upGrad

# Data Ingestion with Apache Sqoop and Apache Flume - Session 2

**Course:** DS - DE

**Lecture On:** Apache Sqoop - I

**Instructor:** Hitesh Hasija

upGrad

# Segment - 01
# Session Overview

# Session Overview

1. Apache Sqoop and its Architecture
2. Case studies for the demonstration of Apache Sqoop
3. The setup of Apache Sqoop and Databases
4. Using Sqoop to export and import data
5. Using various arguments of the Sqoop Import command

upGrad

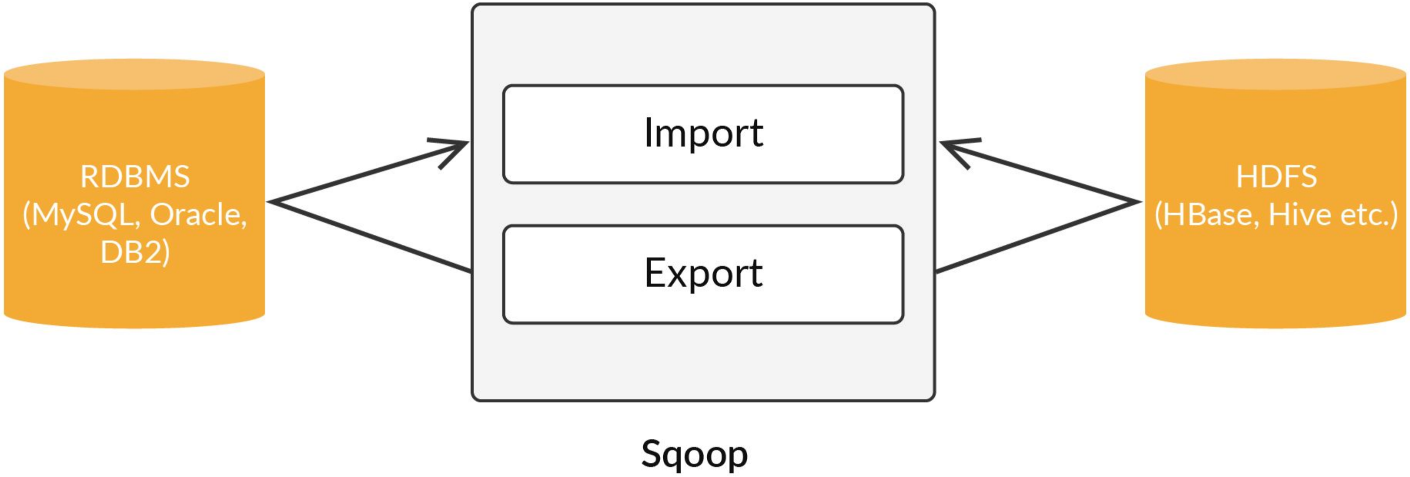
# Segment - 02
# Introduction to Sqoop and its Architecture

# Learning Objectives

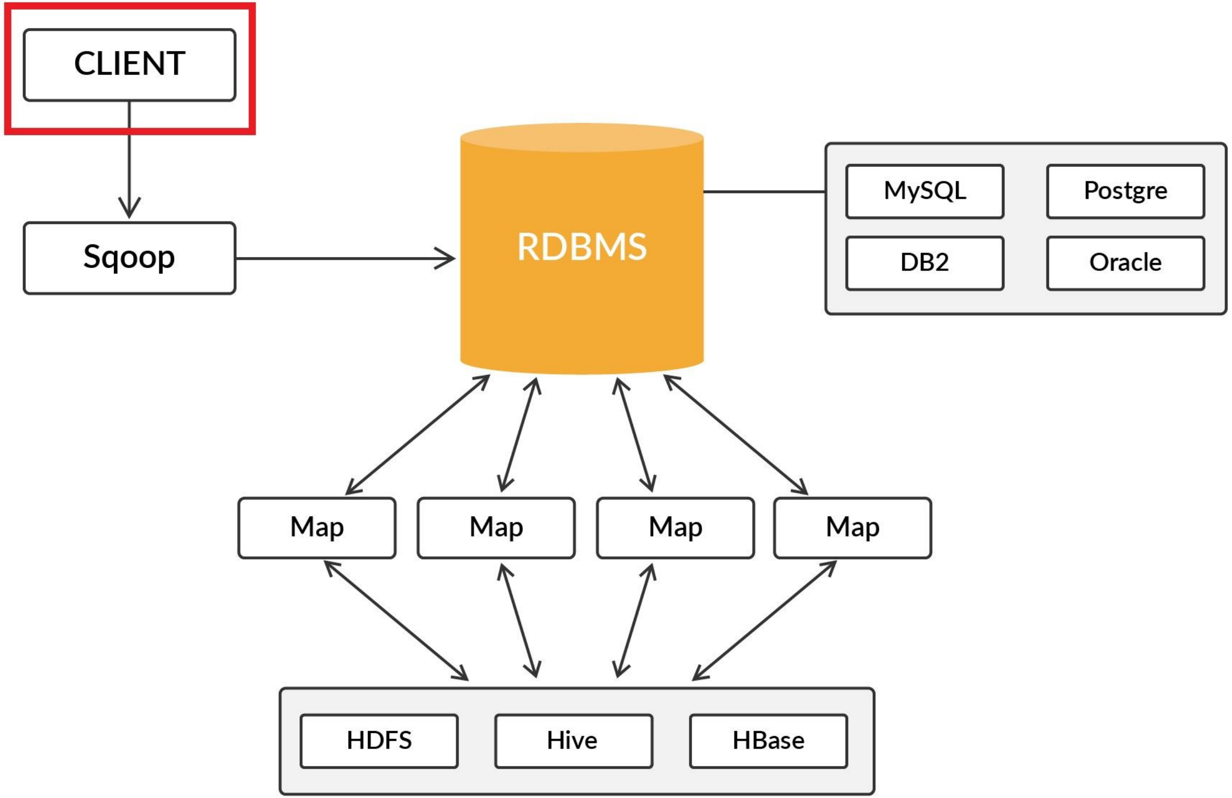
1. Introduction to Apache Sqoop
2. Architecture of Apache Sqoop
3. Advantages of Apache Sqoop

# Introduction to Sqoop and its Architecture

**Sqoop connects an RDBMS, such as MySQL and Oracle, with the HDFS and provides efficient, bidirectional data transfer between them in parallel.**

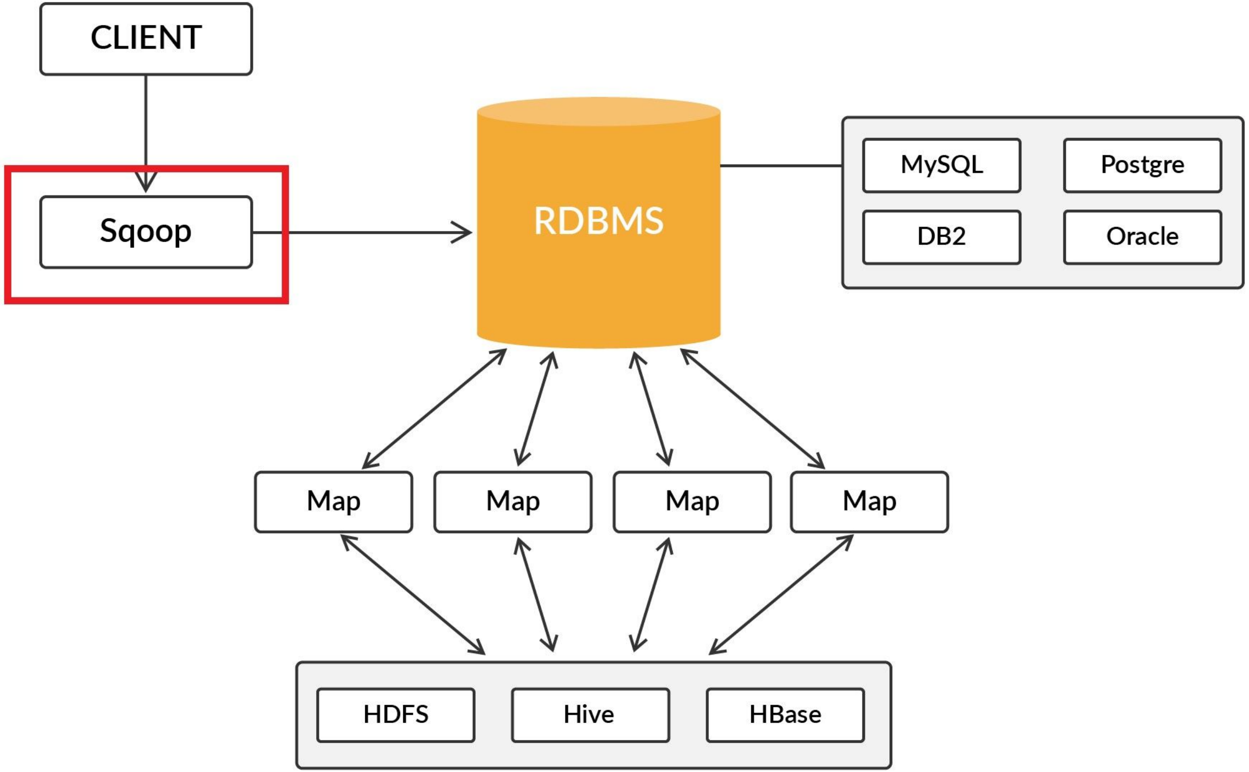


# Introduction to Sqoop and its Architecture

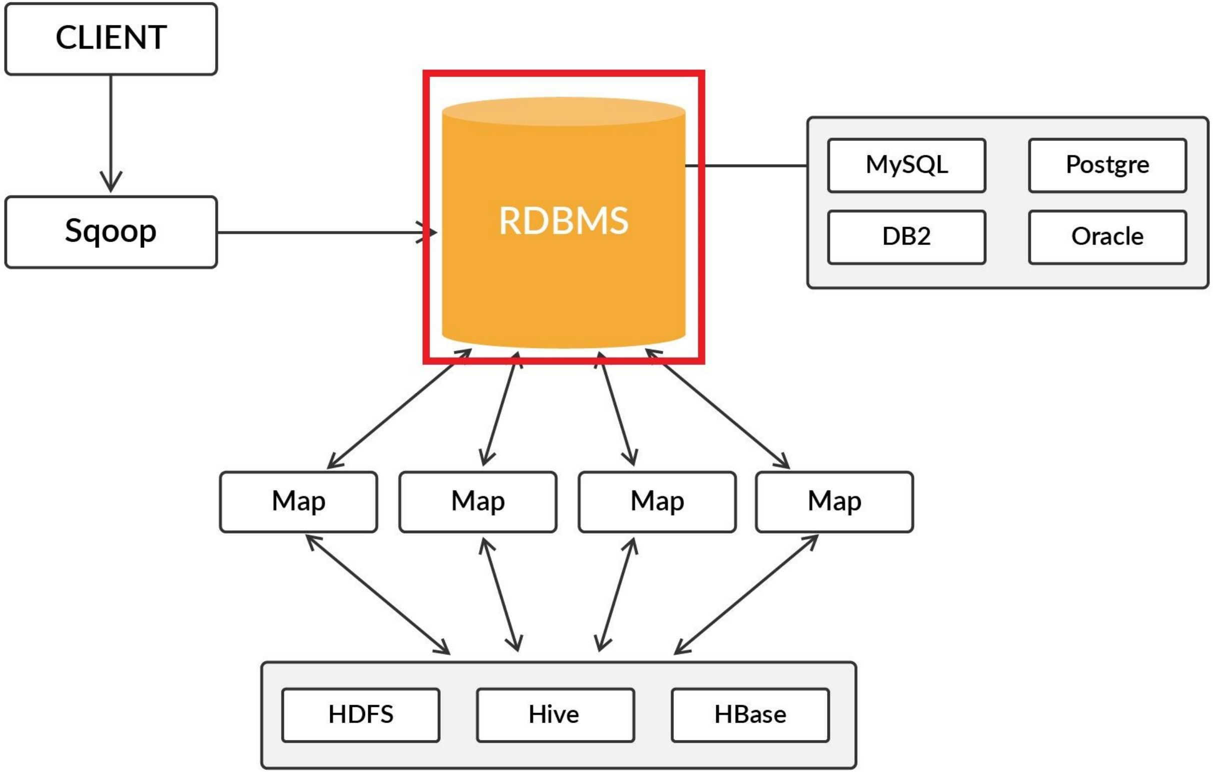


# Introduction to Sqoop and its Architecture

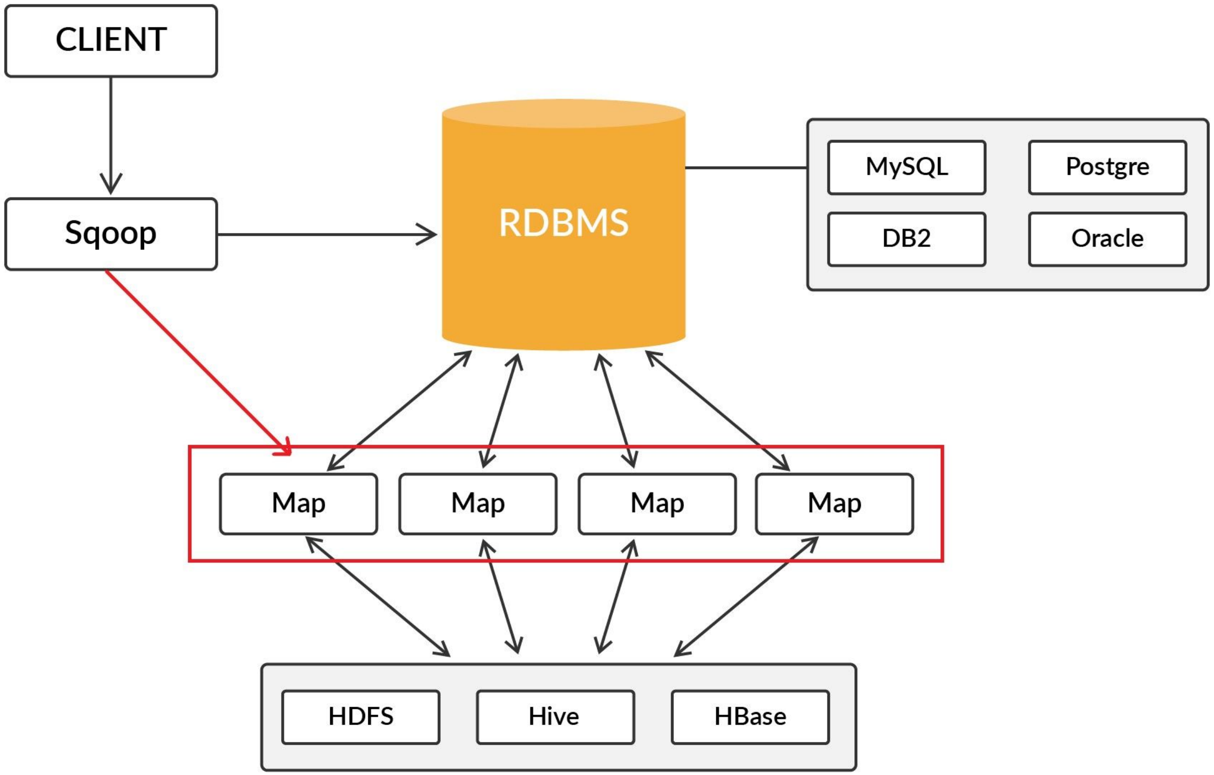
upGrad



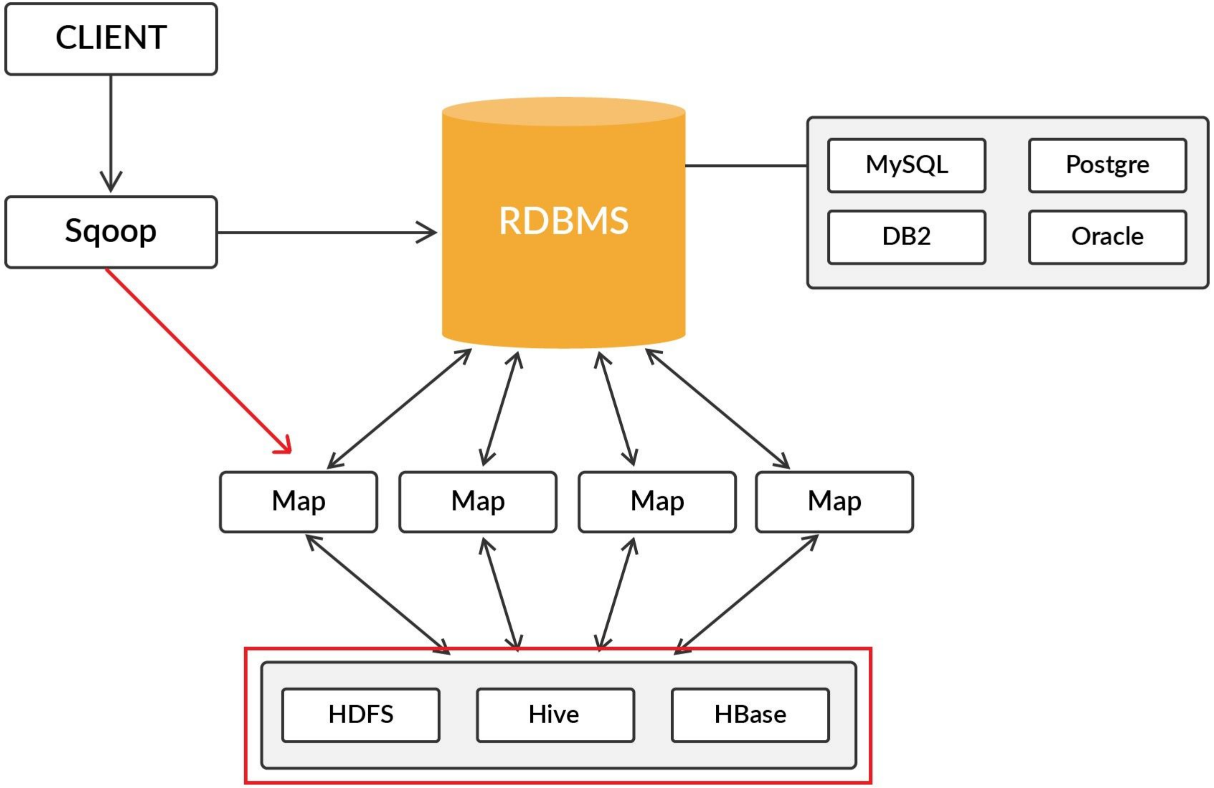
# Introduction to Sqoop and its Architecture



# Introduction to Sqoop and its Architecture



# Introduction to Sqoop and its Architecture



# Introduction to Sqoop and its Architecture

## Advantages of Sqoop

1. Provides stable and reliable interactions
2. Moves bulk data between Hadoop and RDBMSs efficiently
3. Performs tasks such as ETL
4. Cost-efficient

# Segment Summary

1. Discussed Apache Sqoop in brief
2. Discussed the architecture of Apache Sqoop
3. Learnt about the advantages of Apache Sqoop

# Segment - 03

Case Study 1: Flights Data Set

Case Study 2: Retail Data Set

Case Study 3: Employee Data Set

# Learning Objectives

1. Introduction to case studies used for Apache Sqoop
2. Data sets used in these case studies

# Case Study :- Flights Data Set

- The flights data set contains the following three columns: Destination, Origin and Count.
  1. The **destination** consists of the destination country name where the flight is going to land.
  2. The **origin** country name consists of the name of the country of origin from where the flight will take off.
  3. The **count** contains the number of flights between the specified destination and the name of the country of origin.

This flights data set could be used to perform various kinds of analysis as follows:

1. It could be used to determine the airports where maximum flights landed in a year.
2. It could be used to determine the airports from where maximum flights took off in a year.
3. This data set could be used to determine the busiest airport in a year.

# Case Study: Retail Data Set

- The retail data set contains the following eight columns:

  1. The **invoice number** is the column to track every invoice and all the orders associated with it.
  2. The **stock code** is basically the commodity code for a particular item.
  3. The **description** is a column that describes a commodity.
  4. The **quantity** is the amount of commodity available in stock or in the warehouse for sale.
  5. The **invoice date** is the date on which the invoice was generated.
  6. The **unit price** is the price of a single unit of that commodity.
  7. The **customer ID** is the ID of that customer who purchased the corresponding item.
  8. The **country** is the name of the country from the corresponding item is to be sold.

The online retail data set is maintained by a company in order to keep a track of all the goods being sold to its different customers across different countries.

# Case Study :- Employee Data Set

- The employee data set contains the following four columns:
    1. The **employee ID** is the primary key of this table that is used to keep track of every employee.
    2. The **first name** is the first name of the corresponding employee.
    3. The **designation** is the designation of that employee in the company.
    4. The **salary** is the salary paid by the company to the corresponding employee.

The employee data set is used, or this kind of pattern is usually maintained by different companies in order to keep track of all their employees. Further analyses could be performed on this data set as follows:

1. How many employees have a salary of less than a particular threshold?
2. How many employees belong to a designation?
3. What is the first name/last name of an employee belonging to some employee id value.

# Segment Summary

1. Case studies for Apache Sqoop were introduced
2. Data sets for these case studies were introduced

# Session Summary

1. Introduced Sqoop and learnt about its architecture and working
2. Introduced the case study on which the Sqoop commands will be used
3. Learned about the setup of Apache Sqoop
4. Introduced Sqoop Export and Import along with its various arguments and options using a case study
5. Used other options of the Import command for its customisation

# Thank You