ETL JARAYONLARINI AVTOMATLASHTIRISH: APACHE AIRFLOW

Dilmurodov Shoxjahon Sirojiddin o‘g‘li
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti magistiranti
E-mail: shohjahondilmurodov56@gmail.com

Abstract: *This article discusses the automation of ETL (Extract, Transform, Load) processes using Apache Airflow. The main stages of the ETL process are described, and a detailed example is provided, demonstrating the use of Airflow for effective data pipeline management.*

Key words: *ETL, Apache Airflow, data automation, DAG.*

Аннотация: *В данной статье рассматривается автоматизация ETL (извлечение, преобразование, загрузка) процессов с использованием Apache Airflow. Описаны основные этапы процесса ETL, а также приведён подробный пример использования Airflow для эффективного управления конвейером данных.*

Ключевые слова: *ETL, Apache Airflow, автоматизация данных, DAG.*

Annotatsiya: *Ushbu maqolada Apache Airflow yordamida ETL (Extract, Transform, Load) jarayonlarini avtomatlashtirish ko‘rib chiqilgan. ETL jarayonining asosiy bosqichlari yoritilgan va Airflow yordamida ma’lumotlar oqimini boshqarish bo‘yicha batafsil misol keltirilgan.*

Kalit so‘zlar: *ETL, Apache Airflow, ma’lumotlarni avtomatlashtirish, DAG.*

Introduction (Kirish). Bugungi kunda ma’lumotlar har qanday tashkilot yoki korxonaning strategik qarorlar qabul qilishida muhim o‘rin tutadi. Ularning samarali tahlili orqali bozor tendensiyalarini aniqlash, jarayonlarni optimallashtirish va innovatsion yechimlar ishlab chiqish mumkin. Ammo ma’lumotlar odatda turli manbalarda — veb-saytlar, databaselar, log fayllar yoki API’larda saqlanadi va ular strukturasi, sifati hamda formati jihatidan bir-biridan keskin farq qiladi. Shu sababli, bu ma’lumotlarni yagona tahlil uchun mos holga keltirish murakkab va ko‘p bosqichli jarayonni talab etadi. ETL (Extract, Transform, Load) jarayonlari ana shu ehtiyojni qondirishga xizmat qiladi. Bu jarayonlar orqali ma’lumotlar manbalardan olinadi (Extract), kerakli tarzda qayta ishlanadi (Transform) va ma’lumotlar omboriga yoki tahlil tizimlariga yuklanadi (Load). Katta hajmdagi ma’lumotlar bilan ishlaganda, ushbu jarayonlarni qo‘lda boshqarish samarasiz bo‘lib, inson xatoliklariga sabab bo‘lishi mumkin.

Shu nuqtayi nazardan qaralganda, ETL jarayonlarini avtomatlashtirish zamonaviy ma’lumotlar muhitida zaruratga aylanmoqda. Apache Airflow — bu ma’lumotlar pipeline’larini avtomatik rejalashtirish, bajarish va monitoring qilish imkonini beruvchi kuchli va kengaytiriladigan orkestratsiya platformasidir. Ushbu maqolada Apache Airflow vositasi yordamida ETL jarayonlarini qanday avtomatlashtirish mumkinligi, uning asosiy imkoniyatlari va amaliy qo‘llanilishi tahlil qilinadi.

Methodology (Adabiyotlar tahlili va metodlar). Zamonaviy axborot tizimlarida ma’lumotlarni samarali boshqarish uchun ETL (Extract, Transform, Load) jarayonlari muhim ahamiyatga ega. Ushbu jarayonlar ma’lumotlarni turli manbalardan yig‘ish, ularni standartlashtirish, tozalash va ma’lumotlar omboriga joylashtirishni o‘z ichiga oladi. So‘nggi yillarda ETL jarayonlarini avtomatlashtirishda Apache Airflow ochiq kodli platformasi keng qo‘llanilib, murakkab ish oqimlarini boshqarishda samarali yechim sifatida tan olinmoqda.

Ilmiy manbalarga ko‘ra, Apache Airflow ma’lumotlar oqimlarini rejalashtirish, kuzatish va xatolarni boshqarishda muhim vosita sifatida ajralib turadi. Platforma Python tilida ishlab chiqilgan bo‘lib, ish jarayonlarini Directed Acyclic Graph (DAG) shaklida tashkil qiladi. Airflow real vaqtli ma’lumotlar oqimlarini boshqarish, masalan, sensorlardan kelgan ma’lumotlarni tozalash va integratsiya qilish kabi vazifalarda qulaylik yaratadi. Apache Airflow hujjatlari (2024) platformaning arxitekturasi va DAG’lar bilan ishlash mexanizmlarini batafsil yoritadi.

Ushbu tadqiqotda ikki xil ob-havo API’sidan (Weatherstack va WeatherAPI) ma’lumotlarni yig‘ish, ularni yagona formatga keltirish va ma’lumotlar bazasiga joylashtirishga qaratilgan ETL tizimi ishlab chiqildi. Jarayonlar Apache Airflow yordamida to‘liq avtomatlashtirildi. Metodologiya quyidagi bosqichlardan iborat:

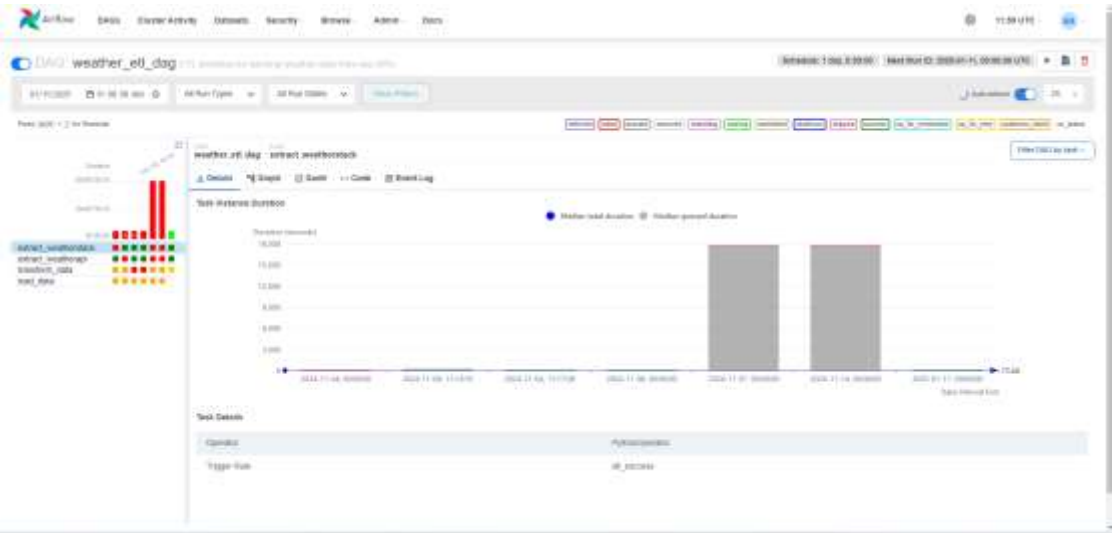
Extract (Ma’lumotlarni olish): Ob-havo ma’lumotlari ikki API’dan HTTP so‘rovlari orqali olindi. Har bir API’dan olingan ma’lumotlar (harorat, namlik, shamol tezligi, ob-havo holati) alohida fayllarga saqlandi.

Transform (Ma’lumotlarni qayta ishlash): Ikkala manbadan olingan ma’lumotlar o‘qildi, yagona formatga keltirildi va birlashtirildi. Ma’lumotlar tozalanishi va har bir qatorga manba identifikatori qo‘shildi.

Load (Ma’lumotlarni yuklash): Birlashtirilgan ma’lumotlar SQLite ma’lumotlar bazasiga yozildi. Bu jarayon ma’lumotlarni doimiy saqlash va keyingi tahlillar uchun tayyorlashni ta’minladi.

Ish oqimini avtomatlashtirish (DAG): Yuqoridagi bosqichlar Apache Airflow’da DAG sifatida tashkil qilindi. Har bir bosqich alohida vazifa sifatida belgilandi va ular ketma-ketlikda bajarilishi ta’minlandi. DAG har kuni avtomatik ravishda ishga tushadigan tarzda rejalashtirildi, bu esa foydalanuvchi ishtirokisiz to‘liq avtomatlashtirilgan ETL pipeline’ini yaratdi.

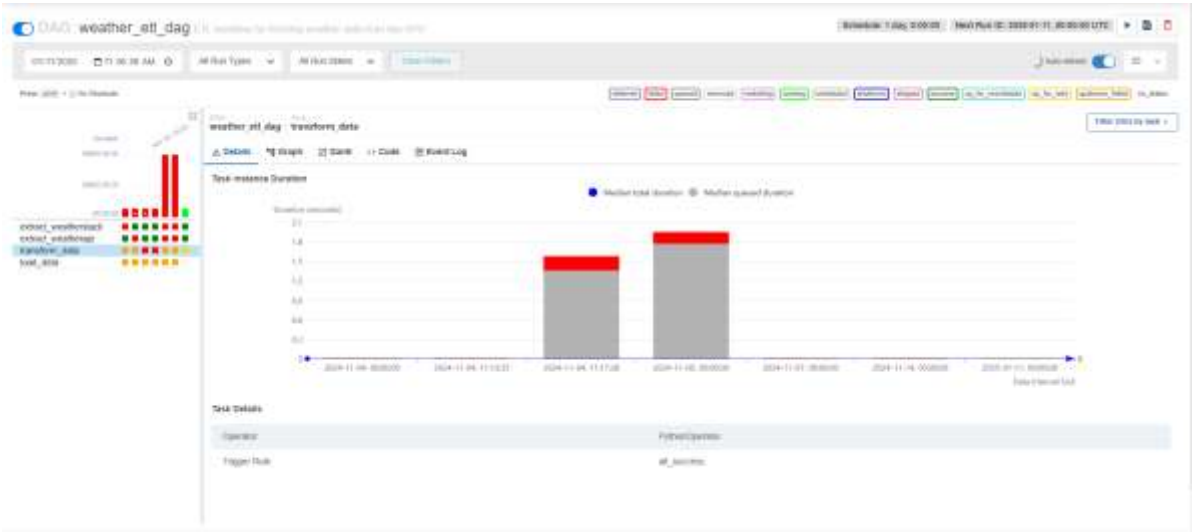
Results (Natijalar). Tadqiqot davomida Apache Airflow platformasidan foydalanib, ETL (Extract, Transform, Load) jarayonlarini avtomatlashtirishning samaradorligi chuqur tahlil qilindi. Olingan natijalar va amalga oshirilgan tajriba shuni ko‘rsatdiki, yaratilgan tizim murakkab ma’lumotlar oqimlarini avtomatik tarzda boshqarish va transformatsiya qilishda muvaffaqiyatli ishladi. Avvalo, Airflow yordamida ikki xil ob-havo ma’lumot manbasidan (Weatherstack va WeatherAPI) har kuni avtomatik ravishda ma’lumotlar olish, ularni yagona formatda birlashtirish va SQLite bazasiga saqlash jarayonlari muvaffaqiyatli amalga oshirildi. Har bir bosqichning bajarilishi Airflow'ning vizual interfeysi orqali real vaqt rejimida kuzatildi, bu esa tizimning ishonchliligini ta’minlashga yordam berdi. DAG (Directed Acyclic Graph) yordamida jarayonlar boshqarilib, har bir taskning muvaffaqiyatli bajarilganligi yoki xatolik yuzaga kelganligi haqida aniq ma’lumotlar taqdim etildi.



1-rasm. Ma’lumotlarni olish jarayoni natijasi.

Tizimning ishlashida muhim o‘rin tutgan jihatlardan biri — ma’lumotlarni transformatsiya qilish jarayoni bo‘ldi. Ikkita alohida ob-havo API’dan olingan ma’lumotlar muvaffaqiyatli yagona struktura shaklida birlashtirildi va saqlanishga tayyorlandi. Bu jarayon davomida, har bir manbadan olingan ma’lumotlar

formatlanib, bitta CSV faylga birlashtirildi, bu esa keyingi tahlil va vizualizatsiya jarayonlarini soddalashtirdi.



2-rasm. Transformatsiya jarayoni natijasi.

Ma’lumotlar bazasiga yuklash bosqichi ham muvaffaqiyatli amalga oshirildi. Birlashtirilgan ob-havo ma’lumotlari SQLite ma’lumotlar bazasiga saqlandi, bu esa ularni keyinchalik tizimda foydalanish uchun qulay shaklga keltirdi. Tizimning avtomatik ishlashi yordamida, inson aralashuvisiz har kuni ma’lumotlar yangilandi va saqlandi, bu esa tizimni yuqori samaradorlik bilan ishlashini ta’minladi.

Monitoring va xatoliklarni aniqlashda Apache Airflow’ning imkoniyatlari katta ahamiyatga ega bo‘ldi. Har bir taskning bajarilishi monitoring qilindi va agar biror xato yuzaga kelsa, tizim avtomatik ravishda qayta urinishni amalga oshirdi. Bu xatoliklarni aniqlash va tuzatish jarayonlarini sezilarli darajada soddalashtirdi. Shuningdek, tizimni kengaytirish va qo‘shimcha ma’lumot manbalarini integratsiya qilish imkoniyatlari mavjud. Boshqa ob-havo manbalaridan yoki turli xil sanoat ma’lumotlaridan foydalanib, tizimning qamrovini kengaytirish mumkin. Bu esa kelajakda tizimni yanada optimallashtirish va ko‘proq ma’lumotni samarali tarzda boshqarishga yordam beradi.

Umuman olganda, Apache Airflow asosidagi ETL tizimi nafaqat ma’lumotlarni boshqarish, balki ularni qayta ishlash va tahlil qilish jarayonlarini avtomatlashtirishda yuqori samaradorlikni namoyon etdi. Tizimning barqarorligi, avtomatlashtirilgan jarayonlarning takrorlanishi va kengaytirish imkoniyatlari uni murakkab ma’lumotlar tizimlarida qo‘llash uchun samarali va qulay vosita qiladi.

Discussion (Muhokama). Apache Tadqiqotda Apache Airflow platformasidan foydalanish orqali ETL (Extract, Transform, Load) jarayonlarini avtomatlashtirish samaradorligi o‘rganildi. Yaratilgan tizimning muvaffaqiyatli ishlashi, ma’lumotlarni avtomatik ravishda olish, transformatsiya qilish va saqlashni soddalashtirdi. Shuningdek, tizimning imkoniyatlari va kamchiliklari ham aniqlanib, ularning har biri tahlil qilindi. Airflow platformasining asosiy afzalliklaridan biri uning kengaytirilgan monitoring va vizualizatsiya imkoniyatlari bo‘lib, bu tizimni boshqarishda sezilarli darajada yordam beradi. DAG (Directed Acyclic Graph) yordamida har bir jarayonning bajarilishini kuzatish, xatoliklar yuzaga kelganda avtomatik ravishda qayta urinishni amalga oshirish imkonini berdi. Bu xususiyat tizimni ancha ishonchli va barqaror qildi. Shuningdek, tizimning modulyarligi va qo‘shimcha ma’lumot manbalarini integratsiya qilish imkoniyati uning kelajakdagi rivojlanishiga katta umidlar bog‘lashga imkon yaratadi. Turli xil ob-havo API-larini qo‘shish yoki boshqa sanoat ma’lumotlarini tizimga integratsiya qilish orqali tizimning ishlashini kengaytirish mumkin.

Biroq, tizimda ba’zi kamchiliklar ham mavjud. Misol uchun, ma’lumotlarni transformatsiya qilish jarayonida, har bir API’dan olingan ma’lumotlar formatlari o‘rtasida kichik farqlar mavjud edi. Bunday farqlarni aniqlash va ularni birlashtirish jarayonida qo‘shimcha vaqt sarflandi. Ma’lumotlarni qayta ishlash va integratsiya qilishda tizimda formatlashdagi noaniqliklar va nomuvofiqliklar keltirib chiqarilishi mumkin, bu esa ba’zi hollarda xatoliklarni keltirib chiqaradi. Bunday holatlar tizimning ishonchliligini kamaytirishi mumkin, ammo bu

muammolarni oldini olish uchun ma’lumotlarni oldindan yaxshilab formatlash va moslashtirish zarur.

Shuningdek, tizimning resurslarga ehtiyoji ham muhim masaladir. Har bir API so‘rovi, ma’lumotlarni yig‘ish va transformatsiya qilish jarayonlari server resurslariga ortiqcha yuk tushirishi mumkin. Ma’lumotlar miqdori ko‘paygan sari tizimning ishlash tezligi va samaradorligi pasayishi mumkin. Kengaytirilgan ma’lumotlarni ishlash uchun tizimning imkoniyatlarini yanada oshirish mumkin. Masalan, ma’lumotlarni tahlil qilish va vizualizatsiya qilish uchun qo‘shimcha vositalar integratsiya qilish, tizimdan olinadigan ma’lumotlarning sifatini oshirishi mumkin. Ma’lumotlarning vizual tahlili orqali foydalanuvchilar tezda ma’lumotlar bazasidagi o‘zgarishlarni ko‘rishlari va tezkor qarorlar qabul qilishlari mumkin. Bu esa tizimning umumiy samaradorligini oshiradi. Apache Airflow yordamida ETL jarayonlarini avtomatlashtirish juda samarali va qulay vosita bo‘lib, ma’lumotlarni to‘plash, transformatsiya qilish va saqlash jarayonlarini soddalashtirdi.

Conclusion (Xulosa). Ushbu tadqiqot Apache Airflow platformasi yordamida ETL jarayonlarini avtomatlashtirishning samaradorligini muvaffaqiyatli namoyish etdi. Yaratilgan tizim ikki xil ob-havo API’sidan ma’lumotlarni avtomatik yig‘ish, ularni yagona formatga keltirish va SQLite ma’lumotlar bazasiga saqlashni muammosiz amalga oshirdi. Airflow’ning kuchli monitoring, vizualizatsiya va xatolarni boshqarish imkoniyatlari jarayonlarni samarali tashkil qilishda muhim rol o‘ynadi. Tizimning modulyar tuzilishi, avtomatik qayta urinish mexanizmlari va yangi ma’lumot manbalarini integratsiya qilish qulayligi uning asosiy afzalliklari sifatida ajralib turdi.

Biroq, ma’lumotlar formatidagi noaniqliklar va resurs talabgorligi kabi cheklovlar kelajakda yaxshilanishi zarur bo‘lgan jihatlardir. Tizimni yanada optimallashtirish uchun ma’lumotlarni oldindan standartlashtirish, bulutli yechimlardan foydalanish va qo‘shimcha tahlil vositalarini integratsiya qilish

tavsiya etiladi. Apache Airflow’ning moslashuvchanligi va kengaytirilishi mumkin bo‘lgan arxitekturasi uni nafaqat ob-havo ma’lumotlari, balki turli sohalardagi ma’lumotlar oqimlarini avtomatlashtirish uchun ideal vositaga aylantiradi. Ushbu tadqiqot natijalari Airflow’ning zamonaviy ma’lumotlar boshqaruvi tizimlaridagi muhim o‘rnini yana bir bor tasdiqlaydi va kelajakda yanada kengroq qo‘llanilish imkoniyatlarini ochib beradi..

References (Foydalanilgan adabiyotlar)

1. Harish Goud Kola. Optimizing ETL Processes for Big Data Applications: International Journal of Engineering and Management Research, Volume-14, Issue-5 (October 2024). https://doi.org/10.5281/zenodo.14184235

2. Matt Palmer, Understanding ETL Data Pipelines for Modern Data Architectures, O Reilly, pp. 1-107, 2024. [Online]. Available: https://www.databricks.com/sites/default/files/2024-03/oreilly-technical-guide-understanding-etl.pdf

3. Airflow Documentation. (2024). *Apache Airflow: The platform to programmatically author, schedule and monitor workflows.* Retrieved from https://airflow.apache.org/

4. https://www.bigdataframework.org/knowledge/etl-in-data-engineering/