**Paper Title: Exploring Apache Airflow for Enhanced Workflow Management and Error Handling in Scheduling Apache Spark Jobs: A Case Study in Stock Prediction**

Course: MET CS777 A1

Assignment: Term Paper Proposal

Team Members: Chuong Nguyen – Zhen Cao

**Abstract:** This paper investigates the utility of Apache Airflow for scheduling Apache Spark jobs, emphasizing workflow optimization and error management. Focusing on stock prediction as a case study, we explore Airflow's Directed Acyclic Graphs (DAGs) and demonstrate their effectiveness in automating task scheduling and error handling. By integrating data collection, preprocessing, and machine learning model deployment, we showcase how Airflow streamlines workflow management, facilitating more informed investment decisions. Through this research, we highlight the potential of Airflow in enhancing efficiency and reliability in big data analytics workflows.

Table of Contents

[I. Introduction 1](#_Toc162472508)

[A. Background 1](#_Toc162472509)

[B. Motivation 1](#_Toc162472510)

[C. Objectives 2](#_Toc162472511)

[II. Apache Airflow 2](#_Toc162472512)

[A. Overview of Workflow Management Systems 2](#_Toc162472513)

[B. Apache Airflow 2](#_Toc162472514)

[C. Limitations 5](#_Toc162472515)

[D. Conclusion 5](#_Toc162472516)

[III. Demo: Stock Prediction 5](#_Toc162472517)

[A. Overview of the demo 5](#_Toc162472518)

[B. Dataset 6](#_Toc162472519)

[C. Environment Setups 6](#_Toc162472520)

[D. Apache Airflow DAG 12](#_Toc162472521)

[E. Running DAG in Airflow UI 13](#_Toc162472522)

[F. Results 14](#_Toc162472523)

[IV. Conclusion 15](#_Toc162472524)

[V. References 17](#_Toc162472525)