Customer Segmentation and Analysis using Yelp Reviews Dataset

Akanksha Tyagi
Department of Applied Data Science
San Jose State University

Anjali Ojha
Department of Applied Data Science
San Jose State University

Sakshi Mukkirwar
Department of Applied Data Science
San Jose State University

Swati
Department of Applied Data Science
San Jose State University

Keerthana Raskatla
Department of Applied Data Science
San Jose State University

***Abstract*—In today's data-driven world, restaurants strive to acquire a competitive advantage by analyzing diner preferences and behavior. Yelp, as a platform for user-generated reviews and ratings of businesses, offers a rich source of data that can be leveraged to understand customer behaviors and preferences. The Yelp restaurant reviews dataset, which contains millions of diner reviews, ratings, and restaurant profiles, customers' profile, their social networks, is a significant resource for such customer insights. This research highlights our intention to use Big Data Analytics to do consumer segmentation utilizing this rich Yelp dataset. As the size of data grows, identifying a segment with desired attributes is a core problem of any Big Data Marketing application. As part of our Big-Data Applications project, we are going to explore the Yelp Restaurant Reviews dataset and find out how we can make segmentation more effective and fast. As the data size keeps increasing, querying a large amount of data becomes very time-consuming and computationally expensive. For any modern marketing system having a great segmentation engine increases the time to market.**

***Index Terms*—Yelp, Big Data Analytics, Consumer segmentation, Restaurant**

I. INTRODUCTION

In an increasingly competitive business landscape, understanding customer's preferences, behaviors, and sentiments has become pivotal for companies striving to provide personalized experiences and targeted services. Customer segmentation involes segregating customer into separate groups that are homogeneous in themselves [12]. The Yelp dataset, a rich repository of user-generated reviews and ratings, offers an extensive platform for businesses to glean valuable insights into customer sentiments and behaviors. In today's data-driven world, businesses strive to acquire a competitive advantage by analyzing customer's preferences and behavior. Through the extensive data available on Yelp, this study employs advanced data analytics techniques to decode customer behaviors, preferences, and sentiments related to their experiences, and Yelp gave users a platform to register their views. Although Yelp has data collected from different market segments, but the food, dining, and restaurant segment dominates that. We can see the same in the figure below and we are only plotting the top 30 categories using 1.0 percent of the data sample.

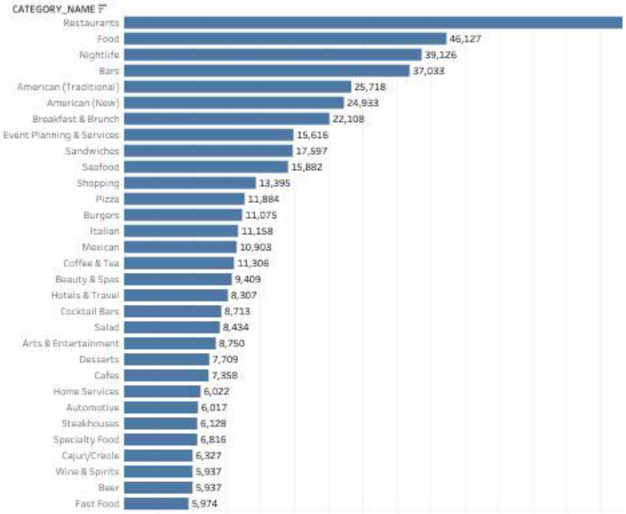


Fig. 1. Market Segment Distribution, for the top 30 categories using 1.0 percent of the data sample

This study utilizes advanced analytics to uncover intricate customer behaviors, preferences, and what they think of regarding dining in the Yelp dataset. It goes beyond conventional analysis to implement sophisticated segmentation, identifying distinct customer groups based on nuanced preferences. The aim is two-fold, not only deciphering complex customer sentiment patterns but also generating actionable insights to help businesses tailor their offerings. This facilitates increased satisfaction and loyalty in a dynamic market. In essence, the research strives to both decode subtle dining preferences through advanced segmentation and translate the findings into practical guidance that enables businesses to better serve diverse customers [13].

II. SIGNIFICANCE TO THE REAL WORLD

The project illustrates the real-world significance of Big Data technologies by showing how they empower businesses to make data-driven decisions, improve operational efficiency, and gain a competitive edge [1]. Adopting reliable and highly resilient programs like Kafka and Hadoop, helps businesses cope with enormous amounts of information, therefore making operations more efficient [2] [3]. Data warehousing from the cloud can do on-demand scalability cheaply, as well as process massive amounts of information. Using tools such as Apache Spark and Tableau helps one get actionable insights that can build customer customer-centric strategy that ultimately drives innovation. Ultimately, the research shows how large-scale, diverse Yelp data must be handled with efficient data management techniques. Overall, this project has the potential to make a significant impact on the restaurant industry and beyond by providing valuable insights into customer behavior and preferences.

III. MOTIVATION

The primary objective for this research project lies in the ever-increasing significance of data-informed decision-making within the contemporary business landscape. The Yelp dataset encapsulates invaluable insights regarding businesses, user reviews, and user profiles. It has a comprehensive array of business details, encompassing names, locations, attributes, and categories, as well as user reviews containing both star ratings and textual content. Moreover, user profiles furnish detailed reviewer information, inclusive of names, review counts, and voting statistics. The dataset further extends to cover check-in records, tips, and business-related photos. This comprehensive repository presents businesses with an unparalleled opportunity to leverage data insights for a nuanced understanding of customer behaviors and preferences.

IV. LITERATURE SURVEY

The surge of online platforms and user-generated content has granted researchers and businesses unprecedented access to valuable data sources. Among these, the Yelp dataset, primarily housing restaurant reviews and ratings, has garnered significant attention in the realm of Big Data analytics. According to [4], their research focused on developing effective recommendation systems to combat the prevalent information overload witnessed on review websites. Employing a modified Latent Aspect Rating Analysis [8] technique, the study identified five notable features. Particularly noteworthy was their emphasis on the influential role of ”restaurant value” and ”food and drinks” in shaping sentiment and satisfaction levels. The authors advocated for dynamic recommendation systems that account for attribute-specific evaluations, aiming to enhance the user experience while navigating through extensive review data. In the research done by [5], findings underscored the influence of reviewer credibility on reader perception, demonstrating the potential of big data analytics in extracting insights from extensive online review databases to guide consumer decision-making. The study performed by [6] revealed the significance of framing, argument quality, and moderate ratings in eliciting reader engagement. This exploration shed light on the amplifying role of heuristics in the impact of evaluations and the consequential value they add to the platform.

According to a study by [7], a pioneering market segmentation approach was introduced, utilizing online consumer reviews to profile both customers and businesses. Their innovative methodology offered comprehensive insights for focused segmentation strategies on social media, capitalizing on publicly available consumption details embedded within online reviews. The research by [11] concentrates on utilizing recommendation system principles to construct a predictive model for forecasting customer ratings of un-visited businesses. It harnesses Yelp's expansive data to extract collaborative and content-based features for discerning customer and restaurant profiles. Various models including generalized regression, ensembles, collaborative filtering, and factorization machines are applied. Root Mean Squared Error(RMSE) evaluation enables comparative analysis of model effectiveness. For cold start issues [14], segmentation ensembles and three imputation techniques - mean, random, and predicted values - tackle missing information.

In summary, this comprehensive study leverages recommendation approaches on rich Yelp data to predict ratings via an array of models, evaluated through RMSE. Segmentation and imputation address cold start and missing data challenges. The study in [9] investigates the convergence of Online Marketing, Customer Segmentation, and Big Data Analytics amidst today's highly competitive online business landscape. It highlights the growing importance of online marketing strategies for customer engagement in the digital realm. The study underscores performing customer segmentation based on online data and addresses the formidable challenges posed by the massive data volumes, which can be managed and analyzed through Big Data technologies. An Online Customer Segmentation (OCS) framework is presented to demonstrate how Big Data tools can support online marketing goals. The framework outlines key online marketing objectives, contrasts offline and online customer attributes, defines OCS categories, and introduces relevant Big Data concepts and tools [1]. A hypothetical business scenario applies the OCS framework on an online customer dataset to illustrate its implementation. In summary, this paper provides an OCS framework exemplifying how Big Data analytics can enable effective online marketing and customer segmentation amidst expansive online data.

These studies collectively highlight the potential of the Yelp dataset, showcasing its relevance in understanding customer segmentation and behaviors, and its transformative impact on businesses through data-driven insights. This literature survey serves as a foundational cornerstone for our project.

V. PROJECT OVERVIEW AND ARCHITECTURE

In the project, we employ a robust Big Data technology stack that helps in processing and analyzing complex and huge Yelp datasets. The most critical aspect of our system design is the cloud, and we are using Amazon Web Services for all our needs. It is the place where the data is kept and processed. With regards to handling and processing large volumes of varied data, distributed computing with Spark enables this. Kafka brokers provide an effective and prompt mechanism for handling streaming data in real time. This is how all types of data, such as reviews, check-ins, and tips come in the form of streams and we incrementally process those datasets.

The user-friendly feature of Amazon S3 is its ability to scale up and down. It is also a distributed file system and it's an extension of the Hadoop File System. This way provides us with a secure and reliable methodology for storing our static and processed files that comprise all data concerning businesses and users. Data processing is done using Apache Spark, recognized for its high-speed in-memory computing. Spark helps us in handling our tasks related to data processing such as aggregation or trait extraction by speeding up the work. The data gets stored in a data store of Snowflake for faster analytics and also in parquet file format, which will later be used in the hive tables to extract final segments.

Snowflake offers cloud-based data warehousing capabilities while Hive allows users to do queries and processing on extremely big datasets using a SQL-like language. We use Tableau, which visualizes our data; and it matches well with our data stores. The dashboard allows us to view and communicate with visualizations, which enables an understanding of how the various types of customers relate. All intermediate datasets are also stored in cluster, but in parquet format for easy retrieval when needed. This is necessary in order to achieve speedy and effective data processing.

We also built a Streamlit app to visualize how each slicing and dicing of the data changes the desired segment. We put a widget to capture all the filters and data exploration in the form of a query, and later that query can be executed against the raw parquet files to extract the desired segment.

VI. DATA, TOOLS AND TECHNOLOGIES

A. Dataset

This project utilizes over six million reviews on thousands of restaurants in eleven cities taken out of Yelp. This is a key dataset consisting of information from 1,987,897 customers contributed through purchases that have been packaged in compressed **.tar** files approximately weighing 11.8GB. In the JSON format, each file contains an isolated object type. The JSON objects are in one line. This comprehensive database provides details on different restaurant profiles offering detailed information on served foods, atmosphere, and others. In addition, the dataset goes into depth about customers' demographics, behavioral characteristics, and many more aspects. This dataset should be noted that it is not an artificial and simulated one, but it has been obtained from real business and reviews. Thus, we have the chance to look at how restaurants and businesses are struggling in the competitive situation and environment of online reviews and customer engagement [10]. The realism of the dataset becomes the central issue as we move forward to analyze it. It is this realism that ensures that our findings and conclusions rest on the authentic experience of all of us.

Table VI-A has all the dataset level details [].

TABLE I
YELP DATA DETAILS

Dataset Name	Number of Records
yelp_academic_dataset_business.json	150346
yelp_academic_dataset_user.json	1987897
yelp_academic_dataset_checkin.json	131930
yelp_academic_dataset_review.json	6990280
yelp_academic_dataset_tip.json	908915

B. Data Cleaning and Processing

All the datasets are in JSON format and 1 file for each dataset mentioned above. It was fairly easy to load the data with spark into the dataframe, but on further analysis, we can see how the different attributes for different entities are being captured and we have to bring each of the attributes for the entity as a common format. This common format helped in extracting the desired attributes easily. As the data was quite big, we had to use sampling while maintaining the relations with other entities to avoid data sparsity.

C. Feature Engineering

The Yelp data is quite extensive and so much can be learned from this data alone. The other data sets on top of the reviews help to understand user behaviors. For our application, we explore users' likes and dislikes, user demographics (where they live), their behaviors (where they frequently travel, how they express their sentiments, how they rate each business category, and their social network). For customer segmentation, all this information is helpful for marketers to reach their targeted audience.

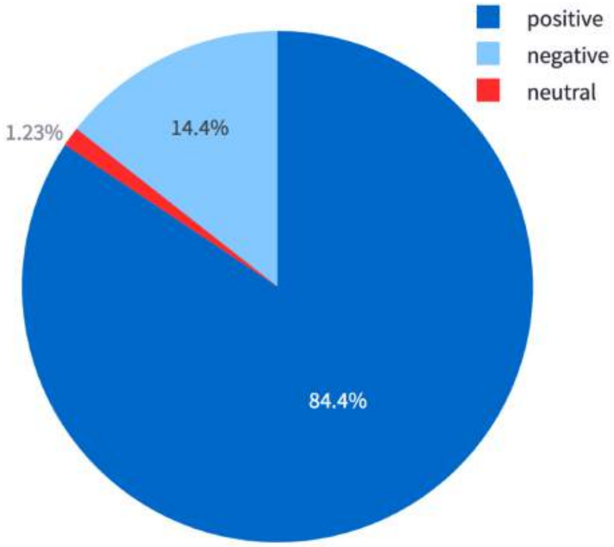


Fig. 3. Distribution of the sentiments from the review

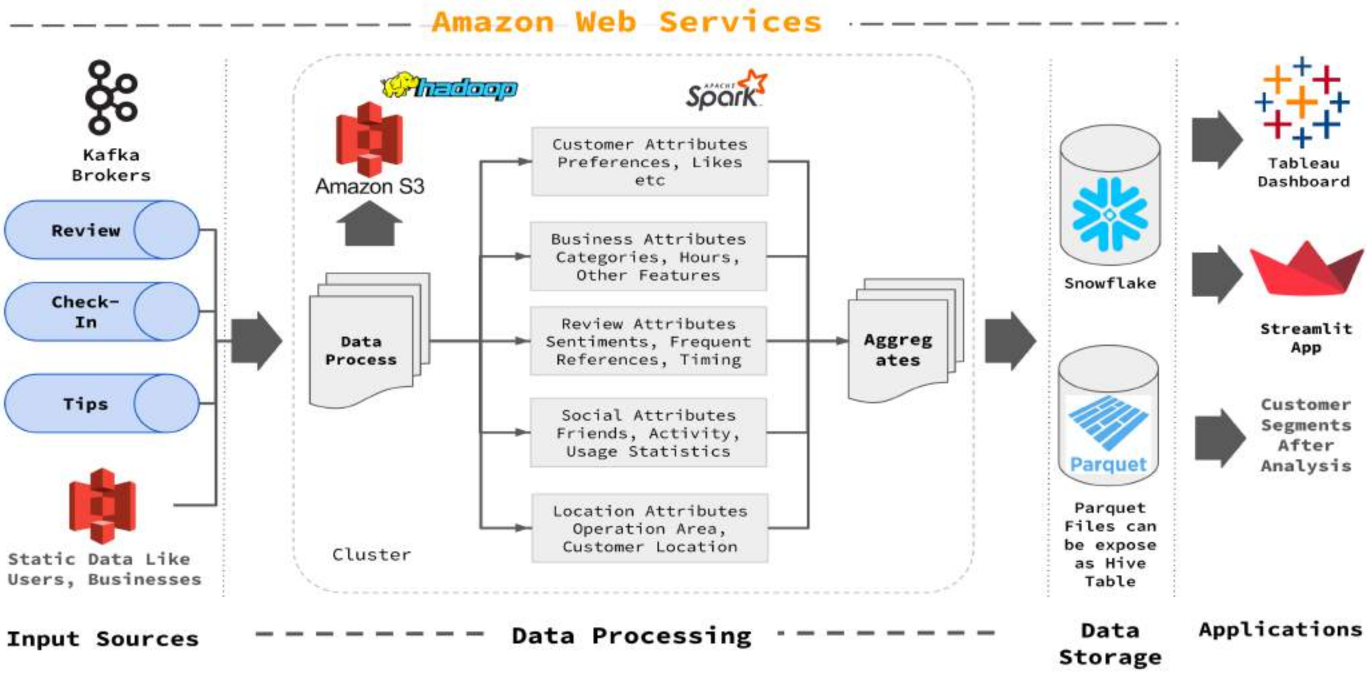

Amazon Web Services
Kafka Brokers
Review
Check-In
Tips
Static Data Like Users, Businesses
hadoop
Spark
Amazon S3
Data Process
Customer Attributes Preferences, Likes etc
Business Attributes Categories, Hours, Other Features
Review Attributes Sentiments, Frequent References, Timing
Social Attributes Friends, Activity, Usage Statistics
Location Attributes Operation Area, Customer Location
Cluster
Aggreg ates
Snowflake
Parquet
Parquet Files can be expose as Hive Table
Tableau Dashboard
Streamlit App
Customer Segments After Analysis
Input Sources
Data Processing
Data Storage
Applications

Fig. 2. System Architecture Diagram And Setup in cluster

D. Exploratory Data Analysis

Once we extracted the desired features for our use case, we did the data analysis using Tableau. As the size of the data was huge, we picked the top 1 percent (based on the number of reviews) of the users and analyzed their behaviors. This shows how the data is skewed for the food and restaurants compared to other categories, and how people tend to give higher ratings. We can see the distribution of the reviews in the figure-VI-C.

E. Visualizations

We created multiple visualizations using Tableau and Streamlit. We use these visualizations to find the patterns in the data and understand the data. In figure VI-E, we show the change of sentiments over time. In figure VI-E, we show how many reviews are received in each year. Figure VI-E shows the frequency distribution of review counts by user activity. We also show the correlation of different sentiments of reviews in figure VI-E.

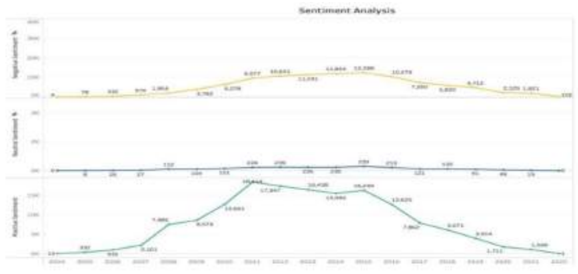


Fig. 4. Distribution of the sentiments every year.

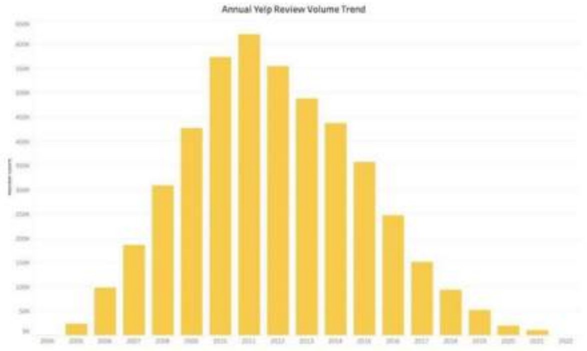


Fig. 5. Distribution of the review over time.

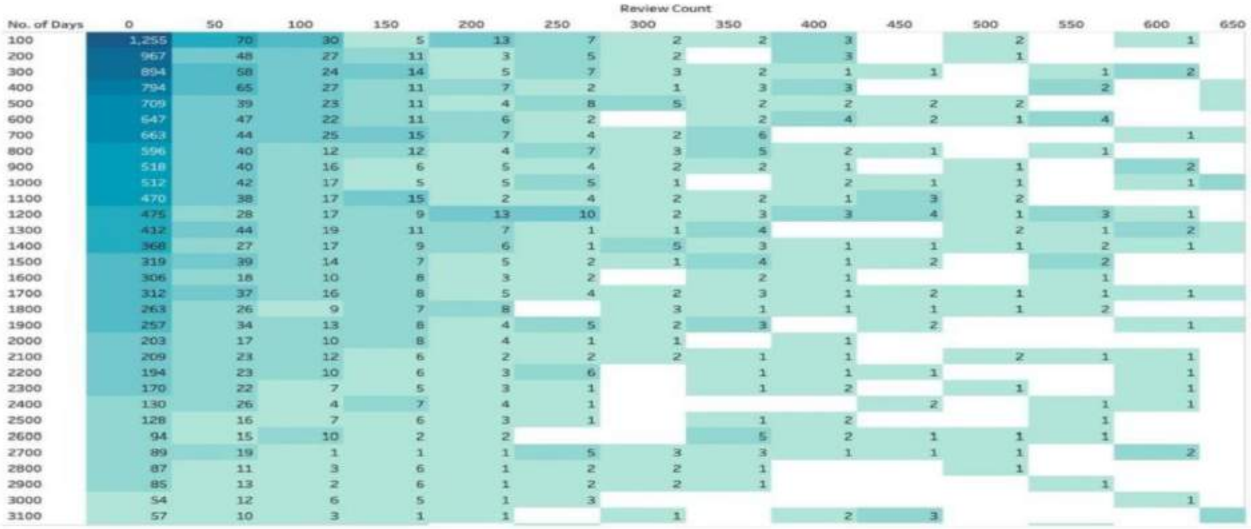
Table II had the full list of tools and technologies we used for this project. i

VII. TECHNICAL DIFFICULTIES

The Yelp dataset, characterized by its user-generated and customer-facing nature, presents analytical challenges. Even though the data was given in proper JSON form, it had its own challenges. The overarching challenge lies in balancing preprocessing, efficiency, and scalability to extract meaningful insights from the voluminous and diverse dataset. We can put these difficulties in the following categories.

A. Scale of the Data

The sheer scale necessitates strategic sampling to glean meaningful insights within system constraints before scaling up to the full dataset. The data size of 10+ GB at disk, once loaded in the memory for processing it grows even more,



						Review Count							
No. of Days	0	50	100	150	200	250	300	350	400	450	500	550	600	650
100	1,255	70	30	5	13	7	2	2	3		2		1	
200	967	48	27	11	3	5	2		3		1			
300	894	58	24	14	5	7	3	2	1	1		1	2	
400	794	65	27	11	7	2	1	3	3			2		
500	709	39	23	11	4	8	5	2	2	2	2			
600	647	47	22	11	6	2		2	4	2	1	4		
700	663	44	25	15	7	4	2	6					1	
800	596	40	12	12	4	7	3	5	2	1		1		
900	518	40	16	6	5	4	2	2	1		1		2	
1000	512	42	17	5	5	5	1		2	1	1		1	
1100	470	38	17	15	2	4	2	2	1	3	2			
1200	475	28	17	9	13	10	2	3	3	4	1	3	1	
1300	412	44	19	11	7	1	1	4			2	1	2	
1400	368	27	17	9	6	1	5	3	1	1	1	2	1	
1500	319	39	14	7	5	2	1	4	1	2		2		
1600	306	18	10	8	3	2		2	1			1		
1700	312	37	16	8	5	4	2	3	1	2	1	1	1	
1800	263	26	9	7	8		3	1	1	1	1	2		
1900	257	34	13	8	4	5	2	3		2			1	
2000	203	17	10	8	4	1	1		1					
2100	209	23	12	6	2	2	2	1	1		2	1	1	
2200	194	23	10	6	3	6		1	1	1			1	
2300	170	22	7	5	3	1		1	2		1		1	
2400	130	26	4	7	4	1				2		1	1	
2500	128	16	7	6	3	1		1	2			1		
2600	94	15	10	2	2			5	2	1	1	1		
2700	89	19	1	1	1	5	3	3	1	1	1		2	
2800	87	11	3	6	1	2	2	1			1			
2900	85	13	2	6	1	2	2	1				1		
3000	54	12	6	5	1	3							1	
3100	57	10	3	1	1		1		2	3				

Fig. 6. Frequency distribution of review counts by user activity.

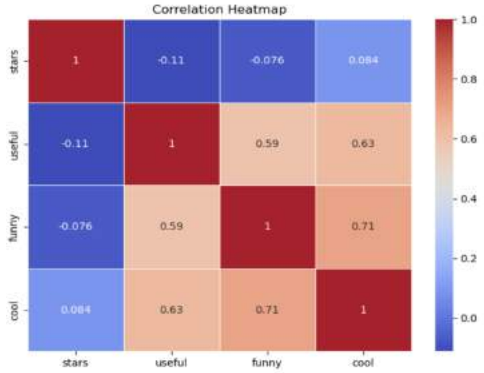


Fig. 7. Correlation of different sentiments.

and processing that entire on local development instances was difficult.

B. Data Sparsity

The data here is collected using many users and businesses and it a generalized data for different market segments which makes it sparse for all the attributes. Also, there is further sparsity resulting from incomplete samples requires focused analysis of smaller, filtered segments to facilitate pattern identification. To get a good sample of have to check it against all the other datasets and extract meaningful insights needs a lot more work. Even after good sampling, other important fields were missing.

TABLE II
TOOLS AND TECHNOLOGY USED

Tools	**Usage**
Python	Main coding language for all the development tasks
Spark	For all data processing
Kafka	To ingest the review data as stream
Snowflake	For data warehousing, for the purpose of Visualization and EDA. Use SQL to fetch the desired data.
Tableau	Data Visualization and Data Analysis
Streamlit	To build a data app for demo purposes and hosting.
AWS	As main cloud infrastructure where all jobs run. EC2 machine for the deployment.
S3 Hadoop	For data storage in the cluster.
Parquet	We use Parquet format for all the intermediate data storage
GitHub and Copilot	For code collaboration and Pair programming
Overleaf (Latex) and Grammarly	Report Writing and Proof Reading
Prezi and Google Slides	Presentation and different diagram creation
Trello	For Agile practice

C. Unstructured Data

The unstructured JSON format introduces complexities, demanding custom transformations for analytical accessibility, especially in handling categories and business attributes with varying structures. We have carefully processed multiple records to get a standard schema for the datasets. For example, the attributes field in the business dataset has many values and each value changes the context of the business. We have to write manual schemas to convert each to a common format.

D. Data Representation

As the Yelp dataset has a lot of features we were adding more derived attributes like Sentiment, social group size, frequent words, etc., on top of it, which resulted in added complexity. We are creating a data cube by merging all attributes in one big table, which creates higher dimensional data, to save the same in the traditional storage system, we have to use complex data structures and it helped us keep the size manageable but impacting the data reads for visualization and analysis.

E. Resource Constraints

As the Yelp data is big in size and is also skewed for popular users or businesses, we have limited processing capabilities, and we have to face a lot of *Java Out Of Memory exceptions* while processing the data. We show the skewness of data in figure **??**. As we flattened the data across multiple categories, it was calculated more than 100 million combinations, and these issues are very hard to debug.

F. Integration and Cloud Setup

As there are many tools involved in the project, integrating all of them into a single project has its own challenges. Finding the right versions for each dependency is difficult, in our project, we have to deal with Kafka, Snowflake, AWS S3, Spark, Python, NLTK, streamlit, and Hadoop dependencies, we have to do a lot of searching and trial and error to address these issues. Compounding these issues with the setup in the cloud makes it more challenging.

VIII. TEAM WORK

During the course of the project, each member was involved during each step of the project development, design and implementation. But we divide the responsibilities that who will lead what aspect of the project and the person will be responsible for the delivery of that module.

IX. NOVELTY UNIQUENESS

This project uniquely leverages real-world Yelp data to benefit businesses through an exploration extending beyond analysis. Our approach is tailored to businesses' distinct needs and goals. Focusing on restaurants, we recognize their specific challenges and opportunities with online reviews and customer engagement. The project encompasses all customer-related data, including demographics and behaviors, structured as data cubes. Diverging from conventional analyses, our project aims to deliver actionable, tailored strategies to empower restaurants in effectively targeting and engaging their audience. Rather than mere analytics, our solution is designed end-to-end to provide restaurants with practical insights and guidance by delving into their nuanced customer data.

TABLE III
TEAM MEMBERS AND ROLES

Team Member	**Role and Responsibilities**
Akanksha Tyagi	• Conceptualization and Formal Analysis • Exploratory data analysis with Spark. • Set up Kafka jobs for consuming streaming data and storage. • Jira Board (Trello) Management.
Sakshi Manish Mukkirwar	• Visualization with Tableau. • Streamlit app setup with different graphs and queries. • Presentation slides.
Swati Verma	• Investigation of data to identify useful attributes for targets. • Generate Features for Segmentation. • Writing – Original Draft, and overleaf setup.
Keerthana Raskatla	• Data Curation and Refining Data. • Wrote jobs for data cleaning and processing. • System architecture and other diagrams.
Anjali Himanshu Ojha	• Methodology - created a pipeline to bring all processes together. • Resources - setup cluster, GitHub, and Streamlit app deployment. • Writing – Review Editing, and experiment with Github-Copilot.

X. RELEVANCE TO THE COURSE

Throughout the project, we adhered to a structured process aligning with course teachings. The project began with a thorough assessment of the Yelp dataset to understand its structure and contents, aligning with the course's focus on managing and analyzing large data volumes. We then leveraged Apache Kafka for data ingestion and we utilized Amazon S3 and Hadoop for storage, with Apache Spark for data processing, practicing the distributed and parallel processing emphasized in the course. This comprehensive use of big data technologies culminated in the deployment of our analytics on Snowflake and the creation of insightful visualizations with Tableau, ensuring that every phase of the project was a practical application of the big data concepts and tools discussed in the course.

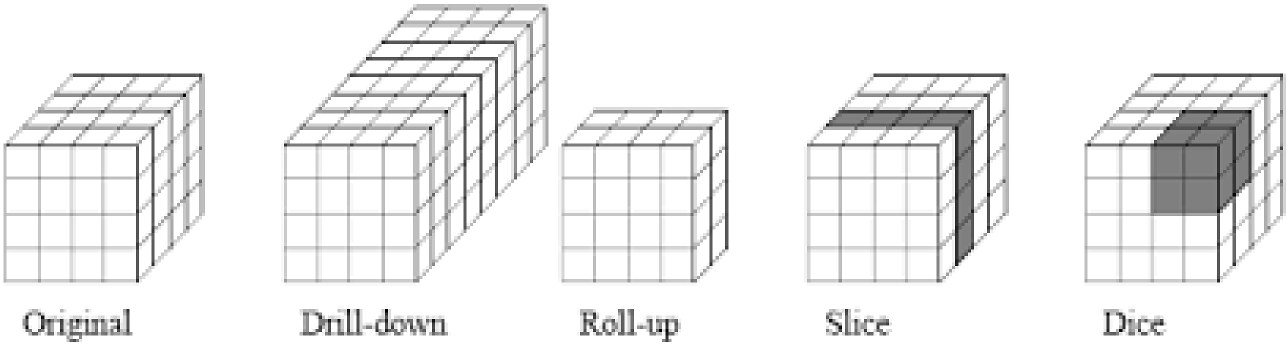


Fig. 8. Data Cube Slicing and Dicing

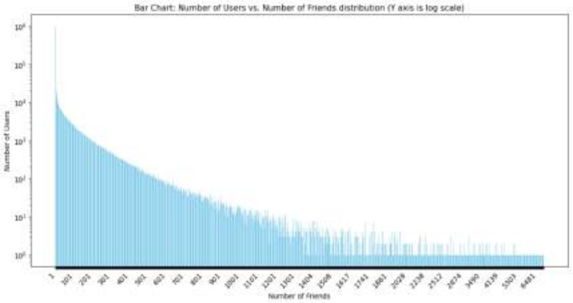


Fig. 9. Data skewness

XI. INNOVATION

The project's novelty is derived from its adept incorporation of several contemporary data processing technologies for the purpose of analyzing and interpreting enormous quantities of disparate data. The system acquires raw data from several input sources, including Kafka Brokers and static data sets. It leverages the storage and processing capabilities of Amazon Web Services, which offer robustness and scalability. Hadoop and Apache Spark are utilized for the purpose of executing distributed data processing, hence facilitating the effective management of intricate analytics activities.

The project distinguishes itself through the consolidation of many data kinds, encompassing client preferences, social traits, and location information. These data are then saved in Snowflake, a cloud-based data platform renowned for its exceptional performance and user-friendly interface. The processed data is utilized in various programs such as the Tableau Dashboard, which enables visual analytics, and Hive, which supports data warehousing. These apps play a crucial role in supporting valuable consumer segmentation and enhancing corporate intelligence. The integration of different technologies in an effective way signifies a notable progression in frameworks for making decisions based on data, indicating a comprehensive comprehension of the immense potential of Big Data in revolutionizing company strategies and results.

XII. IMPACT

The project leverages a robust big data infrastructure to efficiently process diverse data streams, utilizing Kafka, Hadoop, Spark, and cloud storage solutions like Amazon S3 and Snowflake which is used to organize and analyze large amounts of data, making it easier for businesses to understand their customers and make more informed decisions. The ultimate goal was to provide businesses with more knowledge about what's going on in their industry so they could better service their consumers, close deals, and stay ahead of the competition. Modern technology, which can process large amounts of data and expand with the needs of the company, was used to do all of this.

XIII. LESSON LEARNT

For any big data project data and how we are going to use it are two main ideas that drive all the requirements around it. If data is used for visualization, then aggregates will be required and any relational database solution will work. But in our project, we are allowing users to create segments after drilling down the data and its analysis. The final output has to be a segment with a list of users who can be targeted.

While don't do the development we can't work with the full scale of the data, so we have to sample it. A good sample goes a long way keeping all the relations in the data, it is small enough that can be handled on the laptop but diverse enough to capture all the nuances of the data. It saves a lot of time and computing resources, and we can so many ideas with the small data. However, creating such a diverse sample is itself challenging and it takes a lot of time and a full understanding of the domain and data.

Real-world data is not a CSV file which nicely curated and we have to use it in our processing. A large chunk of effort goes into the cleaning and curation of the data itself. We learn this by analyzing the Yelp data set and what will be the best format for our use case. So we have to spend a lot of time on the data cleaning and putting it in the standard format. Once data is cleaned it's easy to work with.

This project taught us that Big Data technologies require a variety of tools and methods to evaluate vast and complicated datasets. Using Kafka for data ingestion with Amazon S3 and Hadoop for storage and processing allows the system to manage massive volumes and maintain fault tolerance. In-memory processing makes Apache Spark suitable for complicated, large-scale analytics workloads by transforming and aggregating data. The integration of Snowflake shows the scalability, administration, and cost-effectiveness of cloud-based data warehousing. Tableau for visual analytics emphasizes the importance of turning data into visual insights for business decision-making. This shows the necessity of choosing the correct Big Data technologies to construct a durable, efficient, and comprehensive analytics platform for data-driven decision-making.

We also learn to work with contained resources. While working with small data we get careless about how the system gonna handle that, but during this project, we learn we are not always gonna get endless resources and we need to optimize our jobs to handle that. We also learned to deploy a full-fledged app to the different environments. We deployed a Streamlit app to demonstrate how the segmentation will work.

XIV. Pair Programming

Pair programming improved our project's progress. We carefully designed and reviewed our data processing application code in pairs to ensure reliability and quality. Integrating Kafka, Hadoop, and Spark was complicated, therefore this collaborative approach was invaluable. We ensured data analysis algorithm correctness and accelerated problem-solving by switching between driver and navigator roles. Pair programming helped us exchange knowledge and insights regarding our project's data architecture, boosting team competence and cohesion.

XV. Discussions and Conclusions

As we conclude our project, our data processing platform has greatly improved enterprises' data interaction and understanding. This project has shown us the power of real-time data processing and analytics. We've built a solid infrastructure that satisfies our initial goals and scales for our data-driven demands using Kafka, Hadoop, Spark, Amazon S3, and Snowflake. Pair programming helped us maintain high-quality code and collaborate. It helped us easily combine multiple technologies and efficiently handle complicated data structures. Our data research has helped organizations make better decisions and better meet client needs. But every enterprise has its obstacles. We solved data consistency and integration problems with adaptive problem-solving and collective expertise. These experiences have deepened our grasp of big data difficulties and prepared us for future complications. Moving forward, we see room for improvement. Machine learning methods for predictive analytics and advanced data visualization techniques could provide value.

References

[1] Srinivasa, S., Bhatnagar, V. (2012) Big Data Analytics. Berlin: Springer.

[2] Agneeswaran, V. S. (2012). Big-Data : Theoretical, Engineering and Analytics Perspective. In Big Data Analytics (pp. 8-15). Berlin Heidelberg: Springer.

[3] MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[4] Luo, Y., Tang, L. (Rebecca), Kim, E., and Wang, X. (2020). Finding the reviews on Yelp that actually matter to me: Innovative approach of improving recommender systems.

[5] Lee, M., Kwon, W., and Back, K.-J. (2021). Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making.

[6] Meek, S., Wilk, V., and Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews.

[7] Moon, S., Jalali, N., and Erevelles, S. (2021). Segmentation of both reviewers and businesses on social media.

[8] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). Association for Computing Machinery, New York, NY, USA, 783–792. https://doi.org/10.1145/1835804.1835903.

[9] Fotaki, Georgia & Spruit, Marco & Brinkkemper, Sjaak & Meijer, Dion. (2014). Exploring Big Data Opportunities for Online Customer Segmentation. International Journal of Business Intelligence Research. 5. 58-75. 10.4018/ijbir.2014070105.

[10] Fotaki, G., Gkerpini, N., Triantou, A, I., Brinkkemper, S. (2012). Online Customer Engagement Management. Utrecht University.

[11] Ting, J., & Ramaswamy, S. I. (2013). Yelp Recommendation System.

[12] Tsiptsis, K., & Chorianopoulos, A. (2009). Data Mining Techniques in CRM: Inside Customer Segmentation. Wiley.

[13] Lee, J., & Park, S. (2005). Intelligent profitable customers' segmentation system based on business intelligence tools. Expert Systems with Applications, 29(1), 145–152.

[14] Padilla, N., & Ascarza, E. (2021). Overcoming the Cold Start Problem of Customer Relationship Management Using a Probabilistic Machine Learning Approach. Journal of Marketing Research, 58(5), 981-1006. https://doi.org/10.1177/00222437211032938.

Appendix

A. Rubric

Abstract - The abstract conveys an overview of the key aspects of a document. It serves as a brief yet comprehensive representation to quickly grasp the essential content.

Motivation-It serves as a catalyst for sustained effort in the face of challenges, influencing both the initiation and persistence of actions.

Literature Survey-The literature survey involves a comprehensive review and analysis of existing scholarly works and publications within a specific field.

Methodology - The methodology details our infrastructure for the project along with the step-by-step of the project's progress.

Team members and their roles - Team members with their respective responsibilities are present.

Relevance to the Course - This project is aligned to the course Big Data.

Technical Difficulty - We faced multiple challenges which were resolved.

Novelty Uniqueness - We used strategies to empower restaurants in effectively targeting and engaging their audience.

Lesson Learned- Learnt variety of tools and methods to evaluate vast and complicated datasets.

Grammarly - We used Grammarly for perfection in our writing. The screenshot is included in the document shared.

Agile - We included a Trello screenshot in the document shared. https://trello.com/b/s04B5dOQ/msda228bigdata

Version Control - We used GitHub and copilot. The screenshot is included in the document shared.

Impact - Leveraged robust big data infrastructure to efficiently process diverse data streams with the help of many tools to handle vast amounts of data.

Github https://github.com/anjali-ojha/customer_segmentation