IBM SPSS Analytic Server
Версия 3.1.1

Обзор

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Замечания” на стр. 5.

Информация о продукте

Это издание применяется к версии 3, выпуску 1, модификации 1 IBM SPSS Analytic Server и ко всем последующим выпускам и модификациям, пока в новых изданиях не будет указано иного.

Содержание

Обзор

IBM® SPSS Analytic Server - это решение для анализа данных большого объема, в котором, благодаря сочетанию технологии IBM SPSS с системами больших данных, можно, работая со знакомыми пользовательскими интерфейсами IBM SPSS, решать задачи ранее недоступного масштаба.

Почему важен анализ больших данных

Объемы данных, собираемые организациями, растут экспоненциально; например, финансовые и торговые организации сохраняют все транзакции клиентов в течение года (или двух лет, или десяти лет), провайдеры телекоммуникаций хранят записи данных вызова (call data record, CDR) и показания сенсоров устройств, а интернет-компании хранят результаты веб-индексации.

Анализ больших данных нужен там, где есть:

- Большой объем данных (терабайты, петабайты, экзабайты), особенно если это смесь структурированных и неструктурированных данных
- Быстро меняющиеся или накапливающиеся данные

Кроме того, анализ больших данных полезен, когда:

- Строится большое число (тысячи) моделей
- Модели строятся или обновляются с высокой частотой

Вызовы

Организации, собирающие большие объемы данных, часто сталкиваются с трудностями при использовании своих же данных по ряду причин:

- архитектура традиционных продуктов анализа данных непригодна для распределенных вычислений и
- существующие статистические алгоритмы не предназначены для работы с большими данными (в этих алгоритмах предусматривается прием данных, но для больших данных такое перемещение слишком затратно), и поэтому
- анализ больших объемов данных при помощи существующих программ требует от аналитика новых умений и тесного знакомства с работой систем больших данных. Немногие аналитики обладают такой компетенцией.
- Аналитические решения, в которых данные загружаются в память, работают для задач среднего размера, но плохо масштабируются для по-настоящему больших данных.

Решение

Analytic Server обеспечивает:

- Архитектуру, ориентированную на данные, которая использует возможности систем больших данных, таких как Hadoop Map/Reduce с данными в HDFS.
- Специальный интерфейс, куда встроены новые статистические алгоритмы, переносящие обработку ближе к данным.
- Привычные пользовательские интерфейсы IBM SPSS, скрывающие подробности сред больших данных, чтобы аналитик мог сосредоточиться на анализе данных.
- Решение, которое можно масштабировать под задачи любого размера.

Архитектура

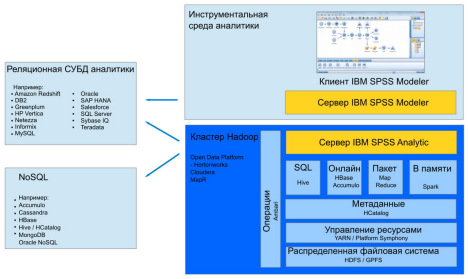

Рисунок 1. Архитектура

Analytic Server находится между клиентской программой и облаком Hadoop. В предположении, что данные находятся в облаке, общая схема работы с Analytic Server выглядит так:

1. Определить источники данных Analytic Server для данных в облаке.
2. Определить анализ, который вы хотите выполнить в клиентской программе. Для текущего выпуска клиентская программа - это IBM SPSS Modeler.
3. Когда вы запускаете анализ, клиентская программа передает требование выполнения Analytic Server.
4. Analytic Server организует задание для выполнения в облаке Hadoop и сообщает результаты клиентской программе.
5. Вы можете использовать эти результаты для дальнейшего анализа, и цикл повторяется.

Spark и Analytic Server

Для повышения производительности Analytic Server интегрируется с Apache Spark.

Когда Spark используется и когда не используется

Если установить Spark как службу Ambari в кластере Hadoop, то Analytic Server использует эту службу для обработки заданий больших данных. Ниже приведены указания, как узнать, в каких случаях Spark не используется.

1. Если набор данных меньше 128 Мбайт, то Analytic Server использует встроенную функцию MapReduce в Analytic Server JVM и не использует Spark или кластер Hadoop.
2. Если не установить Spark в кластере, то Analytic Server будет использовать MapReduce v2.

3. Analytic Server использует MapReduce v2 для построения моделей PSM. Если задание завершается построением модели PSM, то Analytic Server использует Spark для обработки задания на всех шагах, предшествующих построению модели, затем записывает нужные данные на диск, а затем использует MapReduce для построения модели PSM. Например, если задание включает в себя операцию соединения (join), а затем построение модели PSM, это соединение выполняется в Spark, а PSM работает с соединенными данными в MapReduce.

Как используется Spark

После того, как служба Analytic Server запускается и обнаруживает доступность Spark, она инициирует задание "Spark Hadoop job", чтобы поддерживать обмен информацией между распределенными задачами в кластере. Это задание выполняется в течение всего времени работы службы Analytic Server и используется для всех запусков Analytic Server. Такой подход повышает производительность по сравнению с руководством несколькими заданиями MapReduce Hadoop, поскольку устраняет дополнительные затраты на повторную загрузку всех компонентов Analytic Server для каждого задания Hadoop.

Spark поддерживает запуск заданий MapReduce. Благодаря этому Analytic Server может использовать "собственные" алгоритмы Spark, такие как соединение (join), сортировка (sort) и объединение (union), где это доступно. В то же время Analytic Server может выполнять в Spark существующие алгоритмы SPSS Map и Reduce без непосредственного использования API Hadoop.

Что нового в версии 3.1.1

Версия 3.1.1

Платформа

- Поддержка Cloudera 5.11 и 5.12
- Поддержка Ubuntu Linux 16.04 (с Hortonworks Data Platform 2.6 и Cloudera 5.11)
- Cloudera 5.8 и 5.9 более не поддерживаются
- Big Insights 4.1, 4.2 и 4.2.5 более не поддерживаются
- MapR 5.0 более не поддерживается

Источники данных

- Поддержка Apache Hive 2.1
- MongoDB 2.6 более не поддерживается
- MySQL 5.1 более не поддерживается

Улучшение производительности

- Для HDP теперь доступна более автоматизированная процедура автономной установки. Дополнительную информацию смотрите в разделе Автономная установка
- Поддержка нескольких кластеров. Поддержка нескольких кластеров - это усовершенствование возможности высокой доступности IBM SPSS Analytic Server, и она обеспечивает улучшенную изоляцию в мультиарендных средах. По умолчанию при установке службы Analytic Server (и на Ambari, и на ClouderaManager) создается определение единственного сервера аналитики.
- Теперь вы можете сконфигурировать отдельные очереди YARN для каждого арендатора Analytic Server. Дополнительную информацию смотрите в разделе Конфигурирование отдельных очередей YARN для каждого арендатора Analytic Server - HDP или Конфигурирование отдельных очередей YARN для каждого арендатора Analytic Server - Cloudera.
- Теперь обеспечивается поддержка для режима pushback SQL выборки.
- Теперь доступны следующие усовершенствования Spark ML:
  - Поддержка алгоритма bisecting K-means
  - Поддержка XGBoost
  - Поддержка изотонической регрессии

- Перекрестный источник данных Spark RDD теперь можно использовать совместно с различными заданиями ASL.

- Раздел Рекомендуемые приемы и рекомендации IBM SPSS Analytic Server обновлен и содержит теперь информацию, относящуюся к Cloudera.

Самую свежую информацию о требованиях к системе смотрите в подробных отчетах о требованиях к системе на сайте технической поддержки IBM: http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html. На этой странице:

1. Введите в качестве имени продукта SPSS Analytic Server и нажмите кнопку **Search** (Поиск).
2. Выберите нужную версию и область отчета, затем нажмите кнопку **Submit** (Передать).

Замечания

Эта публикация разрабатывалась для продуктов и услуг, предлагаемых в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

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

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

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

КОРПОРАЦИЯ INTERNATIONAL BUSINESS MACHINES ПРЕДОСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ ГАРАНТИЙ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПОДРАЗУМЕВАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ КАКОЙ-ЛИБО КОНКРЕТНОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в данной информации на сайты, не принадлежащие IBM, приводятся только для удобства и никоим образом не означают поддержки этих сайтов. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

IBM может использовать или распространять предоставленную вами информацию любым способом, как фирма сочтет нужным, без каких-либо обязательств перед вами.

Если обладателю лицензии на данную программу понадобится информация о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

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

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Упоминаемые данные о производительности и примеры клиентов представлены только для иллюстративных целей. Фактические результаты производительности могут быть иными в зависимости от определенных конфигураций и конкретных условий.

Информация, касающаяся продуктов других компаний (не IBM) была получена от поставщиков этих продуктов, из опубликованных ими заявлений или из прочих общедоступных источников. IBM не проводила тестирования этой продукции и не может подтвердить или опровергнуть информацию о точности ее работы и совместимости, а также другие заявления относительно продуктов других производителей (не IBM). Вопросы относительно возможностей продуктов других компаний (не IBM) следует адресовать поставщикам этих продуктов.

Утверждения, касающиеся намерений и планов IBM, могут быть изменены без предварительного предупреждения; они приведены здесь только для обозначения целей и задач IBM.

Все упоминаемые цены IBM - это рекомендуемые розничные цены IBM на текущий момент; они могут быть изменены без уведомления. Цены дилеров могут отличаться.

Эта информация приводится только для целей планирования. Приведенная информация может быть изменена до того, как описанные продукты станут доступными.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия вымышлены и любое их сходство с реальными именами и названиями компаний полностью случайно.

ЛИЦЕНЗИЯ НА КОПИРОВАНИЕ:

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия вымышлены и любое их сходство с реальными именами и названиями компаний полностью случайно.

Каждая копия или каждая часть этих примеров программ или работы, основанной на них, должна содержать следующее замечание об авторских правах:

© (название вашей компании) (год). Части этого кода получены из примеров программ IBM Corp.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM можно найти в Интернете в разделе "Copyright and trademark information" ("Информация об авторских правах и товарных знаках") по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

IT Infrastructure Library представляет собой зарегистрированный товарный знак Central Computer and Telecommunications Agency, являющейся теперь частью Office of Government Commerce.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

ITIL - зарегистрированный товарный знак и зарегистрированный товарный знак сообщества секретариата кабинета министров (Minister for the Cabinet Office) Великобритании, и он зарегистрирован в Бюро по регистрации патентов и торговых марок США (U.S. Patent and Trademark Office).

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Cell Broadband Engine является товарным знаком корпорации Sony Computer Entertainment в Соединенных Штатах и других странах, и используется по лицензии, выдаваемой там.

Linear Tape-Open, LTO, логотип LTO Logo, Ultrium и логотип Ultrium являются товарными знаками корпораций HP, IBM и Quantum в США и других странах.

Напечатано в Дании