Capstone project

Objective: Analyze the restaurants data by applying the business rules defined.

Outputs

1. Solution should ensure all that standard and coding best practices are followed, designed keeping performance considering the larger dataset and complex business rules to be applied in future
2. Demo of the working solution

Key concepts to be used

- Spark RDD & DataFrame
- Transformations & actions
- windowing functions
- udf functions
- use persist/cache where ever required

Dataset:

Download the data from https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

and convert the data to Avro (with nested avro schema for composite columns) or parquet with nested format.

Process and rules

First do clean up :

1)Remove all non-ascii characters from all columns : Name, Location etc

2) Remove restaurants with no ratings.

From the dataset:

1) Filter out records which have invalid restaurant links (use regex to take main part of the url) – which means that the restaurant is most probably closed now.
2) Group by Address location for the closed restaurants and find out which area has the most restaurants getting closed.
3) For the restaurants which are still working, group by restaurant type and location and find out the restaurants which have highest rating for each cuisine type.

4) For the reviews list column, find the distribution of star rating, on the condition that there are - at-least 30 ratings for that restaurant.
5) Group by location for individual cost buckets (for 2 people) : [<=300, 300-500, 500-800, >= 800] and take the 5 highest rated restaurants in each location and each cost bucket and save as parquet file.
6) For the restaurants which are not closed, publish the data into individual kafka topics based on location, so that downstream processes can customize promotions based on this.

Technical considerations:

- Setup cluster
- Configuration
- Calculate number of executors and executor memory is required to run the spark job.
- Use funsuite or scalatest to test functionality in local with multiple test cases with sample data.