Firstly, the architecture will be built on a **cloud platform** that can address the **accessibility**, **scalability, flexibility, and security requirements** of the project. To determine the cost, a suitable costing strategy must be implemented, considering the utilization of the platform and tool preferences. We are also utilizing a structured analytical database that may support structured or unstructured formats like Big Query or Azure Synapse.

**ETL (Extract, Transform, Load) Processing** will be performed to ingest the data.

**Source Layer –** This is the stage where databases, files, APIs, or web pages are to be scraped and extracted. The data sources can be internal or external. Typically, an ETL Tool is recommended for easier maintenance and scheduling, including full loads, batch processing, near-real-time updates, and real-time integration. For fetching data from databases, files, and APIs, an ETL tool can be utilized. Additionally, custom scripts written in languages such as Python, Java, or C# can be employed to scrape data from web pages. Ingestion from internal sources (craigslist, user clickthrough, user profile), and external (car prices, external market, social media) can be easily ingested with proper access.

**Staging –** This is the stage where the extracted data is initially loaded. In some architectures, the data is dumped into files and then loaded into the Landing Zone Tables. From there, it is transferred to Transformation Tables, where it undergoes processing based on data profiling. At this stage, various data transformations, cleansing, de-duplication, and imputations can be performed to ensure that the data meets the defined Data Quality standards from the performed Data Profiling. Then finally pushed to Staging Tables before pushing to the final tables in the Data Warehouse.

**Data Warehouse –** This stage stores the Final Tables. Strategies may involve preserving the model architecture from the source and creating pipelines for it or models can already join tables already. The primary Fact Table should be at a granular level for detailed aggregation and calculations. It serves as the main data source for BAU reporting, data analytics, predictive modeling, and deep learning. Processed Data Models for Data Science requirements are also stored in the Data Warehouse. All data sources we have ingested can be combined in this area for further utilization.

**Data Marts –** the tables here can address security concerns, restrict object access, and simplify complexity. They provide pre-aggregated and calculated data based on specific requirements, improving performance, and allowing users to focus on analysis. A data refresh pipeline can be set up, incorporating dimensions such as region, manufacturer, or time. Alternatively, the baseline cube architecture can be adapted for analytical tools like SAS or SSAS, connecting with Power BI or Tableau for advanced data analysis.

We can integrate data from our Data Warehouse with other platforms to enhance functionality. For instance, by combining user data and available cars, we can improve our recommendation engine and share the results through APIs, adding value to the platform.

**Data Governance Strategy –** It is crucial to apply the following aspects to ensure effective governance. Data Classification: Understand data sensitivity and democratize its usage through data classification. Access Controls and Data Security: Determine user access levels and ensure secure data handling. Data Quality Management: Define rules for processing to ensure dataset accuracy, integrity, completeness, and consistency. Data Privacy: Comply with government regulations and protect individual privacy through data anonymization or pseudonymization. Data Retention and Archiving Policies: Determine data lifecycle based on business needs and policies. Comprehensive Data Documentation: Maintain comprehensive documentation, including data dictionaries, data lineage, and metadata, for platform usability and maintainability. Data Ethics and Responsible AI: Establish guidelines based on ethical frameworks (beneficence, autonomy, explicability, justice, non-maleficence) for data usage. Data Governance Committee: Implement these practices in the Data Warehouse with the oversight of a Data Governance Committee. Lastly, through training and enablement, we can achieve user democratization for the utilization of data.

Appendix A

