

tp big data : cassandra

Sylvain Lefebvre

2015-03-12 jeu.

Contents

1	Cassandra	1
1.1	Installation	1
1.1.1	Via moodle	1
1.1.2	Ou bien	2
1.2	Configuration de cassandra	2
1.2.1	création de dossiers:	2
1.2.2	ouvrir le dossier décompressé dcs-cassandra	2
1.2.3	lancer cassandra	3
1.3	Modélisation et création de la base	3
1.3.1	Création d’une base de données	4
1.3.2	Le Cassandra Query Language	5
1.3.3	Clef composite	5
1.4	Requêtes	6
1.4.1	Analyse pour concevoir une base cassandra	6

1 Cassandra

1.1 Installation

1.1.1 Via moodle

- Connectez vous sur le cours de Big Data (module IX.2415)
- Téléchargez le fichier “bundle cassandra + spark”

1.1.2 Ou bien

Télécharger Cassandra:

- Pour linux: [<http://planetcassandra.org/cassandra/?dlink=http://downloads.datastax.com/comm/cassandra-2.0.12-bin.tar.gz>]
- Pour windows: [http://planetcassandra.org/cassandra/?dlink=http://downloads.datastax.com/c/community-32bit_2.0.12.msi]

1.2 Configuration de cassandra

1.2.1 création de dossiers:

- créer un dossier cassandra

1.2.2 ouvrir le dossier décompressé dcs-cassandra

Dans le dossier conf: ouvrir le fichier cassandra.yaml, modifier les lignes ci-dessous comme suit:

```
# Directories where Cassandra should store data on disk.  Cassandra
# will spread data evenly across them, subject to the granularity of
# the configured compaction strategy.
data_file_directories:
  - cassandra/data

# commit log
commitlog_directory: cassandra/commitlog
....

....
# saved caches
saved_caches_directory: cassandra/saved_caches
```

Dans le meme dossier ouvrir le fichier log4j-server.properties, et modifiez la ligne suivante:

```
# Edit the next line to point to your logs directory
log4j.appender.R.File=cassandra/system.log
```

1.2.3 lancer cassandra

```
$ bin/cassandra -f
```

```
...
```

```
...
```

```
Listening for thrift clients...
```

```
$ bin/cqlsh
```

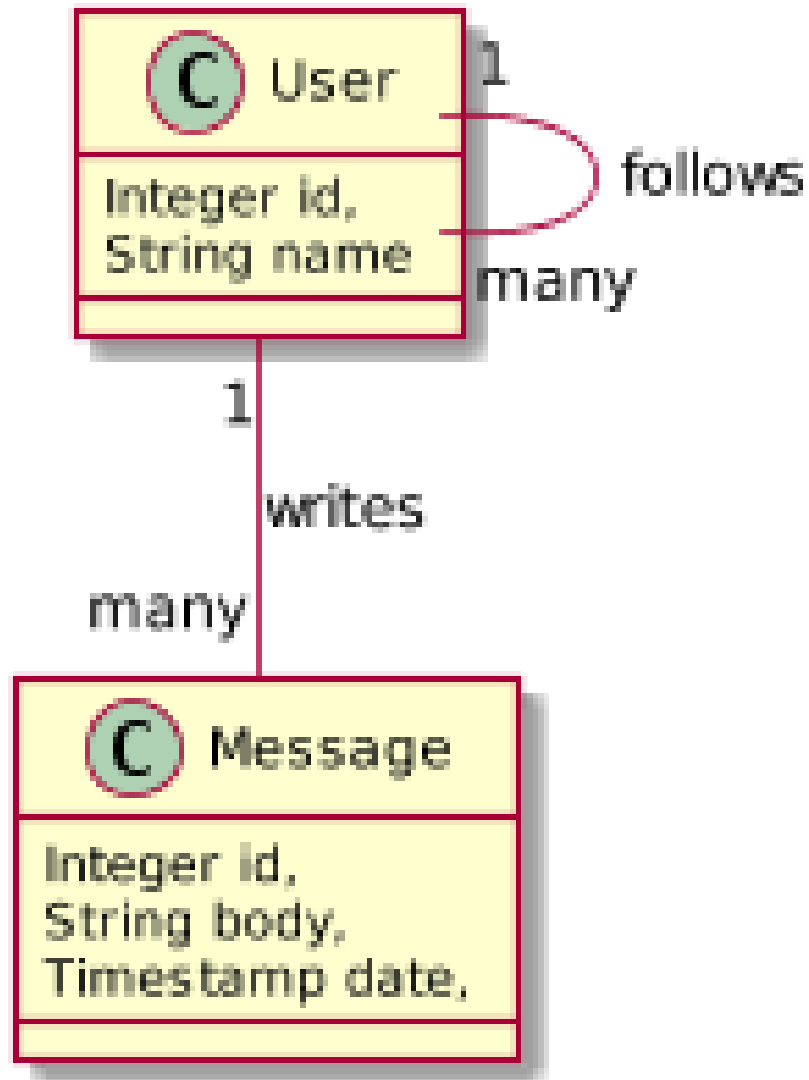
OK maintenant nous avons une base de donnée cassandra qui fonctionne !

1.3 Modélisation et création de la base

Les modèles de données à utiliser avec Cassandra sont **dénormalisés**. C'est la nature *distribuée* et *performante* de cassandra qui nécessitent cette dénormalisation. Ainsi, il n'est plus possible de faire de jointures, ou d'aggregats. De plus, il n'est pas possible de filtrer les données de la même manière que dans une base de données html. Nous allons explorer les méthodes de transformation et de requêtages usuelles de cassandra dans l'exemple qui suit.

- <https://www.hakkalabs.co/articles/cassandra-data-modeling-guide>

Le principe général de la modélisation des bases de données Cassandra est de profiter de la haute performance du système en écriture afin de simplifier au maximum les requêtes en lecture. Ainsi la base de données suivante:



1.3.1 Création d'une base de données

Nous allons créer cette base de données de la manière traditionnelle. Tout d'abord cassandra fonctionne par espace de clefs, nous allons donc créer l'espace de clefs TP:

```
CREATE KEYSPACE TP
WITH REPLICATION = {
```

```
'class' : 'SimpleStrategy',  
'replication_factor' : '1' };
```

```
use TP;
```

1.3.2 Le Cassandra Query Language

CQL permet d'avoir une syntaxe proche du SQL pour l'utilisation de cassandra. Cependant cette simplification masque une grande part de la complexité des table sous-jacentes.

Référence

Créons donc nos tables:

```
create table user ( user_id bigint PRIMARY KEY, user_name varchar);  
create table message ( msg_id bigint PRIMARY KEY, writer_id bigint, body text);  
create table follows (follower_id bigint PRIMARY KEY, followed_id bigint);
```

Ajoutons y des utilisateurs et des messages:

```
insert into user (user_id, user_name) values (1,"sylvain");  
insert into user (user_id, user_name) values (2,"georges");  
  
insert into follows (2,1);  
insert into message (msg_id,writer_id,body)  
values (1,1,"hello world");
```

1.3.3 Clef composite

- Insérer l'utilisateur "Patrick" dans la base de données et faite le suivre par Georges. Afficher la liste des gens suivis par l'utilisateur 2. Que s'est-il passé ?
- Supprimez la table "followers " (drop table) et recréez la comme suit:

```
drop table follows;  
create table follows (  
  follower_id bigint,  
  followed_id bigint,  
  PRIMARY KEY (follower_id, followed_id) );
```

- Rejouez les requêtes permettant de faire suivre sylvain et patrick à georges. Que constatez vous ?

1.4 Requêtes

- Ecrire la requete qui affiche les messages qui seront vus par l'utilisateur n° 2 (Georges) à sa connexion.
- Quelle est votre conclusion ?

1.4.1 Analyse pour concevoir une base cassandra

Lister les vues nécessaires au fonctionnement du mini twitter.

- définissez une table par vue, de sorte qu'aucune requête ne nécessite de jointure.