IBM SPSS Analytic Server
Version 1.0.1

User's Guide

Note
Before using this information and the product it supports, read the information in "Notices" on page 11.

Product Information

This edition applies to version 1, release 0, modification 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

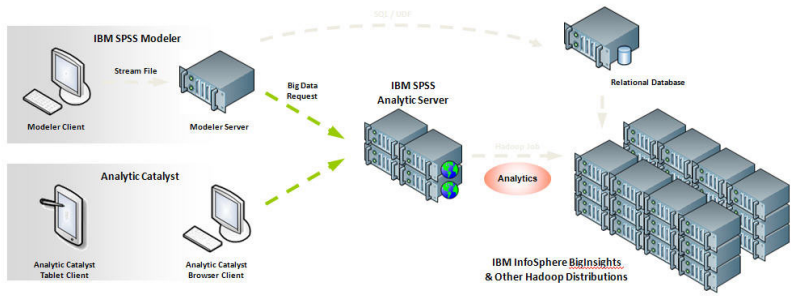

Figure 1. Architecture

The Analytic Server sits between a client application and Hadoop cloud. Assuming that the data resides in the cloud, the general outline for working with the 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 applications are IBM SPSS Modeler and IBM SPSS Analytic Catalyst.
3. When you run the analysis, the client application submits an Analytic Server execution request.
4. The 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 analyses, and the cycle repeats.

Chapter 2. Analytic Server console

Analytic Server provides a thin client interface for managing data sources and projects.

1. Enter the URL of the Analytic Server in your browser's address bar. This can be obtained from your server administrator.
2. Enter the user name with which to log on to the server.
3. Enter the password associated with the specified user name.

After login, the default accordion is the Data sources accordion.

Navigating the console

The Analytic Server console has four components:

- The header displays the product name, the link to the help system, and the name of the currently logged in user. The name of the currently logged in user is the head of a dropdown list that includes the logout link and a link to general information about the product.
- The left column displays the available accordions, or functional groupings. The selected accordion determines what is shown in the content area.
- The content area displays the controls associated with the currently selected accordion. Details of each accordion's contents follow in the sections below.
- The footer displays the installed version of the Analytic Server.

Data sources

A data source is a collection of records, plus a data model, that define a data set for analysis. The source of records can be a file (delimited text, fixed width text, Excel) on HDFS, a database, or HCatalog. The data model defines all the metadata (field names, storage, measurement level, and so on) necessary for analyzing the data. Data source owners can grant or restrict access to data sources.

Left column

The left column displays the existing data sources under the accordion heading.

- Select a data source to display its details in the content area and edit its properties. Typing in the search area filters the listing to display only data sources with the search string in their name.
- Click the **New data source** button to create a new data source with the name and content type you specify in the **Add New Data Source** dialog.
  - Data source names must be case-sensitive. Leading and trailing white space is ignored. Certain names are rejected to protect against SQL injection.
  - The available content types are File, Database, and HCatalog.

    Note: The content type cannot be edited once selected.
- Click the **Delete data source** button to remove the data source. This action leaves all files associated with the data source intact.

Content area

The content area is divided into several sections, which can depend on the content type of the data source. After you specify the settings for the data source, click Preview and Metadata to check the data source specifications, then click **Save** to save your work.

Data Source Properties

Settings common to all content types.

Name An editable text field that displays the name of the data source.

Description

An editable text field to provide explanatory text about the data source.

Is public

A check box that indicates whether anyone can see the data source (checked) or if users and groups must be explicitly added to the owners list (cleared).

Sharing

You can share ownership of a data source by adding users and groups as authors.

- Typing in the text box filters on users and groups with the search string in their name. Click the **Add participant** button to add them to the list of authors.
- To remove an author, select a user or group in the Author list and click the **Remove participant** button.

Note: Authors will only have read access to the data source, unless they are Owners on the Project to which the data source has write access.

Note: Administrators have read and write access to every data source, regardless of whether they are listed as an Author on the data source or Owner on the Project.

File Input

Settings that are specific to defining data sources with file content type.

File Viewer

Shows available files for inclusion in the data source. Select **Projects** mode to view files within the Analytic Server project structure, or **HDFS** to view the rest of the Hadoop distributed file system. You can browse either folder structure, but HDFS is not editable at all, and the Analytic Server folder structure is only editable within defined projects. That is, you cannot add files, create folders, or delete items at the root level of the **Projects** mode. To create, edit, or delete a project, use the Projects accordion.

- Clicking the **Upload file to HDFS** button uploads a file to the current project/subfolder.
- Clicking the **Create a new folder** button creates a new folder under the current folder, with the name you specify in the New Folder Name dialog.
- Clicking the **Download file to the local filesystem** button downloads the selected files to the local file system.
- Clicking the **Delete the selected file(s)** button removes the selected files/folders.

Files included in data source definition

Use the move button to add selected files to or remove them from the data source. For each selected file in the data source, click Settings to define the specifications for reading the file.

File Output

Data sources with file content type can be appended to by output from streams that are run on Analytic Server. Select **Make writeable** to enable appending and choose an output folder where the new files are written. Select **Newlines can be escaped** to cause newlines in the data to be written as the string "\n" in the output file and the string "\n" to be written as "\\n" in the output file. If unselected, the string "\n" is written as "\n" in the output file and the presence of a newline will cause an error.

Database Selections

Specify the connection parameters for the database that contains the record content.

Database
Select the type of database you want to connect to. Choose from: DB2, Oracle, SQL Server, TeraData, or Netezza.

Server address
Enter the URL of the server that hosts the database.

Server port
The port number that the database listens on.

Database name
The name of the database you want to connect to.

Username
If the database is password-protected, enter your user name.

Password
If the database is password-protected, enter your password.

Table name
Enter the name of a table from the database that you want to use.

Maximum concurrent reads
Enter the limit on the number of parallel queries that can be sent from Analytic Server to the database to read from the table specified in the data source.

Database Output
Data sources with database content type can be appended by output from streams that are run on Analytic Server. Select **Make writeable** to enable appending and choose an output database table where the output data are written.

HCatalog Selections
Specify the parameters for accessing data that are managed under Apache HCatalog.

Database
The name of the HCatalog database.

Table name
Enter the name of a table from the database that you want to use.

Filter The partition filter for the table, if the table was created as partitioned table. HCatalog filtering is supported only on Hive partition keys of type string.

Note: Filtering is applied when you click **Preview Raw Data**; however, filtering does not take effect until you have clicked **Save** to save the data source. If you enter an invalid filter value and no data is returned when previewing the raw data, you need to remove the filter values and click **Save** in order to preview the unfiltered raw data.

HCatalog Schema
Displays the structure of the specified table. HCatalog can support a highly structured data set. To define an Analytic Server data source on such data, the structure must be flattened into simple rows and columns. Select an element in the schema and click the move button to map it to a field for analysis. Not all tree nodes can be mapped. For example, an array or map of complex types is considered a "parent" and cannot be mapped. These nodes can be identified by the label in the tree ending in ...:array:struct, or ...:map:struct.

HCatalog Field Mappings
Displays the mapping of an element in HCatalog to a field in the data source. Click **Preview Raw Data** to see the records as they are stored in HCatalog; this can help you determine how to map the HCatalog schema to fields.

HCatalog Element
Double-click a cell to edit. You must edit the cell when the HCatalog element is an array or map. With an array, specify the integer corresponding to the member of the array you want to map to a field. With a map, specify a quoted string corresponding to the key you want to map to a field. See Figure 2 for an example of how the raw data preview can be used to determine the string corresponding to the map index.

Mapping Field
The field as it appears in the Analytic Server data source. Double-click a cell to edit. Duplicate values in the Mapping Field column are not allowed and result in an error.

Storage
The storage of the field. Storage is derived from HCatalog and cannot be edited.

Note: When you click Preview and Metadata to finalize an HCatalog data source, there are no editing options.

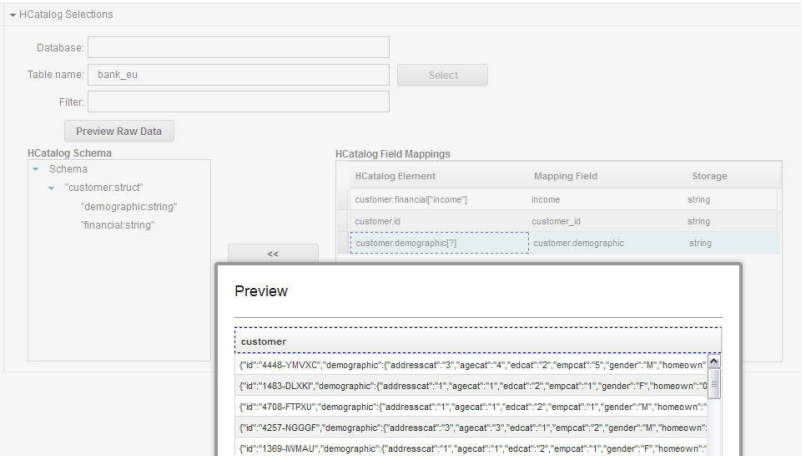

Figure 2. Data sources accordion, defining an HCatalog data source

Settings (file data sources)

Character set encoding
The character encoding of the file. Select or specify Java charset name such as "UTF-8", "ISO-8859-2", "GB18030". The default is **UTF-8**.

Locale Defines a locale. Optional. Defaults to the server locale. The locale string should be specified as: <language>[_country[_variant]], where:

language
A valid, lower-case, two-letter code as defined by ISO-639. Required.

country
A valid, upper-case, two-letter code as defined by ISO-3166. Optional.

variant
A vendor or browser-specific code. Optional.

Trim white space
Removes white space characters from the beginning and/or end of the string fields. Defaults to **None**. The following values are supported:

None Does not remove white space characters.

Left Removes white space characters from the beginning of the string.

Right Removes white space characters from the end of the string.

Both Removes white space characters from the beginning and end of the string.

Grouping symbols
Sets whether or not the locale-specific character used for the thousands separator should be used.

Field delimiters
One or more characters marking field boundaries. Each character is taken as an independent delimiter. For example, if you select **Comma** and **Tab** (or select **Other** and type `,\t`), it means that either a comma or a tab marks field boundaries. If control characters delimit fields, the characters specified here are treated as delimiters in addition to control characters. Default is "," if control characters do not delimit fields; otherwise the default is the empty string.

Control characters delimit fields
Sets whether ASCII control characters, except LF and CR, are treated as field delimiters. Defaults to **No**.

First row contains field names
Sets whether to use the first row to determine the field names. Defaults to **No**.

Number of initial characters to skip
The number of characters at the beginning of the file to be skipped. A non-negative integer. Default is `0`.

Merge white space
Sets whether to treat multiple adjacent occurrences of space and/or tab as a single field delimiter. Has no effect if neither space nor tab is a field delimiter. Default is **Yes**.

End-of-line comment characters
One or more characters that mark end-of-line comments. The character and everything following it on the record are ignored. Each character is taken as an independent comment marker. For example, "/*" means either a slash or an asterisk starts a comment. It is not possible to define multi-character comment markers like "//". The empty string signals that no comment characters are defined. If defined, comment characters are checked for before quotes are processed or initial characters to skip are skipped. Default is the empty string.

Invalid characters
Determines how invalid characters (byte sequences that do not correspond to characters in the encoding) are to be handled. An empty string indicates they are to be discarded. A non-empty string (usually a single character) indicates they are to be replaced by the contents of the string. Default is the empty string.

Single quotes
Specifies handling of single quotes (apostrophes). Default is **Keep**.

Keep Single quotes have no special meaning and are treated as any other character.

Drop Single quotes are deleted unless quoted

Pair Single quotes are treated as quote characters and characters between pairs of single quotes lose any special meaning (they are considered quoted). Whether single quotes themselves can occur inside single-quoted strings is determined by the setting **Quotes can be quoted by doubling**.

Double quotes
Specifies handling of double quotes. Default is **Pair**.

Keep Double quotes have no special meaning and are treated as any other character.

Drop Double quotes are deleted unless quoted

Pair Double quotes are treated as quote characters and characters between pairs of double

quotes lose any special meaning (they are considered quoted). Whether double quotes themselves can occur inside double-quoted strings is determined by the setting **Quotes can be quoted by doubling**.

Quotes can be quoted by doubling
Indicates whether double quotes can be represented in double-quoted strings and single quotes can be represented in single-quoted strings when set to **Pair**. If **Yes**, double quotes are escaped inside double-quoted strings by doubling and single quotes are escaped inside single-quoted strings by doubling. If **No**, there is no way to quote a double quote inside a double-quoted string or a single quote inside a single-quoted string. Default is **Yes**.

Newlines can be escaped
Indicates whether the parser interprets the string "\n" as a new line when reading a file. If newlines are not escaped, then "\n" is simply read as a string. If newlines are escaped, then "\n" is read as and ASCII newline character and "\\n" is read as the string "\n". Default is **Yes**.

Preview and Metadata (data sources)

Clicking **Preview and Metadata** displays a sample of records and the data model for the data source. Here you have a chance to review the basic metadata information.

Preview
The Preview tab shows a small sample of records and their field values.

Edit

The Edit tab displays the basic field metadata. For data sources with Files content type, the data model is generated from a small sample of records, and you can manually edit the field metadata on this tab. For data sources with HCatalog content type, the data model is generated based upon the HCatalog Field Mappings, and you cannot edit the field storage on this tab.

Field Double-click on the field name to edit it.

Measurement
This is the measurement level, used to describe characteristics of the data in a given field.

Role Used to tell modeling nodes whether fields will be Input (predictor fields) or Target (predicted fields) for a machine-learning process. Both and None are also available roles, along with Partition, which indicates a field used to partition records into separate samples for training, testing, and validation. The value Split specifies that separate models will be built for each possible value of the field. Frequency specifies that a field values should be used as a frequency weight for each record. Record ID is used to identify a record in the output.

Storage
Storage describes the way data are stored in a field. For example, a field with values of 1 and 0 stores integer data. This is distinct from the measurement level, which describes the usage of the data, and does not affect storage. For example, you may want to set the measurement level for an integer field with values of 1 and 0 to Flag. This usually indicates that 1 = True and 0 = False.

Projects

Projects are workspaces for storing inputs and accessing outputs of jobs. They provide the top-level organizational structure for containing files and folders. Projects can be shared with individual users and groups.

Left column

The left column displays the existing projects under the accordion heading.

- Select a project to display its details in the content area and edit its properties. Typing in the search area filters the listing to display only projects with the search string in their name.
- Click **New Project** to create a new project with the name you specify in the Add New Project dialog. Names are case-sensitive, ignore leading and trailing white space, and protect against SQL injection.
- Click **Delete Project** to remove the project. This action leaves all files that are associated with the data source intact.

Content area

The content area is divided into **Settings**, **Owners**, and **Version** tabs.

Settings

Project description
An editable text field to provide explanatory text about the project.

Is public
A check box that indicates whether anyone can see the project (checked) or if users and groups must be explicitly added to as owners (cleared).

Clicking **Save** saves the current state of the settings.

Project data sources
A non-editable area that lists all the data sources that are associated with the project.

Project structure pane

The right pane shows the project/folder structure for the currently selected project. You can browse the folder structure, but it is not editable, except through the buttons.

- Click **Download file to the local filesystem** to download a selected file to the local file system.
- Click **Delete the selected file(s)** to remove the selected file/folder.

Owners

Owners are full members of a project, and can modify the project as well as the folders and files within it.

The Available users and groups list displays users and groups in the active tenant that are not currently associated with this project.

- Typing in the search pane filters on users and groups with the search string in their name.
- Selecting the users icon above the list shows the available users. Clearing the icon hides the users. This icon is selected by default.
- Selecting the groups icon above the list shows the available groups. Clearing the icon hides the groups. This icon is selected by default.

Users and groups can be moved to the Project Users and Groups using the move button. These users and members of these groups have write (Analytic Server Export node) access to this project when connecting to Analytic Server through IBM SPSS Modeler.

Note: Changes made on the Owners tab are immediately and automatically applied.

Note: Administrators have read and write access to every project, regardless of whether they are specifically listed as an owner.

Versions

Projects are versioned based on changes to the file and folder contents. Changes to a project's attributes, such as the description, whether it is public, and with whom it is shared, do not require a new version. Adding, modifying, or deleting files or folders does require a new version.

Project versioning table

The table displays the existing project versions, their creation and commit dates, the users responsible for each version, and the parent version. The parent version is the version upon which the selected version is based.

- Click **Lock** to make changes to the selected project version contents.
- Click **Commit** to save all changes that are made to a project and make this version the current visible state of the project.
- Click **Discard** to discard all changes that are made to a locked project and return the visible state of the project to the most recently committed version.
- Click **Delete** to remove the selected version.

Automatically clean up when number of versions exceeds

Automatically deletes the oldest committed project version when the number of versions exceeds the specified number. The default is 25.

Notices

This information was developed for products and services offered in the U.S.A.

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
Armonk, NY 10504-1785
U.S.A.

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.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: 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 states 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 Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply 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 Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All 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 the names and addresses used by an actual business enterprise is entirely coincidental.

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

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 the names and addresses used by an actual business enterprise is entirely coincidental.

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

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

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.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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.

Printed in USA