Hewlett Packard
Enterprise

HPE Security ArcSight Investigate

Software Version: 1.0 (BETA)

ArcSight Investigate HDFS Feature Tech Note

April 17, 2017

Legal Notices

Warranty

The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

The information contained herein is subject to change without notice.

The network information used in the examples in this document (including IP addresses and hostnames) is for illustration purposes only.

HPE Security ArcSight products are highly flexible and function as you configure them. The accessibility, integrity, and confidentiality of your data is your responsibility. Implement a comprehensive security strategy and follow good security practices.

This document is confidential.

Restricted Rights Legend

Confidential computer software. Valid license from Hewlett Packard Enterprise required for possession, use or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Copyright Notice

Follow this link to see a complete statement of copyrights and acknowledgements:
https://www.protect724.hpe.com/docs/DOC-13026

Support

Contact Information

Phone	A list of phone numbers is available on the HPE Security ArcSight Technical Support Page: https://softwaresupport.hpe.com/documents/10180/14684/esp-support-contact-list
Support Web Site	https://softwaresupport.hpe.com
Protect 724 Community	https://www.protect724.hpe.com

Contents

Introduction 4

Requirements 5

- Support matrix 5
- Database configuration and requirements 5
- HDFS configuration and requirements 6
- Network requirements 6

What the HDFS feature cannot do 8

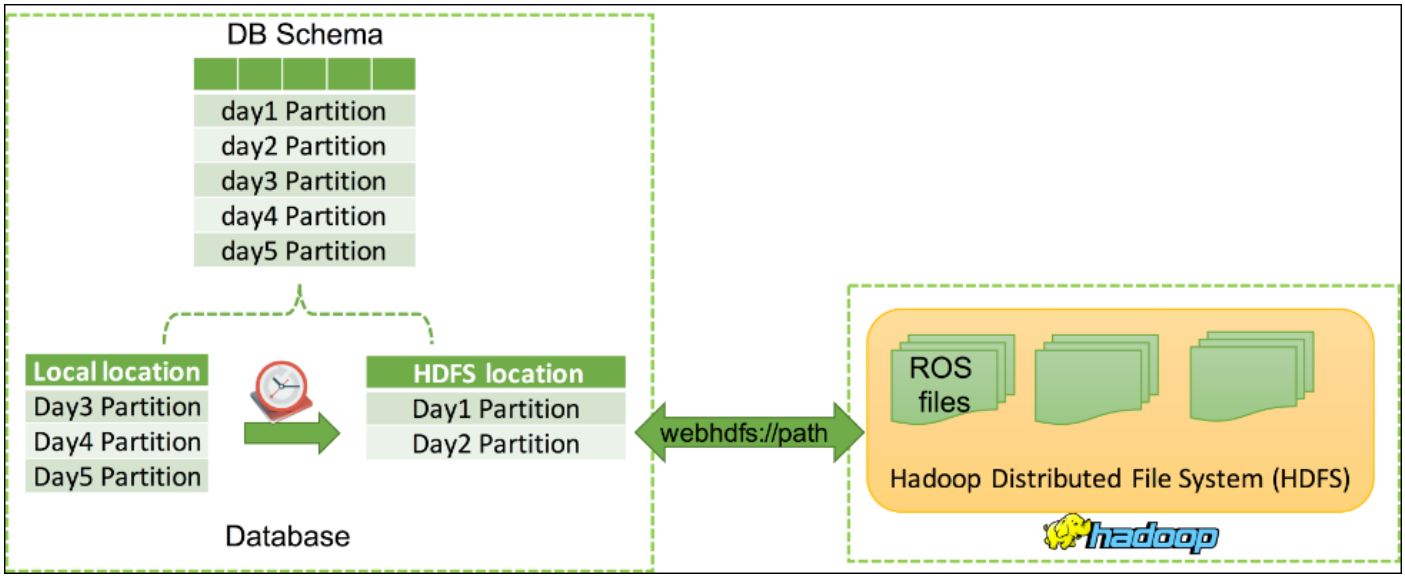
HDFS feature configuration and usage 9

Stopping the HDFS feature 12

Send Documentation Feedback 13

Introduction

HDFS feature enables user to automatically archive old data, e.g., older than 90 days, to HDFS and query them through Investigate as needed. This feature will enable Investigate admin to periodically archive data based on a retention period chosen beforehand. This tool is CLI based tool. Database stores data in its native format, ROS, in local file system storage locations. This is a default feature for storage. Admin can choose to enable HDFS location optionally. You would typically use HDFS storage locations for old-data as cold storage data. Doing so frees space on your DB cluster for higher-priority newer data. This is an optional feature for users who have a mature HDFS cluster.



Requirements

Before enabling this feature the Admin needs to make sure all these requirements are met.

Support matrix

Source: [vertica 8.0 _support_matrix]

Hadoop distribution version	Hadoop version	Supported OS
Cloudera (CDH) 5.6 Supported JDK to set up CDH 5.6 are (Use jdk1.7) . Don't use Jdk 1.8 version.	till 2.6 (HDFS version)	RHEL : 7.1, 6.7, 6.6, 6.5,5.10 , 5.7 CentOS:7.1,6.7,6.6,6.5,6.4,5.10,5.7
Cloudera (CDH) 5.7	till 2.6 (HDFS version)	Included all above and 7.2 (RHEL, CentOS)
HortonWorks Data Platform (HDP) 2.3	till 2.7.1 (HDFS version)	64-bit CentOS 6.x 64-bit CentOS 7.x 64-bit Red Hat Enterprise Linux (RHEL) 6.x 64-bit Red Hat Enterprise Linux (RHEL) 7.x
HortonWorks Data Platform (HDP) 2.4	till 2.7.1 (HDFS version)	64-bit CentOS 6 64-bit CentOS 7 64-bit Red Hat Enterprise Linux (RHEL) 6 64-bit Red Hat Enterprise Linux (RHEL) 7

Database configuration and requirements

To store Vertica's data on HDFS, verify the following:

- Your Hadoop cluster has WebHDFS enabled.
- If you have proxy, you should put all your HDFS nods as no_proxy nodes on each DB node. This can be achieved through using .bashrc file.
- All of the nodes in your Vertica cluster can connect to all of the nodes in your Hadoop cluster. Any firewall between the two clusters must allow connections on the ports used by HDFS. See Testing Your Hadoop WebHDFS Configuration for a procedure to test the connectivity between your Vertica and Hadoop clusters.

- You have a Hadoop user whose username matches the name of the Vertica database administrator (usually named dbadmin). This Hadoop user must have read and write access to the HDFS directory where you want Vertica to store its data.
- Your HDFS has enough storage available for Vertica data. See HDFS Space Requirements below for details. See: https://my.vertica.com/docs/8.0.x/HTML/index.htm#Authoring/AdministratorsGuide/StorageLocations/MovingDataStorageLocations.htm?Highlight=moving%20storage%20locations
- Data on HDFS should not be modified or changed, otherwise it will cause not reversible data lose for HDFS data.

HDFS configuration and requirements

- Hadoop should be a stable mature cluster with tens of nodes as minimum, e.g. 20 nodes
- You can adjust the number of duplicates stored by HDFS by setting the HadoopFSReplication configuration parameter. This value should be minimum 2.
- Increase Linux os process and file limits: see: https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_installing_manually_book/content/ref-729d1fb0-6d1b-459f-a18a-b5eba4540ab5.1.html
- Recommended heap size for Hadoop: https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.4.2/bk_installing_manually_book/content/ref-80953924-1cbf-4655-9953-1e744290a6c3.1.html
- Hadoop node hostnames need to be resolvable using DNS
- The user need to provide enough space based on its ingestion rate into Vertica, record size, and retention period, for example if your ingestion is 50 K EPS with record size=1.5 KB, and planning to archive for 1 year you will need minimum of 500 TB space available on HDFS, this is example just to clarification.
- Make sure your HDFS firewall is configured probably and allows WebHDFS ports.
- For best performance, set the following parameters with the specified minimum values:

Parameter	Minimum value
HDFS block size	512 MB
Namenode Java Heap	1 GB
Datanode Java Heap	2 GB

Network requirements

- The network is a key performance component of any well-configured cluster. When Vertica stores data to HDFS it writes and reads data across the network.

- Database/Hadoop Shared Network: Each Vertica node must be able to connect to each Hadoop data node and the Name Node. Hadoop best practices generally require a dedicated network for the Hadoop cluster. This is not a technical requirement, but a dedicated network improves Hadoop performance. Vertica and Hadoop should share the dedicated Hadoop network.
- Customer should have 10GB NIC between Vertica nodes and HDFS.
- HDFS network connection should Highly available if HDFS is not accessible through network, this will cause DB failure and potential HDFS data loss.

What the HDFS feature cannot do

- HPE uses the storage locations to store ROS containers in a proprietary format, MapReduce and other Hadoop components cannot access your Vertica data stored in HDFS. Never allow another program that has access to HDFS to write to the ROS files. Any outside modification of these files can lead to data corruption and loss.
- HA enabled clusters is not supported. This is a feature should be available in the next release.
- Customers cannot stop or remove storage location temporarily if they want remove it permanently you need to contact support.
- Recommend that user backs up Vertica database and HDFS together. Vertica provides comprehensive documentation. See Backup Vertica DB bellow
- Vertica does not support secure webhdfs, currently. Customer could create ssh tunnel to address this issue. There is a story to track the possibility of this option.

HDFS feature configuration and usage

It is assumed that you have already obtained and installed your nodes, HDFS and Vertica and want to run the archiving tool.

1. On HDFS, create a folder, e.g., /vertica, and give read and write permission of your Vertica admin on it. See link (you can get more info if you want)

```
-bash-4.2$ hdfs dfs -mkdir /investigate
-bash-4.2$ hdfs dfs -chown dbadmin:dbadmin /investigate
-bash-4.2$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x - dbadmin dbadmin 0 2017-03-31 23:25 /investigate
```

2. Make sure that your HDFS ports are open on your HADOOP cluster:

```
[root@master ~]# firewall-cmd --zone=public --add-port=50070/tcp --
permanent
success
[root@master ~]# firewall-cmd --reload
success
```

On Vertica do the following:

3. Test whether your webhdfs is working, see link.
4. Install the tool under "dbadmin" account home, by doing the following:

```
[dbadmin@localhost ~]$ ./investigate-hdfs-0.5.0-installer
HDFS Support Command Line Installer
vsqloh-ros/
vsqloh-ros/bin/
vsqloh-ros/config/
vsqloh-ros/README.md
vsqloh-ros/release-info
vsqloh-ros/scripts/
vsqloh-ros/util/
vsqloh-ros/util/setup_schema
vsqloh-ros/scripts/commons
vsqloh-ros/scripts/enforce_storage_policy
vsqloh-ros/config/config.cfg
vsqloh-ros/bin/scheduler
Installation complete.
```

5. Configure the tool:

```
[dbadmin@localhost ~ ]$ cd ~/investigate-hdfs/vsqloh-ros
[dbadmin@localhost vsqloh-ros ]$ vim ./config/config.cfg
[dbadmin@localhost vsqloh-ros]$ vim ./config/config.cfg
#!/bin/bash
##
## Configuration file
##
## Author: Loai Zomlot
##
###################
## HDFS variables
###################
## HDFS name_node ip or name
hdfs_name_node="192.168.93.13"
## data folder name on HDFS
vertica_target_folder="investigate"
########################
## Database parameters
########################
db_ip='localhost'
db_database_name='investigate'
db_schema='investigation'
db_table_name='events'
db_user_name='dbadmin'
db_user_pass='dbadmin'
db_port='5433'
db_storage_location_name="titan"
####################
## Data retintion period
####################
partition_by='day'
db_retintion_period=90 # If partition by is day db_retintion_period will
be 5 days and scheduler should work every day at mid-night
########################
## Archiving transaction mode
########################
```

```
## true: Run the data archiving as a transaction, i.e., wait till the job
is done
## false: Run the data archiving as a background job
transaction_mode="true"
```

6. After configuring the tool, do the following:

```
[dbadmin@localhost vsqloh-ros ]$ ./util/setup_schema
```

7. Run the schduler script at the configured time:

```
[dbadmin@localhost vsqloh-ros]$ ./bin/scheduler -c
Scheduler has been created.
```

For more options,

```
[dbadmin@localhost vsqloh-ros]$ ./bin/scheduler [options]
options:
--help, show brief help
-l, --list, list tasks
-r, --remove, remove schedule
-c, --create create scheduler based on db schema partition size in the
config file"
```

8. On HDFS, you can check your Vertica folder on HDFDS, after some time.

```
hadoop fs -ls /investigate/
hadoop fs -du -s /investigate
```

Stopping the HDFS feature

If the admin chooses stop using this feature he or she should call the below command. But data that was moved to cold storage cannot move back to Vertica local Linux storage, however the user can stop the archiving process for any new data using the above command.

```
[dbadmin@192 vsqloh-ros]$ ./bin/scheduler -r
```

Send Documentation Feedback

If you have comments about this document, you can contact the documentation team by email. If an email client is configured on this system, click the link above and an email window opens with the following information in the subject line:

Feedback on ArcSight Investigate HDFS Feature Tech Note (Investigate 1.0)

Just add your feedback to the email and click send.

If no email client is available, copy the information above to a new message in a web mail client, and send your feedback to arc-doc@hpe.com.

We appreciate your feedback!