Hive Overview

Contents

[SQL 1](#_Toc503200566)

[Tools 2](#_Toc503200567)

[Aginity 2](#_Toc503200568)

[Installation steps 2](#_Toc503200569)

[Hue 3](#_Toc503200570)

[Data import and export 3](#_Toc503200571)

[Export 3](#_Toc503200572)

[Small Datasets 3](#_Toc503200573)

[Large Datasets 4](#_Toc503200574)

[Imports 5](#_Toc503200575)

[Small Datasets 5](#_Toc503200576)

[Large Datasets 5](#_Toc503200577)

# SQL

1. ANSI SQL compatible.
2. Columnar database. As such performance will faster by explicitly choosing the required columns instead of using ‘\*’
3. When using GUI or command line tools, try bringing over summarized data. Data transfer takes time and such if bringing millions of rows in may take hrs.
4. Use “limit" clause to limit number of rows.

Ex.

select

dp\_lname, dp\_zip, dp\_hh\_key

from ddg\_ddg

where dp\_state='NY' and dp\_hh\_age\_code=6

limit 1000;

# Tools

We recommend using following tools

1. Aginity
2. Hue

Hue gives a quick connectivity without any setup on user side. Aginity is more analyst friendly.

## Aginity

### Installation steps

EDP cluster is uses Kerberos authentication to provide enterprise level security. To use any desktop based tool, Kerberos needs to be installed first on the PC. Please follow the below steps carefully to install all the required components.

1. Use the attached installation guide to install Kerneros.



1. Download Aginity from Aginity website. <https://www.aginity.com/workbench/hadoop/>

PS – Currently it is getting blocked when reached from Altice network. Reach out to DEV team and we can provide an existing executable.

1. Install the software by running the executable.

* It is a freeware. Please provide your email and other information when asked to register.

1. Once Aginity is installed please use the below instructions to setup the connection.



#### Tips:

1. Set the current database using the “Database” drop down. Doing this will then not require you to qualify tablename with the database name. Will still need it when referring table in another database.
2. Use view->”user query history” to look at previously run queries.
3. Aborting a running query takes time and not always possible.
4. Use tool->option->result set option->limit number of rows…. to limit number of rows returned per query. Prevents long running queries due to data transfer.

### Hue

Hue is a simple browser based tool. It does not require Kerberos. Just login information.

Prod Link: <http://cvlphdpen1.cablevision.com:8000>

Dev Link: [http://cvldhdpan2.cscdev.com:8000](http://cvldhdpan2.cscdev.com:8000/)

# Data import and export

## Export

### Small Datasets

Small datasets can easily be exported over using Aginity.

1. Run the query
2. Right click on the data grid and choose the option needed.
3. Good for upto few million rows
4. Make sure you remove the row limit from “preferences”

### Large Datasets

Use file transfer mechanism. Tables on Hive are stored as files on HDFS and these files can be accessed using HDFS commands. Think of these as external tables.

Steps:

1. Using Aginity, create a table that uses CSV serde and stores data in plain text.

CREATE TABLE work\_ntelligis.naeem\_export\_ex1(

Col1 string,

Col2 string )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"escapeChar" = "\\"

)

STORED AS TEXTFILE location '/apps/hive/warehouse/work\_ntelligis/naeem\_export\_ex1;

1. Using Aginity, set the output compression as GZIP.

SET mapred.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec;

If need uncompressed data, then run the below

set hive.exec.compress.output=false;

1. Using Aginity, insert required data in the table

Insert into work\_ntelligis.naeem\_export\_ex1

Select col1,col2 from smith\_ntelligis.source\_tbl;

1. SSH to the edge node, do kinit and then run the below command to copy data from HDFS to S3 bucket

hadoop distcp -Dhadoop.security.credential.provider.path=jceks://hdfs/user/nali2/aws1.jceks dfs://althdpdev/apps/hive/warehouse/ work\_ntelligis/Naeem\_export\_ex1 s3a://BUCKET/FOLDER

In above example we are using jceks files to pass S3 credentials. Refer below link to create a jceks file.

<https://community.hortonworks.com/articles/59161/using-hadoop-credential-api-to-store-aws-secrets.html>

1. Instead of S3 bucket, the files can be transferred to the edge node and then SFTP’ied over. Use below command to transfer to edge node from HDFS. Run below from edge node

hdfs dfs -get /apps/hive/warehouse/work\_ntelligis/naeem\_export\_ex1 /user/nali2/exportdir

## Imports

### Small Datasets

Using Hue to import smaller datasets. Hue will basically copy the data to edge node and then provide the command to copy it into Hive table. Same can be done using SFTP and Aginity.

1. Log into Hue or Aginity and create the required table if not exists

CREATE TABLE work\_ntelligis.naeem\_import\_ex1(

Col1 string,

Col2 string )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"escapeChar" = "\\"

)

STORED AS TEXTFILE location '/apps/hive/warehouse/work\_ntelligis/naeem\_export\_ex1;

1. In hue browse to the table
2. Choose option “import data”
3. Follow the instructions.

### Large Datasets

1. Can still use SFTP or Hue
2. S3 is another option.
3. Create the table if not already exists
4. Use below command to copy data from S3 buket to HDFS directory for the table.

hadoop distcp -Dhadoop.security.credential.provider.path=jceks://hdfs/user/nali2/aws1.jceks s3a://BUCKET/FOLDER/ dfs://althdpdev/apps/hive/warehouse/ work\_ntelligis/Naeem\_export\_ex1