Benchmarking Distributed Stream Processing Frameworks for Classical Machine Learning Applications

A THESIS

submitted by

Merlin Sundar
(S15011)

Guided by

Dr. Timothy A. Gonsalves
Dr. Sriram Kailasam

for the award of the degree of

Master of Science(by Research)

School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
India - 175005

April 2021

"But as for you, be strong and do not give up, for your work will be rewarded."

– 2 Chronicles 15:7

I dedicate this thesis:

To God Almighty,

To my Guides,

To my Family,

And To my Friends.

Declaration

I hereby declare that the work incorporated in this thesis is the outcome of the studies accomplished by me in the **School of Computing and Electrical Engineering, Indian Institute of Technology Mandi**, under the supervision of **Dr. Timothy A. Gonsalves** and **Dr. Sriram Kailasam**. This work has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on the finding of other investigators. In addition, I certify that no part of this work will, in future, be used for submission in my name, for the award of any other degree at any university.

Place: Mandi **Merlin Sundar**

Date:

Thesis Certificate

This is to certify that the thesis titled **"Benchmarking Distributed Stream Processing Frameworks for Classical Machine Learning Applications"**, submitted by **Merlin Sundar**, at the Indian Institute of Technology Mandi for the award of Master of Science (by research) is a bonafide record of the research work carried out by her under our supervision. To the best of our knowledge, the content of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Dr. Timothy A. Gonsalves
(Thesis Supervisor)

Dr. Sriram Kailasam
(Thesis Supervisor)

Acknowledgement

"It is though Him, that I receive my strength". Without God, my MS would not have reached to completion. Hence, I thank our Lord, Father in Heaven for blessing me abundantly with this lifetime opportunity.

Next, I would like to profusely thank both my guides Prof. Timothy A. Gonsalves and Dr. Sriram Kailasam for showing faith in me and making this work possible. Your ideas, discussions with me, expertise in the domain and perennial support helped me achieve the goal. Thank you for being so patient with me and helping me grow, inside and outside of work. You two's words of wisdom will always remain with me in every step of my life.

I am also very thankful to my course work teachers – Dr. Dileep AD, Prof. Yoder, Prof. Markus, Dr. Shayamashree, Dr. Dericks and Dr. Timothy A Gonsalves, with whom I got a chance to learn new and interesting concepts. I also extend my thanks to all my APC members – Dr. Astrid Kiehn, Dr. Anil Sao (earlier SCEE chairman), Dr. Bharat (also earlier SCEE chairman) and Dr. Samar Agnihotri whose timely feedback and guidance pushed me to achieve my goals better.

In addition to all the faculty, I would like to thank the following members of Manas Lab – Sujeet, Ganesan, Shaifu, Debashis, Naveksha, Jyoti Nigam, Prabhjot who extended their support whenever I needed it. I am very glad to have worked with each one of you.

During my time here in IIT Mandi, I have made many a friends, who turned more into my family away from home without whom this journey would have been monotonous and boring. Huge thanks to Sandhya, Jyoti and Snehal family, my church family and Shruti Kaushik. You guys made my life memorable here. Also thanks to friends supporting me from distance, you guys cheered for me every time -

Odlin Mendez, Palvi aggarwal, Surabhi Soni, Himani Kakkar, Reena, Chandan and Himanshu.

I will be forever grateful to my mom and dad, your prayers and encouragement made this possible. Thank you for constantly nudging me to achieve more. And to my family in Palakkad, mother-in-law, father-in-law, Shreya, Pramod, and Shwetha, thank you for all the love, support and care you showered on me.

A special thanks to the Medical Unit, Dr. Chander Singh and various nurses who helped me cope better with sickness. I also want to thank the mess workers, who cooked food for us during times, when cooking was a task at home.

I am very grateful to NMSWorks Pvt.Ltd. who jointly funded our project with MHRD. Especially Mrs. Usha whose guided motivation and direction helped in shaping my thesis and also for all the resources provided to me for my experiments.

Lastly, I would like to thank my partner Pavin S Samuel. You stood by me like a rock, a sponge, a punchbag, whatever I required you to be and I am thankful to my pet pup Noah, who steered me through some of the darkest nights. Much appreciation and love to you both.

– Merlin Sundar

ABSTRACT

In India, the large telecom service providers each serve 100 million - 400 million subscribers. Where, each telecommunications network may contain hundreds or more different types of network devices transmitting data among each other and to the customer subscriber. In this scenario, a Network Management System (NMS) may collect millions of records/sec of data. There can be lots of network faults that keep happening in real-time. Some of these may be of low priority while others may be of high priority. Hence, it becomes imperative to analyze the data in real-time to manage the high priority network faults in real-time too. In view of the growing complexity and rapid changes in the demands on the network, machine learning (ML) techniques are being used for advanced NMS. ML models are typically computationally intensive, involving training and testing phases. To handle the huge volume of data streaming at high velocity, we not only require powerful machines but also mechanisms to distribute the computation involved across multiple nodes. There are several open-source distributed stream processing frameworks such as Apache Storm, Apache Flink, Apache Spark and Confluent Kafka for building real-time machine learning applications. Prior works benchmarked some of these platforms using low-level operations like filters, joins, windowed computations etc.

In this thesis, we first survey multiple Distributed Stream Processing Frameworks qualitatively for choosing appropriate frameworks and also Message Queuing Application for ordered message delivery. Once the platforms are decided, we benchmark our four chosen DSPFs for their applicability to execute classical machine learning models. For variety in complexity of computation, we have chosen three classical machine leaning models - Online K-Means, Online Linear Regression and Online Logistic Regression. We study the following quantitative metrics of evaluation: throughput, latency, CPU utilization, memory usage and Input/Output usage. The experiments were conducted in both standalone and clusters setups to determine the scalability of the models. In this study, we found that all four frameworks are comparable, except Apache Spark performs marginally better than the others for standalone setup for all algorithms. Whereas, for cluster node setup the best performing framework varies between Apache Storm and Apache Spark. We have also

observed speedup across different setups. These results can help system designers choose the right model and the right framework, given a specific configuration of streaming data.

Later in the thesis, we also discuss a direct application based on the benchmarking experiment, called the Configuration Planner. This Configuration Planner is designed to make recommendations to a telecom network administrator for a server configuration based on the network size to be managed. We describe in detail the parameters involved, design overview and also the data structures of each component of this planner. This thesis covers the design aspect of the planner and also a possible user interface of the Planner.

Contents