



**EBook Gratis**

# APRENDIZAJE

---

## sqoop

Free unaffiliated eBook created from  
**Stack Overflow contributors.**

**#sqoop**

# Tabla de contenido

Acerca de.....	1
<b>Capítulo 1: Empezando con sqoop.....</b>	<b>2</b>
Observaciones.....	2
Examples.....	2
Instalación o configuración.....	2
Instalar dependencias para SQOOP.....	2
Iniciar y detener los servicios del servidor Sqoop.....	3
Pasos de configuración del cliente Sqoop.....	3
<b>Capítulo 2: Conexión de Sqoop a otras bases de datos / almacenes de datos.....</b>	<b>4</b>
Introducción.....	4
Examples.....	4
Cargar controlador JDBC.....	4
Validar la conexión.....	4
Importar tabla a nuevo catálogo.....	4
Importe los resultados de una consulta desde una base de datos relacional a HDFS:.....	5
Importar datos directamente en Hive Warehouse.....	5
Importar datos desde RDBMS a la tabla HBase.....	5
<b>Capítulo 3: Exportación De Sqoop.....</b>	<b>7</b>
Examples.....	7
Ejemplo básico de Sqoop Export.....	7
<b>Capítulo 4: fusionar conjuntos de datos importados mediante importación incremental utiliz.....</b>	<b>8</b>
Observaciones.....	8
Examples.....	8
Importar nuevos datos - modo de añadir.....	8
Importar datos nuevos y actualizados, modo Lastmodified.....	8
<b>Capítulo 5: Importación de Sqoop.....</b>	<b>10</b>
Sintaxis.....	10
Observaciones.....	10
Examples.....	10
Importar tabla RDBMS a HDFS.....	10

Importar a un directorio particular en HDFS.....	11
Especifique el directorio HDFS principal para el trabajo de Sqoop.....	11
Importar subconjunto de tabla RDBMS a HDFS.....	11
Usando la etiqueta - --where.....	11
Uso de consulta de forma libre.....	12
Importar todas las tablas.....	12
<b>Creditos.....</b>	<b>14</b>

---

## Acerca de

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: [sqoop](#)

It is an unofficial and free sqoop ebook created for educational purposes. All the content is extracted from [Stack Overflow Documentation](#), which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official sqoop.

The content is released under Creative Commons BY-SA, and the list of contributors to each chapter are provided in the credits section at the end of this book. Images may be copyright of their respective owners unless otherwise specified. All trademarks and registered trademarks are the property of their respective company owners.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to [info@zzzprojects.com](mailto:info@zzzprojects.com)

---

# Capítulo 1: Empezando con sqoop

## Observaciones

Los archivos de configuración del servidor SQOOP se almacenan en el directorio `server / config` del artefacto distribuido junto con otros archivos de configuración de Tomcat (para alojar el servidor SQOOP).

El archivo `sqoop_bootstrap.properties` especifica qué proveedor de configuración se debe usar para cargar la configuración para el resto del servidor Sqoop. El valor predeterminado `PropertiesConfigurationProvider` debería ser suficiente.

El segundo archivo de configuración `sqoop.properties` contiene las propiedades de configuración restantes que pueden afectar al servidor Sqoop. El archivo está muy bien documentado, así que verifique si todas las propiedades de configuración se ajustan a su entorno. El ajuste por defecto o muy poco debe ser suficiente para la mayoría de los casos comunes.

## Examples

### Instalación o configuración

Sqoop se distribuye como un paquete binario, sin embargo, está compuesto de dos partes separadas de cliente y servidor. Necesita instalar el servidor en un solo nodo en su clúster. Este nodo servirá como un punto de entrada para todos los clientes Sqoop que se conectan. El servidor actúa como un cliente mapreduce y, por lo tanto, Hadoop debe instalarse y configurarse en la máquina que aloja el servidor Sqoop. Los clientes pueden instalarse en cualquier número arbitrario de máquinas. El cliente no actúa como un cliente mapreduce y, por lo tanto, no necesita instalar Hadoop en nodos que actuarán solo como un cliente Sqoop.

Copie el artefacto Sqoop en la máquina donde desea ejecutar el servidor Sqoop. Esta máquina debe tener instalado y configurado Hadoop. No es necesario ejecutar ningún servicio relacionado con Hadoop, sin embargo, la máquina debe poder actuar como un cliente Hadoop.

```
# Extract Sqoop tar
tar -xf sqoop-<version>-bin-hadoop<hadoop-version>.tar.gz

# Move decompressed content to any location
  (you can also setup soft links to sqoop directory)
mv sqoop-<version>-bin-hadoop<hadoop version>.tar.gz /opt/apache/sqoop

# Change working directory
cd /opt/apache/sqoop
```

## Instalar dependencias para SQOOP

Debe instalar las bibliotecas de Hadoop en el archivo de guerra del servidor Sqoop. Sqoop

proporciona el script de conveniencia `addtoward.sh` para hacerlo.

Si ha instalado Hadoop en la ubicación habitual en `/usr/lib` y `hadoop` ejecutable está en su ruta, puede usar el procedimiento de instalación automática de Hadoop:

```
./bin/addtoward.sh -hadoop-auto
```

En caso de que tenga Hadoop instalado en una ubicación diferente, deberá especificar manualmente la versión de Hadoop y la ruta a las bibliotecas de Hadoop. Puede usar el parámetro `-hadoop-version` para especificar la versión principal de Hadoop,

```
./bin/addtoward.sh -hadoop-version 2.0 -hadoop-path /usr/lib/hadoop-common:/usr/lib/hadoop-hdfs:/usr/lib/hadoop-yarn
```

- Se instalaron los tarros JDBC necesarios para que `sqoop` se conecte a la base de datos

```
./bin/addtoward.sh -jars /path/to/jar/mysql-connector-java-*-bin.jar
```

## Iniciar y detener los servicios del servidor Sqoop

```
./bin/sqoop.sh server start
./bin/sqoop.sh server stop
```

## Pasos de configuración del cliente Sqoop

Copie el artefacto de distribución de Sqoop en la máquina de destino y descomprímalo en la ubicación deseada. Puede iniciar el cliente con el siguiente comando:

```
bin/sqoop.sh client
```

El cliente de Sqoop 2 tiene la capacidad de cargar archivos de recursos de manera similar a otras herramientas de línea de comandos. Al comienzo de la ejecución, el cliente Sqoop verificará la existencia del archivo `.sqoop2rc` en el directorio de inicio del usuario registrado actualmente. Si tal archivo existe, será interpretado antes de cualquier acción adicional. Este archivo se carga en modo interactivo y por lotes. Se puede utilizar para ejecutar cualquier comando compatible con lotes.

Ejemplo de archivo de recursos:

```
# Configure our Sqoop 2 server automatically
set server --host sqoop2.company.net

# Run in verbose mode by default
set option --name verbose --value true
```

Lea [Empezando con sqoop en línea](https://riptutorial.com/es/sqoop/topic/1050/empezando-con-sqoop): <https://riptutorial.com/es/sqoop/topic/1050/empezando-con-sqoop>

---

# Capítulo 2: Conexión de Sqoop a otras bases de datos / almacenes de datos

## Introducción

Muestra cómo se podría usar un script de sqoop para importar datos de varios almacenes de datos / bases de datos.

## Examples

### Cargar controlador JDBC

Para acceder a la base de datos de MS SQL Server, Sqoop requiere un controlador JDBC adicional que se puede descargar de Microsoft. Los siguientes pasos instalarán el controlador JDBC del servidor MSSQL en Sqoop:

```
wget 'http://download.microsoft.com/download/0/2/A/02AAE597-3865-456C-AE7F-613F99F850A8/sqljdbc_4.0.2206.100_enu.tar.gz'

tar -xvzf sqljdbc_4

cp sqljdbc_4.0/enu/sqljdbc4.jar /usr/hdp/current/sqoop-server/lib/
```

### Validar la conexión

Para comprobar que la conexión al servidor es válida:

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

Antes de hacer esto, se recomienda verificar si la configuración de SqlServer permite el acceso remoto al puerto 1433.

Abra **Configuration Manager => Configuración de red de SQL Server => Protocolos para MSSQLSERVER** y verifique que el Protocolo esté habilitado y que la IP y el puerto necesarios estén habilitados y estén activos.

### Importar tabla a nuevo catálogo

Para importar datos de SQL Server a Hadoop:

```
sqoop import --table TestTable
--connect "jdbc:sqlserver://192.168.1.100:1433;database=Test_db"
--username user
--password password
```

```
--split-by id
--target-dir /user/test
```

- dividido por - usado obligatorio si no hay clave primaria
- target-dir - nuevo catálogo, que aún no existe

## Importe los resultados de una consulta desde una base de datos relacional a HDFS:

Se puede usar la consulta en lugar de la tabla en la operación de importación:

```
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'
```

- donde \$ CONDICIONES - obligatorio incluso cuando no exista una condición
- dividido por - obligatorio, especifica la columna para la operación de dividir. Se utiliza para dividir tareas en la importación de trabajo MapReduce

## Importar datos directamente en Hive Warehouse

Los datos se pueden importar directamente en Hive:

```
sqoop import --hive-import
--table EventLog
--connect "jdbc:sqlserver://192.168.1.99:1433;database=Test_db"
--username user
--password password
--split-by id
--fields-terminated-by '\t'
```

## Importar datos desde RDBMS a la tabla HBase

El siguiente comando sqoop se usará para importar los datos de la tabla RDBMS a la tabla HBase, si la tabla no existe en HBase, creará una nueva tabla e importará los datos a esta tabla

```
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
```

[Lea Conexión de Sqoop a otras bases de datos / almacenes de datos en línea:](#)

<https://riptutorial.com/es/sqoop/topic/3930/conexion-de-sqoop-a-otras-bases-de-datos---almacenes-de-datos>

---

# Capítulo 3: Exportacion De Sqoop

## Examples

### Ejemplo básico de Sqoop Export

La herramienta de exportación exporta un conjunto de archivos de HDFS a un RDBMS. La tabla de destino ya debe existir en la base de datos. Los archivos de entrada se leen y analizan en un conjunto de registros de acuerdo con los delimitadores especificados por el usuario.

Ejemplo:

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

Lea Exportacion De Sqoop en línea: <https://riptutorial.com/es/sqoop/topic/6999/exportacion-de-sqoop>

---

# Capítulo 4: fusionar conjuntos de datos importados mediante importación incremental utilizando Sqoop

## Observaciones

La importación incremental de Sqoop entra en escena debido a un fenómeno llamado CDC, es decir, la **captura de datos de cambio**. Ahora que es CDC?

CDC es un patrón de diseño que captura cambios de datos individuales en lugar de tratar con todos los datos. En lugar de descargar nuestra base de datos completa, utilizando CDC, podríamos capturar solo los cambios de datos realizados en la base de datos maestra.

Por ejemplo: si estamos lidiando con un problema de datos, digamos, 1 lakh de entradas de datos que ingresan diariamente al RDBMS y tenemos que obtener estos datos en Hadoop diariamente, entonces solo queremos obtener los datos recién agregados, como importación. La información diaria de RDBMS completa para Hadoop será una sobrecarga y también retrasará la disponibilidad de los datos. Para una explicación detallada vaya a través de este [enlace](#).

## Examples

### Importar nuevos datos - modo de añadir

Si solo está agregando nuevas filas en su RDBMS ( *no está actualizando los datos existentes* )

Necesita dos parámetros adicionales:

- `--check-column` : un nombre de columna que debe verificarse para los datos recién agregados. `Integer` sería un tipo de datos adecuado para esta columna.
- `--last-value` : el último valor que se importó con éxito a Hadoop. Todos los datos recién agregados después de este valor serán importados.

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username sqoop \  
--password sqoop \  
--table employee \  
--incremental append \  
--check-column id \  
--last-value 100
```

### Importar datos nuevos y actualizados, modo Lastmodified

Si está agregando nuevas filas y actualizando datos existentes.

Necesita dos parámetros adicionales:

- `--check-column` : un nombre de columna que debe verificarse para los datos recién agregados y actualizados. `date` , `time` , `datetime` y `timestamp` son tipos de datos adecuados para esta columna.
- `--last-value` : el último valor que se importó con éxito a Hadoop. Todos los datos recién agregados y actualizados después de este valor serán importados.

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username sqoop \  
--password sqoop \  
--table employee \  
--incremental lastmodified \  
--check-column last_update_date \  
--last-value "2015-10-20 06:00:01"
```

Lea fusionar conjuntos de datos importados mediante importación incremental utilizando Sqoop en línea: <https://riptutorial.com/es/sqoop/topic/5673/fusionar-conjuntos-de-datos-importados-mediante-importacion-incremental-utilizando-sqoop>

---

# Capítulo 5: Importación de Sqoop

## Sintaxis

- `<rdbms-jdbc-url>` // URL JDBC de RDBMS
- `<username>` // Nombre de usuario de la base de datos RDBMS
- `<password>` // Contraseña de la base de datos RDBMS
- `<table-name>` // tabla de base de datos RDBMS
- `<hdfs-home-dir>` // directorio de inicio HDFS
- `<condition>` // Condición que puede expresarse en forma de una consulta SQL con una cláusula WHERE.
- `<sql-query>` // Consulta SQL
- `<target-dir>` // Directorio de destino HDFS

## Observaciones

Sqoop es una herramienta de línea de comandos de Hadoop que importa una tabla desde una fuente de datos RDBMS a HDFS y viceversa. Genera una clase Java que nos permite interactuar con los datos importados. Cada fila de una tabla se guarda como un registro separado en HDFS. Los registros se pueden almacenar como archivos de texto o en representación binaria como Avro o Archivos de secuencia. Hay 2 versiones de sqoop:

Sqoop1 y Sqoop2

Sqoop1 es la herramienta ampliamente aceptada y se recomienda para entornos de producción. Encuentre la comparación entre Sqoop1 y Sqoop2 como se indica en [el sitio web de Cloudera](#) .

## Examples

### Importar tabla RDBMS a HDFS

```
sqoop import \  
--connect <rdbms-jdbc-url> \  
--username <username> \  
--password <password> \  
--table <table-name>
```

#### Ejemplo con Mysql:

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username root \  
--password root \  
--table employees
```

El archivo CSV con los datos importados se creará en el directorio de **empleados** en el directorio

principal.

Inspeccionar utilizando el comando:

```
hadoop fs -cat <hdfs-home-dir>/employees/part-m-*
```

## Importar a un directorio particular en HDFS

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username root \  
--password root \  
--table employees \  
--target-dir /dev/data/employees
```

Esto generará un archivo CSV en el directorio `/dev/data/employees`.

## Especifique el directorio HDFS principal para el trabajo de Sqoop

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username root \  
--password root \  
--table employees \  
--warehouse-dir /dev/warehouse/
```

`--warehouse-dir` tag en el comando anterior cambiará su directorio de inicio a `/dev/warehouse/`

## Importar subconjunto de tabla RDBMS a HDFS

### Usando la etiqueta - `--where`

```
sqoop import \  
--connect <rdbms-jdbc-url> \  
--username <username> \  
--password <password> \  
--table <table-name> \  
--where "<condition>"
```

Ejemplo con Mysql:

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username root \  
--password root \  
--table employees \  
--where "country = 'INDIA'"
```

Cualquier función especial, conversiones o incluso funciones definidas por el usuario se pueden usar en la cláusula `--where` .

## Uso de consulta de forma libre

```
sqoop import \  
--connect <rdbms-jdbc-url> \  
--username <username> \  
--password <password> \  
--query <sql-query> \  
--target-dir <target-dir>
```

### Ejemplo con Mysql:

```
sqoop import \  
--connect jdbc:mysql://mysql.example.com/testdb \  
--username root \  
--password root \  
--query "SELECT employees.name, \  
        address.city \  
        FROM employees \  
        JOIN address USING(emp_id) \  
        WHERE \$CONDITION" \  
--target-dir /dev/data/employees
```

## Importar todas las tablas

SQOOP proporciona facilidad para importar todas las tablas

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

Puntos importantes a tener en cuenta sobre las diferencias entre la importación y la exportación de todas las tablas:

Debe proporcionar el nombre de la base de datos `--warehouse-dir = // stage.db` como parámetro de entrada para descargar todas las tablas en una base de datos. En la importación de sqoop proporcionaremos solo `--target-dir` no el `--warehouse-dir`

### Ejemplo:

```
sqoop import-all-tables --connect="jdbc:mysql://serverip:3306/dbname" \  
--username=xxx --password=yyy \  
-m 1 --hive-import \  
--hive-overwrite \  
--create-hive-table
```

```
--hive-database dbname  
--hive-home /user/hive/warehouse  
--warehouse-dir=/user/hive/warehouse/retail_stage.db
```

Lea Importación de Sqoop en línea: <https://riptutorial.com/es/sqoop/topic/5057/importacion-de-sqoop>

# Creditos

S. No	Capítulos	Contributors
1	Empezando con sqoop	<a href="#">Community</a> , <a href="#">dev ツ</a> , <a href="#">Venkata Karthik</a>
2	Conexión de Sqoop a otras bases de datos / almacenes de datos	<a href="#">Alex</a> , <a href="#">Ani Menon</a> , <a href="#">Prasad Khode</a>
3	Exportacion De Sqoop	<a href="#">Kannan Kandasamy</a>
4	fusionar conjuntos de datos importados mediante importación incremental utilizando Sqoop	<a href="#">dev ツ</a> , <a href="#">NeoWelkin</a>
5	Importación de Sqoop	<a href="#">dev ツ</a> , <a href="#">Kannan Kandasamy</a> , <a href="#">NeoWelkin</a>