A Review on Text Data Format for Data Science

STEVEN SHIDI ZHOU EMIL STÅHL
shidi | emilstah @kth.se

September 9, 2022

1 Allocation of responsibilities

Steven Shidi Zhou is responsible for writing the benchmark programs and systematically testing the data set formats in question, writing the Introduction, Background, and Results Analysis Section in the report, and presenting the project plan. Additionally, Emil Ståhl is responsible for collecting data using the benchmark programs produced by Steven Shidi Zhou, conducting a qualitative assessment, writing the Results and Conclusion sections in the report, and presenting the ethical and sustainability aspects of the project. Both authors are responsible for presenting the final project and the final version of the report.

2 Organization

This project is organized as a two-person project and build on the basis of some previous works, such as [1] and [2]. Qualitative and quantitative assessments will run simultaneously.

3 Background

The modern society produces a vast amount of data every day across many different types of industries and pipelines. Until processed, this data is merely bits placed on a storage medium. The field of data science is focusing on extracting useful knowledge from such data with the goal of gaining a deeper understanding of a given field. Data can be stored in various different formats such as csv, xlsx, Parquet, and Avro. Processing this data requires a lot of operations to be applied such as read, write, and other mathematical operations. Given the value that data science brings puts heavy requirements on the performance and stability of the data formats used in the processing and storage pipelines. Related work has been made by Plase et al. which in their work ”A comparison of HDFS compact data formats: Avro versus Parquet” have benchmarked the amount of disk space these formats consumes as well as their performance in High Energy Physics (HEP) analysis which uses lots of numerical data. However, the work does not benchmark read/write and simple operations, nor does it benchmark the performance on text data. Understanding the performance advantages of different data formats in terms of read/write speed on text data is therefore of paramount importance in the data science community[1].

4 Problem statement

Analysing the performance and stability of different data formats is key to tackle the scalability and performance challenges in which data science pipelines operates. Today, there exists over 44 zettabytes of data with around 2.5 quintillion bytes worth of data generated each day. (source https://explodingtopics.com/blog/big-data-stats). Data science pipelines must be highly optimized performance-wise in order to be able to process data at sufficient speeds and to ensure stability and job efficiency. Since the different data formats are implemented using different technologies, they may differ in how they handle different sizes of data which affects the performance when processing and storing the data. There are currently no conventions for when to use a particular data format for a given data size to optimize for performance and file stability. Since csv, xlsx, Parquet, and Avro are all widely used in data science it is important to have an understanding of their behavior.

5 Purpose

The purpose of this work is to analyze the performance of the csv, xlsx, Parquet, and Avro data formats by quantifying their read/write speed, file stability, and other operations such as summation. By analysing the performance of these data formats in different pipeline conditions, such as file size, one can use this work for tailoring data science pipelines in order to optimize for performance. The results are potentially of interest for general data scientists, developers of big data frameworks, and the open source community.

6 Research question and contributions to the state of the art

Considering the background and the problem discussion, the paper answers the following research question:

RQ1 How does the csv, xlsx, Parquet, and Avro data formats differ in read/write performance on as well as resilience against file errors when storing text data?

The main contributions of our work are:

- 1
- 2

7 Hypothesis

There is an advantageous data format for a given file size, a row-based Avro format should provide advantages performance with regard to simple operations such as read/write speeds when compared to csv, xlsx, and Parquet. For larger data sets, the Apache Parquet format should provide better performance.

8 Goal

The goal of this work is to run a set of experiments where data of various sizes stored in the csv, xlsx, Parquet, and Avro formats is processed and exposed to random bit flips. The experiments are going to measure metrics such as read/write speed and file stability with the goal of producing an analytical model showing when a particular data format may be advantageous. This model can be used to invalidate the hypothesis. The results are then going to be described, analyzed, and discussed in order to determine if there are any bottlenecks regarding the different data formats ability to ensure performance and resilience against file errors.

9 Tasks

A collection of different data with various sizes will be converted into different data formats such as csv, xlsx, Parquet, and Avro. Our test will be measuring the time it takes to read/write such data, as well as some basic transform operations, with each of the data formats. To measure how well the different data formats handle error, a file stability experiment will be conducted by randomly flipping file bits to simulate a corrupt data file. [2]. Finally, qualitative methods will be prepared to gain more insight into how popular or user friendly each format is.

10 Method

The project will use the empirical method [2] for the data format stability experiment, as well as the general performance test. We, the authors, will also use an analytical method (e.g., interview or survey) to try to get a full image; however, the analytic part of this study will be shortened due to limited time period and resource of the assignment.

11 Milestone chart (time schedule)

The project will start on 7 September and end at 16:59 on 28 October. There will be the following milestones and deliverables:

8 September Presentation of the proposed research: Ethics & Sustainability.

14 September Research plan: First draft of the research plan, presentation, and peer review.

21 September Completion of benchmark programs.

28 September Quantitative analysis done.

5 October Result collection and analysis done (Qualitative interview).

10 October Final report: First draft and presentation with peer review of the draft report and presentation.

14 October Written opposition: before the final seminar - with peer review.

28 October Final report submission (the report will have been written in parallel with each of the above steps).

References

[1] D. Plase, L. Niedrite, and R. Taranovs, “A Comparison of HDFS Compact Data Formats: Avro Versus Parquet,” *Science future of Lithuania*, vol. 9, no. 3, pp. 267–276, 2017. doi: 10.3846/mla.2017.1033 Place: Vilnius Publisher: Gediminas Technical University.

[2] J. Blomer, “A quantitative review of data formats for HEP analyses,” *Journal of Physics: Conference Series*, vol. 1085, p. 032020, Sep. 2018. doi: 10.1088/1742-6596/1085/3/032020 Publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1742-6596/1085/3/032020

12 Acronyms

HEP High Energy Physics