

# **DSC-650 FINAL PROJECT-PROCESS AND ANALYZE HOTEL BOOKING DATA USING BIG DATA COMPONENTS**

**CHITRAMOY MUKHERJEE**

## **Introduction:**

The hotel industry generates huge amounts of data from various sources, including reservations, guest profiles, and operational activities. Analyzing this data using big data components provides an opportunity to gain valuable insights into booking trends (Monthly/weekly Booking Count, Seasonal booking pattern etc), booking preferences (such as best time of year to book a hotel, what type of hotels preferred during what time of the year etc) and operational efficiency (occupancy rate, booking lead time, Revenue per Available Room etc). By leveraging technologies such as HDFS, Apache Hive, spark and solr we can process and analyze large-scale hotel booking datasets efficiently.

## **Objective:**

The objective of a big data project using hotel booking data can vary based on the goals and priorities of the stakeholders involved. From travelers perspective, this data can be used to identify preferences and patterns of the user, Optimize Pricing and Discounts, travel planning assistance and Real-Time Availability Information. From hotel Business perspective, this data can be used to identify how to optimize revenue management, forecasting and Planning, Competitive Benchmarking, Data-Driven Decision Making and Risk management.

## **Problem Statement:**

The overarching aim of the project is to utilize big data analytics on hotel booking data to address these challenges, extract actionable insights, and develop strategies that enhance operational efficiency, improve customer satisfaction, and maximize revenue for the hotel business. The project aims to address the following key challenges and questions using hotel booking data.

Booking Pattern Analysis.

Operational Efficiency.

Cancellation Prediction.

Predictive Analytics for Demand Forecasting.

Customer Segmentation.

Revenue Management.

## Scope and Complexity:

hotel\_booking.csv dataset appears to cover a comprehensive set of features related to hotel bookings, including guest details, booking history, room information, payment details, reservation status, customer information, and additional preferences. The complexity of the big data project would depend on the size of the dataset, the specific analysis or tasks to be performed, and the goals of the project. Analyzing trends, predicting cancellations, and optimizing room allocations are potential areas of exploration within this dataset.

## Dataset Source and Content:

hotel\_booking.csv data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

All personally identifying information has been removed from the data. Credit card information related to each booking is encrypted or masked (last 4 byte of CC # has been visible).

Below is the snippet of dataset columns with data type and count of no-null rows,

```
# Column Non-Null Count Dtype
---
0 hotel 119390 non-null object
1 is_canceled 119390 non-null int64
2 lead_time 119390 non-null int64
3 arrival_date_year 119390 non-null int64
4 arrival_date_month 119390 non-null object
5 arrival_date_week_number 119390 non-null int64
6 arrival_date_day_of_month 119390 non-null int64
7 stays_in_weekend_nights 119390 non-null int64
8 stays_in_week_nights 119390 non-null int64
9 adults 119390 non-null int64
```

10 children 119386 non-null float64  
11 babies 119390 non-null int64  
12 meal 119390 non-null object  
13 country 118902 non-null object  
14 market\_segment 119390 non-null object  
15 distribution\_channel 119390 non-null object  
16 is\_repeated\_guest 119390 non-null int64  
17 previous\_cancellations 119390 non-null int64  
18 previous\_bookings\_not\_canceled 119390 non-null int64  
19 reserved\_room\_type 119390 non-null object  
20 assigned\_room\_type 119390 non-null object  
21 booking\_changes 119390 non-null int64  
22 deposit\_type 119390 non-null object  
23 agent 103050 non-null float64  
24 company 6797 non-null float64  
25 days\_in\_waiting\_list 119390 non-null int64  
26 customer\_type 119390 non-null object  
27 adr 119390 non-null float64  
28 required\_car\_parking\_spaces 119390 non-null int64  
29 total\_of\_special\_requests 119390 non-null int64  
30 reservation\_status 119390 non-null object  
31 reservation\_status\_date 119390 non-null object  
32 name 119390 non-null object  
33 email 119390 non-null object  
34 phone-number 119390 non-null object  
35 credit\_card 119390 non-null object

## Technology Integration:

Planning to use HDFS, Spark (Scala, pyspark), hive and solr.

- **Load hotel\_booking.csv into HDFS :**

Provided screenshot of the steps and validation done during loading **hotel\_booking.csv** the file into HDFS.



- **Spark scala for Booking Patterns, Customer Demographics :**

Execute below queries in SparkSQL with Scala to identify the Booking Patterns, Customer Demographics and Average daily rate from the hotel\_booking df.



The provided Spark SQL query calculates and outputs the total number of bookings, the number of successful bookings (not canceled), and the booking rate as a percentage. 2<sup>nd</sup> Spark SQL provides the cancellation rate.



Executed below sql to identify the most common countries of origin as one of the key customer demographics for guests from the hotel booking table using Spark Scala.



To identify customer types (e.g., transient, contract, group) distribution from a DataFrame (df) in Spark using Scala, below the Spark SQL API. Below is an example of how you can achieve this:



- **Pyspark queries for Revenue Metrics and Customer Demographics :**

Executed the below sql in pyspark to calculate the average booking duration from the df.



Below SQL query selects the hotel type and calculates the total revenue per booking by summing the **adr** (average daily rate) column. The **WHERE** clause filters out canceled bookings (**is\_canceled = 0**), and the **GROUP BY** clause groups the results by hotel type.



- **HIVE Table and Queries for Temporal Metrics and Waitlist Metrics :**

Create statement to create the hotel\_booking table in hive.



To identify seasonal booking patterns from the **hotel\_booking** table in Hive, a Hive SQL query with appropriate aggregations and grouping based on the time-related columns. Assuming that the **arrival\_date\_year** and **arrival\_date\_month** columns are relevant for identifying seasons and based on that executed the below query.

Below query displays the month and year wise booking count filtering out cancelled booking.



Executed the below hive query to identify day-of-the-week trends from the 'hotel\_booking' table in Hive. Used **DAYOFWEEK** function along with other relevant aggregation functions. Want to analyze trends based on the 'arrival\_date\_year', 'arrival\_date\_month', 'arrival\_date\_day\_of\_month' columns.



To calculate the average days in the waiting list from the 'hotel\_booking' table in Hive based on hotel type, used the below query.



Create the mqt\_resort\_bookings using the below query data. mqt\_resort\_bookings table can be used for querying and analysis. We can integrate this table into your reporting or analysis tools to quickly access precomputed results.

We may want to periodically refresh the data in the MQT table to ensure it stays up-to-date. For this, we can recreate the table or update the existing data based on your requirements.

```
SELECT lead_time, COUNT(*) AS total_bookings_resort, (COUNT(*) / SUM(COUNT(*)) OVER  
( )) * 100 AS percentage_bookings_resort FROM hotel_booking WHERE hotel_type = 'Resort  
Hotel' GROUP BY lead_time ORDER BY
```

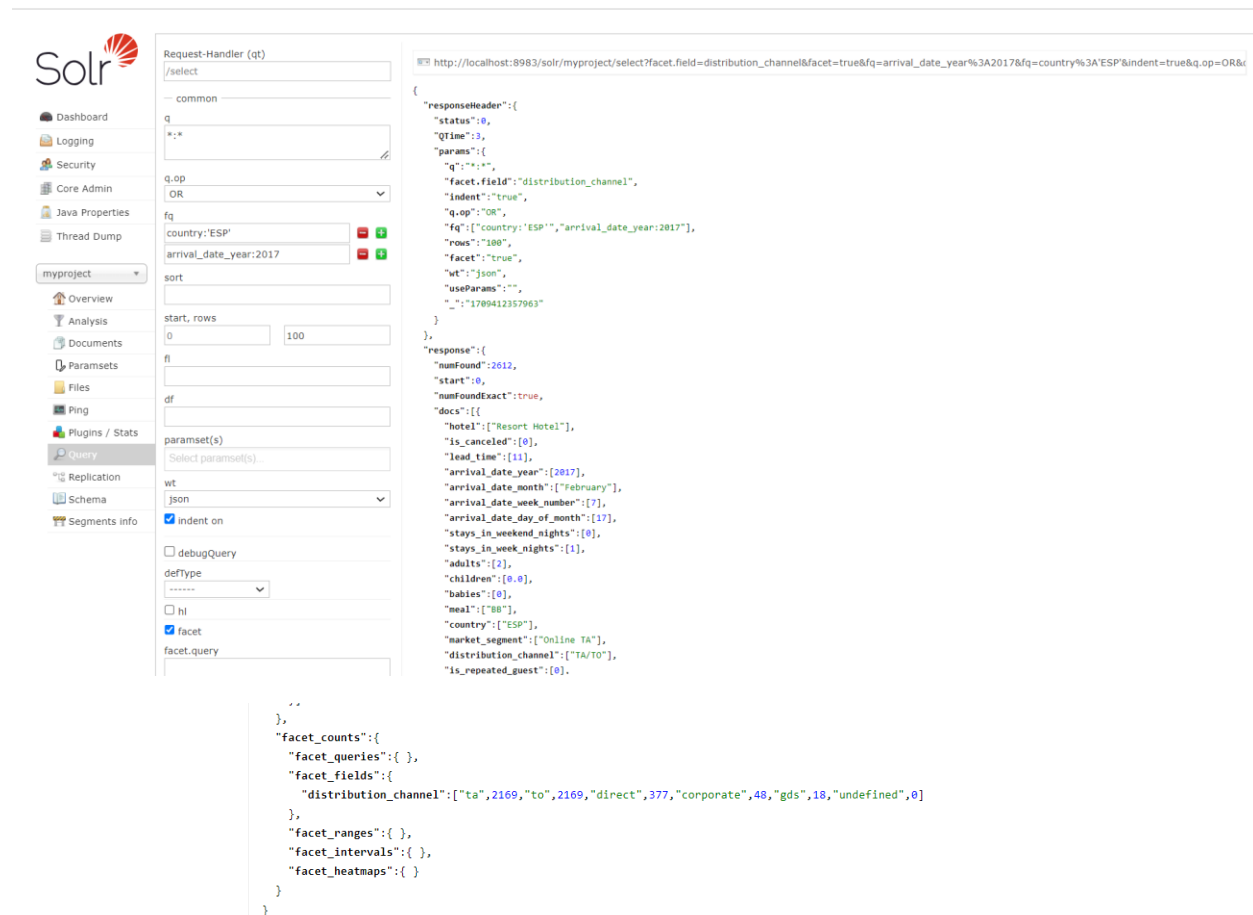


```

    }
  },
  "facet_counts":{
    "facet_queries":{ },
    "facet_fields":{
      "country":["esp",2612,"abw",0,"ago",0,"aia",0,"alb",0,"and",0,"are",0,"arg",0,"arm",0,"asm",0,"ata",0,"atf",0,"aus",0,"aut",0,"aze",0,"bdi",0,"bel",0,"be
    },
    "facet_ranges":{ },
    "facet_intervals":{ },
    "facet_heatmaps":{ }
  }
}

```

Executed below query in solr to identify the how much booking is coming through which distribution channel for arrival year 2017 and for country ESP.



The screenshot displays the Solr Admin interface. On the left, the 'Query' tab is selected under the 'myproject' dropdown. The 'Request-Handler (qt)' is set to '/select'. The 'q' field is empty. The 'q.op' is set to 'OR'. The 'fq' field contains 'country: \'ESP\'' and 'arrival\_date\_year:2017'. The 'sort' field is empty. The 'start' is 0 and 'rows' is 100. The 'wt' is set to 'json'. The 'indent on' checkbox is checked. The 'facet' checkbox is checked, and 'facet.query' is set to 'distribution\_channel'. The 'responseHeader' section shows 'status': 0, 'QTime': 3, and 'params' including 'q', 'facet.field', 'indent', 'q.op', 'fq', 'rows', 'wt', 'useParams', and 'n'. The 'response' section shows 'numFound': 2612, 'start': 0, 'numFoundExact': true, and a list of documents. The first document is a hotel booking record for 'Resort Hotel'.

```

{
  "responseHeader":{
    "status":0,
    "QTime":3,
    "params":{"q":"","facet.field":"distribution_channel",
      "indent":"true",
      "q.op":"OR",
      "fq":["country: \'ESP\'","arrival_date_year:2017"],
      "rows":"100",
      "facet":"true",
      "wt":"json",
      "useParams":"","n":"1709412357963"}
  },
  "response":{
    "numFound":2612,
    "start":0,
    "numFoundExact":true,
    "docs":[
      {
        "hotel":["Resort Hotel"],
        "is_canceled":["0"],
        "lead_time":["11"],
        "arrival_date_year":["2017"],
        "arrival_date_month":["February"],
        "arrival_date_week_number":["7"],
        "arrival_date_day_of_month":["17"],
        "stays_in_weekend_nights":["0"],
        "stays_in_week_nights":["1"],
        "adults":["2"],
        "children":["0.0"],
        "babies":["0"],
        "meal":["BB"],
        "country":["ESP"],
        "market_segment":["Online TA"],
        "distribution_channel":["TA/TO"],
        "is_repeated_guest":["0"]
      }
    ]
  }
}

```

- **Data cleaning and visualization on hotel booking data:**

Followed the steps provided in <https://codelabs.developers.google.com/codelabs/spark-jupyter-dataproc#0> to install apache spark and jupyter notebook on cloud dataproc but setup failed due to the configuration.

Provided the IPYNB file for data cleaning and create visualization.

## Conclusion:

From the query execution on spark-scala and hive, we have analyzed the booking ratio and cancellation ratio. The booking ratio is a measure of the success in converting inquiries or potential reservations into actual bookings. It is calculated as the ratio of the number of successful bookings to the total number of inquiries or attempts. The cancellation ratio measures the percentage of bookings canceled relative to the total number of bookings.

The booking and cancellation ratios provide valuable metrics for assessing the efficiency of booking processes, customer satisfaction, and revenue management. Regular monitoring and analysis of these ratios enable businesses to make data-driven decisions and optimize their operations.

Customer Demographics, such as average number of adults, children, and babies per booking or most common countries of origin for guests has been identified by the spark query. analyzing the average number of adults, children, and babies per booking, along with identifying the most common countries of origin for guests, empowers hotels to tailor their services, target marketing efforts effectively, and enhance the overall guest experience.

Average daily rate (ADR) has been calculated, which is based on the total revenue per booking or total revenue per hotel type is a very important metric from business perspective.

Booking trends over time and average days in waiting list being measured to identify which is the peak booking month or date of the year.

Using solr, we have demonstrated the distribution channel for the booking done for country code 'ESP' for arrival year 2017.

In conclusion, leveraging queries on hotel booking data provides significant benefits for users and businesses, enhancing the overall experience and operational efficiency.

However, addressing privacy concerns, ensuring data quality, and adapting to the dynamic nature of the industry are essential for successful implementation. Continuous monitoring and a strategic approach are crucial for deriving actionable insights and staying competitive in the hospitality sector.

## Limitations and Challenges:

Privacy Concerns:

Collecting and analyzing guest data may raise privacy concerns. Businesses must adhere to data protection regulations and implement secure data handling practices.

- Data Quality and Completeness:



The effectiveness of the analysis heavily relies on the quality and completeness of the data. Inaccuracies or missing information can lead to unreliable insights.

- **Dynamic Nature of the Industry:**

The hotel industry is dynamic, with factors like seasonal variations and external events affecting booking patterns. Historical data might not always accurately predict future trends.

- **Integration Challenges:**

Integrating data from various sources, such as booking platforms, customer feedback, and operational data, can be challenging but is crucial for a holistic analysis.

- **Continuous Monitoring and Adaptation:**

Trends and user behaviors evolve. Continuous monitoring and adaptation of strategies are necessary to stay relevant and competitive.

- **Technical Expertise:**

Extracting meaningful insights requires technical expertise in data analysis and interpretation. Businesses may face challenges if they lack the necessary skills or resources.

- **Interpreting Causation vs. Correlation:**

Understanding the cause-and-effect relationships between different variables can be complex. Correlation does not always imply causation, and misinterpretation may lead to incorrect business decisions.

In conclusion, leveraging queries on hotel booking data provides significant benefits for users and businesses, enhancing the overall experience and operational efficiency. However, addressing privacy concerns, ensuring data quality, and adapting to the dynamic nature of the industry are essential for successful implementation. Continuous monitoring and a strategic approach are crucial for deriving actionable insights and staying competitive in the hospitality sector.