Sustainable Big Data Analytics Process Pipeline Using Apache Ecosystem

S

Jane Cheng
UBS, USA

Peng Zhao
https://orcid.org/0000-0003-1458-8266
INTELLIGENTRABBIT LLC, USA

INTRODUCTION

Big data analytics is an automated process which uses a set of techniques or tools to access large-scale data to extract useful information and insight. This process involves a series of customized and proprietary steps. it requires specific knowledge to handle and operate the workflow properly. Due to 4V nature of big data (Volumes, Variety, Velocity and Veracity), it is required to build a robust, reliable and fault-tolerant data processing pipeline . The proposed approach will help application developers to conquer this challenge.

Apache Airflow is a cutting-edge technology for applying big data analytics, which can cooperate the data processing workflows and data warehouses properly. Apache Airflow was developed by Airbnb technical engineers, aiming to manage internal workflows in a productive way. In 2016, Airflow became affiliated by Apache and was made accessible to users as an open source. Airflow is a framework that can conduct the various job of executing, scheduling, distributing, and monitoring. It can handle either interdependent or independent tasks. To operate each job, a directed acyclic graph (DAG) definition file is required. In this definition file, a collection is included for developers to run and sectionalized by relationships and dependencies.

The sustainable automation can consolidate all tasks of ETL, data warehousing, and analytics on one technology platform. The upstream vendor data will be ingested into a data lake, where source data is maintained and gone through the data processing of cleaning, scrubbing, normalization and data insight extraction. In the next step of data mining tasks, data could be processed to perform study analytics for end users. Motivated by the current demand in big data analytics and industrial applications, this chapter is proposed to illustrate and investigate a novel sustainable big data processing pipeline using a variety of big data tools. The proposed data pipeline starts at the standard data processing workflow using Apache Airflow, using GitLab for source code control to facilitate peer code review, and uses CI/CD for continuous integration and deployment. Apache Spark has been used for the data computer process scaling with standardizing data in the data warehousing procedure. With data persistent in HDFS/ADLS, downstream system can choose either data visualization tool or API to access data. The objectives of this chapter are:

- investigating most recent big data tools for constructing the novel data workflow architecture.
- illustrating the major functional components of the proposed system architecture.
- initializing a state-of-the-art data workflow architecture design that can be used in the industrial applications.

DOI: 10.4018/978-1-7998-9220-5.ch073

BACKGROUND

Due to the fast revolutionary of information technologies and systems, avalanche-like growth of data has prompted the emergence of new models and technical methods for distributed data processing, including MapReduce, Dryad, Spark (Khan et al., 2014). For processing large graphs, special purpose systems for distributed computing based on the data-flow approach were introduced (Gonzalez et al., 2014). Some systems focused on batch (offline) data processing, while other systems and services can handle the real-time (online) data processing, such as Storm, Spark Streaming, Kafka Streams, and Apache, which attract more attentions due to the users' demands on its ability of rapid and smart responding to the incoming data (Zaharia et al., 2012). These systems can implement the distributed data processing operations, so that to support large volumes of incoming data and to fulfill high speed of data delivery. For distributed data processing, a crucial feature of existing data-driven software systems is the abstraction of the programmer from the details of the implementation of computations by using ready-made primitives. For example, distributed data-flow-operators use map and reduce. This makes simplification of writing programs possible, which can fit into the proposed model of computations. However, it may be still difficult to implement other classes of applications. MapReduce-based systems may not be an optimal choice for performing iterative algorithms and fully connected applications. Many professional solutions for diverse kinds of applications have been established to figure out the limitations of existing distributed data processing models and technologies (Suleykin & Panfilov, 2019a).

In recent years, the open-source methods have become increasingly popular. Hadoop stack that promoted data processing of MapReduce is one of the most commonly used technologies for big data storage. Hortonworkers Data Platform stack provides 100% open-source global data management and related services, for the customers to manage the full lifecycle of the data. Many large industrial companies widely used this stack for data processing, storage, analysis, and visualization. The technical applications of open-access big data, based on HDP Hadoop ecosystem stack, have discussed and analyzed in current research. HDP Hadoop ecosystem stack can establish data processing workflow with all job dependencies and proceed various jobs from one workflow orchestrator. Based on sample industrial KPIs data, which shows the adaptability of suggested methodology for all the possible real-world data with specific formats, the workflow can be implemented and simulated. A set of interconnected jobs for workflow include Spark jobs, shell jobs and PostgreSQL query commands. Additionally, all the workflow steps can be connected and patterned in one data pipeline.

Most recent studies are concentrated on industrial applications with hybrid approaches using open-source technologies, such as Apache ecosystems and other analytical tools. Suleykin & Panfilov (2019b) introduced a big data processing workflow using Apache Hadoop, Postgre SQL, and Apache Airflow. Such a system architecture can be performed with stages through multiple storage spaces for industrial KPIs of millions of records. A novel big data workflow has been proposed for the scalable execution of data transaction, along with a scalability comparison of the proposed method with that of Argo Workflows (Dessalk et al., 2020). Ramanan et al. (2020) illustrated the features and strengths of a new data workflow framework with real-world deep learning processing using Apache Spark, Beam, Swift/T, and Apache Airflow. Such a system can be applied in terms of ease of authoring, efficiency, scalability, and fault recovery. Similar studies have been represented in the form of discussions of vast applications and system architecture designs using Apache Airflow, ranging from exploration of workflow management (Mitchell et al., 2019), lightweight pipeline decision supporting system (Kotliar et al., 2019), to industrial-level ETL processing with metadata-driven systems (Suleykin & Panfilov, 2020; Panfilov & Suleykin, 2021).