Cloudant

Import data from a Cloudant database into SPPS Modeler

Product: IBM® SPSS® Modeler

Extension type: Utility

Table of Contents

Description:

This node allows you to import data from a Cloudant database to be used in IBM SPSS Modeler. You can retrieve a dump of the whole database, retrieve a document from its ID or perform a query using a preset view. In order to use this node you must have a Bluemix account with the Cloudant service added. By adding this service you will create "host" credentials that are needed by this node. After adding the service in Bluemix you can access the Cloudant dashboard to create or update databases.

Requirements:

- SPSS Modeler v16.0 or later
- R: http://www.r-project.org/
- 'R Essentials for SPSS Modeler' plugin: https://developer.ibm.com/predictiveanalytics/downloads/
- A Bluemix account with Cloudant activated either as a service or as part of an application

Installation:

Close SPSS Modeler. Save the .cfe file in the CDB folder of the IBM SPSS Modeler installation directory for Windows and Linux. The copy should reside in that same folder and not in a sub-folder.

For example, for Windows 7 the default location is "C:\ProgramData\IBM\SPSS\Modeler\16\CDB". If the ProgramData folder is hidden type the path manually.

Restart SPSS Modeler: the node will now appear in the Record Ops palette.

R Packages used:

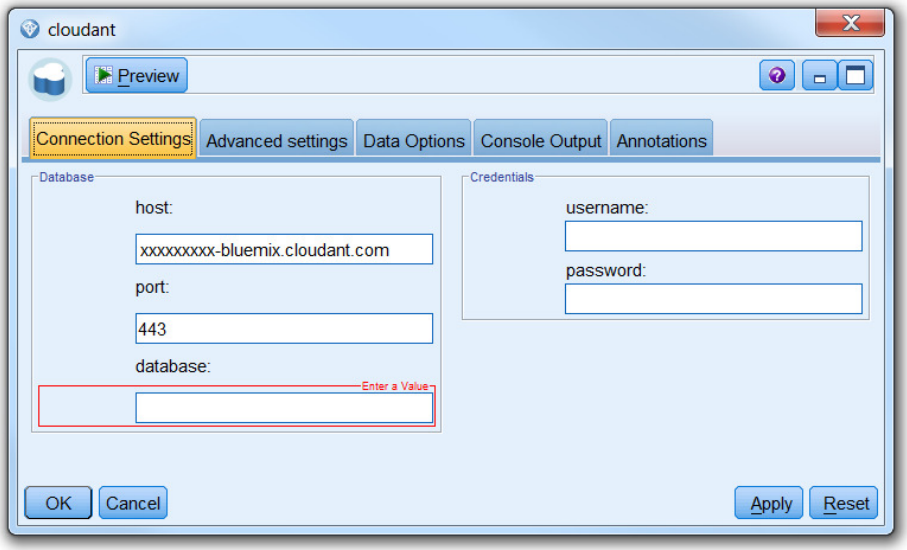
- 'R4CouchDB' package created by Thomas Bock https://cran.r-project.org/web/packages/R4CouchDB/
- 'RCurl' package created by Duncan Temple Lang https://cran.r-project.org/web/packages/RCurl/index.html
- 'plyr' package created by Hadley Wickham https://cran.r-project.org/web/packages/plyr/

User Interface

- Double click on the node to get to the options. There are the following fields:
  - Host: there are two different areas this can be located depending on how you are using Cloudant
    - Service: this is found in the service credentials section on Bluemix shown below
    - Part of application: when accessing Cloudant from the application, this will be the first part of the URL following https://. The host code will look like the example below:

      xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx-bluemix.cloudant.com

  - Port: Specify the server port. Default for Cloudant is 443.
  - Database: Specify the database name. The database must already exist on the Cloudant server.
  - Credentials (optional): This depends on the read permissions on the database being queried. If you need to authenticate, fill in your username and password.



Example

This example will demonstrate connecting to a Cloudant database and viewing the data in a table. The Example directory in this repository contains a Modeler Stream and a dataset that can be used took, test this extension.

Bluemix and Cloudant Set-up:

Prior to starting this tutorial, add the Cloudant database service to your Bluemix account and create a database using the data available in this repository in the Example directory.

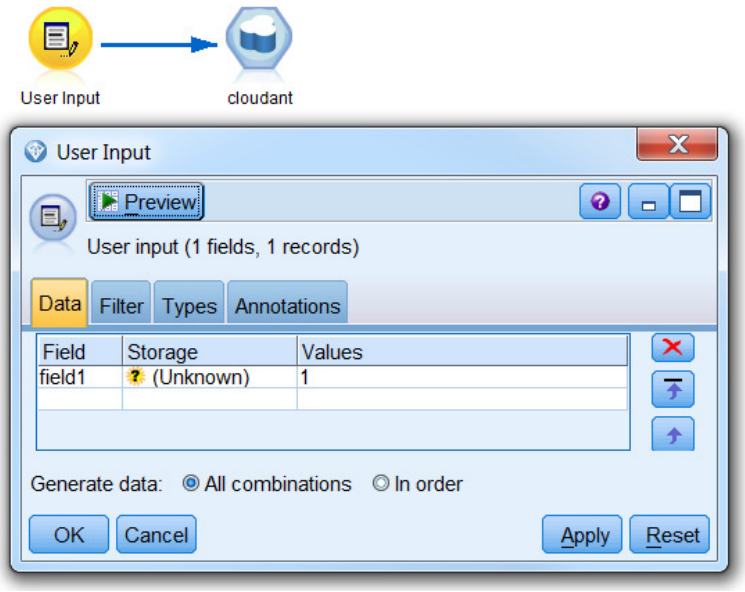
IBM SPSS Modeler Directions:

User Input

1. In the Sources palette, add a User Input node to the stream

   In SPSS Modeler v16.0 and v17.0, it is not yet possible to create proper 'source' nodes with R code.

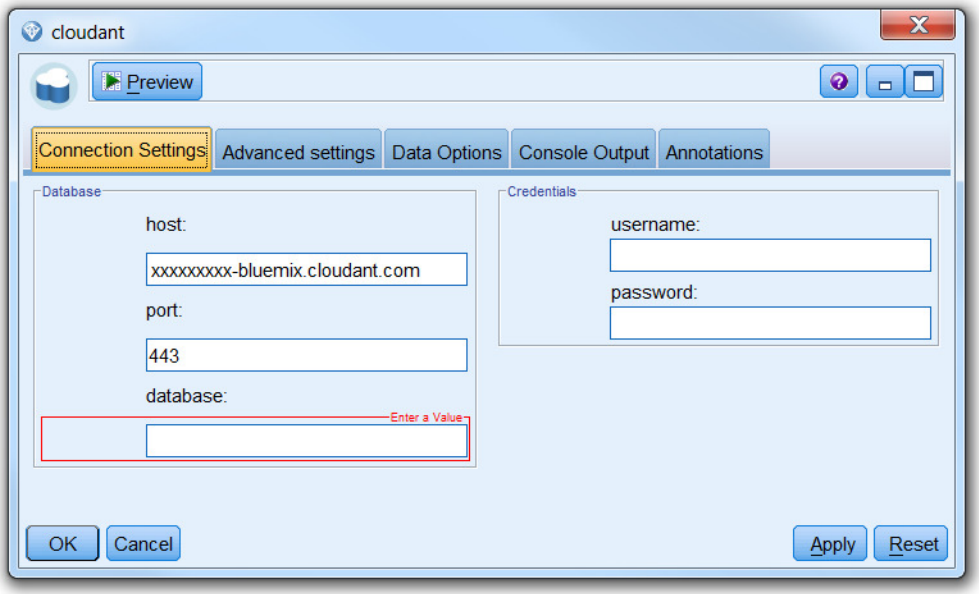
   We will need to give an input to the Cloudant node, but it is only a required trick to start the stream. Create for example a user input with a dummy field, followed by the Cloudant node. The output of the Cloudant node will be the result from the database, regardless of its input.



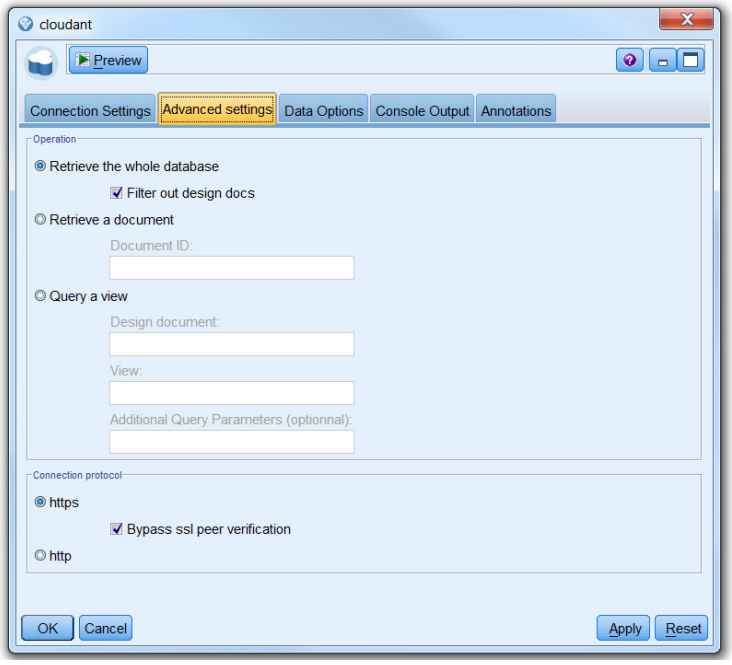
2. Add the Cloudant node to the stream from the Record Ops palette and open

3. Fill in the fields on the Connection Settings tab of the dialog with the following information:

   - Host: This is found in the service credentials section on Bluemix shown below
   - Port: Specify the server port. Default for Cloudant is 443.
   - Database: Specify the database name. The database must already exist on the Cloudant server.
   - Credentials (optional): This depends on the read permissions on the database being queried. If you need to authenticate, fill in your username and password.

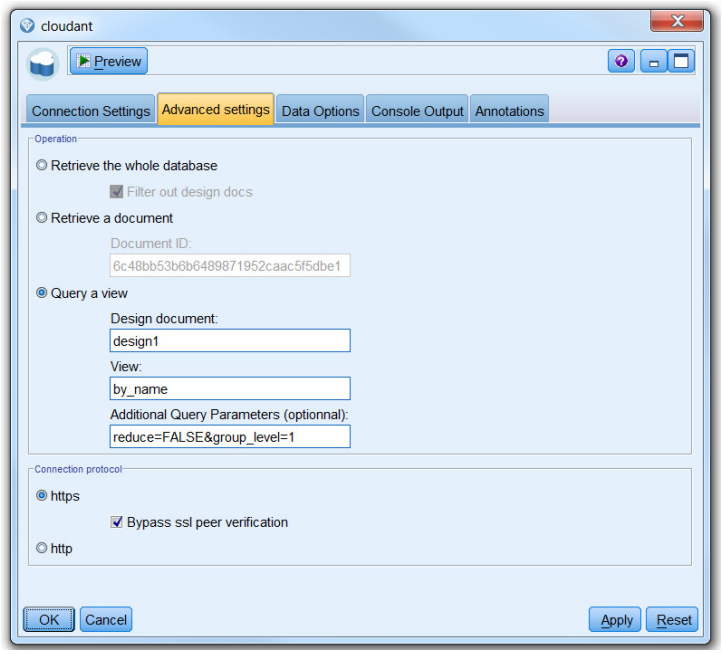


4. By default, the node will fetch all the documents in the database (except design documents). If you want to modify this behavior, you can go to the “Advanced Settings” tab.

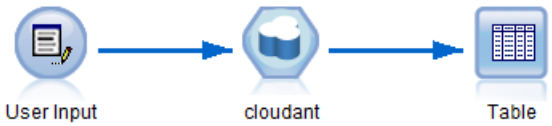


5. Here you can choose:

   - To retrieve the whole database, with the option to filter the design documents
   - To retrieve a precise document by specifying its ID
   - To execute a view that is stored in a design document inside your database. You may add a query parameter with the syntax *param=value*. If you want to add more than one query parameter, you must separate them with *&*.
   - You can also switch the protocol to http instead of https and activate the peer verification during the establishment of the SSL connection. You can connect easily to Blue Mix server with the default settings.
   - See the screenshot on the next page for an example:



6. To visualize the data you gathered, you can add an output node (typically a table):



Important Links

Learn

- Learn more about SPSS software.
- Visit developerWorks Business analytics for more technical analytics resources for developers.
- Create your Bluemix account here: https://console.ng.bluemix.net/solutions/watson
- The Comprehensive R Archive Network is the main site for the R project and each R package. The help pages and manuals that are associated with optimx, nlmrt, and Rcgmin are detailed. Numerous references are provided.
- Read "Do I need to learn R?" (Catherine Dalzell, developerWorks, September 2013) to learn why R is a valuable tool for data analytics that was expressly designed to reflect the way that statisticians think and work.
- "Calling R from SPSS" describes how to use R code inside IBM SPSS Modeler 16.
- Read "Create new nodes for IBM SPSS Modeler 16 using R" to learn how to create new extensions easily.

Discuss

- Visit the IBM SPSS Community to share tips and experiences with other IBM SPSS developers.
- Follow developerWorks on Twitter to be among the first to hear about new resources.