Moduł 8
Zbieranie danych. Obszar roboczy

iBigWorld:
Innovations for Big Data in a Real World

TSNUK team

iBigWorld:2020-1-PL01-KA203-082197

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

Spis treści

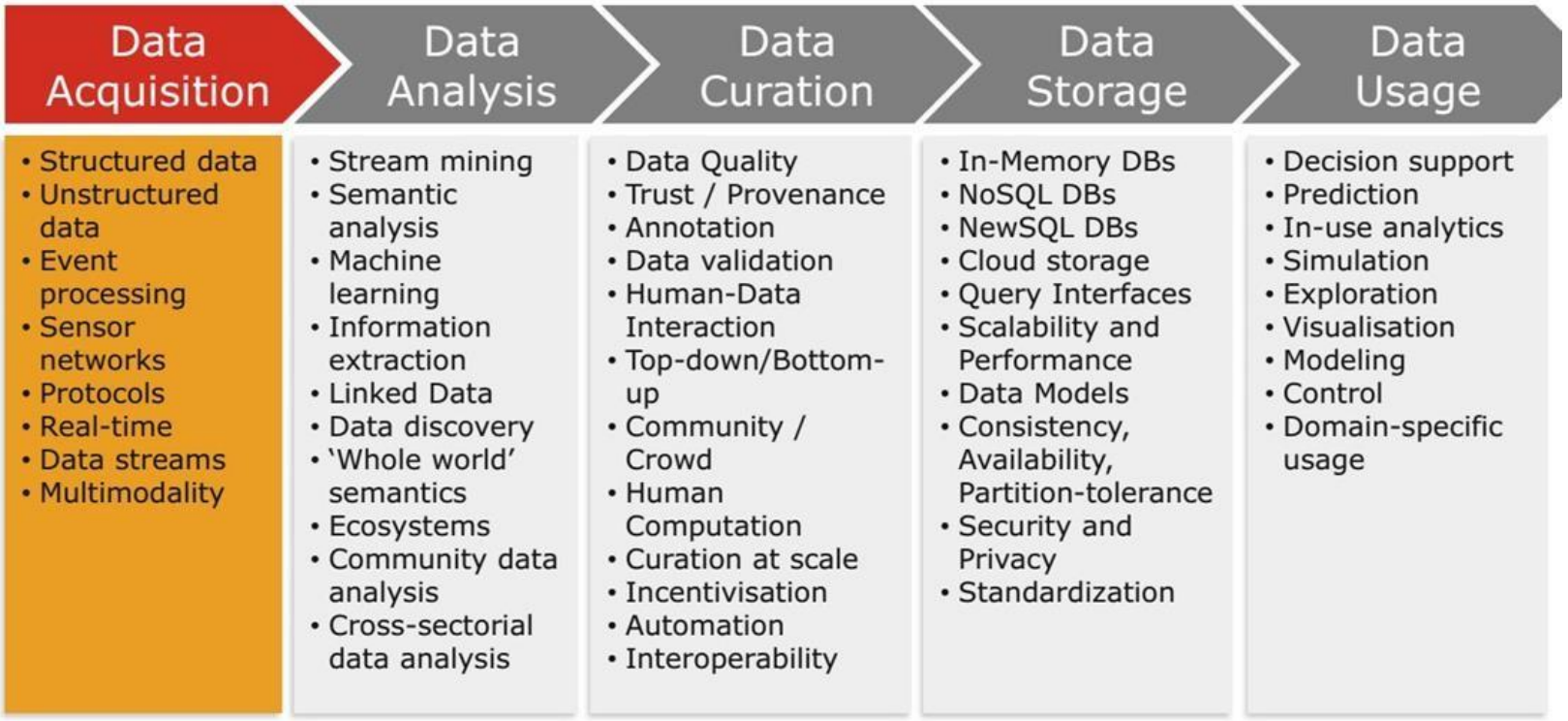
Co-funded by the Erasmus+ Programme of the European Union

Cele tej części są trojakie

- Po **pierwsze**, celem jest zdefiniowanie aktualnych ogólnych wymagań w zakresie gromadzenia danych poprzez przedstawienie otwartych, najnowocześniejszych struktur i protokołów gromadzenia danych big data dla firm.
- Po **drugie**, celem jest nagłośnienie aktualnych podejść stosowanych do gromadzenia danych z różnych sektorów.
- Po **trzecie**, omawia, w jaki sposób obecne podejścia spełniają wymagania dotyczące gromadzenia danych, a także możliwe przyszłe zmiany w tym samym obszarze.

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

Sieć Big Data



Kluczowe spostrzeżenia dotyczące gromadzenia danych big data

- W różnych architekturach przetwarzania danych big data istotą zbierania danych jest zbieranie danych z rozproszonych źródeł informacji w celu przechowywania ich w skalowalnej hurtowni danych obsługującej big data.
- Aby osiągnąć ten cel, konieczne są trzy główne elementy:

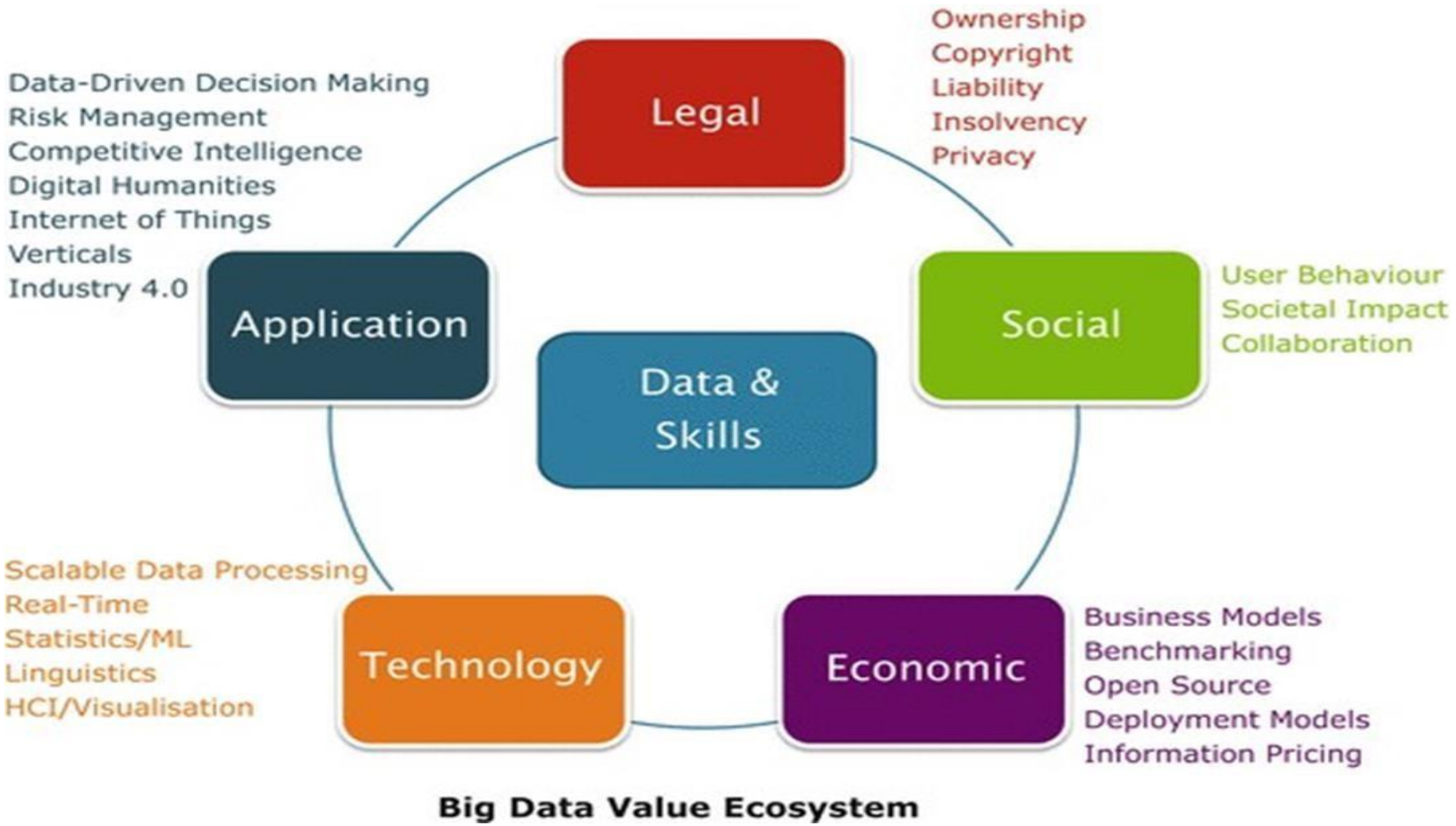
1. Protokoły umożliwiające zbieranie informacji do rozproszonych źródeł danych dowolnego typu (nieustrukturyzowane, częściowo ustrukturyzowane, ustrukturyzowane).
2. Frameworki zbierające dane z rozproszonych

źródła wykorzystujące różne protokoły.

3. Technologie umożliwiające trwałe przechowywanie danych otrzymywanych przez frameworki.

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

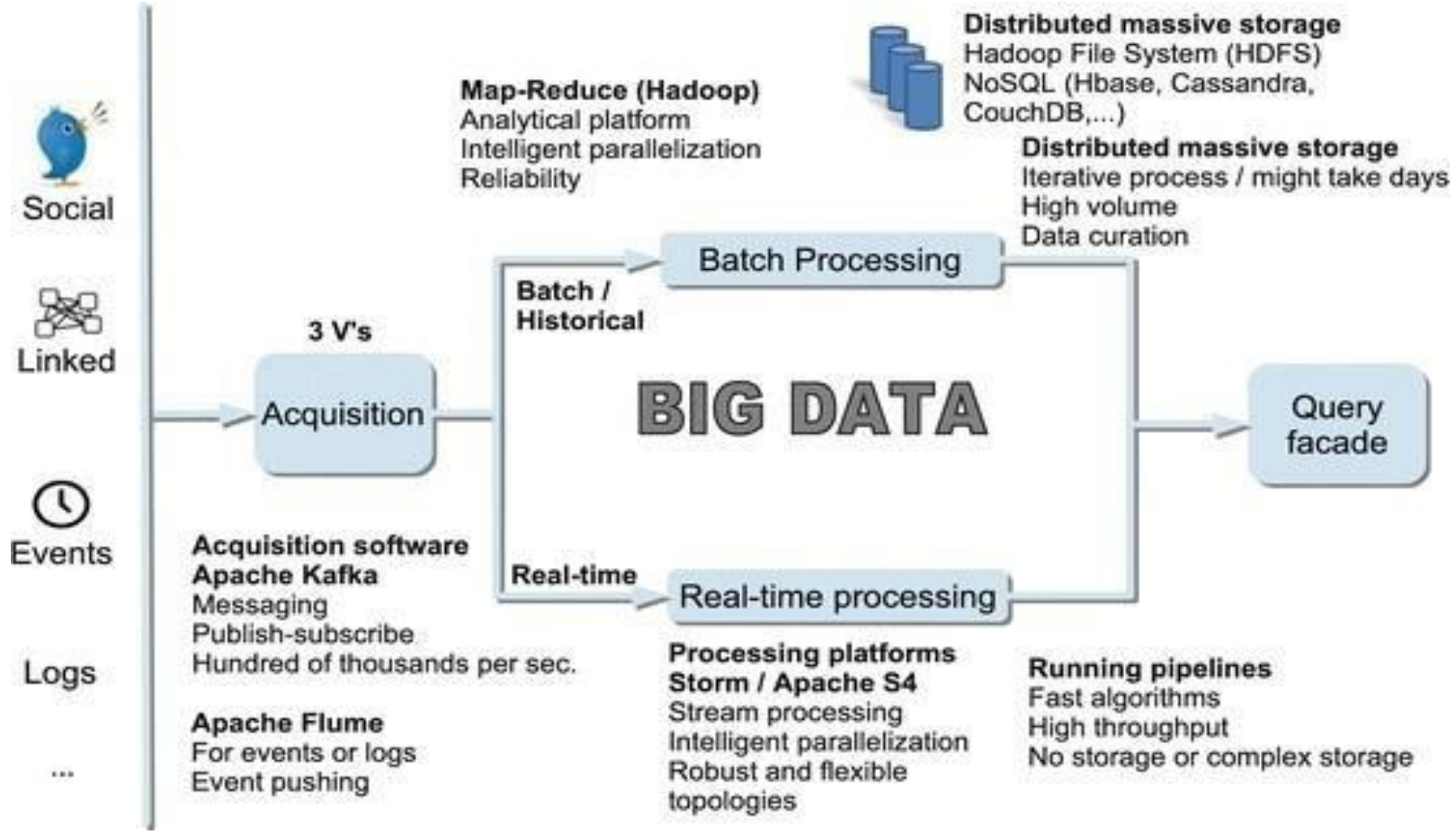
Społeczny i ekonomiczny wpływ gromadzenia dużych zbiorów danych



Big Data Value Ecosystem

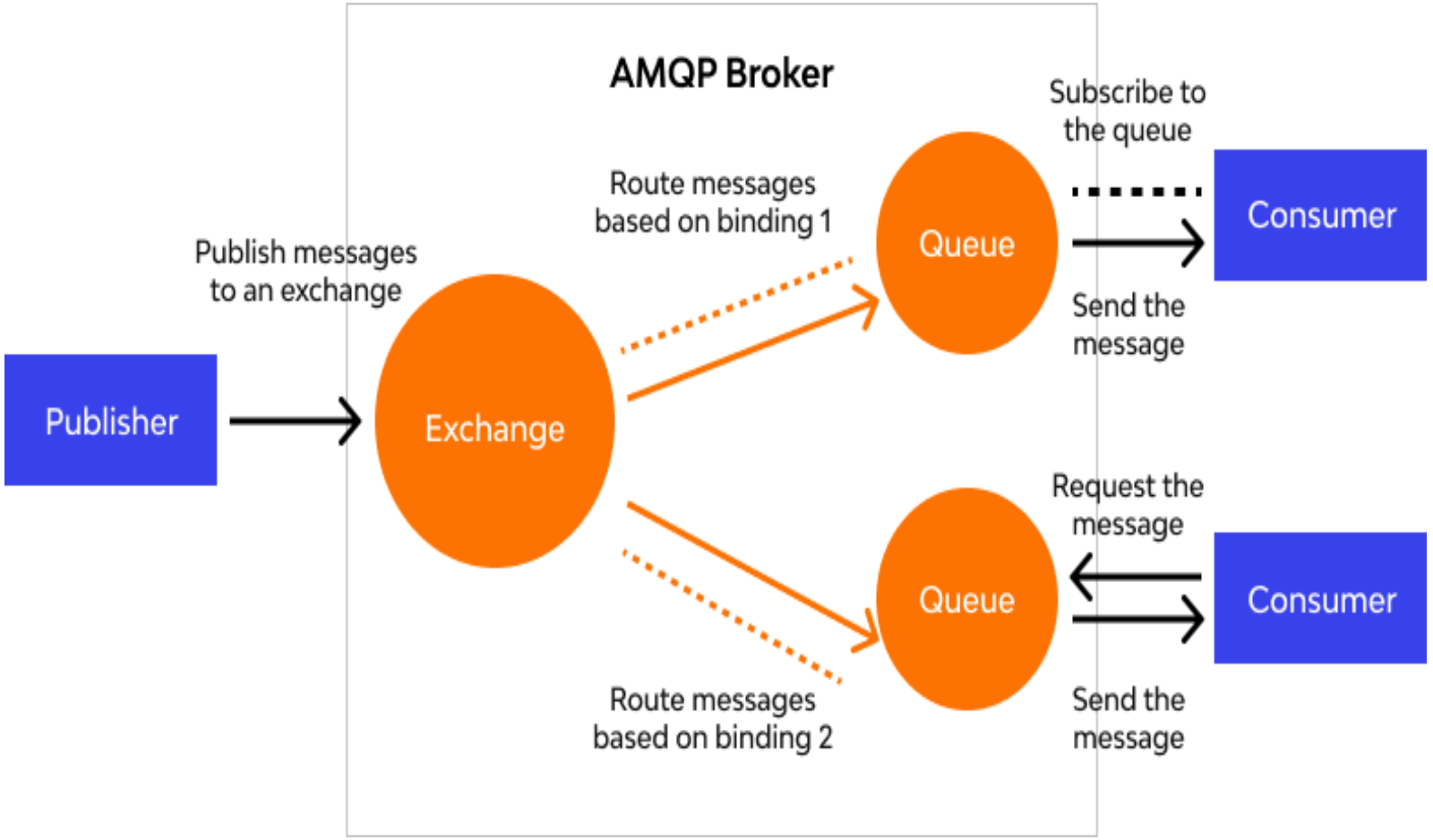
Co-funded by the Erasmus+ Programme of the European Union

Zbieranie Big Data: Stan techniki



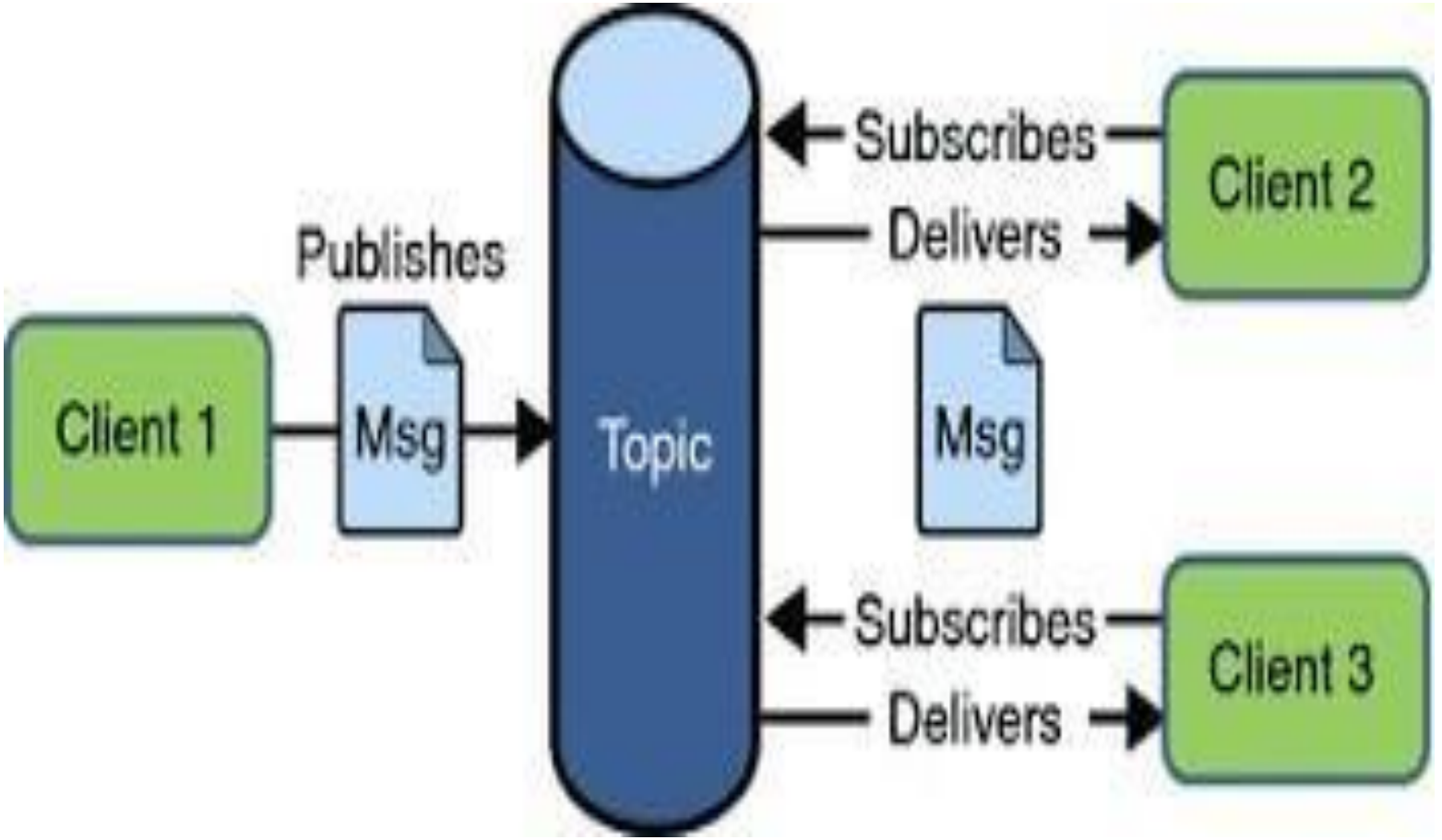
Co-funded by the
Erasmus+ Programme
of the European Union

Protokół AMQP



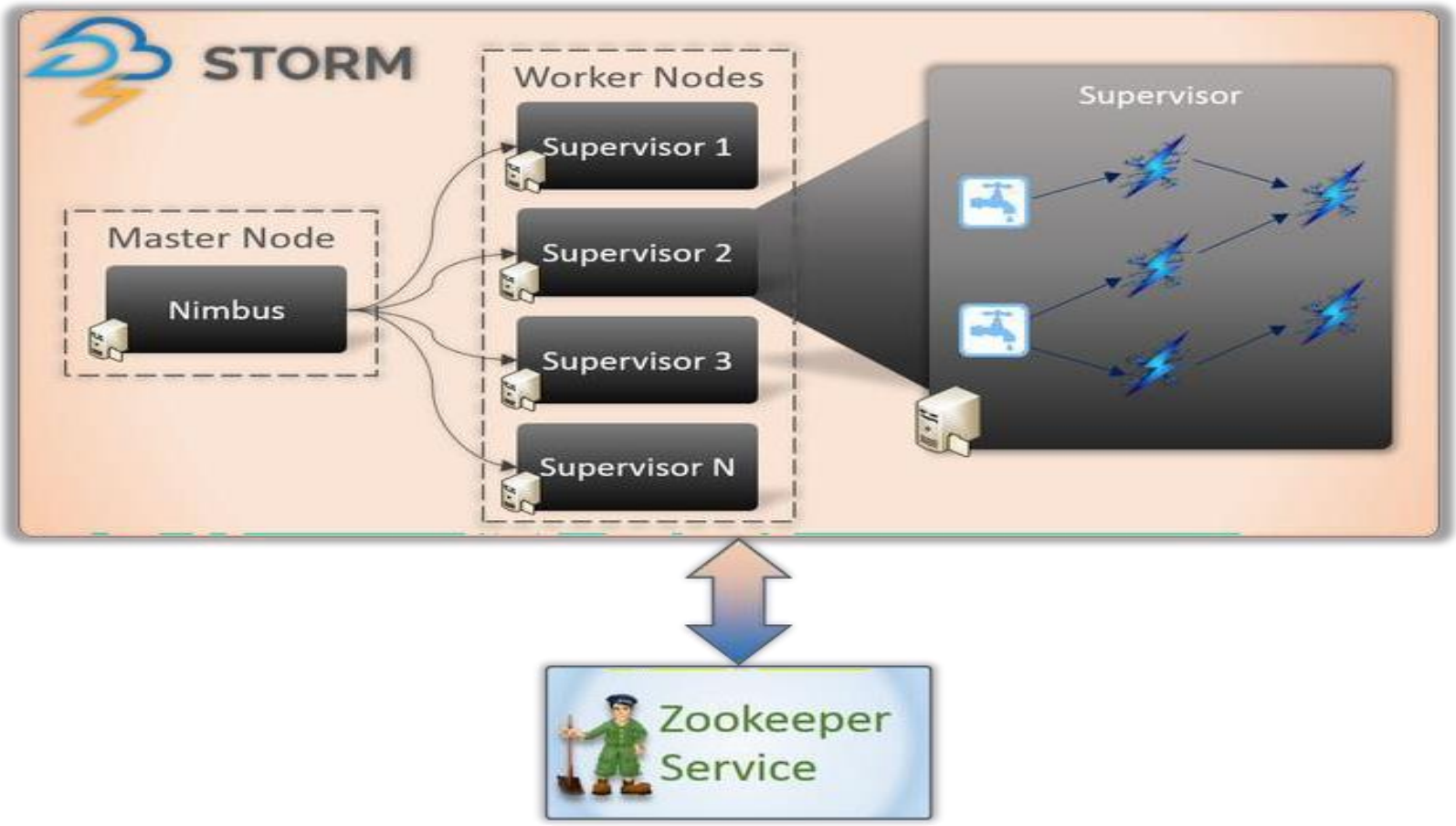
Co-funded by the Erasmus+ Programme of the European Union

Protokół Java Message Service



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

Narzędzia programowe. Obszar roboczy. STORM

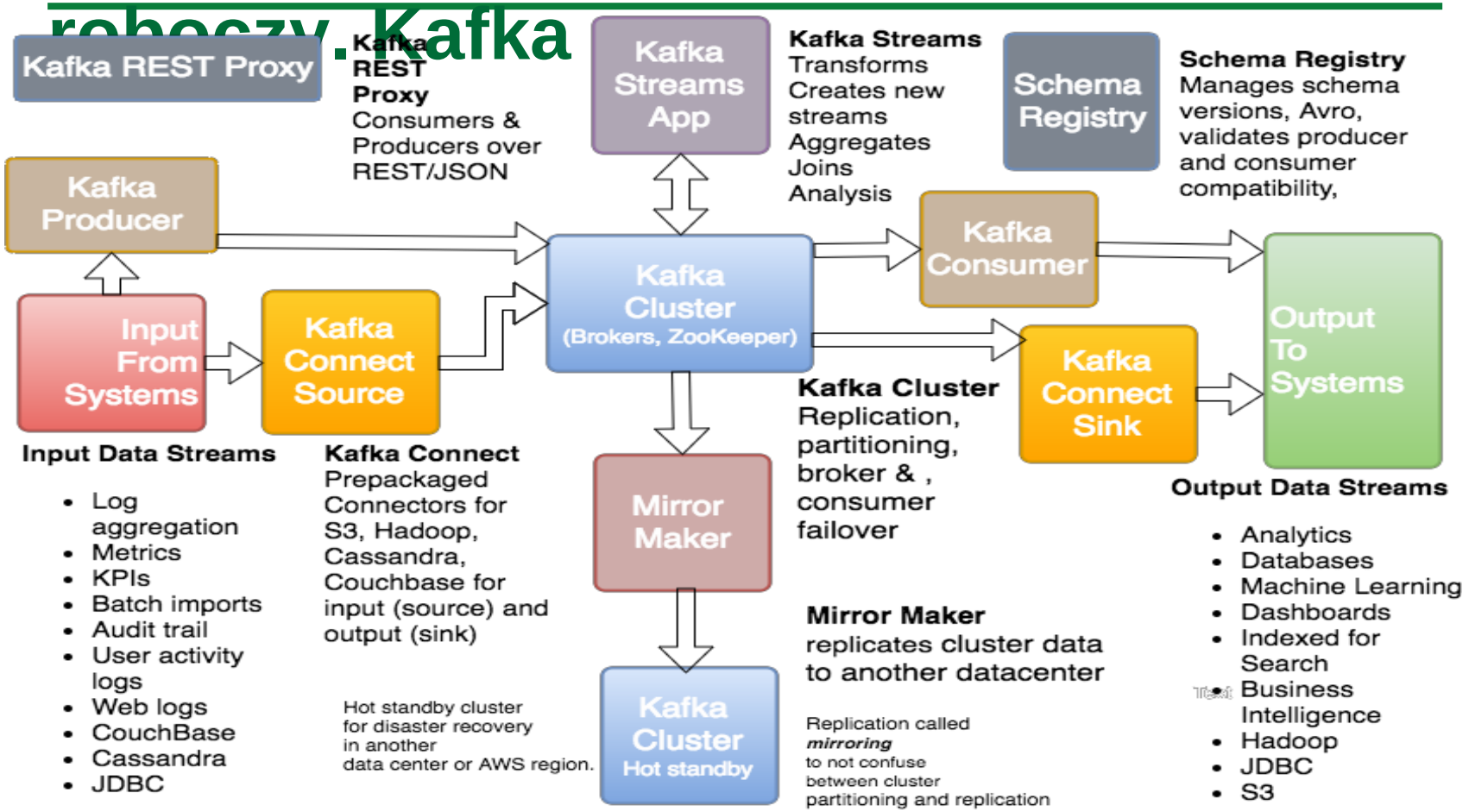


Narzędzia programowe. Obszar roboczy. Kafka

Kafka to rozproszony system przesyłania wiadomości typu publikuj-subskrybuj zaprojektowany do obsługi głównie trwałych wiadomości o wysokiej przepustowości. Kafka ma na celu ujednolicenie przetwarzania offline i online poprzez zapewnienie mechanizmu równoległego ładowania do Hadoop, a także możliwość dystrybucji zużycia w czasie rzeczywistym między klastrem maszyn.

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

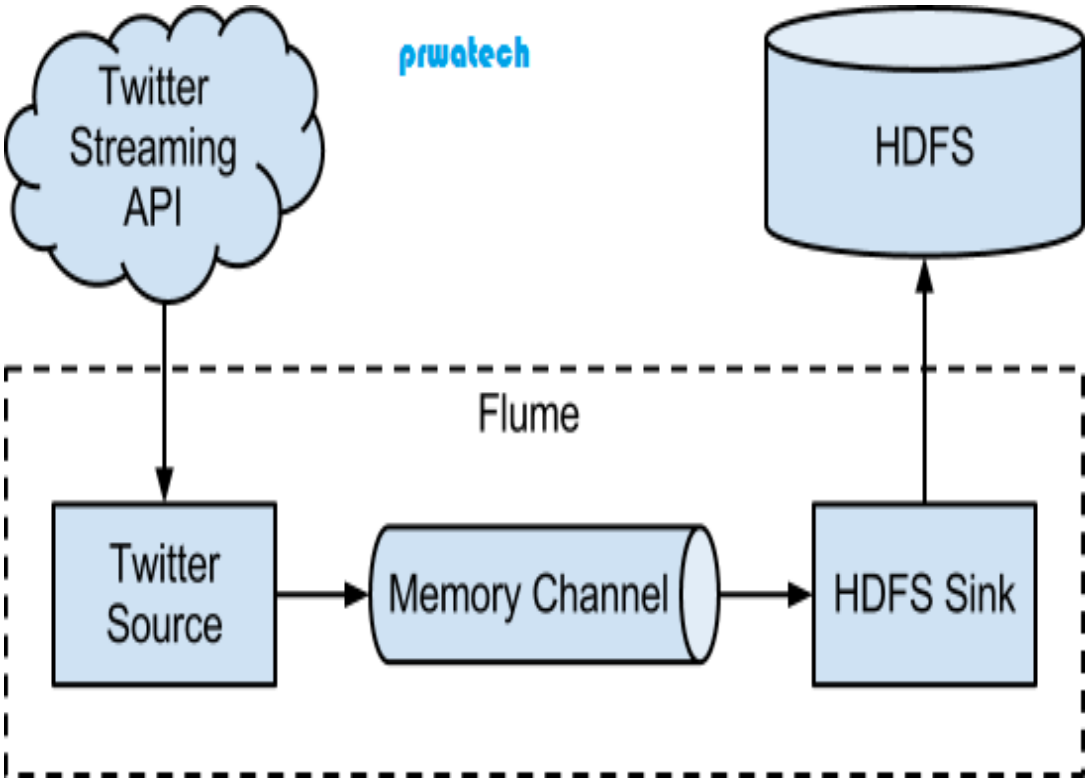
Narzędzia programowe. Obszar roboczy. Kafka



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

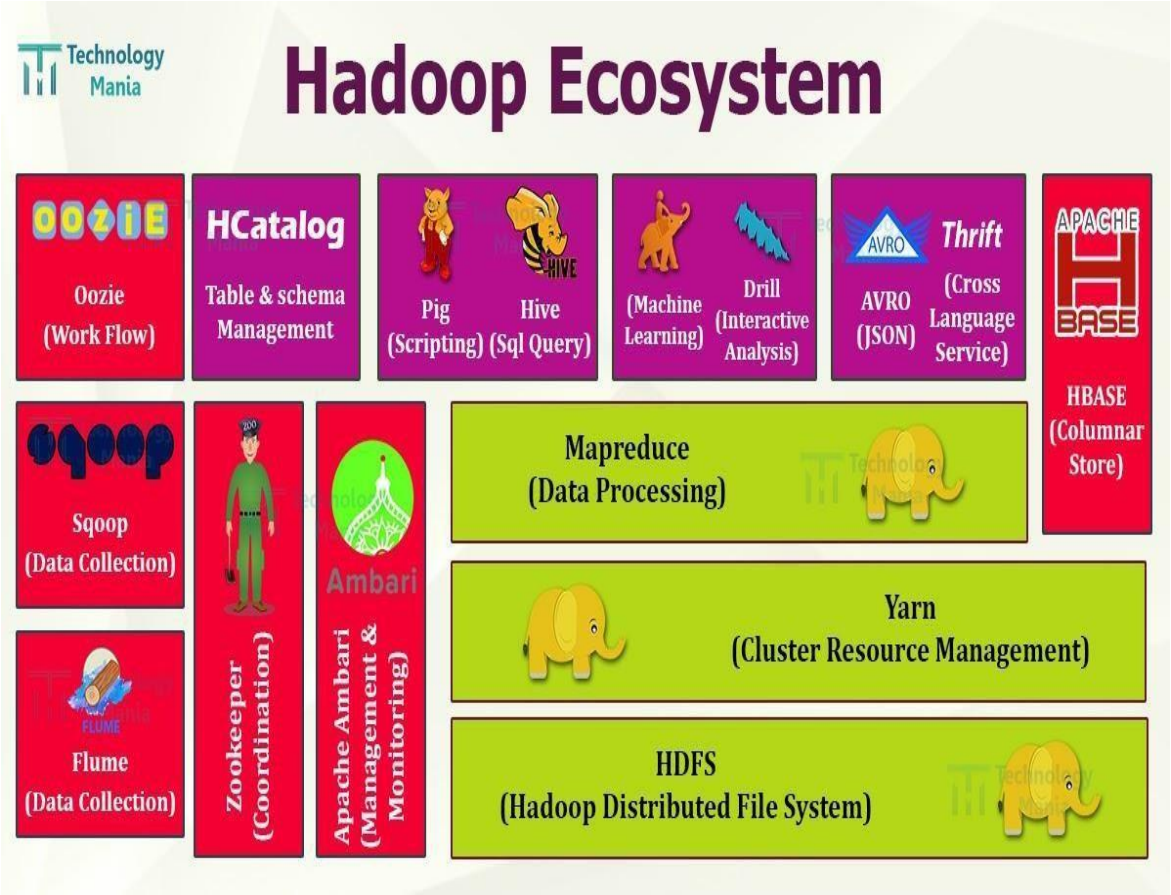
Narzędzia programowe. Obszar roboczy. Flume

Flume to usługa efektywnego zbierania i przemieszczania dużych ilości dane dziennika. Posiada prostą i elastyczną architekturę opartą na strumieniowych strumieniach danych.



Narzędzia programowe. Obszar roboczy. Hadoop

Apache Hadoop to projekt open source, który opracowuje platformę niezawodnego, skalowalnego i rozproszonego przetwarzania na dużych zbiorach danych przy użyciu klastrów gotowego sprzętu. Został wyprowadzony z Google MapReduce i Google File System (GFS) i napisany w JAVA



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

Przyszłe wymagania i pojawiające się trendy w zakresie gromadzenia danych big data

- Narzędzia do gromadzenia dużych zbiorów danych muszą obsługiwać szybkie, zróżnicowane gromadzenie danych w czasie rzeczywistym.
- Narzędzia do gromadzenia danych muszą zapewniać bardzo wysoką przepustowość. Oznacza to, że dane mogą pochodzić z
  - kilka zasobów (sieci społecznościowe, czujniki, eksploracja sieci, logi itp.)
  - o różnej strukturze, nieustrukturyzowane (tekst, wideo, obrazy i pliki multimedialne)
  - z bardzo dużą prędkością (dziesiątki lub setki tysięcy zdarzeń na sekundę).
- Głównym problemem w pozyskiwaniu dużych zbiorów danych jest to, że zapewnienie struktury i narzędzi, które zapewniają niezbędną przepustowość dla bieżącego problemu bez utraty danych w procesie.

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

Przyszłe wymagania i pojawiające się trendy w zakresie gromadzenia danych big data

- Dlatego głównym celem przy określaniu właściwej strategii gromadzenia danych jest zrozumienie potrzeb systemu w zakresie ilości, różnorodności i szybkości danych oraz podjęcie właściwej decyzji o tym, które narzędzie najlepiej zapewni gromadzenie i pożądaną przepustowość.
- Każde rozwiązanie techniczne, które ma na celu pozyskiwanie danych z różnych źródeł, musi radzić sobie z szeroką gamą różnych wdrożeń Konieczne jest zapewnienie mechanizmów łączenia zbierania danych z ich wstępną i końcową obróbką (analizą) oraz przechowywaniem danych, ponieważ zarówno w historii, jak i w czasie rzeczywistym.
- Różnorodność danych wymaga semantyki przetwarzania danych w celu prawidłowego i wydajnego łączenia danych z różnych źródeł podczas przetwarzania. Praca nad semantycznym przetwarzaniem zdarzeń, takim jak aproksymacja semantyczna, tematyczne przetwarzanie zdarzeń i tworzenie znaczników, to nowe podejścia w tej dziedzinie.

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

Literatura

1. José María Cavanillas, Edward Curry, Wolfgang Wahlster. New Horizons for a Data-Driven Economy. 2016. https://link.springer.com/content/pdf/10.1007/978-3-319-21569-3.pdf
2. Gerardus Blokdyk. Data Acquisition System A Complete Guide - 2020 Edition
3. *Gouri Ginde, Rahul Aedula, Snehanshu Saha, Archana Mathur, Sudeepa Roy Dey, Gambhire Swati Sampatrao, BS Daya Sagar.* Big Data Acquisition, Preparation, and Analysis Using Apache Software Foundation Tools. 2017. in Big Data Analytics book *by Arun K. Somani, Ganesh Chandra Deka*
4. Maurizio Di Paolo Emilio. Data Acquisition Systems From Fundamentals to Applied Design. https://www.academia.edu/27733193/Data_Acquisition_Systems_From_Fun damentals_to_Applied_Design

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