

APACHE SQOOP

- ✓ Why we need Sqoop
- ✓ Introduction to Sqoop
- ✓ How it works
- ✓ Sqoop commands



Why we need Sqoop ?



- Designed to efficiently transfer bulk data between Apache Hadoop and structured datastores such as relational databases
- **Allows data imports** from external datastores and enterprise data warehouses into Hadoop
- **Parallelizes data transfer** for fast performance and optimal system utilization
- **Copies data quickly** from external systems to Hadoop
- **Makes data analysis more efficient**
- **Mitigates excessive loads** to external systems.

What is Apache Sqoop ?



- Sqoop is a tool designed to transfer data between Hadoop and relational databases.
- You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.
- Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported.
- Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

- Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.
- It supports a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import.
- Imports can also be used to populate tables in Hive or HBase.
- Exports can be used to put data from Hadoop into a relational database. Sqoop became a top-level Apache project in March 2012

- Sqoop provides a pluggable connector mechanism for optimal connectivity to external systems.
- The Sqoop extension API provides a convenient framework for building new connectors which can be dropped into Sqoop installations to provide connectivity to various systems.
- Sqoop itself comes bundled with various connectors that can be used for popular database and data warehousing systems.

- **Import the data (MySQL table) to HDFS**

```
$ bin/sqoop import -connect jdbc:mysql://localhost:3306/db1 -      username root -  
password root --table tableName --target-dir /user/anish/tableName
```

- Import data to HIVE

```
sqoop import --connect jdbc:mysql://remote-ip/db --username -password xxx --table tb  
--hive-import --hive-table foo.tb
```

```
e.g. sqoop-import --connect jdbc:mysql://localhost:3306/db1 -      username root -  
password password --table tableName --hive-table      tableName --create-hive-  
table --hive-import --hive-home path/to/hive_home --split-by column_name
```


- **Export to Mysql from HDFS**

```
sqoop export --connect jdbc:mysql://foo.com/db --table bar --  
export-dir/hdfs_path/bar_data
```

Thank You

DataFlair Web Services Pvt Ltd

+91-8451097879 / +91-7718877477

info@data-flair.com

<http://data-flair.com>

<https://www.facebook.com/DataFlairWS>