Le SGBD Cassandra

Présentation de SGBD NoSQL

PESTELLE Nicolas / SPECQ Corentin

Université du Littoral Côte d'Opale

16 octobre 2018

Historique de Cassandra

- Développé en interne par Facebook pour sa messagerie.
- Offert à la fondation Apache en 2008.
- Devenu un projet à part entière en 2010.

Utilisation

Aujourd'hui utilisé par Netflix, Spotify, eBay, Twitter.

Architecture

- Collection de nœuds indépendants formant un cluster.
- Pas de nœud maitre.
- Utilise un protocole "Gossip" pour l'état et la localisation des autres nœuds.

Cas d'utilisation

Problématique d'accès importants aux données, rapidité, accessibilité.

Architecture des données

- Base de donnée orientée colonnes de type comparator.
- Une colonne = nom, valeur (optionnelle ou collection) et timestamp.
- Les colonnes sont regroupées en lignes identifiées par une clé.
- Les lignes sont regroupées en column family (tables en relationnel).
- Les familles de colonnes sont regroupées en Keyspaces (schémas en relationnel).

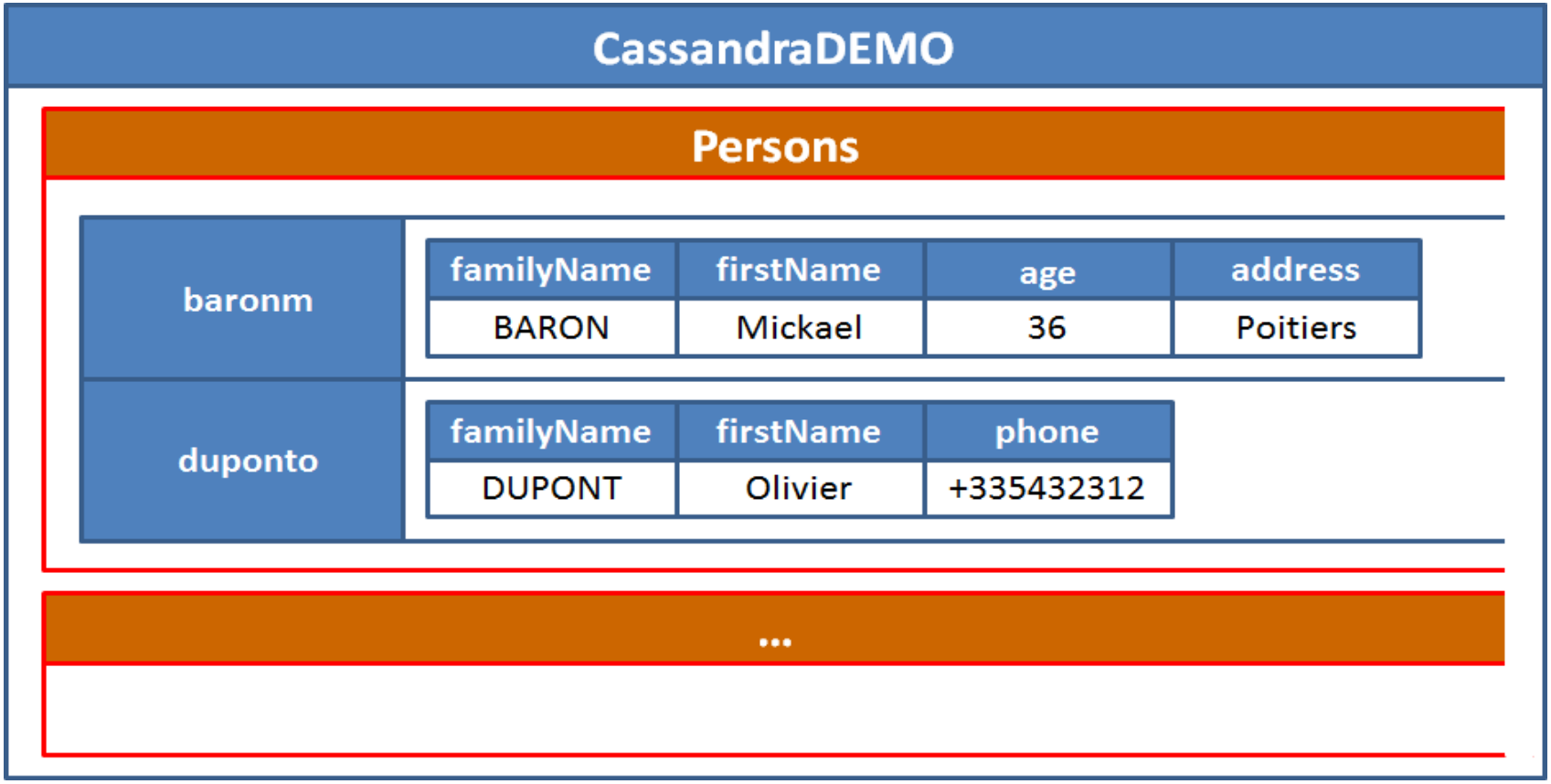
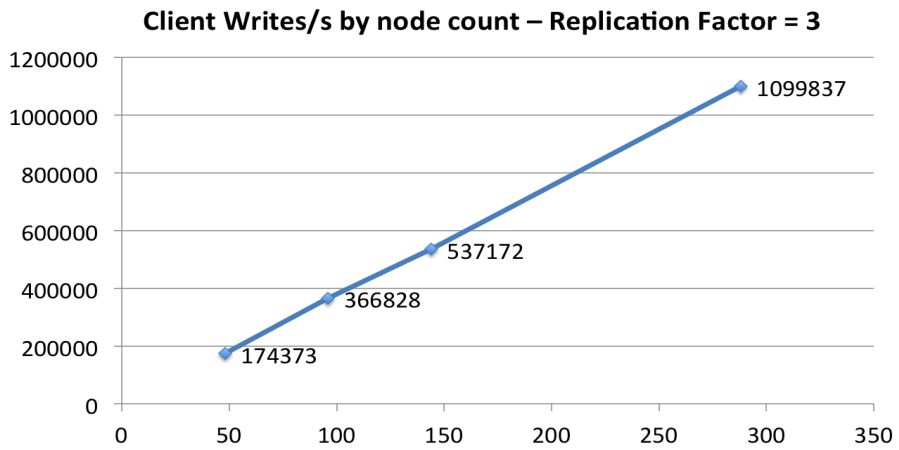


Figure: Représentation d'un Keyspace

Caractéristiques importantes

- Tolérance aux pannes de par la réplication des données.
- Architecture décentralisée sans noeud maitre.
- Modèle de données riche permettant de nombreux cas d'utilisation.
- Haute disponibilité, écriture des données très rapide.

Scale-Up Linearity



Le CQL (Cassandra Query Language)

- Langage de requête spécifique à Cassandra
- Implémentations existantes en Java (JDBC), Python (DBAPI2), Node.js (Helenus), Go (gocql) et C++.

```
CREATE KEYSPACE MonEspaceDeCle
  WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 3 };

USE MonEspaceDeCle;

CREATE COLUMNFAMILY MesColonnes (id text, Nom text, Prenom text, PRIMARY KEY(id));

INSERT INTO MesColonnes (id, Nom, Prenom) VALUES ('1', 'Doe', 'John');

SELECT * FROM MesColonnes;
```

```
 id | nom | prenom
----+-----+--------
  1 | Doe |   John
```

Avantages de Cassandra

- Très rapide pour manipuler un volume important de données.
- Schémas de données flexible
- Possibilité de mettre en cluster plusieurs serveurs Cassandra.
- Réplication des données.
- Passage par une table en mémoire afin d'éviter les latences d'écriture sur disque.

Limites de Cassandra

La limitation principale est sur la taille des colonnes : toutes les données pour une valeur de clé, doivent tenir sur le disque d'une seule machine.