Hadoop & Python (Data Sciences)

Sqoop

Par Houda DAKI

Introduction Sqoop

APACHE SQOOP

Pourquoi Sqoop ?

En big data, Hadoop est le framework le plus utilisé pour gérer et analyser des données. Cependant, il n'est pas fait pour se connecter directement à des données stockées dans des bases SQL.

En effet, Hadoop a été conçu pour utiliser d'autres technologies de stockage, telles que HDFS ou Hive. Or, la base SQL reste la solution de stockage la plus répandue…

cet outil permettant une cohabitation des bases de données (Oracle, mysql…) avec la plateforme Hadoop (Le nom Sqoop est un mot valise constitué de sql et de hadoop)

Sqoop permet d'exporter des données depuis la base de données et de procéder aux traitements en exploitant le cluster Hadoop.

APACHE SQOOP

Pourquoi Sqoop ?

Sqoop est l'outil qui permet de ne pas faire de compromis entre des capacités d'analyses en big data et l'utilisation de bases de données SQL.

En effet, Sqoop permet de relier un cluster Hadoop à une (ou plusieurs) base(s) de données SQL, en vue de transférer des données de l'un à l'autre, dans les deux sens.

il est possible d'exporter le résultat d'un traitement vers une base de données tierce afin qu'il soit exploité par une application (à des fins de restitution par exemple).

Sqoop a été conçu avec comme objectif principal d'assurer des performances élevées pour ces opérations d'import ou d'export massifs.

Sqoop prend des données à la source et les écrit dans une destination !!

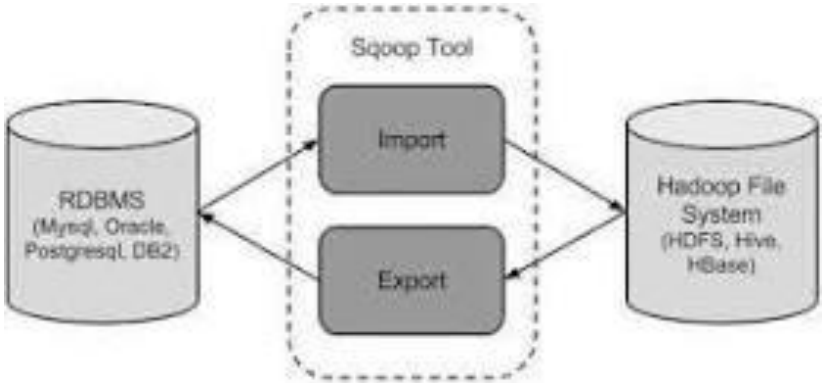
Présentation de Sqoop

APACHE SQOOP

Fonctionnement de Sqoop ?

Rentrons un peu plus dans le concret. En pratique, Sqoop se présente sous la forme d'une boîte à outils en ligne de commandes, codée en Java et accessible avec la commande sqoop.

Comme nous l'avons dit, Sqoop permet de transférer des données ; on retrouve donc deux outils principaux : **sqoop-import et sqoop-export**.



APACHE SQOOP

Import avec Sqoop

La commande **sqoop-import** permet d'importer des données **d'une base de données SQL vers Hadoop**. Pour utiliser cette commande, vous devez spécifier la base de données SQL d'où les données seront importées.

Sqoop supporte les technologies SQL les plus populaires (mysql, Oracle, PostgreSQL…), mais vous pouvez aussi spécifier un driver particulier via le paramètre — driver.

Sqoop faisant partie d'un framework distribué, plusieurs processus s'exécutent en parallèle pour un même import, rendant le processus très efficace.

APACHE SQOOP

Import avec Sqoop

Sqoop propose plusieurs paramètres d'import, notamment :

```
sqoop import --query 'select Id,Message from TestTable where $CONDITIONS'
             --where 'id>100'
             --connect "jdbc:sqlserver://192.168.1.100:1433;database=Test_db
             --username user
             --password password
             --split-by id
             --target-dir /user/test/
             --fields-terminated-by '\t'
```

Par défaut, Sqoop importe vos données vers un stockage HDFS, mais vous pouvez aussi utiliser d'autres technologies big data comme destination, notamment HBase ou Hive.

```
sqoop import \
        --query 'select emp_id, emp_name, emp_sal from employee where $CONDITIONS' \
        --connect "jdbc:sqlserver://192.168.1.99:1433;database=test_db" \
        --username username \
        --password password \
        --hbase-create-table \
        --hbase-table employee_table \
        --hbase-row-key emp_id
```

APACHE SQOOP

Export avec Sqoop

L'outil d'exportation exporte un ensemble de fichiers de HDFS vers un SGBDR avec la commande **sqoop-export**.De la même manière, vous spécifiez une base de données SQL cible.

À noter que la table dans laquelle vous souhaitez exporter les données doit déjà exister. Vous pouvez insérer les données (sql INSERT) ou bien mettre à jour des lignes existantes (sql UPDATE).

À l'instar de l'import, de nombreux paramètres d'export existent et l'export est parallélisé pour gagner en efficacité.

```
sqoop export \
--connect="jdbc:<databaseconnector>" \
--username=<username> \
--password=<password> \
--export-dir=<hdfs export directory> \
--table=<tablename>
```

APACHE SQOOP

Plusieurs autres commandes Sqoop utiles

sqoop-list-databases et sqoop-list-tables pour lister les schémas des bases de données connectées au serveur ainsi que leurs tables.

```
sqoop list-tables --connect "jdbc:sqlserver://<server_ip>:1433;database=<database_name>"
                  --username <user_name>
                  --password <password>
```

Import-All-Table permet d'importer toutes les tables de la base de données spécifiée.

```
sqoop import-all-tables \
--connect <rdbms-jdbc-url> \
--username <username> \
--password <password> \
--hive-import \
--create-hive-table \
--hive-database <dbname> \
--warehouse-dir <warehouse-dir>
```

APACHE SQOOP

Plusieurs autres commandes Sqoop utiles

Eval est utilisé pour exécuter des requêtes SQL définies par l'utilisateur sur une base de données et imprimer les résultats sur la console afin qu'un utilisateur puisse examiner le contenu de la table avant d'effectuer l'opération d'importation.

```
sqoop eval --connect jdbc:mysql://localhost/userdata?serverTimezone=UTC --username root
--password cloudduggu --query "select * from employee«
```

```
sqoop eval --connect jdbc:mysql://localhost/userdata?serverTimezone=UTC --username root
 --password cloudduggu --query "update employee set empname='cloudduggu' where empid=1010;"
```