








# End to End Data Pipeline untuk Melihat Pengaruh Cuaca Terhadap Jumlah Penjualan

## Overview

Masalah ini berfokus pada membangun pipeline dan model prediksi untuk memanfaatkan data cuaca dan data penjualan dalam memprediksi jumlah penjualan.

Proyek ini bertujuan untuk membangun sistem prediksi penjualan berbasis cuaca dengan mengintegrasikan data dari API cuaca dan dataset penjualan untuk membantu bisnis meningkatkan akurasi perencanaan stok dan strategi penjualan. Dengan pipeline otomatis menggunakan Apache Airflow, data cuaca seperti suhu, kelembapan, dan curah hujan akan digabungkan dengan data penjualan untuk analisis dan modeling. Proses ini melibatkan ETL (Extract, Transform, Load) untuk menghasilkan dataset yang terstruktur, diikuti dengan feature engineering untuk menciptakan variabel baru yang relevan.

## Table of contents

-  [Overview](#)
-  [Stakeholders](#)
-  [Sumber](#)
-  [Tujuan & Hipotesis](#)
-  [Latar Belakang](#)
-  [Tujuan Solusi](#)
-  [End to End Workflow](#)
  - [Pipeline](#)
  - [Architecture](#)
  - [Testing](#)
- [Algoritma Predictive Modelling](#)
- [Alur Penggunaannya](#)
  - [Alur](#)
- [Kesimpulan](#)

## Stakeholders

Pihak yang terlibat dalam proyek ini

Data Engineer	Diah Ayu Rahmawati
	Navika Berlianda R
Project Manager	Tsabitah Inayah

## Sumber

Berisi link yang berkaitan dengan dataset serta API yang digunakan sebagai sumber untuk proyek ini.

- <https://openweathermap.org/api>

- [https://github.com/diahayuuu/Airflow/blob/main/retail\\_sales\\_dataset.csv](https://github.com/diahayuuu/Airflow/blob/main/retail_sales_dataset.csv)

## Tujuan & Hipotesis

Bagian ini menjelaskan bagaimana proyek ini ingin diimplementasikan dan diinisiasi untuk menguji hipotesis apa serta tentunya tujuan yang ingin dicapai.

### Hipotesis :

1. **Cuaca memengaruhi pola pembelian konsumen** — kondisi cuaca tertentu (misalnya, hujan deras atau suhu tinggi) akan berdampak pada tingkat penjualan produk tertentu.
2. **Suhu dan curah hujan berhubungan positif atau negatif dengan volume penjualan** — misalnya, saat cuaca panas, penjualan minuman dingin atau es krim cenderung meningkat.
3. **Pola penjualan berulang terkait cuaca** — pola penjualan produk tertentu akan cenderung mengikuti pola cuaca yang berulang (seperti musim panas dan musim hujan).
4. **Hari-hari dengan kondisi cuaca ekstrem** memiliki efek yang lebih besar terhadap penjualan dibandingkan dengan kondisi cuaca normal.
5. **Integrasi data cuaca akan meningkatkan akurasi model prediksi** dibandingkan dengan model yang hanya menggunakan data historis penjualan.

### Tujuan Projek :

1. **Membangun pipeline data otomatis** untuk mengumpulkan, mengolah, dan menyimpan data cuaca dan data penjualan secara terintegrasi.
2. **Mengembangkan model prediksi** untuk memperkirakan jumlah penjualan berdasarkan data cuaca dan faktor-faktor lainnya.
3. **Mengidentifikasi faktor cuaca yang memengaruhi penjualan**, seperti suhu, kelembapan, dan curah hujan.
4. **Meningkatkan akurasi perencanaan penjualan dan stok barang** dengan menggunakan prediksi berdasarkan kondisi cuaca.
5. **Mengurangi risiko kehabisan atau kelebihan stok** dengan perkiraan yang lebih akurat mengenai perubahan permintaan terkait cuaca.



## Latar Belakang

Dalam dunia bisnis, memahami faktor-faktor yang memengaruhi tingkat penjualan merupakan hal yang sangat penting. Salah satu faktor eksternal yang sering kali memengaruhi perilaku konsumen adalah **kondisi cuaca**. Cuaca dapat memengaruhi preferensi pembelian konsumen, seperti meningkatnya permintaan minuman dingin saat cuaca panas atau berkurangnya aktivitas belanja saat hujan deras.

Namun, banyak perusahaan belum sepenuhnya memanfaatkan data cuaca untuk meningkatkan prediksi penjualan mereka. Biasanya, model prediksi hanya mengandalkan data historis penjualan, tanpa memperhatikan faktor eksternal seperti perubahan iklim atau kondisi cuaca yang mungkin berdampak langsung pada permintaan produk.



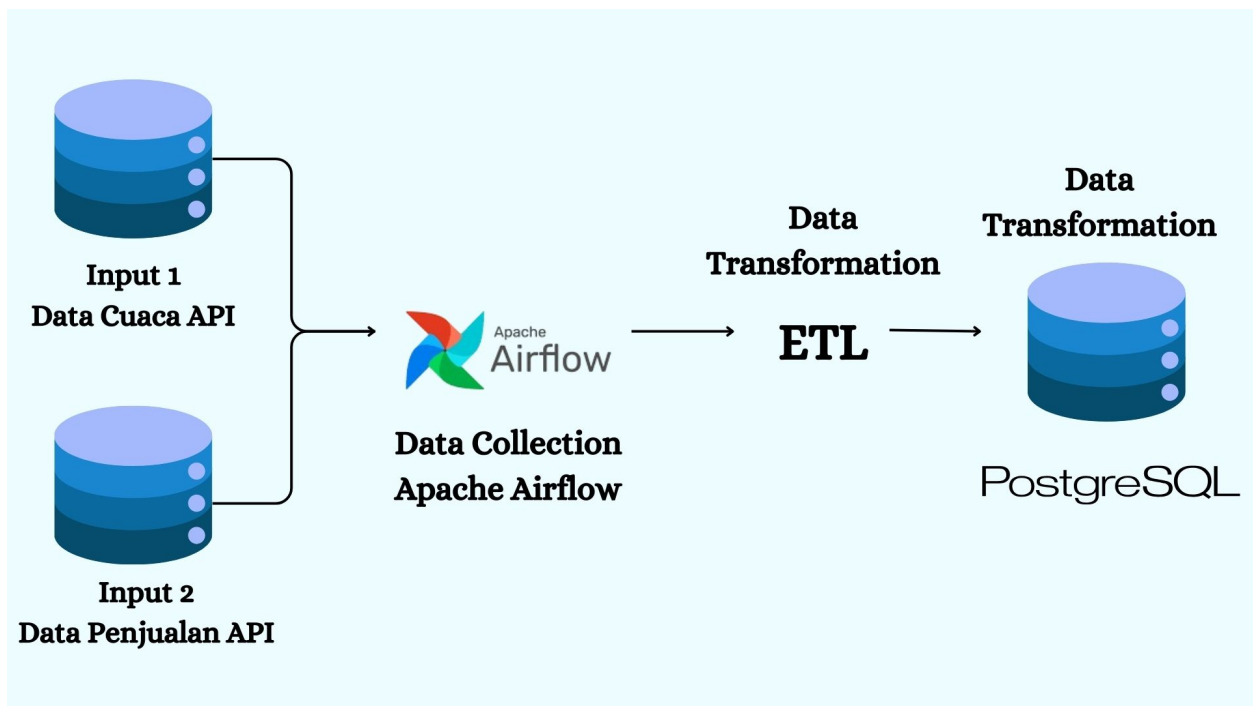
## Tujuan Solusi

1. Membangun Pipeline Data Otomatis untuk Integrasi Data Cuaca dan Penjualan
2. Mengembangkan Model Prediktif Berbasis Cuaca
3. Memperluas Analisis dengan Menambahkan Fitur-Fitur Berbasis Cuaca

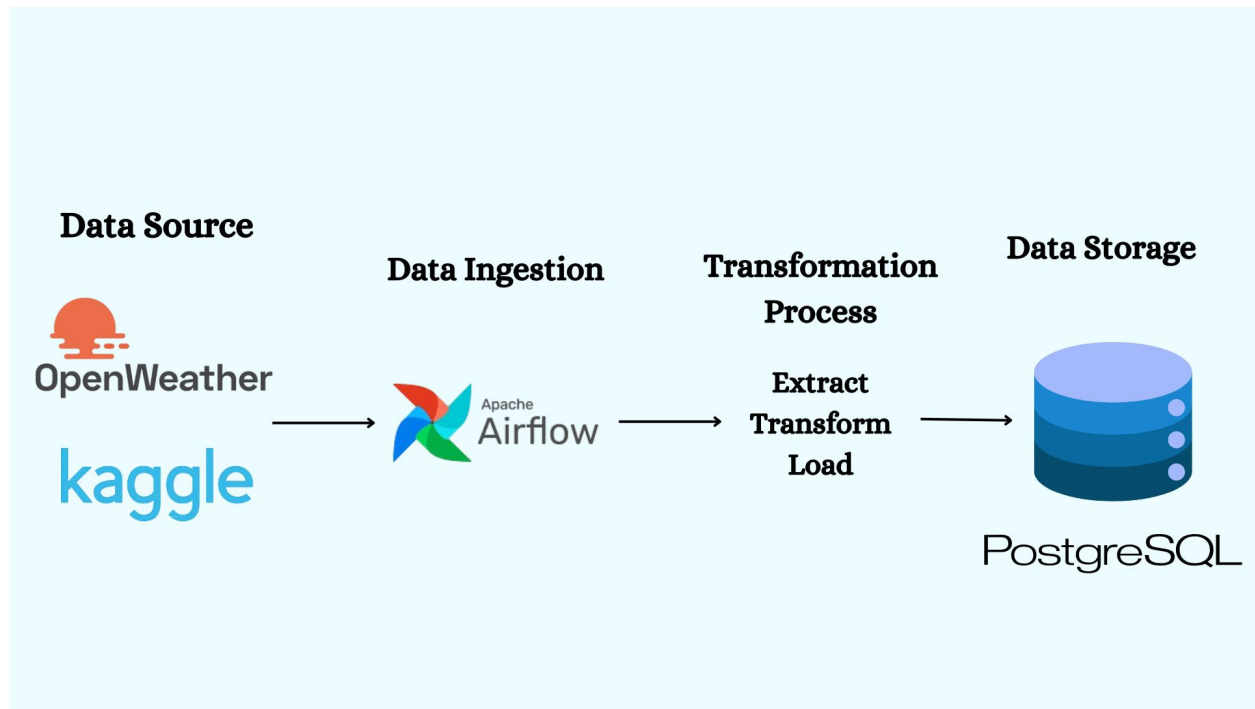
4. Meningkatkan Akurasi Prediksi Penjualan dengan Pendekatan Berbasis Data Eksternal
5. Membantu Pengambilan Keputusan dalam Manajemen Stok dan Persediaan
6. Menyediakan Sistem Prediksi yang Mudah Diakses dan Terintegrasi dengan Proses Bisnis

## End to End Workflow

### Pipeline



### Architecture



## Testing

1. **Unit Testing:** Tes komponen individual, terutama API dan transformasi data.
2. **Integration Testing:** Tes alur data antara komponen, terutama antara Airflow, ETL, dan PostgreSQL.
3. **Data Validation Testing:** Tes kualitas data untuk menjaga integritas data selama pipeline.
4. **Performance Testing:** Tes kecepatan dan ketahanan pipeline terhadap data dalam jumlah besar.
5. **Error Handling and Recovery Testing:** Tes untuk memastikan pipeline dapat pulih dari kegagalan.
6. **Regression Testing:** Tes setelah pembaruan untuk memastikan stabilitas pipeline.

## Algoritma Predictive Modelling

Di dalam konteks proyek Anda, predictive modeling dapat diterapkan untuk memprediksi berbagai aspek, seperti:

Prediksi Cuaca	Prediksi Penjualan
Dengan menggunakan data historis cuaca (misalnya suhu, tekanan udara, kelembaban), kita dapat memprediksi kondisi cuaca untuk periode tertentu di masa depan.	Berdasarkan data penjualan yang terkumpul (seperti produk yang terjual, jumlah penjualan, lokasi penjualan), model prediktif dapat digunakan untuk memprediksi jumlah penjualan di masa depan untuk produk tertentu atau seluruh kategori produk.

## Alur Penggunaannya

Pengambilan Data Cuaca	Pengambilan Data Penjualan	ETL (Extract, Transform, Load)	Jadwal Pemrosesan Data
Fetch Weather Data: Fungsi <code>fetch_weather_data()</code> bertanggung jawab untuk mengambil data cuaca dari API OpenWeather. Data ini kemudian diproses dan disimpan di database PostgreSQL melalui fungsi <code>save_weather_data_to_postgres()</code> .	Fetch Retail Sales Data: Fungsi <code>fetch_sales_data()</code> mengambil data penjualan dari file CSV ( <code>retail_sales_dataset.csv</code> ). Data ini kemudian diproses dan dapat digunakan untuk analisis lebih lanjut.	Extract: Mengambil data dari sumbernya, baik itu API cuaca atau file CSV. Transform: Mengubah data menjadi format yang sesuai dan menyiapkannya untuk dimasukkan ke dalam sistem penyimpanan. Load: Memasukkan data yang telah diproses ke dalam PostgreSQL.	DAG (Directed Acyclic Graph): DAG ini didefinisikan menggunakan Apache Airflow untuk menjalankan pipeline secara otomatis setiap hari (menggunakan <code>@daily</code> ).

Penyimpanan Cuaca ke PostgreSQL: Data yang diambil meliputi nama kota, suhu, dan deskripsi cuaca, yang kemudian dimasukkan ke dalam tabel weather_data di database PostgreSQL.	Penyimpanan Penjualan ke PostgreSQL: Data yang diambil dari file CSV disimpan ke dalam tabel yang sesuai di PostgreSQL.	Task Dependency: weather_task dijalankan terlebih dahulu untuk mengambil dan menyimpan data cuaca, baru kemudian sales_task dijalankan untuk memproses data penjualan.
--	---	--

## Alur

### 1. Prediksi dengan Model

- Setelah data cuaca dan penjualan disimpan di PostgreSQL, model prediksi dapat dibangun dan digunakan untuk memprediksi cuaca atau penjualan di masa depan berdasarkan data historis.

Integrasi Model Prediksi: Anda bisa menambahkan model machine learning di dalam DAG ini untuk memproses data penjualan (seperti prediksi penjualan) dengan menggunakan teknik seperti regresi linier, decision tree, atau model lainnya.

Pengawasan dan Pemeliharaan:

### 2. Integrasi Model Prediksi

- Anda bisa menambahkan model machine learning di dalam DAG ini untuk memproses data penjualan (seperti prediksi penjualan) dengan menggunakan teknik seperti regresi linier, decision tree, atau model lainnya.

### 3. Monitoring

- Airflow memungkinkan pemantauan otomatis untuk memeriksa apakah pipeline berhasil dijalankan atau terjadi kegagalan.

### 4. Pembaruan Model

- Secara periodik, model prediksi dapat diperbarui dengan data baru untuk meningkatkan akurasi.



# Kesimpulan

Kesimpulan dari proyek ini adalah bahwa dengan mengintegrasikan data cuaca dan data penjualan melalui pipeline otomatis menggunakan Apache Airflow, kita dapat mengembangkan sistem prediksi yang lebih akurat dan efisien untuk memproyeksikan penjualan berdasarkan kondisi cuaca. Dengan memanfaatkan model machine learning yang dipadukan dengan feature engineering yang mempertimbangkan faktor-faktor cuaca seperti suhu dan curah hujan, bisnis dapat memperoleh wawasan yang lebih dalam tentang bagaimana cuaca mempengaruhi pola permintaan produk. Sistem ini dapat membantu dalam pengelolaan stok dan perencanaan strategi penjualan yang lebih responsif terhadap perubahan kondisi cuaca, sehingga mengurangi risiko kelebihan atau kekurangan stok, meningkatkan efisiensi operasional, dan pada akhirnya mendukung keputusan yang lebih baik dalam meningkatkan profitabilitas perusahaan.