

Summary Herramientas Consulta

16 junio 2021

Objetivo

Conocer herramientas que facilitan la interacción con el cluster como Hue y Zeppelin que nos permiten realizar consultas de Pig, Hive y Spark

Datos

Los datos que utilizaremos se obtuvieron de la siguiente liga:
<https://www.kaggle.com/open-flights/airline-database>

Cada entrada de los datos tiene la siguiente información:

- Airline ID: identificador único de la aerolínea.
- Name: Nombre de la aerolínea.
- Alias: Alias de la aerolínea.
- IATA 2: Código IATA
- ICAO 3: Código ICAO
- Callsign Airline callsign.
- Country: País o territorio de origen de la aerolínea.
- Active: "Y" o "N" aunque no es tan confiable.
-

Notemos que el valor \N es utilizado para "NULL" para indicar que el valor no esta disponible.

S3

Los datos los subimos a un bucket en S3. En mi caso yo llame al bucket "datos-aerolineas", pero le pueden nombrar como ustedes quieran.

EMR

Vamos a utilizar el servicio de Amazon EMR, por lo que valen la pena unas recomendaciones:

- Para este ejemplo pueden utilizar instancias pequeñas de preferencia o del “free tier” pues la base es pequeña.
- Recuerda TERMINAR las instancias (clúster) siempre que termines de usarlas, de lo contrario sigue generando un costo.

HIVE usando HUE

1. Levantamos el clúster de EMR, asegurando que tengamos HIVE:

The screenshot shows the Amazon EMR console interface. At the top, there are buttons for 'Clonar', 'Finalizar', and 'Exportación de la CLI de AWS'. Below these, the cluster name 'Clúster: summary' is followed by a green status 'Esperando' and a message 'Cluster ready after last step completed.' A navigation bar contains tabs: 'Resumen', 'Historial de aplicaciones', 'Monitorización', 'Hardware', 'Configuraciones', 'Eventos', 'Pasos', and 'Acciones de arranque'. The 'Resumen' tab is active, displaying cluster information: ID 'j-TIZGM9BBUIXW', creation date '2021-06-16 13:45 (UTC-5)', and time elapsed '1 hora, 47 minutos'. It also shows 'Terminar automáticamente: Cluster waits', 'Protección contra la terminación: Desactivado', and 'Etiquetas: --'. A 'DNS público principal' is listed as 'ec2-3-139-93-158.us-east-2.compute.amazonaws.com'. The 'Detalles de las configuraciones' section shows 'Etiqueta de la versión: emr-5.33.0', 'Distribución Hadoop: Amazon 2.10.1', 'Aplicaciones: Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2', 'URI de registro: s3://aws-logs-475341802139-us-east-2/elasticmapreduce/', 'Vista coherente de EMRFS: Deshabilitados', and 'ID de AMI personalizada: --'. The 'Redes y hardware' section shows 'Zona de disponibilidad: us-east-2a', 'ID de subred: subnet-67ada80f', 'Maestro: En ejecución 1 m4.large', 'Principal: En ejecución 2 m4.large', 'Tarea: --', and 'Cluster scaling: Not enabled'.

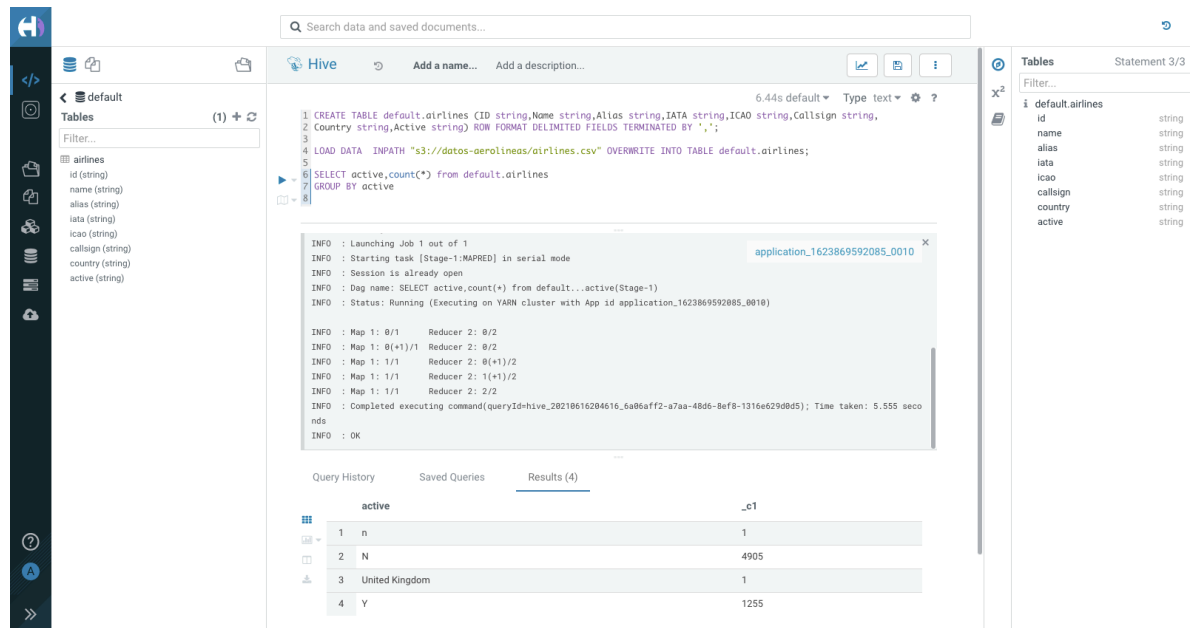
2. Verificamos que podamos realizar conexiones ssh.
3. Ejecutamos en la terminal:

```
ssh -i ~/mykeypair.pem -N -D 8157 hadoop@ec2-###-##-##-###.compute.amazonaws.com
```

Fuente: <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-ssh-tunnel.html>

- Una vez que el túnel esta activo, configuramos un SOCKS proxy. Para realizar esto consulta esta documentación:
<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-connect-master-node-proxy.html>
- Para abrir Hue usamos: `http://master public DNS:8888`
- Si es la primera vez que entras tendrás que crear una cuenta con un usuario y contraseña.

Hue tiene la siguiente estructura:



The screenshot displays the Apache Hue web interface. On the left, a sidebar shows the file system structure with a 'default' directory containing a table named 'airlines'. The main panel shows a Hive query being executed. The query consists of three lines: a CREATE TABLE statement, a LOAD DATA statement, and a SELECT statement. The results of the query are displayed in a table with two columns: 'active' and '_c1'. The results are as follows:

active	_c1
1	n
2	N
3	United Kingdom
4	Y

Para crear la tabla y poblarla utilizamos los comandos de la línea 1 y 2. Una vez creada la tabla podemos ejecutar comandos muy parecidos a SQL.

Si quieres conocer mucho más sobre Hive te recomiendo leer la documentación:

<https://cwiki.apache.org/confluence/display/Hive/Home#Home-UserDocumentation>