What is Sqoop?

* Sqoop is a top level Apache project and is a core member of hadoop ecosystem.
* Sqoop can be used 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 using any engine in Hadoop ecosystem like Spark or MapReduce, and then export the results back into an RDBMS.
* Sqoop is a command line tool and uses MapReduce to import and export the data as it supports parallelism as well as fault tolerance.
* Apache Sqoop is compatible with all JDBC compatible databases
* Sqoop is used by thousands of hadoop users and is now an integral part of ETL and/or data transfer pipelines.

What does the name Sqoop imply?

The name Sqoop is derived from **Sq**l + Had**oop**, mainly due to it's "SQL to Hadoop; Hadoop to SQL" feature.

What is the need for Sqoop?

As industry’s focus on SQL-oriented analytics is on rise, systems like Cloudera Impala, Spark SQL and Dremel-style analytic engines are offering powerful distributed analytics on top of HDFS data using familiar with SQL style and syntax. Sqoop enabled all those users with large amounts of data stored in their existing relational stores to use these new analytic engines built on Hadoop & Yarn with simple export and import utilities.

What are the challenges that Sqoop solves in data transfer pipelines?

Sqoop enables data transfer between most RDBMSs (MySQL, Teradata, Oracle, Netezza etc) and HDFS (stored in variety of formats ). Variety in data sources and targets presents a challenge in setting up robust data transfer pipelines.

Transferring data to and from relational databases is challenging and laborious. Because data transfer requires careful handling. Data engineers will have to script data transfer pipelines between different DBMS sources and different targets on Hadoop ecosystem. This poses below challenges –

* Differences in SQL dialects, treatment of various data types, optimal data transfer parameters among DBMS providers.
* Differences in Hadoop targets – HDFS, Hive, Impala etc
* Different applications need different file formats – parquet, sequence file or csv
* Different compression algorithms to apply.

Apache Sqoop, short for “SQL to Hadoop,” was created to perform bidirectional data transfer between Hadoop and almost any external structured datastore. Sqoop provide a compact but powerful tool that negotiates between these systems and their data layouts.Taking advantage of MapReduce, execution engine, Sqoop performs the transfers in a parallel and fault tolerant manner.

What is structure of Sqoop CLI command?

Sqoop command-line interface has the following structure:

sqoop TOOL PROPERTY\_ARGS SQOOP\_ARGS [-- EXTRA\_ARGS]

TOOL indicates the operation that you want to perform. Most import ones are import and export.

List of available tools:

codegen Generate code to interact with database records

create-hive-table Import a table definition into Hive

eval Evaluate a SQL statement and display the results

export Export an HDFS directory to a database table

help List available commands

import Import a table from a database to HDFS

import-all-tables Import tables from a database to HDFS

list-databases List available databases on a server

list-tables List available tables in a database

version Display version information

PROPERTY\_ARGS are a special set of parameters that are entered as Java properties in the format -Dname=value

SQOOP\_ARGS contain all the various Sqoop parameters

EXTRA\_ARGS are only for specialized connectors. EXTRA\_ARGS parameters must be separated from the SQOOP\_ARGS with a –

How do you transfer a table named Sales from Mysql into Hadoop’s Distributed File System (HDFS)?

Importing one entire table is one of the most common and straightforward use cases of Sqoop.

$ sqoop import \

--connect jdbc:mysql://mysql.example.com/sqoop \

--username sqoop \

--password sqoop \

--table cities