IBM SPSS Analytic Server
Wersja 3.2.1

Przegląd

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 5.

Informacje o produkcie

Niniejsze wydanie publikacji dotyczy wersji 3, wydania 2, modyfikacji1 produktu IBM SPSS Analytic Server oraz wszystkich następnych wersji i modyfikacji do czasu, aż w kolejnym wydaniu publikacji zostanie zawarta informacja o stosownej zmianie.

Spis treści

Przegląd

IBM® SPSS Analytic Server jest rozwiązaniem do analizy wielkich zbiorów danych (big data), które integruje technologię IBM SPSS z systemami wielkich zbiorów danych i umożliwia rozwiązywanie problemów na nieosiągalną dotąd skalę przy wykorzystaniu dobrze znanych interfejsów użytkownika oprogramowania IBM SPSS.

Dlaczego analizy wielkich zbiorów danych są takie ważne

Ilości danych zbierane przez organizacje rosną wykładniczo; na przykład przedsiębiorstwa z sektora finansowego i sektora handlu detalicznego przechowują dane wszystkich transakcji z klientami z ostatniego roku (lub dwóch albo dziesięciu...), operatorzy telekomunikacyjni przechowują rekordy połączeń (CDR) i odczyty z czujników w urządzeniach, firmy internetowe zbierają wyniki automatycznego przeszukiwania sieci przez roboty...

Analizy wielkich zbiorów danych są potrzebne wszędzie tam, gdzie:

- Istnieją duże ilości danych (rzędu terabajtów, petabajtów, eksabajtów), zwłaszcza gdy jest to kombinacja danych ustrukturyzowanych i nieustrukturyzowanych.
- Dane szybko się zmieniają/szybko przyrasta ich objętość.

Analizy wielkich zbiorów danych są także pomocne, gdy:

- Budowana jest duża liczba modeli (rzędu tysięcy).
- Modele są budowane/odświeżane z dużą częstotliwością

Wyzwania

Organizacje, które gromadzą duże ilości danych, napotykają często trudności w ich efektywnym wykorzystaniu. Może to być spowodowane różnymi przyczynami:

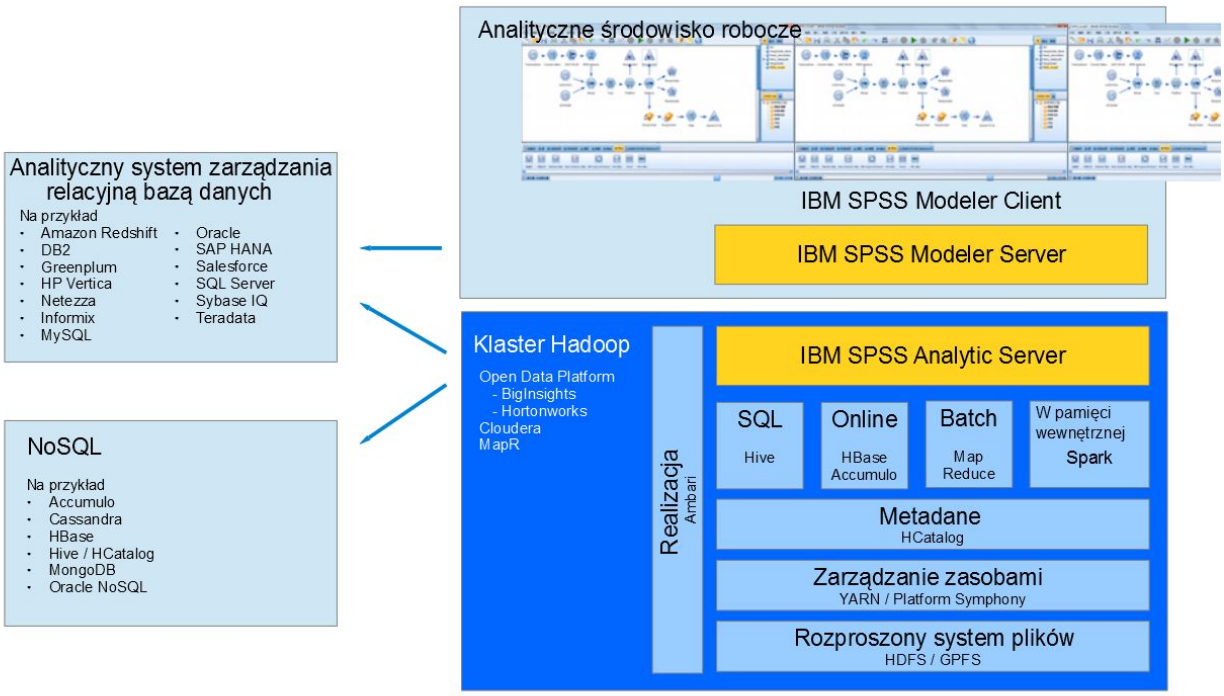
- Architektura tradycyjnych produktów analitycznych nie jest odpowiednia do przetwarzania rozproszonego.
- Istniejące algorytmy statystyczne nie zostały opracowane z myślą o operowaniu na wielkich zbiorach danych (algorytmy te oczekują, że będą wczytywać dane, jednak przemieszczanie wielkich zbiorów danych jest zbyt kosztowne).
- Do stosowania nowoczesnych technik analitycznych względem wielkich zbiorów danych potrzebne są nowe umiejętności i gruntowna znajomość systemów zarządzających takimi danymi. Bardzo niewielu analityków posiada odpowiednie kompetencje.
- Rozwiązania operujące na danych w pamięci wewnętrznej są odpowiednie do rozwiązywania problemów o średniej skali, ale źle poddają się skalowaniu do pracy z naprawdę wielkimi zbiorami danych.

Rozwiązanie

Analytic Server udostępnia:

- architekturę zorientowaną na dane, która korzysta z systemów operujących na wielkich zbiorach danych, takich jak Hadoop Map/Reduce z danymi rezydującymi w systemie plików HDFS;
- ściśle zdefiniowany interfejs umożliwiający uwzględnianie nowych algorytmów statystycznych do analizy danych;
- dobrze znane interfejsy użytkownika oprogramowania IBM SPSS, które ukrywają przed analitykiem szczegóły środowiska wielkich zbiorów danych, pozwalając mu skupić się na badanym problemie;
- rozwiązanie skalowalne praktycznie bez ograniczeń.

Architektura



Rysunek 1. Architektura

Analytic Server działa pomiędzy aplikacją kliencką a chmurą Hadoop. Przy założeniu, że dane rezydują w chmurze, ogólny schemat pracy z produktem Analytic Server obejmuje następujące etapy:

1. Zdefiniowanie źródeł danychAnalytic Server czerpiących z chmury.
2. Zdefiniowanie analiz, które mają być wykonane w aplikacji klienckiej. W bieżącej wersji aplikacją kliencką jest IBM SPSS Modeler.
3. Po uruchomieniu analizy aplikacja kliencka wysyła żądanie wykonania kierowane do Analytic Server.
4. Analytic Server koordynuje wykonanie zadania w chmurze Hadoop i zgłasza wyniki do aplikacji klienckiej.
5. Można wykorzystać te wyniki do zdefiniowania kolejnych analiz i powtórzyć cykl.

Spark i Analytic Server

Analytic Server może współpracować z mechanizmem Apache Spark, by działać wydajniej.

Kiedy mechanizm Spark jest, a kiedy nie jest używany

Jeśli mechanizm Spark jest zainstalowany jako usługa Ambari w klastrze Hadoop, programAnalytic Server używa go do przetwarzania wielkich zbiorów danych (big data). Przy podejmowaniu decyzji, czy mechanizm Spark ma być używany, obowiązują następujące kryteria.

1. Jeśli objętość zbioru danych jest mniejsza niż 128 MB, to Analytic Server używa funkcji MapReduce wbudowanej w maszynę JVM Analytic Server i nie korzysta z mechanizmu Spark ani klastra Hadoop.
2. Jeśli mechanizm Spark nie jest zainstalowany w klastrze, Analytic Server używa MapReduce v2.
3. Analytic Server używa MapReduce 2 do tworzenia modeli PSM. Gdy zadanie kończy się utworzeniem modelu PSM, Analytic Server używa mechanizmu Spark do realizacji wszystkich etapów zadania prowadzących do

powstania modelu, następnie zapisuje te pośrednie wyniki na dysk i używa MapReduce do zbudowania modelu PSM. Na przykład, jeśli zadanie obejmuje łączenie, po którym następuje tworzenie modelu PSM, łączenie realizowane jest przez mechanizm Spark, a następnie połączone dane są przetwarzane w MapReduce celem utworzenia modelu PSM.

Sposób korzystania z mechanizmu Spark

Gdy po uruchomieniu usługa Analytic Server wykryje dostępność mechanizmu Spark, inicjuje „zadanie Spark Hadoop", które umożliwia komunikację między rozproszonymi zadaniami w klastrze. Zadanie to działa tak długo, jak uruchomiona pozostaje usługa Analytic Server, i jest używane do wszystkich wykonań Analytic Server. Ta strategia umożliwia uzyskanie wydajności większej niż w przypadku koordynacji wielu osobnych zadań MapReduce Hadoop, ponieważ eliminuje narzut związany z ponownym ładowaniem wszystkich komponentów Analytic Server dla każdego zadania Hadoop.

Spark może wykonywać zadania MapReduce. Dzięki temu Analytic Server ma możliwość korzystania z własnych algorytmów mechanizmu Spark, takich jak łączenie, sortowanie i unia. Jednocześnie Analytic Server może uruchamiać istniejące algorytmy SPSS Map and Reduce w mechanizmie Spark, nie odwołując się bezpośrednio do interfejsu API Hadoop.

Co nowego w wersji 3.2.1

Wersja 3.2.1

Platforma

Obsługa platformy Hadoop Data Platform (HDP) 3.0 i 3.1.

Obsługa platformy Cloudera 6.0 i 6.1.

Funkcja rank

Funkcja **rank** służy do dzielenia zbioru danych wejściowych na odrębne podzbiory i wygenerowania nowej zmiennej z rangami poszczególnych wierszy podzbiorów. Funkcja ta jest podobna do następujących funkcji Hive: **rank()**, **dense_rank()** i **row_number()**.

Funkcje użytkownika — kierowanie przetwarzania do środowiska Hive

Wprowadzono nowe funkcje zdefiniowane przez użytkownika, przeznaczone dla środowiska Hive. Po zarejestrowaniu w bazie HiveDB funkcji zdefiniowanych przez użytkownika program Analytic Server może kierować przetwarzanie do środowiska Hive, korzystając z nowych funkcji użytkownika.

Najbardziej aktualne wymagania systemowe zawierają raporty ze szczegółowymi wymaganiami dostępne w serwisie WWW Wsparcia technicznego IBM: http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html. Na tej stronie:

1. Wpisz SPSS Analytic Server jako nazwę produktu i kliknij przycisk **Search** (Szukaj).
2. Wybierz żądaną wersję i zakres raportu, a następnie kliknij przycisk **Submit** (Wyślij).

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. Materiał ten jest również dostępny w IBM w innych językach. Jednakże w celu uzyskania dostępu do takiego materiału istnieje konieczność posiadania egzemplarza produktu w takim języku.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przesyłać na adres:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japonia

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów niż IBM zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjobiorcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów firm innych niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z produktami firm innych niż IBM. Pytania dotyczące możliwości produktów innych podmiotów niż IBM należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Wszystkie ceny podawane przez IBM są propozycjami cen detalicznych. Ceny te są aktualne i podlegają zmianom bez wcześniejszego powiadomienia. Ceny podawane przez dealerów mogą być inne.

Informacje te podano tylko w celu planowania. Wszelkie podane tu informacje mogą zostać zmienione zanim opisywane produkty staną się dostępne.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

LICENCJA W ZAKRESIE PRAW AUTORSKICH:

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Każda kopia programu przykładowego lub jakikolwiek jego fragment, jak też jakiekolwiek prace pochodne muszą zawierać następujące uwagi dotyczące praw autorskich:

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript oraz logo PostScript są znakami towarowymi lub zastrzeżonymi znakami towarowymi Adobe Systems Incorporated w Stanach Zjednoczonych i/lub w innych krajach.

IT Infrastructure Library jest zastrzeżonym znakiem towarowym agencji CCTA (Central Computer and Telecommunications Agency), wchodzącej w skład Departamentu Ministerstwa Skarbu Wielkiej Brytanii (Office of Government Commerce).

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium oraz Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i/lub w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

ITIL jest zastrzeżonym znakiem towarowym i zarejestrowanym znakiem wspólnotowym The Minister for the Cabinet Office, a także znakiem zarejestrowanym w U.S. Patent and Trademark Office.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Cell Broadband Engine jest znakiem towarowym Sony Computer Entertainment, Inc. w Stanach Zjednoczonych i/lub innych krajach, używanym tutaj na mocy udzielonej licencji.

Linear Tape-Open, LTO, logo LTO, Ultrium, a także logo Ultrium są znakami towarowymi HP, IBM Corp. oraz Quantum w Stanach Zjednoczonych i innych krajach.

Drukowane w USA