1. **Introduction**
   * Overview of the Project
   * Objectives and Goals
2. **Cluster Setup**
   * Setting Up the Hadoop Cluster
   * Challenges Faced during Cluster Setup
   * Solutions Implemented
3. **Web Crawling**
   * Tools and Technologies Used
   * GitHub Repository Setup
   * Extracting Data from the Cricbuzz Website
4. **Data Processing and Storage**
   * Overview of the Data Processing Pipeline
   * Data Extracted from the Website
   * Saving Data on Hadoop Distributed File System (HDFS)
5. **Failure Handling of Worker Nodes**
   * Challenges in Worker Node Failures
   * Strategies Implemented for Handling Failures
   * Ensuring Fault Tolerance in the System
6. **Database Setup**
   * Choice of Database (PostgreSQL)
   * Database Schema Design
   * Migrating Data from HDFS to PostgreSQL
7. **Running Queries**
   * Executing Queries on the Database
8. **Conclusion**
   * Summary of Key Achievements
   * Lessons Learned
   * Future Improvements
9. **Q&A Session**
   * Open the floor for questions from the audience.
10. **Demonstration**
    * Live or recorded demonstration of key functionalities (if applicable).

**Overview of the Project:**

In this project, we aim to leverage web scraping and data processing techniques to extract valuable cricket-related information from the Cricbuzz website. The primary goal is to create a robust data pipeline that not only collects data efficiently but also stores it in a structured manner. The project revolves around setting up a Hadoop cluster for distributed data storage, utilizing web crawling tools, and implementing a PostgreSQL database for structured data management.

**Objectives and Goals:**

1. **Web Crawling and Data Extraction:**
   * Extract comprehensive cricket statistics, including team rankings, player rankings, match details, and news articles, from the ICC, ESPN, Cricbuzz websites.
   * Develop a web crawling mechanism to navigate through different sections of the website and extract relevant data.
2. **Cluster Setup and Management:**
   * Implement a Hadoop cluster to facilitate distributed storage using HDFS (Hadoop Distributed File System).
   * Address challenges related to the setup, configuration, and maintenance of the Hadoop cluster.
3. **Data Processing Pipeline:**
   * Create an efficient data processing pipeline to clean and transform the extracted data.
   * Implement mechanisms for handling and recovering from worker node failures to ensure fault tolerance.
4. **Database Management:**
   * Choose PostgreSQL as the relational database management system (RDBMS) for storing structured data.
   * Design a suitable database schema to accommodate cricket-related information, including teams, players, matches, and news articles.
5. **Data Migration and Integration:**
   * Develop a process for migrating data from HDFS to PostgreSQL, ensuring data consistency and integrity.
   * Explore strategies for efficiently integrating data from various sources into the relational database.
6. **Query Execution and Optimization:**
   * Run queries on the PostgreSQL database to analyze and retrieve meaningful insights.
7. **Documentation and Presentation:**
   * Document the project setup, configurations, and methodologies used for future reference.
   * Prepare a comprehensive presentation covering key aspects of the project for effective knowledge sharing.

The ultimate objective is to create a robust, scalable, and efficient system that not only extracts and processes cricket-related data but also provides a foundation for future enhancements and integrations. The project's goals align with achieving a seamless data flow from web scraping to database storage, enabling valuable insights and analysis for cricket enthusiasts and analysts alike.

1. **Web Scraping:**
   * **Beautiful Soup:** A Python library for pulling data out of HTML and XML files, facilitating web scraping tasks.
   * **Selenium:** A web testing library used for automating browser actions, enabling dynamic content extraction.
2. **Hadoop Cluster Setup:**
   * **Hadoop Distributed File System (HDFS):** A distributed file system designed to store large volumes of data across multiple machines.
   * **Apache Hadoop:** An open-source framework for distributed storage and processing of big data.
3. **Database Management:**
   * **PostgreSQL:** A powerful open-source relational database management system (RDBMS) known for its reliability and extensibility.
4. **Version Control:**
   * **Git:** A distributed version control system for tracking changes in source code during software development.
5. **Data Processing and Analysis:**
   * **Pandas:** A data manipulation and analysis library for Python, ideal for cleaning, transforming, and analyzing structured data.
   * **NumPy:** A library for numerical operations in Python, commonly used for data manipulation.

**GitHub Repository Setup:**

The project's source code and related files are organized and managed using a Git repository hosted on GitHub. The repository structure includes:

1. **Main Codebase:**
   * Web scraping scripts written in Python using Beautiful Soup and Selenium.
   * Configuration files for Hadoop cluster setup and management.
   * Data processing and migration scripts.
2. **Documentation:**
   * README.md file providing an overview of the project, installation instructions, and usage guidelines.
   * Code comments and inline documentation for better code understanding.
3. **Data Storage:**
   * CSV files containing raw and processed data.
   * SQL scripts for database schema creation and data insertion.
4. **Presentation:**
   * A presentation directory containing slides, images, or any visual aids used in presentations.

**Extracting Data from the Cricbuzz Website:**

The web scraping process involves the following key steps:

1. **URL Navigation:**
   * Selenium is used to automate browser actions and navigate through different sections of the Cricbuzz website.
2. **Data Extraction:**
   * Beautiful Soup is employed to parse HTML content and extract relevant data elements based on their HTML tags and attributes.
3. **Dynamic Content Handling:**
   * For pages with dynamic content loading, Selenium ensures that the script waits for the content to be fully loaded before attempting extraction.
4. **Data Storage:**
   * Extracted data is stored in appropriate data structures (e.g., lists or dictionaries) before being processed or saved to CSV files.

By combining these tools and technologies, the project establishes a seamless pipeline for extracting, processing, and storing cricket-related data from Cricbuzz. The GitHub repository serves as a central hub for collaboration, version control, and documentation.

**Database Setup:**

The choice of a suitable database system is crucial for efficient data storage, retrieval, and management. In this project, PostgreSQL was selected as the relational database management system (RDBMS) due to its robustness, scalability, and support for complex queries.

**Choice of Database (PostgreSQL):**

* **Reliability:** PostgreSQL is known for its reliability and stability, making it a preferred choice for handling critical data.
* **Extensibility:** Its extensible architecture allows the addition of custom functions and features, making it adaptable to various use cases.
* **ACID Compliance:** PostgreSQL adheres to the ACID (Atomicity, Consistency, Isolation, Durability) properties, ensuring data integrity and consistency.
* **Open Source:** Being an open-source database, PostgreSQL provides cost-effective solutions without compromising on performance.

**Database Schema Design:**

Effective database schema design is essential for organizing and structuring data efficiently. In this project, the schema design revolves around key entities such as Team, Player, Match, and NewsArticle.

1. **Team Table:**
   * Attributes: Team\_ID (Primary Key), Team\_Name, Ranking, Points, Game\_Type
2. **Player Table:**
   * Attributes: Player\_ID (Primary Key), Player\_Name, Country, Ranking, Points, Game\_Type, Category
3. **Match Table:**
   * Attributes: Match\_ID (Primary Key), Match\_Name, Venue, Result
4. **NewsArticle Table:**
   * Attributes: Article\_ID (Primary Key), Headline, Content, Author, Publication\_Date

The schema design ensures normalization, minimizing data redundancy and maintaining relationships between entities.

**Migrating Data from HDFS to PostgreSQL:**

1. **Data Extraction from CSV:**
   * CSV files containing extracted and processed data are read using tools like Pandas in Python.
2. **Data Transformation:**
   * Data may be transformed and cleaned as needed to fit the relational schema.
3. **Establishing Database Connection:**
   * Connect to the PostgreSQL database using appropriate credentials and connection parameters.
4. **Inserting Data:**
   * Utilize SQL queries or an ORM (Object-Relational Mapping) library like SQLAlchemy to insert data into respective tables.
5. **Handling Errors:**
   * Implement error handling mechanisms to manage data integrity and address any inconsistencies during migration.
6. **Automation and Scripting:**
   * Scripting languages such as Python can be used to automate the migration process, ensuring efficiency and consistency.

By employing PostgreSQL as the database system and following a well-designed schema, the project ensures a structured and organized approach to data storage. The migration process guarantees seamless transfer of data from HDFS to the relational database, setting the stage for efficient querying and analysis.

**Conclusion:**

In the culmination of this project, we reflect on the key achievements, lessons learned, and potential avenues for future improvements.

**Summary of Key Achievements:**

1. **Successful Cluster Setup:**
   * The establishment of a Hadoop cluster provided a robust foundation for distributed storage and processing.
2. **Web Crawling and Data Extraction:**
   * Effective web crawling techniques were employed to extract valuable cricket-related data from the Cricbuzz website.
3. **GitHub Repository Setup:**
   * A well-organized GitHub repository facilitated collaborative development, version control, and documentation.
4. **Database Setup and Schema Design:**
   * The choice of PostgreSQL as the relational database management system allowed for efficient storage and retrieval of structured cricket data.
5. **Data Migration from HDFS to PostgreSQL:**
   * A seamless transfer of data from the Hadoop Distributed File System (HDFS) to PostgreSQL ensured centralized and organized storage.
6. **Query Execution and Optimization:**
   * Successful execution and optimization of queries enhanced the performance of data retrieval from the PostgreSQL database.

**Lessons Learned:**

1. **Cluster Stability:**
   * Addressing challenges related to cluster stability and worker node failures increased our understanding of distributed systems.
2. **Web Scraping Challenges:**
   * Overcoming obstacles in web scraping highlighted the importance of adapting to changes in website structures and handling dynamic content.
3. **Database Management Best Practices:**
   * Learning and implementing best practices in database management, including indexing and optimization, were pivotal for ensuring efficient query execution.
4. **Collaborative Development:**
   * The collaborative development environment on GitHub reinforced the significance of version control, code reviews, and collaborative documentation.

**Future Improvements:**

1. **Enhanced Web Scraping Robustness:**
   * Implementing more robust web scraping techniques to handle dynamic website changes and ensure continuous data extraction.
2. **Advanced Database Features:**
   * Exploring and implementing advanced database features such as stored procedures, triggers, and materialized views for optimized data handling.
3. **Integration with Real-time Data:**
   * Integrating real-time data feeds to keep the cricket-related statistics up-to-date and reflective of the latest matches and player performances.
4. **User Interface Development:**
   * Building a user-friendly interface for accessing and visualizing the cricket data, providing an enhanced user experience.
5. **Machine Learning Integration:**
   * Exploring the integration of machine learning models for predictive analytics and insights into future cricket match outcomes.

In conclusion, this project has not only met its objectives but has also provided valuable insights and learning experiences. The successful execution of web scraping, database management, and query optimization demonstrates the effectiveness of the chosen technologies and methodologies. As we reflect on the lessons learned and outline future improvements, the project lays the groundwork for continued exploration and innovation in the realm of cricket data analytics.

**Team Collaboration and Management:**

This project was a collaborative effort involving a team of four individuals, and the successful completion of the tasks was greatly influenced by effective teamwork and task management.

**Division of Tasks:**

* The team started by dividing the project tasks based on individual strengths, skills, and interests. This strategic division ensured that each team member could contribute their expertise to the project.

**Task Management:**

* Efficient task management was maintained through regular team meetings, where progress, challenges, and potential roadblocks were discussed. This proactive approach allowed for timely problem-solving and adjustments to the project plan.

**Effective Communication:**

* Clear and open communication was key to the success of the team. Regular updates were shared through various communication channels, fostering an environment where everyone felt informed and connected.

**Learning and Skill Development:**

* The collaborative nature of the project provided an opportunity for team members to learn from each other. Sharing knowledge and skills contributed to the overall growth and development of each team member.

**Strong Group Dynamics:**

* The team dynamic was characterized by mutual respect, trust, and a shared commitment to project success. This positive atmosphere facilitated creativity, idea sharing, and a sense of collective ownership.

**Lessons in Teamwork:**

* Managing a project as a team of four allowed us to experience firsthand the challenges and rewards of effective teamwork. Learning to navigate differences in working styles and leveraging individual strengths enriched the overall team experience.

**Team Unity and Success:**

* The excellent group synergy resulted in the successful completion of project milestones. As a team, we celebrated achievements, learned from setbacks, and collectively contributed to the overall success of the project.

In summary, the team's division of tasks, efficient communication, and positive group dynamics were instrumental in overcoming challenges and achieving project objectives. The experience gained in teamwork and collaboration will undoubtedly contribute to the future success of each team member in their respective endeavors.

**Documentation and Videos:**

Creation of Hadoop clusters and data– Tejas

Webscrapping – Mercy

Data injestion to HDFS and Database and creation of tables - Hanif

Database queries – Muskaan

Failure Test – Team work