Rozproszony system zarządzania danymi w technologiach Hadoop, Kafka, Flume i Spark

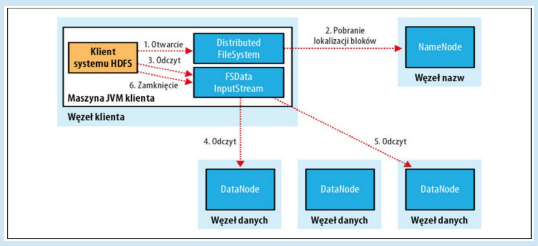
1. CEL PROJEKTU

Celem projektu było stworzenie rozproszonego systemu zarządzania danymi, który charakteryzowałby się wysoką dostępnością i skalowalnością. System powinien być elastyczny w możliwości wykonywania różnych typów zadań i przetwarzania różnych zbiorów danych.

2. UŻYTE TECHNOLOGIE

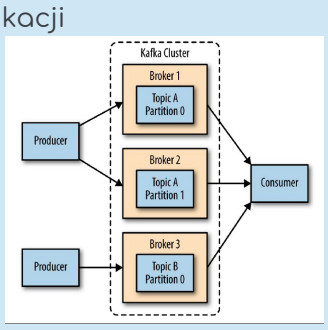
a) Hadoop

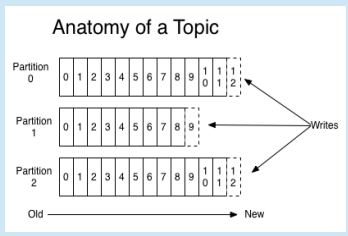
Rozproszony system przechowywania danych. Niezawodny i skalowalny. Wszystkie dane podlegają replikacji.



b) Kafka

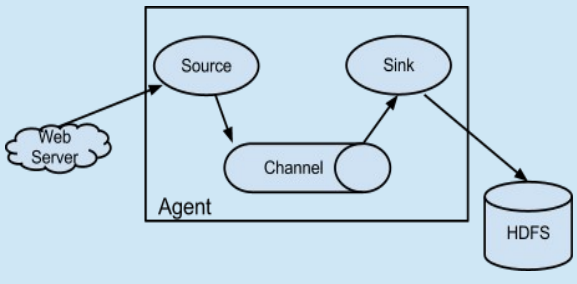
Rozproszony system przekazywania komunikatów. Skalowalny, niezawodny i wysoko dostępny. Przekazywane dane podlegają replikacji





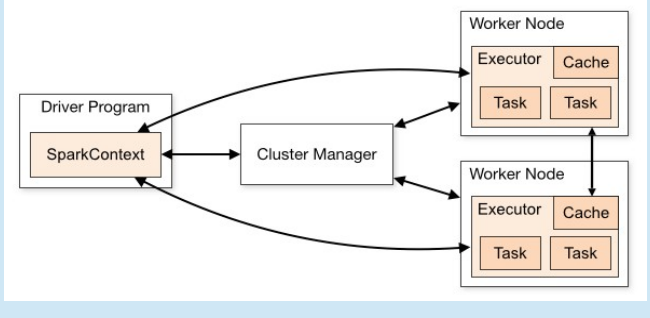
c) Flume

Prosty w konfiguracji i niezawodny system strumieniowania danych, charakteryzuje się dużą niezawodnością.



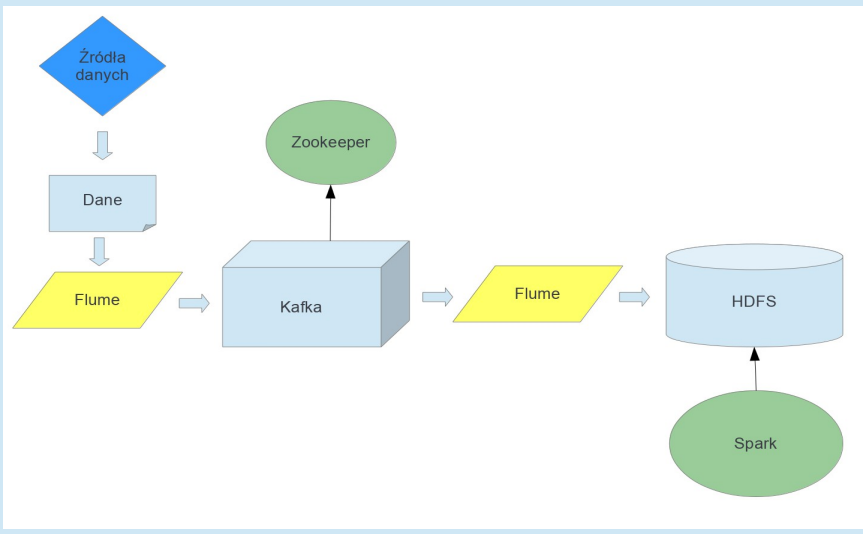
d) Spark

Rozproszony system przetwarzania danych używany do przetwarzania dużych zbiorów danych w systemie wsadowym, bardzo dobrze integruje się z Hadoop'em.



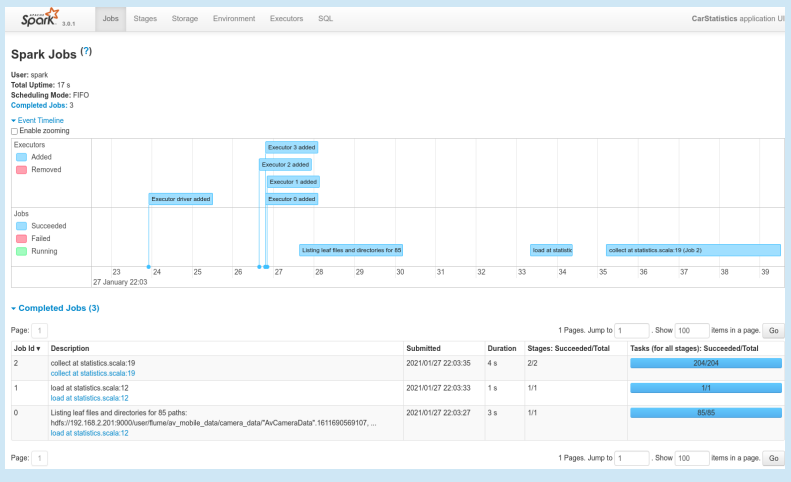
3. REALIZACJA

Ogólny schemat projektu zakłada, że dane będą tworzone i wysyłane przez producentów, na przykład przez pojazdy autonomiczne i następnie przekazywane do agentów Flume'a.
Flume strumieniuje dane do klastra Kafki, gdzie oczekują one na odebranie przez kolejnego wolnego agenta Flume'a, który następnie przekazuje je do *storage'u*. Tam podlegają dalszemu przetwarzaniu przez klaster Spark'a.



4. WYNIKI

Zastosowanie tak połączonych technologii umożliwia stworzenie elastycznego i wydajnego systemu przetwarzania dużych zbiorów danych. System charakteryzuje się dużą skalowalnością i odpornością na awarie dzięki replikacji oraz rozproszeniu systemu danych.



5. ZASTOSOWANIE

Podobne rozwiązania są szeroko stosowane w komercyjnych, produkcyjnych systemach przetwarzania zbiorów danych i stanowią trzon takich architektur jak Lambda i Kappa