Scalable Visualization Techniques for Big Data using Distributed Systems

Jonathan Tong
Texas A&M University/ Computer Science Graduate Student

Satvik Praveen
Texas A&M University/ Data Science Graduate Student

Matt Palmer
Texas A&M University/ Data Science Graduate Student

Kaitlyn Griffin
Texas A&M University/ Computer Under-Graduate Student

Reed Palmer
Texas A&M University/ Visualization Under-Graduate Student

***Abstract*—This project investigates big data visualization techniques applied to large-scale tabular, image, audio, and text data collected from the NHL website. We successfully gathered 11 seasons' worth of data, encompassing player statistics, game records, and multimedia content. For the audio data, we analyzed stadium noise during games and goals by creating MFCCs, waveforms, and spectrograms. In the case of tabular data, we visualized key game attributes, including the number of skaters and goalies per game. Image data was used to generate 2D embeddings, followed by clustering to identify groups of similar-looking player images. The textual data analysis involved plotting frequent words, creating word clouds, visualizing data with t-SNE for dimensionality reduction, and generating a TF-IDF correlation plot. Additionally, we incorporated an interactive Plotly visualization to analyze word frequency in the scraped articles. For parallel processing, we utilized Dask on local CPU cores, with most of the analysis conducted on Google Colaboratory.**

INTRODUCTION With the exponential growth of data in various domains, traditional data visualization techniques often struggle to handle the sheer volume, variety, and velocity of big data [1]. The need for scalable visualization solutions has become more pronounced, as organizations increasingly rely on complex, large-scale datasets to make informed decisions. Scalable visualization techniques aim to address this challenge by enabling effective data exploration and insight extraction, even with massive datasets [2]. However, achieving scalability requires leveraging distributed systems that can efficiently process, analyze, and visualize big data in parallel.

Distributed computing frameworks, such as Apache Spark, Dask, and Apache Hadoop, have become essential tools for handling big data workloads [3] [4]. These frameworks distribute data and computational tasks across multiple processors or nodes, significantly enhancing processing speed and allowing for real-time analysis. By integrating these

frameworks with advanced visualization libraries and tools, it becomes possible to create responsive, interactive visualizations that provide valuable insights into complex datasets.

This project investigates scalable visualization techniques tailored for big data analysis using distributed systems. By leveraging parallel processing and distributed computing, we aim to visualize diverse data types, including tabular, image, audio, and text data, to reveal patterns, trends, and correlations that would be difficult to identify using traditional methods. The focus is on combining the computational power of distributed systems with advanced visualization strategies to handle the high dimensionality and size of big data effectively.

Background

The rapid expansion of data in recent years has introduced significant challenges in data analysis, particularly when dealing with large, heterogeneous datasets. Traditional data visualization methods, which typically focus on smaller and simpler datasets, often fail to scale effectively with the volume and complexity of modern big data [5]. This has necessitated the development of scalable visualization techniques capable of handling the vast amounts of data generated across various domains, including sports analytics.

The National Hockey League (NHL), known for its extensive collection of player statistics, game records, and multimedia content, provides a rich dataset for analysis. The data spans multiple seasons and includes a variety of formats: massive tabular data (e.g., game statistics), image data (e.g., action shots of players), audio data (e.g., stadium noises), and text data (e.g., articles and news reports) [6]. Each of these data types presents unique challenges for visualization due to their size and complexity. For example, tabular data often involve high-dimensional attributes, while image data require advanced feature extraction techniques for meaningful analysis. Similarly, audio data need to be processed for frequency-based features, and text data involve complex linguistic patterns that can be difficult to interpret visually.

Existing visualization libraries, such as Matplotlib, Seaborn, and Plotly, provide robust tools for data exploration but often struggle with performance when handling big data [7]. To address these challenges, distributed computing frameworks like Dask have been employed for parallel processing. Dask, specifically designed to work with Python, offers an efficient way to manage large datasets by distributing tasks across multiple CPU cores. This allows for faster processing and enables the creation of interactive and dynamic visualizations even with limited computational resources.

The goal of this project is to leverage scalable visualization techniques using distributed systems to explore and analyze diverse data types from the NHL. By utilizing parallel processing and dimensionality reduction techniques, the project aims to uncover patterns and insights that would be difficult to achieve with conventional methods. The use of distributed computing not only facilitates the handling of big data but also enhances the efficiency of the visualization process, making it feasible to work with a diverse and complex dataset on platforms like Google Colaboratory.

Methodology

In this section, we outline the comprehensive approach taken to process, analyze, and visualize the diverse datasets collected for this project. The methodology covers the handling of four main data types: audio, image, textual, and massive tabular data. Each type required unique preprocessing and analysis techniques tailored to its specific characteristics and challenges. We leveraged a variety of data collection tools and machine learning techniques, along with distributed computing frameworks, to efficiently process large-scale datasets [4]. The subsequent subsections describe the detailed methods employed for each data type, including data collection, feature extraction, and visualization strategies, highlighting the use of parallel processing and dimensionality reduction techniques to enable scalable visualization and analysis [8].

Massive Tabular Data

Play-by-play data was gathered from the National Hockey League's (NHL) back-end API, which is publicly accessible online. We scraped the data for all games from the 2013 to 2023 seasons. In total, this amounted to 15,361 games, with each game containing upwards of 300 recorded events. A purpose-built web scraper, implemented in Python, was used to collect the data. Initially, we attempted to run the scraper in parallel; however, due to memory constraints, it was more efficacious to run it sequentially, despite the longer processing time. Managing memory load effectively while scraping data in parallel remains an

interesting avenue for future work.

Data Cleaning:
After scraping the data, a thorough cleaning process was undertaken to remove errors and ensure consistency. We identified and excluded games that were not recorded correctly by the NHL, as these often contained only goals and penalties. Including such games could distort the dataset, for instance by inflating shooting percentages, as these games would show a shooting percentage of 100%, while the average NHL shooting percentage on all shot attempts is typically under 10%. Additional fields were also added to the dataset to enhance usability, such as tracking the number of skaters on the ice for each team, adding shot information, and providing context around events (e.g., preceding and following events). The final cleaned dataset consisted of a series of `.csv` files, one for each season, containing detailed event records for every game.

Data Analysis:
The data was cleaned for discrepancies using several custom scripts, applied individually to both the game and player data scraped from the API. The data was primarily split by season (e.g., all games in the 2023-2024 season), allowing for queries on individual teams and players. No additional filtering was applied to the audio-visual data, as models working with this type of data must be robust to noise in a live environment. Furthermore, audio-visual data consists of raw game broadcasts without any editing or processing, simulating real-world conditions.

Audio Data

The analysis of audio data in this project involves handling unique challenges posed by the high volume, velocity, and complexity of sound data. To effectively process and visualize audio data, we employed the following methods:

- **Data Collection and Preprocessing:**
  - The audio dataset was sourced from stadium noises during NHL games, focusing on capturing crowd reactions during key events such as goals. The raw audio files were in formats like `.wav` and `.mp3`.
  - To prepare the audio data for analysis, we performed transformations to convert the sound waves into analyzable formats. Techniques such as *Short-Time Fourier Transform (STFT)* and *Mel-Frequency Cepstral Coefficients (MFCC)* were used for feature extraction. These transformations allowed us to break down the audio into smaller, time-segmented features, revealing underlying patterns in the sound.
- **Feature Extraction:**
  - *Waveform Analysis:* We visualized the raw waveforms to observe the amplitude variations over time, providing a basic understanding of sound intensity and temporal characteristics.
  - *MFCCs:* Mel-Frequency Cepstral Coefficients were computed to capture the spectral properties of the audio. MFCCs effectively represent the timbre of the sound, making them suitable for pattern recognition tasks.
  - *Spectrograms:* Spectrograms were generated using STFT, displaying the frequency spectrum of the audio over time. This visualization helped identify frequency components and anomalies during key events like goals.
- **Distributed Processing:**
  - Given the large size of the audio dataset, we utilized distributed processing techniques with frameworks like *Dask*. Dask enabled parallel processing of audio files across multiple CPU cores, significantly reducing the time required for feature extraction and visualization.
- **Visualization Techniques:**
  - Various visualization techniques were employed to interpret the audio data, including:
    - *Waveform Plots:* Displaying amplitude over time to observe sound variations during different game events.
    - *Spectrograms:* Visual representations of the frequency spectrum, highlighting changes in sound frequency over time.
    - *MFCC Heatmaps:* Heatmaps of MFCC features were plotted to visualize patterns in the spectral characteristics of the audio, aiding in identifying distinct sounds and crowd reactions.

These methods provided a comprehensive approach to analyzing and visualizing complex audio data, enabling us to uncover meaningful patterns and insights from the stadium noise recordings.

Image

For this project, we downloaded action images of 5 NHL players to analyze visual patterns and perform clustering. The image data processing and analysis involved the following steps:

- **Data Collection and Preprocessing:**
  - The images were sourced from online repositories, capturing players in action. The images were standardized to ensure uniform analysis by resizing them to a fixed size and normalizing the pixel values.
  - We applied a series of transformations using *PyTorch* and *PIL (Python Imaging Library)*. The images were resized to 256×256 pixels, center-cropped to 224×224 pixels, converted to tensors, and normalized using standard mean and standard deviation values for RGB channels.
- **Feature Extraction:**
  - We utilized a pre-trained *ResNet-50* model from the *Torchvision* library to extract deep features from each image. The model was set to evaluation mode, and the features were extracted from the final fully connected layer.
  - The extracted features were converted to NumPy arrays for further analysis, capturing the high-level visual representations of the images.
- **Dimensionality Reduction:**
  - Given the high dimensionality of the extracted features, we applied *Principal Component Analysis (PCA)* to reduce the feature space while retaining most of the variance. We determined the optimal number of components based on the dataset size, using a maximum of 20 components.
  - The PCA-reduced features were further processed using *t-Distributed Stochastic Neighbor Embedding (t-SNE)* for visualization. This technique allowed us to visualize the feature embeddings in a 2D space, making it easier to identify clusters.
- **Clustering and Visualization:**
  - We performed clustering on the 2D embeddings using *K-means* clustering, aiming to group similar-looking images. Two distinct clusters were formed based on the visual characteristics of the players' images.
  - The results were visualized using scatter plots, where each point represents an image, and the colors indicate the assigned cluster. This helped us identify patterns and similarities in the player images effectively.

These methods provided a comprehensive approach for analyzing and visualizing image data, leveraging deep feature extraction and dimensionality reduction techniques to uncover meaningful visual patterns.

Textual

For the textual data analysis in this project, we scraped articles from online sources related to NHL games and events. The following steps outline the methods employed for data collection, preprocessing, and analysis:

- **Data Collection:**
  - We used *BeautifulSoup* for web scraping to extract article content from various NHL-related websites. The articles were primarily extracted from paragraph tags (`<p>`), which typically contain the main text content.
  - The data was collected using *concurrent processing* with the `ProcessPoolExecutor` from Python's `concurrent.futures` library. This enabled efficient parallel fetching of articles, reducing the overall data collection time.
- **Data Preprocessing:**
  - The text data was preprocessed to clean and standardize the content. This included converting the text to lowercase, removing punctuation, and eliminating non-alphabetic characters using regular expressions.
  - Stop words were removed using the *NLTK* library's predefined list of English stop words. We also applied *lemmatization* using the `WordNetLemmatizer` from *NLTK* to reduce words to their base forms, aiding in better text analysis.
- **Feature Extraction:**
  - We employed *TF-IDF (Term Frequency-Inverse Document Frequency)* vectorization to extract features from the preprocessed text. This method helps in identifying the most important terms in the corpus by weighing terms based on their frequency and significance across documents.
- **Visualization Techniques:**

– Various visualization techniques were used to analyze the textual data, including:

  * *Frequent Word Plot:* Displaying the most common words in the articles to identify key topics and trends.
  * *Word Cloud:* A visual representation of word frequency, where the size of each word indicates its frequency in the text.
  * *t-SNE Visualization:* We applied *t-Distributed Stochastic Neighbor Embedding (t-SNE)* on the TF-IDF vectors for dimensionality reduction, enabling us to visualize the text data in a 2D space.
  * *TF-IDF Correlation Plot:* A correlation plot was generated using the TF-IDF matrix to show relationships between different terms in the corpus.
  * *Interactive Plotly Visualization:* An interactive plot was created using *Plotly* to analyze word frequency dynamically across the scraped articles.

These methods provided a comprehensive framework for extracting, processing, and visualizing textual data, allowing us to gain insights into the linguistic patterns and themes present in NHL-related articles.

Results

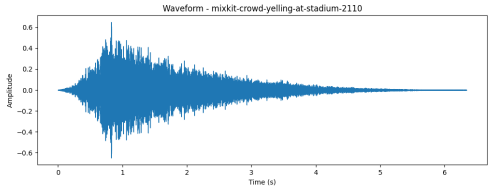
In this section, we present the findings from our analysis of the audio, image, textual, and massive tabular data collected from the NHL. The results showcase the visualizations and key insights obtained using our scalable data processing and analysis techniques.

Audio Data Results

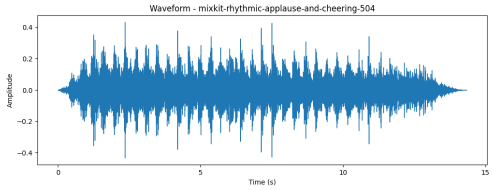
The analysis of audio data provided several key visual insights:

- We generated **waveform plots** to observe variations in sound intensity during different game events. Peaks in the waveforms corresponded to crowd reactions during goals and critical moments.

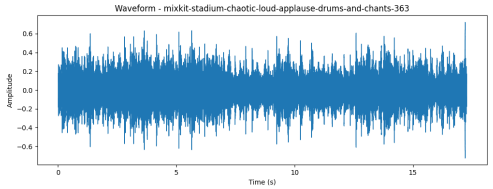
  Waveform plot- Crowd yelling



  Waveform plot- Rhythmic Applause and Cheering

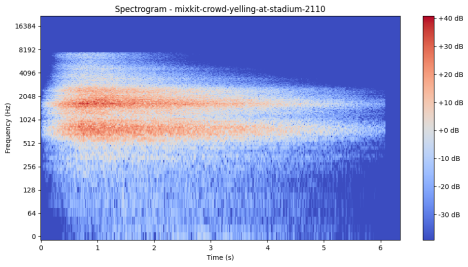


  Waveform plot- Chaotic loud applause: Drums and Chants

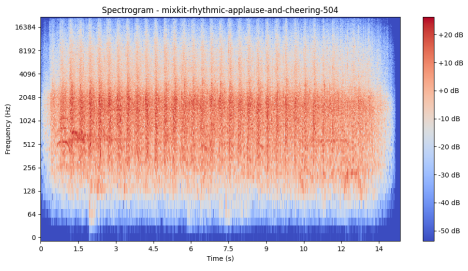


- **Spectrograms** were used to visualize the frequency spectrum of the audio data. Changes in high-frequency components were evident during periods of crowd excitement, highlighting the effectiveness of our audio analysis methods.

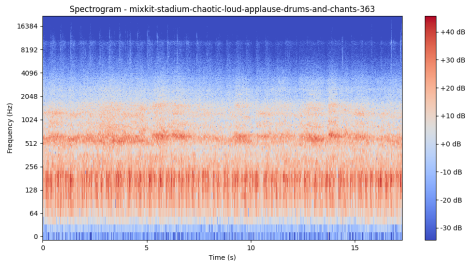
  Spectogram plot- Crowd yelling



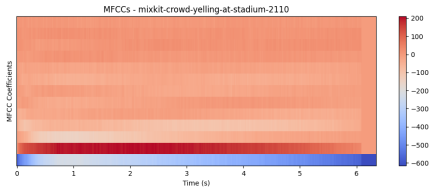
  Spectogram plot- Rhythmic Applause and Cheering



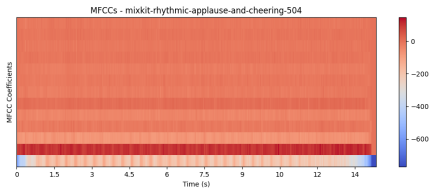
  Spectogram plot- Chaotic loud applause: Drums and Chants



- The **MFCC heatmaps** revealed distinct patterns, distinguishing regular game sounds from intense reactions during key events. **MFCC plot- Crowd yelling**



MFCC plot- Rhythmic Applause and Cheering



MFCC plot- Chaotic loud applause: Drums and Chants

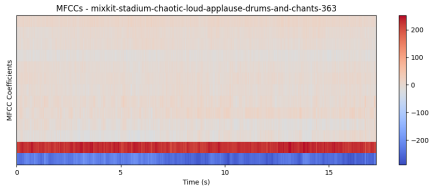
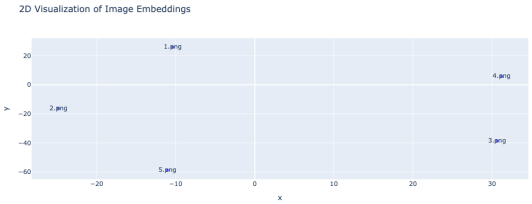


Image Data Results

The image data analysis yielded valuable clustering results:

- Using **t-SNE** for dimensionality reduction, we visualized the high-dimensional image features in a 2D space, revealing clear clusters of similar images.
- **K-means clustering** identified two distinct groups of player images based on their action poses. The clustering results demonstrated the model's ability to capture visual similarities effectively.
- Visual inspection of the clustered images confirmed that the groups corresponded to different types of player actions (e.g., shooting vs. defending).

2D-plot Image Embeddings



Textual Data Results

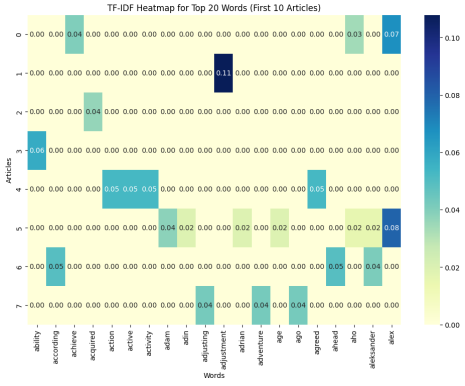
The textual data analysis provided insights into common themes and patterns across the scraped articles:

- **Word cloud visualizations** highlighted frequently occurring terms related to game events and player performance, providing a quick overview of the main topics discussed.

Word Cloud

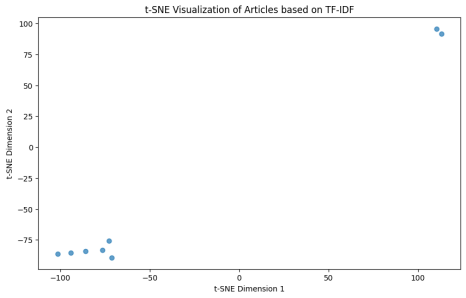
- The **TF-IDF analysis** revealed strong correlations between certain keywords, indicating dominant themes across the text corpus.

TF-IDF



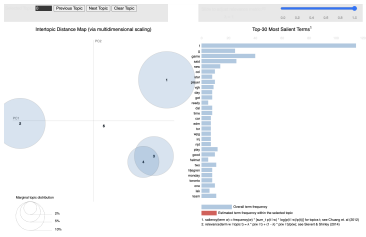
- **t-SNE visualizations** of the text embeddings showed clear clusters of articles based on their content, effectively separating game summaries from player interviews.

t-SNE Visualization

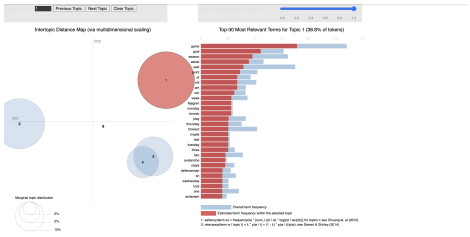


- An **interactive Plotly visualization** was created to analyze word frequency dynamically, allowing for deeper exploration of the text data.

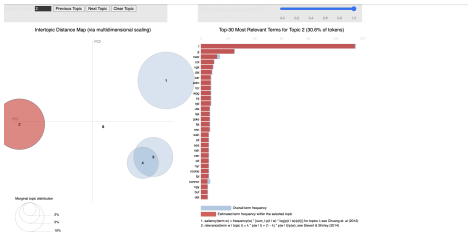
Initial Plot



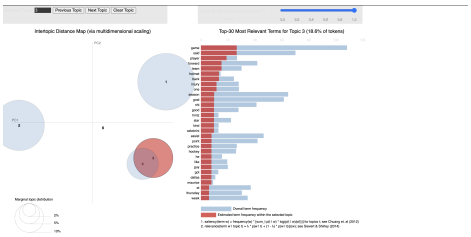
First Plot



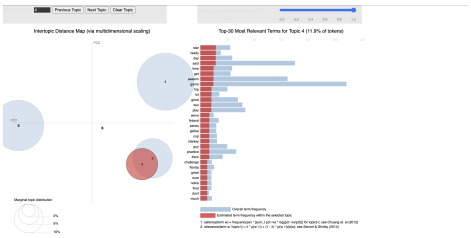
Second Plot



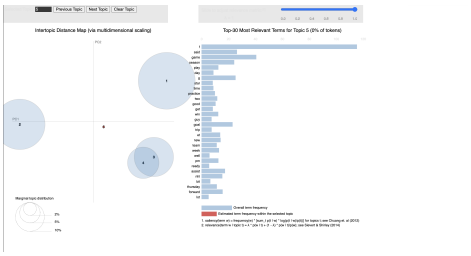
Third Plot



Fourth Plot

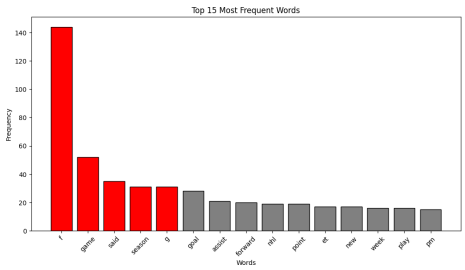


Fifth Plot



- **Frequent Words Plot:** The frequent words plot is a histogram that visualizes the top 15 most frequently occurring words across the scraped articles. As shown in Figure , this plot highlights the words with the highest frequency, such as ”game,” ”goal,” and ”season.” The analysis of word frequency helps in identifying common themes and topics discussed in the NHL-related articles, providing an overview of the most talked-about aspects of the games.

Frequent Words Plot

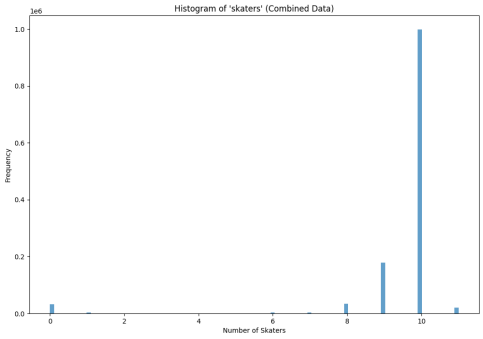


Massive Tabular Data Results

The analysis of the massive tabular dataset uncovered important trends and patterns:

- We visualized key game statistics, including the **number of skaters** and **shot attempts**, revealing consistent trends across the 11 NHL seasons.
- **Shooting percentage analysis** indicated a slight upward trend in recent seasons, aligning with changes in gameplay strategies.
- **Anomaly detection** highlighted irregularities in certain games, such as unusually high shooting percentages, which were further investigated to understand potential data issues or gameplay anomalies.

Plot: Combined Skaters



Discussion

The results of this project demonstrate the effectiveness of scalable visualization techniques in handling large, complex datasets across multiple data types. The analysis of massive tabular data provided valuable insights into NHL game trends, revealing consistent patterns in player statistics and shooting percentages over the 11 seasons. The image data analysis successfully identified visual similarities among player action images, which could be further leveraged for player recognition or automated tagging in sports media. The audio analysis captured significant variations in crowd noise, particularly during key events like goals, providing a novel way to gauge crowd reactions and game excitement. Textual data analysis highlighted dominant themes and topics discussed in NHL-related articles, offering a deeper understanding of the narrative surrounding games and player performances.

Challenges and Limitations

Several challenges were encountered throughout the project. The attempt to scrape data in parallel faced memory constraints, necessitating a sequential approach that increased processing time. Additionally, the high dimensionality of the image and audio data posed difficulties for visualization, which were mitigated through the use of dimensionality reduction techniques like *Principal Component Analysis (PCA)* and *t-Distributed Stochastic Neighbor Embedding (t-SNE)* [8]. However, the exclusion of video data due to computational limitations remains a notable gap that could be addressed in future work.

We faced significant challenges in scraping the data due to the security measures applied to the NHL website. This issue affected both the audio and image data collection, preventing us from scraping the data directly. As a result, we had to manually download the required datasets, which was time-consuming and hindered our progress [3]. Additionally, our attempts to work with video data encountered multiple hurdles. Apart from the security challenges in downloading video files, we faced complexities in processing the video data due to infrastructure limitations. Even with parallelization efforts, the processing of video data proved to be infeasible given our computational resources.

While using *Dask* for handling massive tabular data, we also encountered difficulties. A persistent issue was the disparity between the structure of the scraped CSV files and the expected table format, which led to continuous errors during data loading and analysis. This challenge highlighted the need for better error handling and data validation mechanisms.

Comparison with Existing Methods

Our approach, combining distributed computing with scalable visualization methods, provided a significant improvement over traditional visualization techniques for big data [8] [9]. The use of *Dask* for parallel processing enabled efficient handling of large datasets, while advanced techniques like t-SNE allowed us to effectively visualize high-dimensional data. This approach highlights the potential of distributed systems

in sports analytics, especially for analyzing diverse data types in a unified framework.

Implications of the Findings

The findings from this project have several implications for the field of sports analytics. The insights gained from the tabular data analysis could help teams refine their strategies based on historical performance trends [2]. The image and audio analysis methods could be integrated into media applications for enhanced player recognition and crowd sentiment analysis. Furthermore, the textual analysis offers a unique perspective on public narratives and media coverage, which could be useful for fan engagement and marketing.

Future Work

Future work could focus on overcoming the computational limitations faced during data scraping and analysis. We plan to explore more efficient methods for handling video data, overcoming both security and infrastructure challenges [10]. Additionally, integrating a *Large Language Model (LLM)* could automate the generation of reports based on the visualizations, enhancing the interpretability and accessibility of the results [2].

The current version of the application could be expanded to support the visualization of additional big data sets. A natural extension would involve accommodating the processing and visualization of other sports-related big data. Users could upload large files or provide links to existing datasets available on websites or in repositories. Enhancing the data scraper to be more generalized would allow it to automatically scrape and clean various provided datasets.

Further improvements would include refining the memory management in the current parallel program, enabling it to work efficiently on typical user machines. Enhancing the parallelization of data processing and visualization generation would significantly reduce response times, leading to an improved user experience. These optimizations would result in faster speedups and a more seamless interaction with the visualizations, ultimately increasing the efficiency of the entire analysis process.

Conclusion

In this project, we explored scalable visualization techniques for big data analysis using a diverse set of data types collected from the NHL, including massive tabular, image, audio, and textual data. By leveraging distributed computing frameworks and advanced machine learning methods, we were able to handle the complexity and size of the datasets effectively. The analysis of tabular data provided insights into game statistics and trends across multiple seasons, while the image data analysis successfully grouped player action images based on visual similarities. Audio data visualizations revealed distinct patterns in crowd noise, highlighting key moments during games. The textual data analysis uncovered common themes and topics discussed in NHL-related articles, offering a deeper understanding of public narratives around the games.

Despite the computational challenges, especially with parallel data scraping and handling high-dimensional data, our methodology proved effective in extracting meaningful insights from the large datasets. The use of Dask for parallel processing and dimensionality reduction techniques like PCA and t-SNE played a crucial role in making the visualization process scalable and efficient.

The project demonstrated the potential of combining distributed computing with scalable visualization techniques to analyze and interpret complex big data [4]. However, there are areas for future improvement, such as implementing more robust parallel scraping techniques and incorporating real-time analysis capabilities for live data streams [10]. Overall, this work highlights the importance of scalable approaches in the visualization of big data and sets the groundwork for further exploration in sports analytics and other domains with large-scale, heterogeneous datasets.

REFERENCES

1. A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International Journal of Information Management*, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401214001066
2. M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. De-la Hoz-Franco, and E. De-La-Hoz-Valdiris, “Trends and future perspective challenged in big data,” in *Springer*, 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-16-5036-9_30
3. D. Fisher, “Big data exploration requires collaboration between visualization and data infrastructures,” in *Proceedings of the International Conference on Big Data*, 2016. [Online]. Available:

https://dl.acm.org/doi/abs/10.1145/2939502.2939518
4. Yadranjiaghdam and et.al, “A survey on real-time big data analytics: Applications and tools,” *IEEE Access*, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7881376
5. N. Bikakis, “Big data visualization tools,” *arXiv preprint arXiv:1801.08336*, 2018. [Online]. Available: https://arxiv.org/pdf/1801.08336
6. L. Wang, G. Wang, and C. A. Alexander, “Big data and visualization: Methods, challenges and technology progress,” *International Journal of Digital Technology and Applications*, 2015. [Online]. Available: https://pubs.sciepub.com/dt/1/1/7/
7. S. Saraswathi, G. Deepa, G. Vennila, S. Parthasarathy, and B. Ramadoss, “A survey on big data: Infrastructure, analytics, visualization and applications,” *Journal of Big Data*, 2022. [Online]. Available: https://eds.p.ebscohost.com/eds/pdfviewer/pdfviewer?vid=3&sid=207d837d-32e6-4b30-bc6b-40850ad6a1e6%40redis
8. C. H. Mendhe and colleagues, “A scalable platform to collect, store, visualize, and analyze big data,” *IEEE Transactions on Computational Social Systems*, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9107226
9. A. Arleo, W. Didimo, G. Liotta, and F. Montecchiani, “Large graph visualizations using a distributed computing platform,” *Information Sciences*, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025516318011
10. I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, “Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of covid-19 pandemic cases and contact tracing,” *International Journal of Environmental Research and Public Health*, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/15/5330