Connecting to different databases using SQOOP

Author: Shivshankar Chandankhede

References : https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html#_supported_databases

https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781849519120/1/ch01lvl1sec13/configuring-sqoop-for-microsoft-sql-server

https://www.mssqltips.com/sqlservertip/4426/use-sqoop-to-load-data-from-a-sql-server-table-to-a-hadoop-distributed-file-system/

We have many articles regarding Sqoop which tells us how to connect to Relational Database, But they only connect to MySQL DB and the rest of the Databases remains Ignored.

I am creating this Knowledge Byte(KB) Article to list down different RDBMS and pre-requisites that needs to be satisfied before connecting to the them using SQOOP

Most commonly used Relational DB's in the industry are -

1) Oracle DB
2) PostgreSQL
3) MSSQL
4) MySQL

Sqoop is designed to import tables from a database into HDFS.

To do so, you must specify a connect string that describes how to connect to the database.

The connect string is similar to a URL, and is communicated to Sqoop with the `--connect` argument. This describes the server and database to connect to and you have to specify the port.

You will need to authenticate against the database before you can access it. You can use the `--username` and `--password` or `-P` parameters to supply a username and a password to the database.

Sqoop automatically supports several databases, including MySQL. Connect strings beginning with `jdbc:mysql://` are handled automatically in Sqoop.

Even if Sqoop supports a database internally, you may still need to install the database vendor's JDBC driver in your $SQOOP_HOME/lib path on your client. Sqoop can load classes from any jars in `$SQOOP_HOME/lib on` the client and will use them as part of any MapReduce jobs it runs; unlike older versions, you no longer need to install JDBC jars in the Hadoop library path on your servers.

First, download the appropriate JDBC driver for the type of database you want to import, and install the .jar file in the $SQOOP_HOME/lib directory on your client machine. (This will be /usr/lib/sqoop/lib if you installed from an RPM or Debian package.)

Each driver .jar file also has a specific driver class which defines the entry-point to the driver. For example, MySQL's Connector/J library has a driver class of `com.mysql.jdbc.Driver`. Refer to your database vendor-specific documentation to determine the main driver class. This class must be provided as an argument to Sqoop with --driver.

While JDBC is a compatibility layer that allows a program to access many different databases through a common API, slight differences in the SQL language spoken by each database may mean that Sqoop can't use every database out of the box, or that some databases may be used in an inefficient manner.

When you provide a connect string to Sqoop, it inspects the protocol scheme to determine appropriate vendor-specific logic to use.

If Sqoop knows about a given database, it will work automatically.

If not, you may need to specify the driver class to load via --driver. This will use a generic code path which will use standard SQL to access the database.

Sqoop provides some databases with faster, non-JDBC-based access mechanisms. These can be enabled by specfying the --direct parameter.

1) Oracle DB

Download Oracle JDBC driver from the link mentioned Below and install it in your Sqoop lib path.:
https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html

```
sqoop import \
--connect jdbc:oracle:thin:@database:1521/orcl \
--username MOVIEDEMO \
--password welcome1 \
--table ACTIVITY
```

2) PostgreSQL DB

Link to download the JDBC driver is mentioned below and install it in your Sqoop lib path.:
https://jdbc.postgresql.org/download.html

The connector has been tested using JDBC driver version "9.1-903 JDBC 4" with PostgreSQL server 9.1.

```
sqoop list-databases \
--connect jdbc:postgresql://database.example.com/employees \
--username aaron \
--password 12345
```

3) MSSQL DB

To connect to a SQLServer database, first download the driver from microsoft.com and install it in your Sqoop lib path.

Link to download the JDBC driver is mentioned below:
https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15

```
sqoop import \
--connect jdbc:sqlserver://database:1433/test_db \
--username sqoopLogin
--password password
--table DimCustomer
```

4) MySQL DB

To connect to the MySQL server, you can directly start with writing the sqoop statements. The JDBC connector is already present in the $SQOOP_HOME/lib directory

For example:

```
$ sqoop list-databases \
--connect jdbc:mysql://database.example.com/employees \
--username aaron
--password 12345
```