LEARN.
CONNECT.
ELEVATE.

YCBS 257 - Data at Scale (Winter 2019)

Instructor: Khaled Tannir

McGill | School of Continuing Studies

mcgill.ca
/continuingstudies

School of Continuing Studies

YCBS 257-256 / 257 - Data at Scale (BIG DATA)

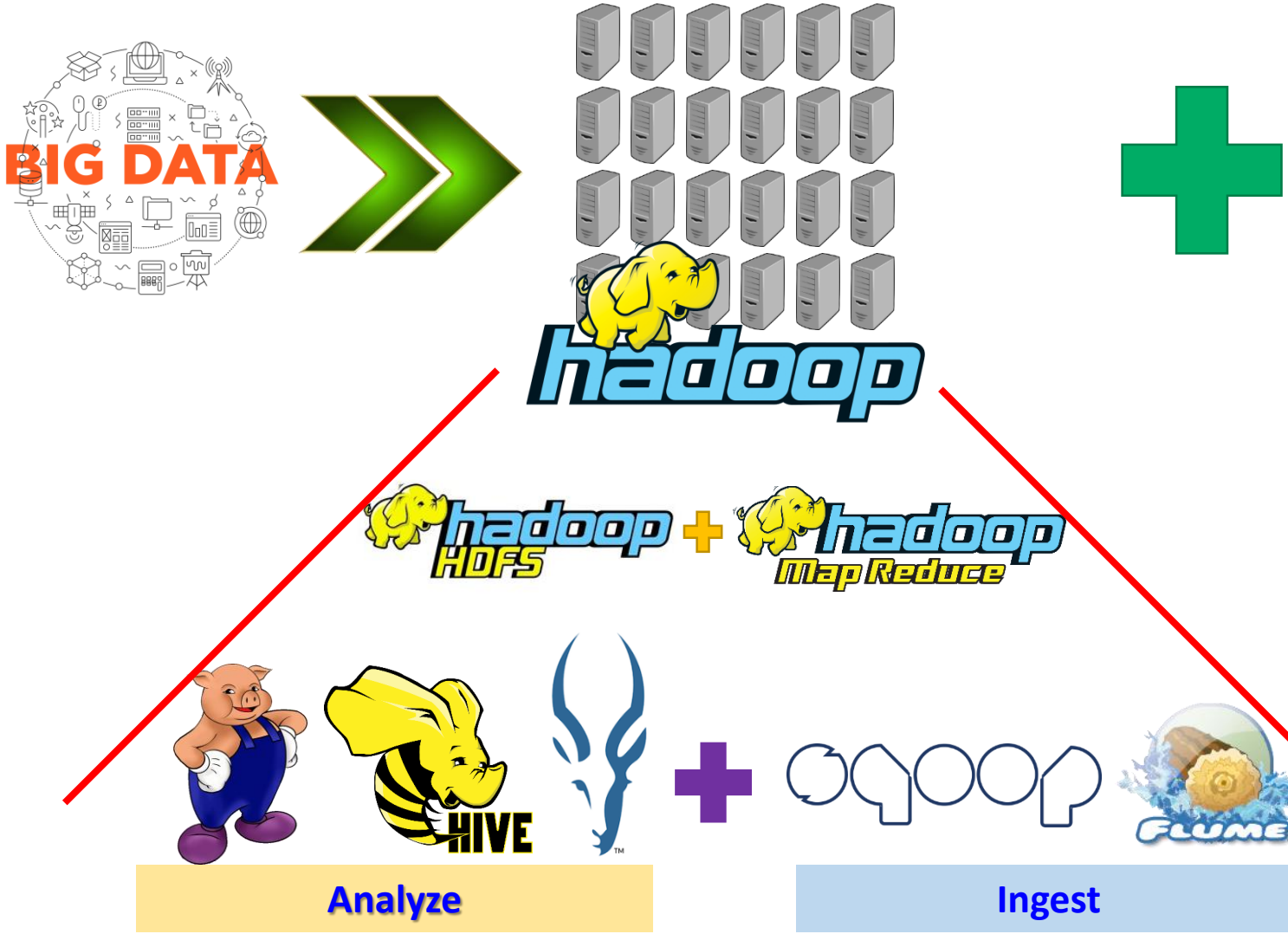
Course 6

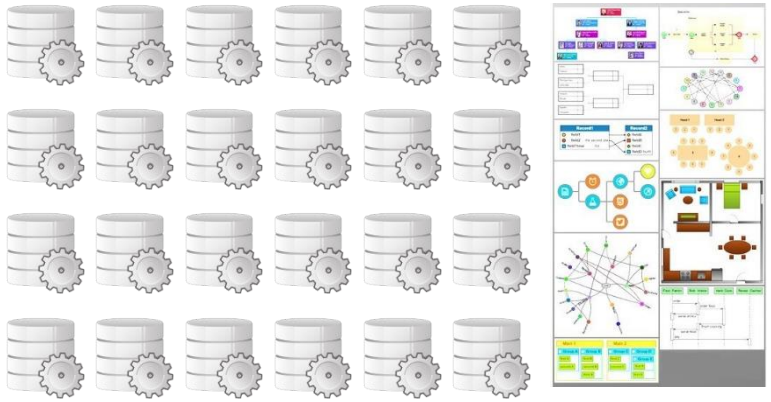
Data Ingestion in Hadoop

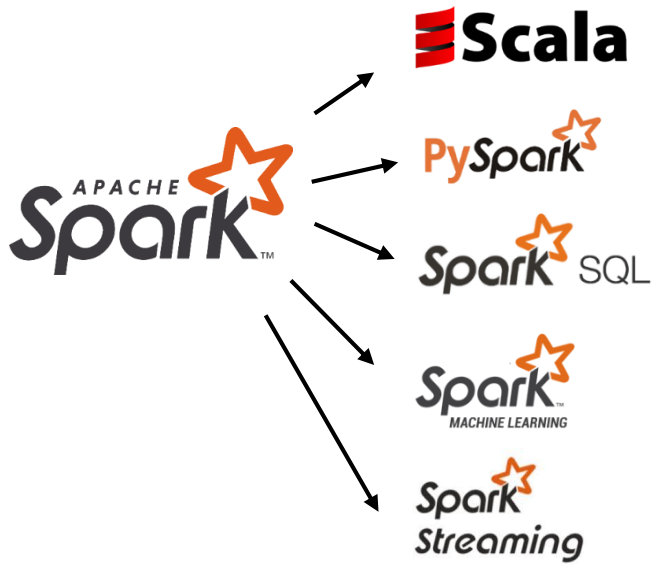
Khaled Tannir

McGill

Machine Learning at Scale







Machine Learning at Scale

Volume

Low Velocity — High Velocity

Distributed Storage

Volume

On Disk — In-Memory

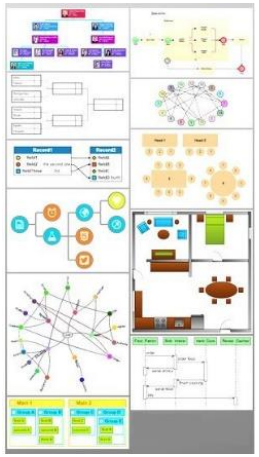
Distributed Processing

Velocity

Low Velocity — High Velocity

Variety

Unstructured — Structured



Theme of this Course

Data Ingestion in Hadoop

- *Apache Sqoop*
  - *Core Concepts*
  - *Import / Export RDBMS data to/from HDFS*

- *Apache FLume*
  - *Core Concepts*
  - *Real-time data ingestion*

Apache Sqoop

Import/Export RDBMS data to/from Hadoop

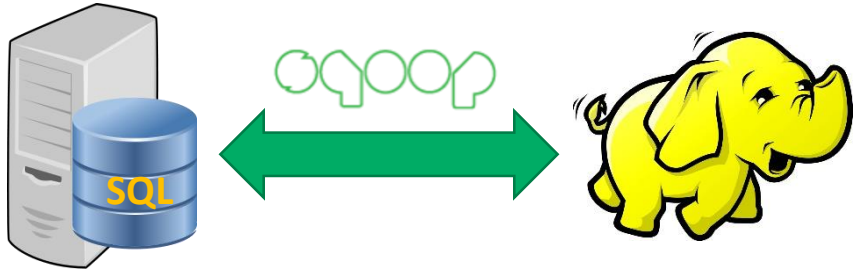
McGill

FALL 2018

What IS Sqoop?

- Sqoop is a hadoop component build on top of hdfs.
- Is a tool to transfer data To/From Hadoop ecosystem
- Sqoop imports data from external structured datastores *(databases, data warehouses, NoSQL)* into HDFS *(Text, Sequence file, Avro, Parquet,…).*
- Sqoop exports data from Hadoop to external structured datastores
- Written in Java and is open source

To/From Teradata, MySQL, PostgreSQL, Oracle, Netezza, …



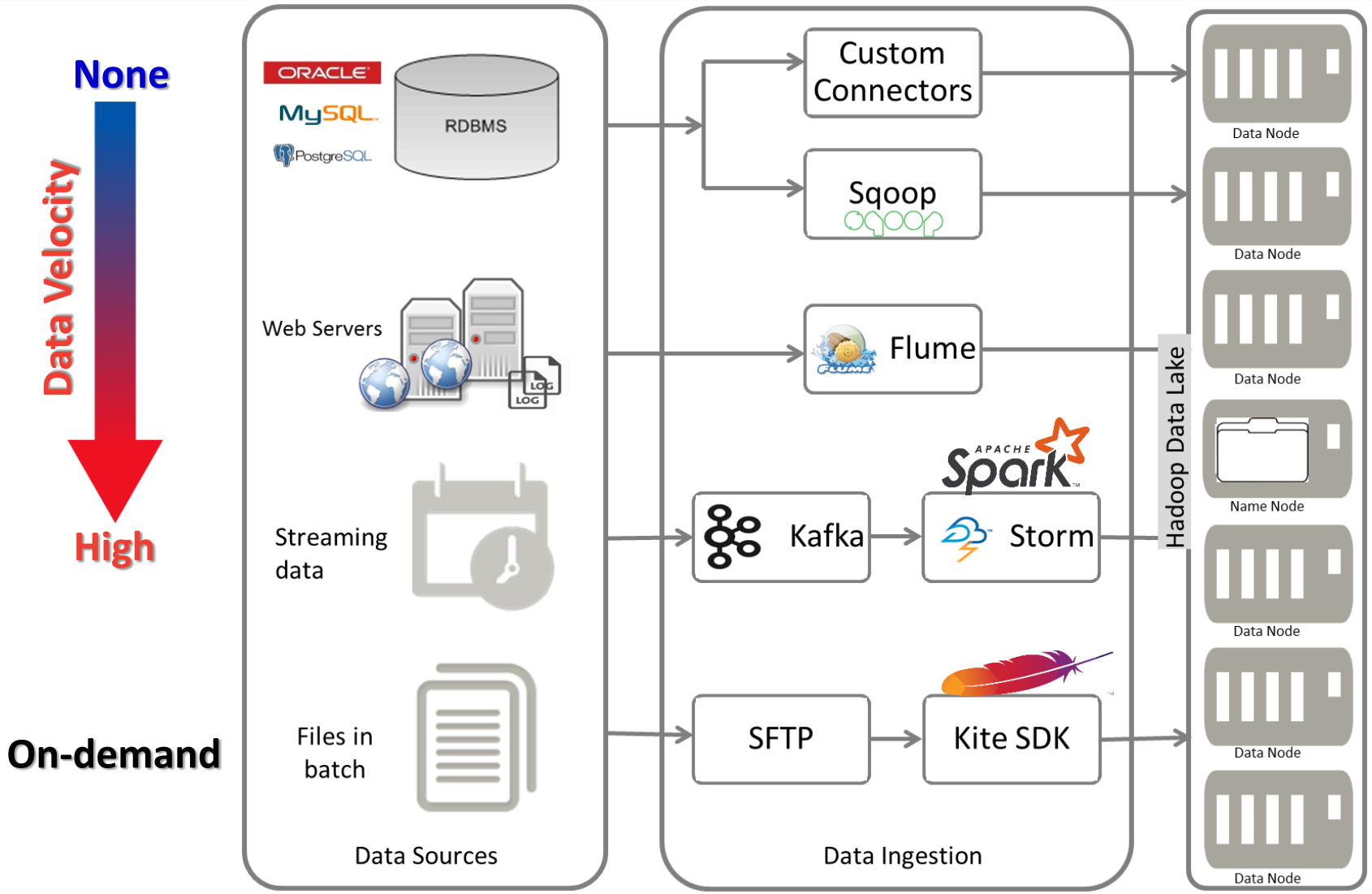
Why do I need Sqoop ?

- Need to transfer large amount of data between Hadoop and existing databases, data warehouses and other data sources
- Transferring data using scripts is a inefficient and time-consuming task
- Accessing production systems data from map-reduce applications is a challenging task

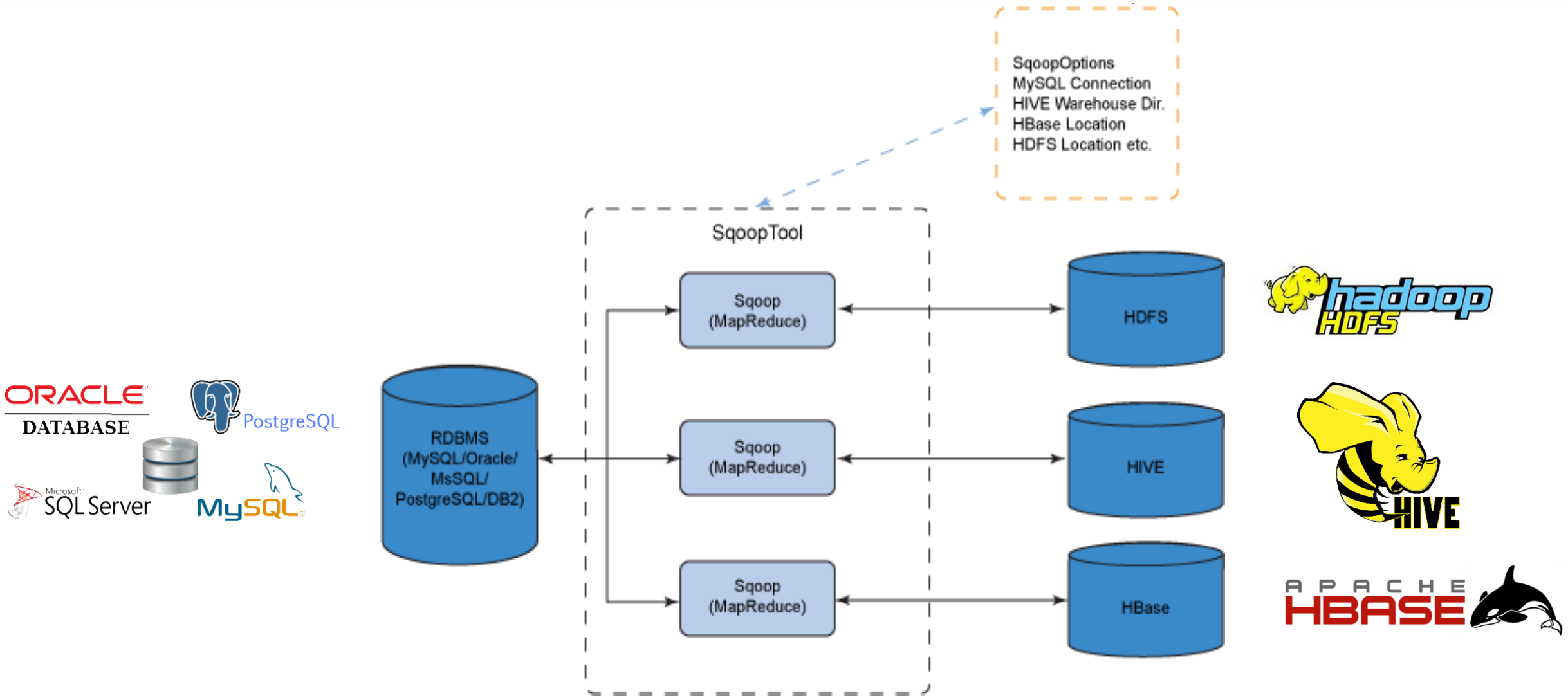
What Sqoop Does

- Uses metadata to infer data structure
- Use MapReduce to parallelize data transfer
- Store data structure into Hive metastore
- Provides a pluggable connector mechanism for optimal connectivity to external systems

Data Ingestion

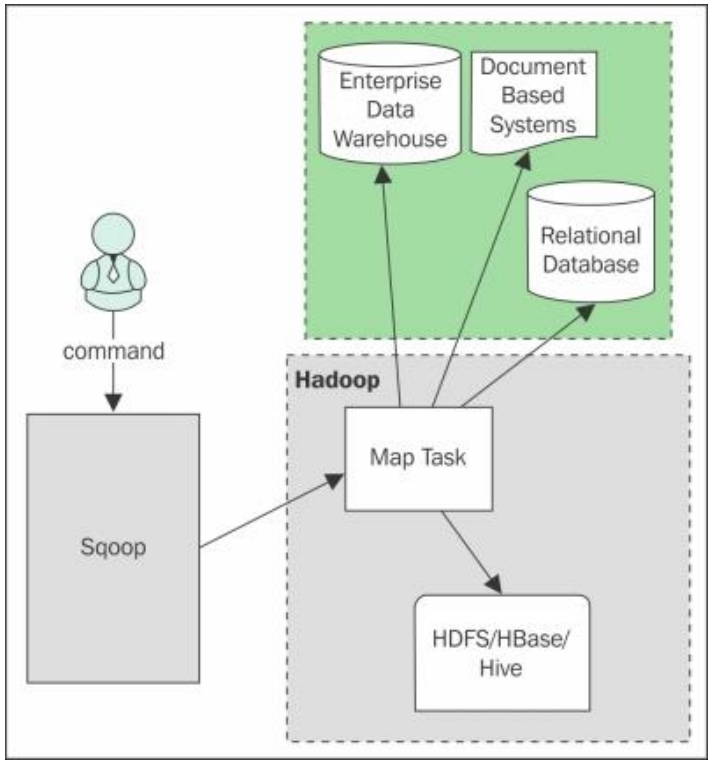


Data Ingestion using Sqoop



Sqoop Architecture

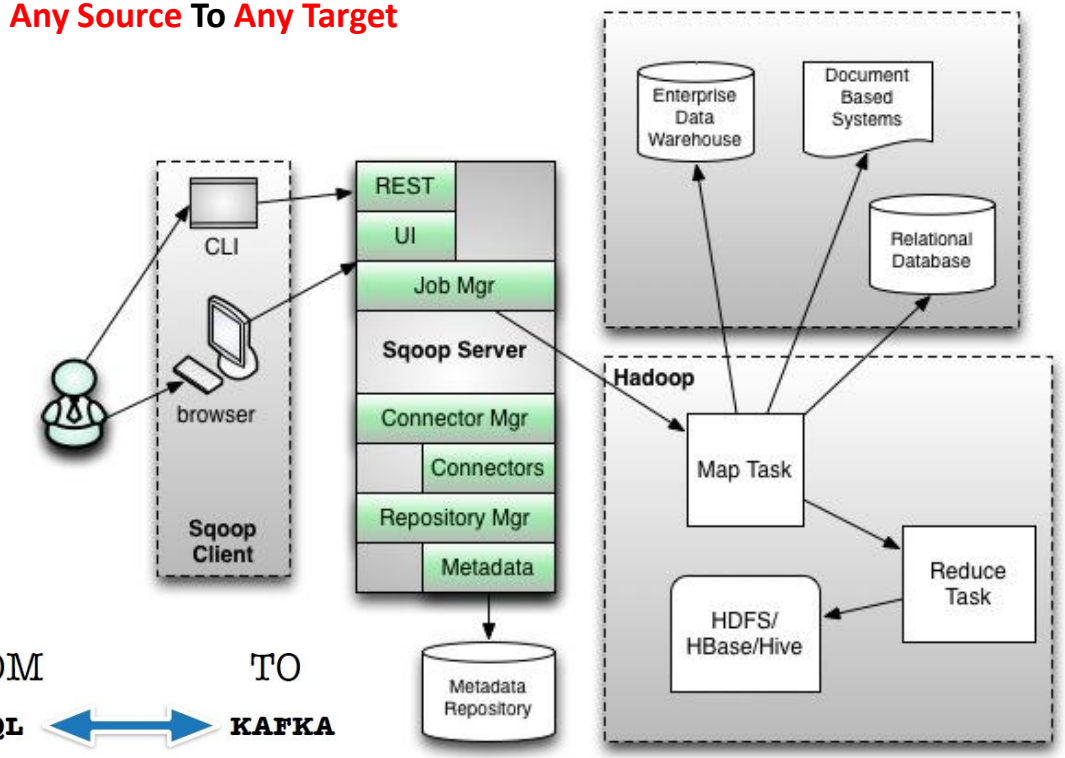
v 1



SQL ⟷ **HADOOP**

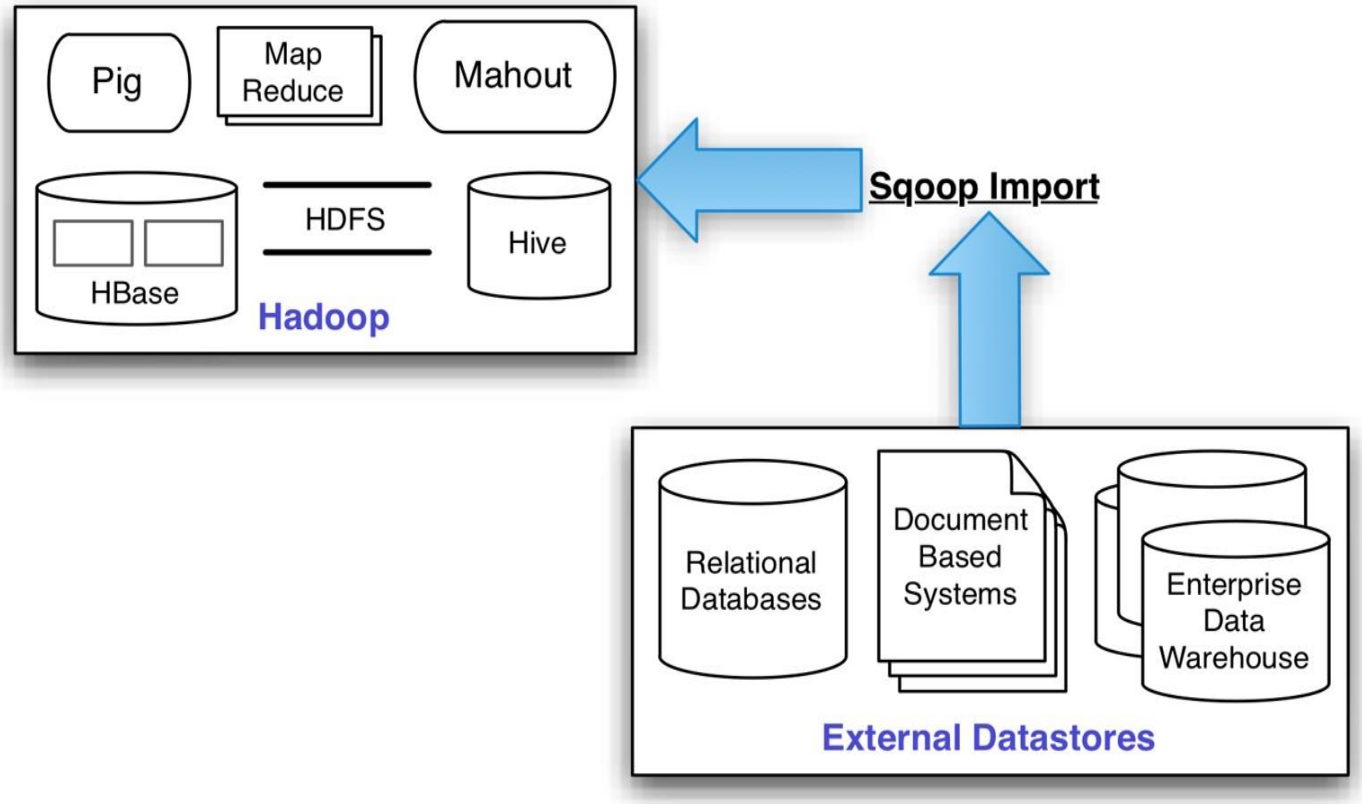
v 2

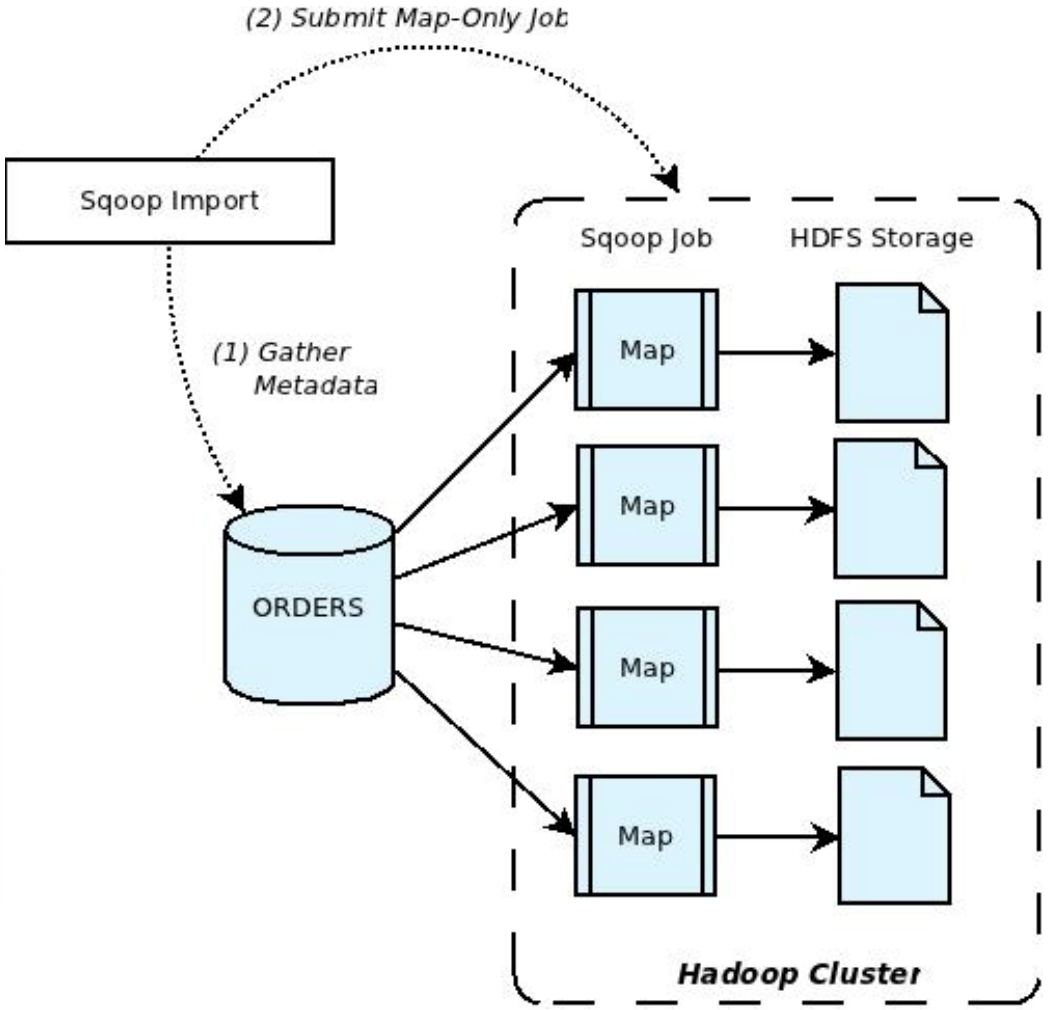
From Any Source To Any Target



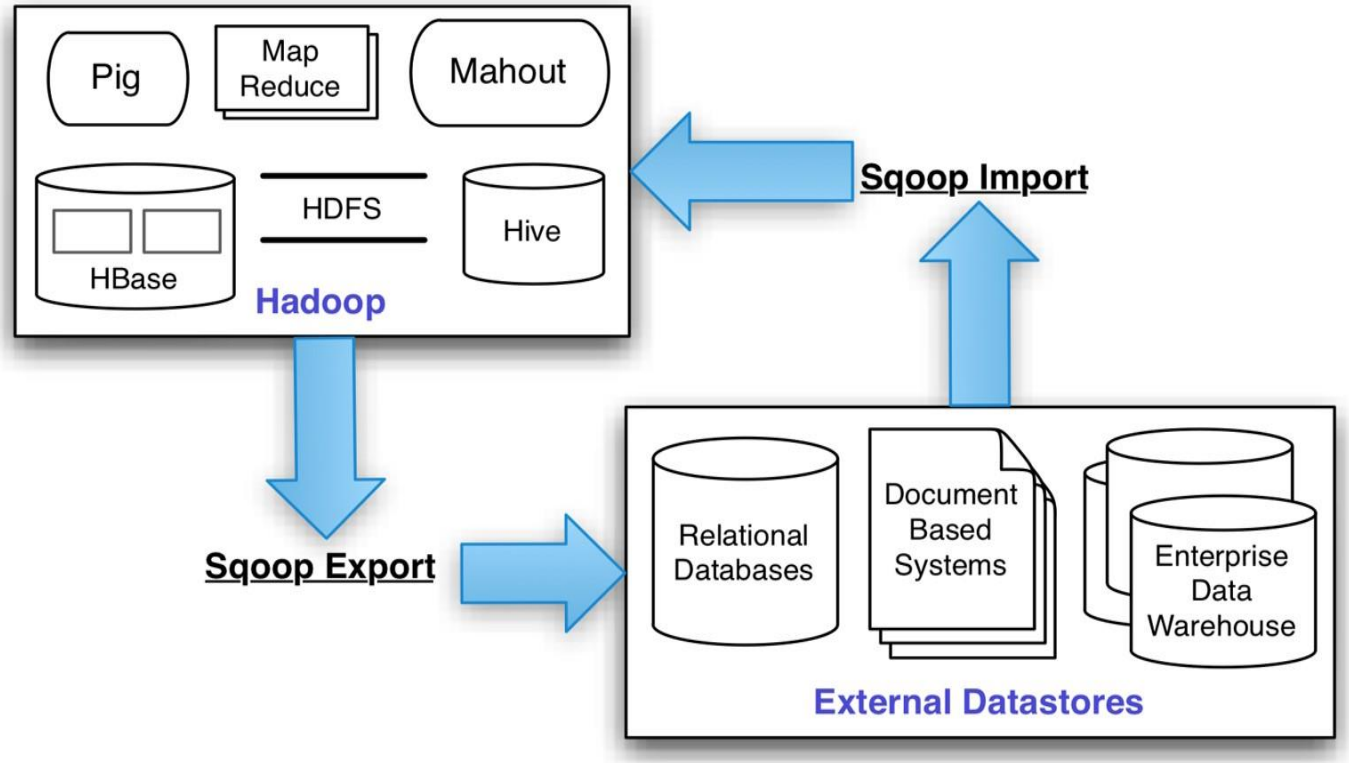
FROM		TO
MYSQL	⟷	KAFKA
HDFS	⟷	MONGO
FTP	⟷	HDFS
KAFKA	⟷	MEMSQL

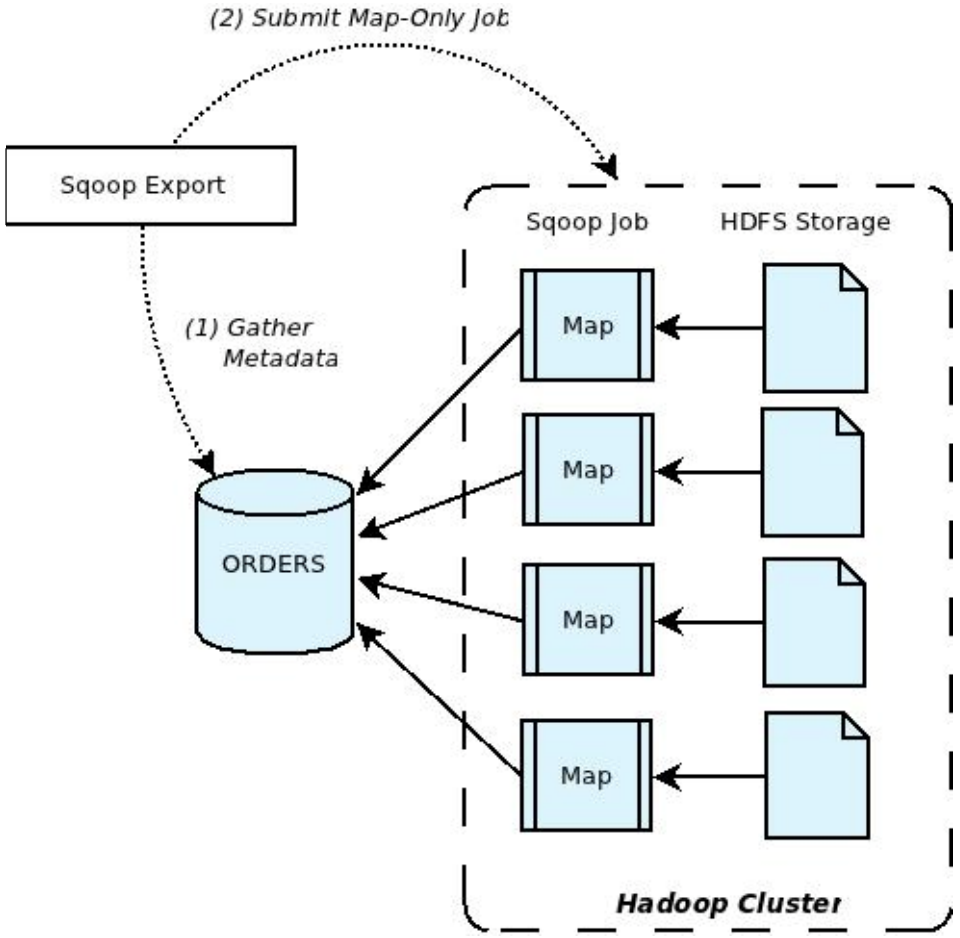
Sqoop Import Process





Sqoop Export Process





How it Works

- Sqoop schedules MapReduce tasks to Import/Export data
- Sqoop requires a specific database connector (for each database) and a JDBC driver

JDBC drivers should be placed into {sqoop home}/lib

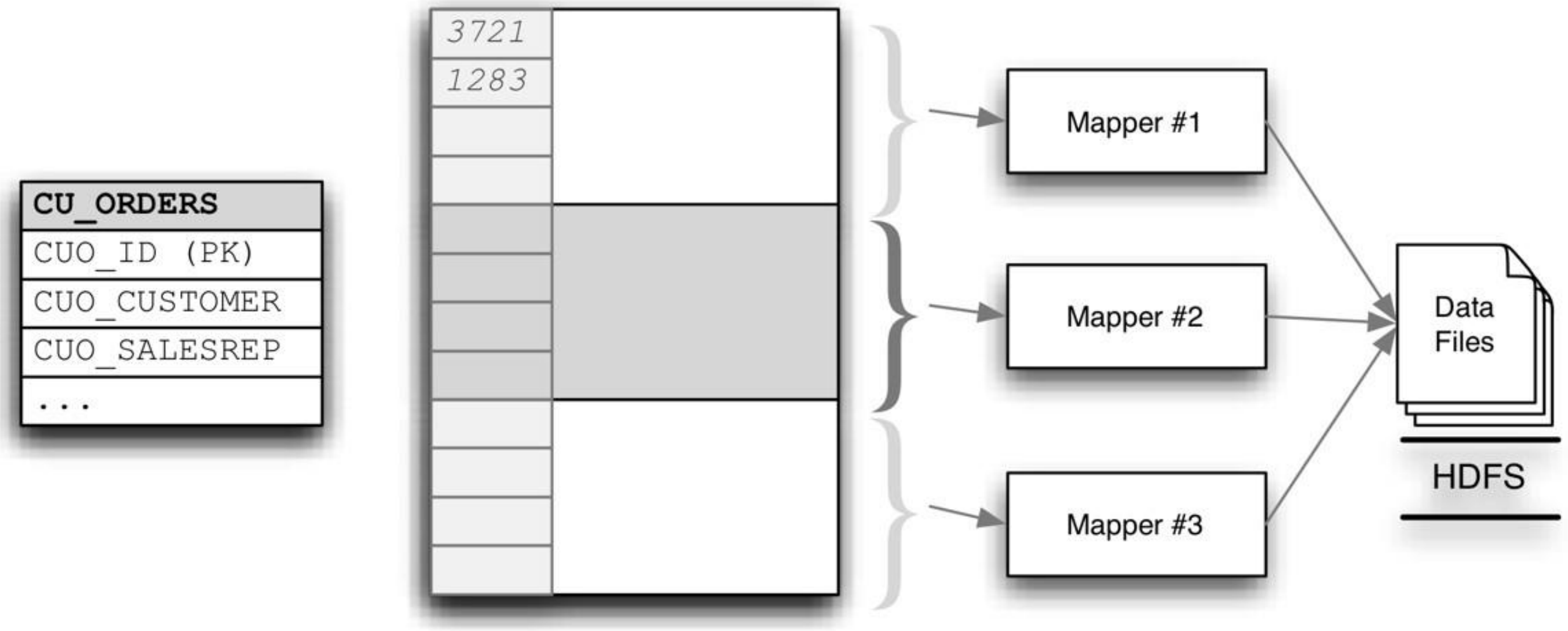
Splitting Data

- By default, the primary key is used.
- Prior to starting the transfer, Sqoop will retrieve the min/max values for this column.
- Changed column with the `--split-by` parameter:
- Required in tables with no index columns or multi-column keys.

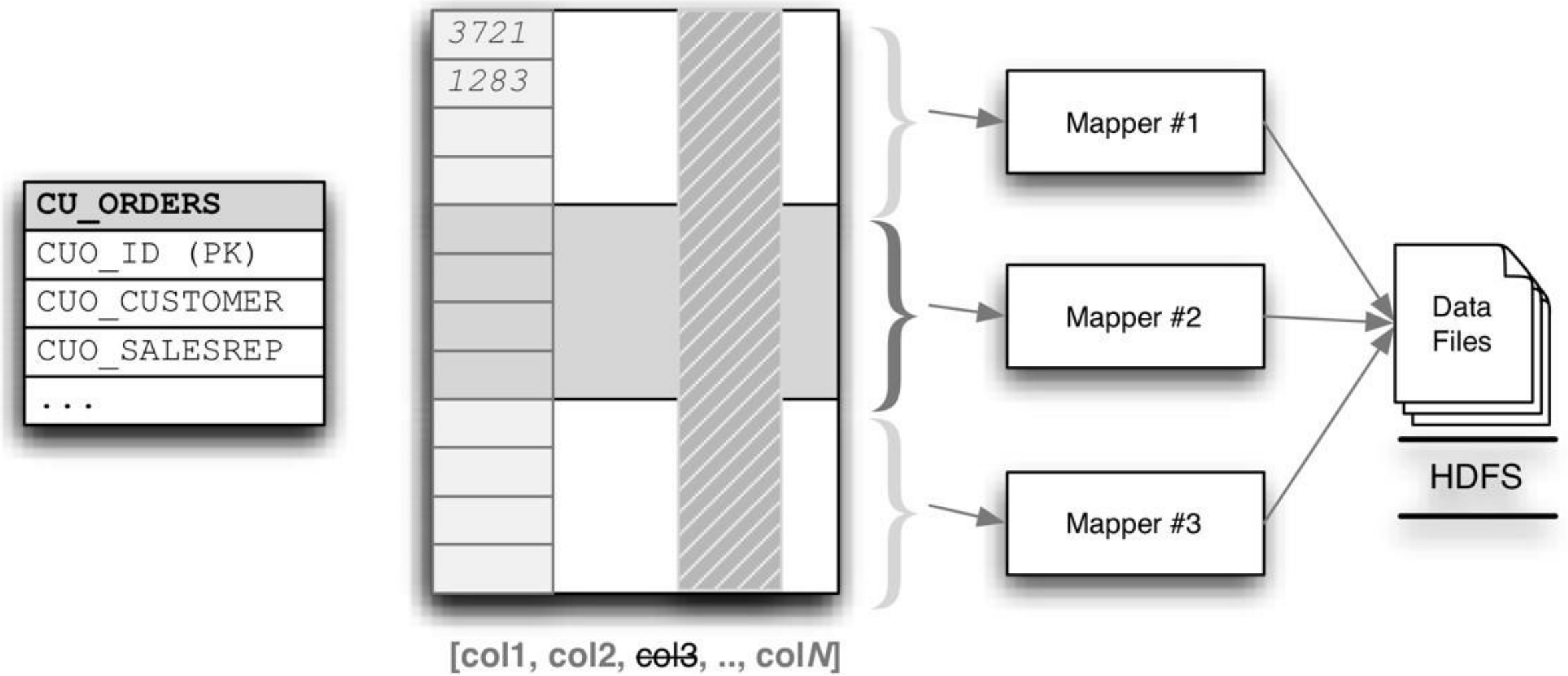
Sqoop To Hive

- `--hive-import` parameter.
  - *If table doesn't exist, Sqoop will create it for you!*
- Override default type mappings with `--map-column-hive`
- Data is first loaded into HDFS and then loaded into Hive..
- Default behavior is append. (`--hive-overwrite`)
- Hive partitions, two parameters:
  - `--hive-partition-key`
  - `--hive-partition-value`

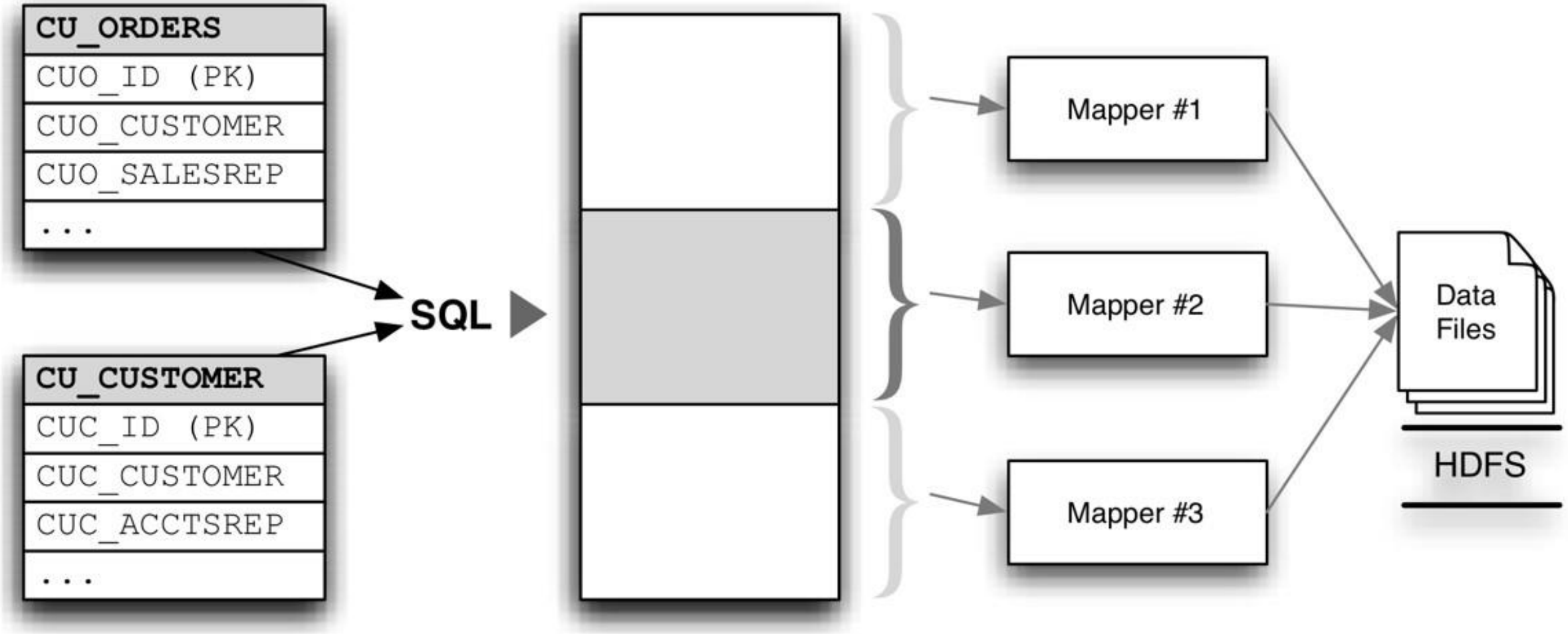
Data Import


CU_ORDERS
CUO_ID (PK)
CUO_CUSTOMER
CUO_SALESREP
...
3721
1283
Mapper #1
Mapper #2
Mapper #3
Data Files
HDFS

Data Import

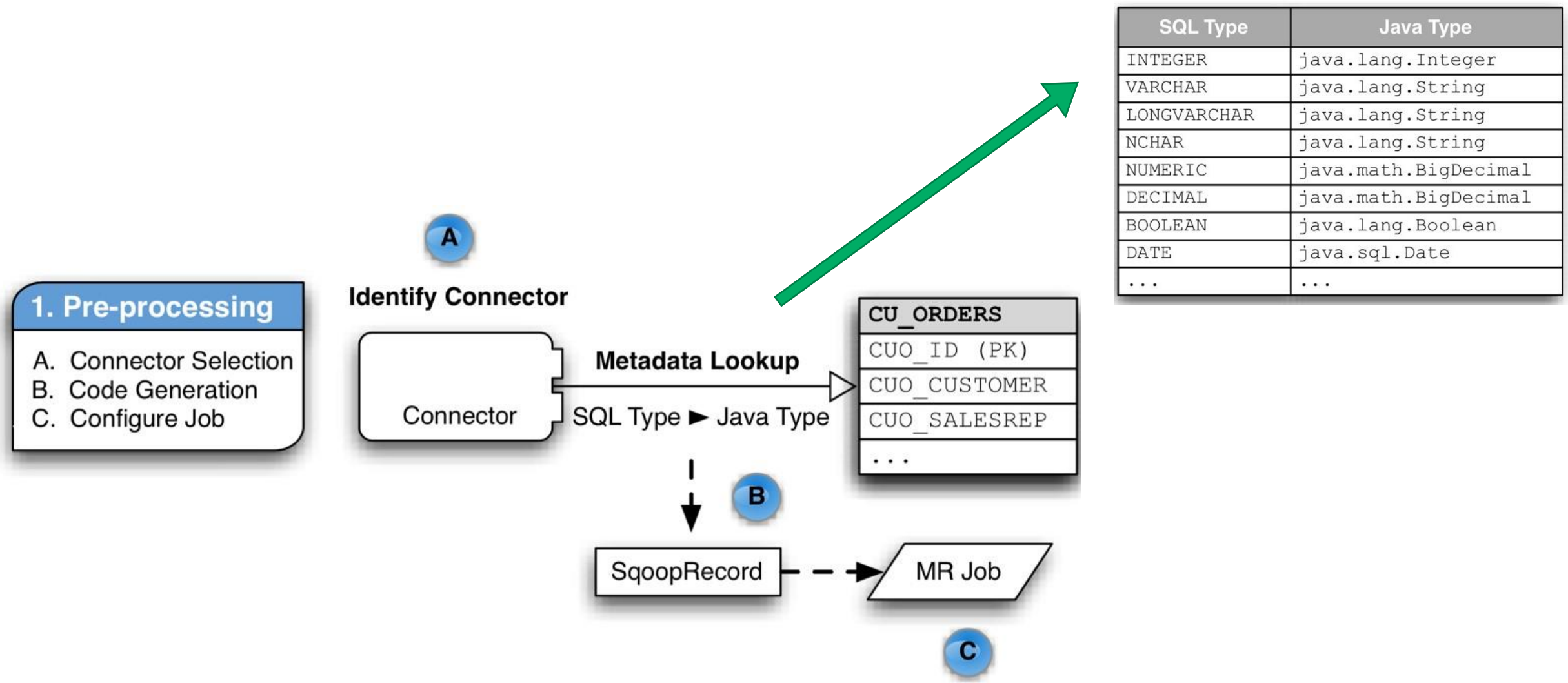

CU_ORDERS
CUO_ID (PK)
CUO_CUSTOMER
CUO_SALESREP
...
3721
1283
[col1, col2, ~~col3~~, .., colN]
Mapper #1
Mapper #2
Mapper #3
Data Files
HDFS

Data Import


CU_ORDERS
CUO_ID (PK)
CUO_CUSTOMER
CUO_SALESREP
...
CU_CUSTOMER
CUC_ID (PK)
CUC_CUSTOMER
CUC_ACCTSREP
...
SQL
Mapper #1
Mapper #2
Mapper #3
Data Files
HDFS

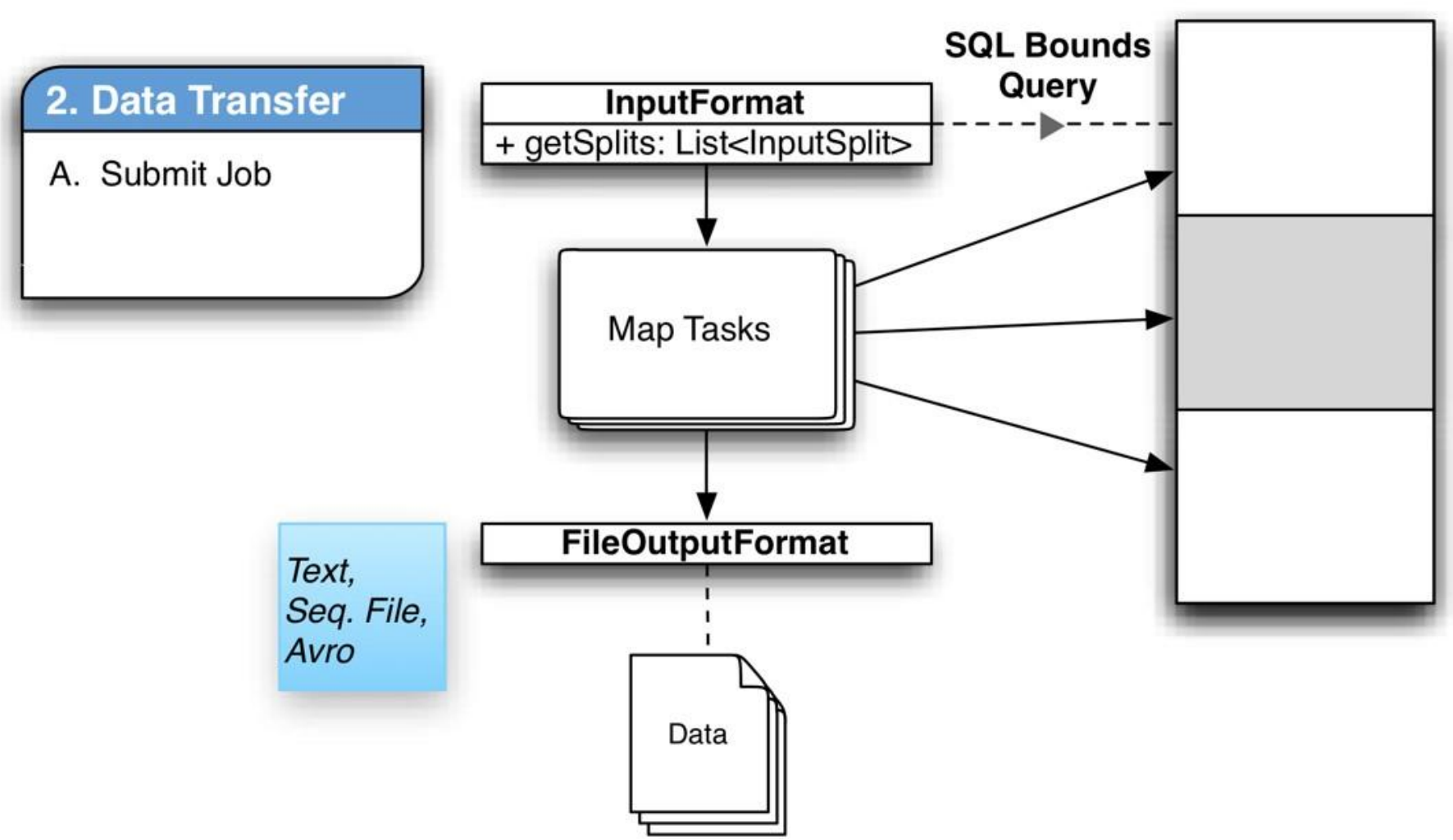
Sqoop Job Anatomy

Sqoop : Pre-Processing

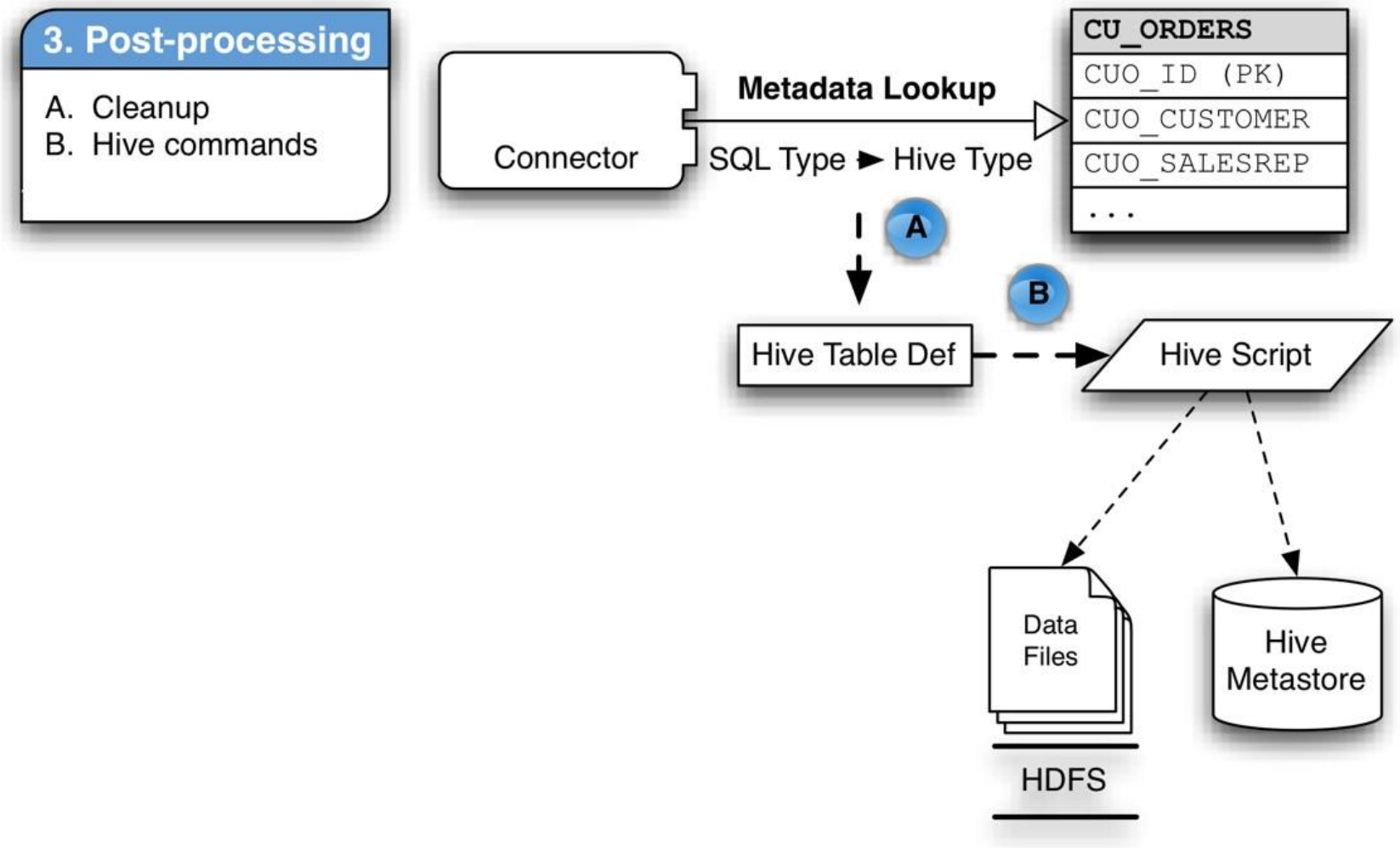


SQL Type	Java Type
INTEGER	java.lang.Integer
VARCHAR	java.lang.String
LONGVARCHAR	java.lang.String
NCHAR	java.lang.String
NUMERIC	java.math.BigDecimal
DECIMAL	java.math.BigDecimal
BOOLEAN	java.lang.Boolean
DATE	java.sql.Date
...	...

Sqoop : Data Transfer



Sqoop : Post-Processing



Sqoop Data Connectors

Connector	Developed By
Oracle	Quest Software
Couchbase	Couchbase
Netezza	Cloudera
Teradata	Cloudera
SQL Server	Microsoft
Microsoft PDW	Microsoft
Volt DB	Volt DB

Common Connectors
MySQL
PostgreSQL
Oracle
SQL Server
DB2
Generic

Direct Connectors
MySQL
PostgreSQL
Oracle
Teradata
And others

--direct parameter for Enhanced Performance!

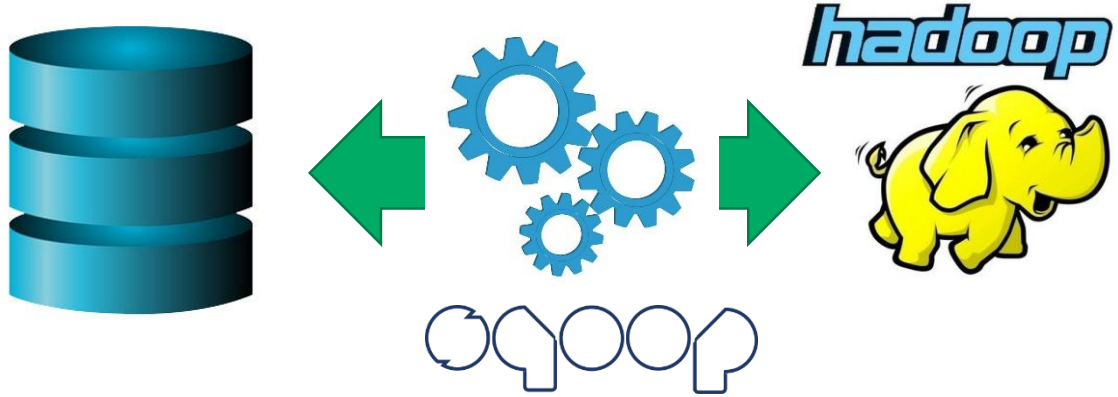
Sqoop Command Line

```
Sqoop TOOL PROPERTY_ARGS SQOOP_ARGS
```

- **TOOL** - the operation to perform (e.g: import, export, etc.)
- **PROPERTY_ARGS** – set of parameters for the Tool and/or Java (format -Dname=value)
- **SQOOP_ARGS** – all other sqoop parameters

Sqoop Operations (Tools)

- LIST-DATABASES
- LIST-TABLES
- IMPORT
- IMPORT-ALL-TABLES
- EXPORT
- EVAL
- CODEGEN
- SQOOP-JOB



Sqoop Examples

List databases and tables:

```
$ sqoop list-databases --connect jdbc:mysql://<<mysql-server>> --username user --password myPassword
```

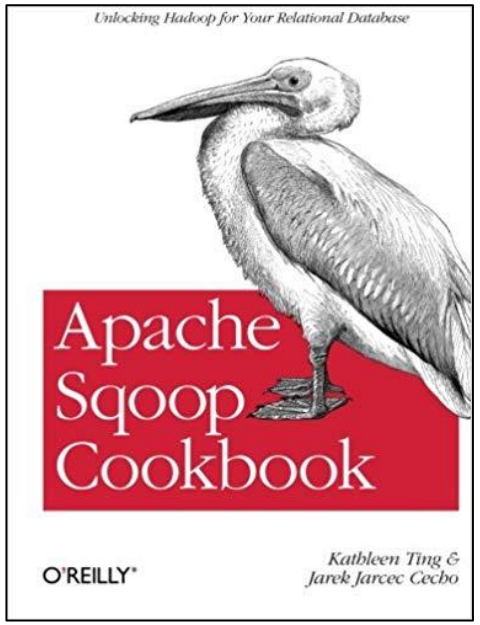
```
$ sqoop list-tables --connect jdbc:mysql://<<mysql-server>>/employees --username user --password myPassword
```

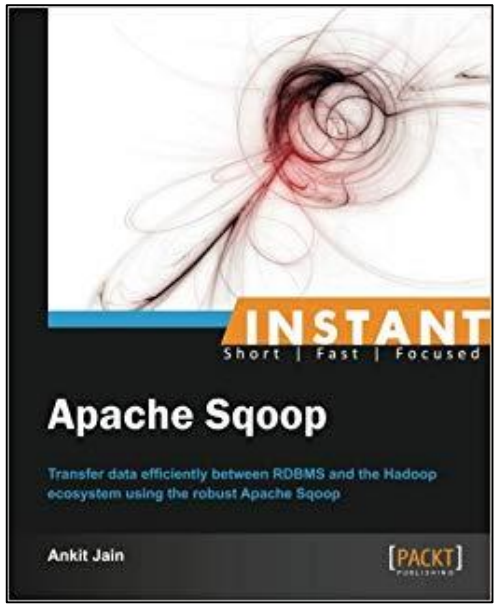
Sqoop Import Command Example

```
sqoop import --connect jdbc:mysql://localhost/userdb
--username hadoop_tutorial -P --num-mappers 1 --table
employees --where «city = 'NYC'» --target-dir
query_result
```

Sqoop Workshop

Resources





Online Resources

http://sqoop.apache.org/

Thank You