Potok do przetwarzania wysokoobjętościowych strumieni danych z pulsarów

Zaawansowane Bazy Danych i Hurtownie Danych

Michał Lubecki
Leszek Gzik

Topologia

Na obecnym etapie prac jesteśmy w stanie stworzyć przepływ pozwalający na:

- wczytanie danych
- wstępny przesiew obiektów poprzez usuwanie duplikatów (np. sygnały o bardzo zbliżonej częstotliwości)
- podanie z góry ustalonego obiektu pulsara do porównywania z dostępną listą znanych gwiazd

Wykonana optymalizacja

- Zoptymalizowano sposób przekazywania argumentów, które miały trafić do wszystkich węzłów topologii. Dawniej każdy węzeł sam tworzył sobie bazową liczbę obiektów, z którą był porównywany obecnie przetwarzany element, teraz gotowa lista jest przekazywana do każdego węzła podczas jego tworzenia.
- Korzystamy z dogodnej cechy Apache Storm która pozwala na wystąpienie błędów w przetwarzaniu obiektów bez zatrzymywania działania sieci. Nie każdy obiekt posiada uzupełnione wszystkie dane, kiedy węzeł będzie chciał go przetworzyć, obiekt zostanie pominięty.

Wykonane zadania - EntryReaderSpout

Dodany został węzeł wejściowy (odpływ - ang. *spout*) o nazwie **EntryReaderSpout**, pozwalający na odczyt danych obiektów astronomicznych z pliku wejściowego **psrcat.db**.

Węzeł ten jako argument przyjmuje ścieżkę do pliku wejściowego, a następnie zamienia każdy wpis w tym pliku na obiekt typu **PSRCatEntry**, który jest wysyłany w postaci krotki do dalszego przetworzenia.

Wykonane zadania - PeriodSiftingBolt

Do topologii dodano także węzeł roboczy (ang. *bolt*) o nazwie **PeriodSiftingBolt**, którego zadaniem jest przesiew zarejestrowanych sygnałów pod kątem okresu.

Ponieważ sygnały o bardzo podobnym okresie najprawdopodobniej pochodzą od tego samego obiektu, węzeł ten odsiewa potencjalne duplikaty, wysyłając na wyjście tylko krotki o unikatowych okresach.

Za “margines podobieństwa” przyjęliśmy **10 ms**.

Wykonane zadania - FreqSiftingBolt

Kolejny węzeł odsiewający, o nazwie **FreqSiftingBolt**, ma za zadanie dalszy przesiew krotek, tym razem pod względem częstotliwości zarejestrowanego sygnału. Jego działanie jest zbliżone do **PeriodSiftingBolt**.

Za “margines podobieństwa” przyjęliśmy tym razem **0.01 Hz**.

Przykład działania przesiewu

Aktualnie program wysyła informację o rezultatach przesiewu wyłącznie na konsolę. Kolejnym krokiem będzie zapis tych danych do pliku wyjściowego dla wygodniejszej analizy.

```
HELLO J2310+6706
FREQ is 1.94478897277861
39737 [Thread-19-sifting-bolt-2] INFO  backtype.storm.daemon.task - Emitting: sifting-bolt-2 default [model.PSRCATEntry@447a9055]
J2310+6706 is not a duplicate
39737 [Thread-19-sifting-bolt-2] INFO  backtype.storm.daemon.executor - Processing received message source: sifting-bolt:3, stream: default, id: {}, [model.PSRCATEntry@2d671068
HELLO J2339-0533
FREQ is 0.0028842267415472283
J2339-0533 is a duplicate!
39737 [Thread-19-sifting-bolt-2] INFO  backtype.storm.daemon.executor - Processing received message source: sifting-bolt:3, stream: default, id: {}, [model.PSRCATEntry@1a38521d
HELLO J2325-0530
FREQ is 0.8687351150250493
39738 [Thread-15-entry-reader-spout] INFO  backtype.storm.daemon.task - Emitting: entry-reader-spout __ack_init [5798661006293326143 -1761292315978577073 2]
getTuple
ret obj
Emitting entry:J1948+3540
39739 [Thread-15-entry-reader-spout] INFO  backtype.storm.daemon.task - Emitting: entry-reader-spout default [model.PSRCATEntry@2ddb685b]
39739 [Thread-15-entry-reader-spout] INFO  backtype.storm.daemon.task - Emitting: entry-reader-spout __ack_init [-2236948727380135894 -3945602349722429471 2]
```

Napotkane i rozwiązane problemy

- Błąd w klasie PSRCatEntry, uniemożliwiający prawidłowe wczytywanie danych z bazy (niektóre wartości w krotkach, np. okres, były NULL-ami).

- Wadliwe biblioteki w folderze ‘lib’ uniemożliwiały uruchomienie topologii na niektórych maszynach.

- Jeśli nawet węzeł odbierze wadliwą krotkę i nastąpi wyjątek, reszta topologii działa nadal bez zarzutu (sprawdzone w testach topologii).

Do wykonania

- Możliwość wydzielenia listy obiektów jako wzorcowych z którymi będą porównywane nowo przychodzące obiekty, w celu ich identyfikacji.

- Zapis wyników pracy topologii do pliku wyjściowego.

Dziękujemy za uwagę