IBM SPSS Analytic Server
Version 3.2.1

Overview

Note

Before using this information and the product it supports, read the information in “Notices” on page 5.

Product Information

This edition applies to version 3, release 2, modification 1, fix 1 of IBM® SPSS® Analytic Server and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Overview

IBM SPSS Analytic Server is a solution for big data analytics that combines IBM SPSS technology with big data systems and allows you to work with familiar IBM SPSS user interfaces to solve problems on a previously unattainable scale.

Why big data analytics matters

Data volumes collected by organizations are growing exponentially; for example, financial and retail businesses have all customer transactions for a year (or two years, or ten years), telco providers have call data records (CDR) and device sensor readings, and internet companies have the results of web crawls.

Big data analytics is needed where there exists:

- A large volume of data (terabytes, petabytes, exabytes), especially when it is a mixture of structured & unstructured data
- Rapidly changing/accumulating data

Big data analytics also assists when:

- A large number (thousands) of models are being built
- Models are frequently built/refreshed

Challenges

The same organizations that collect large volumes of data often have difficulty actually making use of it, for a variety of reasons:

- The architecture of traditional analytic products are not suited to distributed computation, and
- Existing statistical algorithms are not designed to work with big data (these algorithms expect the data to come to them, but big data is too costly to move), thus
- Performing state of the art analytics on big data requires new skills and intimate knowledge of big data systems. Very few analysts have these skills.
- In-memory solutions work for medium-size problems, but do not scale well to truly big data.

Solution

Analytic Server provides:

- A data-centric architecture that leverages big data systems, such as Hadoop Map/Reduce with data in HDFS.
- A defined interface to incorporate new statistical algorithms designed to go to the data.
- Familiar IBM SPSS user interfaces that hide the details of big data environments so that analysts can focus on analyzing the data.
- A solution that is scalable to any size problem.

Architecture

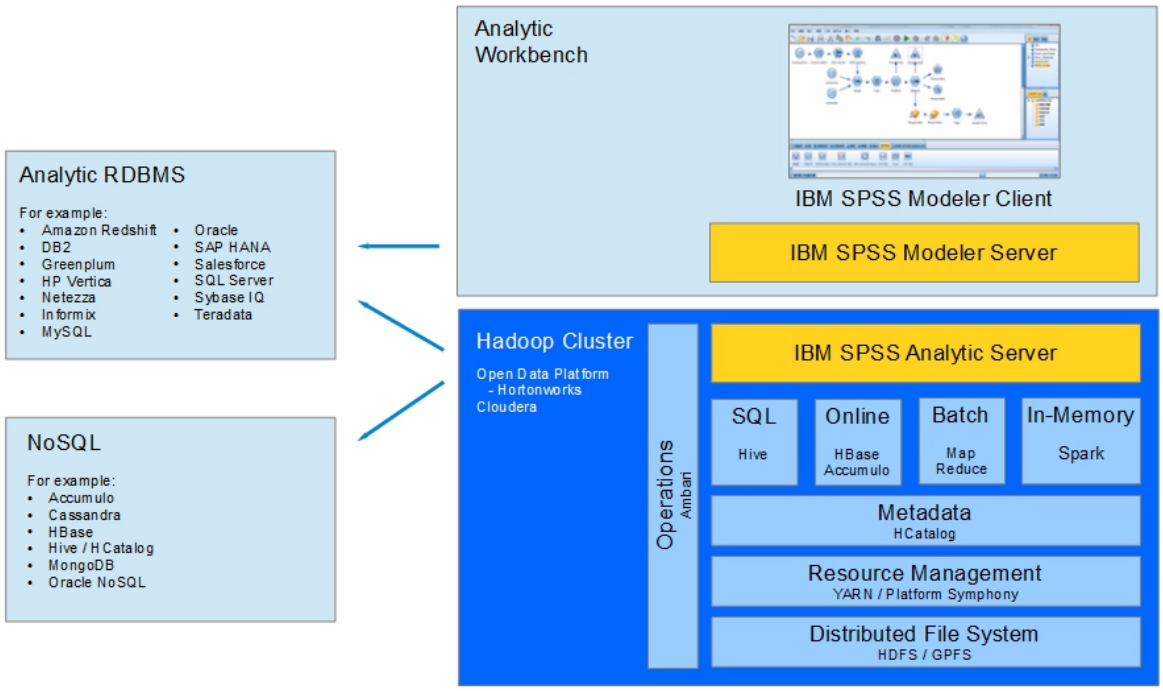

Figure 1. Architecture

Analytic Server sits between a client application and Hadoop cloud. Assuming that the data resides in the cloud, the general outline for working with Analytic Server is to:

1. Define Analytic Server data sources over the data in the cloud.
2. Define the analysis you want to perform in the client application. For the current release, the client application is IBM SPSS Modeler.
3. When you run the analysis, the client application submits an Analytic Server execution request.
4. Analytic Server orchestrates the job to run in the Hadoop cloud and reports the results to the client application.
5. You can use the results to define further analysis, and the cycle repeats.

Spark and Analytic Server

Analytic Server integrates with Apache Spark to increase performance.

When Spark is and is not used

If Spark is installed as an Ambari service on the Hadoop cluster, then Analytic Server uses it to process big data jobs. The following guidelines apply to determine when Spark is not used.

1. If the data set is smaller than 128MB, then Analytic Server uses the embedded MapReduce function in the Analytic Server JVM and does not utilize Spark or the Hadoop cluster.
2. If Spark is not installed on the cluster, then Analytic Server uses MapReduce v2.
3. Analytic Server uses MapReduce v2 to build PSM models. When a job ends with a PSM model build, Analytic Server uses Spark to process the job through all steps leading to the model build, then write to

disk, and then use MapReduce to build the PSM model. For example, if a job includes a join followed by a PSM model build, the join runs in Spark and the PSM runs on the joined data in MapReduce.

How Spark is used

When the Analytic Server service is started and discovers that Spark is available, it initializes a "Spark Hadoop job" that allows communication between distributed tasks across the cluster. This job runs for as long as the Analytic Server service runs, and is used for all Analytic Server executions. This approach improves performance relative to orchestrating multiple MapReduce Hadoop jobs, because it eliminates the overhead of reloading all Analytic Server components for each Hadoop Job.

Spark is capable of running MapReduce jobs. This allows Analytic Server to use "native" Spark algorithms such as join, sort, and union where available. At the same time, Analytic Server can run existing SPSS Map and Reduce algorithms in Spark, and without directly using the Hadoop API.

What is new in version 3.2.1 Fix Pack 1

Version 3.2.1 Fix Pack 1

Note: Refer to IBM SPSS Analytic Server 3.2.1 Fix Pack 1 download document for information on upgrading Analytic Server 3.2.1 to Analytic Server 3.2.1 Fix Pack 1.

Platform
Apache Spark 1.x is no longer supported. 2.x is required.

MultiClient support - configure a separate Dynamic Resource Allocation for each YARN Resource Pool
Separate YARN resource pools are now generated for each Spark job/stream. This means that separate YARN resource pools can now be created for every user (`user` mode) or tenant (`tenant` mode). Spark job impersonation at the YARN level is now based on the actual user (not the `as_user`).

Note: When user mode is used on non-Kerberos systems, a process system user name that matches the name of the signed-in Analytic Server user.

For more information, refer to the *"Configuring a separate Dynamic Resource Allocation for each YARN Resource Pool - HDP"* or *"Configuring a separate Dynamic Resource Allocation for each YARN Resource Pool - Cloudera"* sections in the IBM SPSS Analytic Server 3.2.1 Installation Guide.

Version 3.2.1

Platform
Support for Hortonworks Data Platform (HDP) 3.0 and 3.1.

Note: The **HCatalog** type is not available in the **Admin Console > New data** source dialog for HDP 3.0 and 3.1.

Support for Cloudera 6.0 and 6.1.

Rank function
The **`rank`** function is used to split the input data set into separate partitions and generate a new field that shows the rank of each partition row. The feature is similar to the Hive functions **`rank()`**, **`dense_rank()`**, and **`row_number()`**.

UDF Hive pushback
New Hive UDF functions have been introduced. After Hive UDF is registered to HiveDB, Analytic Server can use the new UDF functions to perform pushback.

For the most up-to-date system requirements information, use the Detailed system requirements reports at the IBM Technical Support site: http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html. On this page:

1. Type SPSS Analytic Server as the product name and click **Search**.
2. Select the wanted version and scope of report, then click **Submit**.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

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

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

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

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

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

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of The Minister for the Cabinet Office, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.