SAS® Contextual Analysis In-Database Scoring 15.1 for Hadoop: Administrator's Guide

SAS® Documentation
October 26, 2018

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS® Contextual Analysis In-Database Scoring 15.1 for Hadoop: Administrator's Guide*. Cary, NC: SAS Institute Inc.

SAS® Contextual Analysis In-Database Scoring 15.1 for Hadoop: Administrator's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

15.1-P1:ctxtindbag

Contents

SAS Contextual Analysis In-Database Scoring 15.1 for Hadoop: Administrator's Guide

Audience

This book is for users of SAS Contextual Analysis who want to use their scoring models in Hadoop. The book provides instructions for deploying the SAS In-Database Scoring for Hadoop and also deploying your models in the Hadoop environment. It is assumed that you know how to use SAS Contextual Analysis.

Requirements

You must have licenses for the following products:

- SAS Contextual Analysis
- SAS In-Database Code Accelerator for Hadoop

Chapter 1

Introduction

Overview of SAS Contextual Analysis In-Database Scoring for Hadoop

Enterprises store large amounts of unstructured text documents, along with other data, in Hadoop. They are seeking tools that enable the analysis of all data within the database in order to avoid excessive consumption of network resources. SAS Contextual Analysis In-Database Scoring for Hadoop enables IT professionals to deploy text analytics models (also called binary models) inside their Hadoop infrastructure. This practice avoids any data movement during scoring and takes advantage of existing data warehouse investments.

About SAS Contextual Analysis

SAS Contextual Analysis is a web-based application that uses natural language processing and machine learning to derive insights from textual data. It derives insights through topic identification, categorization, entity and fact extraction, and sentiment analysis—all from a single interface. Using this application, you can build models (based on training documents) and create taxonomies and rule sets to analyze documents. You can then customize your models for your business domain in order to realize the value of your text-based data.

At the end of the modeling process, SAS Contextual Analysis generates DS2 code and binary models for scoring text data. The score code can be retrieved from the View drop-down menu, which is first seen on the Properties page of any SAS Contextual Analysis project.

SAS Contextual Analysis generates three types of DS2 score code and models corresponding to categorization, concept extraction, and sentiment analysis. The DS2 code can be run within a SAS environment such as SAS Studio or modified to run within your Hadoop cluster using SAS In-Database Code Accelerator for Hadoop. The binary models represent the rule sets for categorization (file extension: .mco), concepts (file extension: .li) and sentiment (file extension: .sam) taxonomies, which are highly optimized to apply all rules in parallel.

For information about using SAS In-Database Code Accelerator for Hadoop, see the SAS In-Database Products: User's Guide.

What is SAS Embedded Process?

SAS Embedded Process is the core of SAS in-database products. It allows the parallel execution of SAS processes inside Hadoop or inside other databases. SAS Embedded Process technology is a portable, lightweight, execution container for SAS DS2 code. SAS Embedded Process is orchestrated by Hadoop MapReduce framework or Spark. Load balancing and resources allocation are managed by YARN.

SAS Embedded Process offers a flexible, efficient way to leverage increasing amounts of data by injecting the processing power of SAS where ever the data lives. SAS Embedded Process can tap into the massively parallel processing (MPP) architecture of Hadoop for scalable performance. Using SAS in-database technologies for Hadoop, you can run scoring models generated by SAS Contextual Analysis.

Why Score Text Analytics Models In-Database?

The SAS In-Database Code Accelerator for Hadoop enables you to publish a DS2 multi-threaded program and its associated files to the database. It enables you to execute that threaded program in parallel within SAS Embedded Process. Benefits of in-database processing include reduced data movement and faster run time. It also gives you the ability to leverage existing data warehousing investments without having to copy data to a secondary location for processing. Examples of threaded programs include large transpositions, computationally complex programs, scoring models, and BY-group processing.

Which Hadoop Platforms Are Supported?

The Cloudera and Hortonworks platforms are supported for this product.

Getting Started with the SAS Contextual Analysis In-Database Scoring for Hadoop

The steps for getting started with this product are as follows:

1. Deploy the SAS Embedded Process for Hadoop. For detailed steps, see the *SAS Embedded Process: Deployment Guide*.
2. Deploy the SAS Contextual Analysis scoring models. For details, see Chapter 2, "Deploying SAS Text Analytics Scoring Models ," on page 5.

For information about using the models, see the SAS Contextual Analysis In-Database Scoring for Hadoop: User's Guide.

Chapter 2

Deploying SAS Text Analytics Scoring Models

Overview of Model Deployment

To deploy a SAS Contextual Analysis text analytics (binary) model, follow these steps:

1. Obtain a text analytics model that includes score code for concept extraction (.li file), categorization (.mco file), or sentiment analysis (.sam).
2. Copy the score code model to the Hadoop master node (NameNode).
3. Use ta_push.sh to deploy the SAS Contextual Analysis score code model in the cluster.

Obtaining a Text Analytics Model

You can obtain the text analytics models from SAS Contextual Analysis. To generate the score code, see the section “Viewing and Downloading Code” in the SAS Contextual Analysis: User’s Guide.

- To locate the binary files for concepts and categories, see the section “Locating the Rules Files (LI and MCO)” in the SAS Contextual Analysis: Administrator’s Guide.

- To locate language-specific SAM (binary) files that are used in SAS Contextual Analysis, see the file **\tktg\sasmisc*nn*-base.sam** (Windows) or **misc/tktg/*nn*-base.sam** (UNIX) under your SAS installation directory. Note that the first two characters (*nn*) of the .sam filename denote the licensed language. For example, in **en-base.sam** , the **en** denotes English.
- To locate sentiment (SAM) files that you created using SAS Sentiment Analysis Studio, see the *SAS Sentiment Analysis Studio: User's Guide*.

Copying the Text Analytics Model to the Hadoop NameNode

After you have obtained a text analytics model, you must copy it to the Hadoop NameNode. It is recommended that you copy the model to a temporary staging area, such as **/tmp/tastage**. You can copy the model to the Hadoop NameNode by using a file transfer command such as FTP or SCP. You can also copy it by mounting the file system where the model is located on the Hadoop NameNode.

The following example shows how you might copy a model that exists on a Linux system to the Hadoop NameNode. The example uses the secure copy with the -r flag to recursively copy the specified directory. For the example, assume the following:

- The host name of the client desktop system where the model is installed is **desktop123**
- The location where the model is located in client desktop system.is **/opt/sas/ta/share**
- The host name of the Hadoop NameNode is **hmaster456**
- The target location on the NameNode is **/tmp/tastage**

To copy the model from the client desktop to the NameNode, issue this command:

```
scp -r /opt/sas/ta/share hmaster456:/tmp/tastage
```

About the ta_push.sh Executable File

SAS Contextual Analysis In-Database Scoring for Hadoop provides the ta_push.sh executable file to enable you to deploy the SAS text analytics models on Hadoop cluster nodes. The ta_push.sh file copies the specified model to a user-specified location on each of the Hadoop nodes. The ta_push.sh file automatically discovers all nodes in the cluster and deploys the model to the specified target model path on each of the cluster nodes.

Executing ta_push.sh

The ta_push.sh file must be run as the root user. The root user becomes the HDFS user in order to detect the nodes in the cluster.

Execute ta_push.sh as follows:

`cd /opt/sas/ep/home/bin`

`./ta_push.sh -s` *source_path* **`-t`** *target_path*

where *source_path* specifies the path on the NameNode where you copied your model, and *target_path* specifies the path on each of the Hadoop nodes where you deployed your model.

Here is an example of the command:

`./ta_push.sh -s /tmp/tastage/en-ne.li -t /opt/sas/ta/model/en-ne.li`

Troubleshooting the Text Analytics Model Deployment

The text analytics model deployment can fail for the following reasons:

Problem: You did not obtain a Kerberos ticket before attempting to run ta_push.sh in a Kerberos environment.
Solution: Obtain the ticket and rerun.

Problem: You executed ta_push.sh from a directory other than the **`/opt/sas/ep/home/bin`** directory.
Solution: Run the ta_push.sh script from **`/opt/sas/ep/home/bin`**.

Problem: There is insufficient space in the **`/tmp`** directory for ta_push.sh to run.
Solution: Clear space and try again.

ta_push.sh: Reference

Overview

The ta_push.sh file is created in the **`/opt/sas/ep/home/bin`** directory when you install the SAS Embedded Process for Hadoop. You must execute ta_push.sh from this directory.

By default, ta_push.sh automatically discovers all nodes in the cluster and deploys the text analytics model to the specified target model path on each cluster nodes. Flags are provided to enable you deploy the model to specific nodes or a group of nodes. If you are expanding your Hadoop cluster by adding new nodes after the initial deployment, consider using one of these flags to deploy the model to these nodes. This practice avoids redeploying to the entire cluster.

Run ta_push.sh as the root user. The root user becomes the HDFS in order to detect the nodes in the cluster.

Note: Only one model can be deployed to each Hadoop node at a time.

Syntax

./ta_push.sh -s source_path -t target_path

Required Arguments

-s source_path
specifies the pathname of the source model on the NameNode.

-t target_path
specifies the pathname of the target model on each of the Hadoop nodes.

Options

-h hostname
specifies the host name of the computers or computers on which to perform the deployment.

-f hostfile
specifies the name of a file that contains a list of host names on which to perform the deployment.

-?
displays usage information

-l logfile
directs status information to the specified log file instead of directing to standard output.

-r
Removes the model from Hadoop nodes.

-v
specifies verbose output.

Examples

To deploy models to one or more nodes that are specified on a command line, execute the following command:

```
./ta_push.sh -h hostname1 [-h hostname2] -s source_path -t
target_path
```

To deploy models using a file that contains a list of node names, execute the following command:

```
./ta_push.sh -f hostfile -s source_path -t target_path
```

To remove a model from the Hadoop nodes, execute the following command:

```
./ta_push.sh -r target_path
```

Recommended Reading

- SAS Contextual Analysis: Administrator's Guide
- SAS Contextual Analysis: User's Guide
- SAS Contextual Analysis In-Database Scoring for Hadoop: User's Guide
- *SAS Embedded Process: Deployment Guide*
- SAS In-Database Products: User's Guide
- SAS Global Forum Paper, "Exploring SAS Embedded Process Technologies on Hadoop"

For a complete list of SAS publications, go to sas.com/store/books. If you have questions about which titles you need, please contact a SAS Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-0025
Fax: 1-919-677-4444
Email: sasbook@sas.com
Web address: sas.com/store/books