German International University of Applied Sciences Informatics and Computer Science
Dr. Nada Sharaf
TA. Mariam Ali
TA. Omaima Ahmed

Big Data & NoSQL Databases, Spring 2025
Assignment 2
Due date is Wednesday, May 7th, 2025 at 11:59 PM Submitted in groups of maximum 2

For this assignment, you'll use Apache Spark to analyze the Hotel Booking Demand dataset, which captures real-world booking details for City Hotel and Resort Hotel. The dataset includes information like booking dates, length of stay, number of guests (adults, children, babies), pricing, parking spaces, and cancellation status. Your task is to preprocess the data, create derived columns (e.g., total stay duration), and answer business questions, such as which countries have the highest cancellation rates, using both SparkSQL and Spark DataFrame API. *Column descriptions are on page 3

You are required to:-

1. **Start a Spark Session and Load the Dataset**
   a) Initialize a Spark session.
   b) Load the egphotelbookings.csv file into a Spark DataFrame.
2. **Preprocessing using Spark DataFrame API**
   a) Explore the data well and clean any noise.
   b) Handle duplicates and missing values appropriately .
   c) Create new derived columns: Total stay duration, Season and Total guests.
3. **Implement the following queries TWICE: once using SparkSQL and once using Spark-Dataframes. Ensure both versions return the same (or very similar) results.**
   a) Compute the cancellation rate as (number of cancellations / total bookings)*100 per country, and list the top five countries have the highest cancellation rates.
   b) Identify which season (Winter, Spring, Summer, Fall) has the highest cancellation rate for bookings with lead time > 100 days. (Define seasons based on arrival_date_month)
   c) Which reserved room types experience the highest mismatch with the assigned room type? (Where they differ).
   d) Which distribution channel shows the lowest cancellation rate and what is its average revenue per booking (ADR x total nights)?
   e) Which meal types are most common among bookings with more than 3 total guests?

Bonus (Optional):

- Show the relationship between **lead time** and **cancellation status** using a chart (e.g., scatter, boxplot).
- You are **not allowed to use a correlation function**. Instead, describe the trend in a sentence.

Deliverables

- Your code is to be submitted to **bigdata602.25@gmail.com** as a zip file containing your notebook (include the names and IDs of the team members in the body of the email with **subject:** "Assignment 2 S25").
- You will be evaluated **based on the submitted notebook** before the deadline only.

Evaluation Guidelines

- If AI was used for any part, you **must bring the exact prompt** used during your evaluation.
- If **heavy reliance on AI is detected** (e.g., copy-pasted queries, uploading file) and **no prompt is provided**, you will receive a **grade of zero**.
- You will be asked for the **reasoning for data cleaning choices** and **query logic.**

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH TEAMS INVOLVED or IF YOU COPIED IT FROM THE INTERNET OR ELSEWHERE (NO. EXCEPTIONS.)!

Column	***Description***
hotel	The name of the hotel (either Renaissance Hotel or JW Marriott Hotel)
is_canceled	Value indicating if the booking was canceled (1) or not (0)
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number of year for arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number of adults
Babies	Number of babies
meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	Country of origin
distribution_channel	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
is_repeated_guest	Value indicating if the booking name was from a repeated guest(1) or not (0)
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons
assigned_room_type	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
agent	ID of the travel agency that made the booking
company	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for designation for anonymity reasons
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	Number of car parking spaces required by the customer
total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor)