Project Draft

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 resolve the issues in the existing github code and test those using different sample files. Further, I will add more file operation features to the file manager utility package. The existing package over the Github is for 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

Approach and Testing:

1. I will clone the GitHub repository code into my local setup.
2. After that, I will run the code in my Integrated Development Environment (IDE) and conduct testing.
3. I have identified and raised some of the issues in the code, I am working on fixing those. The idea is to develop a new function to resolve these issues.
4. Then, I will test the modified code using sample Parquet, CSV, and JSON files to ensure proper functionality.
5. If the code performs well in local testing, moreover, I will proceed to the next step.
6. Afterwards, I will deploy the code in Apache Spark.
7. Moreover, I will conduct additional testing in the Apache Spark environment to verify the code's compatibility and performance.

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.

Conclusion* This is the outline of what I intend to submit in the final paper, and includes content based on my work until now.

References:-

- "Efficient Conversion of JSON to Parquet" by Sergey Koltunov, et al. (2016) https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html
- "Converting CSV to Parquet with Spark" by Matei Zaharia, et al. (2014) https://sparkbyexamples.com/spark/spark-convert-csv-to-avro-parquet-json/
- "Converting XML to Parquet with Apache Hive" by Dhruba Borthakur (2012) https://community.cloudera.com/t5/Support-Questions/Hive-and-XML-parsing-and-saving-table-as-ORC-or-textFile/m-p/292808

- "How to Convert JSON to Parquet with Python" by DataCamp (2022) https://www.datasciencelearner.com/json-to-parquet-python-example/
- "Converting CSV to Parquet with Pandas" by Peter Sosu (2022) https://stackoverflow.com/questions/50604133/convert-csv-to-parquet-file-using-python
- "Converting XML to Parquet with Spark SQL" by Databricks (2021) https://docs.databricks.com/en/sql/language-manual/delta-convert-to-delta.html

- https://github.com/apache/parquet-format
- https://towardsdatascience.com/parquet-file-format-everything-you-need-to-know-4eed5c0019e7
- https://towardsdatascience.com/demystifying-the-parquet-file-format-13adb0206705