DSE Installation & Configuration Guidance

Introduction

DataStax Enterprise (DSE) is a complex persistence store, search and compute engine. It is very demanding on disk performance, memory footprint and processing power.

These instructions assume an Ubuntu-based installation hosted in Amazon Web Services (AWS). They are designed to be supplemental to official DataStax documentation and the DataStax Partner Network (DSPN) resources on GitHub (https://github.com/DSPN).

EC2 Instance Types

Due to the balanced workload requirements of DSE, M4 instances are recommended. Here's a quick indication for each of the sizing options. The **m4.4xlarge** instance type tends to be the preferred one for most production use cases.

Note that in all cases, SSD are strongly preferred to "spinning disks" HDD.

EC2 Type	**vCPU**	**RAM (GB)**	**Usage Recommendation**
m4.large	2	8	Lightweight Cassandra development and testing only
m4.xlarge	4	16	Moderate Cassandra development and testing
m4.2xlarge	8	32	Suitable for Production Cassandra Cassandra + Search ***or*** Cassandra + Graph for dev & test
m4.4xlarge	***16***	***64***	***Best option for Cassandra + Search + Graph + Analytics in production environments on AWS.***
m4.10xlarge	40	160	Can be used in a multi-instance configuration, but resources are usually better spent scaling out with the m4.4xlarge.
m4.16xlarge	64	256	Can be used in a multi-instance configuration, but resources are usually better spent scaling out with the m4.4xlarge.

Chart is based on https://aws.amazon.com/ec2/instance-types/ on March 27, 2014.

Also note the following when using AWS:

- Prefer consistency throughout the cluster by using the same:
    - AMI
    - OS: vendor, version and patches
    - Java vendor and version
    - DSE version
    - OS, JVM, and DSE Configuration
- Start with a single region. If/when expanding to other regions, use Cassandra's configuration capabilities to limit replication traffic and querying across WAN.
- While a lot of ports need to be exposed within the cluster, very few (e.g. SSH, HTTP(S) access) will need to be exposed outside of the AWS VPCs.

Installation & Configuration Notes

Refer first to the DataStax document on Recommended production settings for Linux.

Cluster Size

A development cluster can be run on a single node, but this has no protections against data loss nor does it provide any of Cassandra's scaling benefits. A better size is 4 nodes since this can tolerate the loss of 1 node while configured with a replication factor of 3.

A production cluster should not be less than 5 nodes with a replication factor of 3. This can tolerate the loss of up to 2 nodes.

JVM

The Oracle JVM is preferred to OpenJDK, but it is also more involved and doesn't include the full set of encryption functionality due to export restrictions.

A sample Oracle installation script is in the DSPN installation scripts:

```
# Install add-apt-repository
sudo apt-get -y install software-properties-common

sudo add-apt-repository -y ppa:webupd8team/java
sudo apt-get -y update
echo debconf shared/accepted-oracle-license-v1-1 select true | sudo
debconf-set-selections
echo debconf shared/accepted-oracle-license-v1-1 seen true | sudo
debconf-set-selections
sudo apt-get -y install oracle-java8-installer
```

If using the Oracle JDK and planning to use encryption, the Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy Files should be installed as well. See: Installing Java Cryptography Extension (JCE) Files, or just:

```
sudo apt-get install oracle-java8-unlimited-jce-policy
```

DSE

Access to DataStax Academy is required to download DSE software. The primary DataStax instructions are at Installing DataStax Enterprise 5.0 on Debian-based systems using APT.

Once the software is installed, but *before starting the DSE service*, several configuration matters must be addressed. These include:

File: /etc/default/dse

- Set node type: SOLR_ENABLED=1, SPARK_ENABLED=1, GRAPH_ENABLED=1
- Must be identical within a Cassandra Data Center

File: /etc/dse/dse.yaml (Docs: Configure dse.yaml)

- Configure authentication
- Configure authorization
- System encryption settings

File: /etc/dse/cassandra/cassandra.yaml (Docs: Configure cassandra.yaml)

- **cluster_name**
- **num_tokens** - must be consistent across cluster, generally 64 or 128
- **listen_address** - generally the private IP of the node
- **broadcast_address** - generally the public IP of the node
- **rpc_address** - generally 0.0.0.0
- **broadcast_rpc_address** - generally the public IP of the node
- File locations:
    - **hints_directory** -
    - **cdc_raw_directory** - should be consistent with other file locations
    - **commitlog_directory** - prefer separate partition from data directory
    - **data_file_directories** - prefer separate partition from commitlog directory
    - **saved_caches_directory**
- **endpoint_snitch**
    - **Ec2Snitch** for single-region AWS
    - **Ec2MultiRegionSnitch** for multi-region AWS
- **seed_provider**: **seeds** - prefer at least two for the cluster
- **authenticator** - for enabling authentication
- **authorizer** - for enabling authorization