Środowisko do efektywnego przetwarzania zdjęć satelitarnych w oparciu o technologię Big Data

Juliusz Pukacki, Marcin Krystek

[1)] *Poznańskie Centrum Superkomputerowo-Sieciowe*
office@man.poznan.pl

Na bazie zasobów obliczeniowych Poznańskiego Centrum Superkomputerowo-Sieciowego zaprojektowano i uruchomiono środowisko służące do przetwarzania zdjęć satelitarnych dużej skali. Celem tego projektu było stworzenie ekosystemu usług opartego technologie Big Data mające na celu realizację zaawansowanych scenariuszy analitycznych opartych o dane z ogólnie dostępnych repozytoriów programu Copernicus.

Technologie które można zaliczyć do szeroko rozumianego nurtu Big Data tworzone są ze szczególnym naciskiem na efektywność przetwarzania dużych i różnorodnych wolumenów danych. Rozwijane rozwiązania wspierają składowanie i udostępnianie surowych danych (HDFS), ich efektywne indeksowanie i wyszukiwanie (Elasticsearch, Hbase, Hive), a także współbieżne przetwarzanie (Spark). Wykorzystanie tych technologii pozwoliło na efektywne zastosowanie rozproszonej i skalowalnej mocy obliczeniowej do realizacji procesów związanych z przetwarzaniem danych geoprzestrzennych. W szczególności możliwe stało się skutecznie zaadresowanie wyzwań jakie niesie za sobą wciąż rosnący wolumen dostępnych danych oraz ich różnorodność, przy jednoczesnym skróceniu czasu niezbędnego na realizację przetwarzania. Jednocześnie istniejące narzędzia implementują uznane w świecie geograficznych systemów informacyjnych standardy pozwalając na swobodną interoperacyjność z innymi usługami i narzędziami. Wykorzystanie technologii i doświadczeń z obszaru Big Data w przetwarzaniu danych geoprzestrzennych, umożliwia zatem prowadzenie wszelkiego rodzaju analiz na większą skalę niż tam możliwa przy użyciu tradycyjnych narzędzi. Dodatkowo zapewniając mechanizmy elastycznego dostosowywania wielkości środowiska obliczeniowego do aktualnego zapotrzebowania.

W obszarze danych rastrowych szczególne zastosowanie znajduje framework RasterFrames bazujący na silniku analitycznym Apache Sark. Umożliwia on w szczególności wykonywanie efektywnych analiz czasoprzestrzennych oraz operacji rastrowych z obszaru algebry map. Otwiera również możliwość zastosowania technik z obszaru uczenia maszynowego poprzez bezpośrednie wykorzystanie pakietu Apache SparkML. RasterFrames skutecznie integruje funkcjonalności dostarczane przez komponenty dedykowane szczegółowym zastosowaniom. W zakresie algebry map, operacji przecięcia danych rastrowych i wektorowych oraz efektywnego dostępu do danych źródłowych bazuje na pakiecie GeoTrellis. W obszarze danych wektorowych ich skalowalne składowanie, indeksowanie i przetwarzanie możliwe jest dzięki pakietowi GeoMesa. Dzięki wykorzystaniu technologii takich jak Apache Spark i Kafka możliwe jest zrealizowane zarówno wsadowych jak i strumieniowych scenariuszy przetwarzania. Zgodność ze standardami OGC umożliwia natomiast bezpośrednią integrację z istniejącymi usługami jak np. GeoServer.

Wykorzystanie technik Big Data w przetwarzaniu danych geoprzestrzennych pozwoliło na przekroczenie barier technologicznych wynikających z ograniczonych zasobów dyskowych i pamięciowych pojedynczych komputerów i serwerów. Możliwość horyzontalnego skalowania infrastruktury oraz oparcie implementacji na technologiach efektywne wykorzystujących tak udostępnioną moc obliczeniową, przyczyniło się do znacznego poszerzenia skali możliwych do implementacji scenariuszy analitycznych dla danych geoprzestrzennych.