-eps-converted-to.pdf

The Yelp Effect: From Reviews to Results

Datawarehouse Project

Shristi Kumar, Sujan Govindaraju, Kartheek Alluri, Mahak Gupta

***Abstract*— Yelp is a widely used online platform where users can rate and review local businesses, forming a community-driven directory of establishments in various industries. The platform collects large volumes of user-generated data, including reviews, ratings, and check-ins, which offer valuable insights into customer experiences.**

This project aims to leverage Yelp's business data sets to extract meaningful insights that can help businesses improve customer satisfaction and engagement. By applying ETL (Extract, Transform, Load) techniques and NoSQL data modeling, we will process and analyze the unstructured and semi-structured data to uncover patterns and trends.

The central problem we address is: *'How can businesses in different industries improve customer satisfaction and engagement based on reviews and check-ins?'*

Our objectives include identifying key factors that influence positive customer experiences, detecting common pain points, and recommending actionable strategies for businesses to enhance their service offerings. The insights derived from this analysis can support businesses in making data-driven decisions to improve their operations and customer relationships.

I. Motivation

IN the world of today, the power of 'reviews' is profound. If you pause and think, you would know that the majority of our decisions are review driven, directly or indirectly. For example, to go to a new restaurant, we check for reviews, before booking a hotel or vacation destination, again we check for reviews. In fact, to select a college or even a particular course, we go and look for reviews. There is some truthfulness and accountability in reviews as they come from a community, and hence they are invaluable. The immense importance of reviews in today's digital world excited us with the idea to analyze it for further business solutions. With the knowledge gained from our coursework on ETL pipelines and NoSQL databases, we aim to take advantage of this power of reviews to offer actionable business solutions.

II. Literature for Survey

Customer satisfaction is a key factor in the success of a business, regardless of the industry to which it belongs. Evaluating customer satisfaction and using insights for business growth is a universal goal. Yelp provides a platform where businesses can receive quantifiable feedback through user-generated reviews and check-ins, offering a valuable opportunity to assess and improve customer satisfaction levels.

According to the IEEE article titled *"Using Online Reviews for Customer Sentiment Analysis"* [1], customers who are highly satisfied or dissatisfied are more likely to leave reviews which makes these reviews especially useful for sentiment analysis. Motivated by this insight, our project focuses on analyzing these extreme sentiments to generate targeted recommendations that can inform business strategies across various categories.

A related IEEE publication, *"Sentiment Analysis of Yelp Reviews by Machine Learning"* [2], employs machine learning techniques to classify customer sentiments as positive or negative. While that study leverages supervised learning approaches, our project adopts a different methodology. We utilize ETL (Extract, Transform, Load) techniques and dimensional data warehouse modeling strategies to organize and analyze Yelp's semi-structured data. We are using 'Alteryx' (drag-and-drop) ETL pipeline and ustilised text processing techniques to refine reviews. The data modeling is inspired by Ralph Kimball's seminal work, *"The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling"* [4], which guides our conversion of the Yelp dataset into fact and dimension tables using star and snowflake schemas.

For data storage, we selected NoSQL database—MongoDB as they are well-suited to handle the unstructured nature of Yelp's data. Our choice was informed by the IEEE article *"Cassandra vs MongoDB: A Systematic Review"* [3], which presents a comparative analysis of the two databases. This review helped us identify the strengths and limitations of each system, enabling us to optimize our data storage and processing approach based on our specific use case.

We would like to extend our sincere gratitude to Dr. Vishnu Pendyala for his invaluable guidance throughout the course. We also thank the teaching assistants, Kanchan Ashok Naik and Mayank Kapadia, for their constant support and assistance during the coursework, which greatly contributed to the successful completion of this project

References

[1] S. Rathore and A. Maheshwari, "Using Online Reviews for Customer Sentiment Analysis," *IEEE Xplore*, 2021. Available: https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9512387

[2] R. Rajendran and S. A. Patel, "Sentiment Analysis of Yelp Reviews by Machine Learning," *IEEE Xplore*, 2020. Available: https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9065812

[3] A. Qayyum et al., "Cassandra vs MongoDB: A Systematic Review of Two NoSQL Data Stores in Their Industry Uses," *IEEE Xplore*, 2023. Available:https://www.researchgate.net/publication/384550935_Cassandra_vs_MongoDB_A_Systematic_Review_of_Two_NoSQL_Data_Stores_in_Their_Industry_Uses

[4] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 2nd ed., Wiley, 2002.

III. Methodology

The following steps outline the methods we will use to build our project.

label=.

1) **Data Collection and Exploration**
   The data set is collected from the Yelp Open Dataset, which is available under an educational license. The data set is provided in the form of five large JSON files. For this project and according to our multidisciplinary problem, we have used the following three files: `yelp_academic_dataset_checkin.json`, `yelp_academic_dataset_business.json`, and `yelp_academic_dataset_review.json`. The data set can be accessed at: `https://business.yelp.com/data/resources/open-dataset/`
   (Tool used: **Python**)

2) **Data Cleaning and Preparation (ETL techniques)**

   - **Data Cleaning** Per inspection, that the datasets were cleaned and appropriate for usage. The important fields, like text review and ids were already present in the datasets. However, we noted that the 'Business dataset' had ids, wherein there were no 'category field'. Those rows were dropped as a result.
     (Tool used: **Python**)

   - **Feature Engineering** Our multidisciplinary problem was associated with industry insights. Upon reviewing the business dataset, we observed that each business was associated with multiple categories. For example, the business "Target" was linked to categories such as *Shopping*, *Store*, and others. Using such keywords in the `categories` field, we mapped each business to the 22 distinct and self - defined industry sectors. For eg, Target was mapped to `Retail and Shopping`
     (Tool used: **Python**)

   - **Data Preparation** In a purview of dimension modeling, we created a new dataset called 'Industries' having *id, industry and categories* as it's column. The check-ins file was transformed into more metric oriented dataset which had columns like - *number of check-ins, last and first check-in date and check-in per number of days.*. Please note, a sample of 500,000 dataset was used from the 'review dataset' for this project.
     (Tool used: **Python**)
     The datasets at the end of this transformation:
     - **Review Dataset**: Yelp Reviews (500K sample)
     - **Business Dataset**: Businesses mapped with Industries
     - **Checkin Dataset**: Transformed for numerical insights
     - **Industry Dataset**: Industry mapped with Categories

3) **Text Mining**

   - **Data Merging** We wanted to club the Business dataset to the Review dataset. This mapped the reviews to each specific Industry. The same dataset was then converted into CSV file for further processing.
     (Tool used: **Altryx - Cloud**)

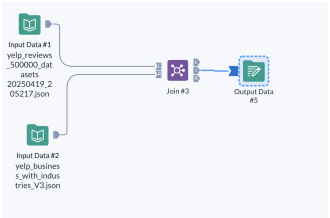


Fig. 1. Alteryx workflow

   - **Text-pre processing and insights**
     We chunked our data via python to 100K to be able to process the big file. We performed several operations on review text for the specific industries. To go through let's take the example of one such pipeline for **'Restaurant and Food' Industry**

Fig. 2. Text mining pipeline for 'Restaurant and Food' Industry

     Let's go through each node to summarize the working:

     - **Node 1: CSV-File Import**: We imported the combined file coming from Altryx (Business + Industry) here.
     - **Node 2: Corpus**: It takes a collection of text data (here texts) and organizes it so you can easily work with it.
     - **Node 3: Selected Rows**: We filtered table with the industry 'Restaurant and Food'
     - **Node 4: Pre-process Text**: We applied filters like *Transformation: Lower casing, Remove accents; Tokenization: Regexp and Filtering: Stop words*
     - **Node 5: Sentiment Analysis**: Tool has pre-built doctionary - Vader. Same was used to calculate sentiment scores.
     - **Node 6: Data Table**: The table was used to visualize sentiment score calculated.

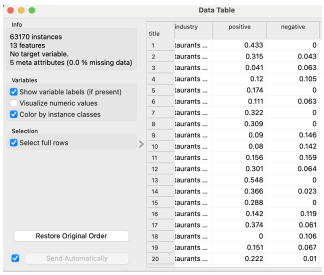


Fig. 3. Sentiment score calculated in Orange

– **Node 7: Formula**: We assigned a column Sentiment label mapping each review text to sentiment Positive, Negative, and Neutral'.
– **Node 8: Pivot table**: We calculated Sentiment label counts to visualise it better for each industry.

Sentiment_label	Count	Negative
	Negative	7294.0
	Neutral	0.0
	Positive	0.0
	Total	7294.0

Fig. 4. Pivot table used in Orange to visualize sentiment by industry

– **Node 9: Distribution**: Finally, we used this node to visualize the sentiment label.

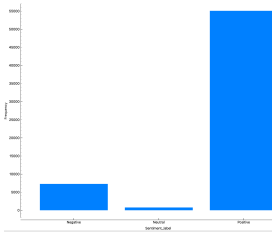


Fig. 5. Distribution graphs to visualize sentiment analysis

– **Node 10: Topic Modeling**: The associated words across 10 topics were generated from this node. The green highlights the most frequent used word and the red highlights the less frequent used word.

Fig. 6. Topic Modeling

– **Word Cloud**: We analyzed the most common words based on their frequency in review text.

Fig. 7. Word Cloud

Tool used: ***Orange Desktop***

4) **ETL Pipeline**

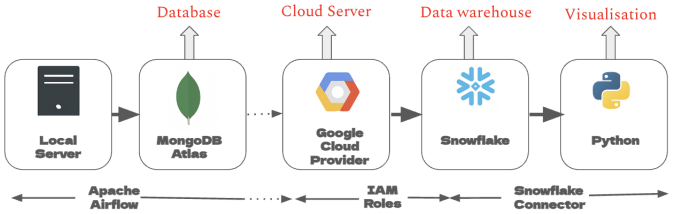


Fig. 8. ETL Pipeline

- **Local Server to MongoDB** This connection was achieved via Apache Airflow. DAGs were created on the local server. The following files were triggered and sent via DAG for further analysis from the local Apache UI setup. The following three files were sent via DAG to MongoDB Atlas:
`Yelp Business and Review combined dataset` *-¿ from Altryx* `Transformed Check in Dataset` *-¿ from Data Preparation* `Industry Dataset` *¿ from Data Preparation*

  However, there is a key thing to note. We had already performed Sentiment Analysis, because of the huge size of dataset, MongoDB was unable to load the datasets in its free tier, which has maximum capacity of only 512 MB.

  To solve this problem we used Apache Airflow's text blob library and ran another DAG to calculate sentiment score for each review text. Once the score was calculated, we removed the review text from the base file. Let us call this file now `Yelp review dataset with sentiment`

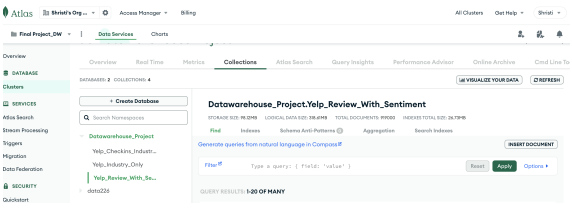


Fig. 9. ETL Phase 1: Local Server Mongo DB

- **MongoDB to GCP** This connection was supposedly assumed to be achieved by similar fashion as stage 1. However, we hit a hard rock here. We tried everything, chunk processing, log processing but there was a wall that did not allow MongoDB files to be written in the GCP Bucket.The DAG kept running for 3-4 hours until it failed. Even the lightest file, Industry Dataset which has only 22 rows and 3 column, could not be written to GCP bucket. As a result files were manually loaded into GCP for further processing.

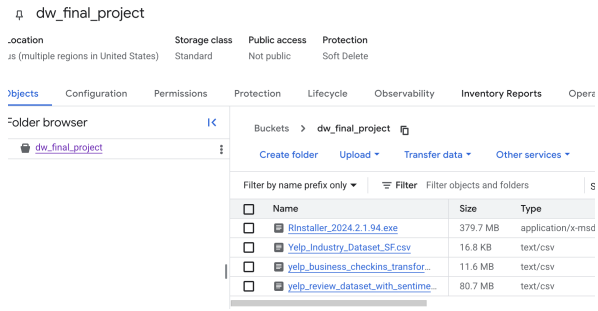


Fig. 10. ETL Phase 2: MongoDB to GCP

- **GCP to Snowflake**: This was fairly easy. Snowflake works around cloud first approach and hence the connection was actually smooth. We had to create a role giving desired access to the bucket. Then we generate an API cloud key for the bucket. In snowflake we ran some commands to built a **a staging area** so that we can copy the files from GCP to Snowflake. Also, we created a schema for the three datasets so that they can be copied directly.

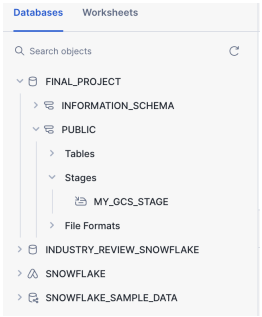


Fig. 11. ETL Phase 3: Snowflake Staging area

- **Snowflake to Python**: The Free-tier of Snowflake does not have an vizulisation tool. The paid tool of Snowflake has in-built tool : Snowsight Interface tool, which is capable of making dashboards and charts but we didn't had the access of it. As a result, we connected python with Snowflake using snowflake-library and made use of Matplotlibrary and Seaborn to gain the insights.

  We will learn more about it in the **Vizualisation** section.

5) **Dimensional Modeling**
Now we have copied data from GCP, we now have three basic files. The goal is to generate a Star schema from the available data. To refresh, we have loaded the following datasets in the Snowflake: `Combined review and business dataset`, `Transformed Checkin Dataset`, and `Industry Dataset`. Now, to achieve dimension modeling, the following were the key considerations:

- **STAR over SNOWFLAKE** – We preferred start over snowflake. There were key reasons for doing that. The foremost being, we wanted the data modeling concepts to be simple and avoid joins.

- **Availability of less Dimension attributes** – We have very less dimension attributes (only available in a business dataset). Hence, modeling should be done accordingly.

- **Fact Constellation Requirement** - Please note that our motive is to get data around industry. Hence, it is necessary to have an industry fact table. However, in case of need, there should always be the review fact table as Yelp dataset is all about review.

- **Creation of new dimensions** - We all know that a dimension table is incomplete without the date dimension. Hence, the date dimension needed to be created. Also, we have the sentiment score, but it is of no use if we cannot attribute it to the sentiment label. Thus, as a result, a new dimension related to the sentiment label was created.

- **Excluding user data** - We have not added the user database and hence there is no need of keeping user id or user related data for now. Thus, user data can be removed.

- **Denormalise wherever is possibility** - The check-ins data is related to the business. Hence, it makes sense to combine the two datasets into one. Please note that the star schema promotes denormalization, unlike the star schema.

Based on the criteria discovered above, our dimension model consisted of 4 dimension tables and 2 fact tables in the fact constellation schema. Dimension tables `Date`, `Industry`, `Business and Review Sentiment` while Fact tables were `Industry Fact and Review Fact`

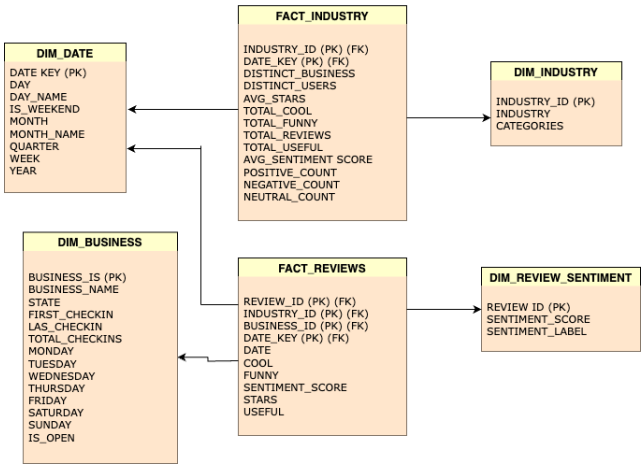


Fig. 12. ER Diagram - Dimensional Modeling

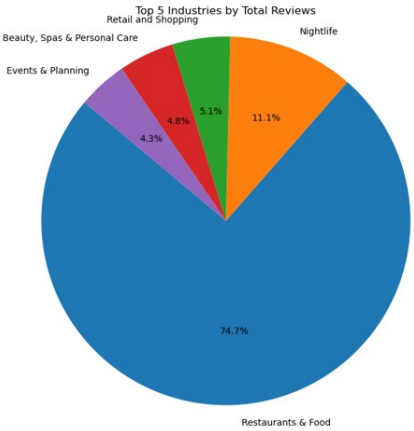


Fig. 14. Top 5 Industries with total reviews

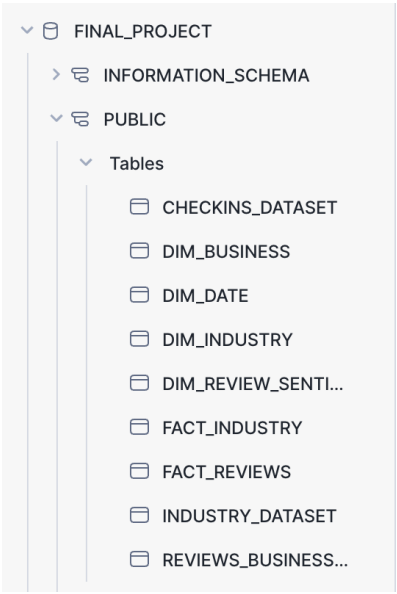


Fig. 13. Fact and Dimension tables created in Snowflake

6) **Data Visualization**

We have connected Python over Snowflake. As we discussed earlier, though Snowflake has in-built tool, but the same could not be used in the free-tier. As a result, we have run snowflake in python and got insights, as is mentioned below:

- **Total Reviews** - What we are not surprised is that Restaurant and Food is on the scale in which reviews are received by this industry. Having 75% capture of reviews is a big thing. It is followed by nighlife industry, but that is like 11%. Surprisingly Retail, Beauty and Spas shares almost same percentage with the Events and Planning. This suggests that the customer/service facing industry is more prone to receiving reviews.

- **Daily Checkins** - What we understood was that the Restaurant Food will have the maximum check-ins considering the number of reviews. However, looking at the numbers, we were surprised that the check-ins were actually more in mid-days than in weekends. We often believe that weekends are the busiest. Yelp also has online business listed and it feels that mid-week food ordering is at the peak.

Fig. 15. Daily Check-ins by Industry

- **Total Reviews** - We calculated the average sentiment score for each industry. Per analysis, we could find out that the no industry lies in a red zone, or so to say, where negative comments are higher than the positive comments. That is commendable, people tend to put their healthy feedback more often. The three industries that were a little far behind were Professional Services and Real Estate Housing. That is the place where key emphasize needs to be played on analyzing what went wrong.

Fig. 16. Avg. Sentiment Score Analysis

- **Analyzing Restaurant Industry** - We now know, that the industry that has captured most number of review market, if that was a term is none other than Restaurant and Food. So, to dig deep we wanted learn about the Restaurant / Industry based on its average sentiment label distribution over the years. .

  a) **Overall Sentiment share** - We noted that the overall positive sentiment in the review text is *84.2* in these many years while *negative* is only *8.0* and neutral is *7.8*. It is a good thing right to have this amazing overall metric in sentiment share.

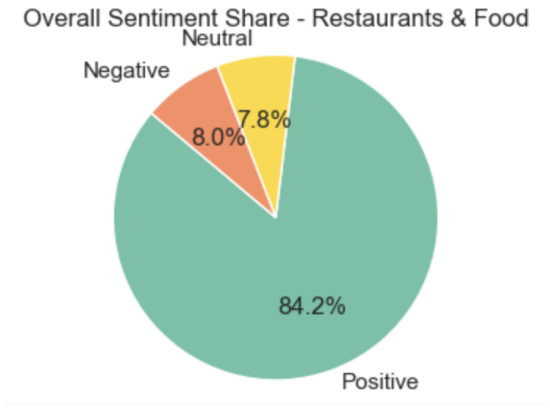


Fig. 17. Overall Sentiment share

  b) **Average Star Rating** The average star rating peaked at 2020 but declined drastically in 2022. There are two reasons for it, one is COVID, people were scared and advised against of eating outside. Hence there were very few reviews or ratings provided for the year 2022.

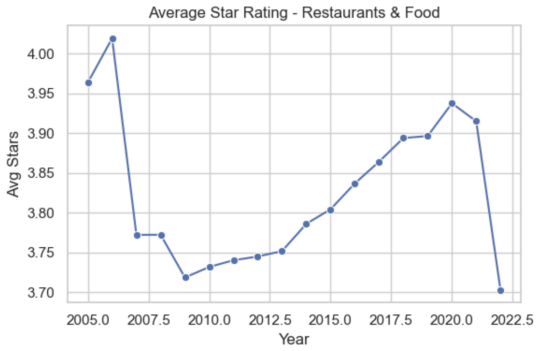


Fig. 18. Average Star Rating

  c) **Negative Reviews by Year** - Looking at the size of the review market, now we know that Restaurant / Food industry is a highly competitive market. So to understand the pain points is of immense high importance. That is what we have done here. The Negative comments kept rising exponentially until it die down in 2020. This indicates high expectations from the customers considering the competitive market.

  The data is visualised both in bar and line chart for the better understanding.

Fig. 19. Negative Reviews by Year(line chart) - Restaurant / Food.png

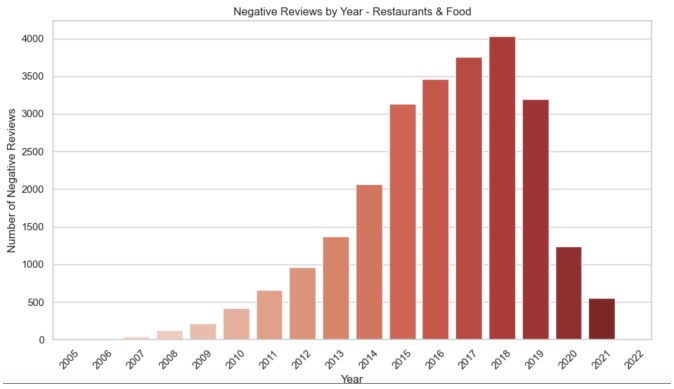


Fig. 20. Negative Reviews by Year(bar chart) - Restaurant / Food.png

IV. Key Learnings / Recommendations

A. Based on Analysis

a) **Based on reviews** - Every customer oriented job/industry is more prone to reviews. Hence, for a customer industry it is of vital importance that for them to make use of text mining tools to get sentiments analyzed. Here, we could see Restaurant and Food industry totally captured the review market

b) **Effect of Covid** - We noted that the Covid affected the overall industry in a major way. The reviews to checkins all fell short. As a result the key metrics like ratings also dropped drastically.

c) **Negative Reviews analysis** - We noted that the negative reviews share is really less in comparison. But this is a competitive world and hence, one should pay attention to negative comments. The word cloud and topic modeling can help.

d) **Others** - The other industries like professional service, travel and real estate are to watch out for. People are here very particular and what they want and anything that's not suitable to them can make them off.

B. Based on Technical Experience

a) **Logical Design is very Important** - One should understand the importance of drafting logical design at the very beginning. The iterative approach is not a best practice and would incur a lot of time and cost.

b) **ETL pipeline automation** - Automation is a best practice, to have an end to end pipeline been created and automated just saves a lot of time and energy. But there should be always a descent checks on every stage so to see everything is working fine at least at the time of building.

c) **New tool exploration** - It is always good practice to read about the tools before having to use it. For example, MaC does not supports Altryx desktop. Reading about the subject matter or new tool before hands on learning is a best practice.

d) **Time management** - In the project is very important to manage time and accomplish task, having scrum planning, discussion road blocks, thinking of alternatives way is of good practice.

e) **Use Google over Chat GpT** - The AI is here to help and speed up the overall process but atleast for the project we got a help from Google instead. Public platforms like Reddit, Stack overflow helped us overcome problem in few seconds that practically Chat GpT wasn't able to do so.

V. Technical Difficulty

f) **Logical Planning** : We had a few hiccups while planning the overall layout of the project. The project was supposed to have all the salient features those were mentioned in the rubrics. We also wanted it to be enough challenging for us to gain the necessary skills and learnings.

g) **Data pre-processing** : We constantly pre-processed our data, it was because we were new to the tools that we there were few hiccups while ultimately loading the data.

h) **Performance issue due to date size** : We constantly had to update codes to be able to process the data. The data was huge so we always performance issue and ETL pipelines took forever to load the data from one infra to another.

i) **Dimension Modeling** : Getting the right Dimension Model is the key to this fact paced world. The motive is to always understand and curate the model in accordance to that. What fits the best is not the solution, what fits most compatible with the use case is the main idea.

j) **Mongo DB and GCP** : MongoDB Atlas does not have a native "push to GCS" option. The Airflow automation crashed while uploading files to GCP. GCP might require further processing or IAM roles configuration. So far we can say that the connect is not very user friendly or easy.

k) **Orange Performance Issue** : Orange got stuck as we increased the dataset. It did not hang but it did could not process the huge amount of files. The UI is pretty user-friendy due its drag and drop feature, however having not able to performance is a key issue.

l) **Free Tier Limitations** : The cloud services offers a number rof features which ware both interesting and exciting. They all mostly offer a free account duration but that nothing but a click bait. The long term features are always paid features. No wonder, all big firms are trying to move to in-house no cloud tool.

VI. Significance to the Real World

As we mentioned in the Section: Motivation as well, reviews is like money in today's era. Everything revolves around review, whether it is online or it's word of mouth. For business's it is a key factor to explore so that they can do their SWOT analysis and thus increase productivity. Below we are noting some of the key features as to why this project is relevant to the real world:

a) **Business Intelligence and Benchmarking** : Companies draw comparisons with industry metrics and identify key areas of improvements and developments.

b) **Robust Recommendation System Creation** : The reviews can help in creating a robust recommendation system and ML models to perform better.

c) **Investment and Risk Analysis** : Investors can pay attention to which industry is growing and at what pace. This equips them to have make better judgements.
d) **Cultural Insights** : What works for East might not work for West, and reviews are the best way to tell that. Industry experts understand the trends and make key points in their global expansions.

VII. Innovation

a) **Figured Automation of ETL pipeline** : From storage in a database to visualization, we automate the whole process in Apache Airflow.
b) **Different tools for the Sentiment Analysis** : We all know the importance of sentiment analysis and every tool sort of gives the same result, but we used the tools to achieve the best, say Orange for topic modeling and word cloud right at the instance while Apache airflow to calculate the sentiment score.
c) **Yelp Dataset - Data modeling** : We chose a fact constellation and created a whole new data medal to analyze industry- specific trends.
d) **Feature Engineering** : We perform feature engineering based on the industry use case analysis keeping in mind.
e) **Innovatives way to handle big size** : We worked in chunks, did log analysis, exception handling and, in fact, transformed files to be able to process them quickly.

VIII. Github

We had made a repository on Github and committed our project findings to it. All the source code are uploaded in the Github. the Python codes were run and Git helped us to maintain the version control for the project practice. Please refer to the below screenshot:

Link : `https://github.com/Mahak202/DataWarehouseProject`

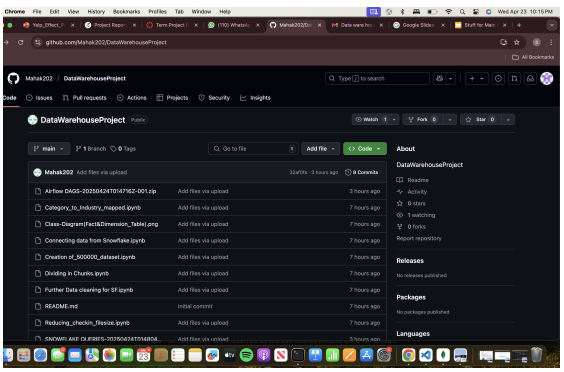


Fig. 21. Git repository

IX. Teamwork

The project was divided among the members of the group in the following way. Please note that , there were overlaps as well in terms of if a team member got stuck, the others came and helped to navigate through the problems. Each problem was discussed and answered by the weekly team meetings. The suggestions were taken from everyone before advancing in the project. Everybody was keen on bringing more to the table and help each other. Overall this project is attributed to each member of the group.

X. CRediT Taxonomy

A. Contributor Roles

- **Literature Survey** Kumar, Shristi
- **Project Proposal** Kumar, Shristi
- **Logical Planning** Gupta, Mahak
- **Methodology** Kumar, Shristi
- **Data Proceesing and Cleaning** Gupta, Mahak
- **Sentiment Analysis** Govindaraju, Sujan
- **ETL Pipeline:** Kumar, Shristi
- **Github Maintainance:** Gupta, Mahak
- **Scrum Planning:** Alluri, Kartheek
- **Dimension Modeling:** Kumar, Shristi
- **Visualisation:** Alluri, Kartheek
- **Validation:** Givindaraju, Sujan
- **Writing – original draft:** Kumar, Shristi
- **Writing – review & editing:** Govindaraju, Sujan

We are proud of our team to have pulled this project with so much sincerity.

XI. Practiced Pair Programming

As we indicated earlier in the report, our group held weekly meetings to determine progress and difficulty at each stage. The meetings were held at the library or at the zoom. If any one gets caught or has some thing to show basically, the ”driver” showed the code to the ”navigator,” who then reviewed the code and explains if anything went wrong or if we need to add up on anything.
We have added a zoom invite screenshot where the two members are going over the Snowflake schema creation codes for the reference here.

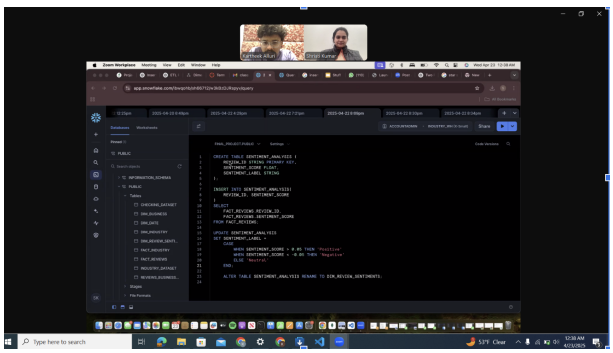


Fig. 22. Code Review in ZOOM

XII. Agile and Scrum Meetings

As we discussed that the meetings happened over in the zoom call and the library. As a we decided on a sprint of 15 days - close to the project submission. We already had the layout as was submitted in the project proposal earlier, However, for the project to be wrapped up in well within the time frame. We set up a sprint of 3 days. `Kartheek` acted as the Scrum Master, and the others were team members. We kept Retrospective call after every two Sprint, to decide on what we can do better. Here is the attached screenshot of our Kanban board:

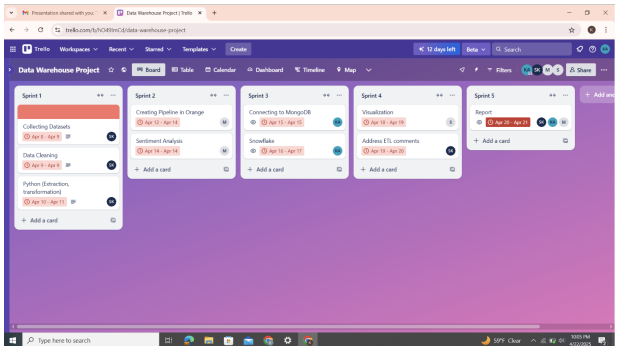
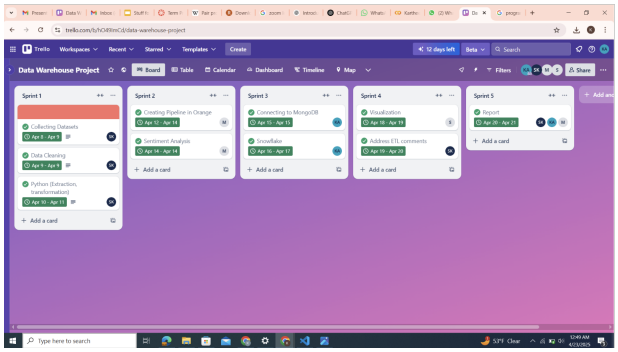


Fig. 23. Kanban board setup



XIII. Analytics component in the project

The project consisted several analytics component we have summarized the same in the below bullet points:

a) **Feature Engineering** : A complete new data set was created within the use case - *Industry dataset*.
b) **Sentiment Analysis** : We performed sentiment analysis both through *Orange* and *Apache Workflow*. The sentiment score was calculated, and the sentiment label was tagged to each review text.
c) **Topic Modeling** : The topic involving all industries was extracted to be useful in gaining insight from industry leaders.
d) **Dimension modeling** : We created a Fact Constellation based on the use case and available datasets.
e) **Infra set up** : ETL pipeline was created where the flow of data and insights was visualized.

XIV. New tool covered in the project

We covered two new tools for this project. Please , refer below for the Salient features noted for the tools.

a) **Altryx** : The tool was user-friendly. It had several suits, Designer , and Intelligence suite that had various potential such as text processing, machine learning, and others. It also , had connection to several clouds and supported various file types. However, the major drawback for Altryx was that the desktop version only supported Windows. This is a huge setback in today's world. The cloud version was accessed and explored, but did not have many free tools (nodes) for us to perform actions.
b) **Orange** : As a result, we moved to 'Orange' to perform drag -and-drop sentiment analysis and get the text mining results with visualizations. The interface is very user-friendly and so is the setup. The key drawback was performance issue as Orange was unable to process heavy datasets and therefore was not appropriate when using big datasets. The nodes are appropriate and in sync with the real world. Another drawback was the cloud connection that the Orange business team can explore and enhance.

XV. Future Work - Project Extention

a) We can extend our project to involve users as well in the dataset and get user specific key insights on the industry as well.
b) We can implement NoSE to automate schema design for the project
c) We can automate MongoDB to GCP portion to automate the ETL pipeline entirely.

Appendix and the Use of Supplemental Files

This report's data, code, and additional materials are available on GitHub at `https://github.com/Mahak202/DataWarehouseProject`. Below is an outline of key components:"
/data/: A Google drive link is added as the data size was too huge. pasting the link here as well `https://drive.google.com/drive/folders/1AGB1XQ3UW9r0diXq9BvX3L_DBLu0a3iG`
/notebooks/: Data pre processing, data exploration, data transformation codes are added /scripts/: ETL pipelines, including Airflow DAGs and MongoDB-GCP automation /Snowflake queries/: The queries history file is added as a zip from the /visualizations/: Final charts and dashboard prototypes are added in the report /docs/: Dimensional modeling diagrams and architecture notes Appendix C: Custom Functions "Refer to scripts/preprocessing.py for the full text cleaning function used prior to sentiment analysis."

APPENDIX I
REFERENCES

- *Literature Paper 1*
  A. Qayyum et al., "Cassandra vs MongoDB: A Systematic Review of Two NoSQL Data Stores in Their Industry Uses," *IEEE Xplore*, 2023.
- *Book*
  R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 2nd ed., Wiley, 2002.
- *Literature paper 3*
  R. Rajendran and S. A. Patel, "Sentiment Analysis of Yelp Reviews by Machine Learning," *IEEE Xplore*, 2020.
- *Snowflake Documentation*
  `https://docs.snowflake.com/` Covering the detailed documentation