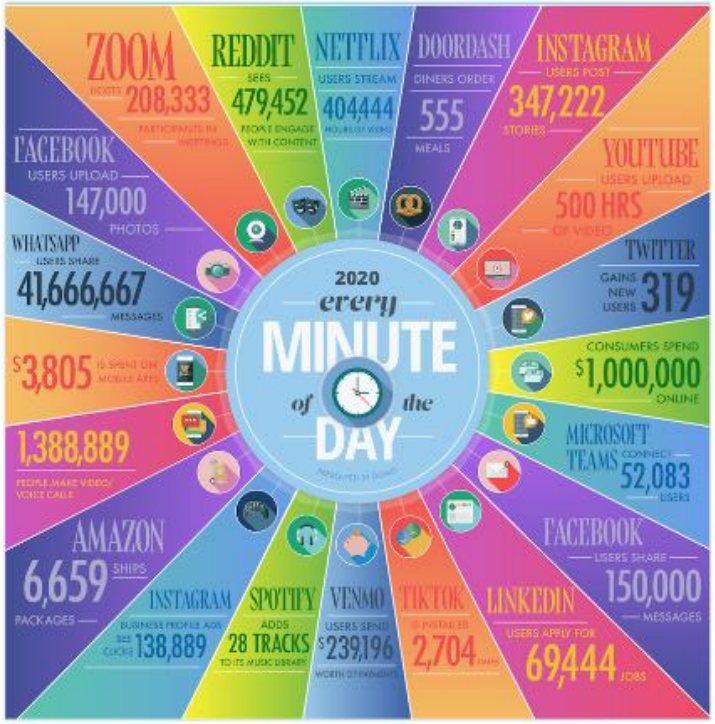
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Generowanie Big Data w 2019 i 2020

- Covid-19 wpływ na nasze życie



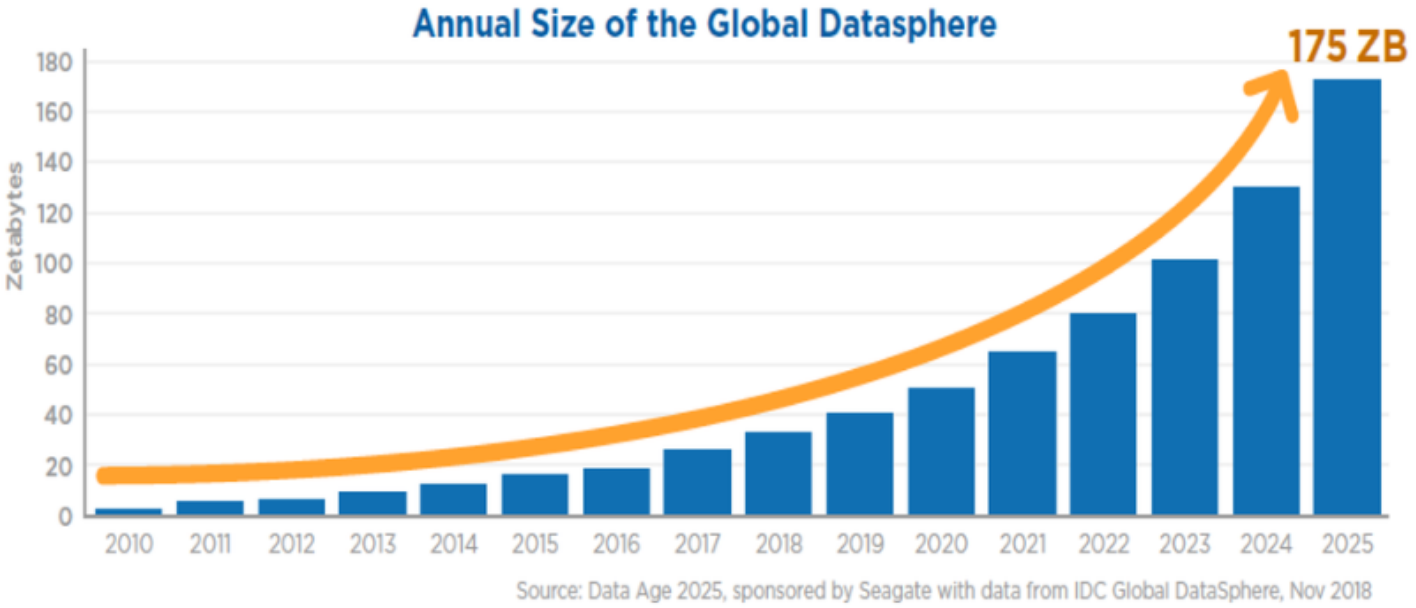


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big data - jakie ilości danych?

- Ostatnia eksplozja ilości danych
  - W 2013 r.: 90% wszystkich danych na świecie zostało wygenerowanych w ciągu ostatnich dwóch lat
- 50-krotny wzrost od 2010 do 2020



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data - liczby

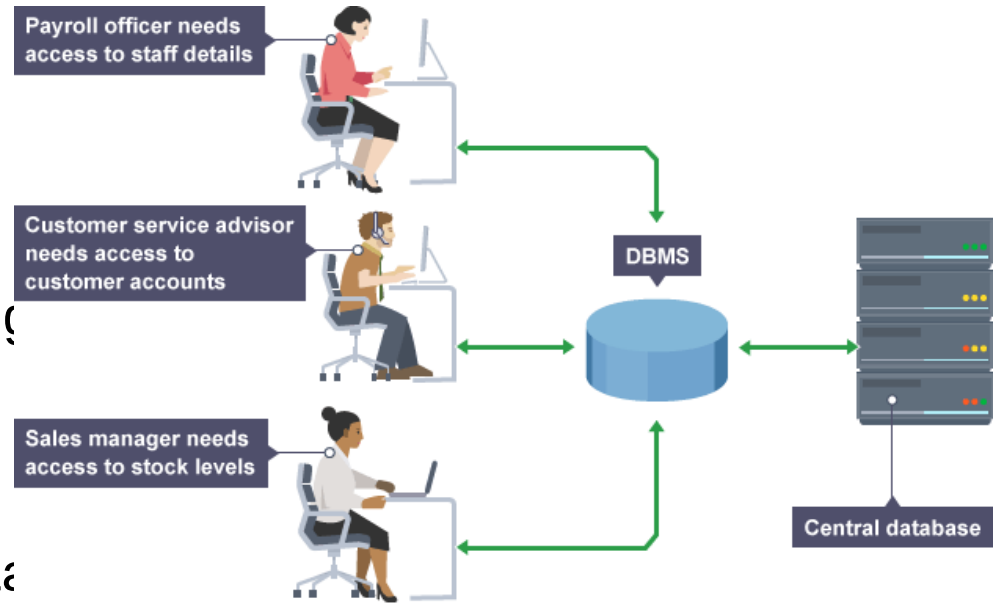
- W 2020r każda osoba generuje 1,7 MB w ciągu sekundy
- Internauci generują około 2,5 EB danych dziennie
- Rynek Big Data i analityki biznesowej ma osiągnąć 274 mld USD do 2022 r. (źródło IDC)
- 91% organizacji inwestuje w Big Data i AI
- Korzystając z Big Data, Netflix oszczędza 1 miliard dolarów rocznie na utrzymaniu klientów

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Problemy z tradycyjnymi bazami danych

- Wydajność
  - Czytanie,
  - Zapis/aktualizacja
- Skalowalność
  - Kolejki wiadomości
  - Replikacja, shardin
- Tolerancja błędów
  - Awaria dysku
  - Dostępność?
- Kompleksowe zarządza
  - Błędy ludzkie i błędy oprogramowania
  - Tolerancja na ludzkie winy
- BD nie jest świadoma swojej rozproszonej natury



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Skalowanie

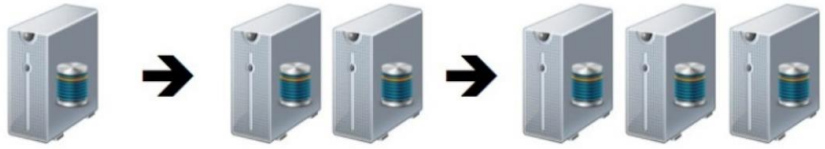
Podejścia do skalowania systemów obliczeniowych i pamięci masowej

Skalowanie w pionie (Skalowanie w górę)

- Powiększ pojedynczą maszynę
- Ograniczone w przestrzeni
- Kosztowne

- Skalowanie w poziomie (Skalowanie poziome)
- Korzystaj z wielu popularnych maszyn i twórz klastry lub sieci komputerowe

Utrzymanie klastra



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Zarządzanie danymi w skali Peta

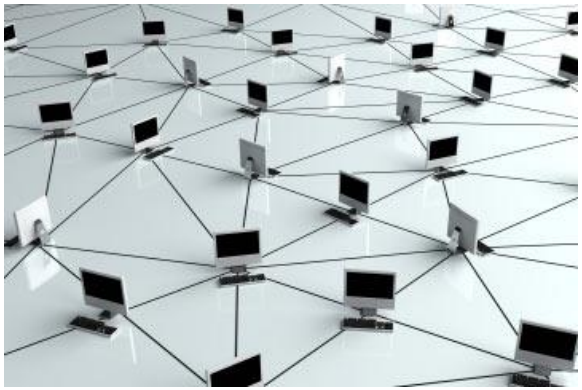
- Zarządzanie danymi w skali Peta
- Przepustowość sieci jest ograniczona
- Za mało pamięci RAM, aby pomieścić wszystkie dane w pamięci
- Dostęp do dysku jest powolny, ale przepustowość dysku jest rozsądna
- Nie przenoś danych do pracowników... przenoś pracowników do danych!
  - Przechowuj dane na lokalnych dyskach węzłów w klastrze
  - Uruchom pracowników w węźle, w którym dane są przechowywane lokalnie
- Rozproszony system plików jest odpowiedzią
  - GFS (system plików Google) dla Google MapReduce
  - HDFS (rozproszony system plików Hadoop) dla Hadoop

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Techniki Big Data

- Dystrybucja zasobów i usług
- Bazy danych i systemy obliczeniowe są świadome swojej rozproszonej natury
- Architektura rozproszona
  - Klastry komputerów, spółdziałają (skala w górę)
- Tolerancja błędów
  - Replikacja zasobów
  - Spójność
  - Tylko dołączanie
- Przetwarzanie rozproszone na dużą skalę
  - Model bez współdzielenia
  - Nowe paradygmaty programowania
  - Nowe oprogramowanie



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Pożądane właściwości systemu Big Data

- Pierwsze zasady
  - System danych odpowiada na pytania w oparciu o dane/informacje, które zostały pozyskane w przeszłości do chwili obecnej *query = function(all_data)*
- Solidność i odporność na uszkodzenia
- Odczyty i aktualizacje o niskim opóźnieniu
- Skalowalność
- Uogólnienie
- Rozszerzalność
- Zapytania ad hoc
- Minimalne zarządzanie
- Debugowalność

iBigWorld:2020-1-PL01-KA203-082197

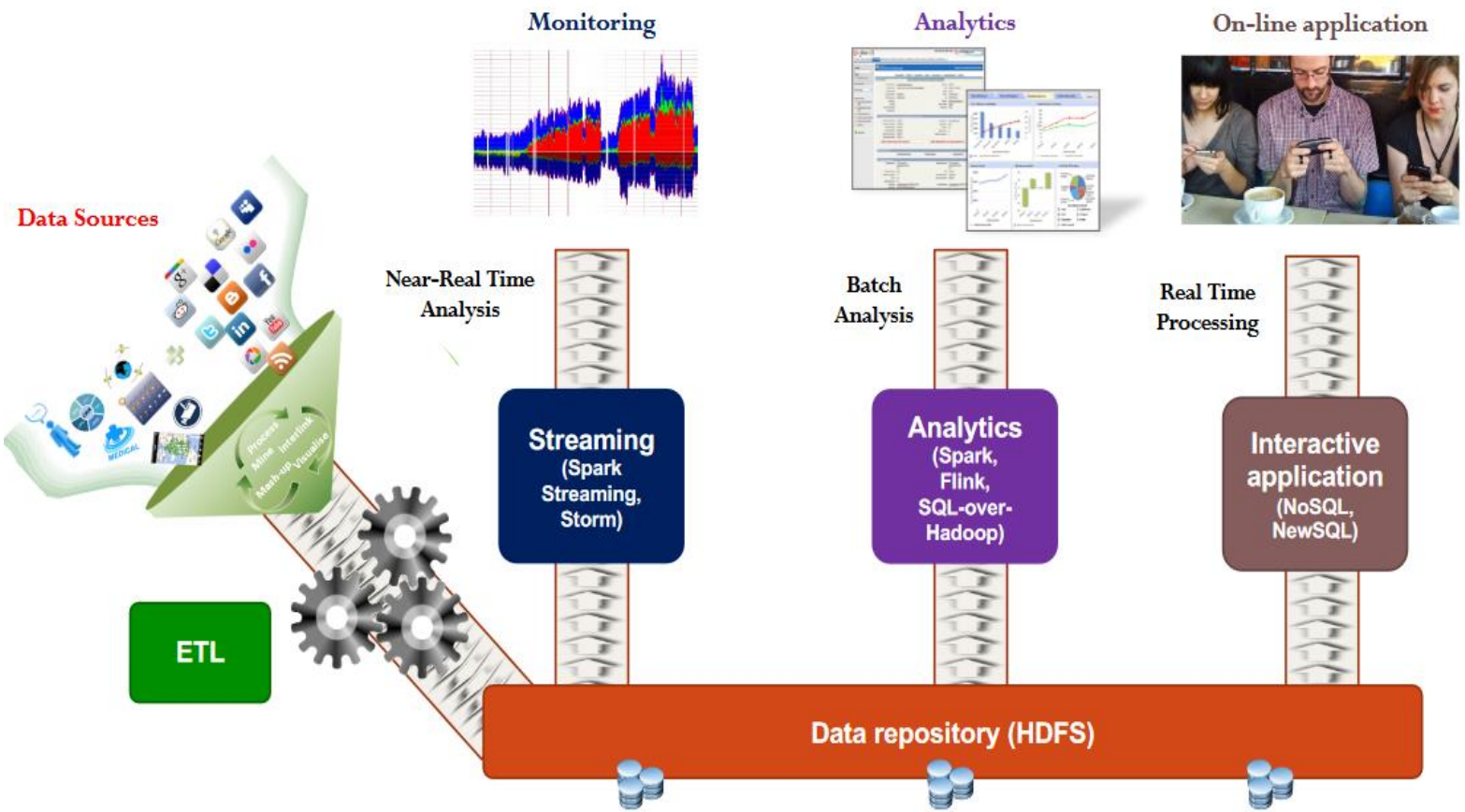
Co-funded by the Erasmus+ Programme of the European Union

Architektura Big Data

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przepływ danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przepływ danych

Analiza (wsadowo)

- Zbieranie, przekształcanie i modelowanie danych w celu odkrywania przydatnych informacji i wspomagania podejmowania decyzji
- Wejście/wyjście tylko do dołączania, niekoniecznie trwałe dane, brak transakcji ACID

Przesyłanie strumieniowe (blisko czasu rzeczywistego)

- Przetwarzanie danych strumieniowych (sekwencji elementów danych udostępnianych w czasie) w celu monitorowania i analizowania danych w locie za pomocą okien czasowych
- Strumień I/O, prawdopodobnie trwałe dane do analizy, brak transakcji ACID

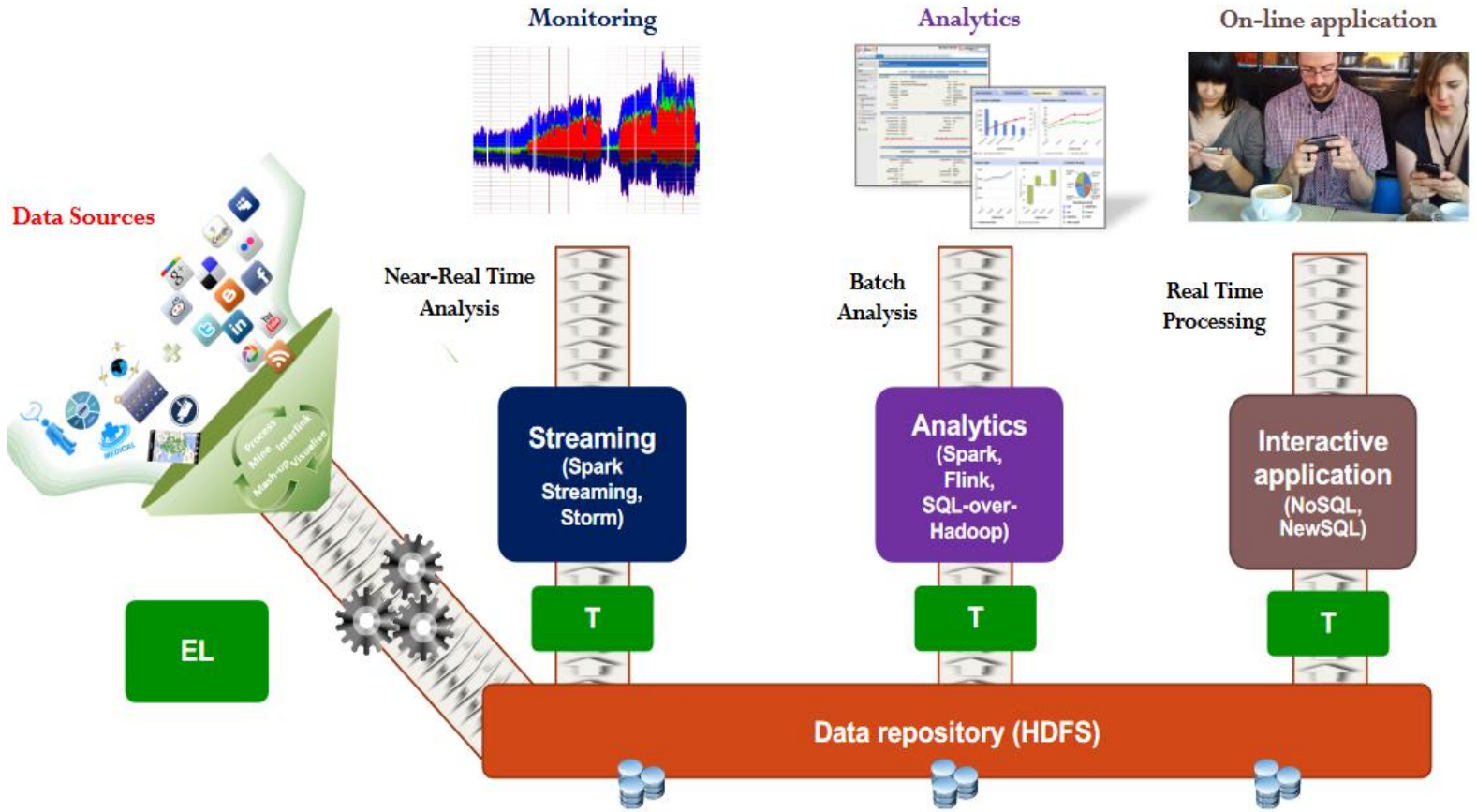
Interaktywny (czas rzeczywisty)

- Przetwarzanie danych i szybkie zwracanie wyników, aby wpłynąć na środowisko w tym czasie (e-commerce, wyszukiwarki, rezerwacje, ...)
- Odczyt/zapis we/wy, trwałe dane, (miękkie) transakcje ACID

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

A recent trend for Data Collection



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

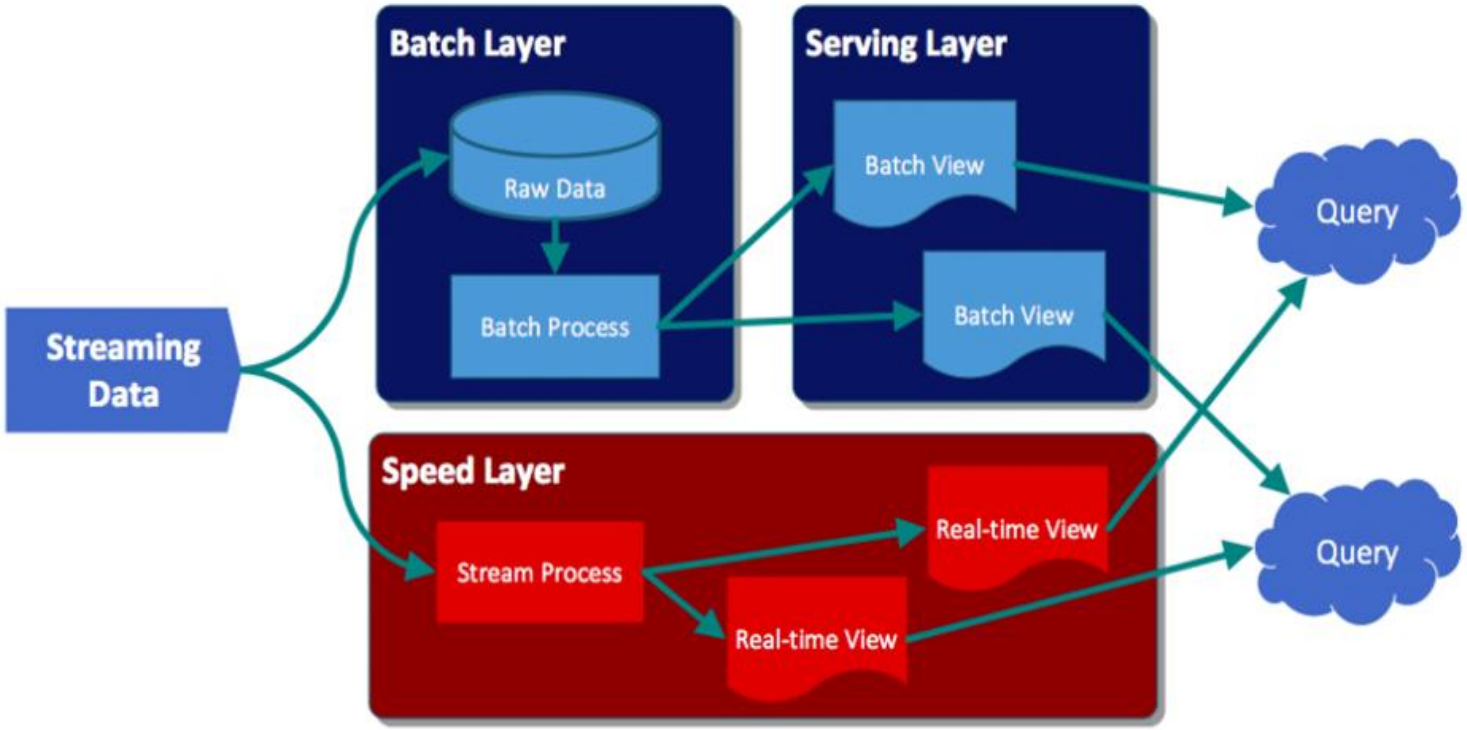
Architektura Lambda

- Nigdy nie zmieniaj/usuwaj danych, przechowuj oryginalne i przekształcone dane
- Rozróżnij warstwę szybkościową i wsadową
  - Warstwa szybkości: indeksuje widok wsadowy w celu uzyskania interaktywnego dostępu
  - Warstwa wsadowa: dzieli wszystkie dane na widoki wsadowe
  - Warstwa obsługująca: aktualizacje z dużą częstotliwością/najnowsze dane
- Na każde zapytanie można odpowiedzieć za pomocą połączenia usługi i warstwy szybkości

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura Lambda



iBigWorld:2020-1-PL01-KA203-082197

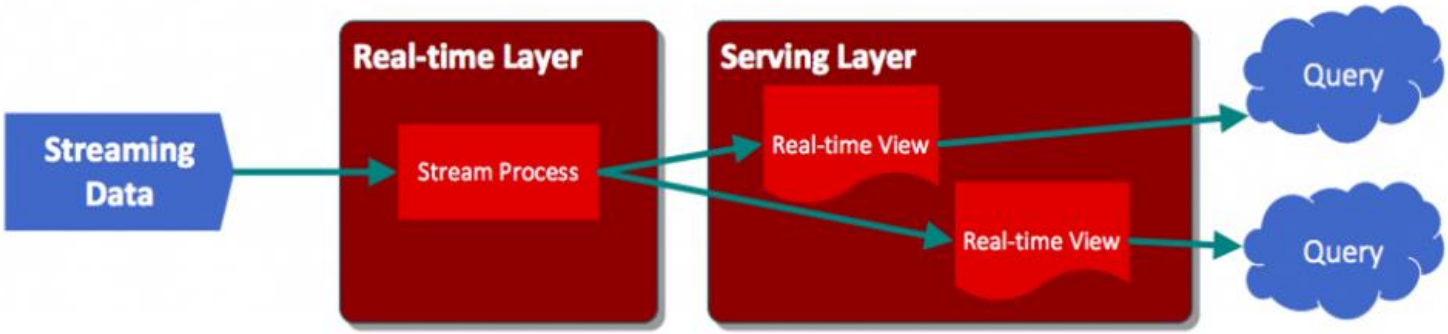
Co-funded by the Erasmus+ Programme of the European Union

Architektura Kappa

Uproszczenie architektury Lambda z usuniętą warstwą przetwarzania wsadowego.

Z dziennika dane są przesyłane strumieniowo przez system obliczeniowy i wprowadzane do magazynów pomocniczych w celu ich obsługi.

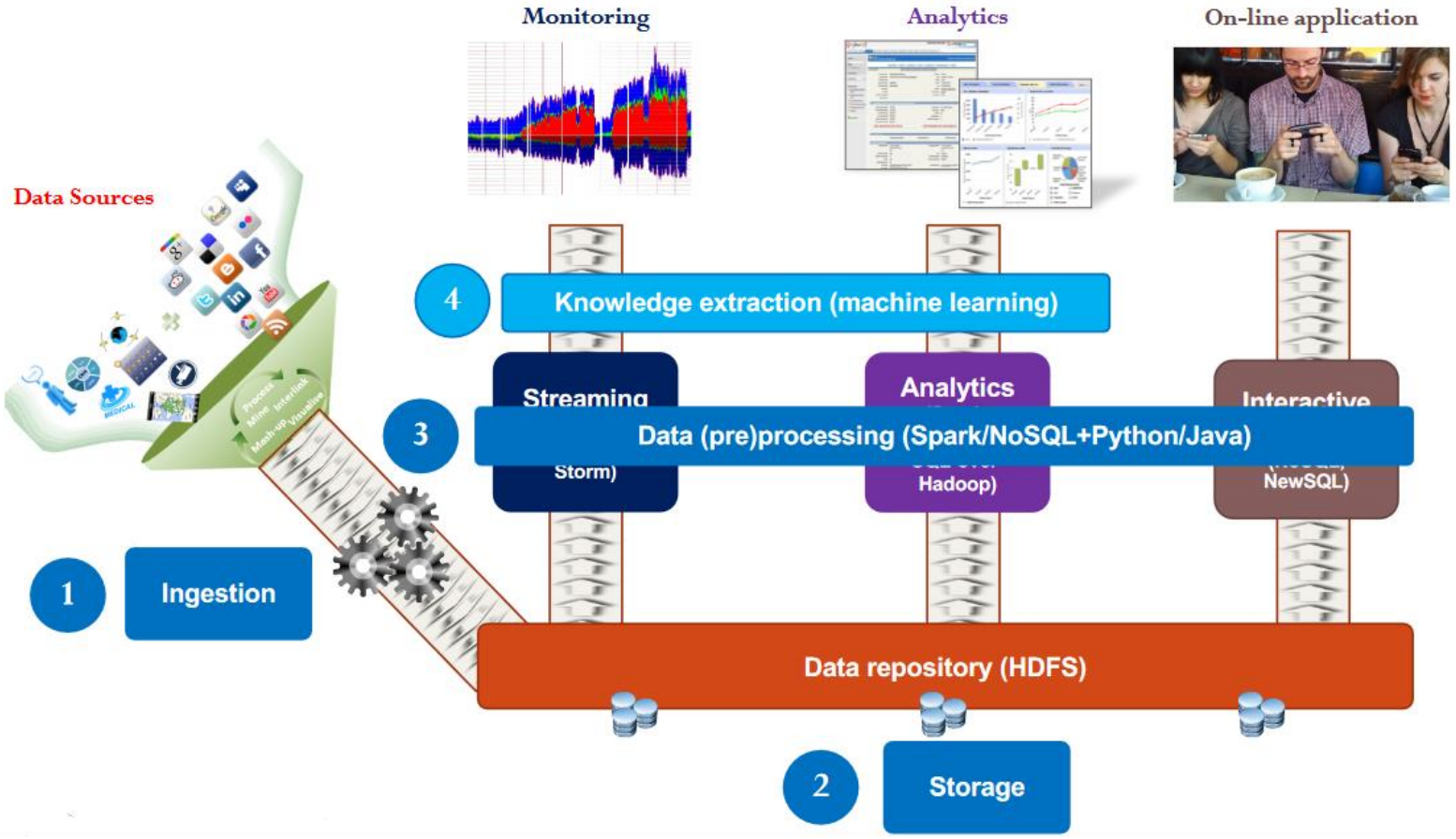
Duże dane są szybko przesyłane przez system przesyłania strumieniowego.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

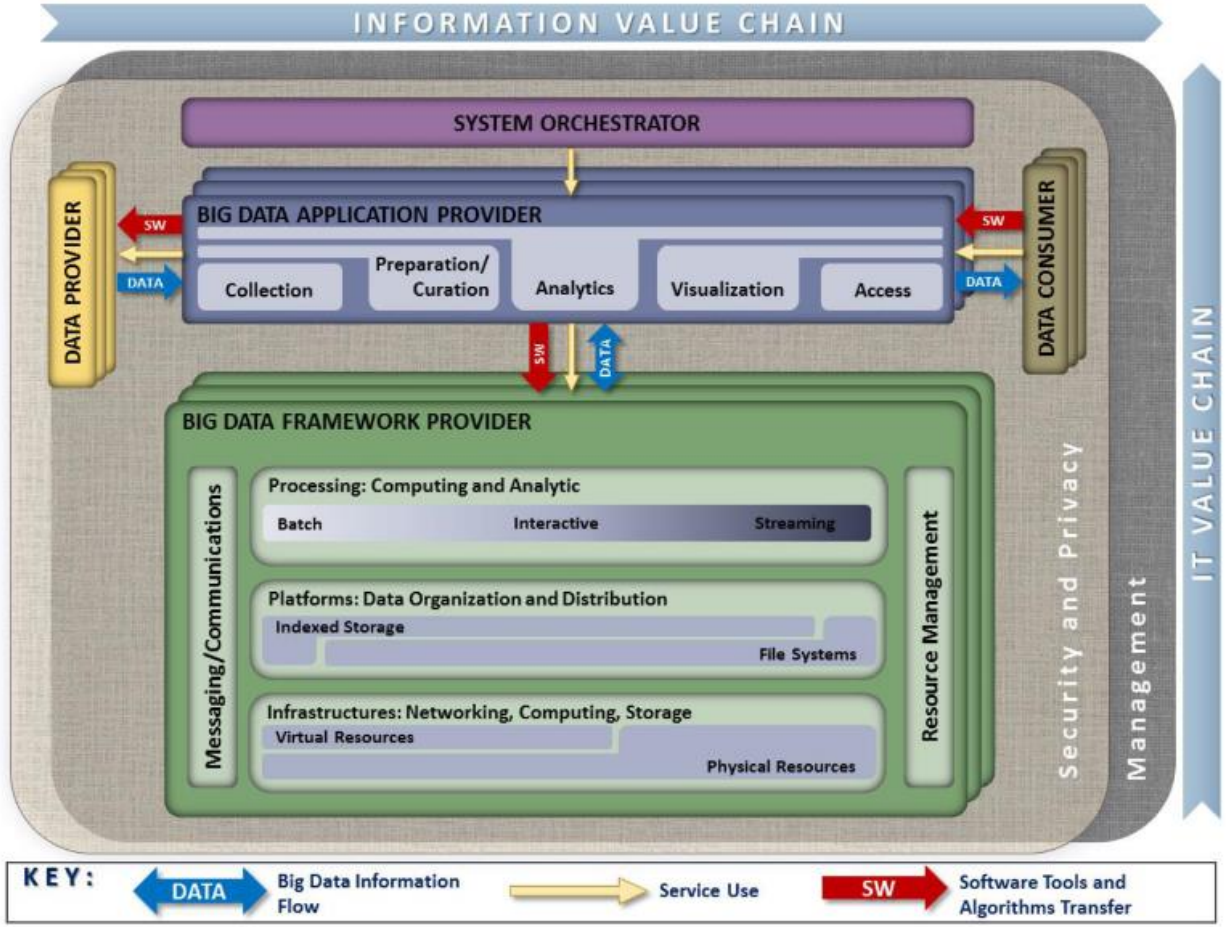
Stos oprogramowania Big Data



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura referencyjna dużych zbiorów danych NIST



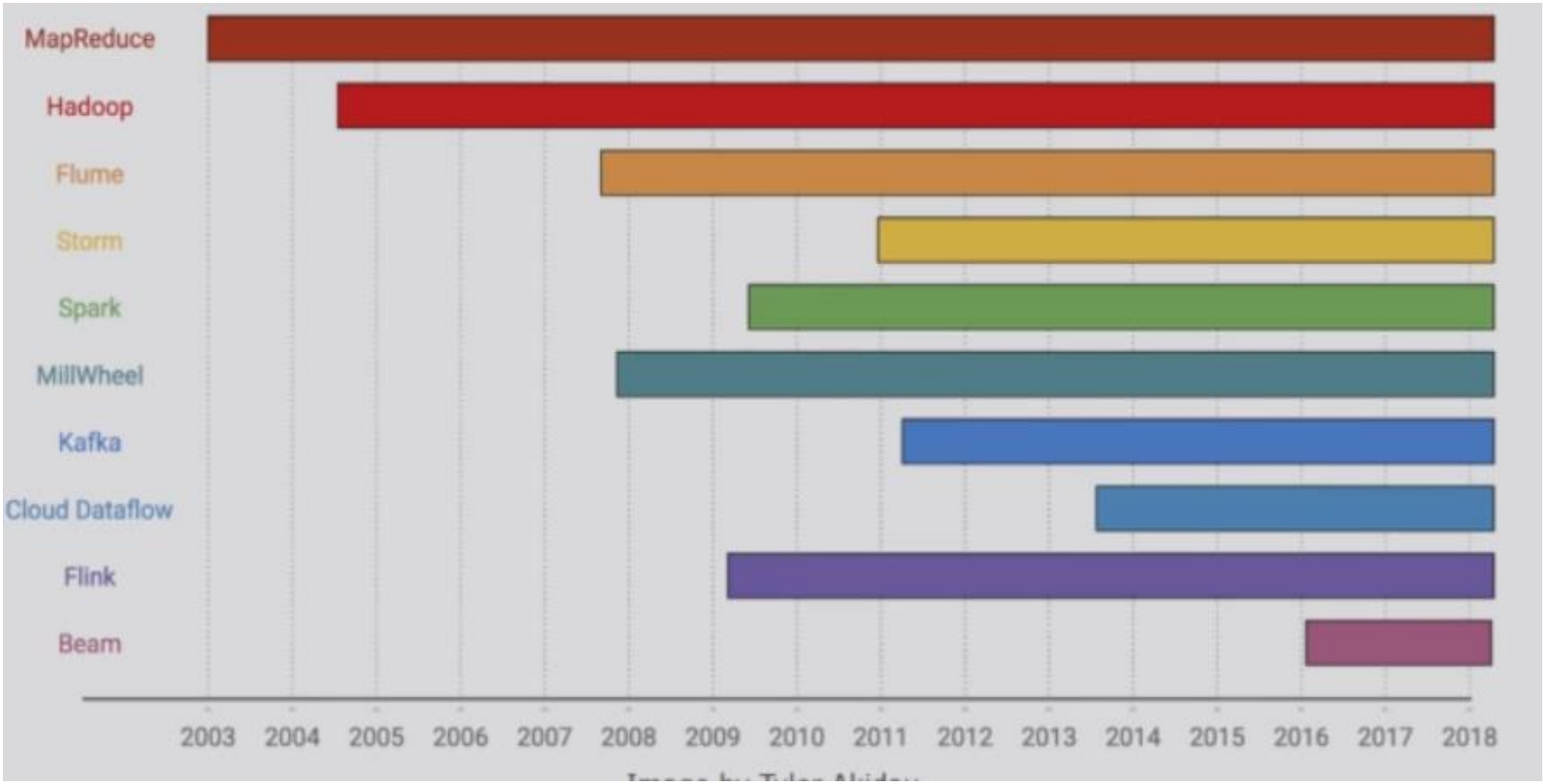
https://doi.org/10.6028/NIST.SP.1500-6r2

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Ewolucja systemu Big Data

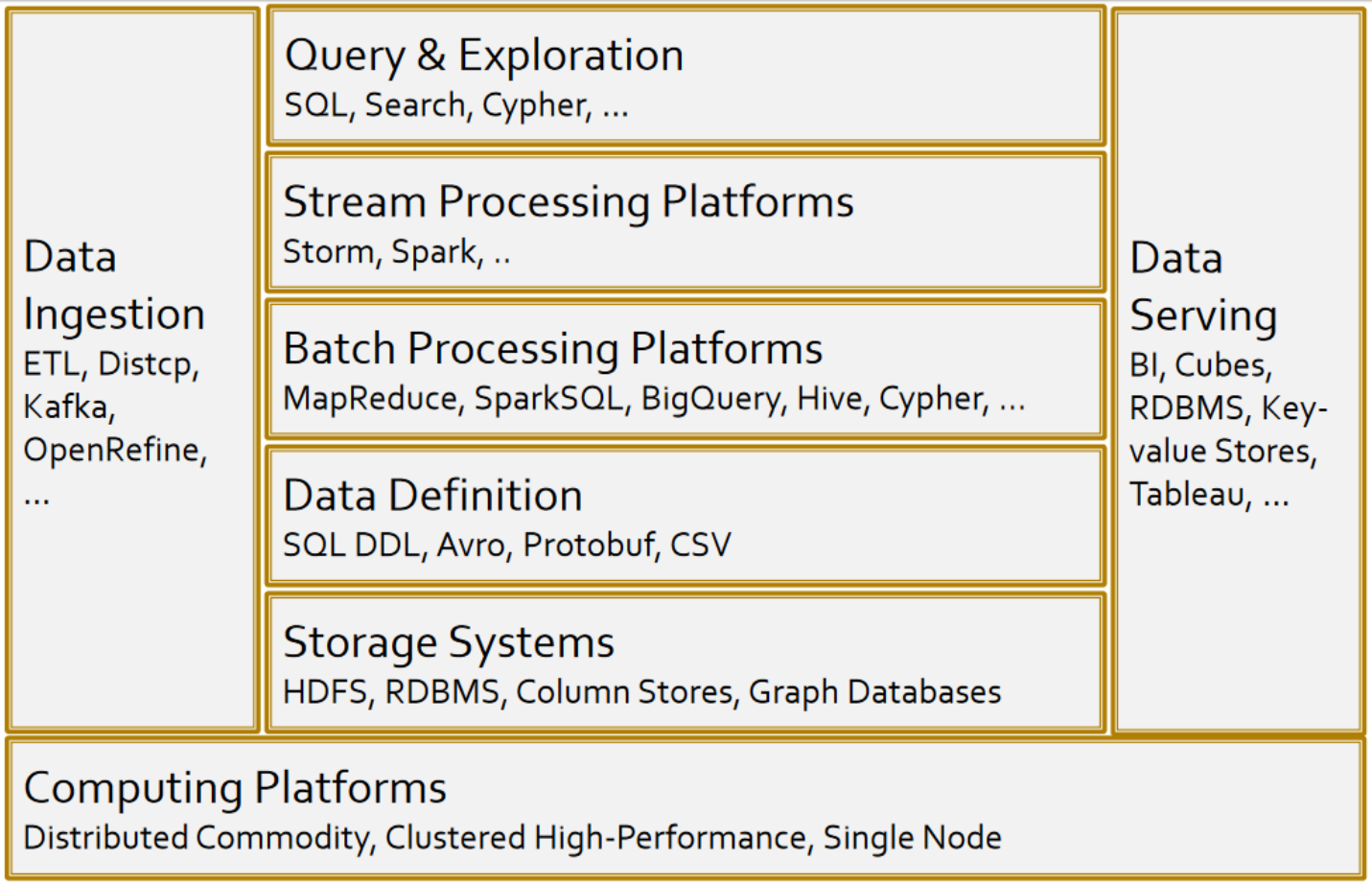
Z punktu widzenia Google



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

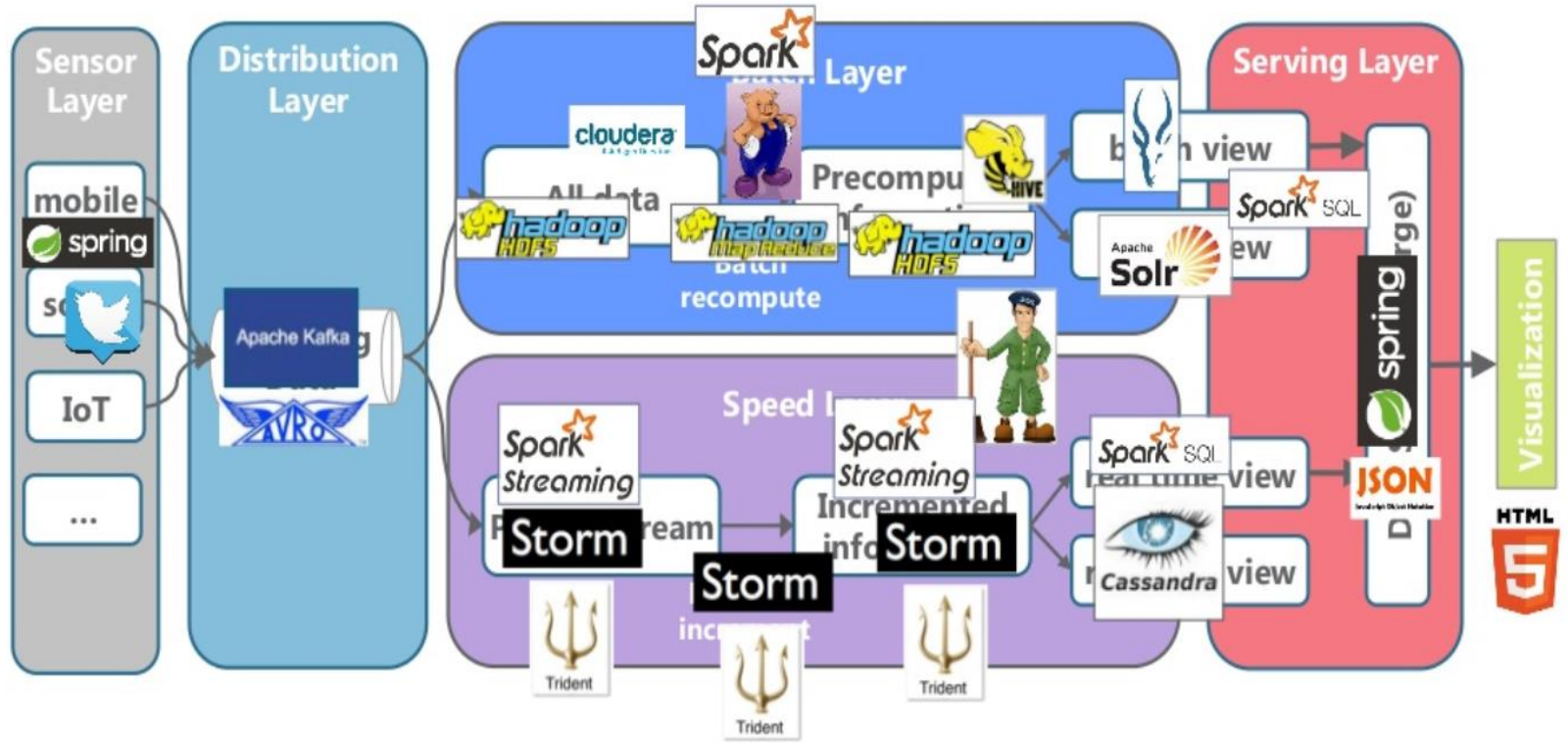
Ecosystem Big Data



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

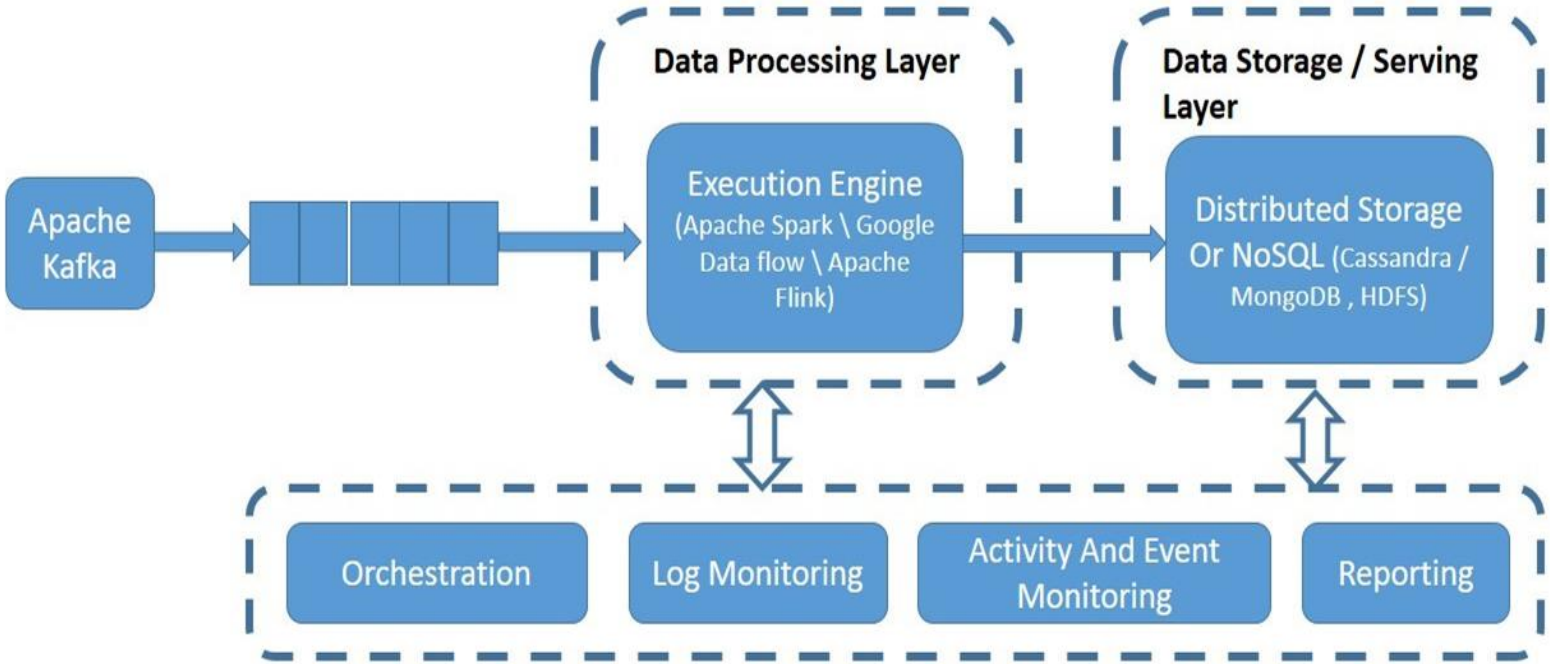
Stos technologii Big Data w architekturze Lambda



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Stos technologii Big Data w architekturze Kappa



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

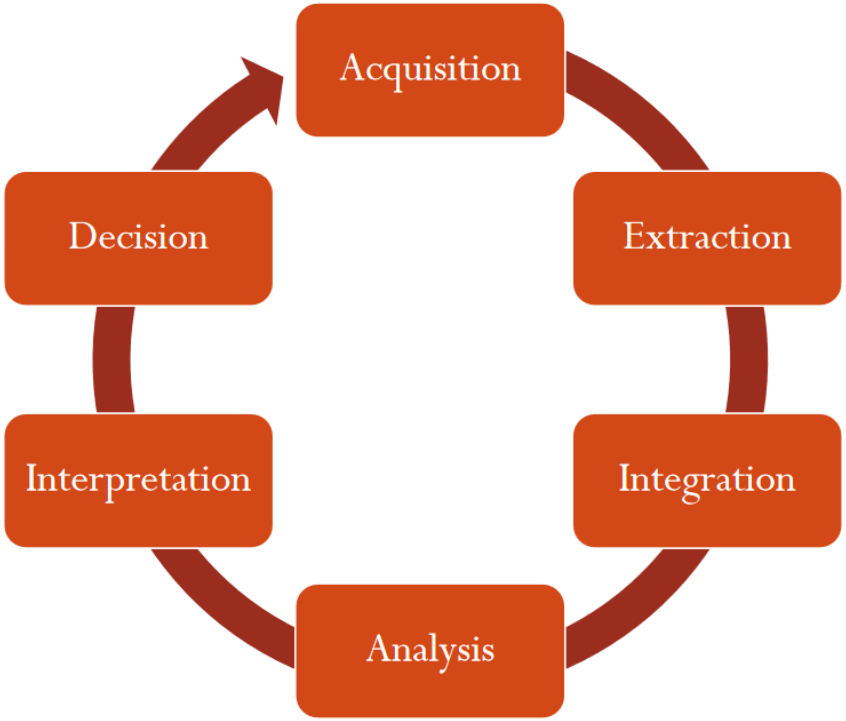
Potok Big Data

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Proces Big Data

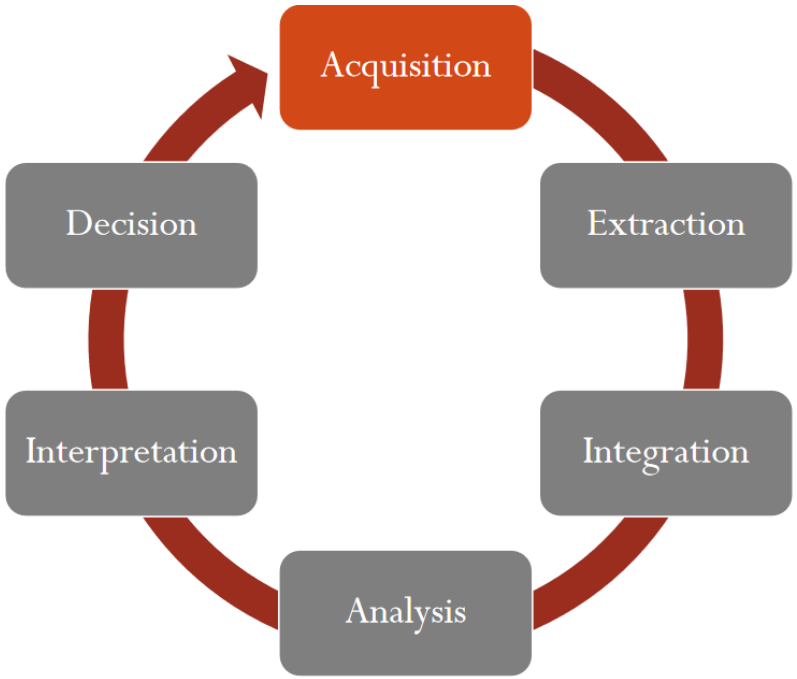
Cel: Podejmowanie skutecznych decyzji strategicznych z wykorzystaniem dostępności Big Data



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

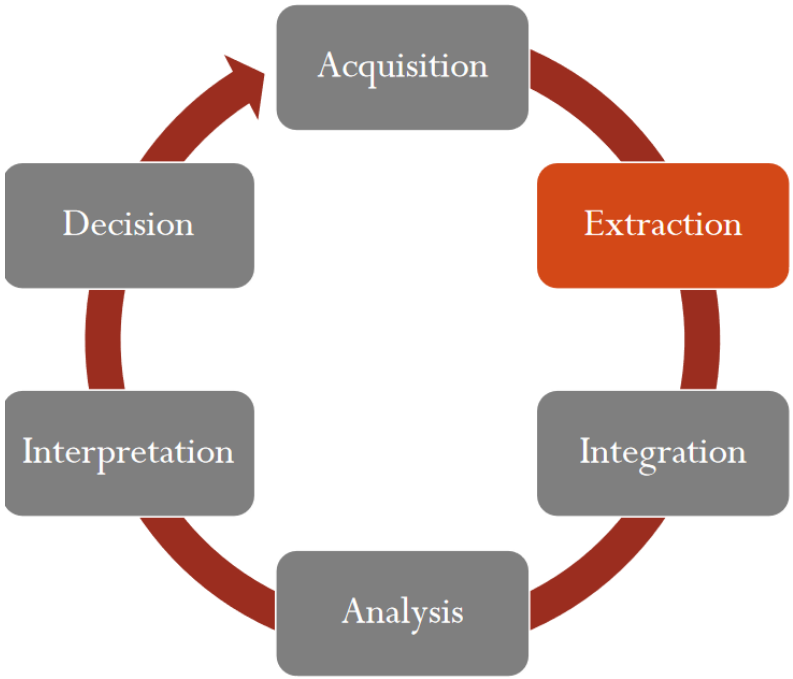


- Wymaga:
  - Wyboru
  - Filtrowania
  - Generowania metadanych
  - Zarządzania pochodzeniem

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

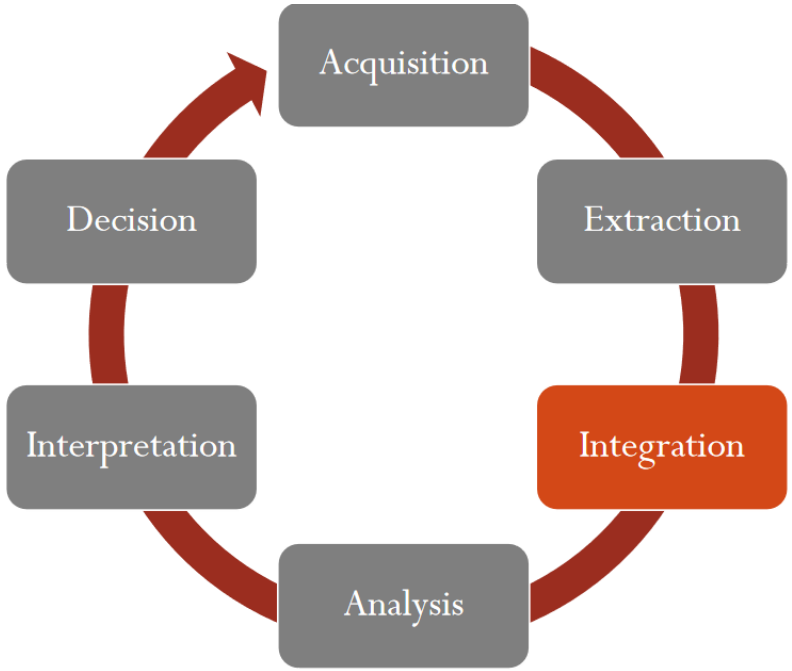


- Wymaga:
  - transformacji
  - normalizacji
  - czyszczenia
  - agregacji
  - obsługi błędów

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

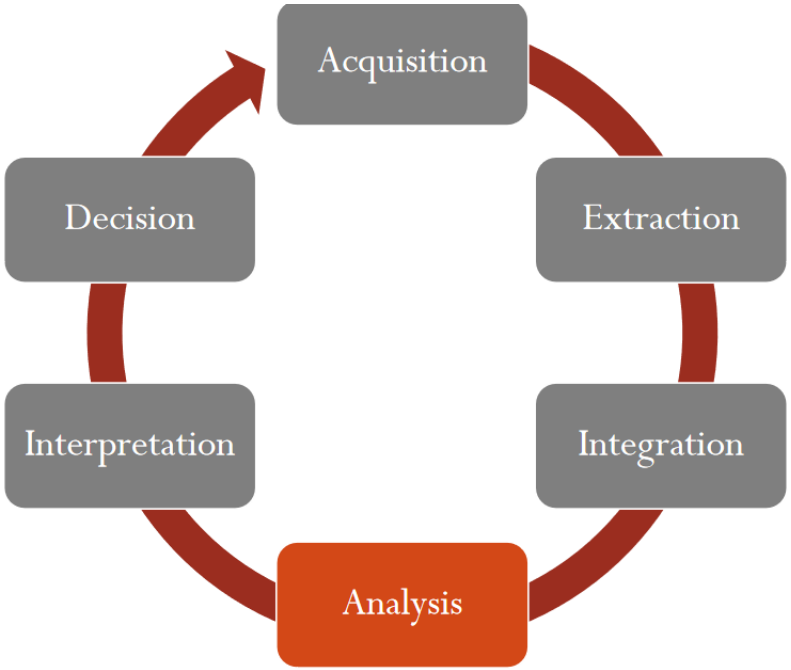


- Wymaga:
  - standaryzacji
  - zarządzania konfliktem
  - rekoncyliacja
  - definicja odwzorowania

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

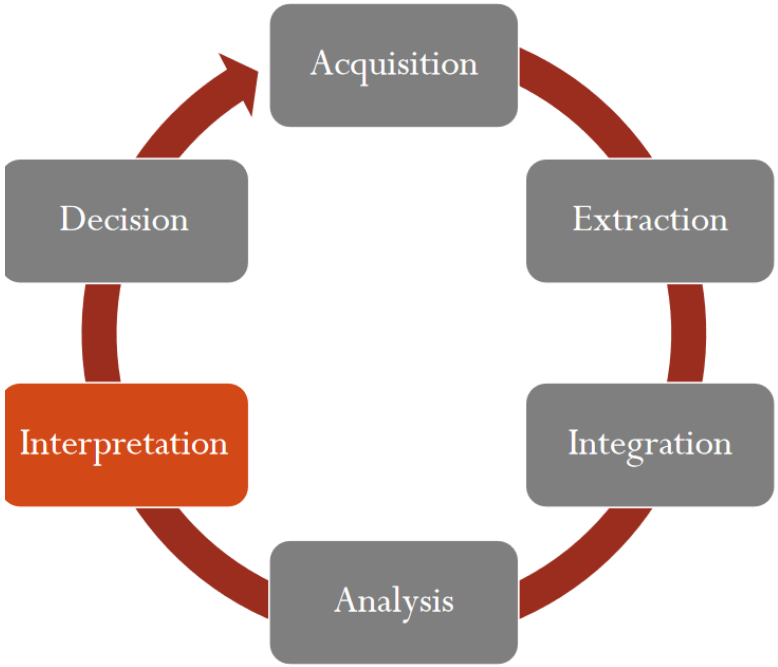


- Wymaga:
  - badania
  - przetwarzania
  - eksploracji danych
  - uczenia maszynowego
  - wizualizacji

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

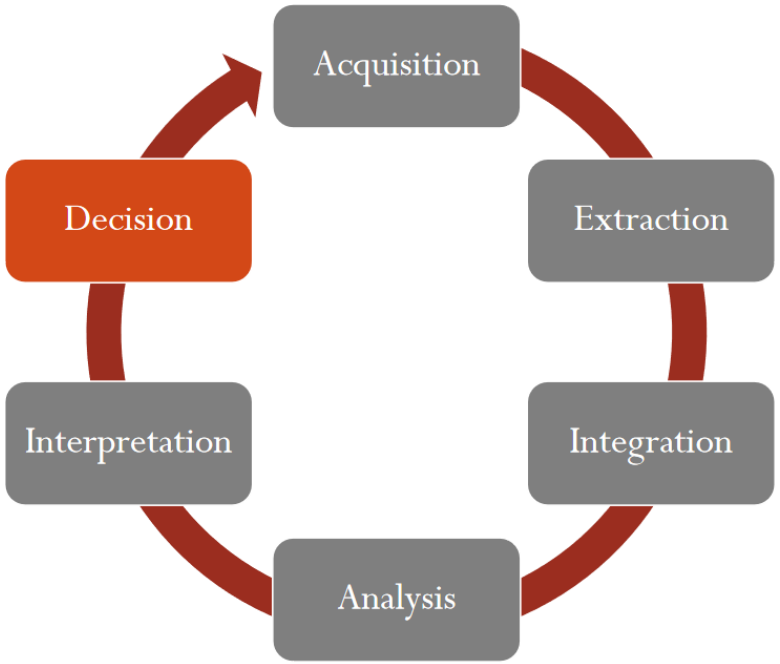


- Wymaga:
  - znajomości domeny
  - znajomości pochodzenia
  - Identyfikacji wzorców
  - elastyczności procesu

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data



- Wymaga:
  - zdolności kierowniczych
  - ciągłego doskonalenia procesu

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Potok danych

Potok danych łączy w całość operację polegającą na gromadzeniu danych, przekształcaniu ich w widoki, trenowaniu modelu, dostarczaniu wglądów, stosowaniu modelu zawsze i wszędzie tam, gdzie trzeba podjąć działania, aby osiągnąć cel biznesowy.

- A system for moving data from one system to another.
- Encompasses ETL as a subsystem
- Transformation of data is optional
- May be processed in real-time or in batch manner

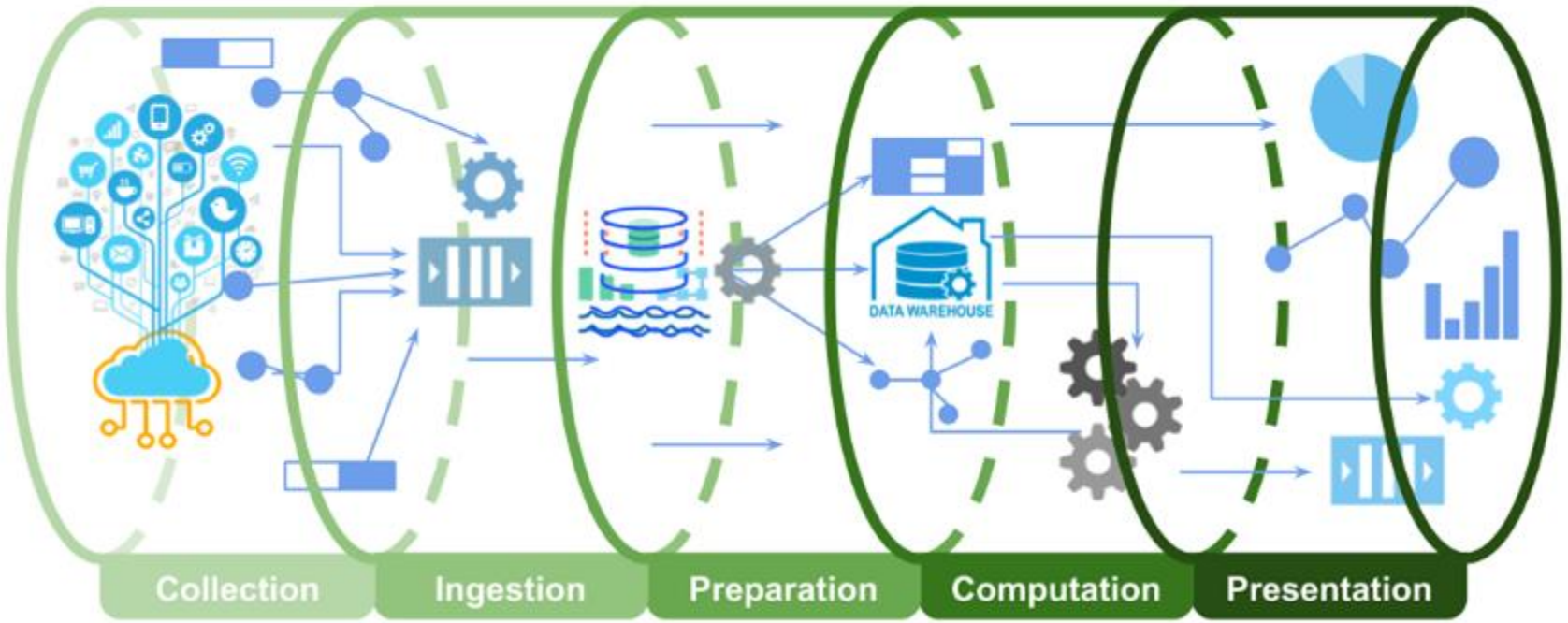
A data pipeline three heads:

- Data Engineering: collection, ingestion, preparation
- Processing / Analytics / Machine Learning: computation
- Delivery: visualization

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Etapy w potoku Big Data



https://towardsdatascience.com/scalable-performance-big-data-analytics-machine-learning-pipeline-architecture-on-cloud

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Etapy potoku danych

- **Gromadzenie danych**: Źródła danych (aplikacje mobilne, strony internetowe, aplikacje internetowe, mikrousługi, urządzenia IoT itp.) są oprzyrządowane do gromadzenia odpowiednich danych.
- **Przyjmowanie danych**: Oprzyrządowane źródła przesyłają dane do różnych punktów wejściowych (HTTP, MQTT, kolejka komunikatów itp.). Mogą również istnieć zadania importowania danych z usług takich jak Google Analytics. Dane mogą mieć dwie formy: pojedyncze zbiory i strumienie. Wszystkie te dane są gromadzone w Data Lake.
- **Przygotowanie**: Jest to operacja wyodrębniania, przekształcania, ładowania (ETL) w celu oczyszczenia, dostosowania, ukształtowania, przekształcenia i skatalogowania obiektów blob danych i strumieni w jeziorze danych; przygotowanie danych do wykorzystania na potrzeby uczenia maszynowego i przechowywanie ich w hurtowni danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Etapy potoku danych

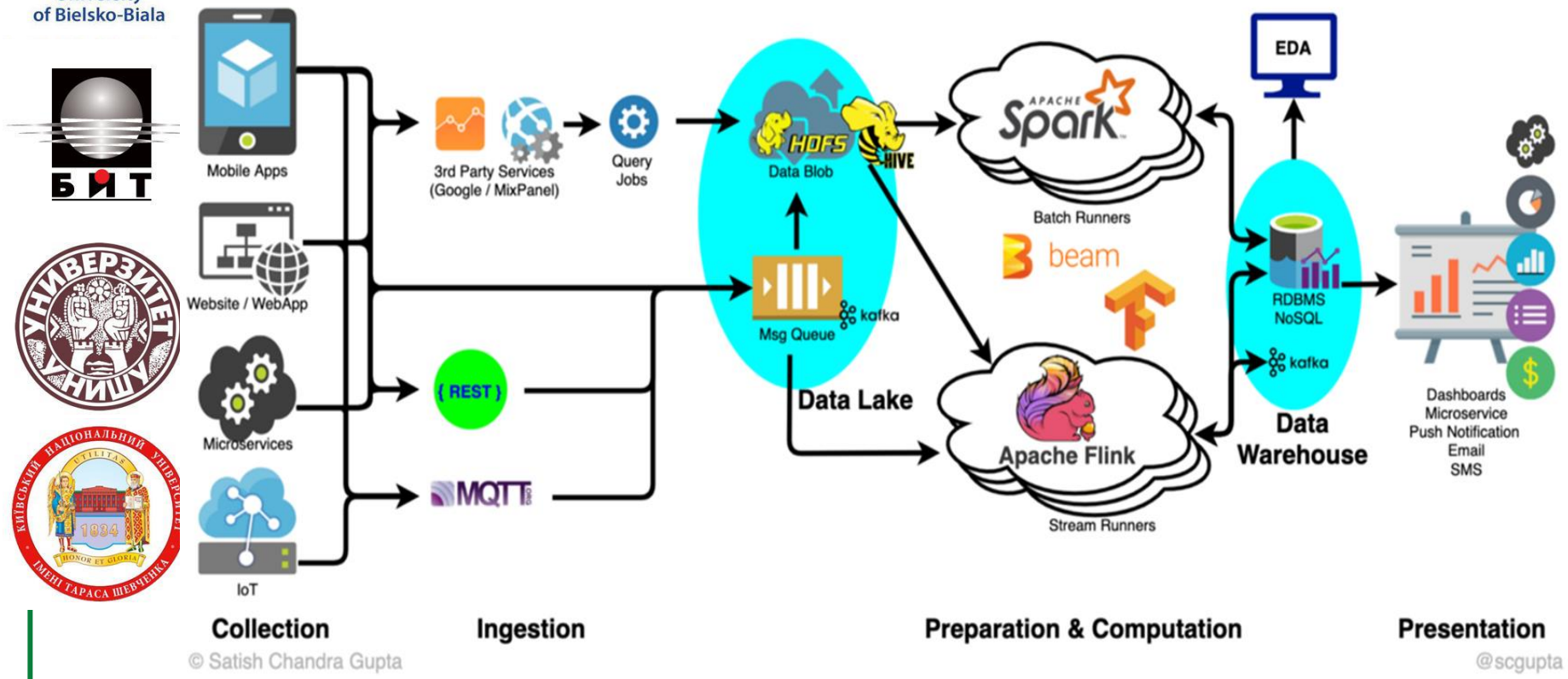
Obliczenie: To tutaj odbywają się analizy, nauka o danych i uczenie maszynowe. Obliczenia mogą być kombinacją przetwarzania wsadowego i strumieniowego. Modele i spostrzeżenia (zarówno dane strukturalne, jak i strumienie) są przechowywane z powrotem w hurtowni danych.

Prezentacja: Informacje są dostarczane za pośrednictwem pulpitów nawigacyjnych, e-maili, SMS-ów, powiadomień push i mikrousług. Wnioski modelu ML są udostępniane jako mikrousługi.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Implementacja potoku Big Data z wykorzystaniem technologii open source



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Potok Big Data – Kluczowe komponenty i technologie

- **Punkty końcowe HTTP / MQTT** do pozyskiwania danych, a także do udostępniania wyników. Istnieje kilka platform i technologii do tego.
- **Kolejka wiadomości Pub/Sub** do pozyskiwania dużych ilości przesyłanych strumieniowo danych. Kafka jest obecnie de facto wyborem. Skalowanie do wysokiego współczynnika pozyskiwania zdarzeń zostało sprawdzone w boju.
- **Niedrogi magazyn danych o dużej objętości** dla data lake (i hurtowni danych),Hadoop HDFSlub przechowywanie obiektów typu blob w chmurze, np. AWS S3.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Potok Big Data – Kluczowe komponenty i technologie

- **Infrastruktura zapytań i katalogów** dla przekształcenia jeziora danych w hurtownię danych, Ul Apachejest popularnym wyborem języka zapytań.
- **Map-Reduce Batch Compate** silnik do przetwarzania wysokoprzepustowego, np. Hadoop Map-Reduce, Apache Spark.
- **Strumień obliczeniowy** do przetwarzania danych wrażliwych o na opóźnienia, np. Apache Storm, Apache Flink. Apache Beam wyłania się jako wybór do zapisywania obliczeń przepływu danych. Można go wdrożyć na urządzeniu do uruchamiania wsadowego Spark lub urządzeniu do uruchamiania strumieni Flink.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Rurociąg Big Data – Kluczowe komponenty i technologie

- **Platformy uczenia maszynowego** dla nauki o danych i uczenia maszynowego. Scikit-Learn,TensorFlow, oraz PyTorch są popularnymi wyborami do implementacji uczenia maszynowego.
- **Magazyny danych o niskim opóźnieniu**do przechowywania wyników. Istnieje wiele ugruntowanych SQL vs. NoSQLwybór magazynów danych w zależności od typu danych i przypadku użycia.

Orkiestracja wdrażaniaopcje Hadoop YARN ,Kubernetesa/Kubeflow.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Potok Big Data – Skala i efektywność

- **Wydajność** zależy od skalowalności pozyskiwania (tj. punktów końcowych REST/MQTT i kolejki komunikatów), pojemności magazynu data lake i przetwarzania wsadowego z redukcją map.
- **Czas oczekiwania** zależy od wydajności kolejki komunikatów, strumienia obliczeniowego i baz danych służących do przechowywania wyników obliczeń.
- **Niezawodność** potoku danych wymaga, aby poszczególne systemy w potoku danych były odporne na awarie. Niezawodny potok danych z wbudowanymi mechanizmami audytu, rejestrowania i sprawdzania poprawności pomaga zapewnić jakość danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

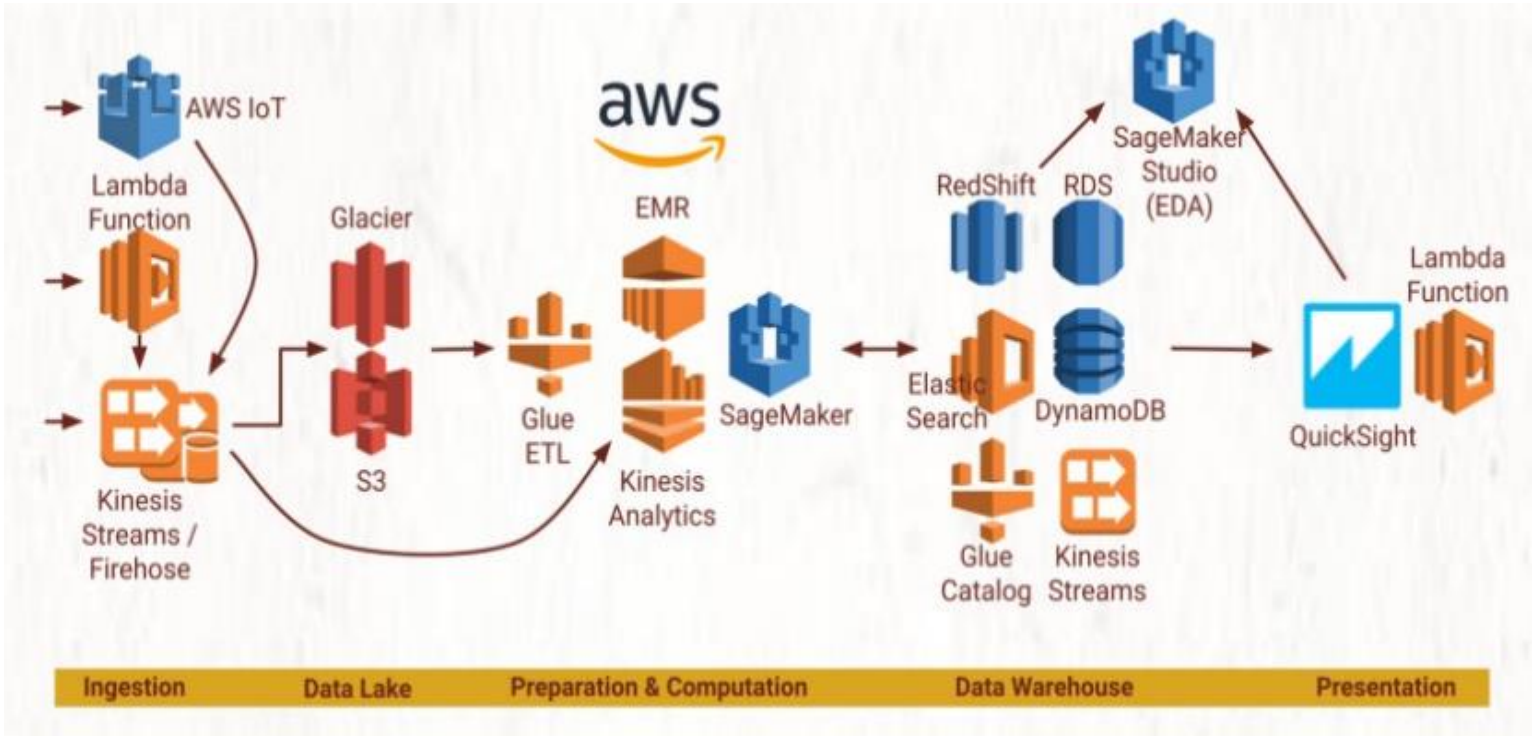
Potok Big Data działa bezserwerowo

- Z nadejściem **przetwarzania bezserwerowego**, można szybko zacząć, unikając DevOps.
- Różne komponenty w architekturze można zastąpić ich bezserwerowymi odpowiednikami od wybranego dostawcy usług w chmurze.
- Typowe bezserwerowe architektury potoków Big Data na platformach chmurowych Amazon Web Services, Microsoft Azure i Google Cloud Platform (GCP)
- Każdy z nich ściśle odwzorowuje ogólną architekturę dużych zbiorów danych omówioną w poprzedniej sekcji.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura potoku Big Data w Amazon Web Services (AWS

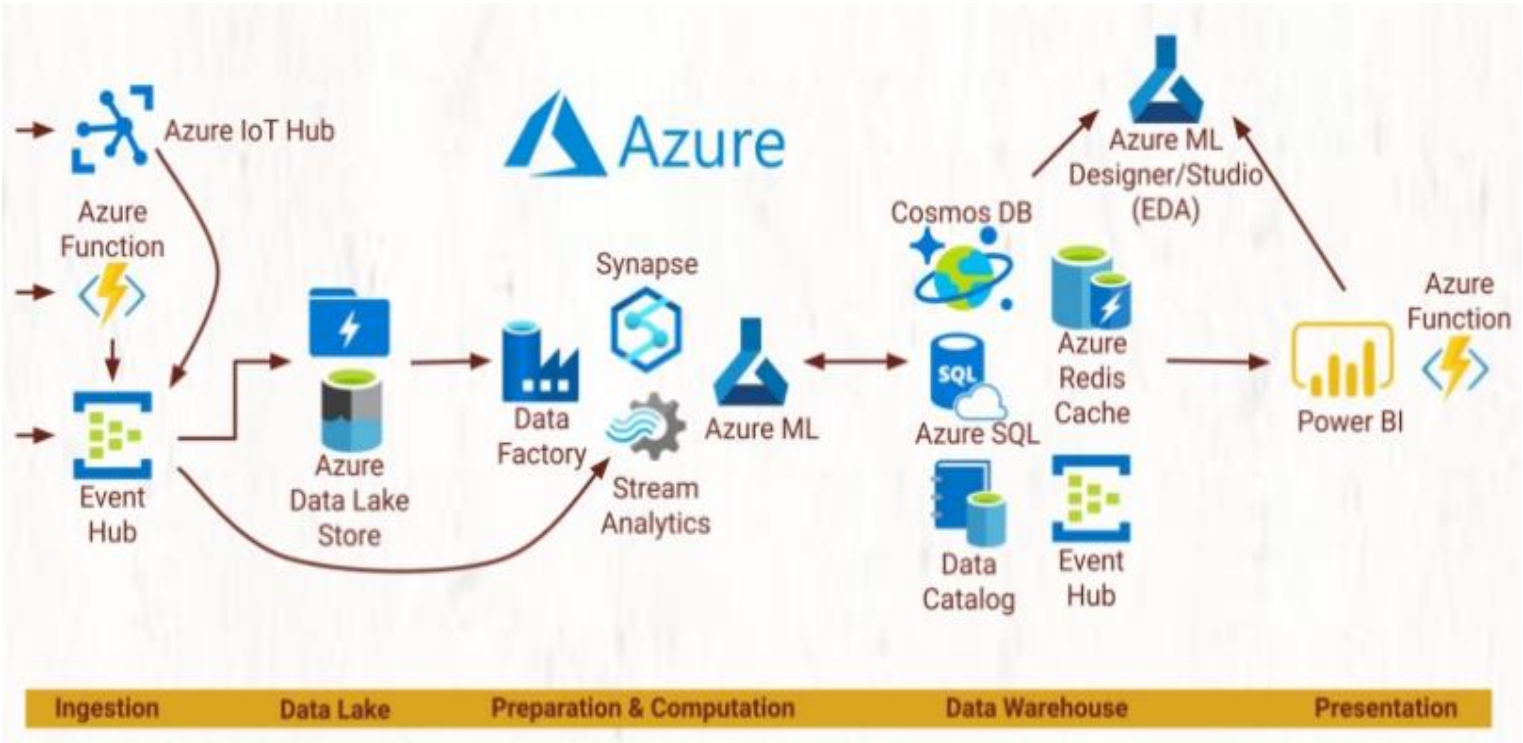


http://scgupta.link/big-data-pipeline

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura potoku Big Data na platformie Microsoft Azure

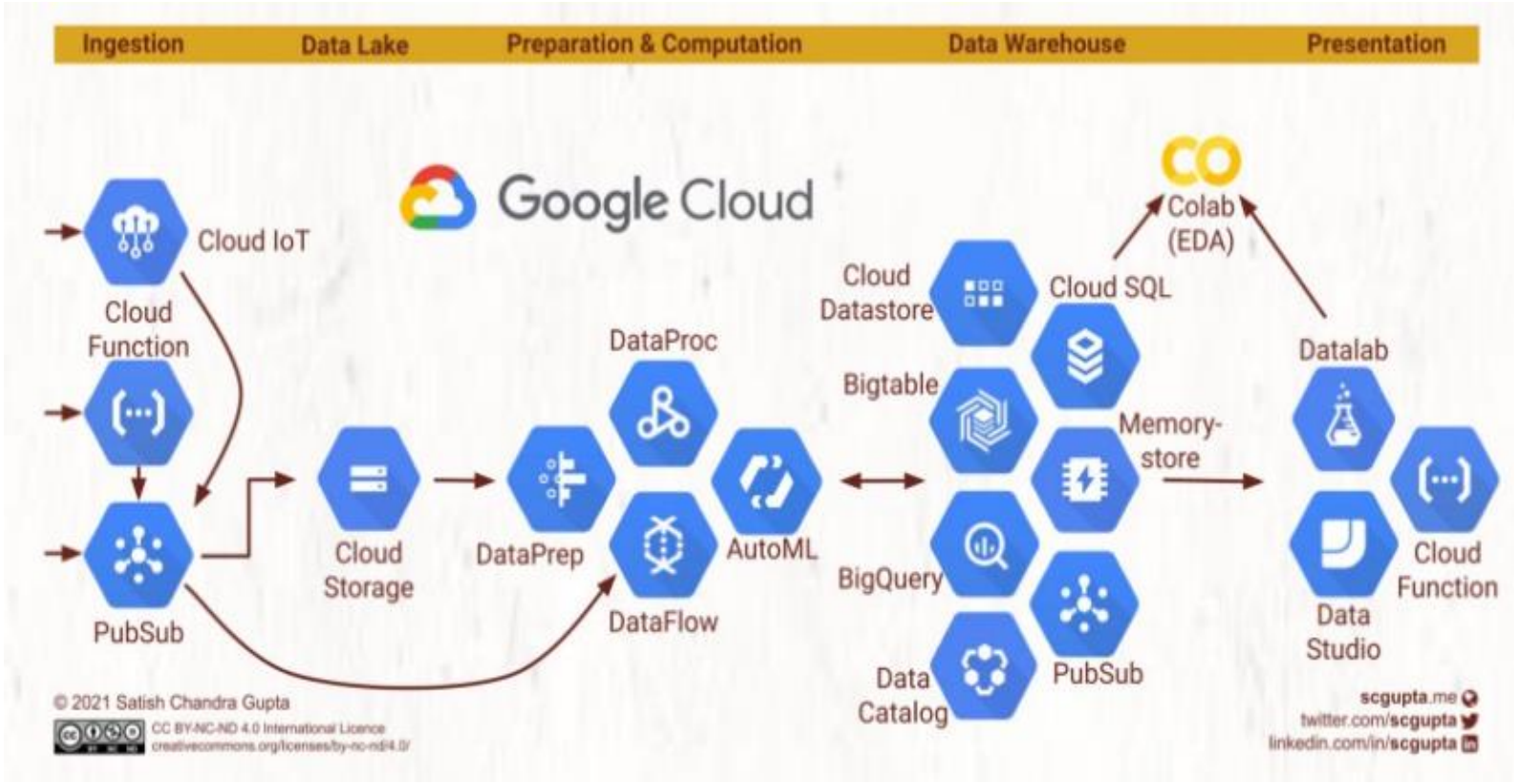


http://scgupta.link/big-data-pipeline

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura potoku Big Data w Google Cloud Platform (GCP)



http://scgupta.link/big-data-pipeline

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Zbieranie i przetwarzanie Big Data

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Zbieranie danych

- Proces gromadzenia, filtrowania i czyszczenia danych
- Jako pierwsza warstwa potoku danych, źródła danych są kluczem do jego zaprojektowania. Bez wysokiej jakości danych nie ma nic do przetworzenia i przeniesienia przez potok.
- Dane mogą być jako:
  - Tekst
  - Audio
  - Wideo
- 5V danych:
  - Objętość (rozmiar danych), Szybkość (jak szybko dane są generowane?), Różnorodność (dane ustrukturyzowane, częściowo ustrukturyzowane, nieustrukturyzowane), Prawdziwość (nieporządek, jakość i dokładność?), Wartość

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Migracja danych

- Przesyłanie danych z jednego systemu pamięci komputera do innego. np :
  - Przesyłanie zdjęć ze smartfona do laptopa
  - Przenoszenie danych ze starego laptopa na nowy
  - Przesyłanie danych z Dysku Google do Dropbox
- Proces selekcji, przygotowania, wydobywania, przekształcania i przenoszenia danych
- Zwykle zaangażowane są tysiące źródeł danych
- Wygenerowane dane mają niewielki rozmiar
- Wyższa częstotliwość generowania danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Migracja danych — wyzwania i zagrożenia

- Utrata danych
- Problemy ze zgodnością:
  - Kompatybilność pamięci masowej
  - Kompatybilność aplikacji
  - Kompatybilność platformy (np. formularz on-premise do chmury),
  - Zgodność z chmurą
- 2 Szerokie kategorie
  - Online i offline
- Czynniki do rozważenia
  - Rodzaj obciążenia, ilość danych, szybkość realizacji

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Pochłanianie dużych ilości danych

- Składniki pozyskiwania potoku danych to procesy odczytujące dane ze źródeł danych
- Proces wyodrębniania odczytuje dane z każdego źródła przy użyciu interfejsów programowania aplikacji (API) udostępnianych przez źródło danych.
- Zanim jednak napiszesz kod, który wywołuje interfejsy API, musisz dowiedzieć się, jakie dane chcesz wyodrębnić w procesie zwanym profilowaniem danych — badając dane pod kątem ich cech i struktury oraz oceniając, jak dobrze pasują do celu biznesowego.
- Po profilowaniu danych są one pozyskiwane jako partie lub przez przesyłanie strumieniowe.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Pozyskiwanie wsadowe i pozyskiwanie strumieniowe

- **Przetwarzanie wsadowe** ma miejsce, gdy zestawy rekordów są wyodrębniane i obsługiwane jako grupa. Przetwarzanie wsadowe jest sekwencyjne, a mechanizm pozyskiwania odczytuje, przetwarza i generuje grupy rekordów zgodnie z kryteriami ustalonymi wcześniej przez programistów i analityków. Proces nie obserwuje nowych rekordów i nie przenosi ich w czasie rzeczywistym, ale zamiast tego działa zgodnie z harmonogramem lub działa w oparciu o zewnętrzne wyzwalacze.
- **Przesyłanie strumieniowe** to alternatywny paradygmat pozyskiwania danych, w którym źródła danych automatycznie przekazują poszczególne rekordy lub jednostki informacji jeden po drugim. Wszystkie organizacje używają przetwarzania wsadowego dla wielu rodzajów danych, podczas gdy przedsiębiorstwa używają przetwarzania strumieniowego tylko wtedy, gdy potrzebują danych w czasie zbliżonym do rzeczywistego do użytku z aplikacjami lub analizami, które wymagają minimalnego możliwego opóźnienia.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przygotowanie i przechowywanie Big Data

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Transformacja Big Data

- Po wydobyciu danych z systemów źródłowych może zaistnieć potrzeba dostosowania ich struktury lub formatu.
- Transformacje obejmują mapowanie zakodowanych wartości na bardziej opisowe, filtrowanie i agregację.
- Czas wszelkich przekształceń zależy od tego, jakiego procesu replikacji danych przedsiębiorstwo zdecyduje się użyć w swoim potoku danych: **ETL**(wyodrębnij, przekształć, załaduj) lub **ELT**(wyodrębnij, załaduj, przekształć). ETL, starsza technologia używana w lokalnych hurtowniach danych, może przekształcać dane, zanim zostaną załadowane do miejsca docelowego. ELT, używany z nowoczesnymi hurtowniami danych opartymi na chmurze, ładuje dane bez stosowania jakichkolwiek transformacji.
- Odbiorcy danych mogą następnie zastosować własne przekształcenia danych w hurtowni danych lub jeziorze danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przechowywanie dużych zbiorów danych

Baza danych NoSQL oraz **hurtownia danych** są głównymi miejscami docelowymi dla danych pozyskiwanych i przekształcanych przez potok.

Te wyspecjalizowane bazy danych zawierają wszystkie wyczyszczone, wzorcowe dane przedsiębiorstwa w scentralizowanej lokalizacji do wykorzystania w analityce, raportowaniu i analizie biznesowej przez analityków i kierownictwo.

Mniej ustrukturyzowane dane mogą napływać **jeziora danych**, gdzie analitycy danych i naukowcy zajmujący się danymi mogą uzyskiwać dostęp do ogromnych ilości bogatych i możliwych do wydobycia informacji.

Wreszcie przedsiębiorstwo może wprowadzać dane do narzędzia lub usługi analitycznej, które bezpośrednio akceptują źródła danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Rozproszony system plików - założenia

- Wybierz tańszy sprzęt zamiast „egzotycznego” sprzętu
  - Skaluj „w poziomie”, a nie „w pionie”
- Wysokie wskaźniki awaryjności komponentów
  - Niedrogie komponenty towarowe cały czas zawodzą
- „Skromna” liczba ogromnych plików
  - Pliki wielogigabajtowe są powszechne, nawet jeśli są nie zalecane
- W większości są dołączane pliki są jednokrotnego zapisu
  - Być może równolegle
- Duże odczyty strumieniowe zamiast losowego dostępu
  - Wysoka stała przepustowość przy niskim opóźnieniu

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Rozproszony system plików

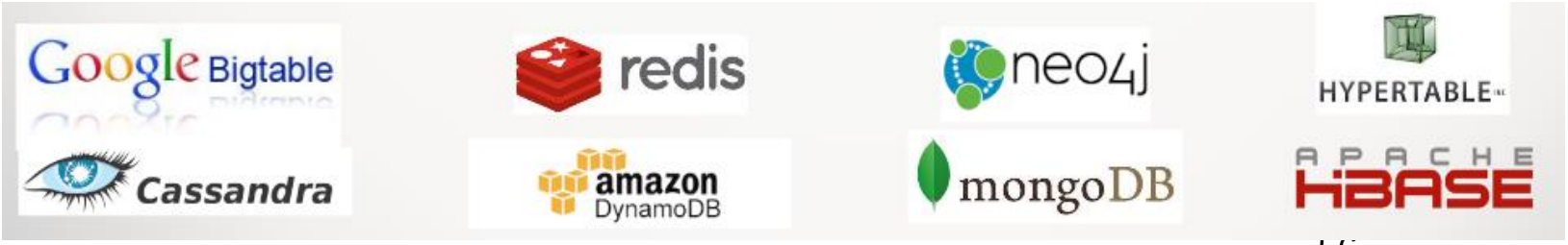
- Pliki przechowywane jako fragmenty
  - Blok o stałym rozmiarze (64 MB, 128 MB)
- Niezawodność dzięki replikacji
  - Każdy blok był replikowany w ponad 3 węzłach
- Pojedynczy master do koordynowania dostępu, zachowaj metadane
  - Proste scentralizowane zarządzanie
- Brak buforowania danych
  - Mała korzyść ze względu na duże zbiory danych, odczyty strumieniowe
- Uprość interfejs API
  - Przekaż niektóre problemy klientowi (np. układ danych)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Bazy danych NoSQL

- Połowa lat 90.: Trend obiektowo-relacyjnego modelu bazy danych
  - Niedopasowanie impedancji
- Połowa 2000 roku: Narodziny NoSQLRuch
  - Problem klastrów komputerowych: relacyjne bazy danych nie skalują się dobrze w poziomie
  - Wielcy gracze, tacy jak Google i Amazon, opracowali własne systemy pamięci masowej: narodziły się bazy danych NoSQL („Not-Only SQL”)
- Dzisiaj: Wiek NoSQL



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Bazy danych NoSQL

- Charakterystyka baz danych NoSQL
  - Skalowalność pozioma (przyjazność dla klastrów)
  - Nierelacyjne
  - Rozpowszechniane
  - Bez schematu
  - Open-source (przynajmniej większość systemów)

koncepcje

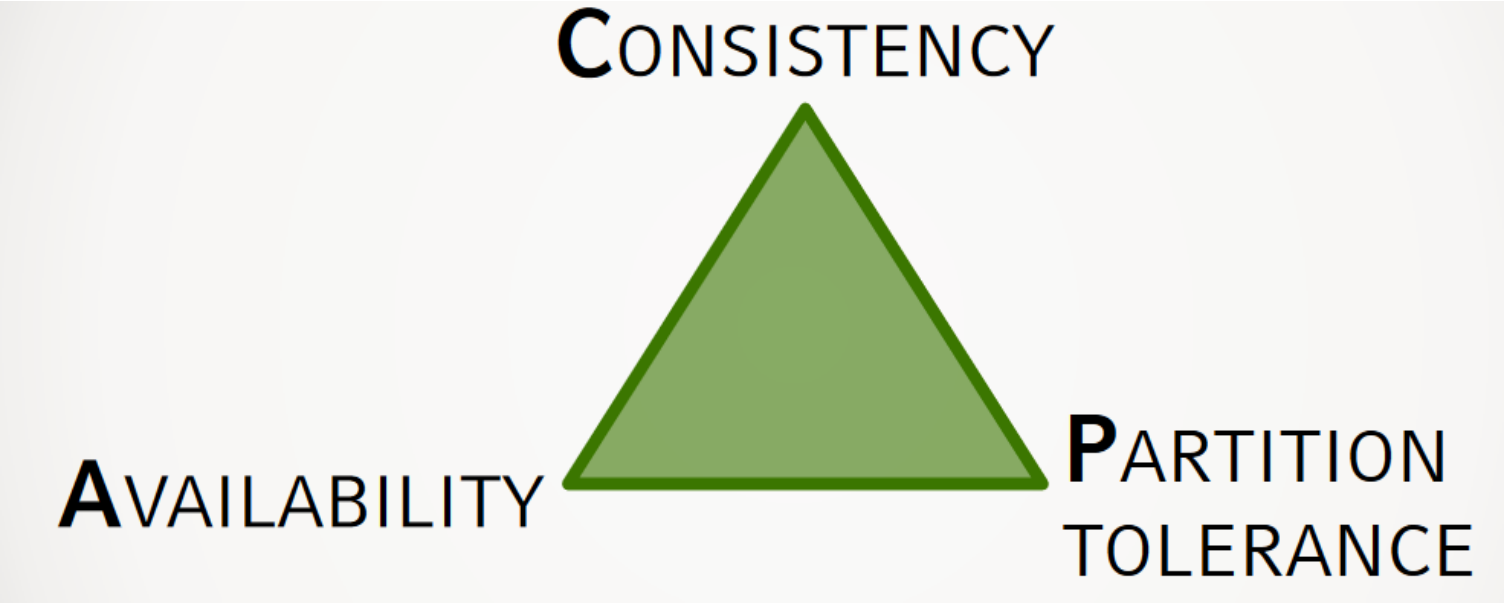
- ACID – Święty Graal RDBMS
- BASE – Sztuczna koncepcja baz danych NoSQL
  - Zasadniczo dostępne
  - Niekoniecznie spójny
  - Docelowa spójność
- BASE jest ustalany po przeciwnej stronie niż ACID, gdy rozważa się „spektrum spójności-dostępności”

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Twierdzenie CAP - Brewer

Każdy sieciowy system współdzielonych danych może mieć co najwyżej dwie z trzech pożądanych właściwości!



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

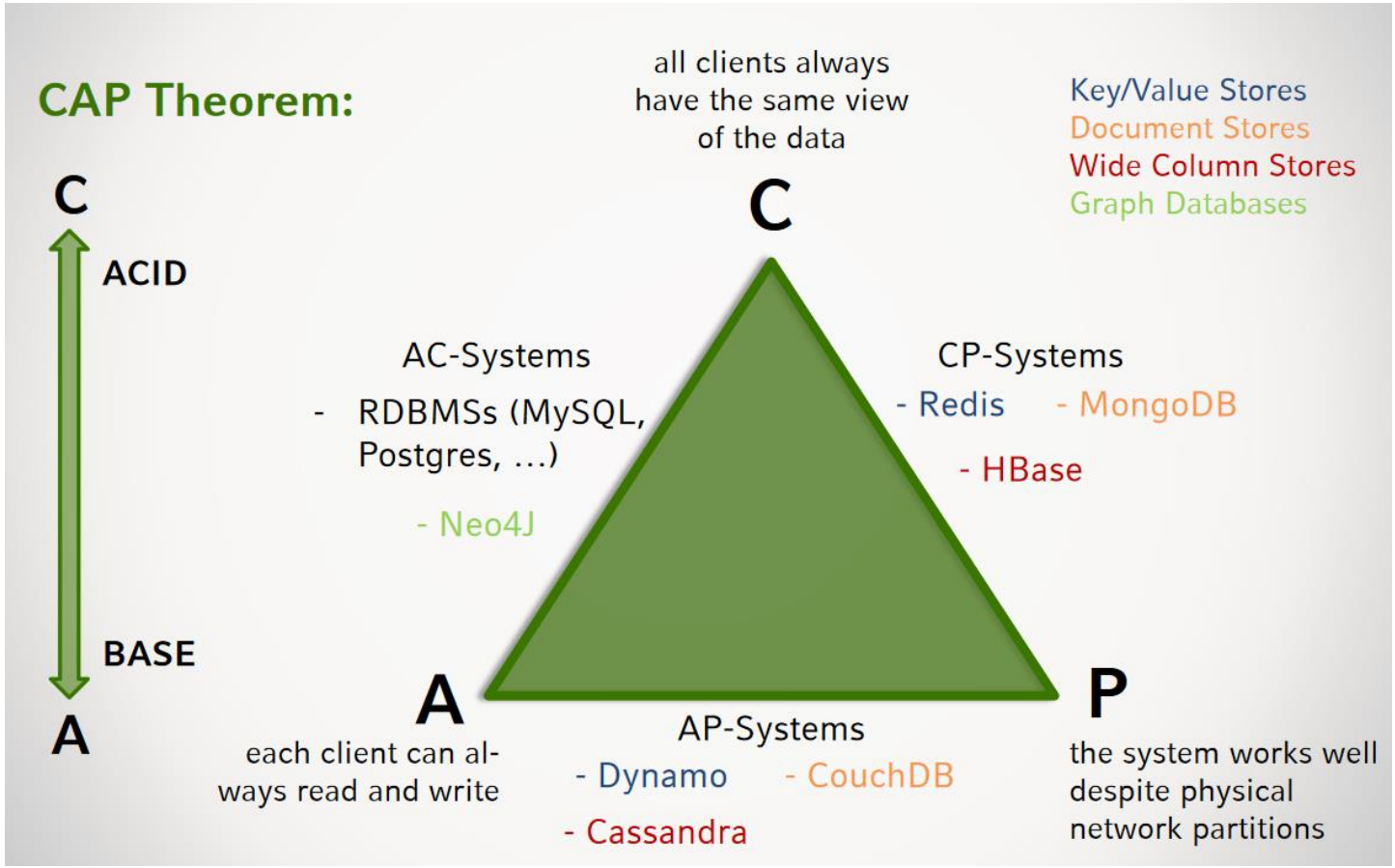
Modele danych NoSQL

- Kluczowa wartość Repozytoria
  - Redis, Riak, DinamoDB, MemcachedDB, Hazelcast, Voldemort,…
- Dokument Repozytoria
  - MongoDB, CouchDB, RethinkDB, RavenDB, Terrastore, OrientDB,…
- Zorientowany na kolumny (Szeroka kolumna) Repozytoria
  - BigTable, Cassandra, HBase, Hypertable, SimpleDB,…
- Wykres Bazy danych
  - Neo4J, Giraph, Pregel,…

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

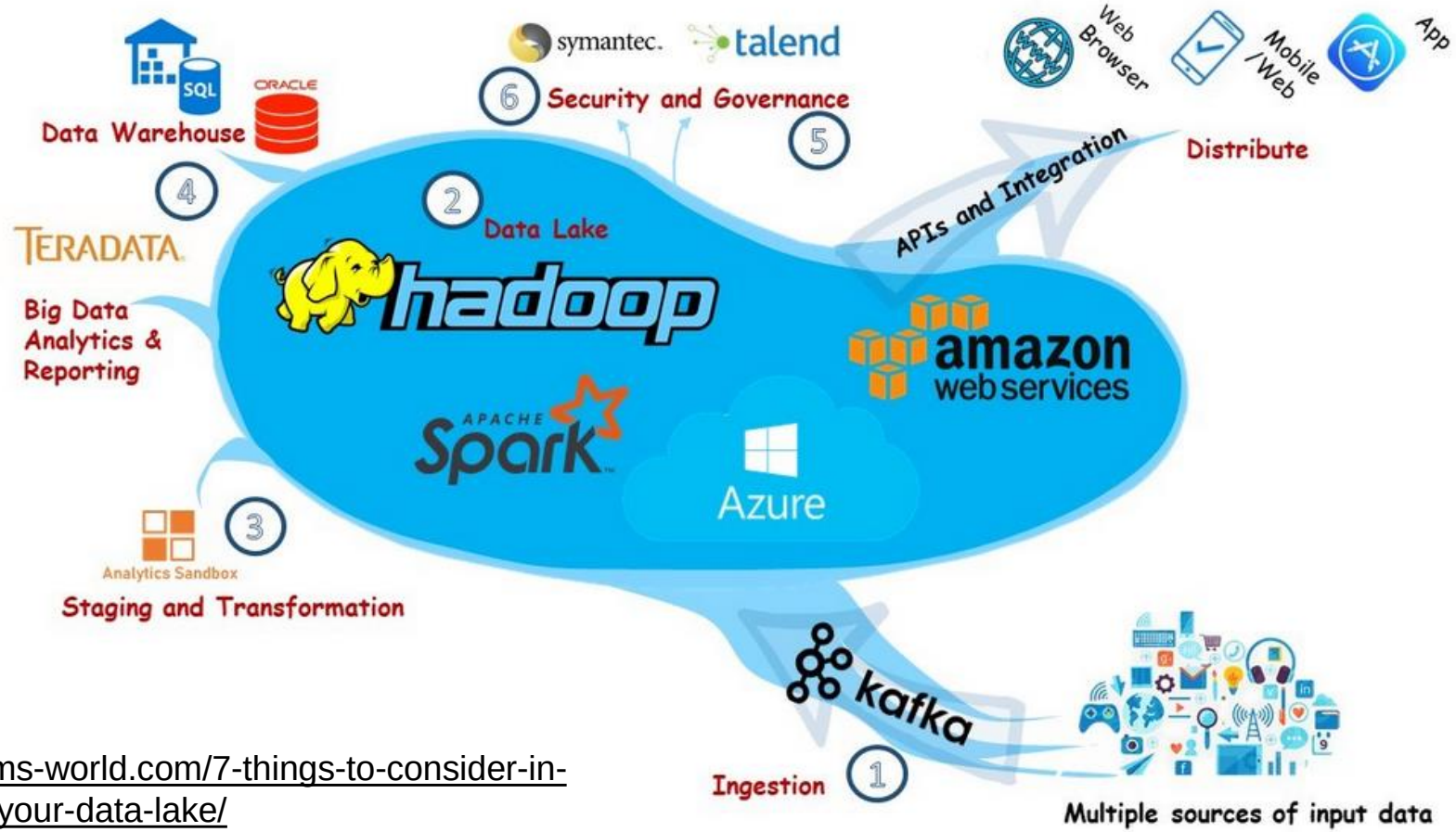
Bazy danych (SQL i NoSQL)



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jezioro danych i jezioro Delta



https://kms-world.com/7-things-to-consider-in-building-your-data-lake/

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jezioro danych a hurtownia danych

- **Jezioro danych** zawiera wszystkie dane w swojej naturalnej/surowej formie, tak jak zostały one odebrane zwykle w obiektach lub plikach.
- **Hurtownia danych** przechowuje oczyszczone i przekształcone dane wraz z katalogiem i schematem.
- Dane w jeziorze i hurtowni mogą być różnych typów: ustrukturyzowane (relacyjne), częściowo ustrukturyzowane, binarne i strumienie zdarzeń w czasie rzeczywistym.
- Jest to kwestia wyboru, kierowana wymaganiami dotyczącymi szybkości i ograniczeń kosztowych, czy jezioro i magazyn są fizycznie przechowywane w repozytoriach, czy też magazyn jest materializowany przez jakiś interfejs (np. zapytania Hive) nad jeziorem.
- Niezależnie od zastosowanego podejścia ważne jest zachowanie nieprzetworzonych danych do celów audytu, testowania i debugowania.

iBigWorld:2020-1-PL01-KA203-082197

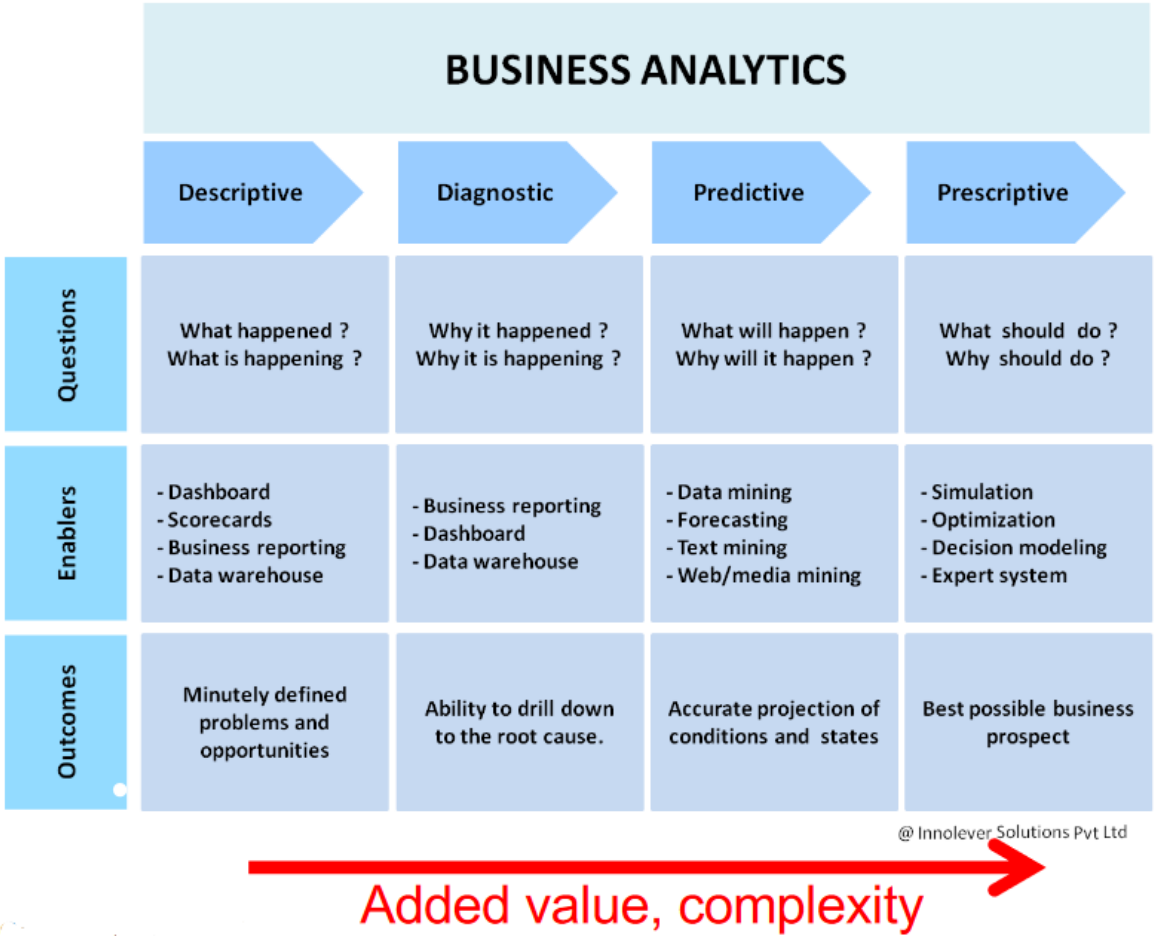
Co-funded by the Erasmus+ Programme of the European Union

Przetwarzanie i analiza dużych zbiorów danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza dużych zbiorów danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Struktury potoków Big Data

- **Apache NiFi**- obsługuje potężne i skalowalne ukierunkowane grafy routingu danych, transformacji i logiki mediacji systemu https://nifi.apache.org/
- **Apache Airflow** - zbudowany przez AirBnB do budowania, monitorowania i modernizacji potoków danych poprzez definiowanie przepływów pracy jako ukierunkowanych grafów acyklicznych (DAG) i zadań, które są tworzone dynamicznie https://airflow.apache.org/
- **Zestawy strumien i**-platforma integracji danych przeznaczona do budowy inteligentnych potoków danych w celu zasilania przetwarzania i obsługi danych w architekturach hybrydowych i wielochmurowych - https://streamsets.com/
- **Luigi** – stworzony przez Spotify dla zespołów zajmujących się analizą danych w celu tworzenia długotrwałych potoków tysięcy zadań, które rozciągają się na dni lub tygodnie https://luigi.readthedocs.io

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Apache NiFi

- Otwarte źródło, w ramach Apache Software Foundation
- Automatyzuje i zarządza przepływem danych pomiędzy systemami
- Internetowy interfejs użytkownika do tworzenia, monitorowania i kontrolowania przepływów danych
- Kluczowe cechy
  - Zarządzanie przepływem: Buforowanie danych, kolejkowanie z priorytetami, gwarantowana dostawa
  - Łatwość użycia: Szablony przepływu, Pochodzenie danych, Szczegółowa historia
  - Bezpieczeństwo: System do systemu, użytkownik do systemu, autoryzacja wielu dzierżawców
  - Rozszerzalna architektura: Rozszerzenie, protokół komunikacyjny Site-to-Site

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Kluczowe koncepcje i komponenty NiFi

Kluczowe idee

- Grupa procesów
- Przepływ
- Edytor
- Flowfile (reprezentuje pojedynczą porcję danych)
- Wydarzenie
- Pochodzenie danych

Główne komponenty

- Procesory (wykonują zadanie) i kolejki (między procesorami)

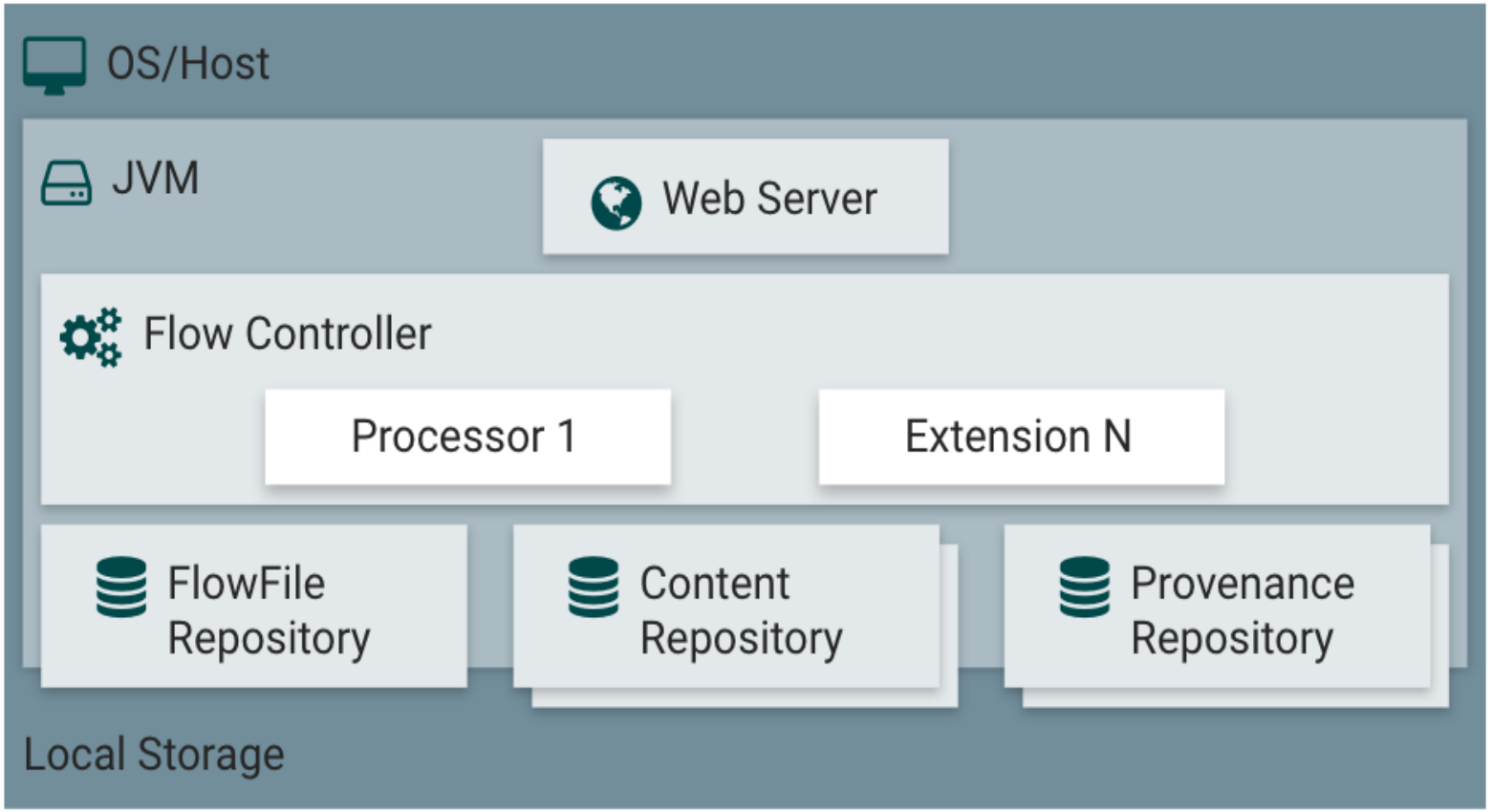
Dodatkowe komponenty

- Port wejściowy, port wyjściowy, grupa procesów, grupa procesów zdalnych, szablon

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura NiFi



https://nifi.apache.org/docs/nifi-docs/html/overview.html

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Systemy przetwarzania Apache Hadoop

- **HDFS**- Rozproszony system plików
- **Hadoop**– Framework programistyczny MapReduce
- **HBase**- Rozproszony, skalowalny magazyn dużych zbiorów danych.
- **Ul**- Hurtownie danych na platformie Hadoop
- **Pig**- Platforma wysokiego poziomu dla Apache Hadoop
- **Yarn**- System zarządzania zasobami i planowania zadań
- **Zookeeper**-Usługa rozproszona, koordynacyjna

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Platformy przetwarzania Big Data

- Apache Spark
- Apache Flink
- Apache Storm/Heron
- Apache Beam
- Apache Kafka i Kafka Sreams
- Apache Samza
- Apache Kudu

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Apache Kafka

- Rozproszony system przesyłania wiadomości typu „publikuj-subskrybuj”, szybki, skalowalny, trwały
- Kluczowe możliwości:
  - Publikuj i subskrybuj strumienie rekordów.
  - Przechowuj strumienie rekordów w trwały sposób odporny na uszkodzenia.
  - Przetwarzaj strumienie rekordów na bieżąco.
- Kafka jest ogólnie używana w dwóch szerokich klasach aplikacji:
  - Tworzenie potoków danych przesyłanych strumieniowo w czasie rzeczywistym, które niezawodnie pobierają dane między systemami lub aplikacjami
  - Tworzenie aplikacji strumieniowych w czasie rzeczywistym, które przekształcają lub reagują na strumienie danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

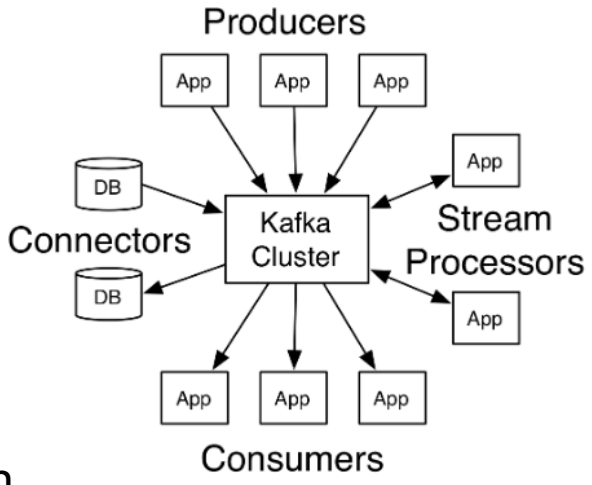
API Kafki

Producer API umożliwia aplikacji publikowanie strumienia rekordów w jednym lub kilku tematach Kafki.

Consumer API pozwala aplikacji subskrybować jeden lub więcej tematów i przetwarzać strumień rekordów dla nich tworzonych.

Streams API aplikacji (procesorowi strumieniowemu) pobieranie strumienia wejściowego z jednego lub większej liczby tematów i tworzenie strumienia wyjściowego do jednego lub większej liczby tematów wyjściowych

Connector API umożliwia połączenie tematów Kafki z istniejącymi aplikacjami lub systemami danych.



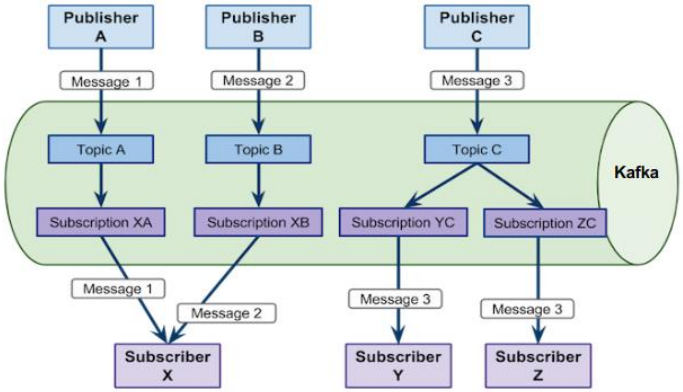
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Publikuj/subskrybuj system przesyłania wiadomości

- Kafka utrzymuje kanały wiadomości w kategoriach zwanych tematami
- **Producrs** publikują wiadomości (rekordy) w jednym lub kilku tematach
- **Consumers** subskrybują tematy i przetwarzają kanał opublikowanych wiadomości

Temat może mieć zero, jednego lub wielu konsumentów, którzy subskrybują zapisane w nim dane.

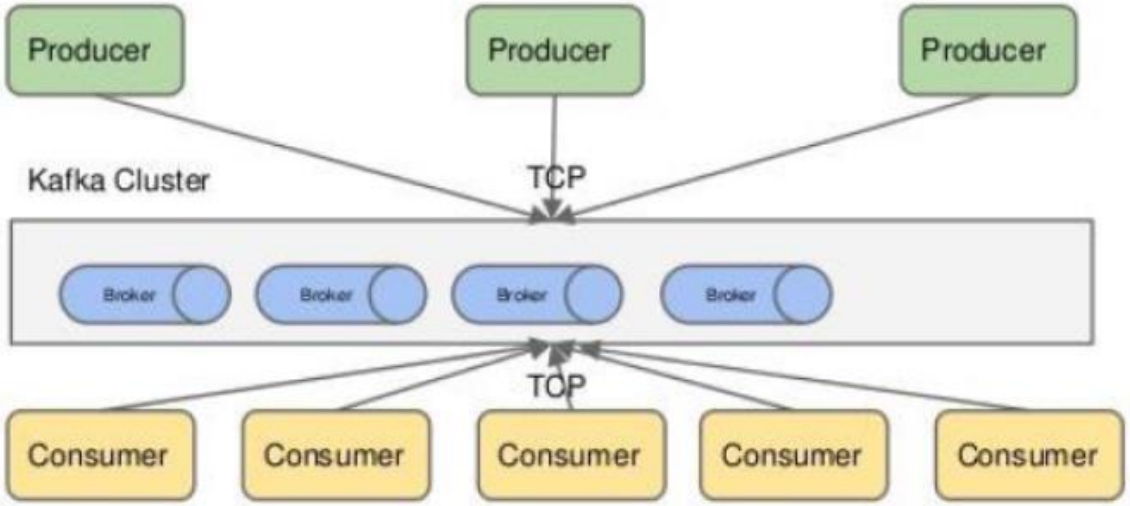


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klaster Kafki

- Ponieważ Kafka jest rozproszona, działa jako klaster.
- Klaster zazwyczaj składa się z wielu serwerów z których każdy nazywany jest brokerem.
- Komunikacja między klientami a serwerami odbywa się za pośrednictwem protokołu TCP



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Kafka – Kluczowe korzyści

- Skalowalny w poziomie
  - Jest to system rozproszony, który można elastycznie i przejrzyście rozbudowywać bez przestojów
- Wysoka przepustowość
  - Wysoka przepustowość jest zapewniona zarówno w przypadku publikowania, jak i subskrybowania, nawet przy wielu terabajtach przechowywanych wiadomości
- Niezawodna dostawa
  - Utrzymuje komunikaty na dysku i zapewnia replikację wewnątrz klastra
  - Obsługuje dużą liczbę abonentów i automatycznie równoważy odbiorców w przypadku awarii

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Kafka - użycia

- Kafka jako system przesyłania wiadomości
  - Przesyłanie wiadomości ma tradycyjnie dwa modele: kolejkowanie i publikowanie-subskrypcja. Koncepcja grupy konsumentów w Kafce uogólnia te dwie koncepcje.
- Kafka jako system przechowywania
  - Dane zapisywane w Kafce są zapisywane na dysku i replikowane w celu zapewnienia odporności na błędy, oddzielając fazę publikowania od fazy konsumpcji.
  - To sprawia, że Kafka jest bardzo dobrym systemem przechowywania.
- Kafka do przetwarzania strumieniowego
  - W Kafce procesor strumieniowy obsługuje wszystko co pobiera ciągłe strumienie danych z tematów wejściowych, wykonuje pewne przetwarzanie na tych danych wejściowych i generuje ciągłe strumienie danych do tematów wyjściowych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wizualizacja dużych zbiorów danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wizualizacja Big Data

- Wizualizacja danych wykorzystuje moc dużych zbiorów danych i chmury, aby zapewnić natychmiastowy wgląd w to, co najważniejsze dla decydentów.
- Dane napędzają decyzje biznesowe, ale dane muszą stać się analizą biznesową, zanim będzie można na nich działać.
- Wizualizacja danych to jeden z najpotężniejszych sposobów pozyskiwania wiedzy z danych i jasnego przekazywania jej innym.
- Wizualizacja danych to sztuka jasnego ilustrowania złożonych korelacji.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Narzędzia do wizualizacji Big Data

- Qlik
- Power BI
- Tableau
- Wykresy Google
- Zoho Analytics
- Kibana
- Grafana

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Monitorowanie potoku Big Data

- Potoki danych to złożone systemy składające się z oprogramowania, sprzętu i komponentów sieciowych, z których wszystkie podlegają awariom.
- Aby potok działał i mógł pobierać i ładować dane, programiści muszą pisać kod monitorowania, rejestrowania i ostrzegania, aby pomóc inżynierom danych zarządzać wydajnością i rozwiązywać wszelkie pojawiające się problemy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Bibliografia

Nathan Marz, James Warren, Big Data Principles and best practices of scalable realtime data systems. Manning Publications Co., 2015.

Martin Kleppmann, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly Media, 2017.

Dean Wampler, Fast Data Architectures for Streaming Applications, 2nd edition, O'Reilly Media, 2019.

A. Bahga, V. Madisetti, Big Data Science and Analytics: A Hands-On Approach, VPT, 2016.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union