

viu
.es

2024 - 2025



ACTIVIDAD

Máster en Big Data y Data Science

01MBID – Sistemas de almacenamiento y gestión Big Data

Nombre : Alex Anthony Prieto Romani

Fecha: 08-10-2024

Tabla de contenido

Introducción	3
Prerrequisitos	3
Paso a Paso de la Migración de Datos.....	3
Transferir el Archivo calidadAire(2).csv y Script CQL al Contenedor de Docker	3
Acceder al Contenedor de Cassandra	4
Crear la Tabla en Cassandra Usando el Script CQL.....	4
Instalamos dsbulk Manualmente en el Contenedor en Ejecución.....	5
Actualizar la Lista de Paquetes e instalar wget y unzip.....	5
Descargar y descomprimir dsbulk desde el Sitio Oficial de DataStax	5
Mover dsbulk a un Directorio Permanente.....	6
Crear un Enlace Simbólico para Facilitar el Acceso a dsbulk.....	6
Usar DataStax Bulk Loader para Migrar los Datos	7
Verificar que la Migración Se Realizó Correctamente	7

Introducción

En esta sección, se detalla el proceso de migración de datos utilizando DataStax Bulk Loader para cargar información desde un archivo CSV a una tabla en Cassandra. A continuación, se describen los pasos realizados para lograr una migración exitosa.

Prerrequisitos

- Docker y Docker Compose instalados
- Contenedores de Cassandra en ejecución
- Archivo calidadAire.csv listo en el sistema de archivos.

Paso a Paso de la Migración de Datos

Transferir el Archivo calidadAire(2).csv y Script CQL al Contenedor de Docker

Para migrar los datos, primero necesitamos que los archivos de calidadAire(2).csv y el script migracion.cql estén disponibles en el contenedor de Cassandra. Por lo que usaremos el siguiente comando para copiar estos archivos desde tu máquina local al contenedor:

Copiar el Archivo CSV

```
docker cp "C:\Users\alexa\OneDrive\Documentos\calidadAire(2).csv"
cassandra_db:/calidadAire.csv
```

Copiar el Script CQL

```
docker cp "C:\Users\alexa\OneDrive\Documentos\migracion.cql" cassandra_db:/migracion.cql
```

```
PS C:\Users\alexa> docker cp --help

Usage:  docker cp [OPTIONS] CONTAINER:SRC_PATH DEST_PATH|-
        docker cp [OPTIONS] SRC_PATH|- CONTAINER:DEST_PATH

Copy files/folders between a container and the local filesystem

Use '-' as the source to read a tar archive from stdin
and extract it to a directory destination in a container.
Use '-' as the destination to stream a tar archive of a
container source to stdout.

Aliases:
  docker container cp, docker cp

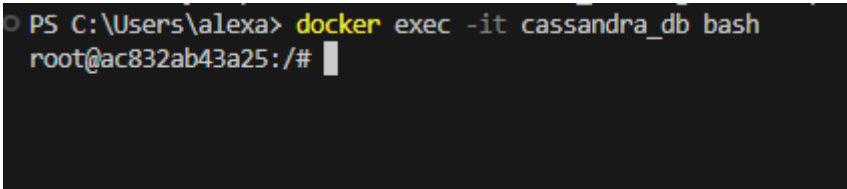
Options:
  -a, --archive           Archive mode (copy all uid/gid information)
  -L, --follow-link       Always follow symbol link in SRC_PATH
  -q, --quiet            Suppress progress output during copy. Progress
                        output is automatically suppressed if no terminal
  -v, --verbose           Show progress during copy. Progress output is
                        automatically suppressed if no terminal
  -x, --exclude EXCLUDE  Exclude files or directories from the copy
  -y, --yes               Assume yes to all prompts
  -z, --compress          Compress files during copy
  -Z, --no-compress       Do not compress files during copy
  -S, --sparse             Copy sparse files
  -s, --no-sparse         Do not copy sparse files
  -p, --preserve PERS     Preserve permissions, ownership, and other
                        metadata
  -P, --no-preserve PERS  Do not preserve permissions, ownership, and
                        other metadata
  -H, --hardlink           Copy files using hardlinks where possible
  -h, --help              Display this help message
  -V, --version           Show version information and exit
  -v, --verbose            Show progress during copy. Progress output is
                        automatically suppressed if no terminal
```

```
>> C:\Users\alexa> docker cp "C:\Users\alexa\OneDrive\Documentos\calidadAire(2).csv" cassandra_db:/calidadAire.csv
Successfully copied 8.46MB to cassandra_db:/calidadAire.csv
PS C:\Users\alexa> docker cp "C:\Users\alexa\OneDrive\Documentos\migracion.cql" cassandra_db:/migracion.cql
Successfully copied 2.05kB to cassandra_db:/migracion.cql
```

Acceder al Contenedor de Cassandra

Entonces una vez ya pasemos los archivos necesarios al Docker, procedemos a entrar al contenedor de cassandra que creamos con el siguiente comando:

```
docker exec -it cassandra_db bash
```

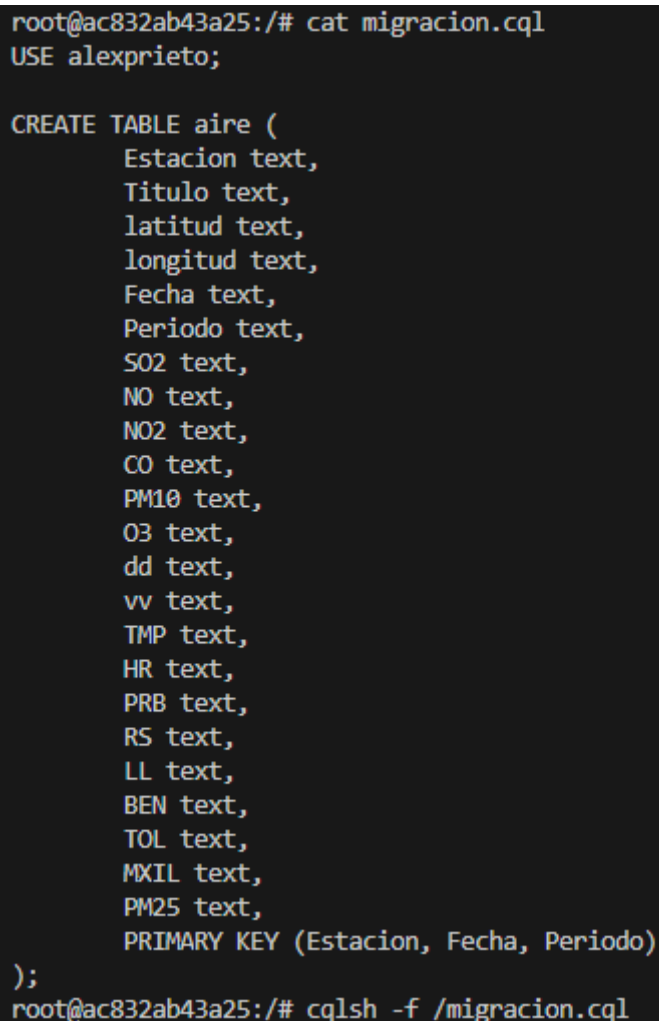


```
PS C:\Users\alexa> docker exec -it cassandra_db bash
root@ac832ab43a25:/#
```

Crear la Tabla en Cassandra Usando el Script CQL

Ahora que el archivo migracion.cql está en el contenedor, debemos ejecutar el script para crear la tabla necesaria. Por lo que usamos cqlsh para ejecutar el script:

```
cqlsh -f /migracion.cql
```



```
root@ac832ab43a25:/# cat migracion.cql
USE alexprietos;

CREATE TABLE aire (
    Estacion text,
    Titulo text,
    latitud text,
    longitud text,
    Fecha text,
    Periodo text,
    SO2 text,
    NO text,
    NO2 text,
    CO text,
    PM10 text,
    O3 text,
    dd text,
    vv text,
    TMP text,
    HR text,
    PRB text,
    RS text,
    LL text,
    BEN text,
    TOL text,
    MXIL text,
    PM25 text,
    PRIMARY KEY (Estacion, Fecha, Periodo)
);
root@ac832ab43a25:/# cqlsh -f /migracion.cql
```

Instalamos dsbulk Manualmente en el Contenedor en Ejecución

Actualizar la Lista de Paquetes e instalar wget y unzip

- **wget:** Esta herramienta se utiliza para descargar archivos desde la web. Se necesitará para descargar el archivo dsbulk de la página de DataStax.
- **unzip:** Aunque dsbulk viene en un archivo .tar.gz, instalar unzip es una buena práctica para asegurar que puedas trabajar con archivos comprimidos si es necesario.

apt-get update

apt-get install -y wget unzip

```
root@ac832ab43a25:/# apt-get update
Get:1 http://archive.ubuntu.com/ubuntu jammy InRelease [270 kB]
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/restricted amd64 Packages [164 kB]
Get:6 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,162 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy/main amd64 Packages [1,792 kB]
Get:8 http://archive.ubuntu.com/ubuntu jammy/multiverse amd64 Packages [266 kB]
Get:9 http://archive.ubuntu.com/ubuntu jammy/universe amd64 Packages [17.5 MB]
Get:10 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [2,377 kB]
Get:11 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [3,205 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/multiverse amd64 Packages [44.7 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2,654 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 Packages [51.8 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [3,283 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,450 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy-backports/main amd64 Packages [81.4 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-backports/universe amd64 Packages [33.7 kB]
Fetched 34.7 MB in 13s (2,643 kB/s)
Reading package lists... Done
root@ac832ab43a25:/# apt-get install -y wget unzip
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
wget is already the newest version (1.21.2-2ubuntu1.1).
Suggested packages:
  zip
The following NEW packages will be installed:
  unzip
0 upgraded, 1 newly installed, 0 to remove and 2 not upgraded.
Need to get 175 kB of archives.
After this operation, 386 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 unzip amd64 6.0-26ubuntu3.2 [175 kB]
Fetched 175 kB in 4s (40.3 kB/s)
debconf: delaying package configuration, since apt-utils is not installed
Selecting previously unselected package unzip.
(Reading database ... 8740 files and directories currently installed.)
Preparing to unpack .../unzip_6.0-26ubuntu3.2_amd64.deb ...
Unpacking unzip (6.0-26ubuntu3.2) ...
Setting up unzip (6.0-26ubuntu3.2) ...
```

Descargar y descomprimir dsbulk desde el Sitio Oficial de DataStax

wget <https://downloads.datastax.com/dsbulk/dsbulk-1.8.0.tar.gz>

Este comando usa wget para descargar el archivo dsbulk-1.8.0.tar.gz desde la URL proporcionada

tar -xvzf dsbulk-1.8.0.tar.gz

Esto extrae los contenidos del archivo dsbulk-1.8.0.tar.gz y coloca los archivos en una nueva carpeta llamada dsbulk-1.8.0.

```
dsbulk-1.8.0/lib/HdrHistogram-2.1.12.jar
dsbulk-1.8.0/lib/esri-geometry-api-1.2.1.jar
dsbulk-1.8.0/lib/json-20090211.jar
dsbulk-1.8.0/lib/jackson-core-asl-1.9.12.jar
dsbulk-1.8.0/lib/jackson-core-2.12.1.jar
dsbulk-1.8.0/lib/jackson-databind-2.12.1.jar
dsbulk-1.8.0/lib/jackson-annotations-2.12.1.jar
dsbulk-1.8.0/lib/reactive-streams-1.0.3.jar
dsbulk-1.8.0/lib/dsbulk-connectors-csv-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-config-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-io-1.8.0.jar
dsbulk-1.8.0/lib/commons-compress-1.19.jar
dsbulk-1.8.0/lib/zstd-jni-1.4.5-6.jar
dsbulk-1.8.0/lib/xz-1.8.jar
dsbulk-1.8.0/lib/dec-0.1.2.jar
dsbulk-1.8.0/lib/dsbulk-connectors-api-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-connectors-commons-1.8.0.jar
dsbulk-1.8.0/lib/univocity-parsers-2.9.1.jar
dsbulk-1.8.0/lib/reactor-core-3.4.2.jar
dsbulk-1.8.0/lib/dsbulk-connectors-json-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-workflow-load-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-workflow-api-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-workflow-commons-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-cql-1.8.0.jar
dsbulk-1.8.0/lib/antlr4-runtime-4.9.1.jar
dsbulk-1.8.0/lib/dsbulk-mapping-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-codecs-text-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-format-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-partitioner-1.8.0.jar
dsbulk-1.8.0/lib/metrics-jmx-4.1.17.jar
dsbulk-1.8.0/lib/caffeine-2.8.8.jar
dsbulk-1.8.0/lib/netty-tcnative-boringssl-static-2.0.31.Final.jar
dsbulk-1.8.0/lib/jul-to-slf4j-1.7.26.jar
dsbulk-1.8.0/lib/logback-classic-1.2.3.jar
dsbulk-1.8.0/lib/logback-core-1.2.3.jar
dsbulk-1.8.0/lib/dsbulk-codecs-api-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-batcher-api-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-sampler-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-executor-api-1.8.0.jar
dsbulk-1.8.0/lib/jctools-core-3.2.0.jar
dsbulk-1.8.0/lib/dsbulk-workflow-unload-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-workflow-count-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-batcher-reactor-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-executor-reactor-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-runner-1.8.0.jar
dsbulk-1.8.0/lib/dsbulk-url-1.8.0.jar
dsbulk-1.8.0/lib/lz4-java-1.7.1.jar
dsbulk-1.8.0/lib/snappy-java-1.1.7.3.jar
dsbulk-1.8.0/lib/jansi-1.18.jar
```

Mover dsbulk a un Directorio Permanente

`mv dsbulk-1.8.0 /opt/dsbulk`

Este comando mueve la carpeta dsbulk-1.8.0 a /opt/dsbulk. El directorio /opt es comúnmente utilizado para instalar aplicaciones de terceros y personalizaciones del sistema, por lo que es un buen lugar para alojar dsbulk.

Crear un Enlace Simbólico para Facilitar el Acceso a dsbulk

`ln -s /opt/dsbulk/bin/dsbulk /usr/local/bin/dsbulk`

Este comando crea un **enlace simbólico** (también conocido como "alias") en /usr/local/bin/ para que puedas ejecutar dsbulk desde cualquier lugar del sistema sin tener que escribir la ruta completa.

Verificamos que dsbulk esté instalado correctamente

Para poder verificar que dsbulk este correctamente instalado; llamados al comando:

`dsbulk`

```
root@ac832ab43a25:/# dsbulk
DataStax Bulk Loader v1.8.0
Usage: dsbulk <command> [options]
       dsbulk help [section]

Available commands:
count:
  Computes statistics about a table, such as the total number of rows, the number of rows per token range, the number of rows per host, or the total number of rows in the N biggest partitions in the table. Run "dsbulk help stats" for more information.
load:
  Loads data from external data sources into DataStax Enterprise or Apache Cassandra (X) databases. This command requires a connector to read data from, the target table, or alternatively, the insert query must also be properly configured. Run "dsbulk help schema" for more information.
unload:
  Unloads data from DataStax Enterprise or Apache Cassandra (X) databases into external data sinks. This command requires a connector to write data to, the source table, or alternatively, the read query must also be properly configured. Run "dsbulk help schema" for more information.

Common options:
Note: on the command line, long options referring to DSbulk configuration settings can have their prefix 'dsbulk:' omitted.
Note: on the command line, long options referring to driver configuration settings can be introduced by the prefix 'datastax-java-driver' or just 'driver'.
-v, --version
  Show program's version number and exit.
-h, --help
  This help text. May be combined with -c <connectorName> to see short options for a particular connector.
-c, --cstrings <file>
  Load options from the given file rather than from 'dsbulk.conf/application.conf'.
-c, --connector <name>
  dsbulk.connector.name cstrings
  The name of the connector to use.
  Default: 'csw'.
-c, --connector.csv.uri <uri>
  dsbulk.connector.csv.uri cstrings
  The URI or path of the resource(s) to read from or write to.
  which URI protocols are available depend on which URI stream handlers have been installed, but at least the file protocol is guaranteed to be supported for reads and writes, and the http and https protocols are guaranteed to be supported for writes.
  The file protocol can be used with all supported file systems, local or remote.
  In the case of a directory, the 'file:dsbulk.conf/application.conf' setting can be used to filter files to read, and the 'recursive' setting can be used to control whether the reading the URI can point to a single file, or to an entire directory.
  In the case of a directory, the 'file:dsbulk.conf/application.conf' setting can be used to filter files to read, and the 'recursive' setting can be used to control whether the reading the URI can point to a single file, or to an entire directory.
```

Usar DataStax Bulk Loader para Migrar los Datos

Una vez que la tabla esté creada, ya podemos usar dsbulk para cargar los datos desde el archivo CSV.

`dsbulk load -k alexprieto -t aire -url /calidadAire.csv -delim ',' -header true`

-k alexprieto: Especifica el keyspace (alexprieto).

-t aire: Especifica el nombre de la tabla (aire).

-url /calidadAire.csv: Ruta al archivo CSV.

-delim ',': Indica que el archivo CSV usa comas como delimitadores.

-header true: Indica que el archivo CSV tiene una fila de encabezados.

```
root@ac832ab43a25:/# head -n 5 calidadAire.csv
estacion,titulo,latitud,longitud,fecha,periodo,so2,no,no2,co,pm10,o3,dd,vv,tmp,hr,prb,rs,so2,titulo,tmx,tol,vv
1,Estacion Avenida Constitucion,43.529886,-5.673428,2022-12-31,24,2,1,11,,29,54,137,1.59,14.1,75,1015,56,0,0.04,0.03,0.03,10
1,Estacion Avenida Constitucion,43.529886,-5.673428,2022-12-31,23,5,1,8,,32,66,202,2.22,14.1,73,1015,56,0,0.02,0.02,0.03,10
1,Estacion Avenida Constitucion,43.529886,-5.673428,2022-12-31,22,2,1,6,,33,78,156,1.86,14.6,69,1014,56,0,0.02,0.02,0.03,12
1,Estacion Avenida Constitucion,43.529886,-5.673428,2022-12-31,21,1,3,13,,36,67,222,2.11,17.7,46,1014,56,0,0.02,0.05,0.03,14

root@ac832ab43a25:/# dsbulk load -k alexprieto -t aire -url /calidadAire.csv -delim ',' -header true
Operation directory: /logs/LOAD_20241027-143542-794383
total | failed | rows/s | p50ms | p99ms | p999ms | batches
43,808 | 0 | 11,913 | 60.02 | 1,035.99 | 1,283.46 | 31.92
Operation LOAD_20241027-143542-794383 completed successfully in 3 seconds.
Last processed positions can be found in positions.txt
```

Verificar que la Migración Se Realizó Correctamente

Para verificar que los datos se han migrado correctamente, ejecutamos una consulta en cqlsh básica:

`cqlsh -e "SELECT * FROM alexprieto.aire LIMIT 10;"`

```
root@ac832ab43a25:/# cqlsh -e "SELECT * FROM alexprieto.aire LIMIT 10;"
```

estacion	fecha	periodo	ben	co	dd	hr	latitud	ll	longitud	mxl	no	no2	o3	pm10	pm25	prb	rs	so2	titulo	tmx	tol	vv
10	2022-01-01	1	null	null	180	89	43.517315	0	-5.672499	null	1	15	37	45	32	1018	2	3	Estacion de Montevill	18	null	0.8
10	2022-01-01	10	null	null	337	89	43.517315	0	-5.672499	null	1	17	30	68	27	1018	7	5	Estacion de Montevill	19.6	null	0.35
10	2022-01-01	11	null	null	236	89	43.517315	0	-5.672499	null	4	24	24	53	32	1018	10	4	Estacion de Montevill	19.8	null	0.47
10	2022-01-01	12	null	null	196	89	43.517315	0	-5.672499	null	4	25	32	55	38	1018	14	3	Estacion de Montevill	21.1	null	0.75
10	2022-01-01	13	null	null	181	89	43.517315	0	-5.672499	null	18	32	26	54	34	1018	24	3	Estacion de Montevill	21.7	null	0.6
10	2022-01-01	14	null	null	180	89	43.517315	0	-5.672499	null	1	10	59	51	36	1017	27	3	Estacion de Montevill	24.7	null	1.7
10	2022-01-01	15	null	null	176	89	43.517315	0	-5.672499	null	1	6	60	32	21	1017	18	2	Estacion de Montevill	24.7	null	1.9
10	2022-01-01	16	null	null	200	89	43.517315	0	-5.672499	null	1	7	60	35	16	1016	10	2	Estacion de Montevill	24.3	null	1.75
10	2022-01-01	17	null	null	188	89	43.517315	0	-5.672499	null	1	8	60	35	21	1016	4	2	Estacion de Montevill	24.3	null	2.1
10	2022-01-01	18	null	null	196	89	43.517315	0	-5.672499	null	1	10	56	38	25	1016	2	2	Estacion de Montevill	23.7	null	1.65