УДК 004.62, 004.75

М. О. Кузьміна

ЗАСТОСУВАННЯ ХМАРНИХ СЕРВІСІВ GOOGLE ТА ПЛАТФОРМИ COLAB ДЛЯ ПОПЕРЕДНЬОЇ ОБРОБКИ І ВІЗУАЛІЗАЦІЇ ВЕЛИКИХ МАСИВІВ ДАНИХ

Донецький національний університет імені Василя Стуса

Анотація

Здійснено аналіз можливостей використання екосистеми хмарних сервісів Google для побудови конвеєра обробки даних. Основна увага приділена інтеграції Google Drive, Google Sheets та середовища Google Colab. Розроблено методику автоматизованого обміну даними між хмарним сховищем та середовищем виконання Python-скриптів. Продемонстровано ефективність використання бібліотек Pandas та gspread для попередньої обробки ("очищення") даних та їх подальшої візуалізації. Отримані результати підтверджують доцільність використання запропонованого підходу для освітніх та наукових задач, що не вимагають розгортання дороговартісної серверної інфраструктури.

Ключові слова: *хмарні технології, Big Data, Google Colab, Google Sheets, візуалізація даних, Python, API, попередні обробка даних.*

Abstract

The paper deals with the application of the Google cloud services ecosystem for building a data processing pipeline. The main attention is paid to the integration of Google Drive, Google Sheets, and the Google Colab environment. A methodology for automated data exchange between cloud storage and the Python script execution environment has been developed. The effectiveness of using Pandas and gspread libraries for data preprocessing ("cleaning") and subsequent visualization is demonstrated. The results confirmed the feasibility of using the proposed approach for educational and scientific tasks that do not require the deployment of expensive server infrastructure.

Keywords: *cloud technologies, Big Data, Google Colab, Google Sheets, data visualization, Python, API, data preprocessing.*

Вступ

В епоху стрімкого накопичення інформації (Big Data) методи її обробки та аналізу стають критично важливими для прийняття рішень у науковій та бізнес-сферах. Традиційні підходи, що передбачають локальне зберігання та обробку даних, часто стикаються з обмеженнями апаратного забезпечення та складністю масштабування [1]. Хмарні обчислення (Cloud Computing) пропонують вирішення цих проблем, надаючи доступ до розподілених ресурсів на вимогу. Серед існуючих рішень особливе місце займає екосистема Google, яка поєднує в собі інструменти для зберігання (Drive), структурування (Sheets) та обробки (Colab) даних. Дослідження можливостей їх безшовної інтеграції є актуальним завданням, оскільки це дозволяє створити доступне середовище для роботи з даними без необхідності фінансових витрат на ліцензійне ПЗ чи серверне обладнання.

Постановка задачі дослідження

При розробці методики використання хмарних сервісів Google для задач Data Science необхідно:

- провести аналіз інструментарію Google Cloud Platform та користувацьких сервісів (SaaS/PaaS) для роботи з даними;
- дослідити методи програмного доступу до файлової системи Google Drive та таблиць Google Sheets через API;
- розробити алгоритм попередньої обробки "сирих" даних, розміщених у хмарі, з використанням мови Python у середовищі Colab;

- реалізувати візуалізацію результатів аналізу та забезпечити зворотну синхронізацію оброблених даних;
- оцінити ефективність запропонованого підходу порівняно з локальними аналогами (Jupyter Notebook, Excel).

Виклад основного матеріалу

Дослідження зосереджене на створенні єдиного інтелектуального циклу обробки даних. Основою архітектури є використання Google Drive як централізованого сховища (Data Lake), куди надходять "сирі" дані у форматах CSV, JSON або XLSX. Для реалізації взаємодії між компонентами системи використано можливості Google Colab — хмарного середовища на базі Jupyter Notebook, що надає безкоштовний доступ до обчислювальних потужностей (CPU/GPU) [2].

Першим етапом є налаштування доступу до файлової системи. Використання бібліотеки google.colab дозволяє монтувати віртуальний диск: from google.colab import drive drive.mount('/content/drive'). Це надає можливість працювати з віддаленими файлами як з локальними об’єктами, що значно спрощує процес завантаження великих датасетів.

Другим, критично важливим етапом, є попередня обробка даних. Часто вхідні дані містять пропуски, дублікати або помилки форматування. Для ручної корекції та верифікації даних доцільно використовувати Google Sheets. Це дозволяє команді дослідників одночасно працювати над розміткою або виправленням даних у реальному часі. Для автоматизованого зчитування цих даних у Python-скрипт використовується бібліотека gspread та oauth2client для авторизації через Google API. Було розроблено скрипт, який виконує такі дії.

1. Авторизується в Google Cloud Console.
2. Зчитує дані з конкретного аркуша Google Sheets у об'єкт DataFrame бібліотеки Pandas.
3. Виконує програмну очистку (видалення NaN, нормалізацію числових значень).

Третій етап — аналіз та візуалізація. Використовуючи потужності Colab, можна обробляти масиви даних, що перевищують можливості стандартних табличних процесорів. За допомогою бібліотек Matplotlib та Seaborn реалізовано побудову складних графіків (теплових карт, діаграм розсіювання), які візуалізують кореляції у даних [3]. Важливим аспектом є використання підходу "Reproducible Research" (відтворювані дослідження). Оскільки код та дані знаходяться у хмарі, іншим дослідникам достатньо отримати посилання на Notebook, щоб повторити експеримент, маючи ідентичне середовище виконання.

Результати дослідження

Проведене тестування розробленої системи на тестовому датасеті обсягом 500 МБ показало високу ефективність інтеграції. Час на розгортання робочого середовища скоротився з годин (у випадку налаштування локального Anaconda/Jupyter) до хвилин. Використання Google Sheets як проміжного шару для "м'якої" очистки даних дозволило залучити до процесу експертів предметної області, які не володіють програмуванням, але можуть редагувати таблиці [4, 5]. Візуалізація даних у Colab продемонструвала високу швидкодію. Завдяки хмарним ресурсам, генерація складних графіків займає секунди, не навантажуючи локальний комп'ютер користувача. Отримані графіки можуть бути автоматично збережені назад на Google Drive у форматі PNG/PDF для подальшого використання у звітах.

Висновки

У ході роботи було обґрунтовано та практично реалізовано методику застосування хмарних сервісів Google для задач Big Data. Поєднання Google Drive (зберігання), Sheets (редагування) та Colab (обчислення) створює повноцінну, безкоштовну та гнучку екосистему для дослідників даних. Запропонований підхід вирішує проблеми доступності ресурсів, спільної роботи та відтворюваності наукових результатів. Він може бути рекомендований для використання у навчальному процесі при вивченні дисциплін "Хмарні технології" та "Аналіз даних", а також для виконання наукових досліджень студентами та аспірантами. Перспективним напрямком подальших досліджень є

автоматизація запуску скриптів за розкладом (Triggers) та інтеграція з BigQuery для роботи з надвеликими масивами даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Raschka S., Mirjalili V. Python Machine Learning. Packt Publishing, 2019. 770 p.
2. Bisong E. Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, 2019. 709 p. URL: https://link.springer.com/book/10.1007/978-1-4842-4470-8
3. VanderPlas J. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2016. URL: https://jakevdp.github.io/PythonDataScienceHandbook/
4. Офіційна документація Google Colaboratory. URL:https://colab.research.google.com/notebooks/intro.ipynb
5. Google Sheets API Overview. URL:https://developers.google.com/sheets/api/guides/concepts

Кузьміна Маргарита Олегівна – студентка кафедри інформаційних технологій, факультет інформаційних та прикладних технологій, Донецький національний університет імені Василя Стуса, м. Вінниця, e-mail: kuzmina.m@donnu.edu.ua

Kuzmina Margarita Olegivna – student of Information Technology Department, Faculty of Information and Applied Technologies, Vasyl Stus Donetsk National University, Vinnytsia, e-mail: kuzmina.m@donnu.edu.ua