ETL-проєкт аналізу даних Uber

Формат даних та джерело

Формат даних: CSV

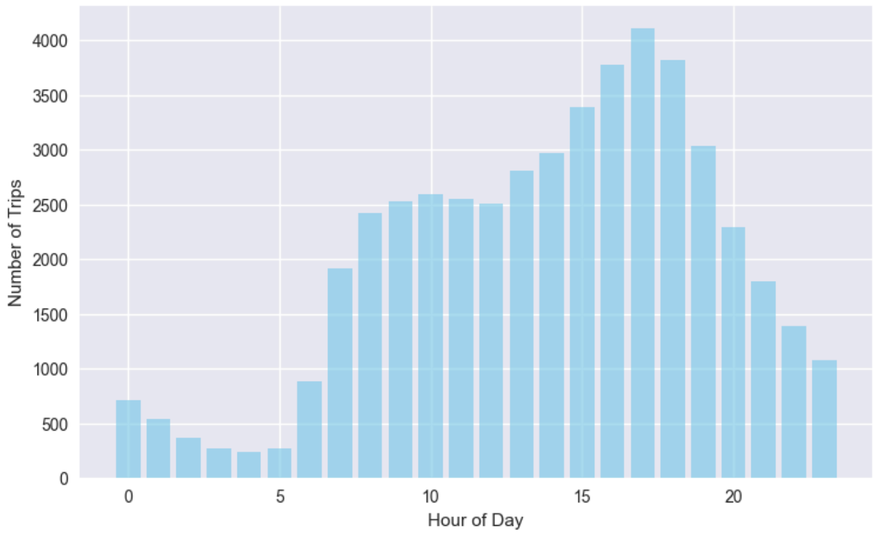
Джерело даних: Публічні дані про поїздки uber таксі Нью-Йорка від NYC TLC

Типові поля: час посадки/висадки, відстань, сума поїздки, тип оплати, кількість пасажирів

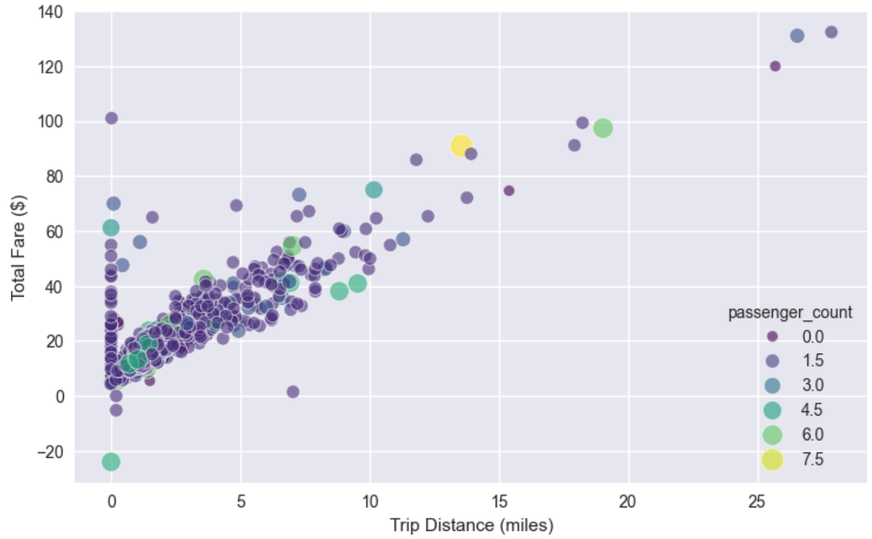
Pickup Time	Distance (miles)	Base Fare ($)	Total Fare ($)	Passengers	Avg Speed (mph)
2025-01-01 00:28	6.97	41.5	52.55	2	10.45
2025-01-01 00:11	0.75	5.8	10.79	5	12.16
2025-01-01 00:48	17.07	87.0	108.9	2	22.66
2025-01-01 00:30	0.78	7.2	11.64	6	10.21
2025-01-01 00:43	3.66	19.8	29.05	5	12.74
2025-01-01 00:43	0.02	3.0	5.5	2	9.0
2025-01-01 01:01	0.01	25.0	30.36	3	4.0
2025-01-31 23:57	11.1	52.7	63.7	2	17.11
2025-01-31 23:02	0.9	7.9	12.48	6	10.45
2025-01-31 23:43	1.32	10.7	15.84	5	8.15
2025-01-31 23:13	1.28	10.0	15.0	2	9.78
2025-01-31 22:57	6.32	26.8	33.3	5	23.8
2025-01-31 23:48	1.49	10.7	15.84	5	8.93

NYC Green Taxi Trip Analysis Dashboard

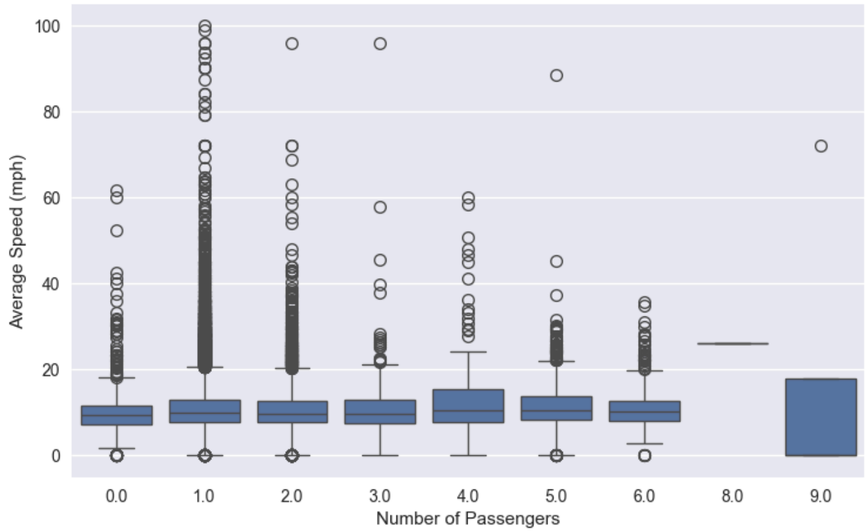
Hourly Trip Distribution



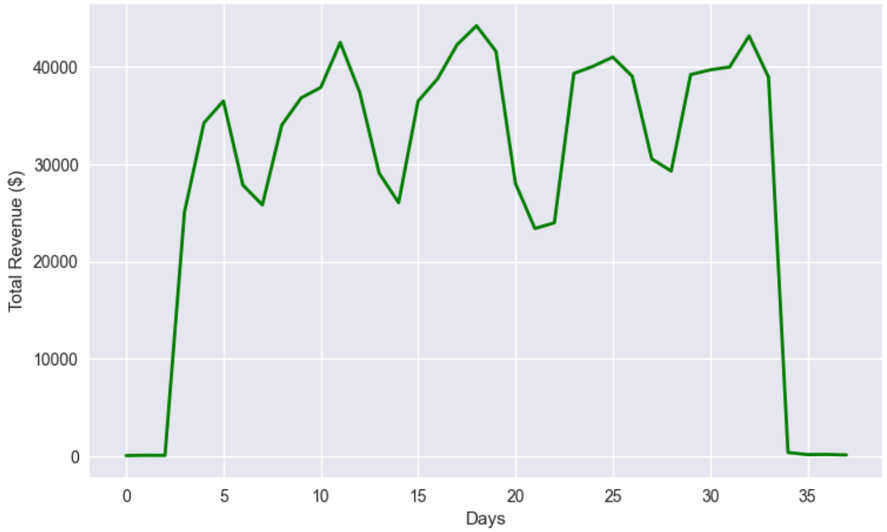
Fare vs Distance by Passenger Count



Trip Speed by Passenger Count



Daily Revenue Trend



Архітектура ETL-процесу

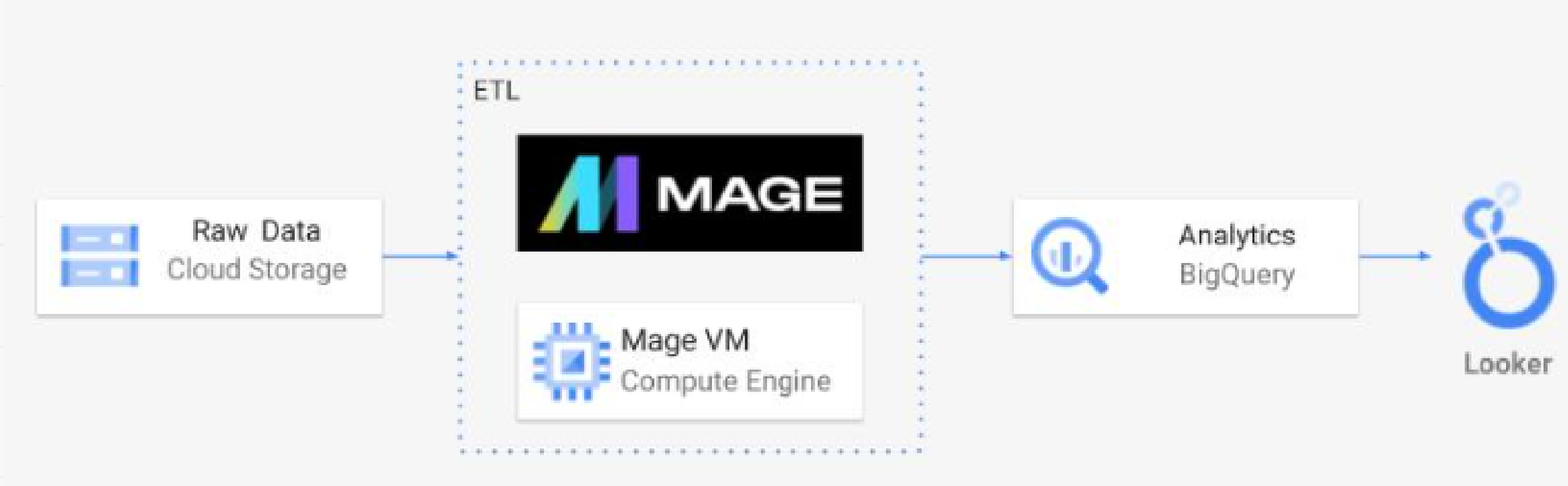
1) Збір даних: Завантаження CSV-файлів з сайту NYC TLC

2) Зберігання: Завантаження даних у Google Cloud Storage

3) ETL-процес: Використання Mage для обробки даних

4) Аналітика: Збереження оброблених даних у Google BigQuery

5) Візуалізація: Створення дашбордів у Looker Studio



Інструменти для використання

01 - Mage

Платформа для створення ETL-пайплайнів. Використовується для автоматизованого витягування, трансформації та завантаження даних. Простий інтерфейс та інтеграція з хмарними сервісами.

02 - Google Cloud Storage (GCS)

Хмарне сховище, де зберігаються початкові CSV-файли з поїздками. Надійне та масштабоване середовище для зберігання великих обсягів даних

03 - BigQuery

Хмарна аналітична база даних від Google. Дозволяє швидко виконувати SQL-запити над великими наборами даних.

04 - Looker Studio

Інструмент для візуалізації даних. Тут створюються дашборди для аналітики поїздок, витрат, маршрутів.

05 - Jupyter Notebook

Середовище для аналізу і попередньої обробки даних на Python. Використовується для дослідження CSV-файлів перед завантаженням.

06 - Apache Airflow

Планувальник та оркестратор ETL-процесів. Дозволяє керувати виконанням ETL-пайплайнів за розкладом (може бути згаданий як альтернатива Mage).

Обґрунтування необхідності розробки ETL-схем

Візуалізація

Зручне представлення даних для бізнес-користувачів

Масштабованість

Можливість обробки великих обсягів даних

Автоматизація

Зменшення ручної роботи та помилок

Аналітика

Швидкий доступ до оброблених даних для прийняття рішень

Висновки

Проєкт демонструє повний цикл обробки даних: від збору до візуалізації

Використання сучасних інструментів забезпечує ефективність та надійність процесу

Модель може бути адаптована для інших джерел даних та бізнес-завдань