CARLSON SCHOOL
OF MANAGEMENT
UNIVERSITY OF MINNESOTA

Data Ingestion with Scoop

MSBA 6330 Prof Liu

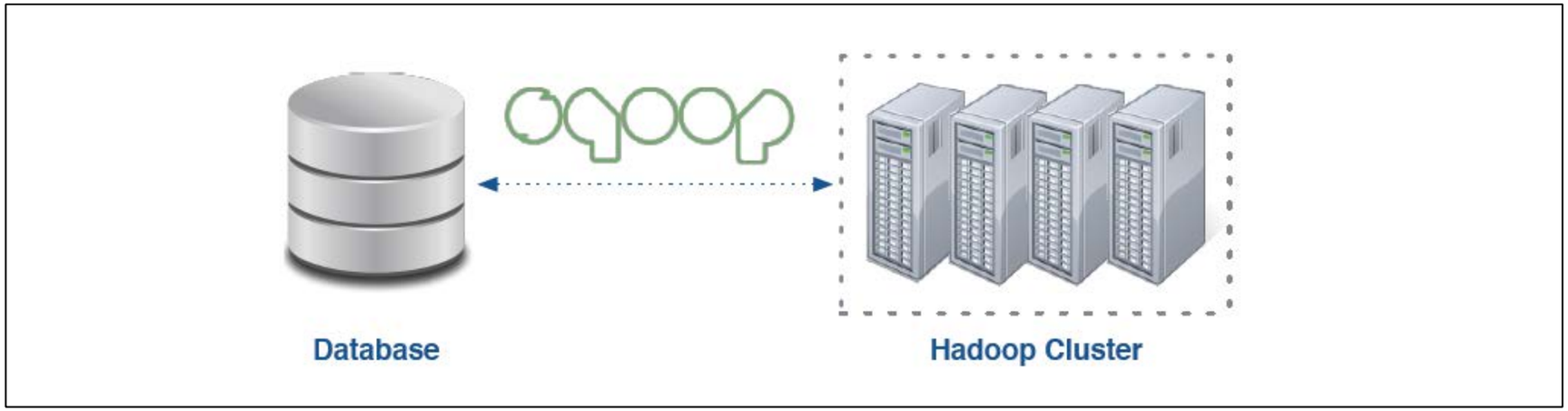
Slides credits: Cloudera Academic Partners Program

Goals

- In this section, you will learn
  - The role of sqoop in Hadoop ecosystem
  - How to use sqoop to import RDBMS tables into HDFS
  - How to use sqoop to export HDFS data into RDBMS tables

Apache Sqoop Overview

- Sqoop exchanges data between an RDBMS and Hadoop
  - Name is a contraction of “SQL-to-Hadoop”
- It can import all tables, a single table, or a portion of a table into HDFS
  - Result is a directory in HDFS containing comma delimited text files (customizable)
- Sqoop can also export data from HDFS back to the database



Apache Sqoop Overview

- Sqoop is a command-line utility with several subcommands, called tools.

  ```
  sqoop toolname [tool-options]
  ```

  There are tools for import, export, listing database contents etc, such as:

  ```
  import
  import-all-tables
  export
  list-tables
  ```

- Imports are performed using **Hadoop MapReduce** jobs
  - Does this efficiently via a throttled, map-only MapReduce job
    - Benefiting from parallel transfers (4 by default)

Importing Tables With Sqoop

- This example imports the customers table from a MySQL database
  - Will create `/mydata/customers` directory in HDFS
  - Directory will contain **comma delimited text** files (by default)
    - One line of data per SQL record (the " \" at the end is for breaking long lines)

```
sqoop import \
     --connect       jdbc:mysql://localhost/company \
     --username      twheeler \
     --password      bigsecret \
     --warehouse-dir /mydata \
     --table         customers
```

- Cloudera offers high-performance custom connectors for many databases.
- By default it uses the id (primary key) column of the table and runs 4 parallel import processes, each importing a range of ids.

Importing an Entire Database With Sqoop

- Import all tables from the database using `import-all-tables` tool (in this example, fields will be tab-delimited)

```
sqoop import-all-tables \
     --connect              jdbc:mysql://localhost/company \
     --username                  twheeler \
     --password                  bigsecret \
     --fields-terminated-by      '\t' \
     --warehouse-dir             /mydata
```

- You want to be careful with the choice of delimiters. If the data contains symbol “\t”, then you should not use “\t” as delimiter.

Importing Partial Tables With Sqoop

- Import only specified columns from the products table

```
sqoop import \
     --connect      jdbc:mysql://localhost/company \
     --username          twheeler \
     --password          bigsecret \
     --warehouse-dir     /mydata \
     --table        products \
     --columns      "prod_id, name, price"
```

- Import only matching rows from the products table

```
sqoop import \
     --connect      jdbc:mysql://localhost/company \
     --username          twheeler \
     --password          bigsecret \
     --warehouse-dir     /mydata \
     --table        products \
     --where        "price >= 1000"
```

Incremental Imports With Sqoop

- What if new records are added to the database?
  - You could re-import all records, but this is inefficient
- Sqoop's incremental append mode imports only new records
  - Based on value of the last record in a specified column
  - Note that this is **an append, not an update**

```
sqoop import \
    --connect          jdbc:mysql://localhost/company \
    --username              twheeler \
    --password              bigsecret \
    --warehouse-dir         /mydata \
    --incremental      append \
    --check-column     order_id \
    --last-value       6713821
```

Handling Modifications With Incremental Imports*

- What if existing records are also modified in the database?
  - Incremental append mode doesn't handle this
- Sqoop's "lastmodified" append mode adds and updates records
  - Caveat: You must maintain a timestamp column in your table

```
sqoop import \
     --connect              jdbc:mysql://localhost/company \
     --username                  twheeler \
     --password                  bigsecret \
     --warehouse-dir             /mydata \
     --incremental          lastmodified \
     --check-column         last_update_date \
     --last-value           "2013-06-12 03:15:59"
```

Beneath the hood: Sqoop runs two standalone MapReduce jobs. One that imports the changed rows into a temporary directory in HDFS. The other that merges the changed rows with the records previously imported into a new set of up-to-date records. So its not a true update in the sense that existing records are modified. Instead, the files containing them are replaced.

Exporting Data From Hadoop To An RDBMS With Sqoop

- We've seen several ways to pull records from an RDBMS into Hadoop
  - It is sometimes also helpful to push data in Hadoop back to an RDBMS
- Sqoop supports this via export

```
sqoop export \
     --connect          jdbc:mysql://localhost/company \
     --username              twheeler \
     --password              bigsecret \
     --export-dir       /mydata/recommeder_output \
     --table            product_recommendations
```

- Typical use of Sqoop
  - Scheduled data import / export

Summary and Other Features

- Sqoop exchanges data between a database and the Hadoop cluster
  - Provides subcommands (tools) for importing, exporting, and more
- Tables are imported using parallel MapReduce jobs
  - These are written as comma-delimited text by default
  - You can specify **alternate delimiters** or **file formats**
  - Uncompressed by default, but you can specify a codec to use
  - It can import data into Hive (we'll see this later)
- Sqoop provides many options to control imports
  - You can select only certain columns or limit rows
  - Supports using joins in free-form queries

- Lab: Hands-On Exercise: Data Ingest With Hadoop Tools

Review Question

- Sqoop functions 

exchanges data between an RDBMS and Hadoop

Scoop Resources

- The following offer more information on topics discussed in this chapter
  - Sqoop User Guide (Your go-to resources for sqoop usage)
    - http://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html
  - Apache Sqoop Cookbook (published by O'Reilly)
    - http://shop.oreilly.com/product/0636920029519.do
  - Sqoop 2 is in development and quite different
    - http://tiny.cloudera.com/adcc05c (Power Point slides)
    - https://sqoop.apache.org/docs/1.99.7/index.html (official Documentation)