Connection d'un Cluster Apache Spark à un Cluster Apache Cassandra

30 août 2018

1 Objectifs

Nous disposons d'un ensemble de 12 serveurs dans chacun desquels nous avons installé Apache Spark et Apache Cassandra. Nous avons configuré les instances Spark pour former un cluster à 12 machines. De même, nous avons configuré les instances Cassandra pour former également un autre cluster à 12 machines. Notre cluster Spark et notre cluster Cassandra cohabitent donc sur le même ensemble de serveurs. Le but de ce document est de montrer comment inter-connecter ces deux clusters de sorte que les données transformées et structurées convenablement par le cluster Spark puissent être intégrées directement dans des tables du cluster Cassandra. Réciproquement, cette connection permettra d'importer dans le cluster Spark les tables du cluster Cassandra et d'y effectuer des requêtes ou des traitements statistiques.

2 Installation des drivers de connection Spark-Cassandra

Deux drivers ont été créés pour permettre d'établir une connection entre un cluster Spark et un cluster Cassandra. Il s'agit des drivers "`spark-cassandra-connector-2.3.0-s_2.11.jar`" et "`jsr166e-1.1.0.jar`". Ces drivers doivent être disponibles dans tous les server du cluster Spark. Nous choisissons de stocker ces drivers dans le repertoire "`/usr/share/spark-2.3.0-bin-hadoop2.7/jars/`" de chaque serveur Spark. On peut donc télécharger et sauvegarder ces deux drivers avec les comandes suivantes :

```
# cd /usr/share/spark-2.3.0-bin-hadoop2.7/jars

# wget http://dl.bintray.com/spark-packages/maven/datastax/spark-cassandra-connector/2.3.0-s_2.11/spark-cassandra-connector-2.3.0-s_2.11.jar

# wget http://central.maven.org/maven2/com/twitter/jsr166e/1.1.0/jsr166e-1.1.0.jar
```

3 Ouverture de la connection Spark-Cassandra

Dans cette section, nous supposons que les clusters Spark et Cassandra sont tous les deux en marche. On peut se référer au document nommé "**Install Spark Cluster**" pour la mise en marche de notre cluster Spark (démarrage des serveurs Zookeeper, des serveurs master de Spark et des serveurs worker ou slave de Spark) et au document nommé "**Cassandra cluster configuration**" pour la mise en marche du cluster Cassandra.

Cette section illustre l'inter-connection Spark-Cassandra à partir du client "`spark-shell`" de Spark. Plus précisément, nous allons nous connecter au cluster Spark via son shell en lui fournissant les informations nécessaires pour accéder automatiquement à notre cluster Cassandra. La commande ci-dessous permet de réaliser cette inter-connection :

```
# spark-shell --master spark://10.10.10.5:7077,10.10.10.6:7077,10.10.10.7:7077 \
--jars /usr/share/spark-2.3.0-bin-hadoop2.7/jars/spark-cassandra-connector-2.3.0-s_2.11.jar \
--jars /usr/share/spark-2.3.0-bin-hadoop2.7/jars/jsr166e-1.1.0.jar \
--conf spark.cassandra.connection.host=10.10.10.13,10.10.10.11,10.10.10.6,10.10.10.8 \
--conf spark.cassandra.auth.username=dba_1 \
--conf spark.cassandra.auth.password=dba_1*pw!
```

Dans cette commande, nous avons indiqué les éléments suivants.

1. Les adresses IP des serveurs master de notre cluster Spark.
2. Les drivers de connections Spark-Cassandra.
3. Les adresses IP de quelques serveurs Cassandra situés dans le même data center de notre cluster Cassandra.
4. Un compte et son mot de passe pour un utilisateur de notre cluster Cassandra.

A ce stade, la connection est établie entre le cluster Spark et le cluster Cassandra. A l'aide de quelques commandes spécifiques, on peut exporter les données de Spark vers Cassandra et on peut importer les tables de Cassandra vers Spark.

4 Conclusion

Dans ce document, nous avons montré comment établir la connection entre un cluster Spark et un cluster cassandra. Dans les prochains documents consacrés à Spark et à cassandra, nous montrerons à travers quelques exemples comment réaliser les échanges de données entre ces deux clusters. Notons que la connection entre Spark et Cassandra peut également être établie directement à l'intérieur d'une application Spark compilée et packagée. Nous présenterons cela dans un futur document.

Références