Harvesting, Storing & Retrieving data.

ADTA 5240 |IPAC 4240 |

**Project: Analyzing and monitoring daily automobile**

**sales in all areas to improve their business.**

**Executive Summary**

**Presented by Group 4**

**Arun Kumar Maram**

**Devika Vadlamudi**

**Harish Kumar Yamsani**

**Phani Kumar Kandepi**

**Sindhu Priya Keshireddy**

**Surya Teja Goud Keesara**

**Venkat Sai Vallala.**

**Executive Summary**

**Introduction:**

In Automotive Industry, Companies constantly analyze and monitor their daily sales and improve their business by applying new strategies and various methods with the help of previous data. Here, we have collected a particular data set and applied all the steps of the data life cycle for better improvement and decision-making. Our data set consists of entire information about the automobile industry its manufacturer, vehicle type, Model year, Its production, horsepower, acceleration, and so on, etc.,

**Explanation:**

As we all know, there are 8 phases in the data life cycle and here we applied every phase of the data life cycle to our data set for a simple understanding and to approach the business insights. The first and foremost phase of a data life cycle is “**Generation**”. Our data is generated from various sources in the Automobile industry. Once the data is generated, we moved to the 2nd phase of the data life cycle i.e., “**Data Collection**” and here we collected the static data set from the provided links that have 5170 rows which are already cleaned except for the minor errors and streaming data set from Twitter. The 3rd phase of the data life cycle is “**Processing**” where data gets cleaned into a structured format without errors and duplicates. Here, with the help of Open Refine, we removed all the minor errors (i.e., The model year in the 47th row has “Prelim 2021” instead of “2021”) and moved to the next phase i.e., “**Storage**”. We stored our cleaned data in the google cloud platform (GCP) for further evaluation.

We moved to the next phase of the data life cycle i.e., “**Management**” Here, we imported our data into BigQuery and performed query analysis for Business agility and insights. We executed Bigquery and checked over cluster monitoring details. Applied queries in Hive and Spark and compared the run-time processing for static data sets. For the streaming data set, we created an API that will fetch data from Twitter based on the criteria. We have written an API to fetch data related to automobiles and entered a query that is related to BMW cars between the start to end times from the given data. Upon the successful run of the API, we inserted the data into BigQuery of GCP and wrote a query to display only BMW data of the cars (sub-category) from the Automobiles (category) and exported the data. From here Data Analysts & Data Scientists analyze the exported data for further evaluations of “**Visualization**” and “**Interpretation**”.

**Summary:**

The project contains the whole process of the data life cycle and the industry takes decisions based on the outputs of the data for their business improvement and production growth. From the visuals of the data, the business concentrate and apply new strategies on low production areas. In this way, the Automobile Industry monitor their day-to-day production sales and expands its revenue.

**References:** [**https://online.hbs.edu/blog/post/data-life-cycle**](https://online.hbs.edu/blog/post/data-life-cycle)

[**https://www.epa.gov/automotive-trends/explore-automotive-trends-data**](https://www.epa.gov/automotive-trends/explore-automotive-trends-data)