IBM SPSS Analytic Server
Version 3.2.1

Übersicht

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 5 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 3, Release 2, Modifikation 1 von IBM SPSS Analytic Server und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Inhaltsverzeichnis

Übersicht

IBM® SPSS Analytic Server ist eine Lösung für die Big Data-Analyse, bei der die IBM SPSS-Technologie mit Big Data-Systemen kombiniert wird und die Ihnen die Arbeit mit vertrauten IBM SPSS-Benutzerschnittstellen ermöglicht, um Probleme in einem zuvor nicht erreichten Maße lösen zu können.

Bedeutung von Big Data-Analysen

Die von Unternehmen erfassten Datenvolumen nehmen exponentiell zu. Dies umfasst bei Finanz- und Einzelhandelsunternehmen beispielsweise die gesamten Kundentransaktionen eines Jahres (bzw. von zwei oder zehn Jahren), bei Telekommunikationsprovidern die Anrufdatensätze und Sensormesswerte und bei Internetunternehmen die Ergebnisse von Websuchen.

Eine Big Data-Analyse ist erforderlich, wenn Folgendes vorliegt:

- Ein großes Datenvolumen (Terabyte, Petabyte, Exabyte), vor allem, wenn es sich um eine Mischung aus strukturierten und unstrukturierten Daten handelt
- Sich schnell ändernde/summierende Daten

Eine Big Data-Analyse ist außerdem in folgenden Situationen hilfreich:

- Es wird eine große Anzahl (Tausende) von Modellen erstellt
- Modelle werden häufig erstellt/aktualisiert

Herausforderungen

Unternehmen, die große Datenvolumen erfassen, haben in der Regel aus den unterschiedlichsten Gründen häufig Schwierigkeiten, einen tatsächlichen Nutzen aus diesen Daten zu ziehen:

- Die Architektur konventioneller Analyseprodukte ist nicht für die verteilte Verarbeitung geeignet.
- Vorhandene Statistikalgorithmen sind nicht für die Arbeit mit großen Datenmengen und großer Datenvielfalt (Big Data) vorgesehen (bei diesen Algorithmen wird erwartet, dass ihnen die Daten zugeführt werden, das Übertragen von Big Data ist jedoch zu kostenintensiv).
- Für die Durchführung modernster Analyseverfahren für Big Data sind neue Kenntnisse und Detailwissen in Bezug auf Big Data-Systeme erforderlich. Sehr wenige Analysten verfügen über solche Kenntnisse.
- Speicherinterne Lösungen funktionieren nur bei Datenvolumen bis zu einer mittleren Größe, sind für wirklich große Datenmengen und -vielfalt aber nicht gut geeignet.

Lösung

Analytic Server bietet Folgendes:

- Eine datenorientierte Architektur, die Big Data-Systeme nutzt, z. B. Hadoop Map/Reduce mit Daten in HDFS.
- Eine definierte Schnittstelle für die Integration von neuen statistischen Algorithmen, die so konzipiert sind, dass sie sich zu den Daten hinbewegen.
- Vertraute IBM SPSS-Benutzerschnittstellen, die die Details der Big Data-Umgebungen ausblenden, damit sich Analysten auf die Analyse der Daten konzentrieren können.
- Eine Lösung, die für Aufgabenstellungen beliebiger Größe skalierbar ist.

Architektur

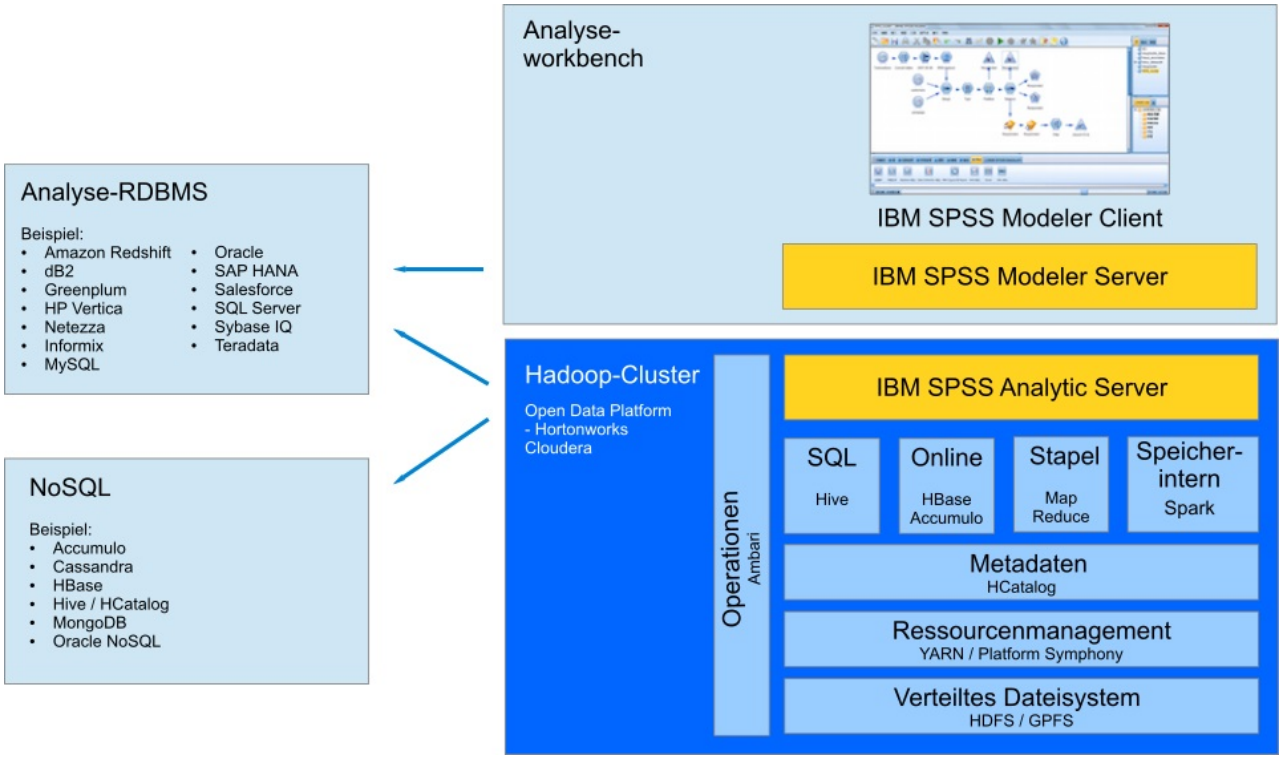

Abbildung 1. Architektur

Analytic Server befindet sich zwischen einer Clientanwendung und einer Hadoop-Cloud. Vorausgesetzt, die Daten befinden sich in der Cloud, gilt für das Arbeiten mit Analytic Server das folgende Schema:

1. Sie definieren Analytic Server-Datenquellen anhand der Daten in der Cloud.
2. Sie definieren die Analyse, die Sie in der Clientanwendung ausführen wollen. Für das aktuelle Release handelt es sich bei der Clientanwendung um IBM SPSS Modeler.
3. Beim Ausführen der Analyse übergibt die Clientanwendung eine Analytic Server-Ausführungsanforderung.
4. Analytic Server koordiniert den Job zur Ausführung in der Hadoop-Cloud und meldet die Ergebnisse an die Clientanwendung.
5. Die Ergebnisse können Sie zum Definieren weiterer Analysen verwenden, wobei sich der Zyklus wiederholt.

Spark und Analytic Server

Analytic Server wird zur Leistungssteigerung in Apache Spark integriert.

Kriterien für die Verwendung von Spark

Wenn Spark als Ambari-Service im Hadoop-Cluster installiert ist, wird es von Analytic Server zum Verarbeiten von Big Data-Jobs verwendet. Die folgenden Richtlinien bestimmen, wann Spark nicht verwendet wird.

1. Wenn das Dataset kleiner als 128 MB ist, verwendet Analytic Server nicht Spark oder den Hadoop-Clustern, sondern die integrierte MapReduce-Funktion in der Analytic Server-JVM.
2. Wenn Spark nicht im Cluster installiert ist, verwendet Analytic Server MapReduce Version 2.

3. Analytic Server verwendet MapReduce Version 2 zum Erstellen von PSM-Modellen. Wenn ein Job mit der Erstellung eines PSM-Modells endet, verarbeitet Analytic Server alle Schritte bis zur Modellerstellung mit Spark, schreibt dann auf die Festplatte und erstellt das PSM-Modell anschließend mit MapReduce. Beispiel: Wenn ein Job einen Join enthält, auf den eine Erstellung eines PSM-Modells folgt, wird der Join in Spark und PSM für die verknüpften Daten in MapReduce ausgeführt.

Verwendung von Spark

Wenn der Analytic Server-Service gestartet wird und erkennt, dass Spark verfügbar ist, initialisiert er einen "Spark-Hadoop-Job", der die Kommunikation zwischen verteilten Aufgaben im Cluster ermöglicht. Dieser Job ist während der gesamten Ausführung des Analytic Server-Service aktiv und wird für alle Ausführungen von Analytic Server verwendet. Dieser Ansatz verbessert die Leistung im Vergleich zum Orchestrieren mehrerer MapReduce-Hadoop-Jobs, da er den Aufwand zum erneuten Laden aller Analytic Server-Komponenten für die einzelnen Hadoop-Jobs beseitigt.

Spark kann MapReduce-Jobs ausführen. Dadurch wird Analytic Server die Verwendung von "nativen" Spark-Algorithmen, wie z. B. Join, Sortierung und Union-Verknüpfung, ermöglicht, wenn sie verfügbar sind. Gleichzeitig kann Analytic Server vorhandene Map- und Reduce-Algorithmen von SPSS in Spark ausführen, ohne direkt die Hadoop-API zu verwenden.

Neuerungen in Version 3.2.1

Version 3.2.1

Plattform

Unterstützung für Hadoop Data Platform (HDP) 3.0 und 3.1.

Unterstützung für Cloudera 6.0 und 6.1.

Rangfunktion

Die Funktion **rank** wird verwendet, um das Eingabedataset auf einzelne Partitionen aufzuteilen und ein neues Feld zu erstellen, in dem der Rang jeder Partitionszeile angezeigt wird. Das Feature ähnelt den Hive-Funktionen **rank()**, **dense_rank()** und **row_number()**.

Hive-UDF-Pushback

Neue Hive-UDF-Funktionen wurden eingeführt. Nach dem Registrieren des Hive-UDF in der HiveDB kann Analytic Server die neuen UDF-Funktionen verwenden, um ein Pushback durchzuführen.

Die aktuellen Informationen zu Systemanforderungen finden Sie in den Berichten mit den detaillierten Systemanforderungen auf der Site des IBM Technical Support unter http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html. Gehen Sie auf dieser Seite wie folgt vor:

1. Geben Sie SPSS Analytic Server als Produktnamen ein und klicken Sie auf **Search**.
2. Wählen Sie die gewünschte Version und den Berichtsumfang aus und klicken Sie dann auf **Submit**.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache kann eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich sein.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler enthalten. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Die angeführten Leistungsdaten und Kundenbeispiele dienen nur zur Illustration. Die tatsächlichen Ergebnisse beim Leistungsverhalten sind abhängig von der jeweiligen Konfiguration und den Betriebsbedingungen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Alle von IBM angegebenen Preise sind empfohlene Richtpreise und können jederzeit ohne weitere Mitteilung geändert werden. Händlerpreise können u. U. von den hier genannten Preisen abweichen.

Diese Veröffentlichung dient nur zu Planungszwecken. Die in dieser Veröffentlichung enthaltenen Informationen können geändert werden, bevor die beschriebenen Produkte verfügbar sind.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corporation in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite "Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

IT Infrastructure Library ist eine eingetragene Marke der Central Computer and Telecommunications Agency. Die Central Computer and Telecommunications Agency ist nunmehr in das Office of Government Commerce eingegliedert worden.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA oder anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

ITIL ist eine eingetragene Marke, eine eingetragene Gemeinschaftsmarke des Cabinet Office (The Minister for the Cabinet Office) und eine eingetragene Marke, die beim U.S. Patent and Trademark Office eingetragen ist.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Cell Broadband Engine wird unter Lizenz verwendet und ist eine Marke der Sony Computer Entertainment, Inc. in den USA und/oder anderen Ländern.

Linear Tape-Open, LTO, das LTO-Logo, Ultrium und das Ultrium-Logo sind Marken von HP, der IBM Corporation und von Quantum in den USA und/oder anderen Ländern.

Gedruckt in Deutschland