Customer Segmentation And Analysis using Yelp Review Dataset

DATA 228 - Project Presentation

Team Members
[illegible]

Project Overview

- The Yelp restaurant reviews dataset, containing millions of [illegible] reviews, ratings, restaurant profiles, customer profiles and social network information, is a significant resource for extracting customer insights.
- The research aims to leverage big data analytics for customer segmentation using the extensive Yelp dataset.
- As data sizes grow identifying segments with specific attributes becomes a vital challenge in the Yelp Marketing applications.

Motivation

Data-Driven Decisions: Recognize the importance of data-driven decisions in business.

Rich Tapestry of Data: [illegible]

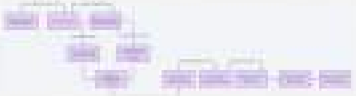
Effective Customer Segmentation: [illegible]

Elevate Customer Experience: [illegible]

Competitive Advantage: [illegible]

Methodology

- Businesses need to reach out to customers, and a good segmentation can help them to reach out to the right audience.
- Data can help to solve this problem.
- We are using the Yelp data sets and derive multiple attributes to create a right segment.



Data Cubes

- **Multidimensional Structure:** Data cubes organize data in a multidimensional structure for efficient analysis.
- **Dimensions and Measures:** Represented by dimensions (attributes) and measures (numerical values).
- **Cuboids for Subsetting:** Subsets of a data cube, called cuboids, allow focused analysis on specific dimensions.
- **Aggregation:** Summarization of measures across dimensions to provide higher-level insights.

Data Cubes [contd.]

Slicing, Dicing, and Pivoting: Operations performed on a data cube to extract specific subsets of data for analysis.

Slicing & Dicing
OLAP data operation

Literature Survey

[illegible]
- [illegible]

Reviews Credibility Influence (Dwarak et al.)
- Highlights the impact of reviews credibility on [illegible]
- Big data analytics [illegible] insights for [illegible] decision-making.

[illegible]
- [illegible]

[illegible]
- [illegible]

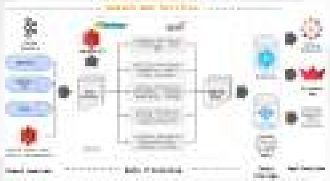
Conclusion

- **Enhanced Data Utilization:** [illegible]
- **Collaborative Deployment:** [illegible]
- **Robust Infrastructure:** [illegible]
- **Overcoming Challenges:** [illegible]

Thank You

Q & A

System Architecture



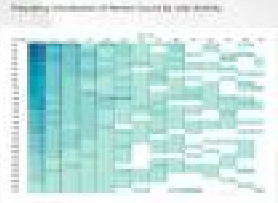
Understand Yelp Data

Dimension -

- Geographical Area (City & State)
- Cuisine: Segments where user shop
- Reason: Reviews on App
- Social Circle: Friends on Platform
- Food Preferences , Photos
- Sentiment

Metric -

- Starts given to a place
- Review about the place
- Tips about the place
- Number of Visits
- Time of visit
- Distance from the place



Attributes for Users

[illegible]

- **Behavioral Attributes:** [illegible]
- **Food Attributes:** [illegible]
- **Predictions:** [illegible]

Key Learning [Contd.]

- **Data Representation Challenges:** [illegible]
- **Resource Constraints Awareness:** [illegible]
- **Integration and Dependency Management:** [illegible]

Key Learning

- **[illegible] Sampling:** [illegible]
- **Handling Unstructured Data:** [illegible]
- **Cloud Setup and Utilization:** [illegible]
- **Iterative Approach:** [illegible]

Just one of the Jobs ...

Technical Difficulties [Contd.]

Data Representation:
- [illegible]

Resource Constraints:
- [illegible]

Integration and Cloud Setup:
- [illegible]

Technical Difficulties

Scale of the Data:
- [illegible]

Data Sparsity:
- [illegible]

Unstructured Data:
- [illegible]

Customer Segmentation And Analysis using Yelp Review Dataset

DATA 228- Project Presention

Team Members
Akanksha Tyagi (016738839)
Anjali Himanshu Ojha (016803033)
Keerthana Raskatla (016780855)
Sakshi Manish Mukkirwar (016794765)
Swati (016702413)

Project Overview

- The Yelp restaurant reviews dataset, containing millions of diner reviews, ratings, restaurant profiles, customer profiles, and social network information, is a significant resource for extracting customer insights.

- The research aims to leverage Big Data Analytics for consumer segmentation using the extensive Yelp dataset.

- As data sizes grow, identifying segments with specific attributes becomes a core challenge in Big Data Marketing applications.

Project Overview

- The Yelp restaurant reviews dataset, containing millions of diner reviews, ratings, restaurant profiles, customer profiles, and social network information, is a significant resource for extracting customer insights.

- The research aims to leverage Big Data Analytics for consumer segmentation using the extensive Yelp dataset.

- As data sizes grow, identifying segments with specific attributes becomes a core challenge in Big Data Marketing applications.

Motivation

Data-Driven Decisions: Recognize the importance of data-driven decisions in business.

Rich Tapestry of Data: Abundance of business, user, and customer profile data.

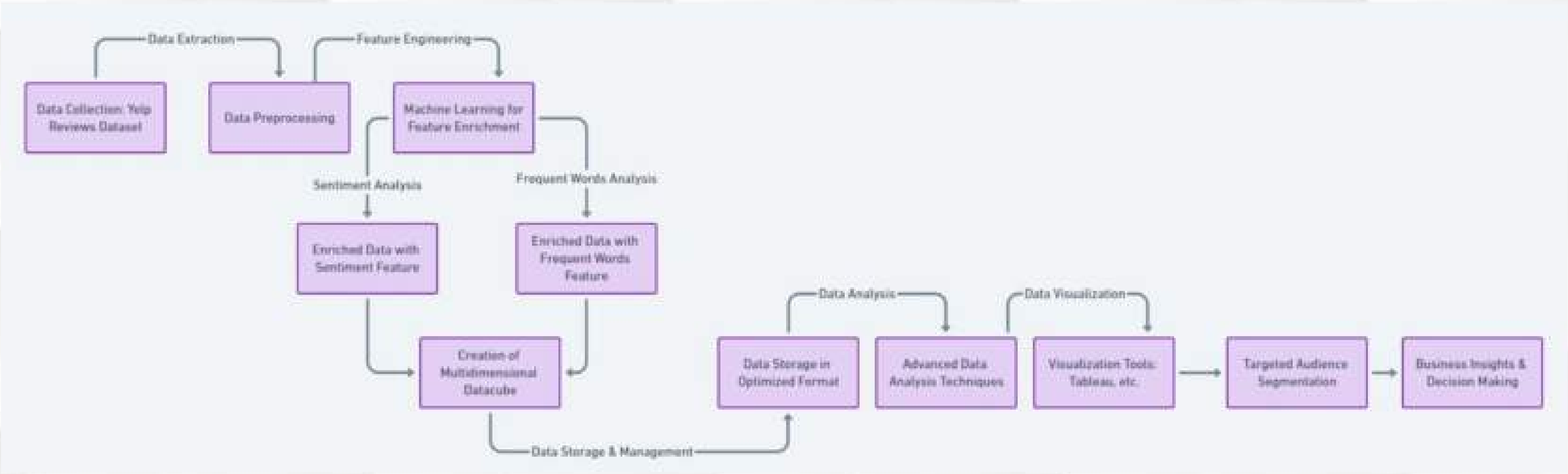
Effective Customer Segmentation: Drive segmentation and analysis for business insights.

Elevate Customer Experiences: Enhance customer experiences through data.

Competitive Advantage: Attain a competitive edge by deciphering customer behaviors and preferences.

Methodology

- Businesses need to reach out to customers, and a good segmentation can help them to reach out to the right audience.
- Data can help to solve this problem.
- We are using the Yelp data sets and derive multiple attributes, to create a right segment.



Data Collection: Yelp Reviews Dataset

Data Extraction

Data Preprocessing

Feature Engineering

Machine Learning for Feature Enrichment

Sentiment Analysis

Frequent Words Analysis

Enriched Data with Sentiment Feature

Enriched Data with Frequent Words Feature

Creation of Multidimensional Datacube

Data Storage & Management

Data Storage in Optimized Format

Data Analysis

Advanced Data Analysis Techniques

Data Visualization

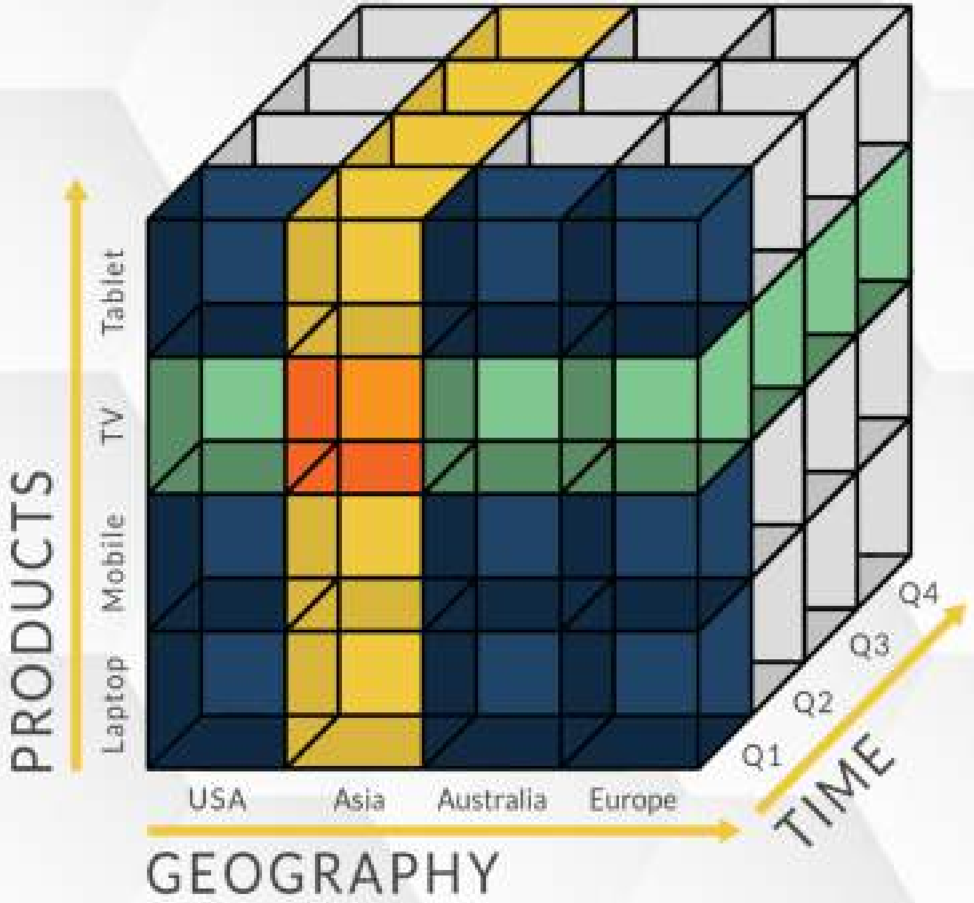
Visualization Tools: Tableau, etc.

Targeted Audience Segmentation

Business Insights & Decision Making

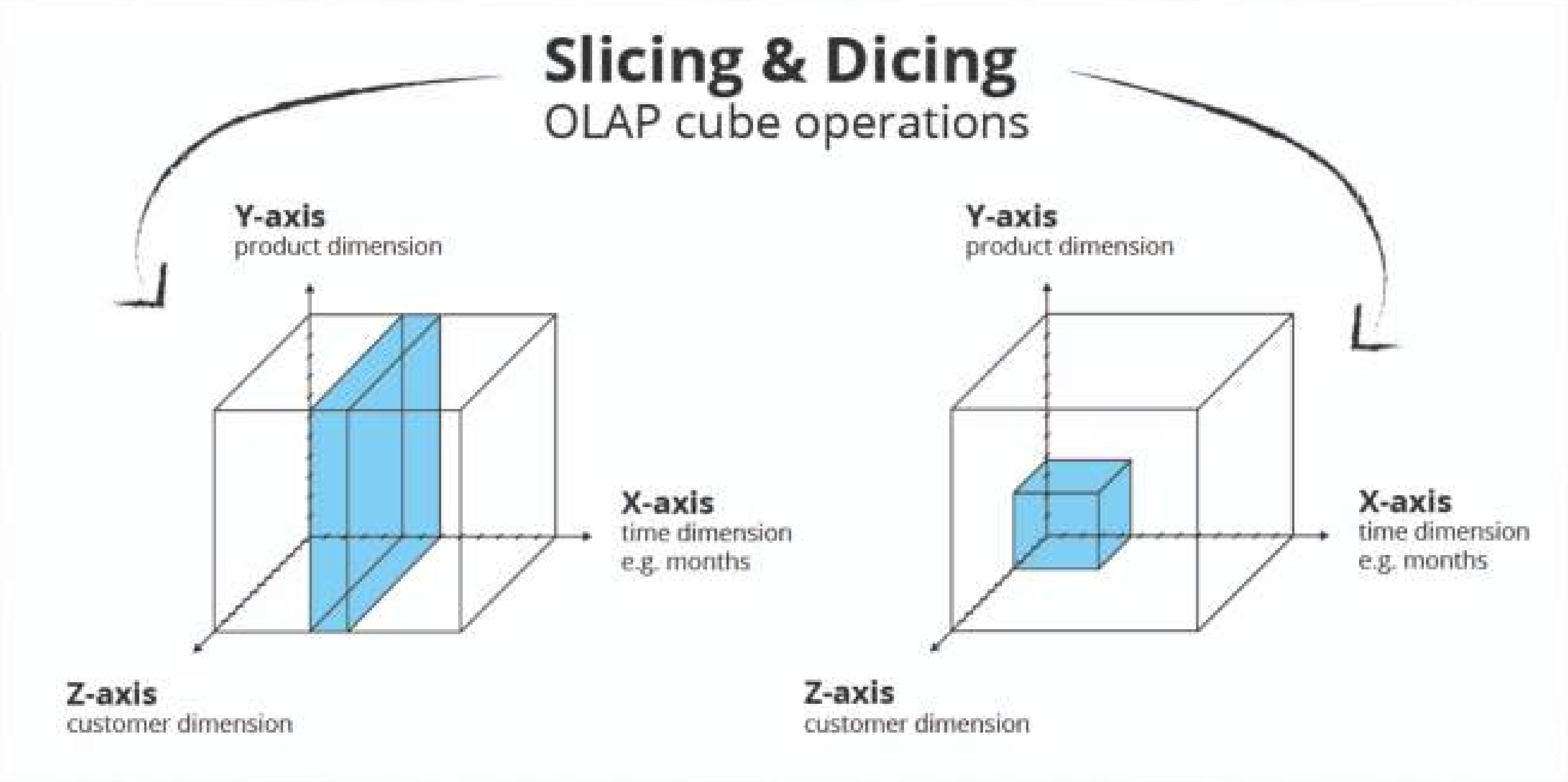
Data Cubes

- **Multidimensional Structure**: Data cubes organize data in a multidimensional structure for efficient analysis.
- **Dimensions and Measures**: Represented by dimensions (attributes) and measures (numerical values).
- **Cuboids for Subsetting**: Subsets of a data cube, called cuboids, allow focused analysis on specific dimensions.
- **Aggregation**: Summarization of measures across dimensions to provide higher-level insights.

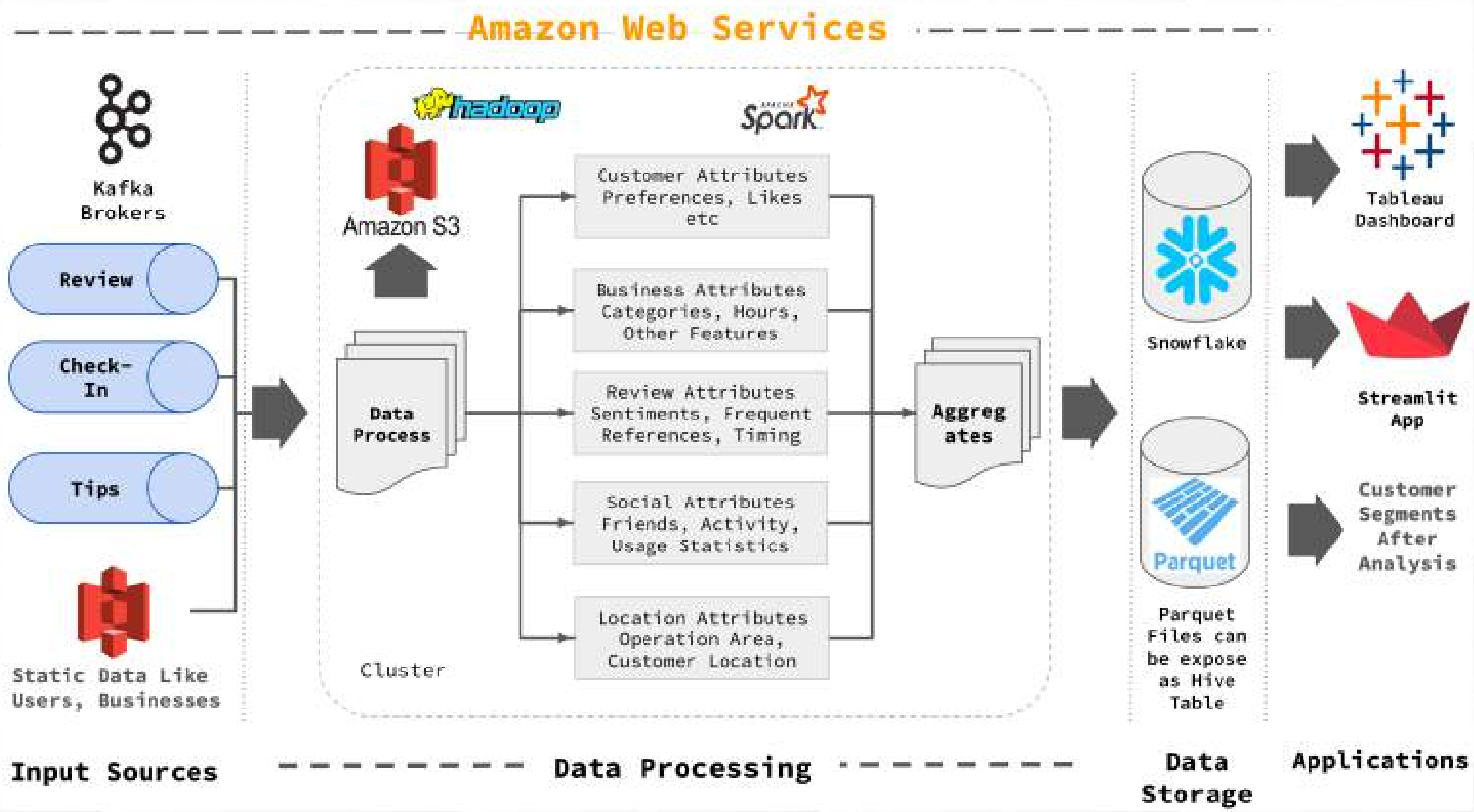


Data Cubes [contd.]

Slicing, Dicing, and Pivoting: Operations performed on a data cube to extract specific subsets of data for analysis.



System Architecture


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
Aggreg ates
Cluster
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

Understand Yelp Data

Dimension -

- Geographical Area (live & travel)
- Market Segments where user shop
- Feature Matters to User
- Social Circle Friends on Platform
- Food Preferences , Photos
- Sentiment

Metric -

- Starts Given to a place
- Review about the place
- Tips about the place
- Number of Visits
- Time of Visit
- Distance from the place

Attributes for Users

Using the dimensiona and metric, we can build our own data cube for the Yelp Dataset. We can devide these features in 3 categories -

- **Behaviorual Attributes**: Attributes which tells us about the user's behavior like and dislikes. Yelp have all the data and it benchmark the user against the world to decode the patterns.
-
- **Fixed Attributes**: Information about the user which are fairly constant like where they live.
-
- **Predictions**: Given all the data, use Machine Learning techiques to understand the pattern. Behavior like propensity or attrition can be easily captured.

Technical Difficulties

Scale of the Data:

- Strategic sampling is necessary due to the large scale (10+ GB) of the data.
- Processing the entire dataset locally is challenging and resource-intensive.

Data Sparsity:

- Collected data is generalized across various market segments, leading to sparsity.
- Incomplete samples require focused analysis of smaller, filtered segments for pattern identification.

Unstructured Data:

- The use of unstructured JSON format introduces complexities in analytical accessibility.
- Custom transformations are needed, especially for handling categories and business attributes with varying structures.

Technical Difficulties [Contd.]

Data Representation:

- Derived attributes like sentiment, social group size, and frequent words add complexity.
- Data cube creation by merging attributes into one table results in higher-dimensional data, impacting data reads for visualization and analysis.

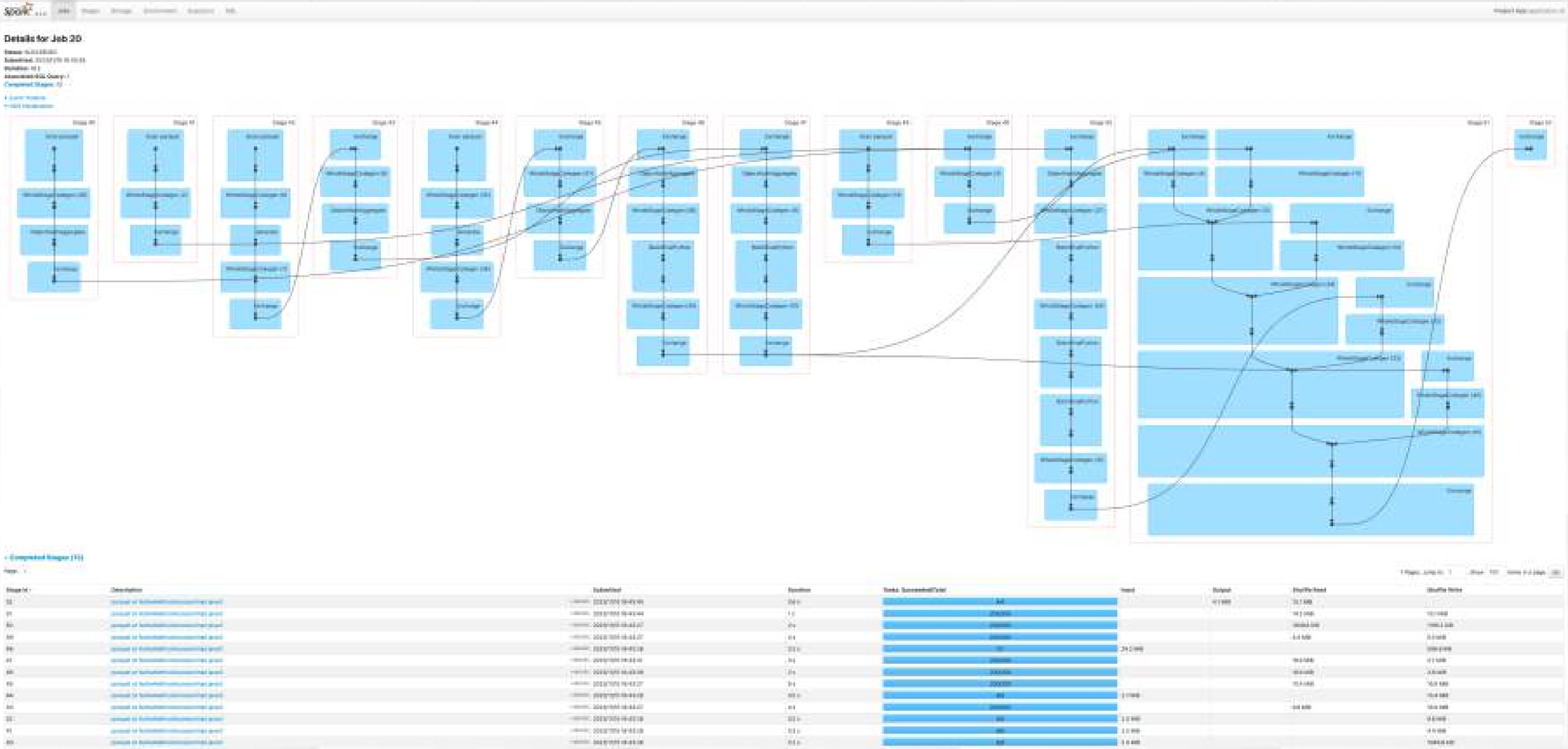
Resource Constraints:

- Size and skewness of Yelp data pose processing challenges.
- Java Out Of Memory exceptions occur during data processing, especially with flattened data across multiple categories.

Integration and Cloud Setup:

- Integrating tools like Kafka, Snowflake, AWS S3, Spark, Python, NLTK, streamlit, and Hadoop presents challenges.
- Finding compatible versions for each dependency requires searching and trial-and-error efforts.

Just one of the Jobs ...



Details for Job 20

Key Learning

- **A Good Sampling** strategy saves a lot of time. FInd a right sample takes time but its paid off while working with the large dataset.
- **Handling Unstructured Data** The data set we use was actual data from yelp and it kind of gives us how a real world data looks like. Its not always a well clean CSV file. Its raw sparse and unstructured.
- **Cloud Setup and Utilization:** Navigating challenges associated with setting up and utilizing cloud services. Balancing the benefits and complexities of utilizing cloud platforms for big data processing.
- **Iterative Approach:**Embracing an iterative approach for project refinement based on continuous learning. Adjusting strategies and techniques based on feedback and evolving project requirements.

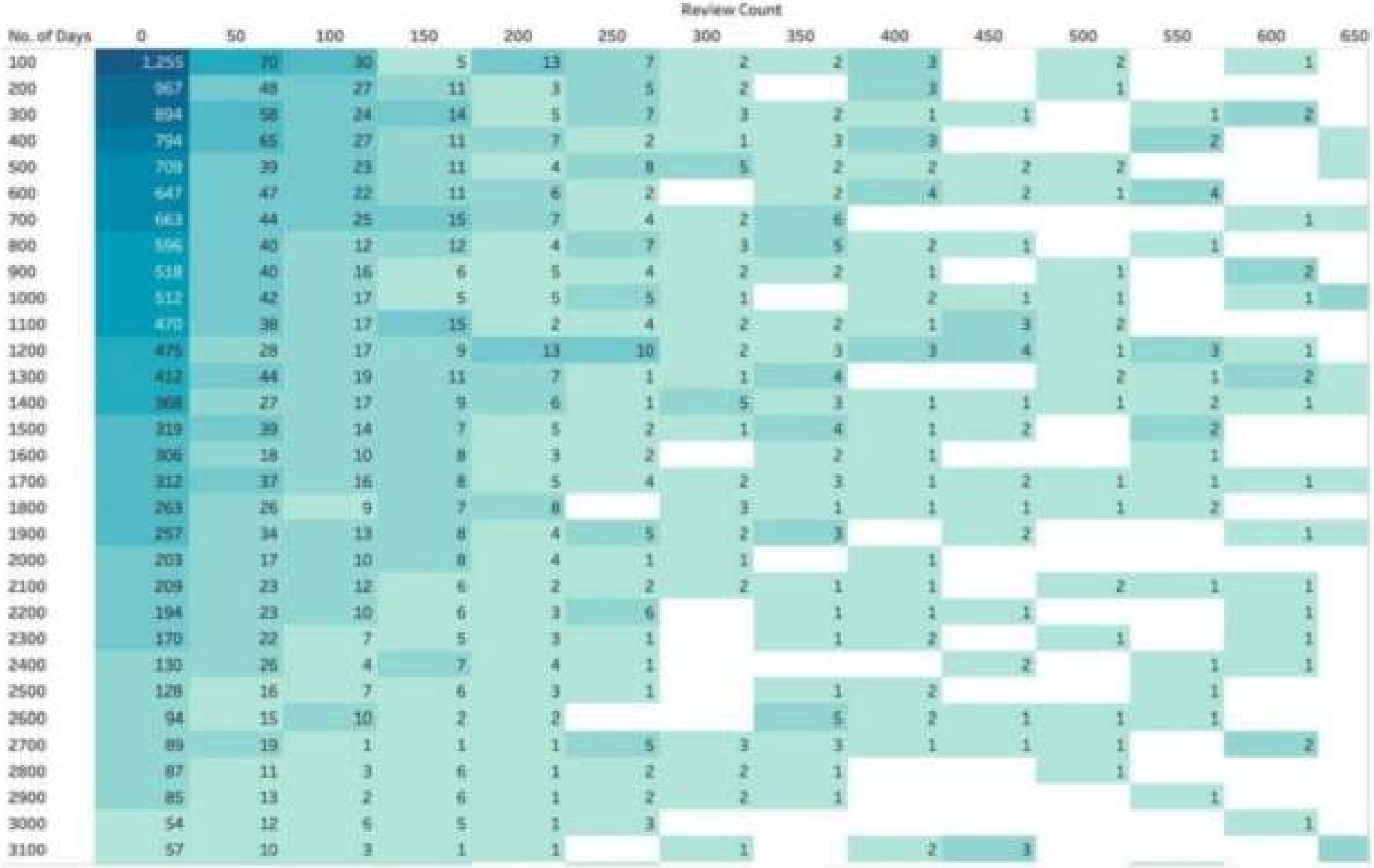
Key Learning [Contd.]

- **Data Representation Challenges:**Managing complexity when adding derived attributes to the data. Balancing the benefits and drawbacks of creating a data cube with higher-dimensional data.

- **Resource Constraints Awareness:**Acknowledging resource constraints, especially when dealing with large and skewed datasets. Addressing memory issues, such as Java Out Of Memory exceptions, during data processing.

- **Integration and Dependency Management:**Recognizing challenges in integrating multiple tools and dependencies within the project. Understanding the importance of finding compatible versions for various dependencies.

Frequency Distribution of Review Count By User Activity



	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				

Literature Survey

Effective Recommender Systems (Luo et al.):

- Focus on combating information overload on review websites.
- Emphasis on "restaurant value" and "food & drinks" influencing sentiment.
- Advocacy for dynamic recommender systems enhancing user experience.

Reviewer Credibility Influence (Kwon et al.):

- Highlights the impact of reviewer credibility on reader perception.
- Big data analytics potential in extracting insights for consumer decision-making.

Factors Affecting Reader Engagement (Meek et al.):

- Significance of framing, argument quality, and moderate ratings in reader engagement.
- Illuminates the role of heuristics in amplifying the impact of evaluations.

Pioneering Market Segmentation (Moon et al.):

- Innovative market segmentation using online consumer reviews.
- Profiling both customers and businesses for focused segmentation strategies.
- Capitalizing on publicly available consumption details within online reviews.

Conclusion

- **Enhanced Data Interation :** Improved entterises data understanding and interation,
- **Collaborative Deployment:** Maintained High-quality code through pair programming, efficiently integrating technologies.
- **Robust Infrastructure :** Establishing a scalable and reliable infrastructure using Kafka, Spark, Hadoop, AWS and snowflake.
- **Overcoming Challenges:** Successfully tackled data consistency and integration issues through adaptive problem solving.

Thank You

Q & A