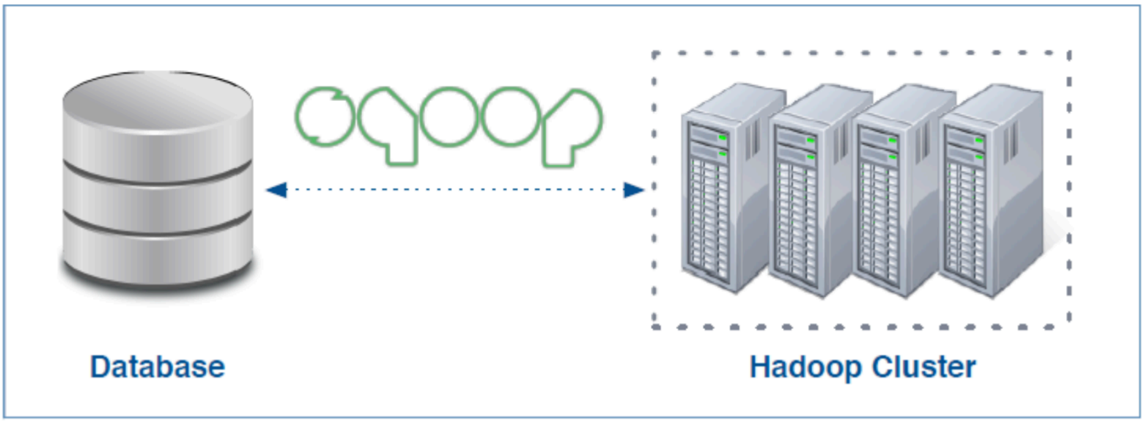
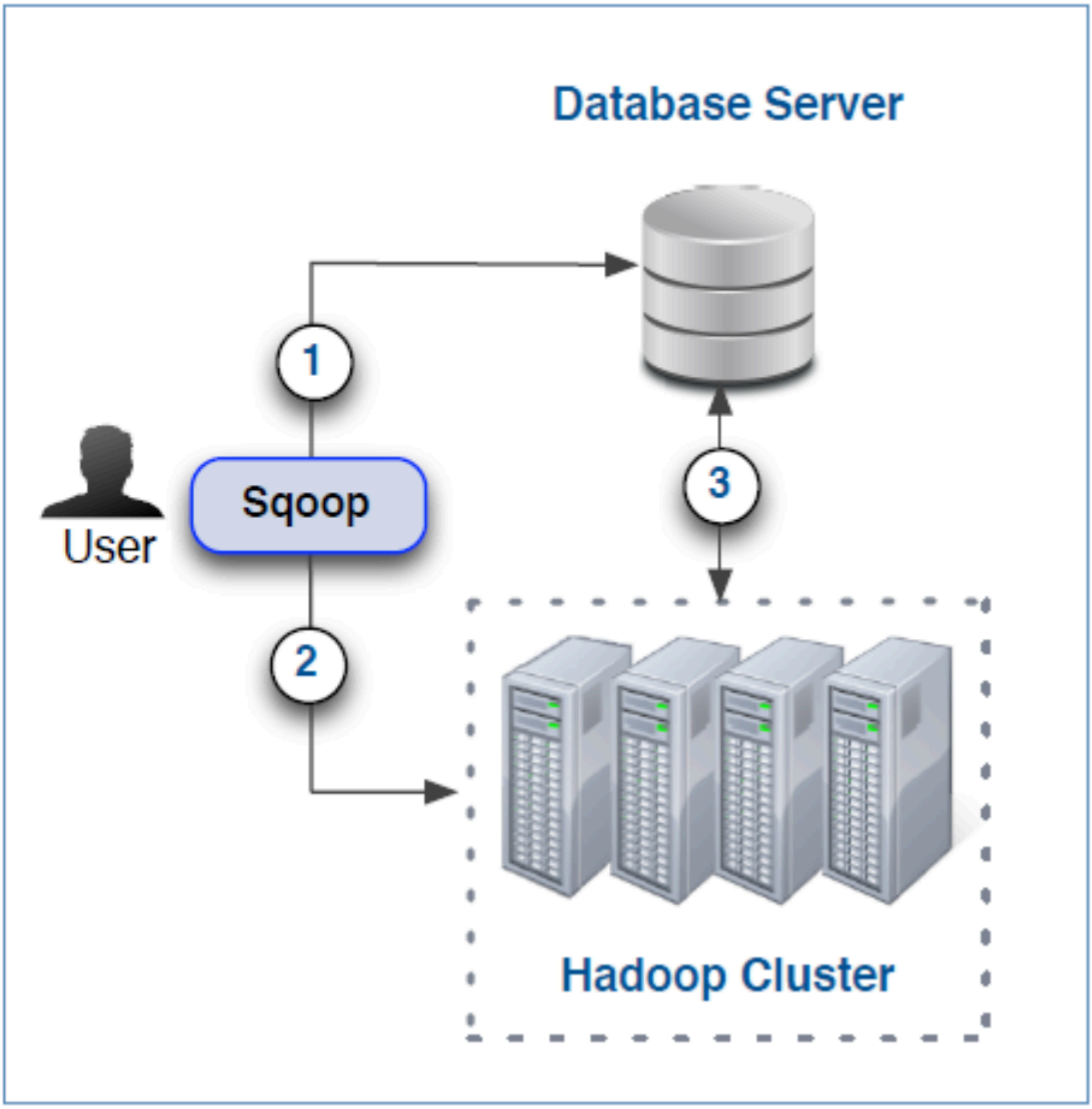
What is Apache sqoop?



- open source Apache project originally developed by Cloudera
- Sqoop exchages data between a database and HDFS

How Does Sqoop Work?

- sqoop is a client side application that imports data useing Hadoop MapReduce
- A basic import involves three steps orchestrated by Sqoop
  - examine table details
  - create and submit job to cluster
  - fetch records from table and write this data to HDFS



- Imports are performed using Hadoop MapReduce jobs
- Sqoop begins by exmaining the table to be imported
    - Determined the primary key if possible
    - Runs a boundary query to see how many records will be imported
    - Devides result of boundary query by the number of tasks (mappers)
        - uses this to confikgure tasks so that they will have equal loads
- Sqoop also generates a java source file for each table being imported
    - it compiles and uses this during the import process
    - the file remains after import, but can be safely deleted

List-tables

```
$ sqoop list-tables --conect jdbc:mysql://localhost/loudcre --username
username --password password
```

import

- map reduce로 진행됨
- warehouse-dir : 저장될 디렉토리
- fields-terminated-by "\t" : delemeter를 사용

```
$ sqoop import --table tablname --conect jdbc:mysql://localhost/loudcre --
username username --password password --warehouse-dir /loudcre --fields-
terminated-by "\t"
```

imcremental imports

- last modified : based on a timestamp in a specified column
    - import new and modified record

```
$ sqoop import --table invoices --conect jdbc:mysql://localhost/loudcre --
username username --password password --incremental lastmodified --check-
column mod_dt --last-value '2015-09-30 16:00:00' --targe-dir
```

- append
    - import only new record based on value of last record in specified column

```
sqoop import --table invoices --conect jdbc:mysql://localhost/loudcre --
username username --password password --incremental append --check-column id -
-last-value 9478306
```

- importing partial tables with sqoop

```
sqoop import --table invoices --conect jdbc:mysql://localhost/loudcre --
username username --password password --coulumns "id,first_name,last-
name,state"
```

- import only matching rows from accounts table

```
sqoop import --table invoices --conect jdbc:mysql://localhost/loudcre --
username username --password password --where "state='ca'"
```

- Using a free-form query
    - must add literal where $condition
    - use split-by identify field used to divide work among mapper
    - target-dir : free form query를 사용하면 필수로 들어가야할 옵션
        - 다른 경우에는 home 디렉토리에 table 이름으로 저장됨(table 이름을 알 수 있어서)

    - Query 절은 무조건 single quote 로 사용해야함 where $CONDITIONS때문에

```
sqoop import --table invoices --conect jdbc:mysql://localhost/loudcre --
username username --password password --target-dir target path --split-by
accounts.id --query 'select ~~~~ from table join table on (table.id =table.id)
where $CONDITIONS'
```

export : hdfs에서 database로 이동

```
sqoop export --conect jdbc:mysql://localhost/loudcre --username username --
password password export-dir export path --update-mode allowinsert --table
tablename
```

Option for database connectivity

- generic (JDBC)
    - compatible with nearly any database
    - over head imposed by jdbc can limit performance
- direct mode
    - use —direct (currently support mysql and Postgres)
    - Can imporve performance
    - 모든 sqoop기능이 지원하지 않음

Controlling Parallelism

- -m 옵션을 통해서 매퍼 개수를 정함, 기본적으로 4개의 병렬 유닛으로 처리가 됨.
- 환경 (node 개수에 따라서) 에 따라 최대개수가 정해짐