



Projet de la matière:

SMB2 I 4: Réseaux ,systèmes embarqués et répartis

Sujet du projet : Apache Cassandra

Préparé par : Sanaa Allaw



Cassandra



Plan du projet

- I. Origine et concept du Cassandra
- II. Caractéristiques principales
- III. Architecture Cassandra
- IV. L'indexation dans Cassandra
- V. Communication entre les nœuds
- VI. Sécurité dans appache cassandra
- VII. Installation et configuration
- VIII. Opération sur les keyspace est sur les tables dans Cassandra

Exemple pratique sur Cassandra

I. Origine et concept du cassandra

Initialement apache Cassandra a été développée en interne par Facebook puis a été diffusée en Open Source.

- Il est écrit en Java
- Utilise un protocole de sérialisation créé par facebook
- Cassandra est une base de données distribuée qui permet de stocker une grande quantité de données grâce à sa scalabilité horizontale
- Fournit un schéma de données dynamique afin d'offrir un maximum de flexibilité et de performance.
- Est une base de données NoSQL appartenant à la famille des bases de données orientées colonnes
- Plusieurs grandes sociétés utilisent Cassandra pour leur application grand public. C'est le cas de Facebook, Twitter
- Mais pour bien comprendre cet outil, il faut tout d'abord bien assimiler le vocabulaire de base.

Colonne : Une colonne est la plus petite unité du modèle de données de Cassandra. C'est un triplet contenant un nom, une valeur et un timestamp. Ce dernier sert à déterminer la mise à jour la plus récente.

Nom
Valeur
<i>Timestamp</i>

ligne : Une ligne est composée d'un ensemble de colonnes. Elle est identifiée par une clé.

Ligne		Colonne		
baronm	familyName	firstName	age	address
	BARON	Mickael	36	Poitiers
duponto	familyName	firstName	phone	Nom colonne
	DUPONT	Olivier	+335432312	Valeur
Clé				

Une famille de colonnes : ou *column family* en anglais est un regroupement logique de lignes. Pour faire le parallèle avec le monde des bases de données relationnelles, une famille de colonnes est en quelque sorte une table.

Persons				
baronm	familyName	firstName	age	address
	BARON	Mickael	36	Poitiers
duponto	familyName	firstName	phone	
	DUPONT	Olivier	+335432312	

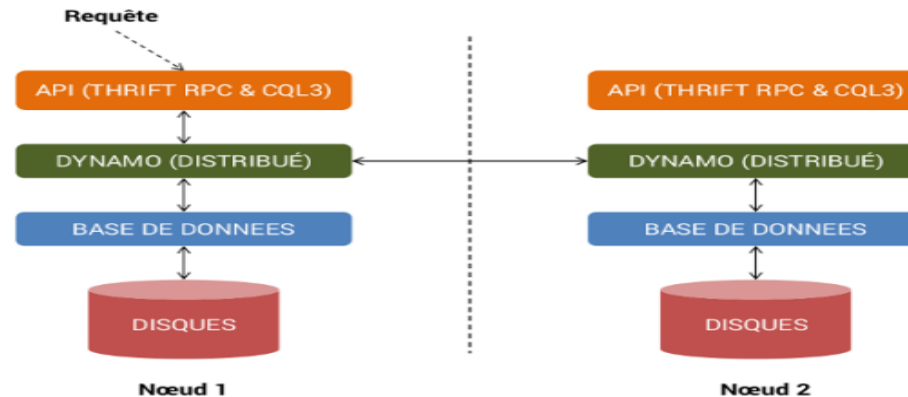
Keyspace :est un regroupement de famille de colonnes. Il s'agit d'une sorte de schéma si on compare au monde des bases de données relationnelles.

CassandraDEMO				
Persons				
baronm	familyName	firstName	age	address
	BARON	Mickael	36	Poitiers
duponto	familyName	firstName	phone	
	DUPONT	Olivier	+335432312	
...				

II. Caractéristiques principales

- **Tolérance aux pannes** : les données d'un nœud (un nœud est une instance de Cassandra) sont automatiquement répliquées vers d'autres nœuds (différentes machines). Ainsi, si un nœud est hors service les données présentes sont disponibles à travers d'autres nœuds. Le terme de facteur de réplication désigne le nombre de nœuds où la donnée est répliquée. Par ailleurs, l'architecture de Cassandra définit le terme de cluster comme étant un groupe d'au moins deux nœuds et un data center comme étant des clusters délocalisés. Cassandra permet d'assurer la réplication à travers différents data center. Les nœuds qui sont tombés peuvent être remplacés sans indisponibilité du service.
- **Décentralisé** : dans un cluster tous les nœuds sont égaux. Il n'y a pas de notion de maître, ni d'esclave, le protocole GOSSIP utilisé pour découvrir la localisation et les informations sur l'état des nœuds d'un cluster.
- **Élastique** : lorsqu'un nouveau serveur est ajouté dans le cluster. Par ailleurs, Cassandra assure qu'il n'y aura pas d'indisponibilité du système ni d'interruption au niveau des applications.
- **Haute disponibilité** : possibilité de spécifier le niveau de cohérence concernant la lecture et l'écriture. On parle alors de *Consistency*. Apache Cassandra ne dispose pas de transaction. L'écriture des données est très rapide comparée au monde des bases de données relationnelles.

III. Architecture Cassandra



Sur le schéma ci-dessus, on distingue la présence de 3 couches métiers :

- API, responsable de recevoir les requêtes venant des clients sous format Thrift (protocole RPC)
 - Dynamo, responsable de la distribution des données entre différents noeuds et du protocole peer-to-peer
 - Base de données, responsable de la persistance des données sur disques
- Cassandra reprend les concepts de 2 bases de données existantes. La première BigTable, créé par Google, pour son modèle de données orienté colonne et son mécanisme de persistance sur disque, et la seconde Dynamo, créé par Amazon, pour son achitecture distribuée sans noeud maître.

IV.l'indexation dans cassandra

Un index est une structure de données qui permet un accès rapide ainsi qu'une recherche de données par rapport à un ensemble de critères donnés.

- Dans les bases de données classiques, une clé primaire est une clé unique utilisée pour identifier chaque ligne d'une table. Cela permet, comme tous les index, d'accélérer l'accès aux données dans une table.
- Dans Cassandra, l'index primaire d'une famille de colonnes correspond à l'index de ses clés de ligne.

Composants de Cassandra

Les principales composantes de Cassandra sont les suivantes -

Node - Il est l'endroit où les données sont stockées.

Cluster - Une grappe est un composant d'un ou de plusieurs noeuds.

Data center - Il est une collection de cluster..

Commit log - Le journal commettre est un mécanisme crash-reprise de Cassandra. Chaque opération d'écriture est écrit dans le journal de validation.

Mem table - Un mem-table est une structure de données résidant en mémoire. Après commettre journal, les données seront écrites dans le mem-table.

SSTable - Il est un fichier de disque sur lequel les données sont rincées de la mem-table lorsque son contenu atteint une valeur de seuil.

V.Communication entre les neouds dans cassandra

- Cassandra utilise un protocole appelé **Gossip** afin de découvrir la localisation et les informations sur l'état des autres noeuds du cluster. Le

protocole **Gossip** est un protocole de communication de type *peer-to-*

peer dans lequel les noeuds échangent périodiquement des informations sur

leur état mais également sur ce qu'ils savent des autres noeuds.

- ❖Avec Gossip, les clients connaissent qui possède quoi. Il s' agit de garder l'information à jour entre les nœuds

- Types de flux:

- 1.Flux datas : enregistrement et lecture de la donnée,

- 2.Flux de services : communication entre les noeuds, gestion du partitionnement, stratégie de la réplication.

VI.Sécurité dans Apache cassandra

Authentification interne

-Gestion des ID de login et des mots de passe dans la base de données.

Gestion de la permission des objects

-Contrôle d'accès aux objets et des actions des utilisateurs dans la base de données.

Encryption client a Noeud

-Protéger les données naviguant vers et depuis le cluster de la base de données.

-S'assure que les données ne peuvent pas être interceptées lors de l'acheminement au serveur.

VII.Installation de Cassandra sur Ubuntu

I – Installation

Étape I - Installation de la machine virtuelle Java

Pour rendre le package Oracle JRE disponibles, vous devrez ajouter personnelles Archives emballage (PPA) en utilisant cette commande:

```
sudo add-apt-repository ppa:webupd8team/java
```

Mettre à jour la base de données du paquet:

```
sudo apt-get update
```

Ensuite, installez le Oracle JRE. L'installation de ce paquet particulier non seulement l'installe mais

aussi rend le JRE par défaut. Lorsque vous êtes invité, acceptez le contrat de licence:

```
sudo apt-get install oracle-java8-set-default
```

Après l'avoir installé, vérifiez qu'il est maintenant le JRE par défaut:

```
java -version
```

Vous devriez voir une sortie similaire à ce qui suit:

```
Output java version "1.8.0_60"Java(TM) SE Runtime Environment (build 1.8.0_60-b27)Java HotSpot(TM) 64-Bit Server VM (build 25.60-b23, mixed mode)
```

Étape 2 - Installation de Cassandra

- Installez les nouvelles mises à jour à l'aide des commandes suivantes

`sudo apt-get update`

`sudo apt-get upgrade`

- Ouvrir `/etc/apt/sources.list` utilisant la commande suivante

`sudo gedit /etc/apt/sources.list`

- Ajouter les lignes suivantes

`deb http://www.apache.org/dist/cassandra/debian 10x main`

`deb-src http://www.apache.org/dist/cassandra/debian 10x main`

- Inscrivez-vous et ajoutez une touche de clés publiques et de mettre à jour à nouveau, notez que vous devrez peut-être modifier la clé conséquence

`gpg --keyserver wwwkeys.pgp.net --recv-keys 4BD736A82B5C1B00`

`sudo apt-key add ~/.gnupg/pubring.gpg`

`sudo apt-get update`

- Installez Cassandra utilisant la commande suivante

`sudo apt-get install cassandra`

VIII. Quelques Operation sur les keyspace dans cassandra

un KEYSPACE appelé *cassandrademocql* via la commande *CREATE KEYSPACE*, en utilisant la stratégie de placement 'SimpleStrategy' et un facteur de réplication à 1.

SimpleStrategy

Utilisez uniquement pour un seul centre de données. SimpleStrategy place la première réplique sur un noeud déterminé par le programme de partitionnement

CQLSH

```
CREATE KEYSPACE cassandrademocql WITH REPLICATION = { 'class' :  
'SimpleStrategy', 'replication_factor' : 1 };
```

•**CQLSH**

```
USE cassandrademocql
```

•**CQLSH**

•**ALTER KEYSPACE** cassandrademocql WITH strategy_class=SimpleStrategy AND
strategy_options:replication_factor=2;

CQLSH

• **DROP KEYSPACE** cassandrademocql

Quelques Operation sur la table

- **CREATE TABLE** emp (emplD int, deptID int, first_name varchar, last_name varchar, **PRIMARY KEY** (emplD, deptID));
- **INSERT INTO** Hollywood.NerdMovies (user_uuid, fan) **VALUES** (cfd66ccc-d857-4e90-b1e5-df98a3d40cd6, 'johndoe')
- **DROP TABLE** worldSeriesAttendees;



Merci pour votre écoute