**UNIVERSITY OF WATERLOO**

Faculty of Mathematics

***A COMPARISON ON DATA FILE FORMATS***

Royal Bank of Canada

Credit Modelling and Methodologies

Toronto, Ontario, Canada

Tan Nguyen

20780276

1C Computer Science

June 2019

1234 Fifth Avenue (your home address)

Anywhere, ON

N0T 4U2

June 24, 2019

John Smith

Supervisor, IT Department

Quick Solutions Inc.

8910 Main Street, Anytown, ON, P0B 2S0

Dear Mr. Smith,

I have prepared the enclosed report, “A comparison on data file formats”, as my 1C work report for the Credit Model and Methodology team at Royal Bank of Canada. This report, the first of four work reports that the Co-operative Education Program requires that I successfully complete as part of my Computer Science Co-op degree requirements, has not received academic credit.

As the Credit Model and Methodology team performs both practical risk modelling and data science research, my position as an Analyst intern required that I do experimentations and build products that help speed up the modelling process. This report is an in-depth study of one particularly interesting topic in my research, the advantages and disadvantages of different data storage files.

The Faculty of Mathematics requests that you evaluate this report for command of topic and technical content/analysis. Following your assessment, the report, together with your evaluation, will be submitted to the Math Undergrad Office for evaluation on campus by qualified work report markers. The combined marks determine whether the report will receive credit and whether it will be considered for an award.

I would like to thank you for your assistance in preparing this report.

Sincerely,

Tan

Encl.

**TABLE OF CONTENTS**

EXECUTIVE SUMMARY

1. INTRODUCTION
2. ANALYSIS
   1. HUMAN-READABLE DATA FILES
      1. XML (eXtensible Markup Language)
      2. JSON (JavaScript Object Notation)
      3. CSV (Comma-separated values)
   2. MACHINE-READABLE DATA FILES
      1. ORC (Optimized Row Columnar)
      2. Parquet
      3. Avro
   3. CASE STUDIES
      1. CSV current popularity in Data Science
      2. The necessary shift to Big Data file formats
3. CONCLUSIONS

REFERENCES

EXECUTIVE SUMMARY

The CMM (Credit Model and Methodology) team of RBC (Royal Bank of Canada) is conducting research on different type of data file storages. The purpose of the analysis is to determine the most efficient way to store and process the increasing amount of data on the company’s servers. This initiation is expected to be useful for other teams within GRM (Group Risk Management) as well as any team within the company who performs analysis on large databases.

The two main type of file storage being considered are

* Human-readable files, which store data raw and can be viewed easily through text editors like Notepad and Word, or tabular data readers such as Excel
* Machine-readable files, most of which are compressed and encoded into binary data for efficiency and requires more specialized software to process them

N

Purpose of the report (done)

Research methods and points of comparison

Conclusion

Recommendation

Why ….

Co-operative Education Program requires that I successfully complete as part of my BMath Co-op degree requirements, has not received academic credit

<https://uwaterloo.ca/math/current-undergraduates/co-op-information/work-report-guidelines/40-writing-tips>

<https://uwaterloo.ca/math/current-undergraduates/co-op-information/work-report-guidelines/appendix-3-work-report-checklist>

<https://uwaterloo.ca/math/sites/ca.math/files/uploads/files/outline_0.pdf>

<https://www.nexla.com/resource/introduction-big-data-formats-understanding-avro-parquet-orc/>

1. INTRODUCTION