Project Proposal Report

Title: Testing the functionality of Parquet File Manager Utility and resolving its unaddressed issues.

Topic Area: In-Memory & File Formats

GitHub Repository used: https://github.com/deepaknairrpf/file-manager

Project Summary:

The current file manager utility present in the GitHub repo is a tool for reading and writing a variety of file formats.
The goal of this project is to add more file operation features and Spark integration to the file manager utility package. The existing package over the Github repo supports reading and writing compressed and uncompressed Parquet , JSON, CSV, XML files in a memory-efficient way.

Issues in the existing Code:

1. An exception in the form of StopIteration is being raised during the operation on a JSON or Parquet file, indicating an interruption or unexpected termination in the file processing.
2. The attempt to read a Parquet file has encountered a failure attributed to an obscure encoding issue, hindering the successful retrieval of data from the specified file.
3. An OSError has been triggered while attempting to read a gzip file, pointing to an invalid argument within the file path. This error stems from problematic characters or inaccuracies in the provided file path, impeding the smooth execution of the file reading operation.

Proposed solution and new features:

1. Implement a corrective measure to address the StopIteration Exception occurring during JSON or Parquet file operations, ensuring uninterrupted and smooth processing.
2. Resolve the Parquet file reading failure caused by an unidentified encoding issue by implementing a solution that accurately handles the encoding intricacies, ensuring successful data retrieval.
3. Rectify the OSError raised during the gzip file read operation by addressing the invalid argument in the file path. Implement a fix to handle problematic

characters or inaccuracies within the file path, enabling error-free reading of gzip files.

4. Deploying the utility on Apache Spark: Creating this parquet project on Apache Spark for in-memory data processing, minimizing operation execution time, and enhancing the efficiency of file operations.

Technologies:

Python and Apache Spark

Future possible areas of enhancement:

- Adding support for additional file formats
- Adding support for additional operations, such as filtering and sorting
- Developing a GUI for the file manager utility package
- Integrating the file manager utility package with other big data processing frameworks.