Big Data Case Study on BMW Sales (2010–2024)

Name: Mohd Yousuf Arif

1. Introduction

In the era of digital transformation, data-driven decision-making has become the backbone of business strategy across industries. This case study focuses on analyzing **BMW's sales data from 2010 to 2024** using Big Data technologies such as **MySQL**, **Hive**, and **Hadoop**. The goal is to uncover insights regarding market performance, model popularity, fuel-type trends, and pricing evolution.

By leveraging **MySQL** for structured data storage, **Hive** for distributed querying, and **Hadoop** for scalable computation, this study demonstrates how data analytics can help BMW understand market dynamics and customer preferences.
The project also explores the shift toward hybrid and electric vehicles in BMW's portfolio and how this transition reflects broader market patterns.

---

2. Details of Dataset

The BMW Sales dataset provides a comprehensive view of the company's global performance between **2010 and 2024**.
It contains detailed information about model types, sales volumes, pricing, fuel categories, transmission types, and regions.
The dataset structure is designed for both SQL-based and Hive-based analytics.

Dataset Attributes

- **Year:** Year of sale (2010–2024).

- **Model:** BMW model name (e.g., X5, 3 Series, i8, M3).
- **Fuel_Type:** Petrol, Diesel, Hybrid, or Electric.
- **Transmission:** Manual or Automatic.
- **Engine_Size:** Engine displacement (liters).
- **Price:** Average selling price in USD.
- **Region:** Sales region (North America, Europe, Asia-Pacific, etc.).
- **Units_Sold:** Total units sold per model per year.
- **Color:** Most popular color for the model in that region.

[Figure 1: Dataset Schema – Placeholder for Diagram]

---

3. Project Scope

This project aims to analyze BMW's global sales data to understand business trends, consumer behavior, and market performance using Big Data technologies.
By combining the capabilities of **MySQL, Hive, and Hadoop**, the study explores key metrics such as annual sales, best-selling models, regional performance, and average pricing trends.

The scope extends to identifying patterns that indicate market shifts — particularly the rise of hybrid and electric vehicles — and providing a data-backed foundation for BMW's strategic planning.

---

4. Goals

1. Analyze yearly sales performance of BMW between 2010 and 2024.
2. Identify the **top 5 best-selling BMW models** globally.
3. Examine the shift from **combustion engines to electric and hybrid models**.
4. Evaluate pricing trends and transmission-type preferences.
5. Utilize **MySQL** and **Hive** to extract data insights efficiently.
6. Employ **Hadoop** for distributed storage and analysis of large datasets.
7. Present visual trends to improve understanding of BMW's business growth.

---

5. Tools and Working Environment

1. MySQL

Description:
MySQL is used to store and manage structured BMW sales data. It allows easy retrieval, aggregation, and manipulation of sales figures, models, and attributes.

Working Environment:
Acts as the primary relational database for loading and validating structured sales information before transferring it to Hive.

---

2. Hive

Description:
Hive is a SQL-like data warehouse tool built on top of Hadoop. It enables querying large datasets using **HiveQL** and is suitable for handling distributed sales records.

Working Environment:
Used to run analytical queries such as total yearly sales, average price by fuel type, and model-based comparisons.

---

3. Hadoop Ecosystem

Description:
Hadoop is an open-source framework that facilitates distributed storage (HDFS) and parallel data processing.

Working Environment:
Provides the backbone for scalable Big Data analysis of BMW's multi-year global sales data.

---

4. Sqoop

Description:
Sqoop is a data transfer tool used to move structured data from **MySQL** to **Hadoop**.

Working Environment:

It bridges MySQL and Hive, ensuring seamless integration of relational and Big Data environments.

Hive queries

01. Creating a table.

```
hive> Create table bmw_sale(model string, year int, region string, color string,
 fuel_type string, transmission string, engine_size string, mileage_km string, p
rice_USD int, sales_vol int, sales_classisfication string)Row format delimited f
ields terminated by ',';
OK
Time taken: 0.138 seconds
```

02. Loading the data into hive.

```
hive> LOAD DATA INPATH '/user/hive/warehouse/BMW_sales.csv' INTO TABLE bmw_sale;
Loading data to table default.bmw_sale
Table default.bmw_sale stats: [numFiles=1, totalSize=3392695]
OK
Time taken: 0.685 seconds
```

03. Total sale volume by year.

Insights:

- Reveals BMW's annual performance trends.
- Helps identify peak and low-performing years (e.g., post-2020 EV surge or 2021 pandemic dip).

```
hive> SELECT year, SUM(sales_vol) AS total_sales
    > FROM bmw_sale
    > GROUP BY year
    > ORDER BY year;
Total MapReduce CPU Time Spent: 4 seconds 500 msec
OK
NULL    NULL
2010    16933445
2011    16758941
2012    16751895
2013    16866733
2014    16958960
2015    17010207
2016    16957550
2017    16620811
2018    16412273
2019    17191956
2020    16310843
2021    16884666
2022    17920946
2023    16268654
2024    17527854
Time taken: 56.463 seconds, Fetched: 16 row(s)
```

04. Top 5 models by total sales.

Insights:

- Highlights the most popular BMW models globally.
- Useful for understanding consumer preferences and production focus.

```
hive>
    > SELECT model, SUM(sales_vol) AS total_sales
    > FROM bmw_sale
    > GROUP BY model
    > ORDER BY total_sales DESC
    > LIMIT 5;
OK
7 Series        23786466
i8      23423891
X1      23406060
3 Series        23281303
i3      23133849
Time taken: 45.81 seconds, Fetched: 5 row(s)
```

05. Sales distribution by fuel type.

Insights:

- Shows the shift from petrol/diesel to hybrid and electric.
- Indicates BMW's alignment with sustainability trends.

```
hive>
    > SELECT fuel_type, SUM(sales_vol) AS fuel_sales
    > FROM bmw_sale
    > GROUP BY fuel_type
    > ORDER BY fuel_sales DESC;
OK
Hybrid  64532097
Petrol  63324154
Electric        63157665
Diesel  62361818
Fuel_Type       NULL
Time taken: 46.684 seconds, Fetched: 5 row(s)
```

06. Average price by transmission type.

Insights:

- Compares pricing between manual and automatic variants.
- Likely shows higher average prices for automatic models due to demand and tech integration.

```
hive> SELECT transmission, ROUND(AVG(price_USD), 2) AS avg_price
    > FROM bmw_sale
    > GROUP BY transmission:
OK
Automatic       75171.41
Manual  74899.47
Transmission    NULL
Time taken: 23.169 seconds, Fetched: 3 row(s)
```

07. Sales classification breakdown.

Insights:

- Breaks down sales by model class (e.g., SUV, sedan, coupe).
- Useful for segment-wise strategy and marketing.

```
hive> SELECT sales_classification, COUNT(*) AS count
    > FROM bmw_sale
    > GROUP BY sales_classification;
OK
High    15246
Low     34754
Sales_Classification    1
Time taken: 22.15 seconds, Fetched: 3 row(s)
```

08. Most popular color by region.

Insights:

- Captures regional aesthetic preferences (e.g., white in Asia, black in Europe).
- Supports inventory and customization planning.

```
hive> SELECT region, color, COUNT(*) AS color_count
    > FROM bmw_sale
    > GROUP BY region, color
    > ORDER BY region, color count DESC;

OK
Africa  Grey    1400
Africa  Red     1398
Africa  White   1393
Africa  Blue    1369
Africa  Silver  1355
Africa  Black   1338
Asia    Black   1460
Asia    Grey    1422
Asia    Red     1409
Asia    Silver  1389
Asia    Blue    1388
Asia    White   1386
Europe  Black   1473
Europe  Blue    1385
Europe  Red     1381
Europe  White   1380
Europe  Grey    1365
Europe  Silver  1350
Middle East     Grey    1429
Middle East     Red     1427
Middle East     Silver  1421
Middle East     White   1389
Middle East     Blue    1375
Middle East     Black   1332
North America   Red     1461
North America   Silver  1435
North America   Grey    1379
North America   Blue    1378
North America   White   1350
North America   Black   1332
Region  Color   1
South America   White   1406
South America   Silver  1400
South America   Red     1387
South America   Blue    1367
South America   Grey    1353
South America   Black   1338
Time taken: 54.32 seconds, Fetched: 37 row(s)
```

09. Engine size vs avg price.

Insights:

- Correlates engine displacement with pricing.

- Larger engines likely command premium pricing; EVs may disrupt this trend.

```
hive> SELECT engine_size, ROUND(AVG(price_USD), 2) AS avg_price
    > FROM bmw_sale
    > GROUP BY engine_size
    > ORDER BY avg_price DESC;
OK
3.6     76263.54
4.2     76170.33
2.1     76116.76
3.9     75923.88
3.4     75800.0
4.4     75710.49
2.5     75667.32
4.7     75613.46
2.6     75416.54
3.8     75356.58
2.8     75336.09
3.5     75323.0
4.1     75313.01
2.4     75293.58
3.3     75277.59
4.9     75260.39
3.2     75227.32
4.6     75176.16
1.9     75174.27
3.0     75102.96
1.7     75102.73
2.0     74812.07
1.6     74801.1
3.7     74767.88
1.5     74680.55
2.2     74631.16
2.9     74621.15
2.7     74527.37
4.3     74455.97
4.0     74391.45
4.8     74313.78
1.8     74099.34
3.1     74072.15
2.3     74028.73
5.0     73261.76
4.5     73109.52
Engine_Size_L   NULL
Time taken: 60.168 seconds, Fetched: 37 row(s)
```

Conclusion

This Big Data case study successfully utilized MySQL, Hive, Hadoop, and Sqoop to analyze BMW's sales from 2010 to 2024.
The results revealed significant patterns, including a shift toward sustainable vehicles, a consistent preference for automatic transmissions, and price differentiation across regions.
The combination of Big Data technologies enables BMW to make informed decisions, optimize production strategies, and anticipate future trends in the automotive market.