Przetwarzanie strumieni danych w systemach Big Data

część 5 – elementy zaawansowane

Krzysztof Jankiewicz

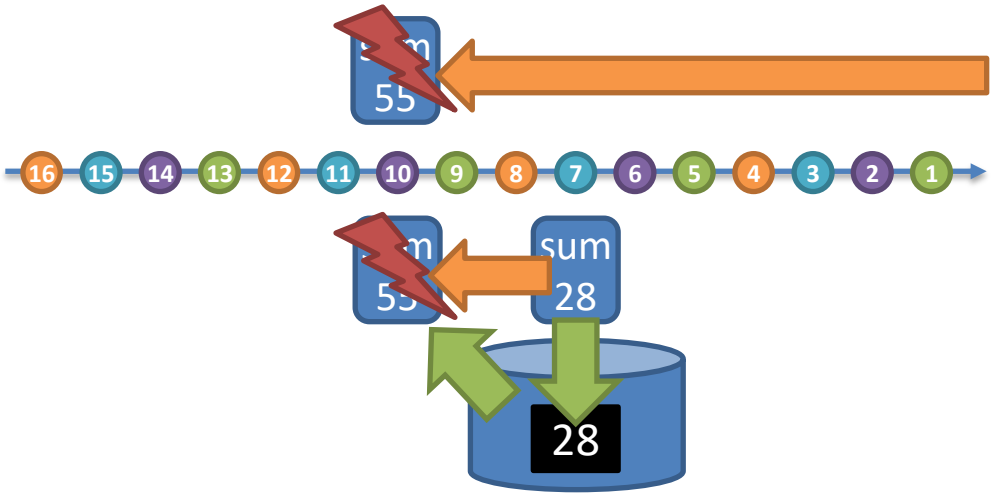
Plan

- Strumienie i ich partycje
- DAG – topologia
  - Logiczne grafy przepływu danych
  - Fizyczne grafy przepływu danych
  - Strategie przesyłania danych
- Transformacje stanowe
  - funkcje widmowe
  - TTL
- Czyściciele (evictors)
- Punkty kontrolne i ich algorytmy
- Punkty zachowania

DO UZUPEŁNIENIA

Awarie w systemach przetwarzania strumieni danych

- Aplikacje przetwarzające strumienie danych działają 24/7 w związku z tym są szczególnie narażone na awarie
- Ponowne przeliczenie danych (o ile w ogóle są one dostępne!), które były przetwarzane przez bardzo długi okres czasu, **w celu odbudowania stanu** przetwarzania, może być bardzo czasochłonne i trwać godziny czy dni
- Dlatego tak bardzo istotnym jest możliwość okresowego utrwalania i zabezpieczania stanu przetwarzania, a następnie przywracania go w przypadku awarii
- Takie zabezpieczanie nosi nazwę punktów kontrolnych (na podobieństwo punktów kontrolnych w bazach danych)



Punkty kontrolne

- **Punkt kontrolny** to zapis stanu przetwarzania aplikacji na **nośnikach** zdolnych przetrwać awarie.
- Punkty kontrolne wykorzystywane są w szczególności z dwóch powodów:
  - aby **materializować stan obliczeń** w celu uniknięcia czasochłonnego przeliczania tego stanu z danych źródłowych
  - **umożliwić ponowne uruchomienie programu sterownika**, który na podstawie zapisów w punkcie kontrolnym będzie wiedział jak odtworzyć stan przetwarzania i to przetwarzanie kontynuować.
- Czasami rozróżnia się dwa typy punktów kontrolnych:
  - punkty kontrolne **metadanych** (*metadata checkpointing*), zawierające:
    - dane dotyczące konfiguracji – wymagane do restartu aplikacji
    - zbiór operacji definiujących aplikację przetwarzającą - DAG
    - "niedokończone" porcje danych – porcje danych, których przetwarzanie nie zostało zakończone
  - punkty kontrolne **danych** (*data checkpointing*), zawierające pośrednie etapy przetwarzania transformacji stanowych, dzięki czemu znacząco może zostać ograniczona:
    - liczba źródłowych porcji danych, oraz
    - czas i zasoby

    Punkty kontrolne danych wymagane są do odzyskania stanu przetwarzania

Punkty kontrolne, punkty zachowania i odtwarzanie stanu

- Przetwarzanie **wsadowe** – **ponowne** pełne obliczenia
- Przetwarzanie **strumieni danych:**
  - **punkty kontrolne** (*checkpointing*) i
  - **ponowne przetwarzanie** strumienia **ostatnich** zdarzeń (*stream replay*)
- Punkty kontrolne wykonywane są **automatycznie** przez system, w celu ich ewentualnego wykorzystania podczas **odtwarzania** systemu po awarii.

- **Spójny punkt kontrolny** – to, dla systemu przetwarzania strumienia zdarzeń opartego na stanach, kopia stanu dla każdej jednostki zadania w momencie, w którym przetworzyły one określony strumień wejściowy.

- **Odtwarzanie** polega na **przywróceniu stanu** wszystkich jednostek zadań, a następnie **dostarczeniu** na wyjście wszystkich **zdarzeń**, które **nie** zostały **uwzględnione** w zapisanych stanach.

- **Punkty zachowania** – to odmiana punktu kontrolnego zapisanego wraz z dodatkowymi metadanymi wykonywana przez administratora systemu
- Celem punktów zachowania jest dla przykładu rekonfiguracja lub migracja przetwarzania

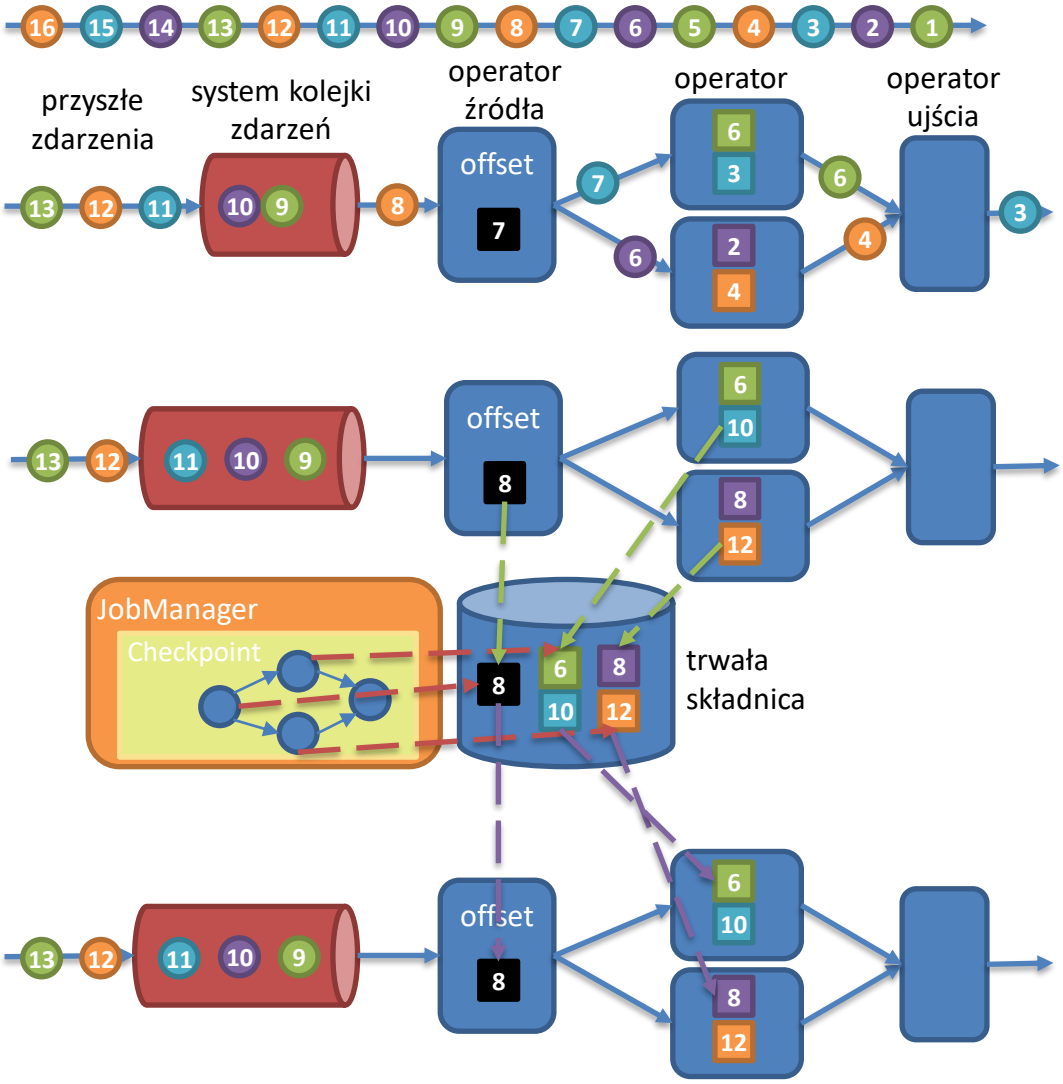
Algorytmy punktów kontrolnych

- Wariant podstawowy – z zatrzymaniem przetwarzania
- Warianty nie wymagające zatrzymania aplikacji (na przykładzie Flinka)
  - Wariant wyrównujący
  - Wariant niewyrównujący

Punkt kontrolny – wariant podstawowy

- Mógłby składać się z następujących kroków
  - Zatrzymujemy źródła i nie przyjmujemy nowych zdarzeń
  - Czekamy, aż wszystkie zdarzenia w systemie zostaną przetworzone
  - Tworzymy punkt kontrolny zapisując stany jednostek przetwarzania w trwałych repozytoriach
  - Czekamy, aż wszystkie jednostki zapiszą swój stan
  - Uruchamiamy ponownie źródła kontynuując obsługę zdarzeń
- Jakie wady takiego podejścia?

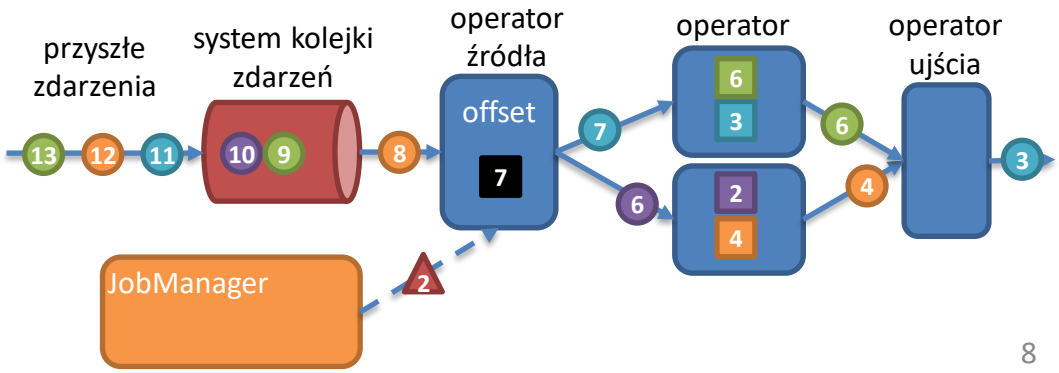
- Kroki odtwarzania
  - Przywrócenie całej aplikacji
  - Przywrócenie stanów wszystkich jednostek zgodnie z zawartością punktów kontrolnych
  - Wznowienie pracy jednostek



Jakie wady takiego rozwiązania?

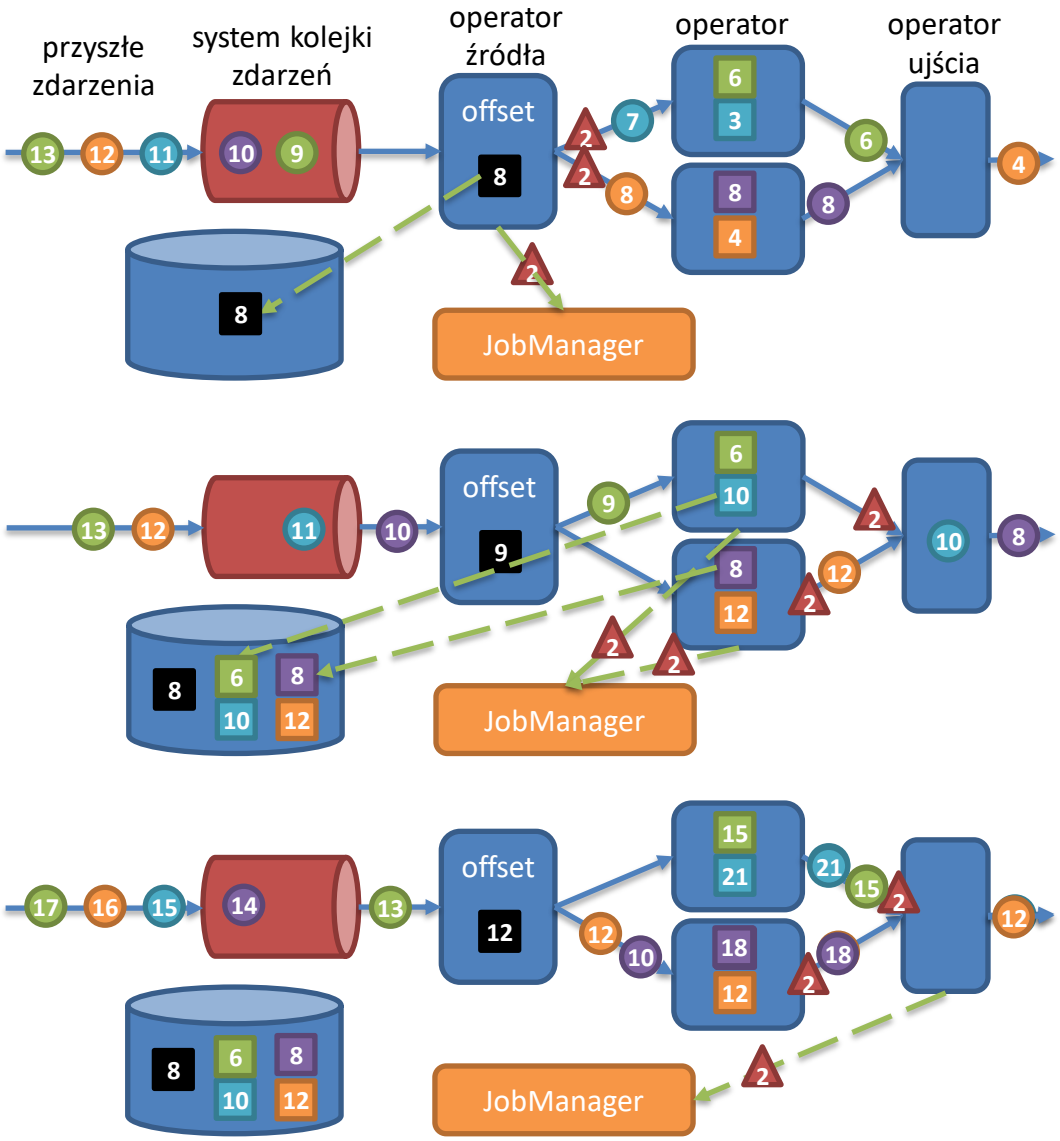
Punkt kontrolny – Flink

- Oparty jest na algorytmie Chandy–Lamport
  patrz: `http://lamport.azurewebsites.net/pubs/chandy.pdf`
- Opisany szczegółowo w: `http://arxiv.org/abs/1506.08603`
- Nie wymaga zatrzymywania całej aplikacji: podczas gdy część z operatorów dokonuje zapisywania swojego stanu, pozostałe dokonują standardowego przetwarzania
- Oparte są na barierach punktów kontrolnych
  - specjalne elementy strumienia zawierające ID punktów kontrolnych, których dotyczą
  - rozdzielają zdarzenia, które zostaną objęte przez poszczególne punkty kontrolne
  - inicjowane przez *JobManagera* poprzez wysłanie nowego ID punktu kontrolnego do każdej jednostki zadań operatora źródła
- Występuje w dwóch wersjach:
  - wyrównującej (*aligned*)
  - niewyrównującej, od wersji 1.11 (*unaligned*)



Punkt kontrolny wyrównujący

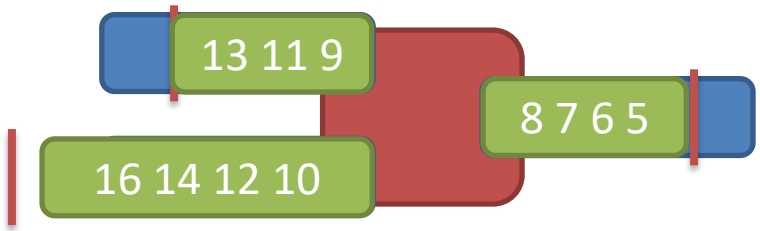
- Bariery punktu kontrolnego
  - nie wyprzedzają zdarzeń
  - są propagowane do wszystkich kolejnych operatorów
- Wyrównanie: operator, który otrzymuje barierę
  - wstrzymuje odbieranie zdarzeń od nadawcy bariery
  - odbiera zdarzenia od pozostałych nadawców i dokonuje ich przetwarzania
  - w momencie gdy otrzyma barierę od ostatniego z nadawców
    - rozpoczyna asynchroniczne zapisywanie swojego stanu
    - propaguje barierę do swoich odbiorców
    - potwierdza wykonanie punktu kontrolnego do *JobManagera*
  - kontynuuje przetwarzanie zdarzeń od wszystkich nadawców
- Koniec punktu kontrolnego następuje w momencie, gdy *JobManager* otrzymuje potwierdzenia od wszystkich operatorów
- Odtwarzanie jest analogiczne do wariantu podstawowego



Jakie wady takiego rozwiązania?

Punkt kontrolny niewyrównujący

- Pojawił się w wersji 1.11, zbliżając rozwiązanie stosowane przez Flinka do algorytmu Chandy-Lamporta
- Sposób obsługi bariery:
  - operator, który otrzymuje po raz pierwszy barierę punktu kontrolnego o określonym ID rozpoczyna od razu jej przetwarzanie
  - natychmiast propaguje ją do odbiorców wyprzedzając zdarzenia zawarte
    - w buforach wejściowych, a także
    - w buforach wyjściowych operatora
  - wszystkie zdarzenia wyprzedzone przez barierę oraz te, które dojdą od pozostałych odbiorców aż do czasu otrzymania od nich tej samej bariery, są oznaczane przez operator i stają się częścią jego stanu zapisanego w składnicy



- Stan punktu kontrolnego obejmuje zatem:
  - oprócz stanu operatora także
  - zdarzenia buforów wyjściowych oraz
  - zdarzenia buforów wejściowych, które zostały wyprzedzone przez barierę
- Odtwarzanie wymaga odtworzenia wszystkich powyższych składowych operatora

- Przydatne dla aplikacji, w których propagowanie zdarzeń przez system może trwać bardzo długo. Patrz:
  `https://nightlies.apache.org/flink/flink-docs-master/docs/concepts/stateful-stream-processing/#unaligned-checkpointing`

Zakres zmian aplikacji a punkty kontrolne

- Wprowadzenie
- Przykładowe akceptowane zmiany
- Zmiany, które mogą być nieakceptowane

Zakres zmian aplikacji a punkty kontrolne

- Aplikacja wznawiająca przetwarzanie na podstawie punktu kontrolnego, w założeniu jest tą samą, która go zapisała
- Możliwy jest jednak przypadek, w którym aplikacja od momentu zatrzymania/awarii uległa pewnym zmianom.
- Zakres tych zmian zależny jest od systemu przetwarzania strumieni danych
- Powody zmian aplikacji mogą być różne
  - poprawka z powodów biznesowych
  - poprawka ze względu na zmiany środowiska po awarii

Przykładowe akceptowane zmiany

- Zmiany w parametrach źródeł nie zmieniające semantyki
- Zmiany typu ujścia (niektóre) w szczególności definiowanych przez użytkownika
- Zmiany w parametrach ujść
- Dodanie, usunięcie operatorów `filter`
- Zmiana w projekcji o ile wynikowy schemat jest identyczny
- Zmiany w funkcjach użytkownika obsługujących przetwarzanie stanowe

Przykładowe zmiany nieakceptowane

- Zmiany typów lub liczby źródeł (np. zmiana tematu Kafki)
- Zmiany w liczbie lub typach kluczy grupujących lub funkcji agregujących
- Zmiany w sposobie eliminacji zduplikowanych wartości
- Zmiany w definicjach operacji połączenia strumieni danych
- Zmiany w postaci stanów oraz sposobie obsługi ich terminów ważności dla funkcji stanowych użytkownika