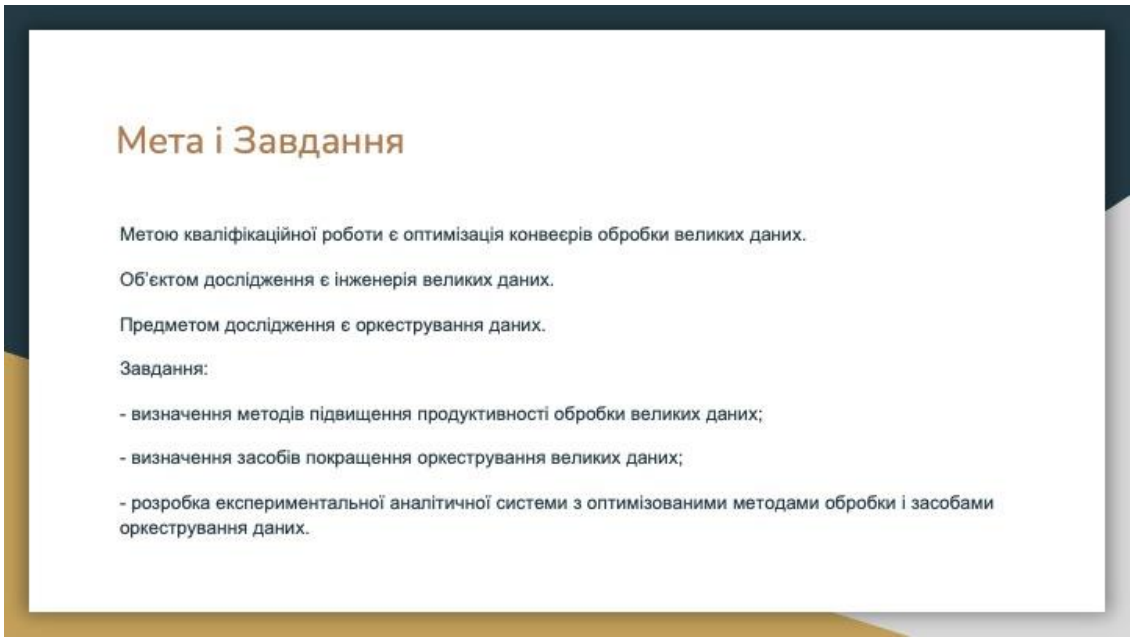
ДОДАТОК А

Графічний матеріал кваліфікаційної роботи


Міністерство освіти і науки України
Харківський національний університет радіоелектроніки
Кафедра ЕОМ
Методи та засоби автоматизації
процесів обробки великих даних
Кваліфікаційна робота
Другий (магістерський) рівень
Автор:
студ.гр.СПм-22-4
Залеський В.Д.
Керівник:
доц.каф.ЕОМ
Федорченко В.М.


Мета і Завдання
Метою кваліфікаційної роботи є оптимізація конвеєрів обробки великих даних.
Об'єктом дослідження є інженерія великих даних.
Предметом дослідження є оркестрування даних.
Завдання:
- визначення методів підвищення продуктивності обробки великих даних;
- визначення засобів покращення оркестрування великих даних;
- розробка експериментальної аналітичної системи з оптимізованими методами обробки і засобами оркестрування даних.

Актуальність

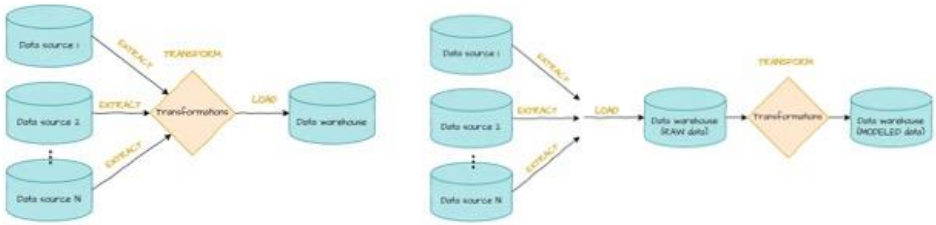
Зі збільшенням кількості даних збільшується потреба в управлінні, синхронізації розкладів та вирішення проблем обробки.

Виникає необхідність зламати бар'єри між джерелами даних та сховищами, щоб по-справжньому використовувати всю інформацію, яка збираються.

Оркестрування даних дозволяє автоматизувати та оптимізувати дані, перетворюючи їх на оперативні активи, щоб цінну інформацію можна було використовувати для прийняття бізнес-рішень у режимі реального часу.

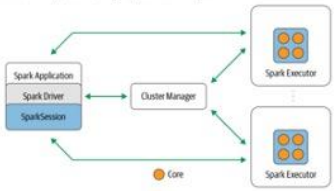
За деякими оцінками, 55% роботи, пов'язаної з аналізом даних, зводиться до збирання та підготовки даних, що означає, що оркестрування даних може скоротити велику кількість часу на обробку та планування.

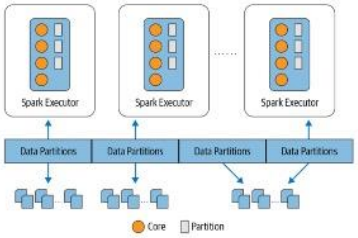
Порівняння ETL та ELT процесів



Обробки великих даних поза сховищем даних з використанням Spark

Apache Spark – це уніфікований механізм, призначений для розподіленої обробки великих даних, як у локальних дата-центрах, так і в хмарі.





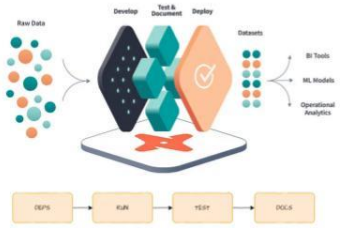
Обробки великих даних безпосередньо у сховищі даних з використанням DBT

DBT - аналітичний інструмент з відкритим кодом.

Основна мета DBT - допомогти перетворити дані сховища даних у простий спосіб – за рахунок виконання SQL-запитів.

На відміну від інструментів ETL, які можуть мати власні обчислювальні механізми, DBT компілює і виконує SQL безпосередньо у цільовому сховищі даних, використовуючи його обчислювальну потужність для перетворень. При цьому цьому зазвичай сховища обробляють дані за MPP методологією.

MPP методологія полягає в розбитті запиту до бази даних або операції на менші завдання, які можуть виконуватися паралельно на декількох обчислювальних ресурсах, таких як ядра процесорів у вузлах кластеру.



Поєднання Spark & MPP для підвищення ефективності обробки даних

Інтегруючи Spark з MPP, можна ефективно покрити різноманітні потреби в обробці даних.

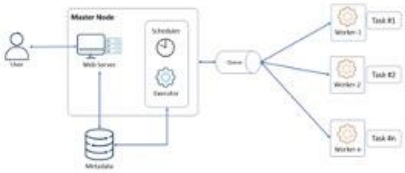
Spark забезпечує ефективну попередню обробку даних і складні трансформації, в той час як MPP може забезпечити необхідну швидкість для запитів і транзакцій в режимі реального часу.

Саме тому така комбінація забезпечує оптимальне управління як дослідницькими(ad-hoc), так і виробничими(prod) навантаженнями, балансуючи між гнучкістю і масштабованістю Spark та продуктивністю MPP сховищ даних.

Оркестрування обробки великих даних

Оркестрування даних часто плутають з оркеструванням робочих процесів. Оркестрування робочих процесів – це процес запуску та моніторингу стану завдань. Її сутність реалізувати аналог event-driven систем на базі пакетної обробки.

Оркестрування даних є підмножиною оркестрування робочих процесів і забезпечує надійну та ефективного синхронізацію даних у продакшен середовищі. На ряду з елементами ETL вона може також включати: елементи управління середовищем (CI/CD для даних або GitOps для даних або безперервна інтеграція та доставка даних), контроль доступу на основі ролей ("RBAC"), оповіщення та моніторинг стану даних.



Порівняння засобів оркестрування даних

Prefect добре підходить , коли не було попереднього досвіду в побудові DAG-ів і потрібний простий в освоєнні інструмент з великою спільнотою, а також гібридна модель.

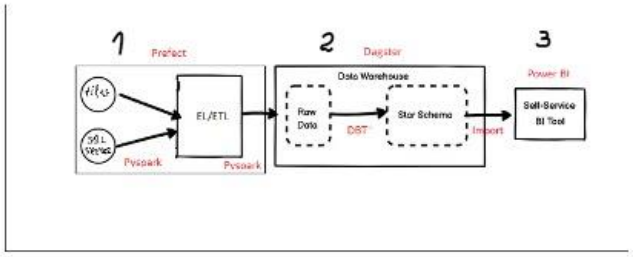
Mage підходить для тих випадків, коли потрібна обробка даних в реальному часі та розгортання в контейнеризованих середовищах.

Kestra чудовий вибір для команд, які не мали попереднього досвіду в побудові DAG-ів але мають гарний досвід з розгортанням інфраструктури з застосування принципів IaC.

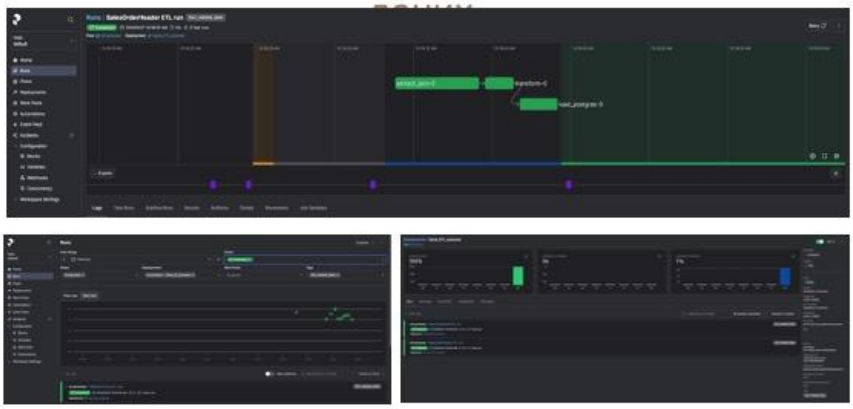
Dagster добре підходить, коли потрібен інструмент для оркестрації dbt процесів та коли надається пріорітет наглядності та простоті у розумінні робочого процесу.

Критерій	Prefect	Mage	Kestra	Dagster
Модель виконання	Гібридна (локальна/хмарна)	Контейнеризована	Універсальна	Універсальна
Підхід до опису робочих процесів	Декоратори Python	Python	YAML	Python
Мови програмування	Python	Python	Python, R, Julia, Node.js, etc.	Python
Спільнота	Велика	Зростаюча	Зростаюча	Зростаюча
Гнучкість	Висока	Середня	Висока	Висока
Масштабованість	Висока	Висока	Висока	Висока
Складність використання	Середня	Низька	Низька	Середня
Підтримка SaaS та хмарних сервісів	Висока	Висока	Середня	Низька
Відображення залежностей	Динамічне	Статичне	Статичне	Статичне
Підтримка контейнеризації	Висока	Висока	Середня	Низька

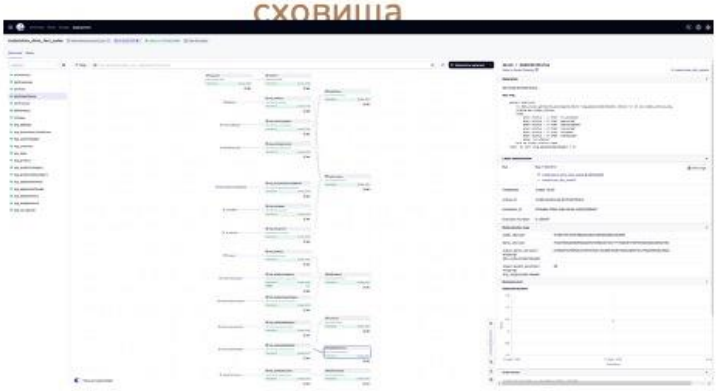
Спрощена архітектура експериментальної аналітичної системи



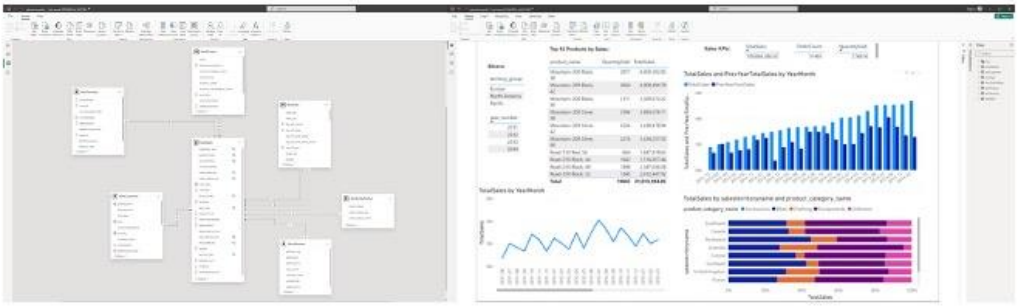
Оркестрування процесу обробки поза сховищем



Оркестрування процесу обробки даних на стороні сховища



Аналітичні панелі на базі даних зі сховища



Апробація

Висновки

У ході виконання кваліфікаційної роботи було проведено дослідження тенденцій методів обробки великих даних та аналіз засобів оркестрування обробки великих даних.

Також була розроблена експериментальна аналітична система з оптимізованими засобами оркестрування та продуктивними методами обробки даних. Ця система складається з трьох компонентів:

- оркестрований процес обробки поза сховищем даних

- оркестрований процес обробки даних на стороні сховища

- аналітичні панелі на базі даних зі сховища.