IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, July 2023

Managing Data Pipeline with Apache Airflow

Mohit Nara[1], Aquila Shaikh[2], Rashmita Pradhan[3]
Student, Master of Computer Application[1]
Assistant Professor, Master of Computer Application[2,3]
Late Bhausaheb Hiray S. S. Trust's Hiray Institute of Computer Application, Mumbai, India

Abstract: *Data orchestration is the process of automating the movement and transformation of data between different systems. It is a key part of any data-driven organization, as it allows businesses to efficiently collect, store, and analyze data from a variety of sources. Nowadays, many applications that run on cluster and cloud resources are workflows. A workflow is represented as a Directed Acyclic Graph (DAG) where each vertex represents a task (i.e., a unit of work) and an edge a computation/data constraint. Apache Airflow has emerged as a powerful open-source tool for data orchestration, offering a scalable and efficient solution for managing complex data workflows. The paper investigates the benefits of using Apache Airflow in terms of workflow management, task scheduling, and monitoring of data processing tasks. Approximately 45% of users are data engineers, 30% are data scientists, and 25% are data analysts who uses the airflow. Also the most common use cases for Apache Airflow are: Scheduling and managing data pipelines (60%), Orchestrating data processing tasks (40%),Monitoring and debugging data pipelines (30%)*

Keywords: ETL, Apache Airflow, Data Orchestration, Data engineering, DAG, data-pipelines

REFERENCES

[1]. M. Beauchemin, (2014) Apache Airflow Project.
[2]. Barika, M., Garg, S., Zomaya, A.Y., Wang, L., Moorsel, A.V., Ranjan, R.: Orchestrating big data analysis workflows in the cloud: research challenges, survey, and future directions. ACM Computing Surveys (CSUR) 52(5), 1–41 (2019)
[3]. F. P. Guimarães and A. C. M. Melo, "User-Defined Adaptive Fault- Tolerant Execution of Workflows in the Grid," in Proceedings of the IEEE CIT, Sep 2011, pp. 356-362.
[4]. 4. L. Li, Z. Miao, L. Yuqing, Q. Liangjuan, "A Survey on Workflow Management and Scheduling in Cloud Computing", Cluster Cloud and Grid Computing (CCGrid) 2014 14th IEEE/ACM InternationalSymposium on, pp. 837-846, 2014
[5]. M. Kotliar et al., "CWL-Airflow: a lightweight pipeline manager supporting common workflow language", bioRxiv, 2018.
[6]. Barker, J. Van Hemert, "Scientific workflow: a survey and research directions", International Conference on Parallel Processing and Applied Mathematics,pp. 746- 753, 2007.
[7]. M. Berger et al., An Evaluation of Workflow Management System, Austria:Institute for Applied Computer Science and Information Systems, University of Vienna, 1997.
[8]. G. Alonso, D. Agrawal, A. El Abbadi, C. Mohan, "Functionality and Limitations of Current Workflow Management Systems", IEEE Expert, vol. 1, no. 9, 1997