The American Journal of Interdisciplinary Innovations and Research
ISSN 2642-7478 | Open Access

Type Original Research
PAGE NO. 16-27
DOI 10.37547/tajiir/Volume07Issue07-03

Check for updates

Hadoop To Bigquery: Migrating Automotive Data Lakes Without Downtime

Vrushali Parate
Department of Computer Science and Engineering, University of Bridgeport, Bridgeport, CT, 06604, USA

OPEN ACCESS

SUBMITED 29 May 2025
ACCEPTED 18 June 2025
PUBLISHED 07 July 2025
VOLUME Vol.07 Issue07 2025

CITATION
Vrushali Parate. (2025). Hadoop To Bigquery: Migrating Automotive Data Lakes Without Downtime. The American Journal of Interdisciplinary Innovations and Research, 7(07), 16–27. https://doi.org/10.37547/tajiir/Volume07Issue07-03

Abstract: The automotive industry is undergoing a tremendous increase in data generation, mostly driven by advancements in vehicle technology, connectivity, and autonomous driving features. The Apache Hadoop data lake was adopted by companies to store and analyze the huge volume, velocity, and variety of automotive data. However, with technological advancement and the need for real-time analytics, operational complexity, scalability, and cost efficiency, Apache Hadoop-based data lakes started presenting challenges. Google BigQuery, on the other hand, is a fully managed, serverless data warehouse and analytics platform that offers a good alternative with its scalable architecture, high performance, ease of use, and integration with advanced analytics and machine learning services. Migrating this massive amount of automotive data from Hadoop to BigQuery needs careful planning and execution, especially while making sure there are fewer disruptions with the ongoing business and avoiding downtime. This paper explores the typical architecture and use case of Hadoop-based data lakes in the automotive sector, explores BigQuery as an alternative option while also considering its benefits, and analyzes various strategies and methods for a seamless migration. Further, it delves into techniques and best practices for achieving zero downtime during the migration of large automotive datasets, addresses the specific challenges and considerations involved in handling automotive data's unique characteristics, examines relevant case studies of successful migrations, investigates methods for ensuring data consistency and integrity, and researches

approaches to optimize data processing and analytics workflows on BigQuery post-migration.

Keywords: Data Lake, Hadoop, BigQuery, Automotive Industry, Migration

1. INTRODUCTION

The automotive industry is undergoing substantial growth with the advancement of technology. Modern vehicles produce tremendous streams of data from telematics, sensors, connected vehicles, and over-the-air (OTA) updates, which in turn contribute to exponential data growth. In addition to connected vehicle data, automobile manufacturers and their tier 1 and tier 2 suppliers use dozens of IoT sensors in their manufacturing equipment to track procurement, production, and shipping processes, which is putting more fuel on the data explosion. This sudden growth is creating unprecedented data challenges, with electric vehicles (EVs) leading the way. Global Electric vehicle (EV) sales reached 17.1 million units in 2024, which is a 25% increase over 2023. The U.S. and Canada registered a 9% year-over-year growth in EV sales, whereas China's growth was 40% [1].

The rapid speed at which this data is being produced in different formats makes it necessary to have robust data management solutions. Data lakes have become a key part of how automotive companies manage their data. A data lake is a centralized repository that can store structured, semi-structured, and unstructured data in its raw format without requiring any transformation [2]. This flexibility to store the data in any format is especially important for the automotive industry since the data gets generated from various sources, such as sensors, telematics, manufacturing logs, part sales revenue, software updates, and multimedia. Data lakes are used by automotive manufacturers for numerous applications, such as predicting equipment maintenance needs, enhancing vehicle design and performance, fleet optimization, personalization, sales, and marketing. By providing a single platform for all enterprise data, data lakes free data silos and make it easier to perform end-to-end analytics, something that is important for getting an overview of automobile industry vehicle performance, customer behavior, and market trends.

Apache Hadoop was traditionally used as a data lake in the automotive industry and has played a significant role in the evolution of big data management. Cost-effectiveness, scalability, and fault-tolerant architecture were the reasons for adopting Apache Hadoop as a data lake by organizations. Hadoop enabled automotive organizations to store and process voluminous data using commodity hardware. However, with the expansion of technology and business needs, Hadoop-based data lakes have encountered limitations and challenges [3]. One of the limitations that Hadoop encountered was not being able to process data in real time. Hadoop was originally built for batch processing through the MapReduce framework, which makes it less ideal for scenarios that demand quick insights, like processing real-time data from connected vehicles. The surge in the use of connected vehicles, electric mobility, and autonomous systems has increased the importance of real-time data processing in the automobile industry. Applications including monitoring the battery health, real-time data analysis for autonomous driving algorithms, and predictive maintenance for critical vehicle components demand low latency and quick response time. However, Hadoop's batch-oriented architecture, with the added complexity of incorporating real-time layers, falls short in meeting these requirements. Moreover, maintaining a Hadoop ecosystem needs significant operational overhead, and a dedicated resource is required to perform regular system updates, performance tuning, and security configurations. These challenges may have led the organizations to explore cloud-native alternatives that offer managed infrastructure, seamless scalability, and integrated real-time analytics capabilities. Cloud platforms like Google BigQuery are increasingly preferred for their ability to support modern data processing needs.

Google BigQuery stands out as a strong alternative to Hadoop for handling and analyzing automotive data. As a fully managed, serverless data warehouse, it removes the need for infrastructure setup and management, letting automotive teams focus on using their data effectively [4]. BigQuery can process petabyte-scale datasets quickly using its distributed engine, making it ideal for the large data volumes generated in this industry. It also supports real-time streaming, which is critical for applications like connected vehicles—something Hadoop struggles with. Its pay-as-you-go

pricing and efficient storage can be more cost-effective than maintaining traditional Hadoop systems [4]. BigQuery also includes built-in tools for security and data governance, which are essential when working with sensitive vehicle and customer data [4]. For automotive companies, BigQuery offers a scalable, flexible, and faster path toward smarter analytics and innovation.

The primary research problem this paper aims to solve is the need for an end-to-end approach to migrate automotive data lakes from on-premises Apache Hadoop environments to Google BigQuery with zero downtime. This migration is critical for the automotive organizations to upgrade the data infrastructure by overcoming the shortcomings presented by Hadoop and leveraging the robust capabilities of BigQuery without disrupting existing operations.

This research paper aims to explore the challenges and opportunities involved in migrating automotive data lakes from Hadoop to Google BigQuery. It begins by analyzing the specific limitations of Hadoop in managing the growing volume, speed, and complexity of automotive data. The paper then explores the features of Google BigQuery that make it particularly suitable for this domain, such as its real-time processing capabilities, scalability, cost-effectiveness, and built-in data governance. Various migration strategies, including lift-and-shift, phased approaches, hybrid models, and full re-platforming are discussed to evaluate their suitability for different scenarios. Special attention is given to best practices that enable zero downtime during migration, which is especially critical for real-time data streams from connected vehicles. Real-world case studies from the automotive sector or related industries are examined to offer practical lessons. The paper also compares the traditional Hadoop-based architecture with BigQuery's cloud-native alternative, highlighting the advantages of the latter in terms of performance and maintainability. Finally, performance benchmarks and cost comparisons between the two platforms are presented to provide a data-driven assessment of BigQuery's value. Through these discussions, the paper serves as a comprehensive guide for automotive data professionals planning a seamless, zero-downtime migration to a modern, scalable data platform.

2. Hadoop Data Lake in the Automotive Industry

Figure. 1 represents a typical architecture of Hadoop in the automotive industry. It contains HDFS (Hadoop Distributed File System) as the primary storage layer, providing a scalable and fault-tolerant repository for large datasets [5]. YARN (Yet Another Resource Negotiator) typically serves as the cluster resource manager, which allocates system resources to various applications and jobs running on the Hadoop cluster [6]. From Figure. 1, the third layer represented the data processing framework layer, such as MapReduce and Apache Spark. MapReduce provides a programming model for processing large datasets in parallel, while Apache Spark provides faster processing due to the concurrent use of batch and stream processing [7].

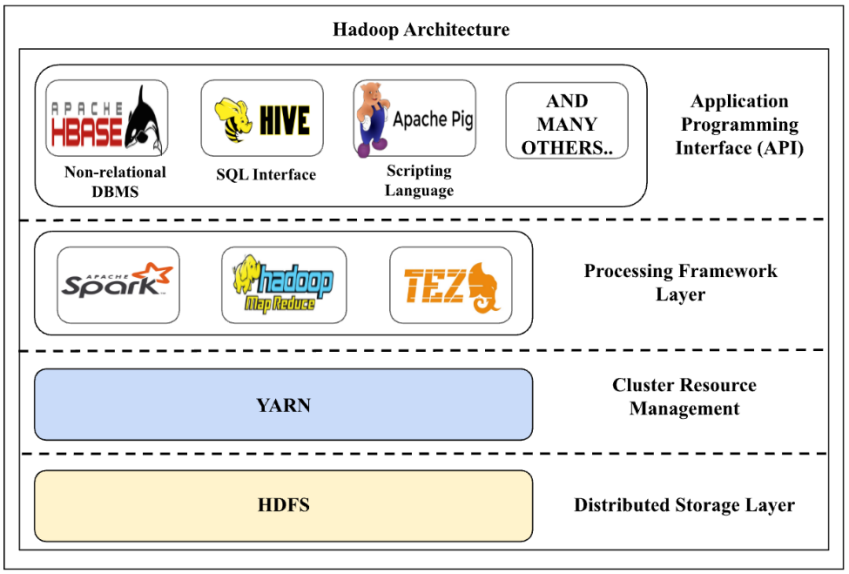

Figure. 1 Hadoop Architecture

The automotive industry generates a vast amount of data from vehicle sensors, software updates, multimedia, and IoT [8]. Sensor data from vehicles is ingested and analyzed to identify potential issues and predict failures. The development of autonomous driving depends on data lakes to store a variety of voluminous data generated from sensors like cameras and radar, which are important for training and validating autonomous algorithms. Connected vehicle data, sales, and service data help to understand customer preferences, personalize marketing efforts, and improve product development. Furthermore, predictive maintenance for manufacturing equipment is an important use case, where sensor data from the production line is analyzed to anticipate potential failures and minimize downtime.

Even though the Hadoop data lake offers diverse applications in the automotive industry, it began to show limitations in this data-dominated world. With this increase in volume of data, scalability can be a major concern. Hadoop architecture may struggle to handle real-time analytics efficiently [9]. Operational costs associated with managing Hadoop clusters and the need for specialized experts can be substantial. Furthermore, ensuring data quality and establishing robust governance mechanisms in Hadoop can be challenging. Security issues and the difficulty of setting up strong protection systems in large Hadoop setups are major concerns for the automotive industry, especially because they deal with sensitive vehicle and customer data.

The automotive industry depends heavily on Hadoop data lakes to manage a wide range of data-driven tasks, highlighting the importance of having a single place to store different types of data. However, Hadoop is starting to show its limits, especially when it comes to handling fast-moving data and the need for more flexible, cost-effective solutions. As managing large and complex automotive data becomes harder and the need for better data quality and governance grows, many companies are now looking to move to platforms like Google BigQuery. These newer tools are built to handle these challenges more efficiently.

3. Google BigQuery and Its Capabilities

Google BigQuery is a fully managed, serverless data warehouse that provides a scalable and cost-effective solution for analyzing data [4]. From Figure. 2, we can see that Google BigQuery contains separate compute and storage layers, which allows each layer to scale independently as needed. The storage system, called Colossus, is a global distributed file system that securely stores massive amounts of data [4]. Data is kept in a columnar format known as Capacitor, which speeds up analysis by allowing specific columns to be accessed without scanning entire rows [4]. BigQuery uses the Dremel query engine and a massively parallel processing (MPP) system to run SQL queries quickly, often processing terabytes of data in just seconds. A high-speed internal network connects storage and compute resources, ensuring fast and efficient data transfer [4].

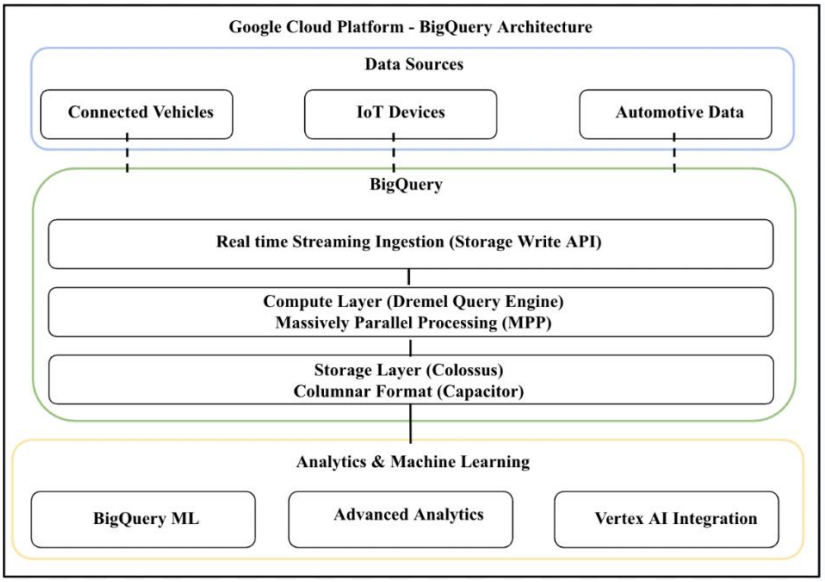

Figure. 2 Google BigQuery Architecture

BigQuery provides several features that make it especially useful for the automotive industry's data needs. It can easily scale to handle the growing amounts of data produced by connected vehicles without needing manual setup or capacity planning [4]. Because it's serverless, companies don't need to manage any servers or infrastructure, which allows their teams to focus more on analyzing data [4]. BigQuery also supports real-time streaming, so data from vehicles and IoT devices can be analyzed immediately. It works with various file formats like Avro, Parquet, JSON, and CSV, which is helpful given the range of data sources in automotive systems [4]. BigQuery integrates with Google Cloud's machine learning tools like Vertex AI and includes built-in machine learning (BigQuery ML), so teams can build models for things like predictive maintenance or detecting fraud directly within the platform. Its pay-as-you-go pricing is often more affordable than maintaining a Hadoop system, especially with tools like clustering and partitioning to make queries more efficient [4].

Compared to traditional data warehouses, BigQuery stands out for its scalability and flexibility, making it well-suited for the large and varied datasets commonly found in the automotive industry [4]. Unlike Hadoop, which requires manual setup and ongoing maintenance, BigQuery's serverless setup simplifies operations and lowers administrative effort. Its real-time data streaming also gives it an edge over Hadoop's batch processing, especially for time-sensitive use cases like vehicle telemetry and diagnostics [4]. While Hadoop is effective for storing large amounts of unstructured data, BigQuery is better optimized for running fast analytical queries and includes built-in tools for data governance and security. Its architecture is purpose-built to address many of the challenges seen with Hadoop, offering the scalability, low maintenance, and real-time capabilities needed for the automotive industry's evolving analytics demands. The built-in machine learning tools further support the development of intelligent solutions without requiring separate platforms [4].

The integration of machine learning capabilities provides an added advantage to the automotive industry. It allows data professionals to query or build ML models and deploy them quickly without the need to move the data to a separate platform, which is often the case with Hadoop. This built-in tool simplifies the process and helps speed up the process in the AI-driven world.

Table 1 shows an overview of the comparison between Hadoop and Google BigQuery based on key features that are important for building and managing the automotive data lakes. It looks at features such as scalability, real-time processing, ease of use, cost, machine learning support, and more. The table also explains how these differences are related to the automotive use cases, such as connected vehicles, data analysis, and predictive maintenance. This helps highlight why a platform like BigQuery might be better suited for the modern needs of the automotive industry.

Table 1: Comparison of Hadoop and BigQuery for Automotive Data Lakes

Feature	Hadoop	Google BigQuery	Why It Matters for Automotive
Scalability	Scales well with added servers	Automatically scales as needed	Both handle large vehicle and operational data, but BigQuery does it with less effort.
Real-time	Needs an extra tool	Real-time streaming	Essential for connected

Processing	like Spark Streaming	built-in	cars and real-time vehicle diagnostics.
Ease of Use	Complex setup and upkeep	Fully managed and serverless	Reduces IT workload - teams focus more on insights, not infrastructure.
Cost	Lower upfront (on-prem servers)	Pay for what you use (storage + compute)	BigQuery can be more budget-friendly for data-heavy tasks and has predictable pricing.
Data Formats	Works with most types of data	Supports formats like Avro, Parquet, JSON	Ideal for handling varied automotive data from sensors, logs, and more.
Machine Learning	Needs separate tools (e.g., Spark ML)	Built-in ML tools (BigQuery ML, Vertex AI)	Makes building models for maintenance, fraud, or customer targeting easier.
Query Performance	Slower for heavy analytics	Very fast on large datasets	Let analysts get quick insights - great for operations and field teams.
Data Governance	Mostly manual setup	Built-in controls with Dataplex integration	Supports privacy, compliance, and cleaner data management.

4. Exploring Various Methods of Migration

Migrating automotive data from one data lake to another requires proper planning and strategy that aligns well with the organization's needs, resources, and tolerance for disruption. There are several techniques available, which have their own advantages and disadvantages:

4.1 Lift and Shift: This is one of the most common strategies, which involves moving existing Hadoop infrastructure, including HDFS data and processing tools like Spark and Hive, to Google Cloud using Google Cloud Dataproc. Dataproc is a managed service that supports Hadoop and Spark workloads allowing teams to migrate their systems with fewer changes to their architecture. While this strategy involves minimal work, it might not fully leverage the benefits and cost efficiencies of BigQuery, as data processing remains in the Hadoop ecosystem [10].

4.2 Phased Migration: This strategy is performed in phases by dividing the data and workloads based on business units or datasets and then migrating from Hadoop to BigQuery, maintaining the Hadoop

environment for ongoing operations. While this migration takes place in phases, it is also crucial to test and validate data in BigQuery, which leaves less room for widespread disruption. This strategic migration enables teams and businesses to learn and adapt to BigQuery while ensuring continuity in the business before a full cutover. This migration can be started with less critical workloads, gradually moving towards more complex workloads as the confidence in working with the BigQuery environment increases [11].

4.3 Hybrid Approaches: Some companies may adopt this approach, where both Hadoop and BigQuery are utilized for different workloads or datasets depending on the use case. Hadoop can be used for data storage purposes, while BigQuery can be used to perform advanced analytics or high-performance workloads. Similar to phased migration, this helps companies to slowly adopt new data platforms such as BigQuery but simultaneously introduces more complexity related to managing two different platforms, which in turn can be resource-consuming [12].

4.4 Replatforming/Rearchitecting: This is a more detailed strategy that involves rearchitecting data models, optimizing processing pipelines, and adapting storage formats to align with BigQuery's architecture. Typically, this includes migrating data from HDFS to Google Cloud Storage, followed by loading it into BigQuery for advanced analytics and machine learning use cases. This would offer long-term value for achieving better performance, is cost-effective, and offers great flexibility in scaling data operations [10].

5. Step-by-Step Process for Planning and Executing Data Migration

Proper planning and execution steps are important to follow, irrespective of the approach followed by the automotive companies in terms of data migration. This involves several key phases:

5.1 Assessment and Planning: This is the first and most important step in data migration. This involves understanding the data in terms of volume, workloads, types, governance policies, and dependencies present in the existing Hadoop ecosystem. This step often includes a proof-of-concept (POC), which involves validating the migration strategy chosen and BigQuery's potential with a sample dataset. Clear goals and success metrics must be defined by automotive companies at this step of migration, which helps in proper resource allocation. Furthermore, appropriate Google Cloud services for different types of workloads need to be selected. For example, batch processing ETL jobs might be targeted for Dataproc and BigQuery, while streaming workloads using Kafka might be designated for Pub/Sub and Dataflow. Data warehousing workloads using Hive might be directly mapped to BigQuery.

5.2 Extraction, Transformation, and Loading of Data: After the completion of the first step, it is time to perform ETL of data. Irrespective of the migration strategy, this is a crucial step in the migration process. For extracting the data from the Hadoop Distributed File System (HDFS), several tools are available to perform this action, among which the most commonly used tool is the Hadoop Distcp (Distributed Copy) utility [13]. It supports distributed copying of large datasets across the environment and can be used to migrate historical data. The Google Cloud Storage connector for Hadoop can also be used by companies to transfer data between Hadoop clusters and Google Cloud Storage (GCS) directly [14]. For handling large-scale migration, Google's Transfer Service may also be used. For incremental loads and ongoing data synchronization, the gsutil command-line tool can be used. Some level of data transformation is often required, depending on the migration strategy and BigQuery's specific requirements [15]. This could potentially include adjusting schemas, cleaning datasets, and enriching the data to improve compatibility and performance. Tools like Google Cloud Dataflow can help build scalable pipelines to automate and manage these transformation tasks. Once the data is transformed, it is time to load the data into Google BigQuery. Typically, historical data, which is large in size, is loaded in batches, while real-time data from software updates or connected vehicle data is ingested using Google BigQuery's streaming systems. Google's BigQuery

Data Transfer Service can automate and streamline data loading directly from sources like Google Cloud Storage, which helps in reducing manual efforts and complexity [15].

5.3 Workflow and Application Migration: After migrating data, it is essential to migrate existing workflows and applications running on Hadoop. Apache Spark workloads running on Hadoop can be migrated to Dataproc. This may also involve rewriting scripts in SQL that were earlier in HiveQL or re-architecting applications using BigQuery's features and API. Workflow orchestration tools like Apache Oozie, commonly used in Hadoop environments, can be migrated to Google Cloud's Cloud Composer (Apache Airflow) [15].

5.4 Testing and Validation: With the entire migration process, it is important to test and validate data to maintain data integrity, accuracy, and quality. This includes comparing data between the Hadoop system and the Google BigQuery environment and testing the performance and reliability of the migrated analytics and reporting capabilities. Automated data validation tools can be useful in ensuring data synchronization and accuracy.

5.5 Cutover and Go-Live: After completion of the above step and making sure the data is accurate and reliable for use by the teams, the final transition takes place from the Hadoop data lake to Google BigQuery. This step is carefully planned to reduce disruption and redirect the traffic to BigQuery while decommissioning the Hadoop systems. A rollback plan should also be in place in case of any unexpected issues taking place during the final cutover.

While this migration takes place from Hadoop to BigQuery, there are certain considerations one needs to apply to different Hadoop components. For HDFS data, the primary method involves copying the data to GCS using tools like DistCp or gsutil. Hive metadata and schemas need to be migrated to BigQuery's data catalog, which can sometimes be done using tools that understand both Hive and BigQuery metadata formats. Spark applications often need to be adapted to run on Dataproc, which might involve changes in configuration or code [15]. Oozie workflows can be migrated to Cloud Composer, which provides a similar workflow orchestration capability using Apache Airflow. For real-time streaming data ingested via Kafka, the migration might involve setting up Google Cloud's Pub/Sub as the message broker and using Dataflow for stream processing and loading into BigQuery. Each of these component-specific migrations requires careful planning and execution to ensure a seamless transition [15].

6. Achieving Zero Downtime: Best Practices and Techniques

The following are some of the best techniques and practices that would help in achieving zero downtime during migration.

6.1 In-depth Analysis of Techniques for Continuous Data Availability: Migrating automotive data, which supports important operational and analytical processes, requires a strategy that guarantees continuous data availability with reduced disruptions to daily business processes. To achieve zero downtime during the transition process from Hadoop to BigQuery, organizations must plan thoroughly and adopt special techniques designed to maintain seamless operations throughout the migration process.

6.1.1 Blue/green Deployment: This is another strategy used to minimize downtime for the migration. It involves setting up and running two identical environments: the old system, Hadoop (blue), and the new system (green) in parallel [16]. Once the new environment - Google BigQuery, is tested and validated, the traffic can then be redirected from the blue environment to the green environment. This leaves enough time for the new environment to be thoroughly tested in a production-like setting before the actual transition takes place. In case of any unexpected issues, the traffic can always be diverted to the blue environment until the issues in the green environment are fixed and tested, ensuring minimal downtime. This method provides a safe and controlled way to migrate to BigQuery and roll back quickly, just in case [16].

6.1.2 Canary Deployment: This strategy is similar to blue/green deployment [16]. Canary releases are the process that can be performed in small portions by moving a small part of the application traffic to

the BigQuery environment, monitoring the performance, and identifying the potential issues before full transition [16]. This allows the industry to test BigQuery with a small amount of their workload and user base, which gives early warnings if there could be potential issues. If this strategy ensures success, gradually the traffic to BigQuery can be increased until the completion of migration. This approach reduces the risk of large-scale outages by allowing early issue detection before full migration takes place.

6.1.3 Data Replication Strategies: Using strong data replication tools and strategies is another crucial step toward achieving a zero-downtime migration. Tools like Cirata Data Migrator for Hadoop are specifically built to enable continuous, real-time replication of both data and Hive metadata from on-premises HDFS to cloud storage solutions like Google Cloud Storage, which BigQuery can then access directly [17]. These tools typically perform an initial one-time migration of the existing data, followed by ongoing replication of any new or updated data. This approach ensures that the BigQuery environment stays continuously synchronized with the live production system. By replicating the data in real-time, teams within the organization can reduce downtime during the transition. Once the migration is complete and verified, the switch to BigQuery can happen smoothly, without any major interruptions to business operations.

6.2 Specific Challenges and Considerations for Automotive Data Migration: With the focus on zero downtime on migrating the automotive data lake, it requires careful attention to the unique characteristics of automotive data. The high-volume data streams generated by connected vehicles make it critical to use BigQuery's streaming ingestion feature to ensure continuous, real-time data capture without any loss. BigQuery's scalable storage is also essential for managing the voluminous sensor data and telemetry from vehicles and manufacturing processes without compromising performance. Maintaining data consistency between the Hadoop and BigQuery environments is important, especially for critical automotive applications like safety systems, predictive maintenance, and autonomous driving development. To achieve this, teams within the automotive industry must implement strong data validation processes and use techniques such as dual-write patterns and continuous replication to ensure data reliability during migration. In addition, careful planning is needed when migrating data processing pipelines to BigQuery. These pipelines must be adapted to handle the specific formats, and processing demands of automotive data, all while ensuring that downstream applications continue to function smoothly without disruptions.

7. Case Studies of Successful Data Migration

Several case studies highlight successful migrations of data lakes to cloud-based solutions, which can provide valuable insights for automotive organizations. One such case study is about a leading telematics provider, Geotab [18]. It leveraged Google BigQuery to capture and analyze telematics data from over 1.4 million vehicles; processing more than 3 billion data records every day. Their ability to analyze raw sensor data in just 5 to 10 seconds underscores BigQuery's capability to handle the high velocity and volume of automotive data [18].

Uber, operating one of the world's largest Hadoop installations, managing over an exabyte of data, embarked on a strategic migration of its batch data analytics and machine learning training stack to Google Cloud Platform (GCP). The primary driver for this move was to enhance scalability, streamline operations, and eventually leverage the benefits of cloud-native services [19].

Uber's initial migration strategy involved a "lift and shift" approach, utilizing Google Cloud Storage (GCS) for their data lake storage and migrating the remainder of their data processing stack to GCP's Infrastructure as a Service (IaaS) [10]. This allowed them to replicate their existing on-premises environment on GCP with minimal disruption to ongoing workflows.

To facilitate this complex transition, Uber developed internal tools such as DataMesh, which helps manage cloud infrastructure and enforce data governance, and a Path Translation Service (PTS) to seamlessly map on-premises HDFS paths to their corresponding cloud-based locations [20].

As of a certain point in their migration journey, Uber had successfully moved over 160 petabytes of data to GCS and was running more than 19% of its analytical workloads on the GCP infrastructure. This migration positions Uber to take full advantage of GCP's elasticity and performance benefits in the future, ultimately creating a more agile, secure, and cost-efficient data ecosystem [20].

These case studies show successful Hadoop-to-cloud migration, which often employs a strategic phased approach. These allow incremental progress, validation at each stage, and reduced risk.

8. Ensuring Data Consistency and Integrity During and After Data Migration

Ensuring data consistency and integrity during and after data migration from Hadoop to BigQuery is important to maintain data accuracy and reliability for business operations. This involves the implementation of robust data validation and cleansing procedures at every step of the migration of automotive data in the BigQuery environment. This is an ongoing, proactive process to reduce errors and guarantee reliable data. Several methods can be implemented and used to maintain data accuracy at various stages of migration.

Before starting the migration, data profiling should be performed to examine the quality and characteristics of the data in the Hadoop data lake, identifying any inconsistencies, errors, duplicates, or missing values that could cause issues in the later process [21]. The next step in this process should be to perform data cleaning so that these identified issues can be corrected, ensuring only accurate and reliable data is migrated to BigQuery.

Implementing data validation protocols is one of the most important steps in the process of ensuring data consistency to help detect any discrepancies between the source and destination data [21]. Checksums or hash comparisons are the techniques used to verify the integrity of the transferred data to ensure no data is lost or corrupted during the migration process [21]. To ensure the correctness and completeness of the data migrated to the target system, a data reconciliation and comparison process is performed between Hadoop and BigQuery. For regular data validation, automated scripts can be written and executed throughout the migration process, which would quickly help in identifying any inconsistencies or mismatches in the data migrated [21].

Post-migration, it is essential to implement ongoing data quality monitoring and governance processes to maintain data integrity in the long term [21]. This includes processes such as setting up automated data quality checks, monitoring data pipelines for errors, and implementing data governance policies to ensure that data remains accurate and reliable over the period. Regular audits and validation checks should also be performed to identify and address any potential issues that may arise after the migration [21].

9. Optimizing Data Processes and Analytics in BigQuery

Optimizing data processes, workflows, and analytics post-successful migration of automotive data is important to fully leverage and utilize the features and functions provided by BigQuery to achieve optimal performance while being cost-efficient.

Designing data models and schemas is a fundamental step. Expensive JOIN operations can be replaced by nested and repeated fields. Using appropriate data types, such as INT64 for join keys, can help query performance and can be storage-efficient. Partitioning large tables by date or other relevant columns, such as vehicle identification number (VIN), can improve query performance and reduce costs by allowing BigQuery to only scan the necessary partitions.

Optimizing SQL queries is another important aspect of maximizing BigQuery’s performance. Best practices, such as using specific columns in SELECT statements, should be enforced instead of SELECT *, which can reduce the amount of data processed. Using proper WHERE conditions to filter data early in the query execution plan can help reduce cost and processing time to a great extent. Another best practice to improve query performance would be to use partitioning and clustering on frequently queried columns to process only relevant subsets of the data.

BigQuery offers advanced analytics and machine learning capabilities, such as BigQuery ML, which allow users to create and train machine learning models using SQL queries within BigQuery, reducing the need for a separate machine learning platform [22]. Geospatial

analysis functions can be used to analyze location-based data from connected vehicles for applications like route optimization and traffic analysis. The BigQuery BI engine is a fast, in-memory analysis service that can be used to accelerate business intelligence and reporting on BigQuery data, providing sub-second query response times for interactive dashboards and reports. Utilizing these features can unlock significant value from the migrated automotive data lake [22].

10. CONCLUSION

This research was able to explore the process of migrating automotive data lakes from Hadoop to Google BigQuery while being able to focus on achieving zero downtime. From the above analysis, it is known that Hadoop has served as a traditional data lake for managing voluminous data generated by the automotive sector, but its limitations in scalability, real-time processing, and complexity are becoming more evident. On the other hand, Google BigQuery, with its seamless architecture, independent scaling, and real-time streaming capabilities, presents a good alternative that directly addresses the shortcomings.

Considering this migration for automotive companies, several practical steps can be offered. Starting with a phased migration to help minimize risk by allowing teams to learn and adjust as they go. Using Google Cloud's managed services, such as Dataproc, Dataflow, and Cloud Composer, the transfer of existing workloads can be simplified. To avoid disruptions when going live, techniques such as blue-green or canary deployment to a new system can also help. The unique challenges involved in migrating large volumes of diverse and evolving data, given strict privacy and security regulations, should also be considered by the automotive industry. Validation checks are a must and should be implemented at every step of the process to maintain data quality. Finally, tuning data models and SQL queries in BigQuery post-migration will help boost performance and be cost-efficient.

A thorough analysis of the data migration journey, including Hadoop's shortcomings and Google BigQuery as a modern alternative, in the automotive industry has been demonstrated by this research paper. It also offers a roadmap with practical steps for companies to upgrade their data infrastructure. The focus on achieving zero or near-zero downtime forms an essential factor to maintain business continuity in a sector that increasingly depends on real-time data to drive efficiency, innovation, and competitive advantage.

REFERENCES

1. Rho Motion. (2025, January 14). *Over 17 million EVs sold in 2024 – record year*. Rho Motion. https://rhomotion.com/news/over-17-million-evs-sold-in-2024-record-year/
2. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12571-12590.
3. Singh, B., Verma, H. K., & Madaan, V. (2023). Performance Challenges and Solutions in Big Data Platform Hadoop. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *16*(9), 27-41.
4. Google Cloud. (n.d.). *Introduction to BigQuery*. Google.
5. Belov, V., & Nikulchev, E. (2021). Analysis of big data storage tools for data lakes based on apache hadoop platform. *International Journal of Advanced Computer Science and Applications*, *12*(8).
6. El Yazidi, A., Azizi, M. S., Benlachmi, Y., & Hasnaoui, M. L. (2021). Apache Hadoop-MapReduce on YARN framework latency. *Procedia Computer Science*, *184*, 803-808.
7. Ibtisum, S., Bazgir, E., Rahman, S. A., & Hossain, S. S. (2023). A comparative analysis of big data processing paradigms: Mapreduce vs. apache spark. *World Journal of Advanced Research and Reviews*, *20*(1), 1089-1098.
8. Ma, C., Zhao, M., & Zhao, Y. (2023). An overview of Hadoop applications in transportation big data. *Journal of traffic and transportation engineering (English edition)*, *10*(5), 900-917.
9. Alwaisi, S. S. A., Abbood, M. N., Jalil, L. F., Kasim, S., Fudzee, M. F. M., Hadi, R., & Ismail, M. A. (2021). A review on big data stream processing applications: contributions, benefits, and limitations. *JOIV: International Journal on Informatics Visualization*, *5*(4), 456-460.

10. Varma, K. M., & Se, G. B. (2022, August). Efficient scalable migrations in the cloud. In *2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* (pp. 3-6). IEEE.

11. Kansara, M. A. H. E. S. H. B. H. A. I. (2022). A structured lifecycle approach to large-scale cloud database migration: Challenges and strategies for an optimal transition. *Applied Research in Artificial Intelligence and Cloud Computing*, *5*(1), 237-261.

12. Hosseini Shirvani, M., Amin, G. R., & Babaeikiadehi, S. (2022). A decision framework for cloud migration: A hybrid approach. *IET software*, *16*(6), 603-629.

13. Apache Software Foundation. (n.d.). *DistCp: Hadoop distributed copy*. Hadoop.

14. Parthi, A. G., Pothineni, B., Jayabalan, D., Banarse, A. R., & Maruthavanan, D. (2024). Efficient Migration of Databases from Teradata to Google BigQuery: A Framework for Modern Data Warehousing. *Journal of Software Engineering (JSE)*, *2*(2), 55-64.

15. Evalueserve. (n.d.). *Modernizing financial data infrastructure: On-premises Hadoop migration to Google Cloud*.

16. Rudrabhatla, C. K. (2020, October). Comparison of zero downtime based deployment techniques in public cloud infrastructure. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 1082-1086). IEEE.

17. Cirata. (n.d.). *Data Migrator*.

18. Google Cloud. (n.d.). *Geotab: Driving innovation with Google Cloud*.

19. Uber Technologies. (2024, July 30). *Enabling security for Hadoop data lake on Google Cloud Storage*. Uber Blog.

20. Masolo, C. (2024, October 12). *Scaling Uber's batch data platform: A journey to the cloud with data mesh principles*. InfoQ.

21. Mohammad, N. (2021). Data integrity and cost optimization in cloud migration. *International Journal of Information Technology & Management Information System (IJITMIS)*, *12*, 44-56.

22. Google Cloud. (2025, May 5). *Introduction to AI and ML in BigQuery*.