# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic)CIS4560-01 Term Project Tutorial

#### Authors: Brian Bao, Karen Alvarez, Natalie Sanchez, Shadman Sayef

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 05/18/2025

**Lab Tutorial**

**U.S. Used Cars Dataset Analysis using Hadoop and Hive**

**Objectives**

In this hands-on lab, you will learn how to:

* Transfer and upload a large dataset (10GB) into a distributed file system (HDFS).
* Create external Hive tables with Beeline for full and simple schemas.
* Run HiveQL queries for statistical analysis.
* Export simplified results from HDFS to local machine.
* Use Microsoft Excel/PowerBI to visualize key trends.

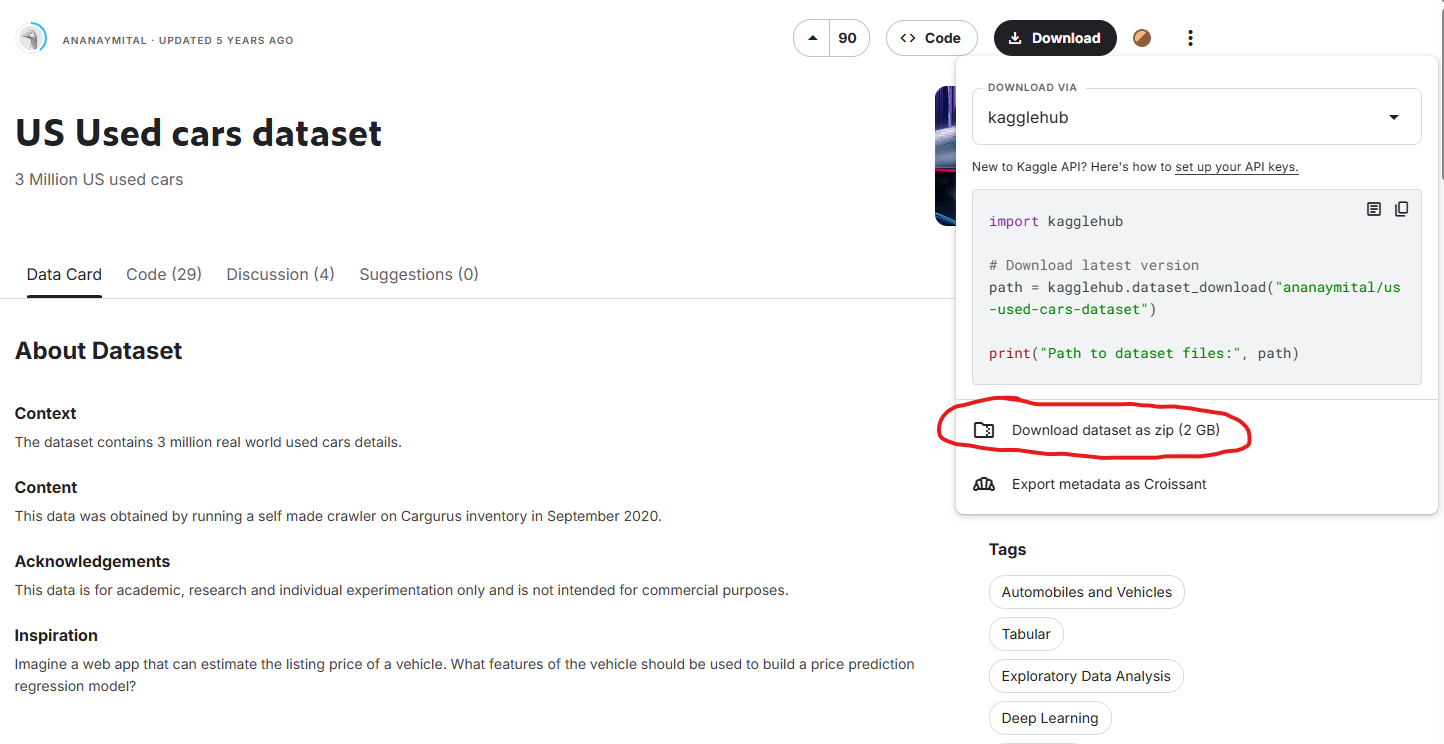
**Platform Spec**

* Hadoop Version 3.1.2
* CPU Speed: 2.45 GHz
* # of CPU cores: 6
* # of nodes: 5 (2 Master, 3 Worker)
* Total Memory Size: 31 GB

Step 1: Transfer Dataset from Kaggle to HDFS

This step is to obtain the unzipped .csv file in HDFS.

1. Download zipped dataset from Kaggle.



1. Transfer zipped dataset from local machine to Linux file system.

(scp used\_car\_data.zip your\_name@144.24.46.199:~)



1. Enter Linux file system with ssh.

(ssh [your\_name@144.24.46.199](mailto:your_name@144.24.46.199))



1. Unzip the dataset into a .csv file.

(unzip used\_car\_data.zip)



1. Create a directory in HDFS.

(hdfs dfs –mkdir UsedCars)



1. Upload unzipped .csv file to HDFS.

(hdfs dfs –put used\_cars\_data.csv UsedCars)



1. Verify upload to HDFS.

(hdfs dfs –ls UsedCars)



Step 2: Create Hive Tables

This step is to create Hive table schemas, one full and one simple for query processing and analysis.

1. Enter beeline.

(beeline)



1. Use your database.

(use your\_name;)



1. Create full table.

(CREATE EXTERNAL TABLE used\_cars\_full (

vin STRING,

back\_legroom FLOAT,

bed STRING,

bed\_height FLOAT,

bed\_length FLOAT,

body\_type STRING,

cabin STRING,

city STRING,

city\_fuel\_economy FLOAT,

combine\_fuel\_economy FLOAT,

daysonmarket INT,

dealer\_zip STRING,

description STRING,

engine\_cylinders INT,

engine\_displacement FLOAT,

engine\_type STRING,

exterior\_color STRING,

fleet BOOLEAN,

frame\_damaged BOOLEAN,

franchise\_dealer BOOLEAN,

franchise\_make STRING,

front\_legroom FLOAT,

fuel\_tank\_volume FLOAT,

fuel\_type STRING,

has\_accidents BOOLEAN,

height FLOAT,

highway\_fuel\_economy FLOAT,

horsepower INT,

interior\_color STRING,

isCab BOOLEAN,

is\_certified BOOLEAN,

is\_cpo BOOLEAN,

is\_new BOOLEAN,

is\_oemcpo BOOLEAN,

latitude FLOAT,

length FLOAT,

listed\_date STRING,

listing\_color STRING,

listing\_id STRING,

longitude FLOAT,

main\_picture\_url STRING,

major\_options STRING,

make\_name STRING,

maximum\_seating INT,

mileage INT,

model\_name STRING,

owner\_count INT,

power FLOAT,

price FLOAT,

salvage BOOLEAN,

savings\_amount FLOAT,

seller\_rating FLOAT,

sp\_id STRING,

sp\_name STRING,

theft\_title BOOLEAN,

torque STRING,

transmission STRING,

transmission\_display STRING,

trimId STRING,

trim\_name STRING,

vehicle\_damage\_category STRING,

wheel\_system STRING,

wheel\_system\_display STRING,

wheelbase FLOAT,

width FLOAT,

year INT

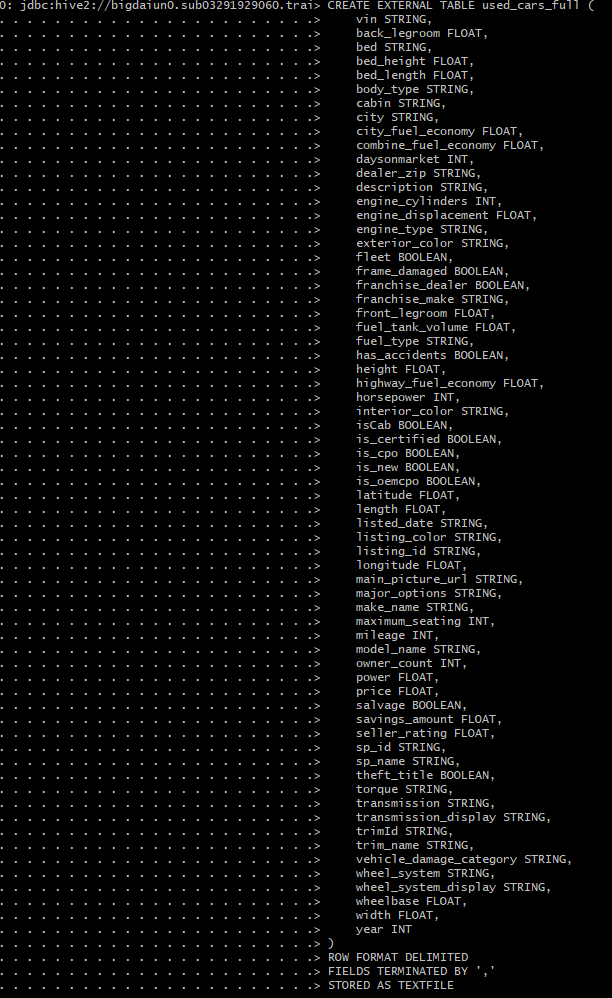
)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/your-user/UsedCars';)



1. Create simple table.

(CREATE EXTERNAL TABLE used\_cars\_simple (

vin STRING,

make\_name STRING,

model\_name STRING,

year INT,

price FLOAT,

mileage INT,

city STRING,

fuel\_type STRING,

body\_type STRING,

transmission STRING

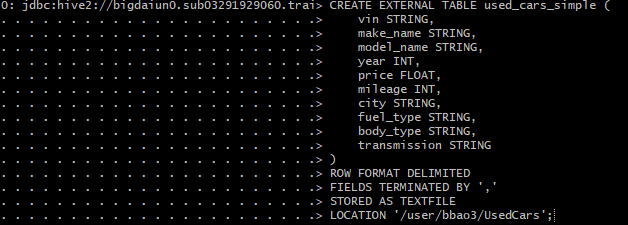
)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/your-user/UsedCars';)



Step 3: Prepare data for export to local machine.

This step is to obtain a cleaned .csv file for analysis and visualization in Excel/PowerBI.

1. Write Hive output to HDFS.

(INSERT OVERWRITE DIRECTORY '/user/your-user/UsedCars'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

SELECT

vin,

make\_name,

model\_name,

year,

price,

mileage,

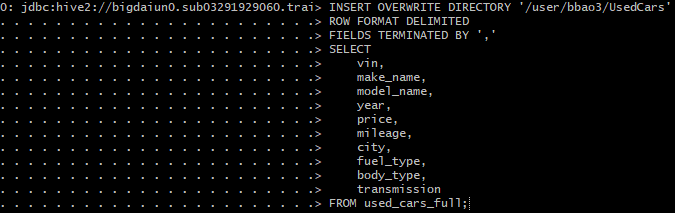
city,

fuel\_type,

body\_type,

transmission

FROM used\_cars\_full;)



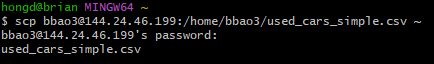
1. Merge created partitions into one cleaned .csv file.

(hdfs dfs -getmerge /user/your-user/UsedCars used\_cars\_simple.csv)



1. Download .csv file to local machine for analysis and visualization in Excel and PowerBI.

(scp [your\_user@your\_ip\_address:/home/your\_name/used\_cars\_simple.csv](mailto:your_user@your_ip_address:/home/your_local_machine/used_cars_simple.csv) ~)



References

* 1. URL of Data Source, <https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset>
  2. URL of your Github, <https://github.com/bbao3-csula/CIS-4560-Used-Cars>