CORSO DI BIG DATA
Primo Progetto
7 maggio 2025

Si consideri il dataset **US Used cars dataset** di Kaggle[1], che contiene circa circa 3 milioni di record contenenti informazioni dettagliate di auto usate in vendita fino al 2020. Il dataset è in formato CSV e ogni record contiene 66 colonne riportate nella prossima pagina.

Dopo avere preparato opportunamente il dataset (per esempio eliminando dati errati o non significativi), progettare e realizzare almeno due delle seguenti analisi in almeno tre tra le seguenti tecnologie: MapReduce, Hive, Spark core e Spark SQL:

1. Un job che sia in grado di generare le statistiche di ciascuna marca di automobile (make_name) presente nel dataset indicando, per ogni marca: (a) il nome della marca e (b) una lista di modelli (model_name) per quella marca indicando, per ciascun modello: (i) il numero di auto presenti nel dataset, (ii) il prezzo (price) minimo, massimo e medio di auto di quel modello nel dataset e (iv) l'elenco degli anni in cui il modello è presente nel dataset.
2. Un job che sia in grado di generare un report contenente, per ciascuna città (city) e per ciascun anno (year): il numero di modelli di auto in vendita quell'anno appartenenti a tre fasce di prezzo (alto: sopra i 50K, medio: tra 20K e 50K, basso: inferiore a 20K) indicando, per ciascuna fascia, oltre al numero di auto in quella fascia, la media dei giorni di presenza delle auto sul mercato (daysonmarket) e le tre parole più frequenti che appaiono nella descrizione delle auto (description).
3. Un job in grado di generare gruppi di modelli di auto che hanno caratteristiche del motore "simili", ovvero per le quali i valori di potenza del motore (horsepower) cilindrata (engine_displacement) differiscono al più del 10%. Per ciascun gruppo va individuato il prezzo medio e modello dell'auto con maggiore potenza.

Per ciascun job bisogna illustrare e documentare in un rapporto finale:

- Le operazioni di preparazione dei dati che sono state eventualmente effettuate;
- La descrizione delle scelte implementative di ciascuna delle tecnologie scelte (testo e/o pseudocodice);
- Le prime 10 righe dei risultati dei vari job;
- Tabella e grafici di confronto dei tempi di esecuzione in locale e su cluster dei vari job con dimensioni crescenti dell'input[2] e, se possibile, del cluster.
- Link a un repository GitHub contenente il codice completo delle soluzioni.

Tutte le specifiche non definite in questo documento possono essere scelte liberamente.

Consegnare il rapporto **entro il 15 giugno 2025** in pdf sul sito moodle del corso disponibile (sperabilmente) all'indirizzo: https://ingegneria.el.uniroma3.it/course/view.php?id=386.

[1] https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset
[2] Si suggerisce di generare porzioni (per dimensioni più piccole del file) e copie (per dimensioni più grandi del file) per generare dataset di dimensione crescente.

Colonne del dataset US Used cars:

1. vin: Type String. Vehicle Identification Number is a unique encoded string for every vehicle.
2. back_legroom: Type String. Legroom in the rear seat.
3. bed: Type String. Category of bed size(open cargo area) in pickup truck. Null usually means the vehicle isn't a pickup truck
4. bed_height: Type String. Height of bed in inches
5. bed_length: Type String. Length of bed in inches
6. body_type: Type String. Body Type of the vehicle. Like Convertible, Hatchback, Sedan, etc.
7. cabin: Type String. Category of cabin size(open cargo area) in pickup truck. Eg: Crew Cab, Extended Cab, etc.
8. city: Type String. city where the car is listed. Eg: Houston, San Antonio, etc.
9. city_fuel_economy: Type Float. Fuel economy in city traffic in km per litre
10. combine_fuel_economy: Type Float. Combined fuel economy is a weighted average of City and Highway fuel economy in km per litre
11. daysonmarket: Type Integer. Days since the vehicle was first listed on the website.
12. dealer_zip: Type Integer. Zipcode of the dealer
13. description: Type String. Vehicle description on the vehicle's listing page
14. engine_cylinders: Type String. The engine configuration. Eg: I4, V6, etc.
15. engine_displacement: Type Float. engine_displacement is the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers.
16. engine_type: Type String. The engine configuration. Eg: I4, V6, etc.
17. exterior_color: Type String. Exterior color of the vehicle, usually a fancy one same as the brochure.
18. fleet: Type Boolean. Whether the vehicle was previously part of a fleet.
19. frame_damaged: Type Boolean. Whether the vehicle has a damaged frame.
20. franchise_dealer: Type Boolean. Whether the dealer is a franchise dealer.
21. franchise_make: Type String. The company that owns the franchise.
22. front_legroom: Type String. The legroom in inches for the passenger seat
23. fuel_tank_volume: Type String. Fuel tank's filling capacity in gallons
24. fuel_type: Type String. Dominant type of fuel ingested by the vehicle.
25. has_accidents: Type Boolean. Whether the vin has any accidents registered.
26. height: Type String. Height of the vehicle in inches
27. highway_fuel_economy: Type Float. Fuel economy in highway traffic in km per litre
28. horsepower: Type Float. Horsepower is the power produced by an engine.
29. interior_color: Type String. Interior color of the vehicle, usually a fancy one same as the brochure.
30. isCab: Type Boolean. Whether the vehicle was previously taxi/cab.
31. is_certified: Type Boolean. Whether the vehicle is certified. Certified cars are covered through warranty period
32. is_cpo: Type Boolean. Pre-owned cars certified by the dealer. Certified vehicles come with a manufacturer warranty for free repairs for a certain time period. Read more at https://www.cartrade.com/blog/2015/auto-guides/pros-and-cons-of-buying-a-certified-pre-owned-car-1235.html
33. is_new: Type Boolean. If True means the vehicle was launched less than 2 years ago.
34. is_oemcpo: Type Boolean. Pre-owned cars certified by the manufacturer. Read more at https://www.cargurus.com/Cars/articles/know_the_difference_dealership_cpo_vs_manufacturer_cpo
35. latitude: Type Float. Latitude from the geolocation of the dealership.
36. length: Type String. Length of the vehicle in inches
37. listed_date: Type String. The date the vehicle was listed on the website. Does not make days_on_market obsolete. The prices is days_on_market days after the listed date.
38. listing_color: Type String. Dominant color group from the exterior color.
39. listing_id: Unique Type Integer. Listing id from the website
40. longitude: Type Float. Longitude from the geolocation of the dealership.
41. main_picture_url: Type String.
42. major_options
43. make_name
44. maximum_seating
45. mileage
46. model_name
47. owner_count
48. power
49. price
50. salvage
51. savings_amount
52. seller_rating
53. sp_id
54. sp_name
55. theft_title
56. torque
57. transmission
58. transmission_display
59. trimId
60. trim_name
61. vehicle_damage_category
62. wheel_system
63. wheel_system_display
64. wheelbase
65. width
66. year