sqoop

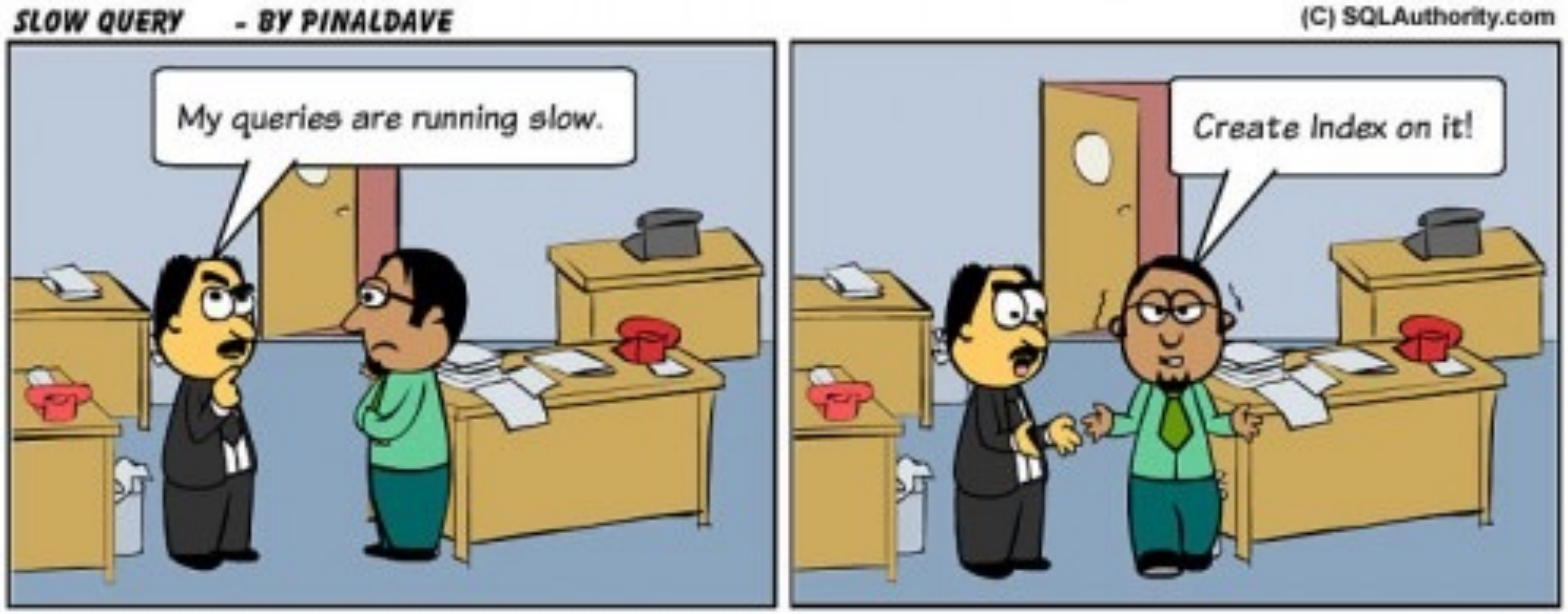
SQOOP

What is SQOOP?

- Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.
- Sqoop imports data from external structured datastores into HDFS or related systems like Hive and HBase.
- Sqoop can also be used to export data from Hadoop and export it to external structured datastores such as relational databases and enterprise data warehouses.
- Sqoop works with relational databases such as: Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB.

SQOOP

Why is SQOOP?



- As more organizations deploy Hadoop to analyse vast streams of information, they may find they need to transfer large amount of data between Hadoop and their existing databases, data warehouses and other data sources
- Loading bulk data into Hadoop from production systems or accessing it from map-reduce applications running on a large cluster is a challenging task since transferring data using scripts is a inefficient and time-consuming task.

SQOOP

Hadoop-Sqoop?

- Hadoop is great for storing massive data in terms of volume using HDFS
- It Provides a scalable processing environment for structured and unstructured data
- But it's Batch-Oriented and thus not suitable for low latency interactive query operations
- Sqoop is basically an ETL Tool used to copy data between HDFS and SQL databases
  - Import SQL data to HDFS for archival or analysis
  - Export HDFS to SQL (e.g : summarized data used in a DW fact table)

SQOOP

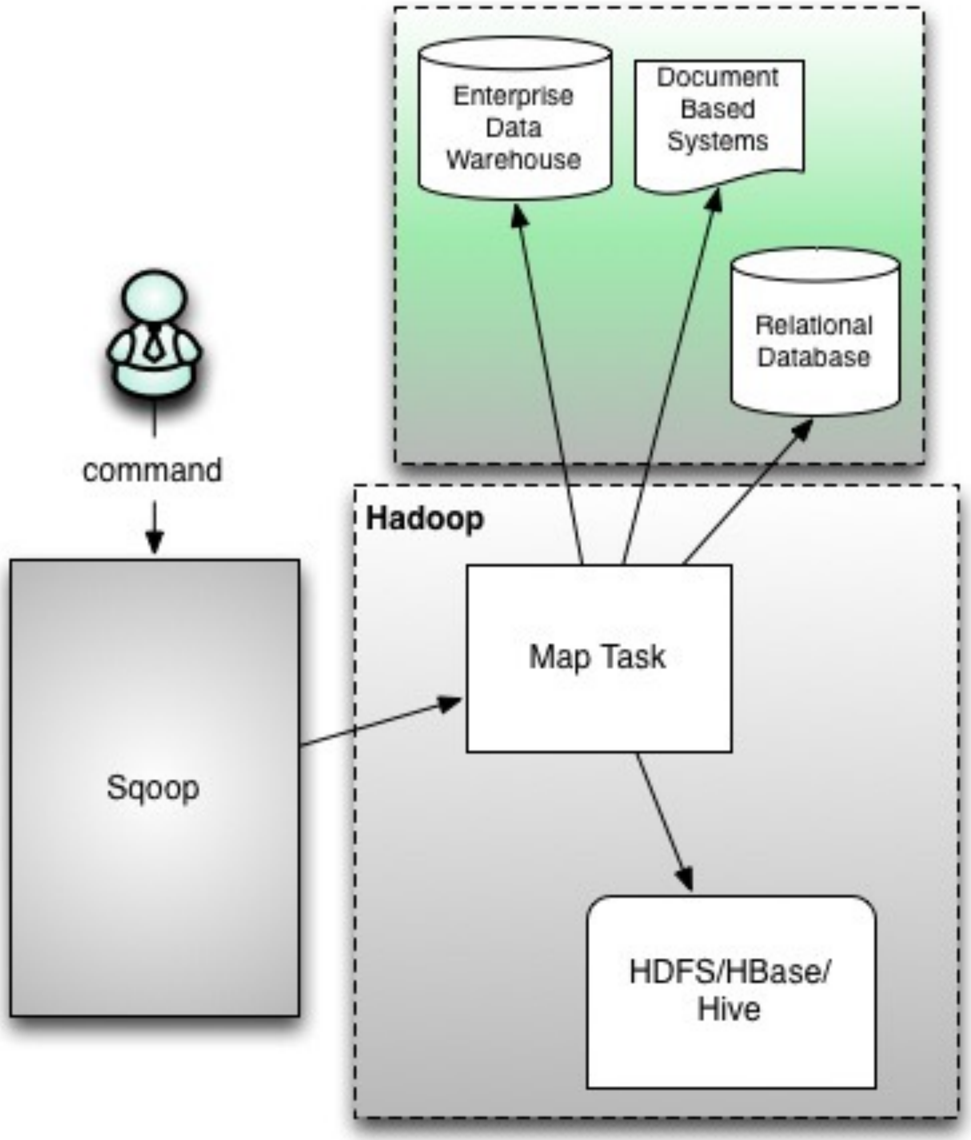
What Sqoop Does?

Designed to efficiently transfer bulk data between Apache Hadoop and structured datastores such as relational databases, Apache Sqoop:

- **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.

SQOOP

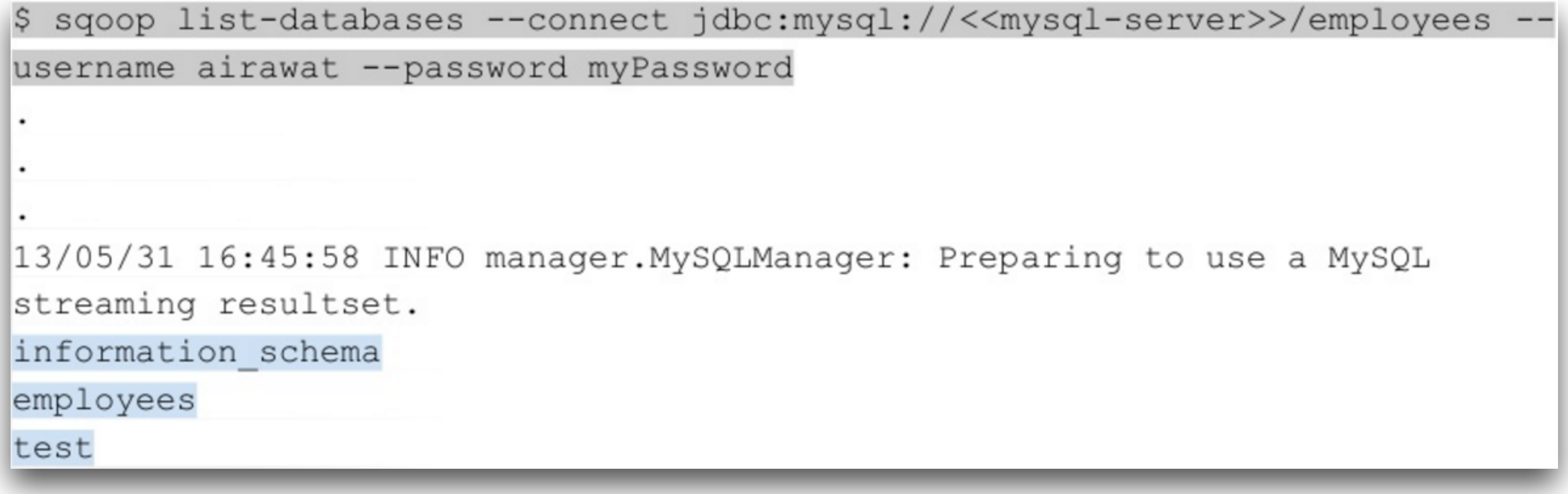
How SQOOP Works?



- 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.

SQOOP

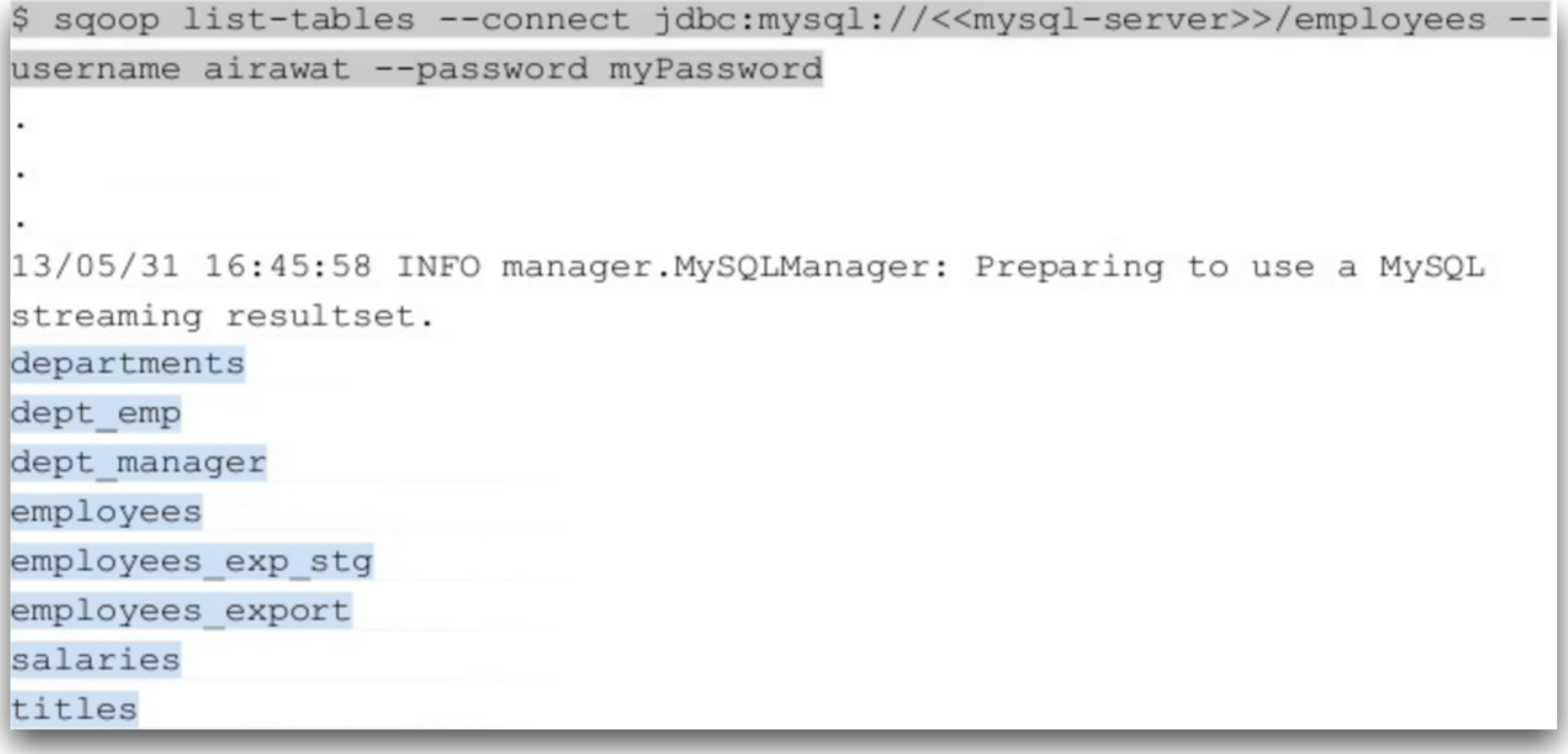
List Databases RDBMS



```
$ sqoop list-databases --connect jdbc:mysql://<<mysql-server>>/employees --
username airawat --password myPassword
.
.
.
13/05/31 16:45:58 INFO manager.MySQLManager: Preparing to use a MySQL
streaming resultset.
information_schema
employees
test
```

SQOOP

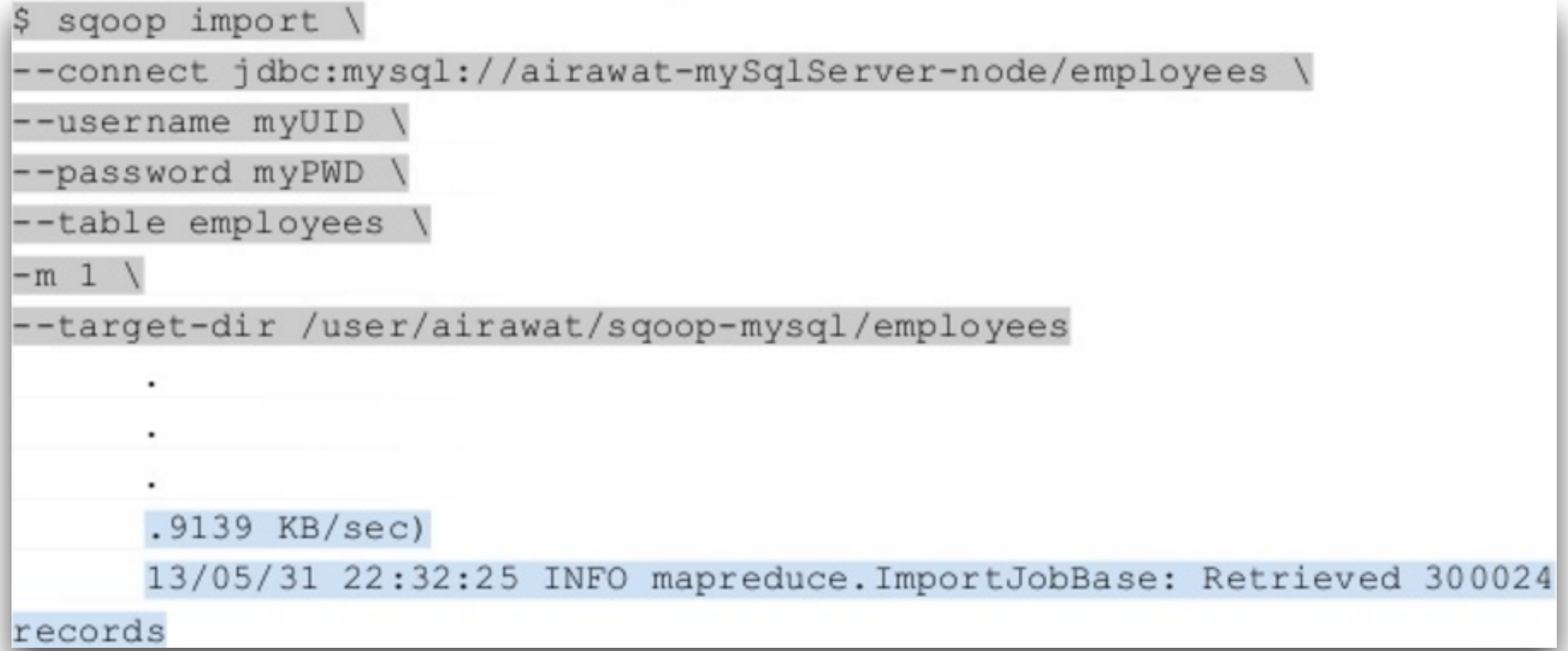
List Tables RDBMS



```
$ sqoop list-tables --connect jdbc:mysql://<<mysql-server>>/employees --
username airawat --password myPassword
.
.
.
13/05/31 16:45:58 INFO manager.MySQLManager: Preparing to use a MySQL
streaming resultset.
departments
dept_emp
dept_manager
employees
employees_exp_stg
employees_export
salaries
titles
```

SQOOP

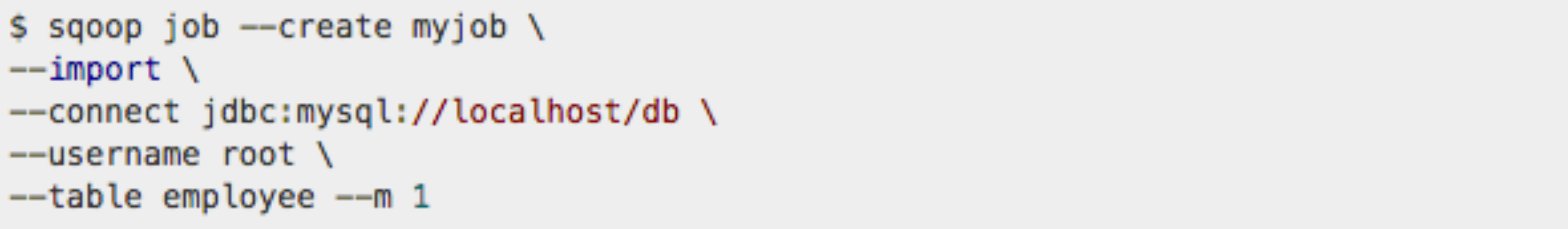
RDBMS To HDFS



```
$ sqoop import \
--connect jdbc:mysql://airawat-mySqlServer-node/employees \
--username myUID \
--password myPWD \
--table employees \
-m 1 \
--target-dir /user/airawat/sqoop-mysql/employees
      .
      .
      .
      .9139 KB/sec)
      13/05/31 22:32:25 INFO mapreduce.ImportJobBase: Retrieved 300024
records
```

SQOOP

JOB

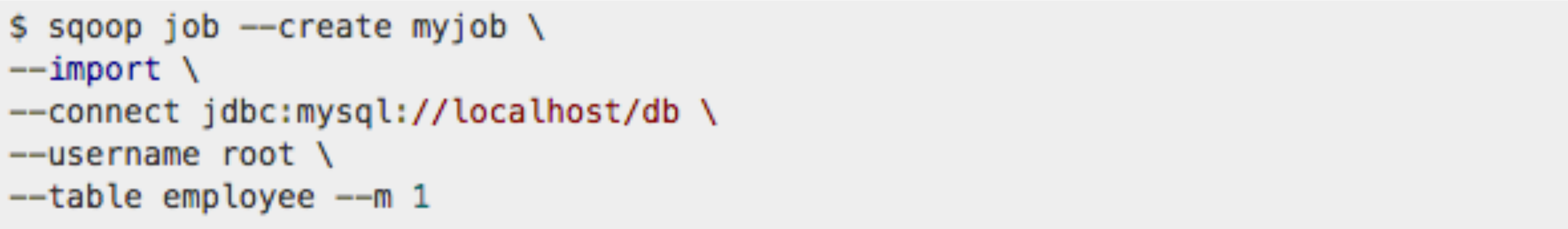


```
$ sqoop job --create myjob \
--import \
--connect jdbc:mysql://localhost/db \
--username root \
--table employee --m 1
```

- sqoop job —list
- sqoop job --show myjob
- sqoop job --exec myjob

SQOOP

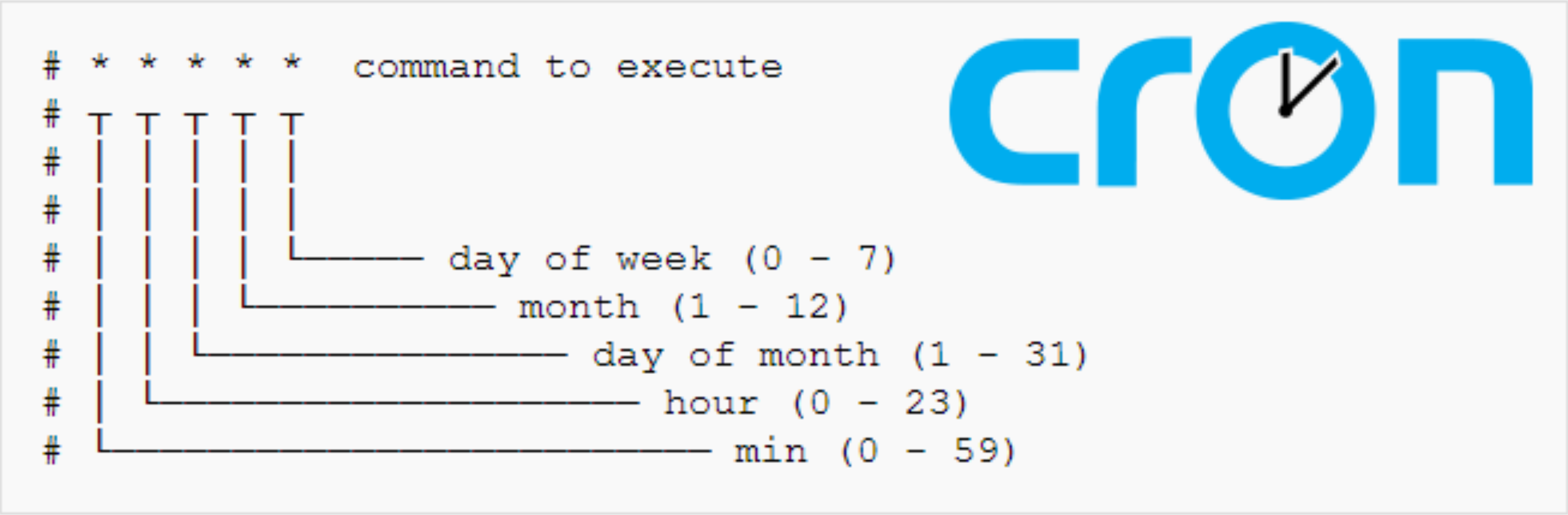
JOB



```
$ sqoop job --create myjob \
--import \
--connect jdbc:mysql://localhost/db \
--username root \
--table employee --m 1
```

- sqoop job —list
- sqoop job --show myjob
- sqoop job --exec myjob

Schedule Your Job



```
# * * * * *  command to execute
# ┬ ┬ ┬ ┬ ┬
# │ │ │ │ │
# │ │ │ │ │
# │ │ │ │ └───── day of week (0 - 7)
# │ │ │ └────────── month (1 - 12)
# │ │ └─────────────── day of month (1 - 31)
# │ └──────────────────── hour (0 - 23)
# └───────────────────────── min (0 - 59)
```