Practical ETL application

Attached is a dataset represents the car theft happened in the US, each row in the dataset represents a the number of thefts happened to a car model (made by a car manufacturer)
Also attached another dataset of different car manufacturer with their country of origin

You are required to build a spark application to do the following,

- Build Hive table to maintain the cars.csv dataset
- The Read the theft dataset file, and extract a file which contains the car model (from the dataset) and the country of origin of this car, use hive table to extract the cars data.
- Do the needed pre-processing to done the previous step.
- Choose a partitioning column and partition your data accordingly, expect that the dataset input for your application to be in Gega bytes (we will collect the history of thefts within the years).
- Save the result dataset as a partitioned internal hive table.
- Use caching properly to optimize the performance, extract a DAG before and after the caching is done.
- Expect to read a file (from another location) with updated records of thefts, you should be able to merge these updates with the original dataset, considering the key of your dataset is a combination of all columns except the rank column.(attached sample file Updated - Sheet1.csv).
- Update your hive table with these updates.
- Extract a csv file contains the most 5 countries from where Americans buy their thefted cars ?

Important Notes:

- Use spark in cluster mode, make sure to optimize spark parallelism capabilities in your application
- Don't forget OOP principles while building your application.

Using Hive

- List the most 5 thefted models in U.S.
- List the most 5 states based on the number of thefted cars.