Double-cliquez (ou appuyez sur Entrée) pour modifier

Project Roadmap: Distributed Spotify Sound Analysis Computation

Phase 1: Setup Distributed Computing Environment

1. **Setup Hadoop Distributed File System (HDFS)**:
   - Installation and configuration of Hadoop for distributed storage.
   - **Estimated Time**: 8 hours
2. **Setup HBase**:
   - Installation of HBase and its integration with HDFS.
   - **Estimated Time**: 8 hours
3. **Setup Spark**:
   - Installation and configuration of Spark for distributed processing.
   - **Estimated Time**: 6 hours

Phase 2: Data Ingestion and Pre-processing

1. **Load .pickle Files into HDFS with MapReduce**:
   - Transfer the 250,000 .pickle files into HDFS using Hadoop.
   - **Estimated Time**: 12 hours
2. **Design HBase Schema**:
   - Planning of tables, column families, and row keys optimized for your access patterns.
   - **Estimated Time**: 7 hours
3. **Data Cleaning and Transformation using MapReduce**:
   - Conversion of .pickle files for HBase and Spark compatibility.
   - Removal of noise and irrelevant data elements.
   - **Estimated Time**: 15 hours

Phase 3: Data Analysis from Processed Images

1. **Visualization of Sound Analysis with Spark**:
   - Creation of images from the sound analysis data.
   - **Estimated Time**: 15 hours
2. **Genre Classification from Images**:
   - **With MapReduce**: Convert images into a matrix representation, making the pixels' intensity/values manageable.
   - **With Spark**: Use clustering methods like K-means in Spark's MLlib on the matrix data to classify genres.
   - **Estimated Time**: 20 hours
3. **Mood and Temporal Analysis from Images**:
   - **With MapReduce**: Quantify color patterns and distributions in images as these may correlate with mood and temporal elements.
   - **With Spark**: Implement classification models in Spark's MLlib to identify moods from patterns and sequence analyses for temporal patterns.
   - **Estimated Time**: 25 hours

Phase 4: Optimization, Refinement, and Evaluation

1. **Enhancement of MapReduce and Spark Tasks**:
   - Profiling and optimization of both MapReduce and Spark operations for better efficiency.
   - Debugging and fixing any bottlenecks or issues in data processing and analysis steps.
2. **Analysis-driven Refinement**:
   - Review preliminary results of the genre, mood, and temporal analyses.
   - Make refinements based on discrepancies or areas of improvement identified.
3. **Evaluation and Validation**:
   - Use a subset of the dataset as a validation set.
   - Evaluate the accuracy and reliability of the genre, mood, and temporal classifications.
   - Fine-tune the models or methods based on the evaluation results.