Transforming Hotel Booking Insights: A Big Data Approach

Term Project - Final Project Submission

Shanthibooshan Subramanian
Bellevue University

Big Data DSC650
Nasheb Ismaily

ss

Table of Contents

1.0 Introduction

The hotel industry is rapidly evolving, and Big Data is playing a transformative role. It is reshaping traditional methods and revolutionizing how hotels manage reservations. The vast amounts of data generated by guests, booking platforms, and market dynamics present hotels with opportunities to gain valuable insights and make informed decisions.

(Figure 1: Data Analytics)

However, amidst this data deluge, hotels face significant challenges in harnessing the power of Big Data to manage reservations and address cancellations effectively. Data's sheer complexity and scale present hurdles in data collection, storage, processing, and analysis, requiring robust infrastructure and sophisticated analytical tools.

In this context, utilizing Big Data analytics holds immense promise for the hotel industry. By leveraging advanced analytical techniques, such as machine learning, predictive modeling, and data visualization, hotels can uncover actionable insights from vast datasets. These insights enable hotels to predict cancellations with greater accuracy, optimize pricing strategies in real time, and personalize guest experiences to drive loyalty and retention.

In this Project, we embark on a comprehensive exploration of the intersection between Big Data and reservation management in the hotel industry. Through an in-depth analysis of industry trends, best practices, and case studies, we seek to illuminate the transformative impact of Big Data analytics on reservation management and provide practical insights for hotels striving to thrive in an increasingly data-driven environment.

1.1 Problem Statement.

The hotel industry faces significant challenges in managing reservations and addressing cancellations, leading to revenue loss and operational difficulties. Traditional methods, such as overbooking, fail to address the unpredictability of guest behavior and external factors. Current reservation systems lack the advanced capabilities needed for comprehensive data analysis, which impedes proactive decision-making.

To overcome these challenges, there's a critical need to leverage data analysis techniques. By analyzing historical booking data, hotels can predict cancellations, optimize booking strategies, and enhance customer satisfaction. However, this requires overcoming technical and organizational hurdles, including data integration and talent acquisition. Embracing data-driven approaches is key to navigating reservation management complexities and maintaining competitiveness in the hospitality sector.

1.2 Significance and Relevance

This project tackles a critical issue in the hospitality industry: managing the growing complexity and volume of hotel bookings from multiple channels. Leveraging Big Data technologies is essential for streamlining operations and enhancing decision-making processes.

Big data analytics enhances operational efficiency by optimizing resource allocation, staffing, and inventory management based on booking patterns. Dynamic pricing strategies, informed by historical data and market trends, maximize revenue by adjusting room rates in real time.

Accurate forecasting through big data supports strategic planning, informing decisions about expansions, renovations, and market adaptation. This strategic foresight helps hotels stay competitive by identifying emerging trends and market shifts.

1.3 Importance/usefulness of solving the problem.

Analyzing hotel booking data is essential for revenue optimization, enhancing guest experiences, improving operational efficiency, refining marketing strategies, and aiding in forecasting and planning. By efficiently harnessing insights from booking data, hotels can maximize revenue, tailor services to guest preferences, manage resources effectively, target the right audience, and make informed strategic decisions. Leveraging big data tools for this analysis provides a significant competitive edge in the hospitality industry, driving growth and ensuring long-term success.

1.4 Stakeholder(s) alignment to proceed.

To proceed with implementing Big Data analytics for reservation management in the hotel industry, alignment among key stakeholders is crucial:

Hotel Management and Owners: Understand the benefits of Big Data for revenue optimization and operational efficiency.

IT and Data Teams: Provide necessary infrastructure and expertise for data collection and analysis.

Marketing and Sales Teams: Utilize data insights for targeted marketing and pricing strategies.

Guest Services and Operations: Implement changes based on data to enhance guest experiences and optimize operations.

External Partners and Vendors: Ensure seamless integration of data sources and compliance with regulations.

Clear communication, collaboration, and a shared understanding of the value of Big Data are essential for stakeholder alignment.

1.5 Dataset Overview.

The dataset used in this project comprises detailed hotel booking records, encompassing various attributes that provide comprehensive insights into reservation dynamics and customer behavior.

Here's a detailed summary of the categories of attributes/features included in the dataset:

- hotel: This attribute denotes the type of hotel, specifically whether it's a Resort Hotel.

- is_canceled: Indicates whether the booking was canceled, represented by binary values (0 for not canceled, 1 for canceled).

- lead_time: Represents the number of days between the booking date and the arrival date, providing insights into booking lead times.

- arrival_date_year/month/week_number/day_of_month: These attributes capture the temporal aspects of the arrival date, including the year, month, week number, and day of the month, facilitating temporal analysis of booking patterns.

- stays_in_weekend_nights: Indicates the number of weekend nights (typically Friday and Saturday) spent by the guests.

- stays_in_week_nights: Represents the number of weeknights (Sunday to Thursday) spent by the guests.
- adults/children/babies: These attributes specify the number of adults, children, and babies included in the booking, offering insights into guest demographics.

- meal: Describes the type of meal booked (e.g., breakfast, half board, full board), providing information on guest preferences and booking packages.

- country: Denotes the country of origin of the guests, facilitating analysis of international booking trends and guest diversity.

- market_segment: Indicates the market segment designation for the booking (e.g., corporate, group, online travel agent), offering insights into the distribution of bookings across different segments.

- distribution_channel: Specifies the booking distribution channel (e.g., direct, online travel agent, corporate), highlighting the source of bookings.

- is_repeated_guest: Indicates whether the booking is from a repeated guest, aiding in customer segmentation and loyalty analysis.

- previous_cancellations/previous_bookings_not_canceled: These attributes capture the history of cancellations and previous bookings not canceled, providing insights into guest behavior, and booking reliability.

- reserved_room_type/assigned_room_type: These attributes denote the code of the room type reserved and the room type assigned, respectively, offering insights into room allocation and preferences.

- booking_changes: Represents the number of changes made to the booking, reflecting booking modifications and flexibility.

- deposit_type: Describes the type of deposit made for the booking (e.g., no deposit, refundable deposit, non-refundable deposit).

- agent/company: These attributes represent the ID of the booking agent and the company/organization associated with the booking, facilitating partner analysis and attribution.

- days_in_waiting_list: Indicates the number of days on the waiting list before confirmation, highlighting demand fluctuations and booking constraints.

- customer_type: Specifies the type of customer (e.g., transient, contract, group), aiding in customer segmentation and marketing strategies.

- adr: Represents the Average Daily Rate, providing insights into pricing strategies and revenue management.

- required_car_parking_spaces: Denotes the number of car parking spaces required, reflecting guest preferences and hotel amenities.

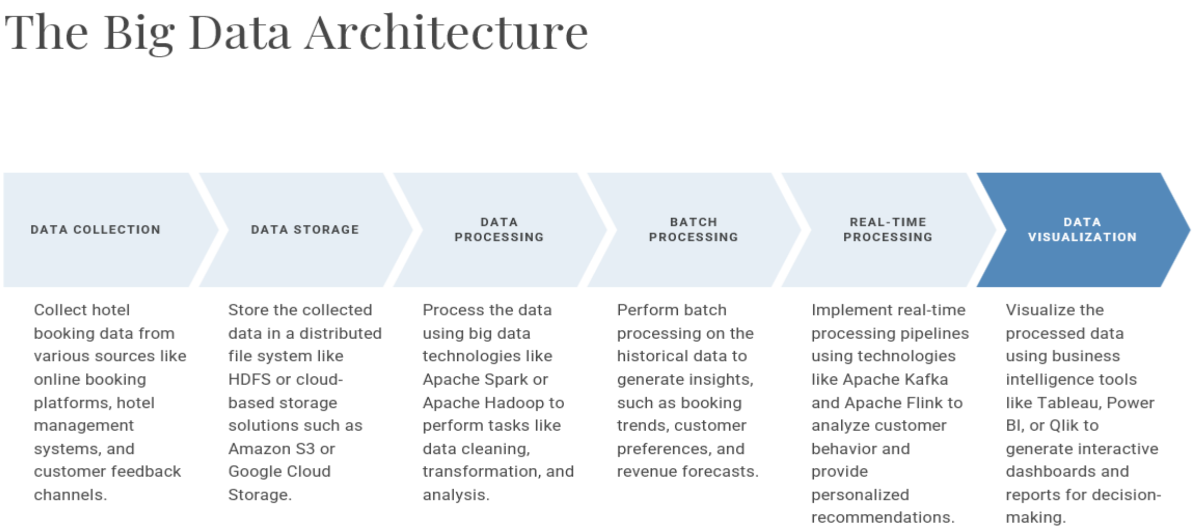
- total_of_special_requests: Indicates the total number of special requests made by guests, offering insights into service customization and guest satisfaction.

- reservation_status: Describes the reservation status (e.g., checked-in, canceled, no-show), reflecting the booking's lifecycle stage.

2.0 Big Data Technology Used

To effectively manage, process, and analyze the hotel booking data, a robust technological stack has been employed. The chosen technologies ensure efficient data storage, real-time data streaming, comprehensive data processing, and seamless data ingestion. Below are the details of the technologies used.

2.1 Cutting-Edge Big Data Architecture for Hotel Booking Analysis



The Big Data Architecture

DATA COLLECTION

Collect hotel booking data from various sources like online booking platforms, hotel management systems, and customer feedback channels.

DATA STORAGE

Store the collected data in a distributed file system like HDFS or cloud-based storage solutions such as Amazon S3 or Google Cloud Storage.

DATA PROCESSING

Process the data using big data technologies like Apache Spark or Apache Hadoop to perform tasks like data cleaning, transformation, and analysis.

BATCH PROCESSING

Perform batch processing on the historical data to generate insights, such as booking trends, customer preferences, and revenue forecasts.

REAL-TIME PROCESSING

Implement real-time processing pipelines using technologies like Apache Kafka and Apache Flink to analyze customer behavior and provide personalized recommendations.

DATA VISUALIZATION

Visualize the processed data using business intelligence tools like Tableau, Power BI, or Qlik to generate interactive dashboards and reports for decision-making.

2.2 Technology used

HDFS (Hadoop Distributed File System)

HDFS is utilized for storing large datasets securely and reliably. It provides a scalable and fault-tolerant storage solution, which is crucial for handling vast amounts of booking data generated from various sources. HDFS allows data to be distributed across multiple nodes, ensuring high availability and reliability. This distributed nature of HDFS makes it ideal for managing the large volumes of data involved in this project.

Apache Hive

Apache Hive is a vital component of our Big Data infrastructure, enabling efficient storage, querying, and analysis of large datasets stored in HDFS. Its SQL-like interface simplifies data management and allows for seamless integration with other Hadoop ecosystem tools. Through Hive, we organize our raw booking data into structured tables, facilitating easy access and management. Overall, Apache Hive empowers us to extract valuable insights and patterns from our hotel booking data, driving informed decision-making, and enhancing operational efficiency.

Apache Spark

Spark is employed for data processing and analysis due to its speed and efficiency in handling large-scale data. It offers in-memory computing capabilities, which significantly accelerates data processing tasks compared to traditional disk-based processing. Spark supports a variety of operations such as data cleaning, transformation, and complex analytical tasks, making it an ideal choice for extracting insights from the booking data.

Apache Kafka

Kafka is used for real-time data streaming, enabling the system to handle continuous data inflow from various booking channels. Kafka's distributed streaming platform ensures that data can be published and subscribed to in real time, providing a robust mechanism for real-time data processing and integration. This real-time capability is essential for maintaining up-to-date booking records and promptly responding to changes in reservation statuses.

Apache NiFi

NiFi is utilized for data ingestion, offering an easy-to-use interface for data flow management. It provides real-time data ingestion capabilities, allowing for the seamless and secure transfer of data from multiple sources into HDFS. NiFi’s ability to handle diverse data formats and its scalability ensures that data from various booking platforms and other sources are efficiently ingested into the system for further processing.

2.3 Detailed Methodology.

- **Data Ingestion with Apache NiFi**: The process begins with data ingestion using NiFi. Data from various booking channels, such as direct bookings, OTAs, and corporate partners, are collected in real-time. NiFi ensures the data is securely transferred and properly formatted before being stored in HDFS.

- **Data Storage in HDFS**: Once ingested, the data is stored in HDFS. The distributed nature of HDFS ensures that the data is stored reliably and can be accessed efficiently for processing. This step is crucial for managing the large volumes of booking data.

- **Data Organization and Querying with Apache Hive:** Apache Hive organizes the data into structured tables and simplifies querying and analysis, providing an SQL-like interface for users to interact with the data stored in HDFS, thereby enhancing efficiency, and enabling easy access to insights.

- **Data Processing and Analysis with Apache Spark**: Spark processes the ingested data, performing various tasks such as data cleaning, transformation, and complex analytics. The in-memory processing capabilities of Spark allow for rapid analysis and extraction of insights. This stage involves applying machine learning models to predict cancellations, optimize booking strategies, and personalize guest experiences based on the processed data.

- **Real-Time Data Streaming with Apache Kafka**: Kafka handles the real-time streaming aspect of the data, ensuring that any updates or new bookings are immediately available for processing. This continuous data flow is critical for maintaining an up-to-date dataset and enabling real-time analytics.

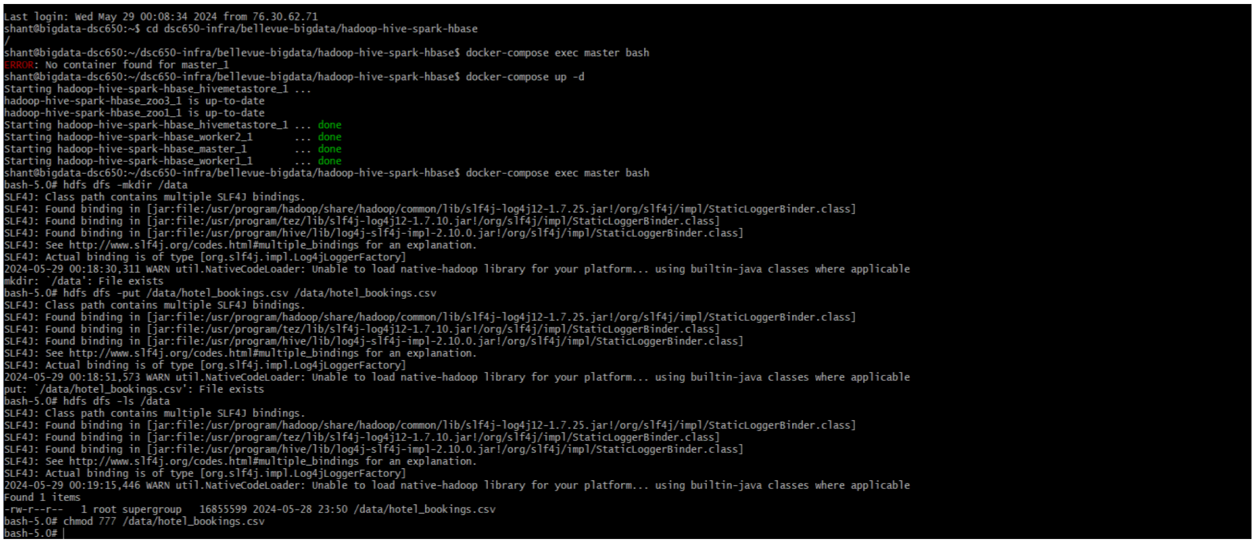
- **Insight Generation and Decision Support**: The final step involves generating actionable insights from the processed data. These insights help in predicting cancellations, optimizing pricing strategies, improving customer satisfaction, and supporting strategic decision-making. The results are used to inform hotel management and operations, ensuring data-driven decisions enhance overall performance.

2.4 Seamless Integration and Implementation of Big Data Solutions

ROBUST DATA STORAGE IN HDFS (Hadoop Distributed File System)

In the Hotel Booking project, we utilized the Hadoop Distributed File System (HDFS) to efficiently manage and store large volumes of booking data. Initially, we created a structured directory in HDFS, which allowed us to organize our data files systematically. We then securely uploaded the hotel_bookings.csv dataset into this HDFS directory. This ensured that the booking data was centralized and stored reliably. To verify the successful upload, we used the command hdfs dfs -ls /data, which listed the contents of the directory, confirming the presence of the hotel_bookings.csv file.

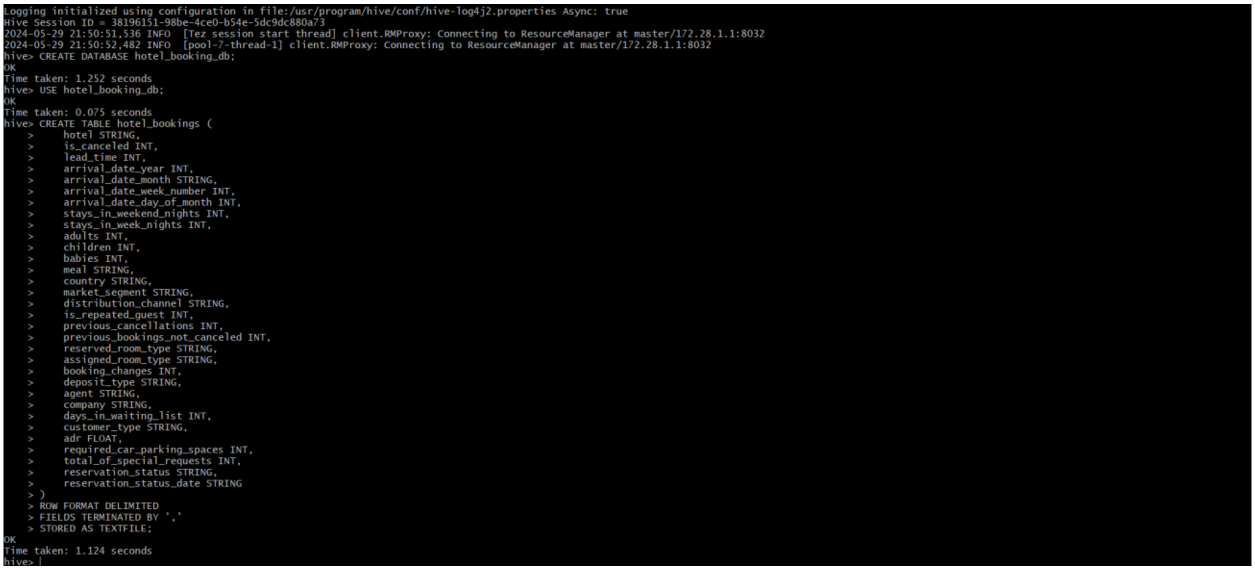
HDFS provided key benefits for Big Data utilization. Its distributed architecture ensured scalability, efficiently handling increasing volumes of booking data. By replicating data across multiple nodes, HDFS guarantees high availability and fault tolerance, even in case of hardware failures. Its high throughput capabilities enabled efficient data to read and write operations, ideal for large-scale processing tasks. Additionally, HDFS seamlessly integrated with tools like Apache Spark and Hive, enhancing data processing and analysis. Leveraging HDFS, we established a robust, scalable, and reliable data storage solution essential for comprehensive Big Data analytics in the project.



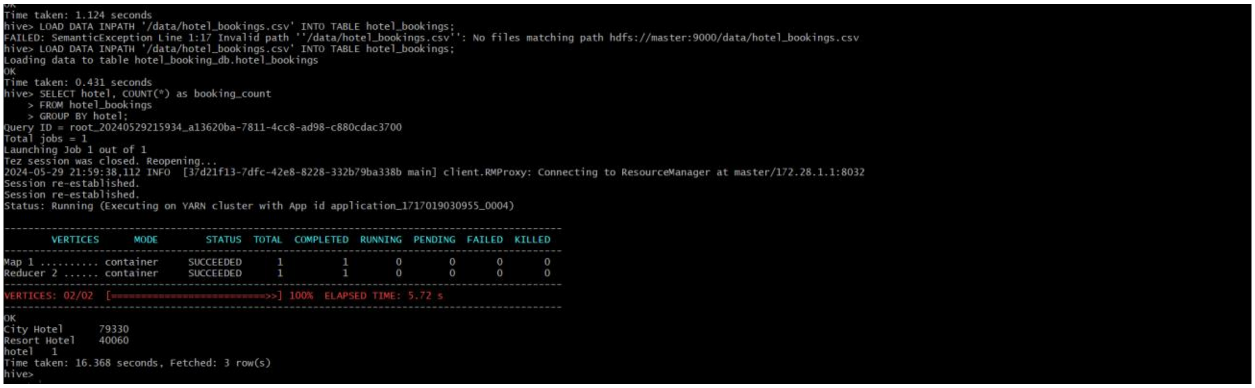
```
Last login: Wed May 29 00:08:34 2024 from 76.30.62.71
shant@bigdata-dsc650:~$ cd dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
/
shant@bigdata-dsc650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose exec master bash
ERROR: No container found for master_1
shant@bigdata-dsc650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose up -d
Starting hadoop-hive-spark-hbase_hivemetastore_1 ...
hadoop-hive-spark-hbase_zoo3_1 is up-to-date
hadoop-hive-spark-hbase_zoo1_1 is up-to-date
Starting hadoop-hive-spark-hbase_hivemetastore_1 ... done
Starting hadoop-hive-spark-hbase_worker2_1       ... done
Starting hadoop-hive-spark-hbase_master_1        ... done
Starting hadoop-hive-spark-hbase_worker1_1       ... done
shant@bigdata-dsc650:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose exec master bash
bash-5.0# hdfs dfs -mkdir /data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-05-29 00:18:30,311 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: `/data': File exists
bash-5.0# hdfs dfs -put /data/hotel_bookings.csv /data/hotel_bookings.csv
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-05-29 00:18:51,573 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: `/data/hotel_bookings.csv': File exists
bash-5.0# hdfs dfs -ls /data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-05-29 00:19:15,446 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 root supergroup   16855599 2024-05-28 23:50 /data/hotel_bookings.csv
bash-5.0# chmod 777 /data/hotel_bookings.csv
bash-5.0# |
```

EFFICIENT DATA ORGANIZING AND QUERYING APACHE HIVE

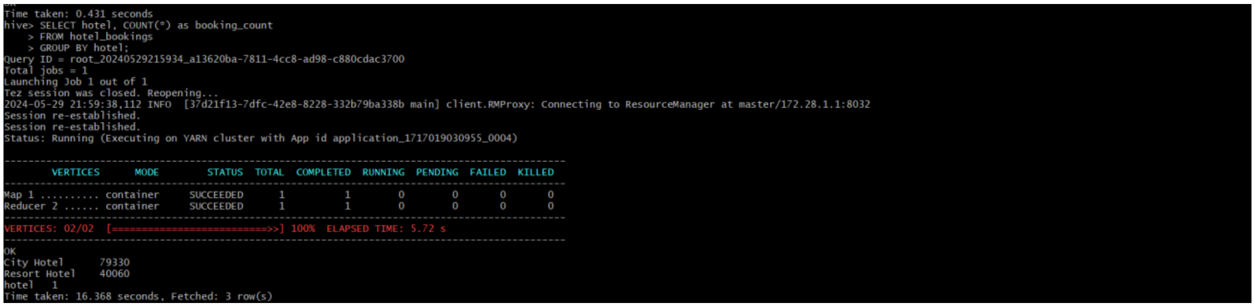
Apache Hive can significantly enhance data analysis and querying capabilities. Integrating will streamline the process of managing and querying Hotel booking's large datasets stored in HDFS. Hive's SQL-like query language simplifies data exploration and transformation, while its integration with Spark enhances analytical capabilities. This setup ensures a scalable, reliable, and efficient data processing environment, crucial for comprehensive Big Data analytics.



```
Logging initialized using configuration in file:/usr/program/hive/conf/hive-log4j2.properties Async: true
Hive Session ID = 38196151-98be-4ce0-b54e-5dc9dc880a73
2024-05-29 21:50:51,536 INFO  [Tez session start thread] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
2024-05-29 21:50:52,482 INFO  [pool-7-thread-1] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
hive> CREATE DATABASE hotel_booking_db;
OK
Time taken: 1.252 seconds
hive> USE hotel_booking_db;
OK
Time taken: 0.075 seconds
hive> CREATE TABLE hotel_bookings (
    >     hotel STRING,
    >     is_canceled INT,
    >     lead_time INT,
    >     arrival_date_year INT,
    >     arrival_date_month STRING,
    >     arrival_date_week_number INT,
    >     arrival_date_day_of_month INT,
    >     stays_in_weekend_nights INT,
    >     stays_in_week_nights INT,
    >     adults INT,
    >     children INT,
    >     babies INT,
    >     meal STRING,
    >     country STRING,
    >     market_segment STRING,
    >     distribution_channel STRING,
    >     is_repeated_guest INT,
    >     previous_cancellations INT,
    >     previous_bookings_not_canceled INT,
    >     reserved_room_type STRING,
    >     assigned_room_type STRING,
    >     booking_changes INT,
    >     deposit_type STRING,
    >     agent STRING,
    >     company STRING,
    >     days_in_waiting_list INT,
    >     customer_type STRING,
    >     adr FLOAT,
    >     required_car_parking_spaces INT,
    >     total_of_special_requests INT,
    >     reservation_status STRING,
    >     reservation_status_date STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 1.124 seconds
hive> |
```



```
Time taken: 1.124 seconds
hive> LOAD DATA INPATH '/data/hotel_bookings.csv' INTO TABLE hotel_bookings;
FAILED: SemanticException Line 1:17 Invalid path ''/data/hotel_bookings.csv'': No files matching path hdfs://master:9000/data/hotel_bookings.csv
hive> LOAD DATA INPATH '/data/hotel_bookings.csv' INTO TABLE hotel_bookings;
Loading data to table hotel_booking_db.hotel_bookings
OK
Time taken: 0.431 seconds
hive> SELECT hotel, COUNT(*) as booking_count
    > FROM hotel_bookings
    > GROUP BY hotel;
Query ID = root_20240529215934_a13620ba-7811-4cc8-ad98-c880cdac3700
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-05-29 21:59:38,112 INFO  [37d21f13-7dfc-42e8-8228-332b79ba338b main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1717019030955_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.72 s
----------------------------------------------------------------------------------------------
OK
City Hotel      79330
Resort Hotel    40060
hotel   1
Time taken: 16.368 seconds, Fetched: 3 row(s)
hive>
```

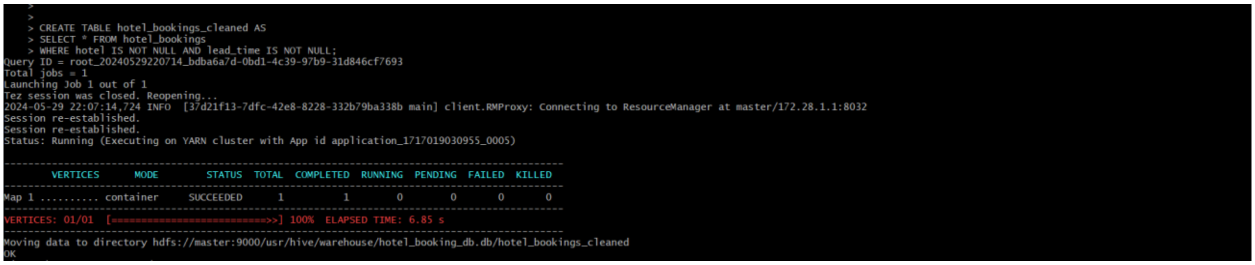


```
Time taken: 0.431 seconds
hive> SELECT hotel, COUNT(*) as booking_count
    > FROM hotel_bookings
    > GROUP BY hotel;
Query ID = root_20240529215934_a13620ba-7811-4cc8-ad98-c880cdac3700
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-05-29 21:59:38,112 INFO  [37d21f13-7dfc-42e8-8228-332b79ba338b main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1717019030955_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.72 s
----------------------------------------------------------------------------------------------
OK
City Hotel      79330
Resort Hotel    40060
hotel   1
Time taken: 16.368 seconds, Fetched: 3 row(s)
```

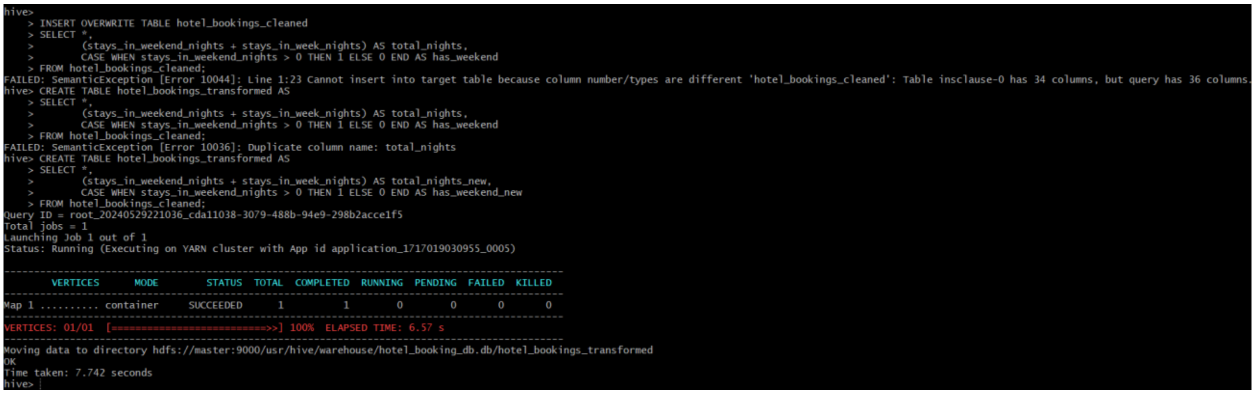
Apache Hive Data Transformation

To transform the data, we clean the data by removing or handling missing, inconsistent, or incorrect booking information to ensure data quality. We do this by creating a new table called hotel_bookings_cleaned and selecting only the rows from the original table where the hotel and lead_time columns are not null. We then alter the lead_time column to ensure it is in the correct integer format. Additionally, we add two new columns, total_nights and has_weekend, to the hotel_bookings_cleaned table.

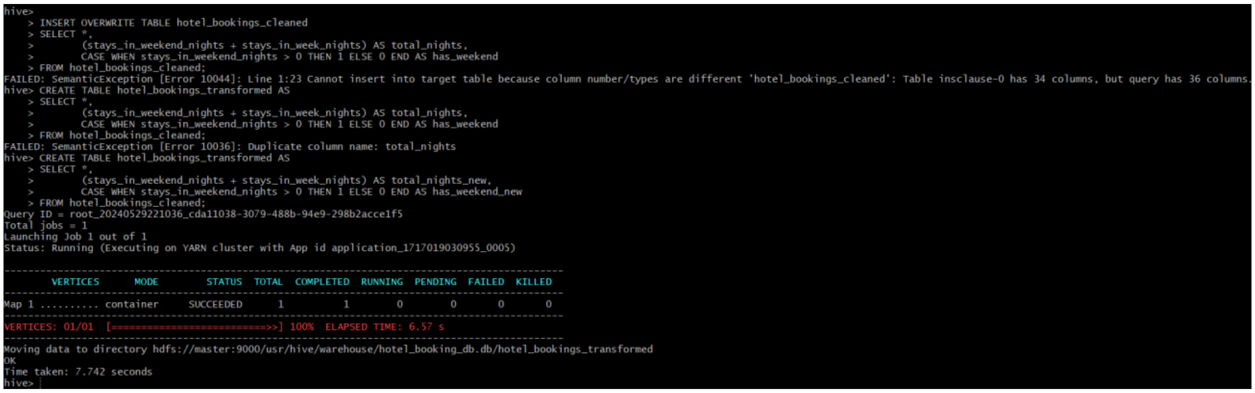
We created a new table called hotel_bookings_transformed, which is a transformed version of the hotel_bookings_cleaned table. The new table includes all the columns from the original table, plus two additional columns: total_nights_new, which is the sum of stays_in_weekend_nights and stays_in_week_nights, and has_weekend_new, which is 1 if stays_in_weekend_nights is greater than 0, and 0 otherwise.



```
    >
    > CREATE TABLE hotel_bookings_cleaned AS
    > SELECT * FROM hotel_bookings
    > WHERE hotel IS NOT NULL AND lead_time IS NOT NULL;
Query ID = root_20240529220714_bdba6a7d-0bd1-4c39-97b9-31d846cf7693
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-05-29 22:07:14,724 INFO  [37d21f13-7dfc-42e8-8228-332b79ba338b main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.85 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://master:9000/usr/hive/warehouse/hotel_booking_db.db/hotel_bookings_cleaned
OK
```



```
hive>
    > INSERT OVERWRITE TABLE hotel_bookings_cleaned
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend
    > FROM hotel_bookings_cleaned;
FAILED: SemanticException [Error 10044]: Line 1:23 Cannot insert into target table because column number/types are different 'hotel_bookings_cleaned': Table insclause-0 has 34 columns, but query has 36 columns.
hive> CREATE TABLE hotel_bookings_transformed AS
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend
    > FROM hotel_bookings_cleaned;
FAILED: SemanticException [Error 10036]: Duplicate column name: total_nights
hive> CREATE TABLE hotel_bookings_transformed AS
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights_new,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend_new
    > FROM hotel_bookings_cleaned;
Query ID = root_20240529221036_cda11038-3079-488b-94e9-298b2acce1f5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.57 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://master:9000/usr/hive/warehouse/hotel_booking_db.db/hotel_bookings_transformed
OK
Time taken: 7.742 seconds
hive>
```



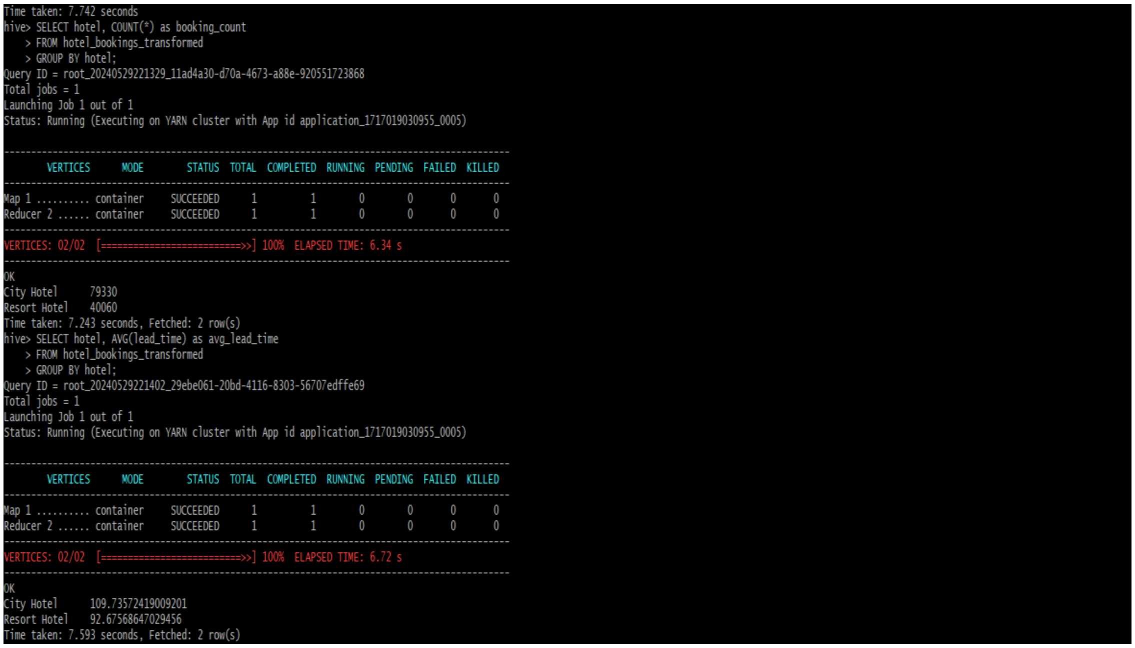
```
hive>
    > INSERT OVERWRITE TABLE hotel_bookings_cleaned
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend
    > FROM hotel_bookings_cleaned;
FAILED: SemanticException [Error 10044]: Line 1:23 Cannot insert into target table because column number/types are different 'hotel_bookings_cleaned': Table insclause-0 has 34 columns, but query has 36 columns.
hive> CREATE TABLE hotel_bookings_transformed AS
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend
    > FROM hotel_bookings_cleaned;
FAILED: SemanticException [Error 10036]: Duplicate column name: total_nights
hive> CREATE TABLE hotel_bookings_transformed AS
    > SELECT *,
    >        (stays_in_weekend_nights + stays_in_week_nights) AS total_nights_new,
    >        CASE WHEN stays_in_weekend_nights > 0 THEN 1 ELSE 0 END AS has_weekend_new
    > FROM hotel_bookings_cleaned;
Query ID = root_20240529221036_cda11038-3079-488b-94e9-298b2acce1f5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.57 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://master:9000/usr/hive/warehouse/hotel_booking_db.db/hotel_bookings_transformed
OK
Time taken: 7.742 seconds
hive>
```

Apache Hive Data Exploration

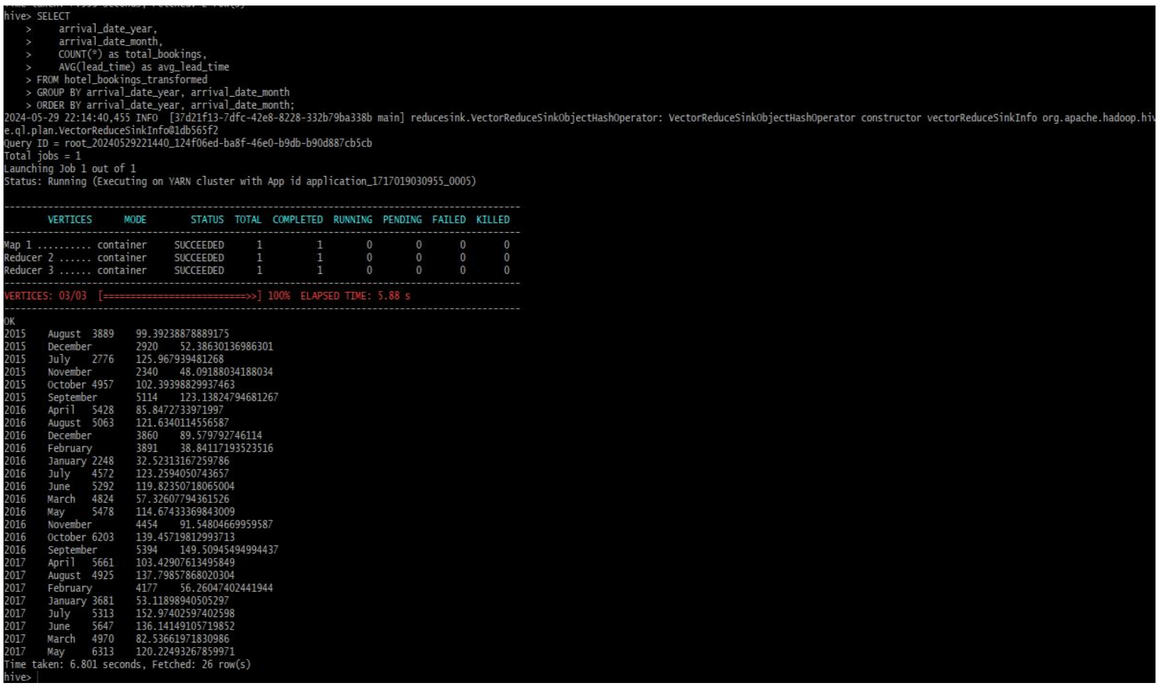
The hotel bookings data shows that the City Hotel has a higher number of bookings (79,330) compared to the Resort Hotel (40,060). The average lead time for the City Hotel is approximately 109.74 days, while the Resort Hotel has an average lead time of around 92.68 days.

When analyzing the data by arrival date year and month, we see that the number of bookings varies throughout the year, with a higher number of bookings in the summer months (July and August) and a lower number of bookings in the winter months (December and January). The average lead time also varies by month, with a longer lead time in the summer months and a shorter lead time in the winter months.

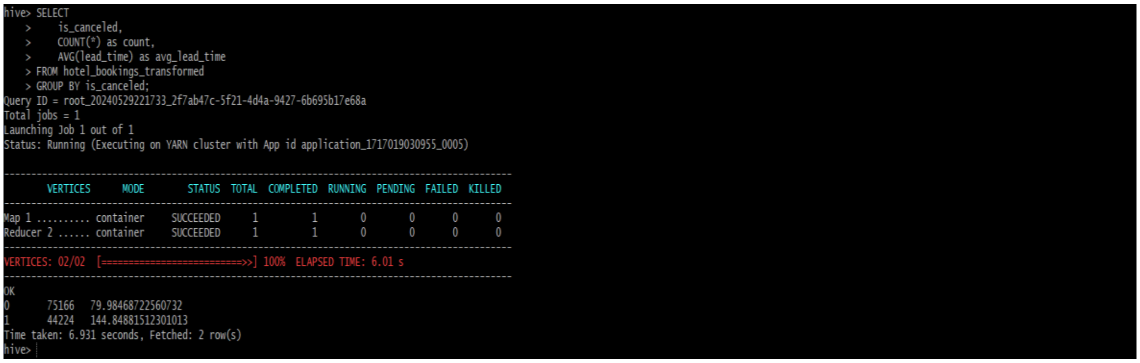
Finally, when looking at the data by cancellation status, we see that most bookings (75,166) were not canceled, while a significant number (44,224) were canceled. The average lead time for non-canceled bookings is approximately 79.98 days, while the average lead time for canceled bookings is around 144.85 days.



```
Time taken: 7.742 seconds
hive> SELECT hotel, COUNT(*) as booking_count
    > FROM hotel_bookings_transformed
    > GROUP BY hotel;
Query ID = root_20240529221329_11ad4a30-d70a-4673-a88e-920551723868
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.34 s
----------------------------------------------------------------------------------------------
OK
City Hotel      79330
Resort Hotel    40060
Time taken: 7.243 seconds, Fetched: 2 row(s)
hive> SELECT hotel, AVG(lead_time) as avg_lead_time
    > FROM hotel_bookings_transformed
    > GROUP BY hotel;
Query ID = root_20240529221402_29ebe061-20bd-4116-8303-56707edffe69
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.72 s
----------------------------------------------------------------------------------------------
OK
City Hotel      109.73572419009201
Resort Hotel    92.67568647029456
Time taken: 7.593 seconds, Fetched: 2 row(s)
```



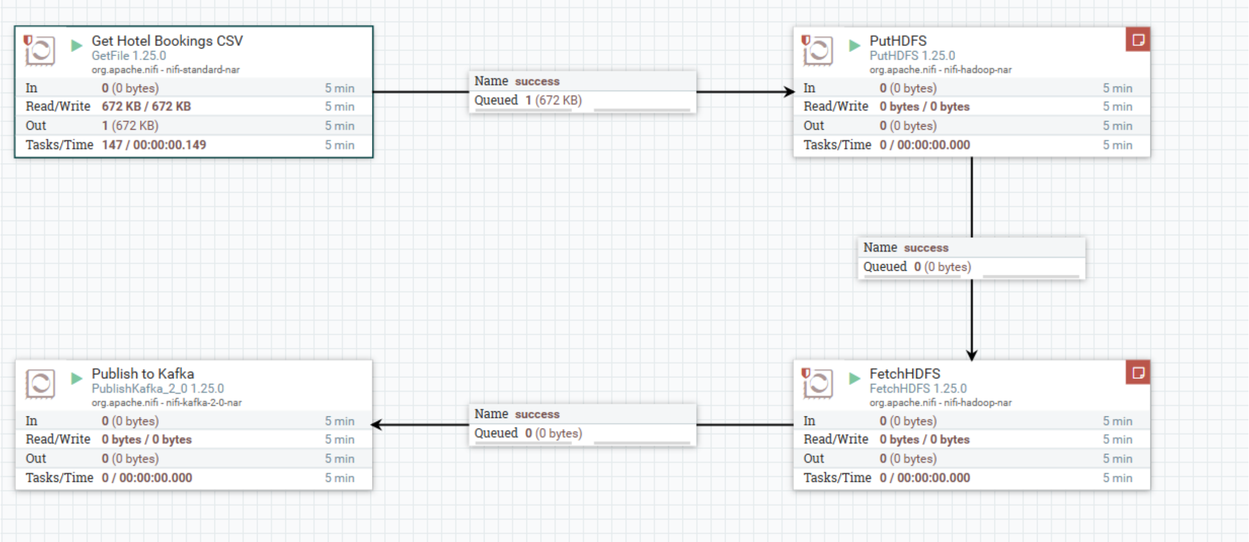
```
Time taken: 7.593 seconds, Fetched: 2 row(s)
hive> SELECT
    >     arrival_date_year,
    >     arrival_date_month,
    >     COUNT(*) as total_bookings,
    >     AVG(lead_time) as avg_lead_time
    > FROM hotel_bookings_transformed
    > GROUP BY arrival_date_year, arrival_date_month
    > ORDER BY arrival_date_year, arrival_date_month;
2024-05-29 22:14:40,455 INFO  [37d21f13-7dfc-42e8-8228-332b79ba338b main] reducesink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObjectHashOperator constructor vectorReduceSinkInfo org.apache.hadoop.hiv
e.ql.plan.VectorReduceSinkInfo@1db565f2
Query ID = root_20240529221440_124f06ed-ba8f-46e0-b9db-b90d887cb5cb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.88 s
----------------------------------------------------------------------------------------------
OK
2015    August  3889    99.39238878889175
2015    December        2920    52.38630136986301
2015    July    2776    125.967939481268
2015    November        2340    48.09188034188034
2015    October 4957    102.39398829937463
2015    September       5114    123.13824794681267
2016    April   5428    85.8472733971997
2016    August  5063    121.6340114556587
2016    December        3860    89.579792746114
2016    February        3891    38.84117193523516
2016    January 2248    32.52313167259786
2016    July    4572    123.2594050743657
2016    June    5292    119.82350718065004
2016    March   4824    57.32607794361526
2016    May     5478    114.67433369843009
2016    November        4454    91.54804669959587
2016    October 6203    139.45719812993713
2016    September       5394    149.50945494994437
2017    April   5661    103.42907613495849
2017    August  4925    137.79857868020304
2017    February        4177    56.26047402441944
2017    January 3681    53.11898940505297
2017    July    5313    152.97402597402598
2017    June    5647    136.14149105719852
2017    March   4970    82.53661971830986
2017    May     6313    120.22493267859971
Time taken: 6.801 seconds, Fetched: 26 row(s)
hive>
```



```
hive> SELECT
    >     is_canceled,
    >     COUNT(*) as count,
    >     AVG(lead_time) as avg_lead_time
    > FROM hotel_bookings_transformed
    > GROUP BY is_canceled;
Query ID = root_20240529221733_2f7ab47c-5f21-4d4a-9427-6b695b17e68a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717019030955_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.01 s
----------------------------------------------------------------------------------------------
OK
0       75166   79.98468722560732
1       44224   144.84881512301013
Time taken: 6.931 seconds, Fetched: 2 row(s)
hive>
```

NIFI INGESTION DATA TO HDFS AND KAFKA

Apache NiFi is set up to handle the ingestion of hotel booking data from a CSV file, store it in HDFS, and subsequently publish it to an Apache Kafka topic. The data flow involves three main steps: ingesting data from a local file using the GetFile processor, storing the data in HDFS with the PutHDFS processor, and fetching the data back from HDFS using the FetchHDFS processor to then publish it to Kafka using the PublishKafka_2_0 processor. The connections between these processors are named and configured to handle successful data flow, ensuring that data transitions smoothly from one stage to the next. This setup leverages NiFi's robust data flow management capabilities to facilitate real-time data processing and integration within a big data architecture.



Get Hotel Bookings CSV
GetFile 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 672 KB / 672 KB 5 min
Out 1 (672 KB) 5 min
Tasks/Time 147 / 00:00:00.149 5 min

Name success
Queued 1 (672 KB)

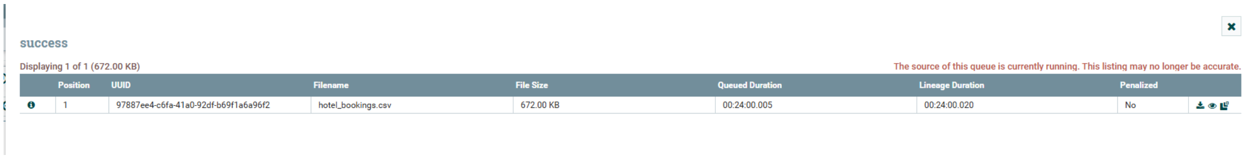
PutHDFS
PutHDFS 1.25.0
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

FetchHDFS
FetchHDFS 1.25.0
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

Publish to Kafka
PublishKafka_2_0 1.25.0
org.apache.nifi - nifi-kafka-2-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

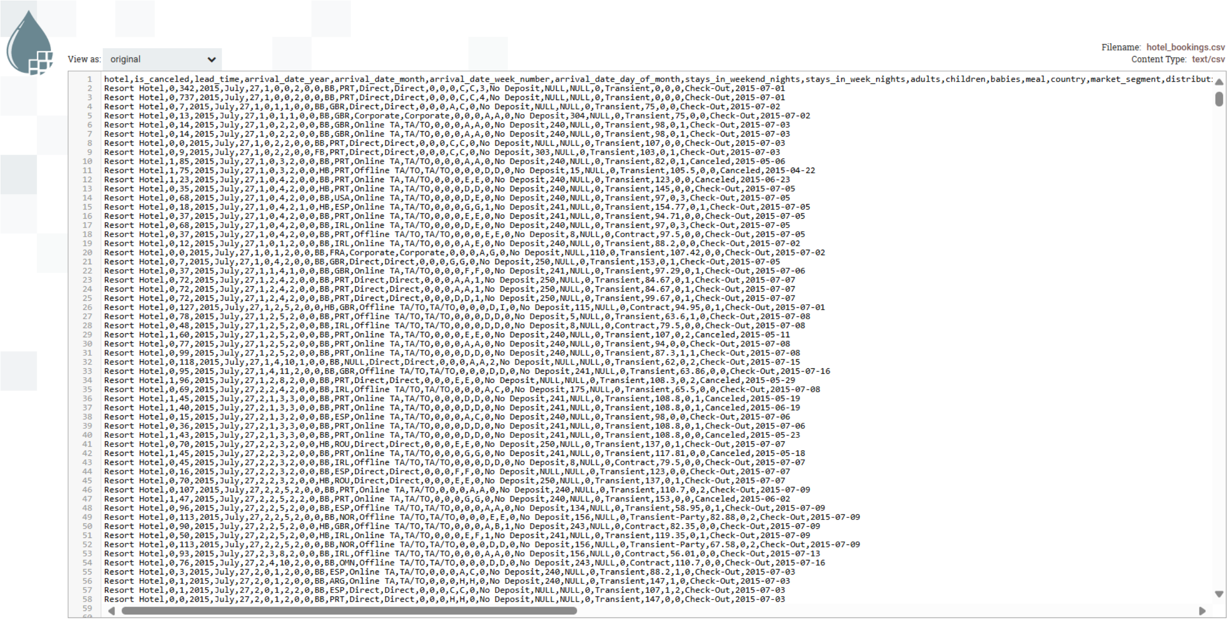


success

Displaying 1 of 1 (672.00 KB)

The source of this queue is currently running. This listing may no longer be accurate.

	Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	
ℹ	1	97887ee4-c6fa-41a0-92df-b69f1a6a96f2	hotel_bookings.csv	672.00 KB	00:24:00.005	00:24:00.020	No	



View as: original

Filename: hotel_bookings.csv
Content Type: text/csv

```
hotel,is_canceled,lead_time,arrival_date_year,arrival_date_month,arrival_date_week_number,arrival_date_day_of_month,stays_in_weekend_nights,stays_in_week_nights,adults,children,babies,meal,country,market_segment,distribut
Resort Hotel,0,342,2015,July,27,1,0,0,2,0,0,BB,PRT,Direct,Direct,0,0,0,C,C,3,No Deposit,NULL,NULL,0,Transient,0,0,0,Check-Out,2015-07-01
Resort Hotel,0,737,2015,July,27,1,0,0,2,0,0,BB,PRT,Direct,Direct,0,0,0,C,C,4,No Deposit,NULL,NULL,0,Transient,0,0,0,Check-Out,2015-07-01
Resort Hotel,0,7,2015,July,27,1,0,1,1,0,0,BB,GBR,Direct,Direct,0,0,0,A,C,0,No Deposit,NULL,NULL,0,Transient,75,0,0,Check-Out,2015-07-02
Resort Hotel,0,13,2015,July,27,1,0,1,1,0,0,BB,GBR,Corporate,Corporate,0,0,0,A,A,0,No Deposit,304,NULL,0,Transient,75,0,0,Check-Out,2015-07-02
Resort Hotel,0,14,2015,July,27,1,0,2,2,0,0,BB,GBR,Online TA,TA/TO,0,0,0,A,A,0,No Deposit,240,NULL,0,Transient,98,0,1,Check-Out,2015-07-03
Resort Hotel,0,14,2015,July,27,1,0,2,2,0,0,BB,GBR,Online TA,TA/TO,0,0,0,A,A,0,No Deposit,240,NULL,0,Transient,98,0,1,Check-Out,2015-07-03
Resort Hotel,0,0,2015,July,27,1,0,2,2,0,0,BB,PRT,Direct,Direct,0,0,0,C,C,0,No Deposit,NULL,NULL,0,Transient,107,0,0,Check-Out,2015-07-03
Resort Hotel,0,9,2015,July,27,1,0,2,2,0,0,FB,PRT,Direct,Direct,0,0,0,C,C,0,No Deposit,303,NULL,0,Transient,103,0,1,Check-Out,2015-07-03
Resort Hotel,1,85,2015,July,27,1,0,3,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,A,A,0,No Deposit,240,NULL,0,Transient,82,0,1,Canceled,2015-05-06
Resort Hotel,1,75,2015,July,27,1,0,3,2,0,0,HB,PRT,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,15,NULL,0,Transient,105.5,0,0,Canceled,2015-04-22
Resort Hotel,1,23,2015,July,27,1,0,4,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,E,E,0,No Deposit,240,NULL,0,Transient,123,0,0,Canceled,2015-06-23
Resort Hotel,0,35,2015,July,27,1,0,4,2,0,0,HB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,240,NULL,0,Transient,145,0,0,Check-Out,2015-07-05
Resort Hotel,0,68,2015,July,27,1,0,4,2,0,0,BB,USA,Online TA,TA/TO,0,0,0,D,E,0,No Deposit,240,NULL,0,Transient,97,0,3,Check-Out,2015-07-05
Resort Hotel,0,18,2015,July,27,1,0,4,2,1,0,HB,ESP,Online TA,TA/TO,0,0,0,G,G,1,No Deposit,241,NULL,0,Transient,154.77,0,1,Check-Out,2015-07-05
Resort Hotel,0,37,2015,July,27,1,0,4,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,E,E,0,No Deposit,241,NULL,0,Transient,94.71,0,0,Check-Out,2015-07-05
Resort Hotel,0,68,2015,July,27,1,0,4,2,0,0,BB,IRL,Online TA,TA/TO,0,0,0,D,E,0,No Deposit,240,NULL,0,Transient,97,0,3,Check-Out,2015-07-05
Resort Hotel,0,37,2015,July,27,1,0,4,2,0,0,BB,PRT,Offline TA/TO,TA/TO,0,0,0,E,E,0,No Deposit,8,NULL,0,Contract,97.5,0,0,Check-Out,2015-07-05
Resort Hotel,0,12,2015,July,27,1,0,1,2,0,0,BB,IRL,Online TA,TA/TO,0,0,0,A,E,0,No Deposit,240,NULL,0,Transient,88.2,0,0,Check-Out,2015-07-02
Resort Hotel,0,0,2015,July,27,1,0,1,2,0,0,BB,FRA,Corporate,Corporate,0,0,0,A,G,0,No Deposit,NULL,110,0,Transient,107.42,0,0,Check-Out,2015-07-02
Resort Hotel,0,7,2015,July,27,1,0,4,2,0,0,BB,GBR,Direct,Direct,0,0,0,G,G,0,No Deposit,250,NULL,0,Transient,153,0,1,Check-Out,2015-07-05
Resort Hotel,0,37,2015,July,27,1,1,4,1,0,0,BB,GBR,Online TA,TA/TO,0,0,0,F,F,0,No Deposit,241,NULL,0,Transient,97.29,0,1,Check-Out,2015-07-06
Resort Hotel,0,72,2015,July,27,1,2,4,2,0,0,BB,PRT,Direct,Direct,0,0,0,A,A,1,No Deposit,250,NULL,0,Transient,84.67,0,1,Check-Out,2015-07-07
Resort Hotel,0,72,2015,July,27,1,2,4,2,0,0,BB,PRT,Direct,Direct,0,0,0,A,A,1,No Deposit,250,NULL,0,Transient,84.67,0,1,Check-Out,2015-07-07
Resort Hotel,0,72,2015,July,27,1,2,4,2,0,0,BB,PRT,Direct,Direct,0,0,0,D,D,1,No Deposit,250,NULL,0,Transient,99.67,0,1,Check-Out,2015-07-07
Resort Hotel,0,127,2015,July,27,1,2,5,2,0,0,HB,GBR,Offline TA/TO,TA/TO,0,0,0,D,I,0,No Deposit,115,NULL,0,Contract,94.95,0,1,Check-Out,2015-07-01
Resort Hotel,0,78,2015,July,27,1,2,5,2,0,0,BB,PRT,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,5,NULL,0,Transient,63.6,1,0,Check-Out,2015-07-08
Resort Hotel,0,48,2015,July,27,1,2,5,2,0,0,BB,IRL,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,8,NULL,0,Contract,79.5,0,0,Check-Out,2015-07-08
Resort Hotel,1,60,2015,July,27,1,2,5,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,E,E,0,No Deposit,240,NULL,0,Transient,107,0,2,Canceled,2015-05-11
Resort Hotel,0,77,2015,July,27,1,2,5,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,A,A,0,No Deposit,240,NULL,0,Transient,94,0,0,Check-Out,2015-07-08
Resort Hotel,0,99,2015,July,27,1,2,5,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,240,NULL,0,Transient,87.3,1,1,Check-Out,2015-07-08
Resort Hotel,0,118,2015,July,27,1,4,10,1,0,0,BB,NULL,Direct,Direct,0,0,0,A,A,2,No Deposit,NULL,NULL,0,Transient,62,0,2,Check-Out,2015-07-15
Resort Hotel,0,95,2015,July,27,1,4,11,2,0,0,BB,GBR,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,241,NULL,0,Transient,63.86,0,0,Check-Out,2015-07-16
Resort Hotel,1,96,2015,July,27,1,2,8,2,0,0,BB,PRT,Direct,Direct,0,0,0,E,E,0,No Deposit,NULL,NULL,0,Transient,108.3,0,2,Canceled,2015-05-29
Resort Hotel,0,69,2015,July,27,2,2,4,2,0,0,BB,IRL,Offline TA/TO,TA/TO,0,0,0,A,C,0,No Deposit,175,NULL,0,Transient,65.5,0,0,Check-Out,2015-07-08
Resort Hotel,1,45,2015,July,27,2,1,3,3,0,0,BB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,241,NULL,0,Transient,108.8,0,1,Canceled,2015-05-19
Resort Hotel,1,40,2015,July,27,2,1,3,3,0,0,BB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,241,NULL,0,Transient,108.8,0,1,Canceled,2015-06-19
Resort Hotel,0,15,2015,July,27,2,1,3,2,0,0,BB,ESP,Online TA,TA/TO,0,0,0,A,C,0,No Deposit,240,NULL,0,Transient,98,0,0,Check-Out,2015-07-06
Resort Hotel,0,36,2015,July,27,2,1,3,3,0,0,BB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,241,NULL,0,Transient,108.8,0,1,Check-Out,2015-07-06
Resort Hotel,1,43,2015,July,27,2,1,3,3,0,0,BB,PRT,Online TA,TA/TO,0,0,0,D,D,0,No Deposit,241,NULL,0,Transient,108.8,0,0,Canceled,2015-05-23
Resort Hotel,0,70,2015,July,27,2,2,3,2,0,0,HB,ROU,Direct,Direct,0,0,0,E,E,0,No Deposit,250,NULL,0,Transient,137,0,1,Check-Out,2015-07-07
Resort Hotel,1,45,2015,July,27,2,2,3,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,G,G,0,No Deposit,241,NULL,0,Transient,117.81,0,0,Canceled,2015-05-18
Resort Hotel,0,45,2015,July,27,2,2,3,2,0,0,BB,IRL,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,8,NULL,0,Contract,79.5,0,0,Check-Out,2015-07-07
Resort Hotel,0,16,2015,July,27,2,2,3,2,0,0,BB,ESP,Direct,Direct,0,0,0,F,F,0,No Deposit,NULL,NULL,0,Transient,123,0,0,Check-Out,2015-07-07
Resort Hotel,0,70,2015,July,27,2,2,3,2,0,0,HB,ROU,Direct,Direct,0,0,0,E,E,0,No Deposit,250,NULL,0,Transient,137,0,1,Check-Out,2015-07-07
Resort Hotel,0,107,2015,July,27,2,2,5,2,0,0,BB,PRT,Online TA,TA/TO,0,0,0,A,A,0,No Deposit,240,NULL,0,Transient,110.7,0,2,Check-Out,2015-07-09
Resort Hotel,1,47,2015,July,27,2,2,5,2,2,0,BB,PRT,Online TA,TA/TO,0,0,0,G,G,0,No Deposit,240,NULL,0,Transient,153,0,0,Canceled,2015-06-02
Resort Hotel,0,96,2015,July,27,2,2,5,2,0,0,BB,ESP,Offline TA/TO,TA/TO,0,0,0,A,A,0,No Deposit,134,NULL,0,Transient,58.95,0,1,Check-Out,2015-07-09
Resort Hotel,0,113,2015,July,27,2,2,5,2,0,0,BB,NOR,Offline TA/TO,TA/TO,0,0,0,E,E,0,No Deposit,156,NULL,0,Transient-Party,82.88,0,2,Check-Out,2015-07-09
Resort Hotel,0,90,2015,July,27,2,2,5,2,0,0,HB,GBR,Offline TA/TO,TA/TO,0,0,0,A,B,1,No Deposit,243,NULL,0,Contract,82.35,0,0,Check-Out,2015-07-09
Resort Hotel,0,50,2015,July,27,2,2,5,2,0,0,HB,IRL,Online TA,TA/TO,0,0,0,E,F,1,No Deposit,241,NULL,0,Transient,119.35,0,1,Check-Out,2015-07-09
Resort Hotel,0,113,2015,July,27,2,2,5,2,0,0,BB,NOR,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,156,NULL,0,Transient-Party,67.58,0,2,Check-Out,2015-07-09
Resort Hotel,0,93,2015,July,27,2,3,8,2,0,0,BB,IRL,Offline TA/TO,TA/TO,0,0,0,A,A,0,No Deposit,156,NULL,0,Contract,56.01,0,0,Check-Out,2015-07-13
Resort Hotel,0,76,2015,July,27,2,4,10,2,0,0,BB,OMN,Offline TA/TO,TA/TO,0,0,0,D,D,0,No Deposit,243,NULL,0,Contract,110.7,0,0,Check-Out,2015-07-16
Resort Hotel,0,3,2015,July,27,2,0,1,2,0,0,BB,ESP,Online TA,TA/TO,0,0,0,A,C,0,No Deposit,240,NULL,0,Transient,88.2,1,0,Check-Out,2015-07-03
Resort Hotel,0,1,2015,July,27,2,0,1,2,0,0,BB,ARG,Online TA,TA/TO,0,0,0,H,H,0,No Deposit,240,NULL,0,Transient,147,1,0,Check-Out,2015-07-03
Resort Hotel,0,1,2015,July,27,2,0,1,2,2,0,BB,ESP,Direct,Direct,0,0,0,C,C,0,No Deposit,NULL,NULL,0,Transient,107,1,2,Check-Out,2015-07-03
Resort Hotel,0,0,2015,July,27,2,0,1,2,0,0,BB,PRT,Direct,Direct,0,0,0,H,H,0,No Deposit,NULL,NULL,0,Transient,147,0,0,Check-Out,2015-07-03
```

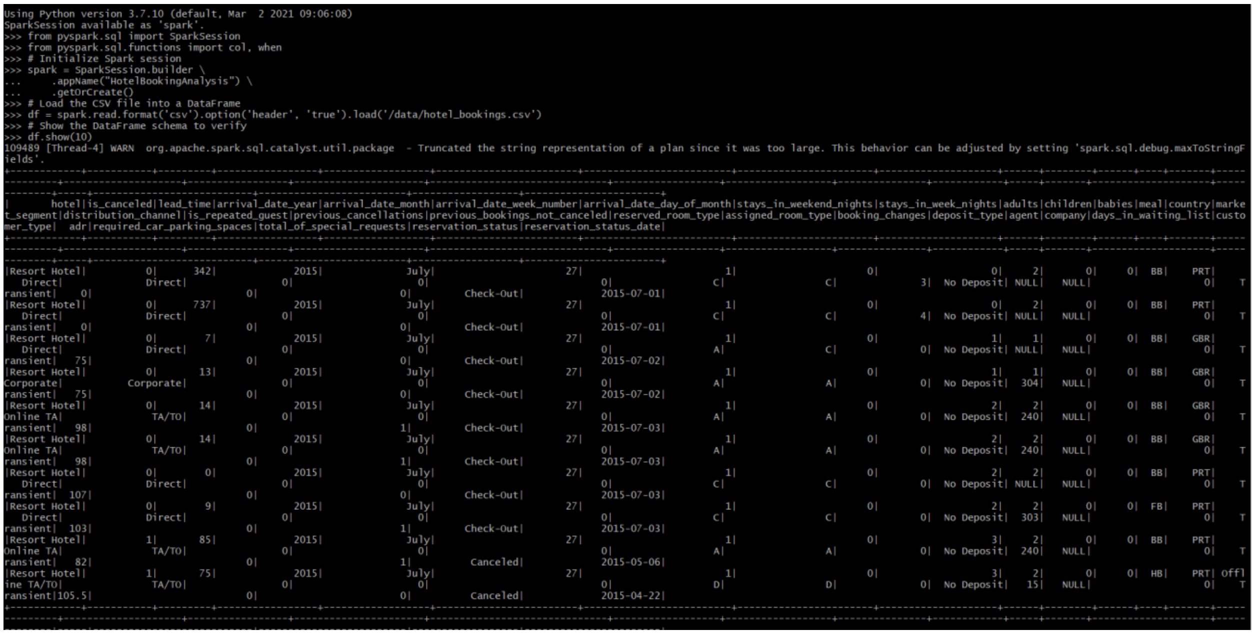
The Hotel Booking project aims to leverage big data technologies to enhance various aspects of hotel management and guest experience. By integrating Apache NiFi for data ingestion, Apache Kafka for real-time data streaming, Apache HDFS for scalable data storage, and Apache Hive for data transformation and querying, the project enables efficient handling of large volumes of booking data. Through detailed configuration and implementation steps, the project ensures seamless data flow from ingestion to analysis, facilitating insights generation, decision support, and operational optimization. Overall, the project demonstrates the power of big data technologies in revolutionizing hotel management and customer service.

DATA PROCESSING AND ANALYSIS WITH APACHE SPARK

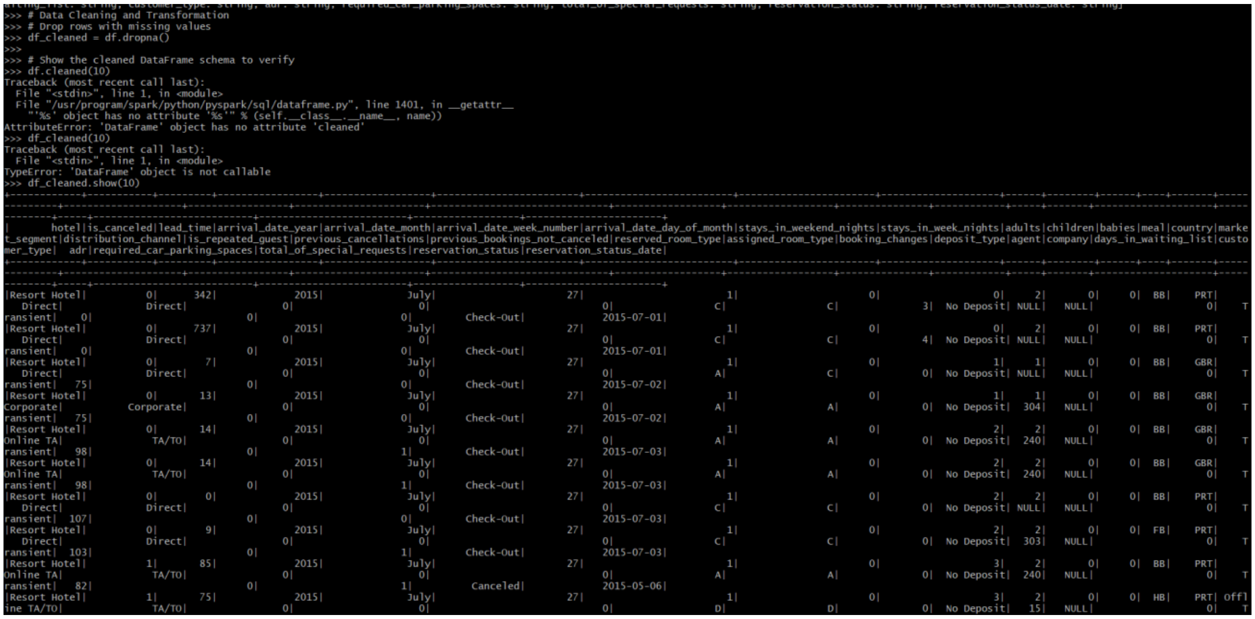
Our analysis involved Data Processing and Analysis with Apache Spark, leveraging its powerful capabilities for handling large datasets efficiently. Spark facilitated various tasks including data cleaning, transformation, and complex analytics. With its in-memory processing capabilities, Spark enabled rapid analysis and extraction of insights from the ingested data.

We utilized Spark to process the dataset, which included tasks such as handling missing values and converting categorical variables to numerical format. Additionally, we performed feature engineering to create new columns that could potentially enhance our predictive models.

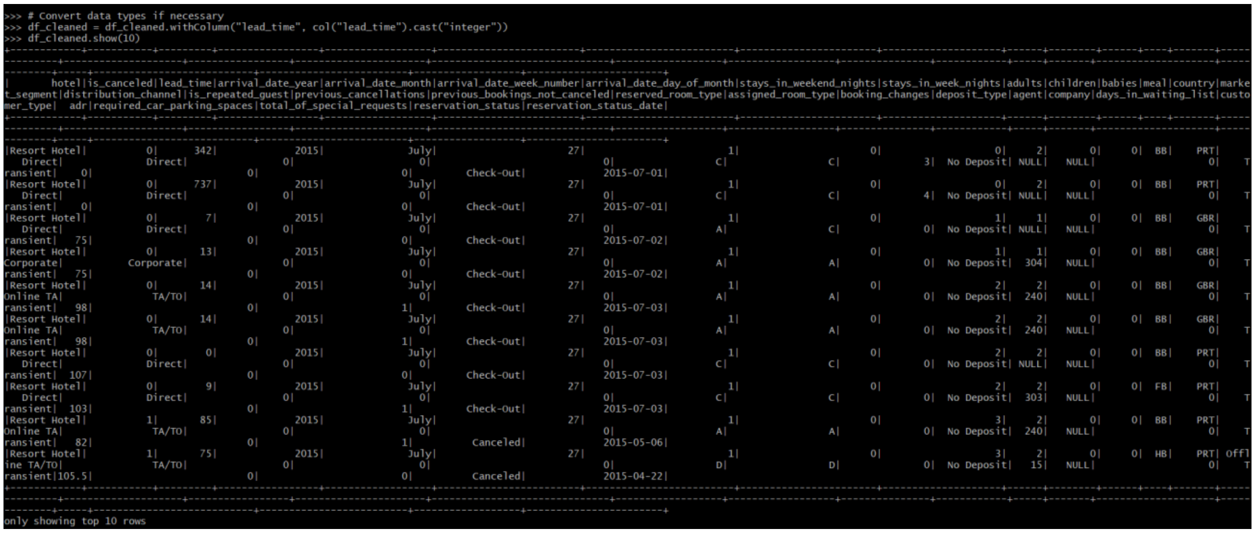
Our main objective was to apply machine learning models to predict cancellations, optimize booking strategies, and personalize guest experiences based on the processed data. To achieve this, we trained a logistic regression model using 50,000 records from the dataset. Features such as lead time, arrival year, and week number were selected for model training.



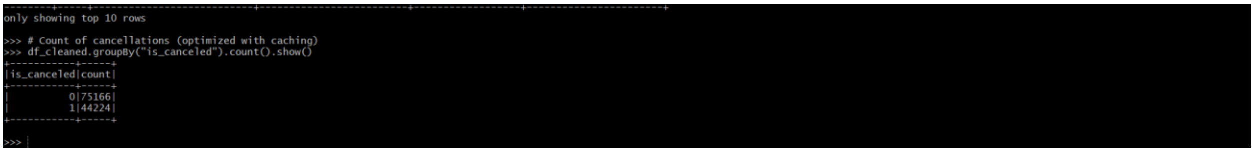
```
Using Python version 3.7.10 (default, Mar  2 2021 09:06:08)
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql.functions import col, when
>>> # Initialize Spark session
>>> spark = SparkSession.builder \
...     .appName("HotelBookingAnalysis") \
...     .getOrCreate()
>>> # Load the CSV file into a DataFrame
>>> df = spark.read.format('csv').option('header', 'true').load('/data/hotel_bookings.csv')
>>> # Show the DataFrame schema to verify
>>> df.show(10)
109489 [Thread-4] WARN  org.apache.spark.sql.catalyst.util.package  - Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringF
ields'.
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+---------------------------+-------------------------+------------------+-----------------------+
|       hotel|is_canceled|lead_time|arrival_date_year|arrival_date_month|arrival_date_week_number|arrival_date_day_of_month|stays_in_weekend_nights|stays_in_week_nights|adults|children|babies|meal|country|marke
t_segment|distribution_channel|is_repeated_guest|previous_cancellations|previous_bookings_not_canceled|reserved_room_type|assigned_room_type|booking_changes|deposit_type|agent|company|days_in_waiting_list|custo
mer_type|  adr|required_car_parking_spaces|total_of_special_requests|reservation_status|reservation_status_date|
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+---------------------------+-------------------------+------------------+-----------------------+
|Resort Hotel|          0|      342|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              3|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                          0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|      737|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              4|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                          0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|        7|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
   Direct|              Direct|                0|                     0|                             0|                 A|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|   75|                          0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       13|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
Corporate|           Corporate|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  304|   NULL|                   0|    T
ransient|   75|                          0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        0|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|  107|                          0|                        0|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        9|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  FB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit|  303|   NULL|                   0|    T
ransient|  103|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          1|       85|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  BB|    PRT|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   82|                          0|                        1|          Canceled|             2015-05-06|
|Resort Hotel|          1|       75|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  HB|    PRT| Offl
ine TA/TO|               TA/TO|                0|                     0|                             0|                 D|                 D|              0|  No Deposit|   15|   NULL|                   0|    T
ransient|105.5|                          0|                        0|          Canceled|             2015-04-22|
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+---------------------------+-------------------------+------------------+-----------------------+
```



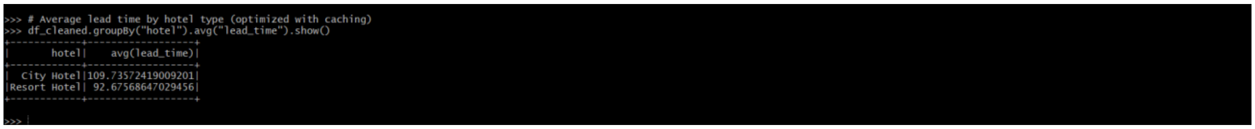
```
aiting_list: string, customer_type: string, adr: string, required_car_parking_spaces: string, total_of_special_requests: string, reservation_status: string, reservation_status_date: string]
>>> # Data Cleaning and Transformation
>>> # Drop rows with missing values
>>> df_cleaned = df.dropna()
>>>
>>> # Show the cleaned DataFrame schema to verify
>>> df.cleaned(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/program/spark/python/pyspark/sql/dataframe.py", line 1401, in __getattr__
    "'%s' object has no attribute '%s'" % (self.__class__.__name__, name))
AttributeError: 'DataFrame' object has no attribute 'cleaned'
>>> df_cleaned(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'DataFrame' object is not callable
>>> df_cleaned.show(10)
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+---------------------------+-------------------------+------------------+-----------------------+
|       hotel|is_canceled|lead_time|arrival_date_year|arrival_date_month|arrival_date_week_number|arrival_date_day_of_month|stays_in_weekend_nights|stays_in_week_nights|adults|children|babies|meal|country|marke
t_segment|distribution_channel|is_repeated_guest|previous_cancellations|previous_bookings_not_canceled|reserved_room_type|assigned_room_type|booking_changes|deposit_type|agent|company|days_in_waiting_list|custo
mer_type|  adr|required_car_parking_spaces|total_of_special_requests|reservation_status|reservation_status_date|
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+---------------------------+-------------------------+------------------+-----------------------+
|Resort Hotel|          0|      342|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              3|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                          0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|      737|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              4|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                          0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|        7|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
   Direct|              Direct|                0|                     0|                             0|                 A|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|   75|                          0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       13|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
Corporate|           Corporate|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  304|   NULL|                   0|    T
ransient|   75|                          0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        0|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|  107|                          0|                        0|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        9|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  FB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit|  303|   NULL|                   0|    T
ransient|  103|                          0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          1|       85|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  BB|    PRT|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   82|                          0|                        1|          Canceled|             2015-05-06|
|Resort Hotel|          1|       75|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  HB|    PRT| Offl
ine TA/TO|               TA/TO|                0|                     0|                             0|                 D|                 D|              0|  No Deposit|   15|   NULL|                   0|    T
```



```
>>> # Convert data types if necessary
>>> df_cleaned = df_cleaned.withColumn("lead_time", col("lead_time").cast("integer"))
>>> df_cleaned.show(10)
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+--------------------------+-------------------------+------------------+-----------------------+
|       hotel|is_canceled|lead_time|arrival_date_year|arrival_date_month|arrival_date_week_number|arrival_date_day_of_month|stays_in_weekend_nights|stays_in_week_nights|adults|children|babies|meal|country|marke
t_segment|distribution_channel|is_repeated_guest|previous_cancellations|previous_bookings_not_canceled|reserved_room_type|assigned_room_type|booking_changes|deposit_type|agent|company|days_in_waiting_list|custo
mer_type|  adr|required_car_parking_spaces|total_of_special_requests|reservation_status|reservation_status_date|
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+--------------------------+-------------------------+------------------+-----------------------+
|Resort Hotel|          0|      342|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              3|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                         0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|      737|             2015|              July|                      27|                        1|                      0|                   0|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              4|  No Deposit| NULL|   NULL|                   0|    T
ransient|    0|                         0|                        0|         Check-Out|             2015-07-01|
|Resort Hotel|          0|        7|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
   Direct|              Direct|                0|                     0|                             0|                 A|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|   75|                         0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       13|             2015|              July|                      27|                        1|                      0|                   1|     1|       0|     0|  BB|    GBR|
Corporate|           Corporate|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  304|   NULL|                   0|    T
ransient|   75|                         0|                        0|         Check-Out|             2015-07-02|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                         0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|       14|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    GBR|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   98|                         0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        0|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  BB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit| NULL|   NULL|                   0|    T
ransient|  107|                         0|                        0|         Check-Out|             2015-07-03|
|Resort Hotel|          0|        9|             2015|              July|                      27|                        1|                      0|                   2|     2|       0|     0|  FB|    PRT|
   Direct|              Direct|                0|                     0|                             0|                 C|                 C|              0|  No Deposit|  303|   NULL|                   0|    T
ransient|  103|                         0|                        1|         Check-Out|             2015-07-03|
|Resort Hotel|          1|       85|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  BB|    PRT|
Online TA|               TA/TO|                0|                     0|                             0|                 A|                 A|              0|  No Deposit|  240|   NULL|                   0|    T
ransient|   82|                         0|                        1|          Canceled|             2015-05-06|
|Resort Hotel|          1|       75|             2015|              July|                      27|                        1|                      0|                   3|     2|       0|     0|  HB|    PRT| Offl
ine TA/TO|               TA/TO|                0|                     0|                             0|                 D|                 D|              0|  No Deposit|   15|   NULL|                   0|    T
ransient|105.5|                         0|                        0|          Canceled|             2015-04-22|
+------------+-----------+---------+-----------------+------------------+------------------------+-------------------------+-----------------------+--------------------+------+--------+------+----+-------+-----
---------+--------------------+-----------------+----------------------+------------------------------+------------------+------------------+---------------+------------+-----+-------+--------------------+-----
--------+-----+--------------------------+-------------------------+------------------+-----------------------+
only showing top 10 rows
```

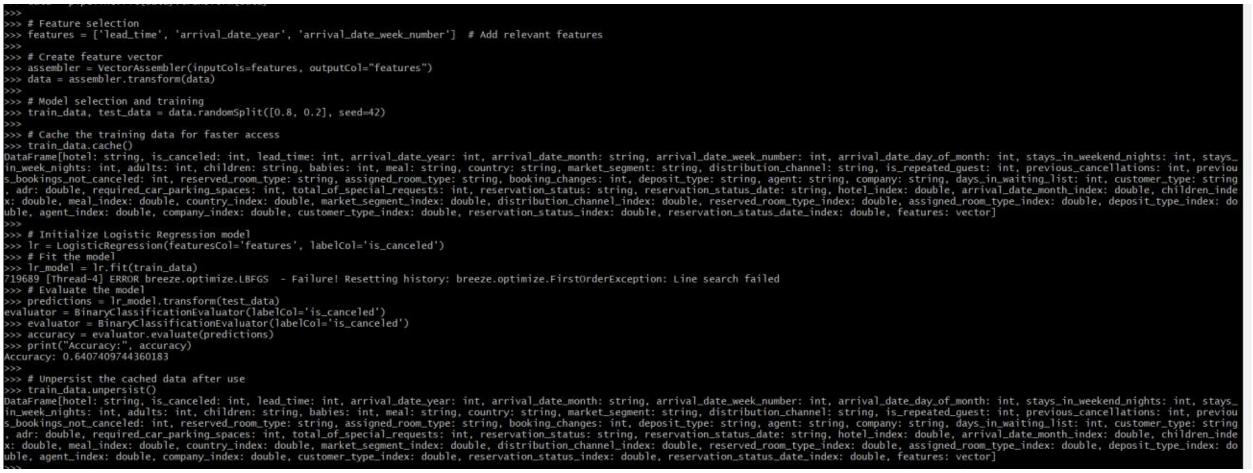


```
--------+-----+--------------------------+-------------------------+------------------+-----------------------+
only showing top 10 rows

>>> # Count of cancellations (optimized with caching)
>>> df_cleaned.groupBy("is_canceled").count().show()
+-----------+-----+
|is_canceled|count|
+-----------+-----+
|          0|75166|
|          1|44224|
+-----------+-----+

>>> |
```



```
>>> # Average lead time by hotel type (optimized with caching)
>>> df_cleaned.groupBy("hotel").avg("lead_time").show()
+------------+------------------+
|       hotel|    avg(lead_time)|
+------------+------------------+
|  City Hotel|109.73572419009201|
|Resort Hotel| 92.67568647029456|
+------------+------------------+

>>> |
```

The logistic regression model achieved an accuracy of approximately 64.07% on the test dataset. Furthermore, we conducted a detailed analysis of the model's performance by examining summary statistics.



```
>>>
>>> # Feature selection
>>> features = ['lead_time', 'arrival_date_year', 'arrival_date_week_number']  # Add relevant features
>>>
>>> # Create feature vector
>>> assembler = VectorAssembler(inputCols=features, outputCol="features")
>>> data = assembler.transform(data)
>>>
>>> # Model selection and training
>>> train_data, test_data = data.randomSplit([0.8, 0.2], seed=42)
>>>
>>> # Cache the training data for faster access
>>> train_data.cache()
DataFrame[hotel: string, is_canceled: int, lead_time: int, arrival_date_year: int, arrival_date_month: string, arrival_date_week_number: int, arrival_date_day_of_month: int, stays_in_weekend_nights: int, stays_
in_week_nights: int, adults: int, children: string, babies: int, meal: string, country: string, market_segment: string, distribution_channel: string, is_repeated_guest: int, previous_cancellations: int, previou
s_bookings_not_canceled: int, reserved_room_type: string, assigned_room_type: string, booking_changes: int, deposit_type: string, agent: string, company: string, days_in_waiting_list: int, customer_type: string
, adr: double, required_car_parking_spaces: int, total_of_special_requests: int, reservation_status: string, reservation_status_date: string, hotel_index: double, arrival_date_month_index: double, children_inde
x: double, meal_index: double, country_index: double, market_segment_index: double, distribution_channel_index: double, reserved_room_type_index: double, assigned_room_type_index: double, deposit_type_index: do
uble, agent_index: double, company_index: double, customer_type_index: double, reservation_status_index: double, reservation_status_date_index: double, features: vector]
>>>
>>> # Initialize Logistic Regression model
>>> lr = LogisticRegression(featuresCol='features', labelCol='is_canceled')
>>> # Fit the model
>>> lr_model = lr.fit(train_data)
719689 [Thread-4] ERROR breeze.optimize.LBFGS  - Failure! Resetting history: breeze.optimize.FirstOrderException: Line search failed
>>> # Evaluate the model
>>> predictions = lr_model.transform(test_data)
evaluator = BinaryClassificationEvaluator(labelCol='is_canceled')
>>> evaluator = BinaryClassificationEvaluator(labelCol='is_canceled')
>>> accuracy = evaluator.evaluate(predictions)
>>> print("Accuracy:", accuracy)
Accuracy: 0.6407409744360183
>>>
>>> # Unpersist the cached data after use
>>> train_data.unpersist()
DataFrame[hotel: string, is_canceled: int, lead_time: int, arrival_date_year: int, arrival_date_month: string, arrival_date_week_number: int, arrival_date_day_of_month: int, stays_in_weekend_nights: int, stays_
in_week_nights: int, adults: int, children: string, babies: int, meal: string, country: string, market_segment: string, distribution_channel: string, is_repeated_guest: int, previous_cancellations: int, previou
s_bookings_not_canceled: int, reserved_room_type: string, assigned_room_type: string, booking_changes: int, deposit_type: string, agent: string, company: string, days_in_waiting_list: int, customer_type: string
, adr: double, required_car_parking_spaces: int, total_of_special_requests: int, reservation_status: string, reservation_status_date: string, hotel_index: double, arrival_date_month_index: double, children_inde
x: double, meal_index: double, country_index: double, market_segment_index: double, distribution_channel_index: double, reserved_room_type_index: double, assigned_room_type_index: double, deposit_type_index: do
uble, agent_index: double, company_index: double, customer_type_index: double, reservation_status_index: double, reservation_status_date_index: double, features: vector]
>>>
```

In summary, our analysis with Apache Spark enabled us to efficiently process and analyze the dataset, train predictive models, and derive insights to optimize booking strategies and enhance guest experiences in the hospitality industry.

3.0 Conclusion

3.1 Implications of the Findings.

Successful implementation of big data analytics in the hotel industry hinges on interdisciplinary collaboration among IT, data science, marketing, and operations teams, ensuring effective communication and goal alignment. Data quality and governance are paramount for reliable analytics outcomes, requiring robust frameworks for data cleansing, normalization, and validation. Continuous improvement through iterative processes and stakeholder feedback is essential to keep analytics solutions aligned with business objectives

3.2 Recommendations.

Implementing big data analytics in the hotel industry requires collaboration among IT, data science, marketing, and operations teams, and maintaining high data quality. Addressing data privacy and bias is crucial. Techniques like NLP for guest reviews and IoT for real-time monitoring offer great potential. Challenges include integrating data, bridging the skill gap, and ensuring security. Embracing a data-driven culture and exploring technologies like blockchain and AI chatbots can drive innovation and differentiation.

3.3 Potential challenges and additional opportunities.

Integrating data from various sources requires robust frameworks and tools. Addressing the skill gap by training data professionals is crucial for successful Big Data projects. Protecting guest data from cyber threats and ensuring regulatory compliance are ongoing challenges. Overcoming resistance to change and fostering a data-driven culture necessitates investment in training and change management. Additionally, exploring blockchain for secure data sharing and AI-driven chatbots for guest interaction offers new opportunities for innovation and differentiation.

3.4 Conclusion.

In conclusion, the intersection of big data analytics and reservation management in the hotel industry holds immense potential for driving operational excellence, enhancing guest experiences, and maximizing revenue generation. By leveraging advanced analytics techniques and robust technological infrastructure, hotels can navigate the complexities of reservation management effectively and stay competitive in a rapidly evolving landscape. However, addressing challenges related to data integration, skills development, security, and organizational change is essential for realizing the full benefits of big data analytics in hospitality. As the industry continues to embrace data-driven approaches, collaboration, innovation, and a commitment to ethical data practices will be key to unlocking the transformative power of big data in hospitality.

4.0 References

- https://www.kaggle.com
- https://365datascience.com/
- https://python.plainenglish.io/
- https://journalofbigdata.springeropen.com/
- https://hadoop.apache.org/
- https://kafka.apache.org/
- https://nifi.apache.org/
- https://spark.apache.org/
- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/
- https://machinelearningmastery.com/
- https://www.soegjobs.com/data-analytics-hospitality-industry