Apache Cassandra

Cristian Andrés Donoso Fernández

Sergio Alejandro López Delgado

Ariel Mejía

Universidad de San Buenaventura

Bases de Datos Distribuidas

Bogotá D.C.

2019

**APACHE CASSANDRA**

* **¿Qué Es?**

Apache Cassandra es una base de datos [NoSQL](https://es.wikipedia.org/wiki/NoSQL) distribuida y basada en un modelo de almacenamiento de clave-valor, de código abierto que está escrita en [Java](https://es.wikipedia.org/wiki/Java_(lenguaje_de_programaci%C3%B3n)), cuyo principal objetivo es la escalabilidad lineal y la disponibilidad.

* **¿Cómo Se Usa?**

Cassandra maneja el lenguaje CQL, el cual ofrece un modelo cercano a SQL en el sentido de que los datos se colocan en tablas que contienen filas de columnas. Por esa razón, los términos (tablas, filas y columnas) tienen la misma definición que tienen en SQL. En Cassandra los datos están desnormalizados de manera que el concepto de joins o subqueries no existe.

Podemos interactuar con Cassandra mediante CQL a través de la shell de CQL, cqlshell. También podemos usar herramientas gráficas como [DevCenter](http://www.datastax.com/what-we-offer/products-services/devcenter) o a través de los drivers soportados para múltiples lenguajes de programación.

* **¿Quiénes Lo Usan?**

Cassandra es una solución brillante para muchos casos de uso que podemos encontrar en el mundo Big Data, es por esto por lo que ha sido utilizada por muchas organizaciones reconocidas, entre las cuales se pueden destacar las siguiente:

* AppScale
* Constant Contact
* Digg
* Facebook
* Twitter
* IBM
* Instagram
* Spotify
* Netflix
* Reddit.
* **Arquitectura**

Implementa una arquitectura Peer-to-Peer, lo que elimina los puntos de fallo único y no sigue patrones maestro-esclavo como otros sistemas de almacenamiento. De esta manera cualquiera de los nodos puede tomar el rol de coordinador de una query. Será el driver el que decida qué nodo quiere que sea el coordinador.

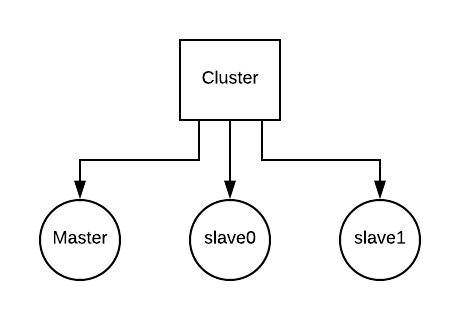
Los datos son repartidos a lo largo del cluster en base a un token único calculado para cada fila por una función hash.

* **Características**
* Cassandra nos proporciona tolerancia a particiones y disponibilidad, pero a cambio de ser eventualmente consistente, tal y como define el [teorema CAP](https://es.wikipedia.org/wiki/Teorema_CAP). El nivel de consistencia puede ser configurado, según nos interese, incluso a nivel de query.
* Es distribuida, lo quiere decir que la información está repartida a lo largo de los nodos del cluster. Además, ofrece alta disponibilidad, de manera que si alguno de los nodos se cae el servicio no se degradará.
* Escala linealmente, lo que quiere decir que el rendimiento de forma lineal respecto al número de nodos que añadamos. Por ejemplo, si con 2 nodos soportamos 100.000 operaciones por segundo, con 4 nodos soportaremos 200.000.
* Escala de forma horizontal, lo que quiere decir que podemos escalar nuestro sistema añadiendo nuevos nodos basados en hardware commodity de bajo coste.
* Implementa una arquitectura Peer-to-Peer.
* Soporta el concepto de data center para agrupar los nodos lógicamente y tener los datos más cerca del usuario.
* Combina propiedades de una base de datos clave-valor y una orientada a columnas.
* **Manual De Instalación En CentOS 7**

1. Selección del motor:

CentOS puede mejorar el rendimiento y el equilibrio de carga de los recursos configurando los equipos para que funcionen de forma colectiva, con un grupo de servidores que comparten un sistema de archivos común y que ofrecen aplicaciones de alta disponibilidad. Además, CentOS 7 goza de una mayor estabilidad a largo plazo con menos errores y agujeros de seguridad en comparación con otros motores.

1. Diseño del cluster:



|  |  |  |
| --- | --- | --- |
| **Nodo** | **IP** | **Host** |
| master | 192.168.37.55 | master.cassandra.bdd.laboratorios.usbbog.edu.co |
| slave0 | 192.168.37.67 | slave0.cassandra.bdd.laboratorios.usbbog.edu.co |
| slave1 | 192.168.37.58 | slave1.cassandra.bdd.laboratorios.usbbog.edu.co |

1. Los 3 nodos tienen que estar conectados a una misma red de Internet.
2. Antes de instalar Cassandra en cada uno de los nodos, primero se deben configurar correctamente sus dominios, para esto usaremos el comando *“gedit /etc/hostname”.*

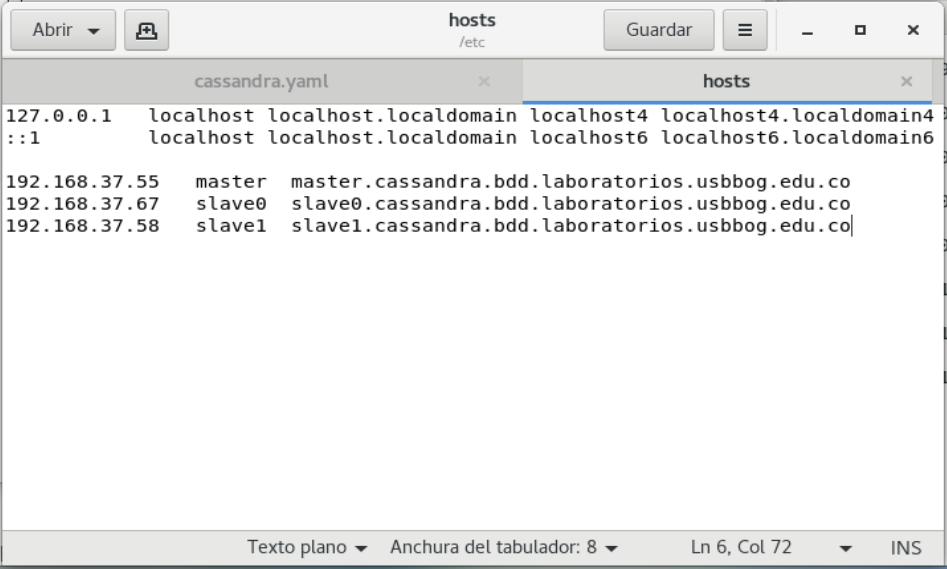
Se debe tener en cuenta que todos los comandos de este manual se deben ejecutar desde el usuario root.





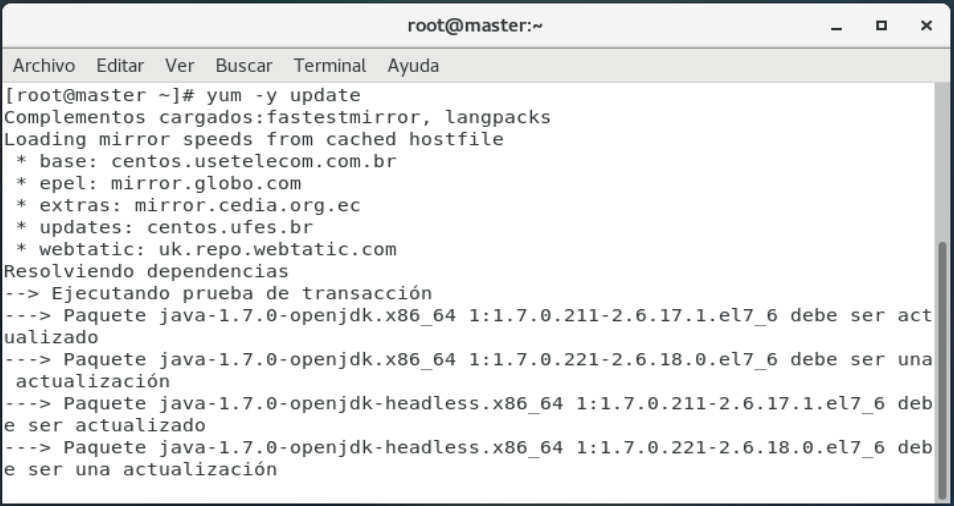
1. Una vez configurados los dominios de los nodos, se deben relacionar todos ellos junto a su correspondiente IP editando el archivo “*/etc/hosts*” en cada máquina de la siguiente forma:



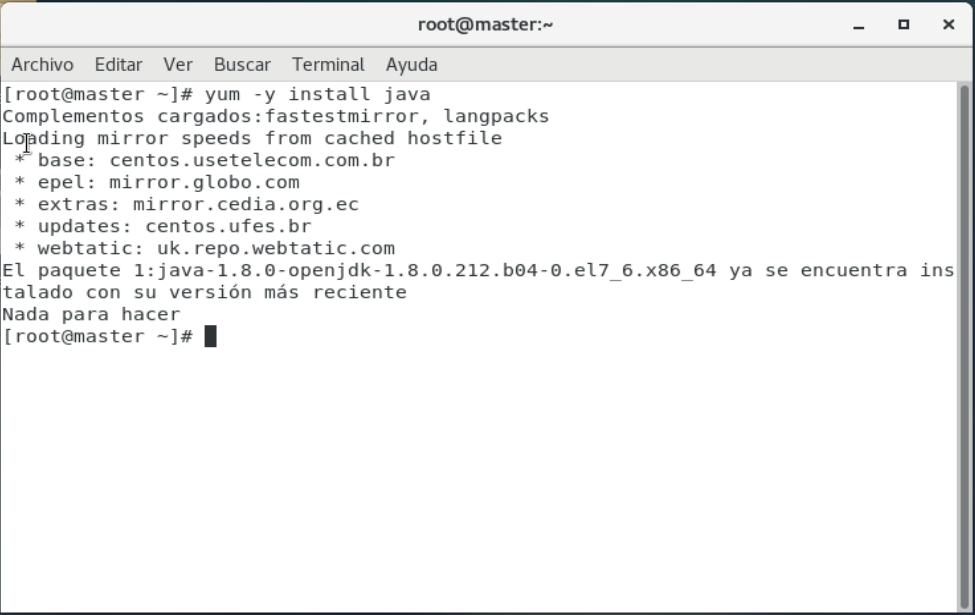


1. Para comenzar la instalación de Cassandra lo primero que se debe hacer es actualizar los repositorios utilizando el comando “*yum – y update”.*





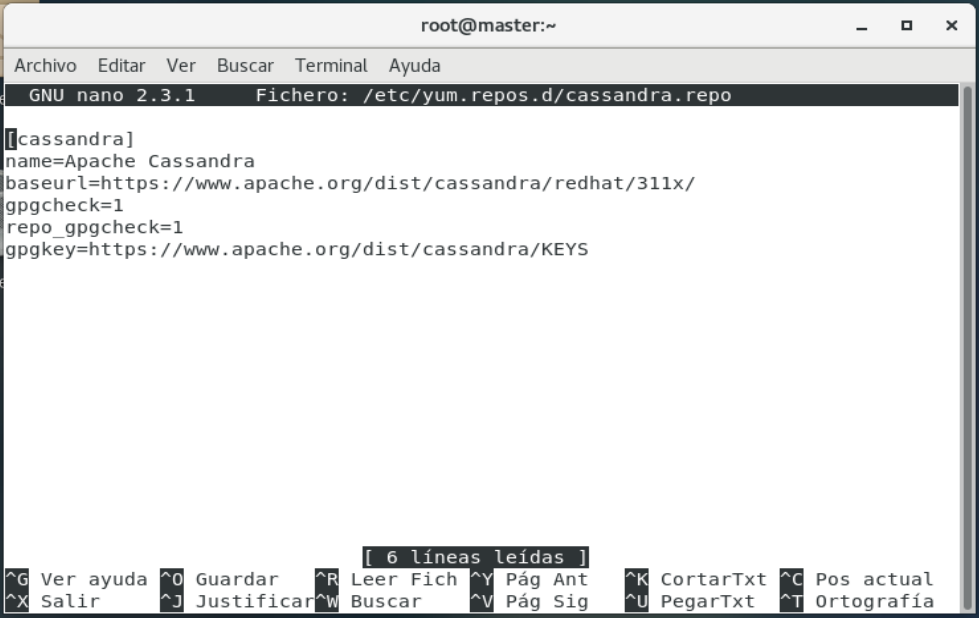
1. Cuando termine la actualización de los repositorios se debe instalar java mediante el comando “*yum -y install java”.*



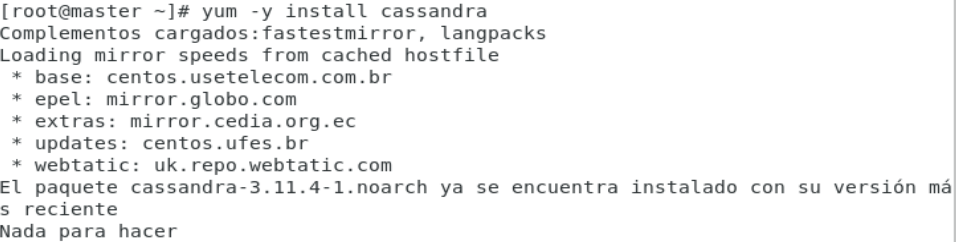
1. A continuación, se debe crear un nuevo repositorio de Cassandra con el comando: *“nano /etc/yum.repos.d/cassandra.repo”.*



*Y* luego de crear el nuevo repositorio se debe editar de la siguiente manera:



1. Ahora instalamos Cassandra con el comando *“yum install cassandra”.*



1. Recargar los demonios del sistema ejecutando *“systemctl daemon-reload”.*



1. Iniciamos el servicio de Cassandra con el comando *“systemctl start cassandra”.*



1. Consultar el estado del servicio ejecutando *“systemctl start cassandra”.*

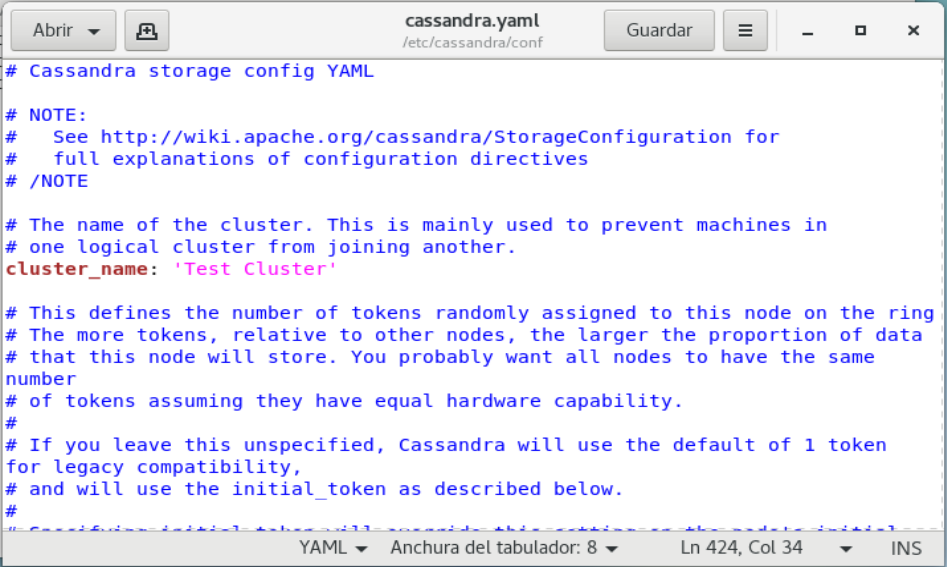


* **Configuración De Cassandra Para Conectar los 3 nodos:**

Este proceso se debe realizar en cada uno de los nodos, cambiando la IP respectivamente para cada uno.

1. Abrir el archivo *“cassandra.yaml”* que se encuentra en el directorio “*etc/cassandra/conf”.*



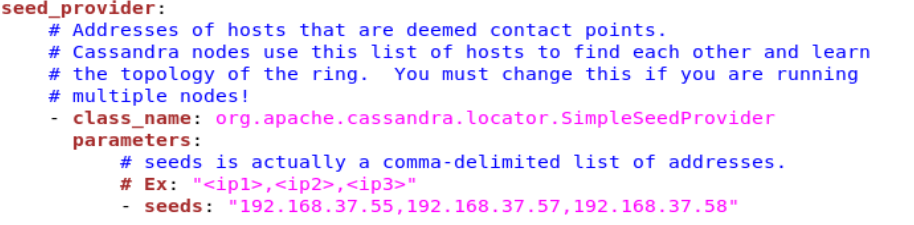


1. Dentro del archivo debemos buscar y editar las siguientes líneas:

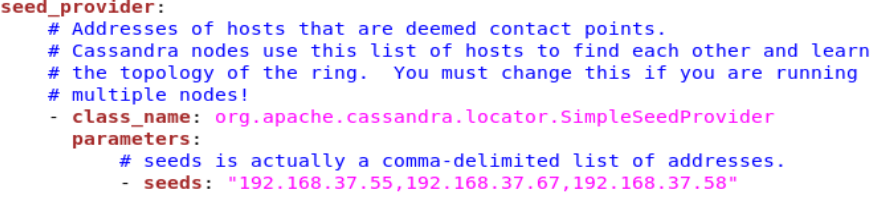
* *cluster\_name:* Darle nombre al cluster y verificar que sea el mismo para todos los nodos.



* *seeds:* Indica las IPs de los servidores asociados al cluster (incluyendo la del nodo en donde se está haciendo esta configuración) separadas por comas.



También se debe quitar toda la línea de *“# Ex”.*



* *num\_tokens:* Indica el peso del servidor a la hora de repartir datos. Se debe descomentar quitando el #.



* *listen\_address:* Se debe colocar la IP del nodo para permitir que otros nodos se comuniquen con este.



* *broadcast\_address:* De igual forma se debe poner la IP de la máquina.



* *rpc\_address:* De igual forma se debe cambiar por la IP del servidor para permitir que otras aplicaciones externas se puedan conectar.



* *start\_rpc:* Se debe colocar en “true”.



* *rpc\_address:* Debe colocarse en *“0.0.0.0”.*



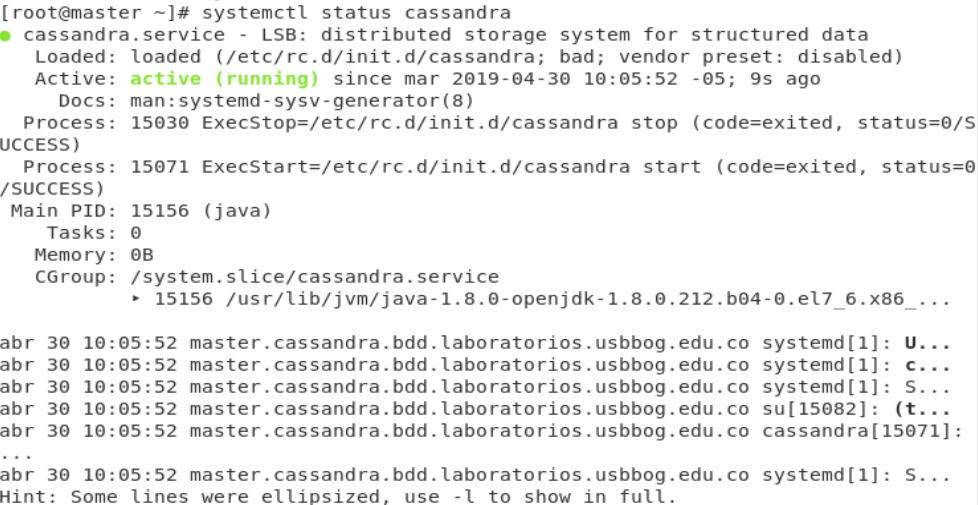
* *broadcast\_rpc\_address:* Esta etiqueta debe estar descomentada y en ella se debe poner la dirección IP de la máquina.



1. Luego de guardar los cambios hechos en el archivo cassandra.yaml, se debe reiniciar el servidor de cassandra ejecutando “systemctl restart cassandra”.



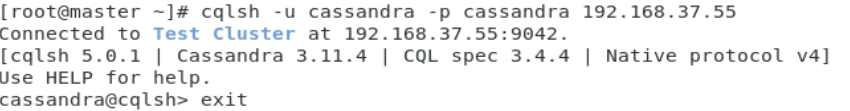
1. Podemos verificar que el servicio efectivamente este activo con el comando *“systemctl status cassandra”*. (En caso de presentar errores, ir al paso 8)



1. Se debe verificar que el firewall este desactivado en todos los nodos. Para detener el firewall se usa el comando *“systemctl stop firewalld”.*

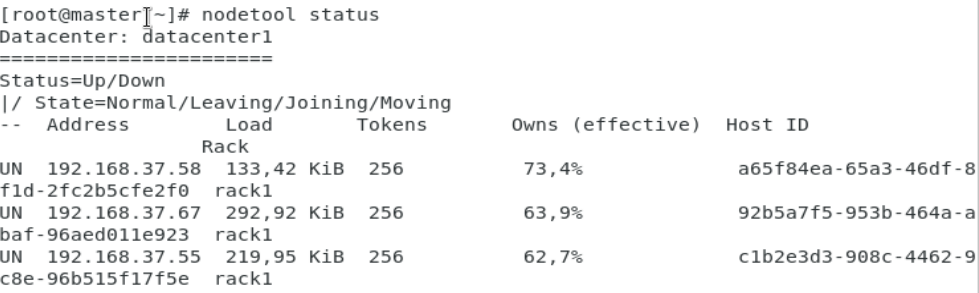


1. Ejecutar el comando *“cqlsh -u cassandra -p cassandra (IP host)”* para poder empezar a manejar Cassandra y que con esto sabremos si se tiene conexión al nodo.

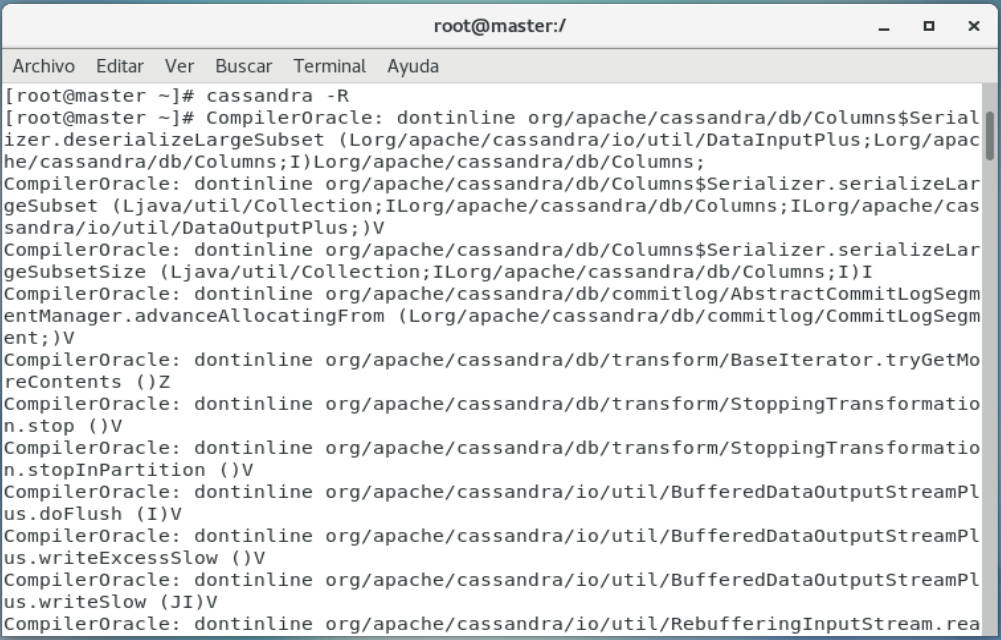


De igual forma se puede verificar la conexión con los otros nodos poniendo sus IPs.

1. Salimos de cqlsh escribiendo *“exit”* y ejecutamos el comando *“nodetool status”* para visualizar los nodos que están conectados . Debe aparecer algo como lo siguiente:



1. Si se llega a presentar algún error en los pasos 4, 6 o 7, utilizar el comando *“cassandra -R”* para visualizar los errores y poderlos solucionar. Si no llega a registrar errores, volver a probar los pasos 4 y 6.



# REFERENCIAS

Cassandra, A. (s.f.). *Cassandra Documentation.* Obtenido de http://cassandra.apache.org/doc/latest/

Chacaltana, G. (21 de Abrl de 2019). *SoloCodigoWeb.* Obtenido de http://www.solocodigoweb.com/blog/2018/05/18/cuando-usar-cassandra-cuando-usar-mongodb/

Chuwiki. (3 de Julio de 2013). Obtenido de http://chuwiki.chuidiang.org/index.php?title=Cluster\_Cassandra\_con\_dos\_servidores

Zaforas, M. (17 de Marzo de 2016). *Paradigma.* Obtenido de https://www.paradigmadigital.com/dev/cassandra-la-dama-de-las-bases-de-datos-nosql/