CREATE DATABASE ProjTweet CHARACTER SET utf8mb4 COLLATE utf8mb4\_unicode\_ci;

>>> The utf8mb4 character set is recommended for full Unicode support, which can be particularly useful if you expect to store multilingual data such as tweets.

Prepare Your Data:

Convert your CSV file into a JSON format, which is more suitable for MongoDB. This can be done using a simple script in Python.

**Big Data Processing and Storage**

1. **Environment Setup**:
   * Install and configure Hadoop and Spark.
   * Set up MongoDB, Hive, and MySQL for data storage and retrieval.
2. **Data Ingestion and Storage**:
   * Import the CSV data into Hive and MongoDB for processing and analysis.
   * Utilize Hadoop's ecosystem for data movement (e.g., using Sqoop to import data into Hive or MySQL).
3. **Data Processing with Spark**:
   * Use Spark for initial data cleaning and transformations.
   * Extract features necessary for sentiment analysis and time series forecasting.
4. **Database Performance Testing**:
   * Conduct a comparative performance analysis using MySQL and MongoDB.
   * Utilize YCSB (Yahoo! Cloud Serving Benchmark) to benchmark database performance.
5. **Architecture Design**:
   * Design and illustrate a comprehensive architecture that includes Hadoop, Spark, MongoDB, Hive, and MySQL.
   * Use diagramming tools to present your design clearly, showing the flow of data through these components.

**Advanced Data Analytics and Time Series Forecasting**

1. **Sentiment Analysis**:
   * Explain and justify the choice of sentiment analysis techniques (e.g., machine learning models vs. lexicon-based approaches).
   * Implement sentiment extraction using Spark or Python libraries.
2. **Time Series Forecasting**:
   * Explore and justify the selection of at least two forecasting methods: one based on neural networks (e.g., LSTM) and one autoregressive model (e.g., SARIMA).
   * Handle the short time series data by techniques such as data augmentation or using external datasets to increase the robustness of the model.
3. **Hyperparameter Tuning**:
   * Discuss the techniques used for tuning the parameters of your machine learning models, including grid search, random search, or Bayesian optimization methods.
4. **Dynamic Dashboard Creation**:
   * Develop a dynamic and interactive dashboard using tools like Tableau, Power BI, or a web-based framework (e.g., Dash by Plotly).
   * Discuss the application of Tufts principles in your dashboard design.

**Report Writing**

1. **Comprehensive Discussion**:
   * Write about the data storage, processing activities, and the use of big data technologies.
   * Discuss the rationale behind choosing specific databases and technologies.
   * Evaluate and justify your choices in EDA, data wrangling, and the machine learning models implemented.
2. **Presentation of Results**:
   * Ensure your results are well-presented in the form of figures, tables, and your interactive dashboard.
   * The report should follow the Harvard citation style and adhere to the specified word count.
3. **Forecasting Results**:
   * Present the sentiment forecasts for 1 day, 3 days, and 7 days, incorporating visual and tabular representations to support your analysis.