# Project Report: Big Data Analytics with Hive and Spark

## Introduction

The goal of the project, "Big Data Analytics with Hive and Spark," is to show how we can use sophisticated big data tools and methods for in-depth research. The primary goal of this project is to process, analyse, and visualise large amounts of data using Apache Hive and Apache Spark.

## Background

Organisations face a distinct mix of opportunities and problems because of the constantly expanding volumes of data. Big data analytics is an effective method for drawing insightful conclusions from massive datasets, which promotes well-informed decision-making. Two essential parts of the big data ecosystem are Apache Hive and Apache Spark, which provide scalable and effective methods for cleaning, processing, and analysing data.

## Data Acquisition and Cleaning

***Dataset Description*:**

For this project, a sizable dataset comprising user interactions with anime content is employed. The dataset contains essential fields, including user\_id, anime\_id, and rating. Below are the top 10 rows of the dataset from hive and spark.

user\_id|anime\_id|rating|

+-------+--------+------+

| 0| 67| 9|

| 0| 6702| 7|

| 0| 242| 10|

| 0| 4898| 0|

| 0| 21| 10|

| 0| 24| 9 |

| 0| 2104| 0 |

| 0| 4722| 8 |

| 0| 6098| 6 |

| 0| 3125| 9 |

A screen shot of a computer

Description automatically generated

## Methodology

***Data Cleaning Using hive***

1. Loading the Data into Hive:

Loading the raw data into Hive tables. This can be done using Hive's LOAD DATA command or external tables pointing to the data location

1. Understanding the Data:

To find problems like missing numbers, outliers, or discrepancies, examine the composition and content of the data. To get summary statistics and investigate the distribution of the data, use Hive queries.

1. Handling Null Values:

Determine which columns have Null values and choose the best course of action for them.

If necessary, use more complex imputation techniques in addition to using default values, mean, and median when utilising Hive queries to substitute null data.

***Challenges and Solutions:***

Challenge 1: Large Datasets

Solution: To effectively manage huge datasets, split data, optimise searches, and take advantage of Hive's parallel processing features.

Challenge 2: Complex Transformations

Solution: Divide difficult cleaning jobs into smaller, more doable phases. If required, make use of custom UDFs (User-Defined Functions) and Hive functions.

Challenge 3: Performance Issues

Solution: To increase performance, fine-tune Hive setups, optimise queries, and take partitioning and indexing into consideration.

Challenge 4: Data Skewness

Solution: Use bucketing, optimise join techniques, and modify data distribution to address data skew.

Challenge 5: Limited Support for Advanced Analytics

Solution: To tackle intricate machine learning or analytics tasks, think about including Apache Spark or exporting data in a format that can be processed further.

## Data Processing and Analysis Using Spark

1. Initialization and Data Loading:

* Started a Spark session and filled a Spark DataFrame with the cleaned data from a CSV file kept in Google Cloud Storage.

1. Exploratory Data Analysis (EDA):

* To perform exploratory data analysis, Spark was used. To comprehend the distribution and features of the data, descriptive statistics were presented, such as the first ten rows and summary statistics.

1. Calculating Average Rating per User:

* The average rating per user was computed using Spark's groupBy and aggregate features. This kind of study offers perceptions into the preferences and actions of users.

1. Categorizing Ratings:

* A new column was created by applying Spark Data Frame transformations to group ratings into three categories: High, Medium, and Low. This stage gives the analysis further level of detail.

1. Further Data Analysis:

* A more thorough data analysis was carried out, which involved looking at the rating distribution, figuring out the average rating for each anime genre, and determining which anime was the highest rated.

1. Visualization:

* Matplotlib and Seaborn were used to convert Spark Data Frames to Pandas Data Frames for visualisation. Histograms and bar plots are examples of visualisations that were made to convey and illustrate the insights gleaned from the data.

**Distribution of Top 10 Ratings**

A graph with a bar

Description automatically generated

## 

**Top 10 Categories By Average rating**A graph of different colored rectangular shapes

Description automatically generated with medium confidence

## Advanced Analytics Techniques:

1. User-Based Average Ratings:
2. Categorization of Ratings:
3. Exploratory Data Analysis (EDA):
4. Top-Rated Anime Identification:
5. Average Rating per Anime Category:

## Key Insights:

1. User Preferences and Behaviour:

* Understanding the preferences and actions of various user segments is possible through the examination of user-based average ratings. Marketing tactics and content recommendation systems can benefit from an understanding of how various consumers evaluate information.

1. Distribution of Ratings:

* The way the ratings are distributed among the three categories (High, Medium, and Low) indicates how people feel about anime content in general. Content producers and platform managers can use this information to determine how well their products are received overall.

1. Top-Rated Anime:

* Finding the highest rated anime series aids in identifying material that has a powerful emotional connection with viewers. To improve the platform's content portfolio, this information is essential for content generation, acquisition, and promotion.

1. Genre-Specific Performance:

* The performance of various genres is revealed by the average ratings for each anime category. Making strategic decisions about content acquisition and platform positioning can be aided by having a clear understanding of which genres frequently garner strong ratings.

1. User Engagement Patterns:

* Marketing plans, user retention campaigns, and the creation of tailored content suggestions can all benefit from patterns in user engagement, such as the distribution of ratings and the content that users find most engaging.

## Conclusion

Hive's data cleaning procedure guaranteed data consistency and quality, laying a solid basis for further research. Spark's distributed computing capabilities made it possible to process data efficiently and use advanced analytics methods. Data professionals must collaborate, be transparent, and be able to repeat results by documenting their questions, problems, and answers.

Hive and Spark worked together to create a potent data processing pipeline that addressed the difficulties of cleaning, converting, and analysing big datasets. Spark's scalability and flexibility were essential for effectively managing complicated data chores in a distributed computing setting. In the field of data analytics, this defined procedure is an invaluable tool for upholding transparency, replicating findings, and encouraging teamwork.