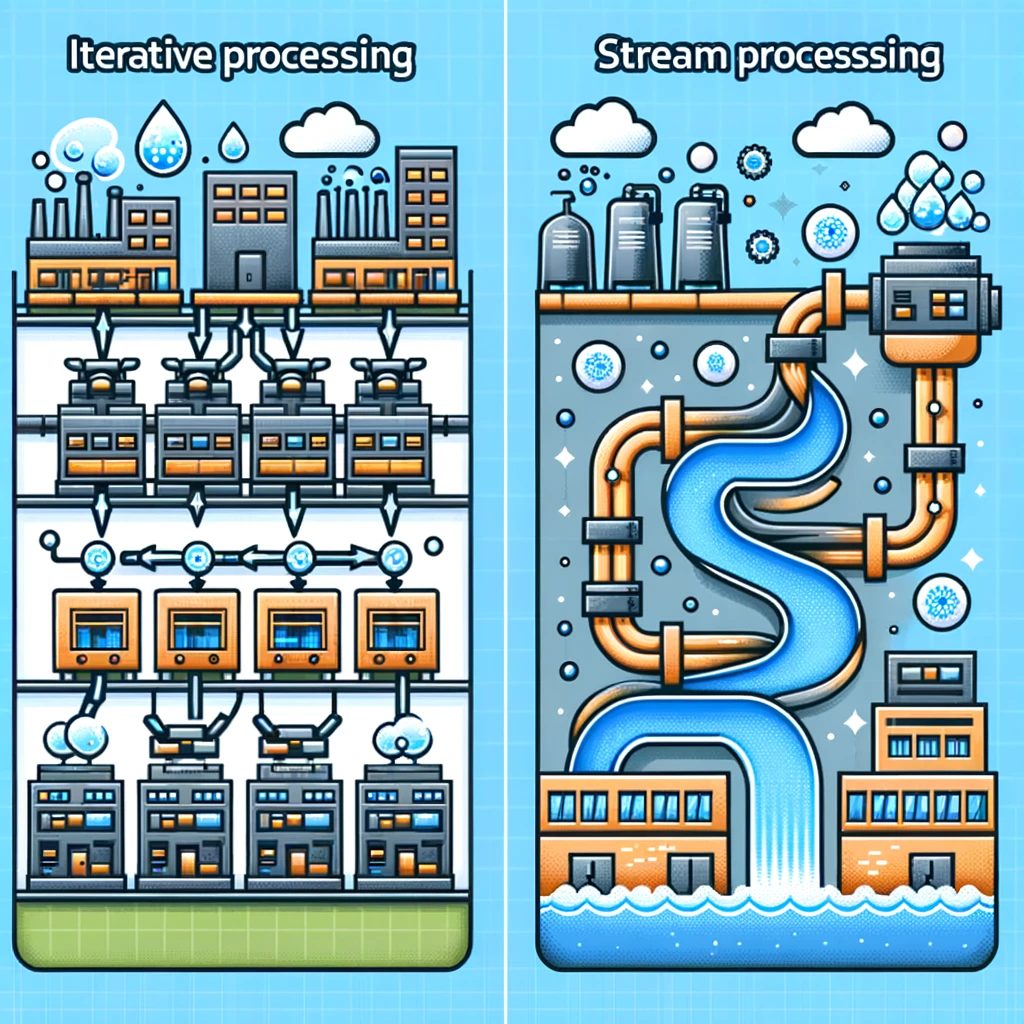
Raport

Przetwarzanie strumieniowe przy zapisie plików Big Data do bazy danych

Adam Szałański

69597

02.01.2024



# Podsumowanie

## Cel Raportu

Celem raportu było zbadanie i porównanie efektywności przetwarzania strumieniowego w kontekście zapisywania dużych zbiorów danych (Big Data) do bazy danych. Raport skupiał się na analizie wydajności przetwarzania danych za pomocą tradycyjnego przetwarzania iteracyjnego w porównaniu do operacji strumieniowych.

## Wykorzystane Technologie

Podczas tworzenia raportu utworzono aplikację opartą o język programowania Java z użyciem Spring Boot, a także z wykorzystaniem bibliotek Lombok, Hibernate JPA, jdbcTemplate oraz MapStruct. Do profilowania wydajności aplikacji użyto Java Flight Recorder (JFR), co umożliwiło dokładną analizę zużycia zasobów i wydajności.

## Przebieg tworzenia projektu, o który oparto raport

Projekt tworzony był zaczynając od podstawowej implementacji, inspirując się popularnymi wynikami z wyszukiwarki Google (głównie platforma Baeldung), a następnie rozwijany o rozwiązania stosowane w pracy komercyjnej autora raportu. W ramach logiki aplikacji utworzono dwie implementacje przetwarzania danych z pliku, mapowania tych danych oraz zapisu ich do bazy danych, różniące się wykorzystaniem przetwarzania iteracyjnego i strumieniowego.

## Kluczowe Wnioski

* Wykorzystanie strumieni wiąże się z koniecznością dostosowania logiki z ukierunkowaniem na przetwarzanie strumieniowe. Może to prowadzić zarówno do uproszczenia, jak i utrudnienia implementacji niektórych rozwiązań.
* W trakcie testów wykazano, że zastosowanie strumieni prowadzi do znacznego zmniejszenia zużycia zasobów komputera.
* Przy większych zbiorach danych, przetwarzanie iteracyjne prowadziło do wyczerpania pamięci operacyjnej, co nie miało miejsca przy przetwarzaniu strumieniowym
* Korzyści z przetwarzania strumieniowego wzrastały wraz z rozmiarem danych testowych

## Zalecenia

* Na podstawie analiz zaleca się stosowanie przetwarzania strumieniowego mimo potencjalnie bardziej wymagającej implementacji
* W związku z koniecznością zapisywania w bazie danych nie wczytanych w pełni, zaleca się zapis danych do tabeli tymczasowej, a następnie, po wczytaniu wszystkich danych, przeniesienie ich do tabeli docelowej prostym poleceniem języka SQL (czas trwania takiego polecenia jest niezauważalny)
* Należy pamiętać, aby podczas przetwarzania strumieniowego nie wykonywać akcji wymagających użycia wszystkich elementów strumienia naraz, jak np. zliczanie elementów strumienia celem wyświetlenia ich liczby

# Wprowadzenie

## Kontekst Biznesowy

W erze cyfryzacji, przedsiębiorstwa z różnych branż coraz częściej polegają na ogromnych zbiorach danych (Big Data) do podejmowania decyzji strategicznych. Efektywne przetwarzanie tych danych staje się zatem kluczowym czynnikiem sukcesu biznesowego.

## Definicja Problemu

Problemem, z którym firma musiała się zmierzyć, było znalezienie metody na szybkie i efektywne przetwarzanie rosnących zbiorów danych, szczególnie w zakresie zapisywania tych danych do bazy danych. To wyzwanie jest istotne w kontekście aplikacji wymagających bieżącej analizy i interpretacji danych.

## Typ Analityki

Projekt skupił się na analityce preskryptywnej i deskryptywnej. Analityka preskryptywna została użyta do określenia najlepszych metod przetwarzania danych, podczas gdy analityka deskryptywna pomogła zrozumieć i opisać charakterystykę danych oraz ich przetwarzanie.

## Uzasadnienie Wyboru Rozwiązania

Wybór przetwarzania strumieniowego przez firmę Software Mind wynikał z potrzeby efektywniejszego zarządzania dużymi zbiorami danych dostarczanych przez klienta z branży telekomunikacyjnej. Przetwarzanie strumieniowe oferuje lepszą wydajność i elastyczność w porównaniu do tradycyjnych metod przetwarzania iteracyjnego, co jest kluczowe w środowiskach Big Data.

## Wady i Zalety

Rozważono zalety przetwarzania strumieniowego, takie jak większa wydajność i niższe zużycie zasobów, oraz wady, w tym złożoność implementacji i potencjalne ograniczenia w zastosowaniach.