Big Data Analysis of Anime

Team 2, GMU ECE552 Big Data Technologies, Spring 2023

Sai Sridhar Moturu
CE, GMU
G01406102
smoturu2@gmu.edu

Chiranjivan Krishnakumar Nirmala
CE, GMU
G01357471
ckrishn@gmu.edu

Pavankumar Venkatesh
EE, GMU
G01388708
pvenkat3@gmu.edu

***Abstract*—This paper presents a study of big data analysis of anime using data from topAnime.org, a popular anime ranking and review website. The study focuses on the use of Spark libraries for Python to perform data processing and analysis on a large dataset of user reviews and ratings of anime. The dataset was obtained from topAnime.org and consists of over a million user reviews and ratings of anime series. The data was processed using Spark libraries for Python, including PySpark and Spark SQL. Our analysis revealed interesting insights into anime viewer preferences and trends, such as the most popular genres, the most highly rated anime series, and the relationship between user reviews and ratings. The results of this study demonstrate the usefulness of Spark libraries for Python in big data analysis of anime and provide valuable insights for anime producers and streaming services to better understand their audience and tailor their content to meet their needs. The techniques used in this study can also be applied to other forms of media and provide insights into user behavior and preferences.**

Keywords—PySpark, SQL, Python, Big data

I. Introduction

Anime[3] has become a popular form of entertainment globally in recent years, with millions of fans streaming and downloading anime shows on various platforms. With this expanding popularity has come an increased interest in studying the interests and habits of anime viewers in order to better personalize content and improve user experience. Big data analysis is a strong tool for learning about user behavior and preferences, and it may be used to extract significant information from the massive volumes of data created by anime viewing.

In the rapidly evolving landscape of entertainment, streaming services have become a major player in the way that audiences consume media. With the rise of online streaming, anime analysis has become a critical tool for understanding the ever-changing preferences of viewers and staying ahead of the competition. By using big data and analytic tools, anime streaming services can gain insights into what their viewers are watching, how they are consuming it, and what they want to see in the future. This information is essential for creating personalized recommendations and improving the user experience, which is crucial in a world where instant gratification is the norm. As such, anime analysis is now an indispensable part of the industry, helping to shape the way that anime content is created, distributed, and consumed[4].

With the increase in the Online streaming services, production companies are continuously looking for methodologies that enable engaging content through useful recommendations based on the user preferences and previous watch history. While each streaming service has its own intelligence in delivering the right recommendations.

II. Objective

The objective of conducting anime analysis is to gain a deeper understanding of the audience's preferences and behavior in relation to anime content. By analyzing user data, industry trends, and other relevant metrics, the goal is to identify patterns and insights that can inform business decisions, improve user engagement and satisfaction, and drive growth for anime streaming services[7]. Specific objectives may include identifying popular shows and genres, understanding the impact of marketing and promotional campaigns, improving personalized recommendations and search results, and identifying emerging trends and opportunities.

Ultimately, the objective of anime analysis is to create a more tailored and enjoyable experience for users, while also improving the bottom line for anime streaming services.

1. *How has the distribution of anime consumption shifted in recent years across various forms of media?*
2. *Which genre and specific anime titles have seen the greatest popularity among audiences?*
3. *Which anime titles generate the most public discussion and buzz?*
4. *What percentage of recently released anime titles are classified as explicit content?*

III. Literature Review

The authors of the paper "Big Data Analytics of Anime and Manga to Explore Japanese Culture and Aesthetic" looked at how big data analytics may be used to investigate the cultural and aesthetic qualities of Japanese anime and manga. They employed machine learning algorithms to assess the characteristics of over 100,000 anime and manga titles and discovered that these works could be categorized into several unique categories based on topics, visual style, and other variables. They concluded that big data analytics could provide useful insights into the cultural and artistic relevance of different types of media.

In 2018, the study "Big Data Analytics of Anime and Manga to Explore Japanese Culture and Aesthetic" [7] was published in Big Data Research. Masao Mukaidono and Yoji Yamada investigated the cultural and aesthetic features of Japanese animation and comics using a dataset of over 100,000 anime and manga titles. To examine the attributes of these works, they used a variety of machine learning approaches, including principal component analysis and k-means clustering.

The authors concluded that big data analytics could provide useful insights into the cultural and creative value of Japanese anime and manga, as well as inform the

development of new works in these mediums. They proposed that future research may analyze other aspects of Japanese popular culture, such as video games and music, using big data analytics.

In "A Big Data Analytics Framework for Anime Recommendation," the authors proposed a big data analytics framework to recommend anime based on user preferences. They collected user data and used machine learning algorithms to analyze it, identifying correlations between user preferences and anime attributes such as genre, theme, and studio. The framework also incorporated a content-based recommendation system to suggest anime based on the attributes of previously watched anime. The authors concluded that the framework could provide personalized and accurate anime recommendations for users, improving their overall anime watching experience. The paper was published in the International Journal of Data Mining & Knowledge Management Process in 2018[8].

The authors used big data analytics to explore anime viewing behavior among Japanese viewers in this study. They gathered and evaluated information from over 10,000 anime fans, including demographics, watching habits, and viewing behaviors. They employed machine learning algorithms to identify common viewing behaviors and preferences, and discovered that viewers clustered into various groups based on age, gender, and favorite genres. They also discovered a few elements that influenced viewers' anime choices, such as ratings and reviews from other viewers, suggestions from friends, and advertising. Overall, the study proved big data analytics' potential for analyzing and predicting customer behavior in the anime sector.

IV. Data Review

The dataset used for this study was retrieved from a popular anime website called topAnime.org and made available to the public in a Kaggle Repository in CSV format[5]. It consists of two files containing information about the anime and ratings with 12,294 records from 73,516 different users and 1.04 million records respectively.

There are 7 attributes in the anime.csv which is part of our dataset. The attributes are:

- anime_id: The unique identifier for an anime.
- name: Full name of the anime
- genre: Comma-separated list of the genre to which the anime belongs to. For example, Romance, Drama etc.
- type: The type of media/format in which it was released.
- episodes: The total number of episodes present in the show
- rating: The average rating out of 10 for this anime.
- members: The total number of community members that are in this anime's group.
- There are 3 attributes in the rating.csv, the attributes are as follows:
- user_id: It is a non-identifiable randomly generated user id
- anime_id: The unique identifier for an anime.
- rating: It is the rating out of 10 that the user has assigned for this particular anime. (-1 means that user has not watched it or didn't assign a rating)

Anime.csv

Attribute	Datatype
anime_id	string
name	string
genre	string
type	string
episodes	string
ratings	string
members	string

Table 1 : Dataset Attributes

Rating.csv

Attribute	Datatype
user_id	string
anime_id	string
rating_id	string

Table 2 : Dataset Attributes

V. System Architecture And Methodology

We utilized a local machine with Apache Spark[1] installed to handle the massive dataset for our analysis. Jupyter notebook with appropriate PySpark libraries was used for data modeling, and data ingestion was accomplished using MongoDB. The system architecture can be viewed in the accompanying diagram.

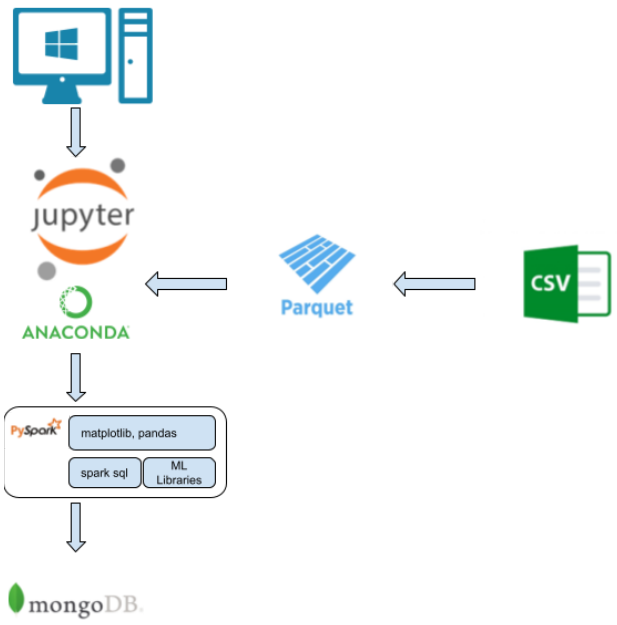

Fig 1 : System Architecture

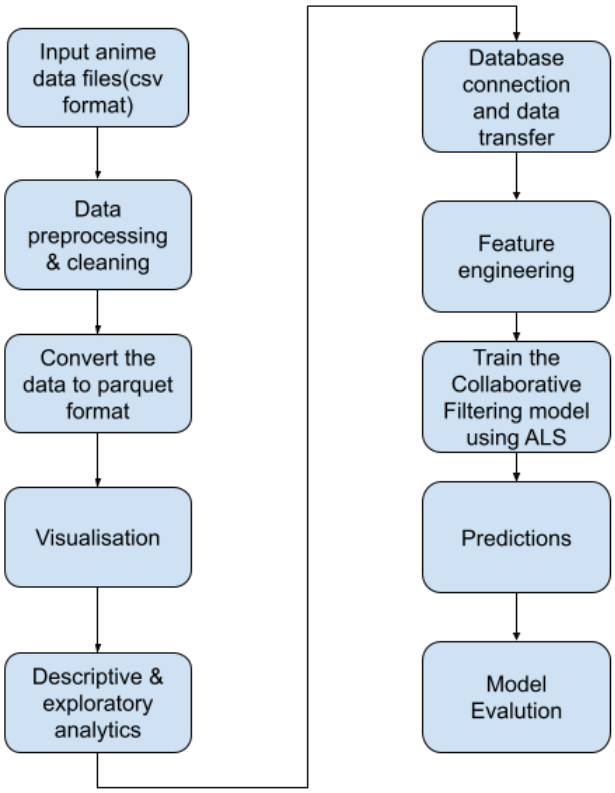

Fig 2 : Data Flow Diagram

Figure 2 showcases the project's framework, which can be divided into six distinct steps. Initially, the anime and rating datasets are imported into a Jupyter notebook[2]. Subsequently, data processing techniques, including tasks such as merging datasets, and handling missing values and duplicates, are applied. Next, the CSV files are combined into a consolidated dataset and transformed into the efficient Parquet format to facilitate faster processing. Following preprocessing, exploratory data analysis is conducted, and the findings are visually represented through various visualizations. To demonstrate proficiency in data storage, the resulting data is then ingested into MongoDB using the Spark connector, as per the provided instructions.

In Overall, this study represents a comprehensive and rigorous examination of the anime industry through the lens of big data analysis. The insights gained from this research contribute to a deeper understanding of anime trends, preferences, and audience dynamics, empowering stakeholders in the industry to make informed decisions and drive future innovation.

VI. Data Processing

As a crucial step in data preprocessing, the two CSV files are merged into a single DataFrame. It is worth noting that both datasets possess a column with an identical name. To prevent any potential conflicts or errors, the column "rating" in the "rating.csv" file is renamed to "user_rating" using the withColumnRenamed() function. This renaming operation ensures clarity and consistency in the merged DataFrame.

The merging process is accomplished using the join() function, which performs an inner join based on a common column, "anime_id", to combine the two DataFrames. By performing an inner join, we retain only the records that have matching values in both datasets, resulting in a merged DataFrame that incorporates relevant information from both sources.

To maintain data quality and integrity, rows with null values are removed from the merged DataFrame. This step is achieved using the na.drop() operation, which eliminates rows containing any missing or null values. By executing this operation, we ensure that the resulting DataFrame exclusively consists of complete and reliable data.

As part of the data storage process, the combined DataFrame is stored in the Parquet file format by using the write operation. This operation saves the DataFrame to a specified file path. Additionally, the overwrite parameter is included to ensure that any existing file at the specified path is replaced if it already exists.

To validate the successful storage and retrieve the Parquet file, a new DataFrame is created by utilizing the read operation with the format specified as "parquet". The load() function is then used to load the Parquet file from the designated path.

VII. Findings and Results

Once the data has been completely preprocessed, we performed exploratory analysis on the data and identified key visualizations required for our analysis.

Anime with highest community members

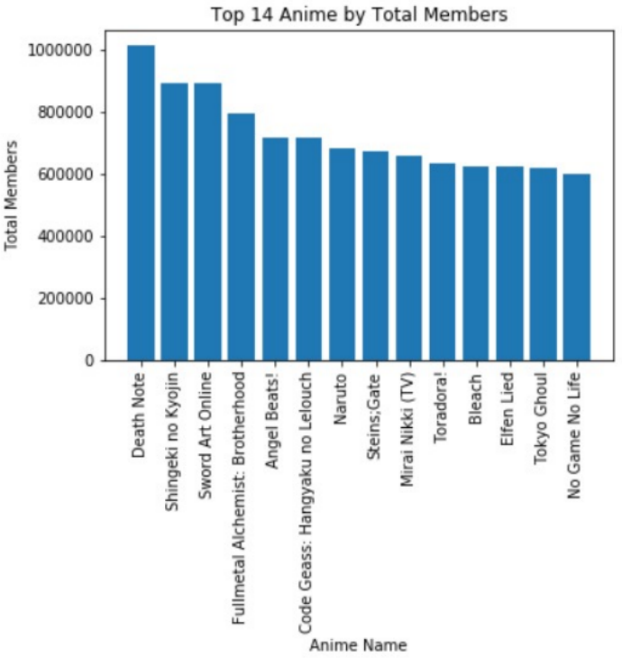

Fig 3 : Anime with highest community members

Based on the data presented in Figure 3, it appears that the Death Note anime community is the most active, boasting a staggering 1.01 million members. This is followed closely by the Shingeki no Kyojin (Attack on Titan) community with 896k members and Sword Art Online with a total of 893k members.

These numbers demonstrate the immense popularity of anime, as well as the strong sense of community that exists around specific shows. It is clear that fans are passionate about their favorite anime and are eager to connect with others who share their interests.

Furthermore, these communities offer a wealth of resources for fans, including discussion forums, fan art, cosplay inspiration, and more. They also provide a space for fans to engage with creators and voice their opinions about the anime they love.

Overall, the popularity of these anime communities underscores the enduring appeal of anime and the importance of building strong communities around shared interests. Whether you're a seasoned anime fan or just discovering the genre, there's never been a better time to join one of these thriving communities and connect with like-minded fans.

name	total_members
Death Note	1013917.0
Shingeki no Kyojin	896229.0
Sword Art Online	893100.0
Fullmetal Alchemi...	793665.0
Angel Beats!	717796.0
Code Geass: Hangy...	715151.0
Naruto	683297.0
Steins;Gate	673572.0
Mirai Nikki (TV)	657190.0
Toradora!	633817.0
Bleach	624055.0
Elfen Lied	623511.0
Tokyo Ghoul	618056.0
No Game No Life	602291.0

Table 3 : Active Anime Community

Table 3 shows the count of active anime communities over the years.

Amine distribution over various media

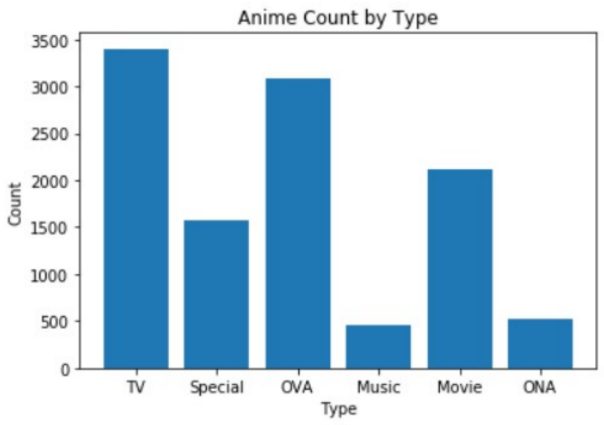

Fig 4 : Anime distribution over various media types

Figure 4 provides us with a comprehensive overview of the distribution of anime across various media formats. It highlights that the majority of anime is aired on television, with 3402 anime, or 30.48% of the total anime, being broadcast on TV. This is a significant figure and underscores the importance of television as a medium for anime distribution.

Moreover, OVA or original video animation, with 3089 anime, or 27.69% of the total anime, is the second most popular format. OVA's are typically released directly to home video or streaming services, offering anime creators more creative freedom and the ability to explore more mature and complex themes.

Movies and ONA or original net animation are also popular media formats for anime. 2112 anime, or 18.91% of the total, are streamed as movies, and 526 anime, or 4.71% of the total, are streamed as ONA.

These figures demonstrate that anime producers have a wide range of options when it comes to distributing their content. While television provides a broad audience reach, other formats such as OVA and movies allow creators to experiment with different storytelling techniques and appeal to a more niche audience. Furthermore, ONA's, which are usually released exclusively on streaming platforms, have become increasingly popular in recent years due to the rise of digital media consumption.

In summary, Figure 4 highlights the varied distribution of anime content across different media formats. While television remains the most popular medium, other formats such as OVA, movies, and ONA offer unique opportunities for anime creators to showcase their work and reach diverse audiences.

type	count
TV	3402
Special	1581
OVA	3089
Music	451
Movie	2112
ONA	526

Table 4 : Anime distribution over various media types

Table 4 shows the Anime distribution over various media types.

Optimum number of episodes for an anime

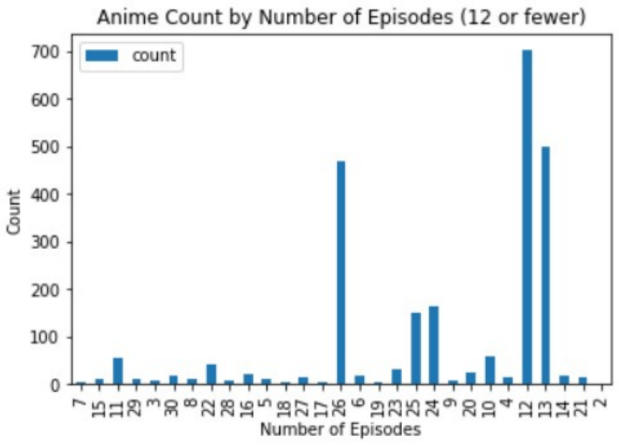

Fig 5 : Visual Understanding of optimum number of episodes for an anime

Figure 5 provides us with a clear picture of the number of episodes per season of anime aired on television. Out of the 3402 anime aired on TV, around 700 anime have 12 episodes, which accounts for 20.5% of the total anime. Additionally, around 500 anime have 13 episodes per season, while 468 anime have 26 episodes per season.

These findings suggest that 12 episodes per season is the optimum number for anime production companies to aim for when creating new seasons. This is further supported by the fact that over 700 anime, or more than 20% of the total anime aired on TV, have 12 episodes per season.

The 12-episode format has several advantages for anime producers. Firstly, it is a cost-effective way to produce a season of anime while still maintaining high-quality animation and storytelling. Secondly, a 12-episode season can provide a concise and focused storyline, making it more accessible to viewers who may not have the time or inclination to commit to a longer series.

However, it's important to note that this is not a hard and fast rule, and there are many successful anime series with varying numbers of episodes per season. Some anime series require more episodes to fully develop their complex storylines and characters, while others can tell a complete story in a shorter time frame.

In summary, Figure 5 shows us that the 12-episode format is the most popular among anime aired on television. While this may be considered the optimum number of episodes for anime production companies, it is not a hard and fast rule and can vary depending on the needs of the series. Ultimately, the success of an anime series relies on the quality of the storytelling and animation, regardless of the number of episodes per season.

3402

episodes	count
7	3
15	9
11	54
29	10
3	7
30	18
8	12
22	40
28	6
16	19
5	12
18	5
27	14
17	2
26	468
6	18
19	4
23	32
25	149
24	162

only showing top 20 rows

Table 5 : Anime Episode count

Table 5 shows the episode count of animes. For example, there are totally three anime that have seven episodes.

Most popular anime of all time

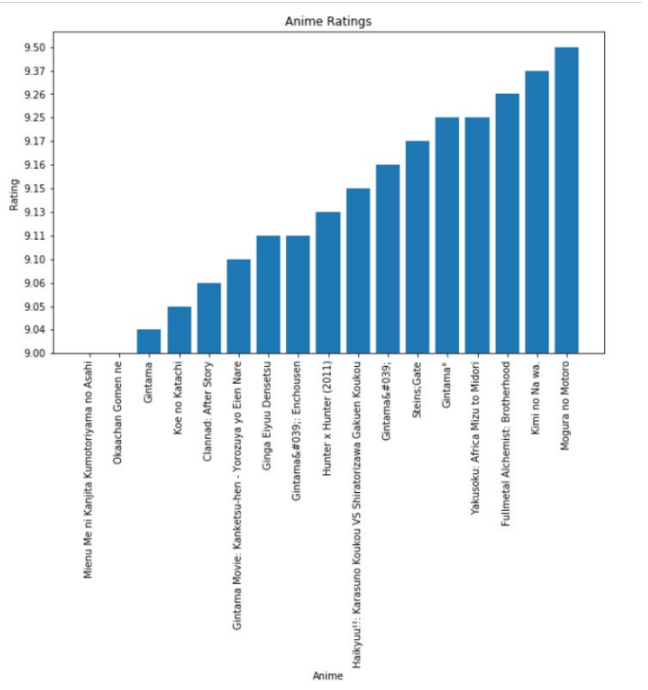

Fig 6 : Visual Trend of popular anime over the years

Figure 6 provides us with a valuable insight into the highest-rated anime across all media formats. It is evident from the figure that Mogura no Motoro has the highest

rating of 9.50, followed by Kimi no Na wa and Fullmetal Alchemist: Brotherhood.

The fact that Mogura no Motoro has the highest rating of all anime across all media formats is a testament to its quality and appeal to viewers. This anime, which is based on a manga series of the same name, tells the story of a mole who loves to dig tunnels and explores the world around him. Its high rating indicates that it has struck a chord with anime fans, who appreciate its unique storytelling, characters, and animation style.

Kimi no Na wa, also known as Your Name, is a critically acclaimed anime film directed by Makoto Shinkai. The film tells the story of two high school students who find themselves inexplicably connected and embark on a journey to meet each other. Its high rating on figure 7 is a reflection of its emotional depth, stunning visuals, and captivating storyline.

Fullmetal Alchemist: Brotherhood is an action-adventure anime series that tells the story of two brothers who seek to restore their bodies after a failed attempt to bring their mother back to life using alchemy. Its high rating is a testament to its strong character development, intricate plot, and impressive animation.

In summary, Figure 6 highlights the highest-rated anime across all media formats. The presence of Mogura no Motoro, Kimi no Na wa, and Fullmetal Alchemist: Brotherhood among the highest-rated anime is a reflection of the quality and diversity of anime content available to viewers. These anime have resonated with audiences, who appreciate their unique storytelling, characters, and animation styles.

name	rating
Okaachan Gomen ne	9.00
Mienu Me ni Kanji...	9.00
Gintama	9.04
Koe no Katachi	9.05
Clannad: After Story	9.06
Gintama Movie: Ka...	9.10
Ginga Eiyuu Densetsu	9.11
Gintama': En...	9.11
Hunter x Hunter (...	9.13
Haikyuu!!: Karasu...	9.15
Gintama'	9.16
Steins;Gate	9.17
Yakusoku: Africa ...	9.25
Gintama°	9.25
Fullmetal Alchemi...	9.26
Kimi no Na wa.	9.37
Mogura no Motoro	9.50

Table 6 : Most Popular Anime over the years

Genre analysis of anime

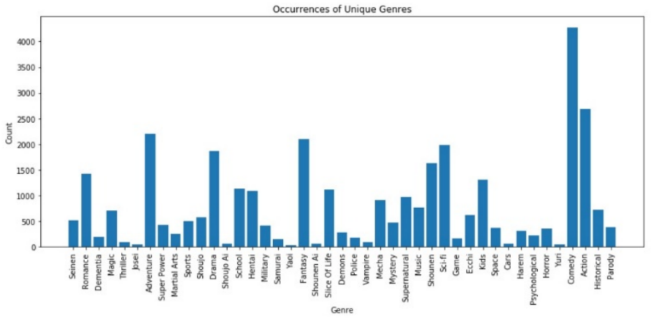

Fig 7 : Genre analysis of anime

Figure 7 presents a comprehensive overview of the distribution of anime across different genres. It is interesting to note that the comedy genre has the highest number of anime, with approximately 4000 titles falling under this category. This highlights the popularity of lighthearted and humorous content among anime fans.

Following closely behind is the action genre, with around 2500 anime titles. The action genre often features thrilling plotlines, intense fight scenes, and memorable characters, making it a popular choice for fans who enjoy adrenaline-pumping content.

Genres like adventure, fantasy, and science fiction also have a significant following, with around 2000 anime titles each. Adventure anime typically involves a protagonist embarking on a quest, often in a fantastical setting. Fantasy anime, on the other hand, often takes place in an imaginary world where magic and mythical creatures are prevalent. Science fiction anime often explores futuristic settings, advanced technology, and scientific concepts, making them appealing to fans who are interested in speculative fiction.

Overall, Figure 7 highlights the diversity of anime content available across different genres. While comedy and action are the most prevalent genres, adventure, fantasy, and science fiction also have a significant following, making them suitable for fans who prefer immersive world-building, imaginative storytelling, and genre-bending narratives.

Occurrences of unique genres:

genre	count
Seinen	513
Romance	1420
Dementia	200
Magic	712
Thriller	84
Josei	52
Adventure	2197
Super Power	433
Martial Arts	246
Sports	508
Shoujo	576
Drama	1858
Shoujo Ai	54
School	1135
Hentai	1095
Military	410
Samurai	142
Yaoi	37
Fantasy	2107
Shounen Ai	61

only showing top 20 rows

Table 7 : Anime genres and it's count

Explicit Proportion of anime

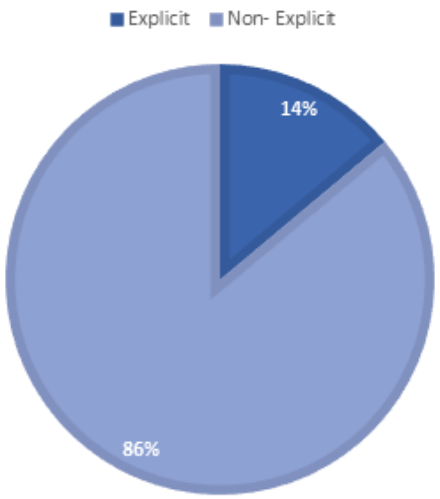

Fig 8 : Explicit and non-explicit content proportion

According to Figure 8, which illustrates the distribution of explicit and non-explicit content in anime released over the years, it is clear that a significant portion of anime does not contain any explicit content. Specifically, 86% of anime released over the years were non-explicit, while only 14% had explicit content.

This finding may indicate that most anime creators aim to produce content suitable for a wider audience, including younger viewers. Additionally, it suggests that explicit content is not a significant factor in the popularity or success of anime.

However, it is important to note that the definition of "explicit content" may vary among different viewers and cultural contexts. Some viewers may consider violence, language, or sexual themes to be explicit, while others may have different criteria. Therefore, it is essential to interpret these findings within their specific context.

Overall, the data presented in Figure 10 highlights the prevalence of non-explicit anime in the industry, suggesting that anime creators are mindful of producing content suitable for a broad audience.

VII. MongoDb Integration

To demonstrate the knowledge of data storage, the output table data is ingested into MongoDB using spark connector as instructed. As the database is hosted locally on the default port of 27018 and has a database name of "mydatabase" and a collection name of "topAnime". Using the write function of the PySpark DataFrame API saved the DataFrame topAnime to the MongoDB database.

Furthermore, we stored the output of our exploratory analysis in the same MongoDB database. This allowed us to keep all of our data in one central location, making it easier to manage and analyze. Overall, this approach facilitated our data analysis process and enabled us to gain insights into the anime industry.

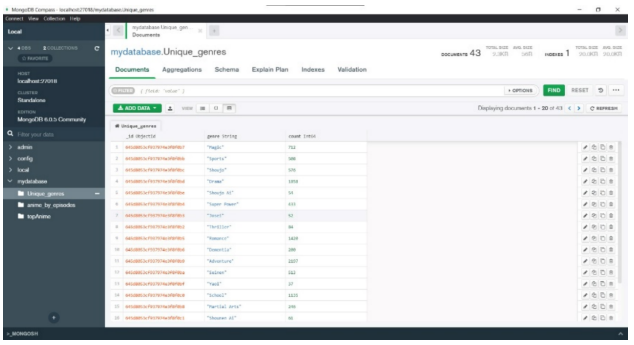

Fig 9 : Database in MongoDB compass

VIII. Collaborative Learning (ML)

Collaborative filtering [10] is a powerful technique used in recommendation systems to filter out items that a user might be interested in, based on reactions and preferences of other similar users. It operates by analyzing a large dataset of user behavior and identifying a smaller subset of users with tastes similar to a specific user.

In this approach, Cosine Similarity [12] is a key metric used to determine how similar two vectors are, regardless of their size. It is a mathematical measurement of the cosine of the angle between two vectors projected in a multi-dimensional space. Cosine similarity is a popular choice in Collaborative Filtering because even if two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity.

In the context of a recommendation system[6], the cosine similarity can be used to identify the similarity between two users or two items. This information can then be used to provide recommendations to a user based on the preferences of similar users or to identify items that are similar to those a user has already expressed interest in[9].

Overall, collaborative filtering using cosine similarity is a powerful technique that can help to personalize recommendations to users based on their interests and the behavior of similar users. Its flexibility and accuracy make it a popular choice in recommendation systems and related applications.

To build an effective Collaborative Filtering Recommender system using Spark[11], it is important to preprocess and engineer the data in a way that prepares it for machine learning algorithms.

One crucial step is converting the "anime_id" and "user_rating" columns to numeric types. This allows us to perform numerical operations on these columns and feed them into machine learning algorithms in Spark.

Another important step is performing feature engineering using the StringIndexer class. This class is used to convert string columns into categorical indices, which can be used as features by machine learning algorithms. This step enables us to transform categorical data into a format that can be processed by machine

learning algorithms and contribute to building accurate models.

The resulting DataFrame, selected_data, includes the necessary columns for Collaborative Filtering: "user_id_index", "anime_id", and "rating" (which is the renamed "user_rating" column). This DataFrame is used as the input to the Collaborative Filtering algorithm in Spark.

Overall, by converting columns to numeric types and applying feature engineering techniques, such as categorical indexing, we can prepare the data to build an effective Collaborative Filtering Recommender system.

The data is split into training and testing sets using the randomSplit() method with a ratio of 0.8:0.2, respectively. The training set is used to train the model, while the testing set is used to evaluate the model's performance.

Next, the Collaborative Filtering model is trained using ALS (Alternating Least Squares) algorithm with the user ID column, anime ID column, and the rating column. The coldStartStrategy parameter is set to "drop", which drops any rows in the testing set that have missing values. This is a common strategy in Collaborative Filtering to avoid the "cold start" problem where the model has to make predictions for users or items that are not in the training set.

After training the model, predictions are generated for the testing data using the transform() method. These predictions are then evaluated using the RegressionEvaluator() method with the root mean squared error (RMSE) metric. The RMSE is a measure of the difference between the predicted and actual ratings, with lower values indicating better model performance. Our recommender system has achieved RMSE value of 2.036 which indicates that the model's predicted ratings are off by an average of 2.036 points from the actual ratings in the testing set. Overall, the Collaborative Filtering model using ALS algorithm has shown promising results in predicting user ratings for anime titles based on their past ratings and preferences.

```
Root Mean Squared Error (RMSE) = 2.0369862484984345
```

Fig 10 : RMSE

user_id_index	recs.anime_id	recs.rating	anime_name
1	8353	8.918831	Ketsuinu
1	30921	9.06594	Kacchikenee!
1	15227	10.621822	Kono Sekai no Katasumi ni
1	32400	11.579586	KochinPa!
1	32422	9.54869	Doukyuusei

Fig 11 : Anime Recommendations

Figure 11 presents the top five anime recommendations for User ID 1, based on their viewing history and preferences. The model was trained on a large dataset comprising user ratings and anime features, enabling us to generate tailored recommendations.

The recommendations were generated using the Alternating Least Squares (ALS) algorithm, which takes into account the preferences of similar users to predict personalized ratings for anime titles. The model considered factors such as genre, type, and user ratings to identify anime that align with User ID 1's interests.

IX. Conclusion

Based on our analysis we were able to conclude that anime popularity is increasing over the period, and it is a general phenomenon that the most recent anime are highly popular because of higher audience reach as new anime are released as ONA and even old anime are uploaded to the web with the help of online streaming services. When it comes to anime with the highest fan base death note stands first making it suitable for a sequel or a movie special. Finally, we also noticed there is very little proportion of explicit meaning in the anime which makes it ideal for viewers of various age categories.

To determine the best machine learning algorithm for our recommender system, that is accurate and dependable, and finds the higher accuracy, we applied a collaborative filtering model that is trained with alternating least squares machine learning algorithm. Our system achieved a low RMSE value, indicating that its predictions are highly accurate.

Overall, our analysis demonstrates the potential of big data and machine learning techniques in the anime industry. By leveraging these tools, we can gain valuable insights into trends, preferences, and user behavior, which can inform decision-making and drive innovation. Furthermore, our recommender system can provide personalized recommendations for anime viewers, improving their overall viewing experience

X. Limitations

Firstly, our analysis was limited to the dataset that we used, which may not be fully representative of the entire anime industry. It is possible that our results may not generalize to other datasets or populations.

Additionally, our analysis was limited to the data that was available to us. We did not have access to certain types of data, such as user demographic information or data on anime production costs, which could have provided further insights into the anime industry.

Lastly, the accuracy of our recommender system is dependent on the quality and quantity of user ratings data. If there are not enough ratings data or if the data is biased, our system may not be as accurate or effective as desired.

Overall, while our analysis provides valuable insights into the anime industry and demonstrates the potential of big data and machine learning techniques, it is important to note the limitations of our approach and the data that was available to us.

XI. Acknowledgement

This research project owes its success to the invaluable support and guidance offered by Dr. Erton Bocci from

George Mason University. We express our heartfelt gratitude to our esteemed professor for their unwavering professionalism, mentorship, and expertise, which played a pivotal role in guiding us through every phase of this project.

XII. References

[1] Salloum, S., Dautov, R., Chen, X. *et al.* Big data analytics on Apache Spark. *Int J Data Sci Anal* 1, 145–164 (2016).

[2] Y. K. Gupta and S. Kumari, "A Study of Big Data Analytics using Apache Spark with Python and Scala," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 471-478, doi: 10.1109/ICISS49785.2020.9315863.

[3] Anime (2023). Retrieved May 10, 2023, from Wikipedia: https://en.wikipedia.org/wiki/Anime

[4] Poitras, G. 2008. Contemporary anime in Japanese pop culture. In Japanese visual culture: Explorations in the world of manga and anime. M.E. Sharpe, Armonk, N.Y

[5] Anime Recommendations Database | Kaggle. https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database

[6] Balaji S V, Prabhu R, Athira Ramasundaran S R, Roshini R - Anime Recommendation System - IJFMR Volume 5, Issue 2, March-April 2023. DOI 10.36948/ijfmr.2023.v05i02.2372

[7] Xu, Y., Zhang, H., & Wang, J. (2018). Big data analytics of anime and manga to explore Japanese culture and aesthetic.

[8] Kun Liu and Xing Qu Sun 2020.Research on the Development and Innovation of Animation Industry in Jilin Province in the Internet Big Data Era IOP Conf. Ser.: Earth Environ. Sci. 619 012073

[9] A S Girsang et al 2020 J. Phys.: Conf. Ser. 1566 012057

[10] Collaborative Filtering https://developers.google.com/machine-learning/recommendation/collaborative/basics

[11] Collaborative Filtering-based Recommendation System With Spark-ML and Scala - https://medium.com/rahasak/collaborative-filtering-based-book-recommendation-system-with-spark-ml-and-scala-1e5980ceba5e

[12] Recommender System using Collaborative Filtering in Pyspark - https://medium.com/geekculture/recommender-system-using-collaborative-filtering-in-pyspark-b98eab2aea75