**MODEL PEMBELAJARAN DAN LAPORAN AKHIR**

**PROJECT-BASED LEARNING**

**MATA KULIAH BIG DATA**

**KELAS A**



**“SPOTIFY ETL MENGGUNAKAN AIRFLOW DAN MONGODB”**

**DISUSUN OLEH KELOMPOK V :**

1. DIENNA ERIES LINGGARSARI ( 21083010092 ) - KETUA
2. AMALIA NUR RAMADHANI ( 21083010017 ) - ANGGOTA
3. RAHMALIA ANINDYA HERDIANTI ( 21083010084 ) - ANGGOTA
4. REGHINA AJENG SANTIKA ( 21083010086 ) - ANGGOTA
5. RIZAL HARJO UTOMO ( 21083010101 ) - ANGGOTA
6. M. AWALUDDIN ILHAM ZUHRI ( 21083010112 ) - ANGGOTA

**DOSEN PENGAMPU:**

TRESNA MAULANA FAHRUDIN, S.ST., MT  ( 199305012022031007 )

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

JAWA TIMUR

2023

# **DAFTAR ISI**

**DAFTAR ISI…………………………………………………………………………………………………………….i**

**DAFTAR GAMBAR…………….……………………………………………………………………………..…….ii**

1. **BAB I : PENDAHULUAN………………………………………………………………………………..….1**
   1. Latar belakang……………………………………………………………………………………………….1
   2. Permasalahan.…………………………...…………………………………………………………………..1
   3. Tujuan…………..……………………………………………………………………………………………….1
   4. Manfaat………...……………………………………………………………………………………………….1
2. **BAB II : TINJAUAN PUSTAKA……...………………………………………………………………..….3**
   1. Teori Penunjang..………………….………….…………………………………………………………….1
   2. Penelitian Terkait……….………………………………………………………………………………….1
3. **BAB III : METODOLOGI PENELITIAN…………..……………………………………………………8**
4. **BAB IV : HASIL DAN PEMBAHASAN……….………………………………………………………10**
5. **BAB V : KESIMPULAN…………………………………………………………………………………….17**
6. **DAFTAR PUSTAKA……………………..…………………………………………………………………18**
7. **LAMPIRAN…………………………………………………………………………………………………….20**

**DAFTAR GAMBAR**

**Gambar 1…………………………………………………………………………………………………………….6**

**Gambar 2…………………………………………………………………………………………………………….11**

**Gambar 3…………………………………………………………………………………………………………….16**

# **BAB I : PENDAHULUAN**

## Latar Belakang

Spotify merupakan salah satu platform streaming musik terbesar dan terkenal di dunia. Dengan jutaan pengguna aktif setiap harinya, Spotify perlu mengelola data secara efisien untuk memberikan pengalaman terbaik mendengarkan lagu kepada para pengguna. Oleh karena itu, diperlukan adanya solusi ETL (Extract, Transform, Load).

Dalam hal ini, Apache Airflow hadir sebagai platform open-source yang populer dalam pengelolaan alur kerja data. Dengan menggunakan Apache Airflow, Spotify dapat dengan mudah membuat dan menjalankan alur kerja ETL yang kompleks. Airflow juga menyediakan fitur pemantauan, penjadwalan, dan penanganan kesalahan yang penting dalam pengelolaan data skala besar. Selain itu, MongoDB, basis data NoSQL yang fleksibel, dipilih sebagai penyimpanan data yang cocok untuk Spotify. Dengan MongoDB, Spotify dapat mengelola data yang tak terstruktur dan dengan mudah meningkatkan skalabilitas sesuai kebutuhan.

Dengan adanya Project based learning ini, Spotify akan memperoleh manfaat berupa pengelolaan data yang efisien, analisis mendalam, dan fleksibilitas dalam mengubah dan memperluas arsitektur data mereka sesuai pertumbuhan bisnis.

## Permasalahan

Dari latar belakang yang telah diuraikan sebelumnya, identifikasi permasalahan yang ada pada penelitian ini yaitu :

1. Belum tersedianya solusi mengatasi permasalahan ketersediaan data dari Spotify untuk proses ETL menggunakan Airflow dan MongoDB.
2. Sulitnya melakukan integrasi yang efektif antara Airflow dan MongoDB dalam konteks ETL Spotify.
   1. Tujuan
3. Mencapai akses yang konsisten dan dapat diandalkan terhadap data Spotify serta memastikan ketersediaan data yang diperlukan untuk proses ETL dapat berjalan secara lancar dan efisien, sehingga tidak terganggu oleh kendala ketersediaan data.
4. Mencapai keterhubungan yang optimal antara kedua platform tersebut, Airflow dan MongoDB dapat saling berinteraksi secara lancar, sehingga memungkinkan transfer data yang efisien dan akurat dalam proses ETL dan menghasilkan data yang konsisten dan akurat.
   1. Manfaat
5. Otomatisasi Proses ETL : Airflow adalah alat manajemen aliran kerja yang kuat. Dalam konteks Spotify ETL, Airflow memungkinkan otomatisasi penuh dari ekstraksi, transformasi, dan pemrosesan data seiring waktu. Dengan Airflow, dapat dijadwalkan dan dijalankan tugas-tugas ETL secara tersusun, teroptimalkan penggunaan sumber daya, dan menghindari tugas-tugas yang berulang.
6. Meningkatkan pemahaman tentang Airflow dan MongoDB dalam konteks ETL : Penelitian ini memberikan wawasan lebih lanjut mengenai potensi penggunaan Airflow dan MongoDB dalam proses ETL, dikhuskan dalam konteks aplikasi musik seperti Spotify. Hal ini dapat memperkaya pemahaman mengenai kedua platform dan memberikan perspektif yang berharga bagi para praktisi dan peneliti pada bidang ini.
7. Menyediakan basis data yang terstruktur dan dapat diakses dengan mudah : Menggunakan MongoDB sebagai basis data, sehingga dapat menyimpan data musik dan metadata dalam struktur yang teroganisir dan fleksibel. Hal ini memungkinkan pengaksesan data yang cepat dan efisien, serta memudahkan pengembangan aplikasi atau analisis lebih lanjut.
8. Memungkinkan pengolahan data musik yang lebih kompleks : Airflow dan MongoDB dapat digunakan untuk memproses data musik yang beragam dan komples dari Spotify. Sehingga, hal ini membuka peluang untuk menganalisis data musik dengan lebih mendalam, termasuk metadata, preferensi pengguna, tren musik, dan lain sebagainya.
9. Mendukung pengembangan aplikasi dan layanan berbasis musik yang lebih canggih: Dengan memiliki ETL yang efektif menggunakan Airflow dan MongoDB, penelitian ini membuka peluang untuk mengembangkan aplikasi dan layanan baru yang berbasis data musik. Data yang diolah dan disimpan dalam MongoDB dapat digunakan untuk membuat rekomendasi musik yang lebih personal, analisis tren musik, pengenalan pola perilaku pengguna, dan pengembangan solusi kreatif lainnya.

**BAB II : TINJAUAN PUSTAKA**

2.1 Teori Penunjang

2.1.1. Spotify

Spotify adalah sebuah layanan penyiaran musik dan siniar yang berbasis di Stockholm, Swedia yang diluncurkan pada 7 Oktober 2008. Platform ini dimiliki oleh Spotify AB, sebuah perusahaan terbuka di Bursa Efek New York sejak 2018 dibawah perusahaan Spotify Technology S.A. yang berbasis di Luksemburg.

2.1.2 AirFlow

Airflow adalah platform sumber terbuka (open-source) yang digunakan untuk mengatur, menjadwalkan, dan mengelola aliran data (data pipelines). Dalam konteks teknologi data, aliran data merujuk pada perpindahan data dari satu sistem atau komponen ke sistem atau komponen lainnya

Airflow dirancang untuk membantu tim pengembangan dan operasi data dalam mengatur alur kerja yang kompleks dan berulang secara efisien. Dengan menggunakan Airflow, Anda dapat dengan mudah membuat, menjadwalkan, dan memantau tugas-tugas yang terkait dengan pengolahan data, seperti pemrosesan data batch, ekstraksi, transformasi, dan pemuatan (ETL), analisis data, atau tugas-tugas lain yang memerlukan alur kerja terstruktur

2.1.3 Mongo Db

MongoDB adalah sistem manajemen basis data (database management system/DBMS) yang bersifat dokumen (document-oriented) dan berbasis NoSQL. Dalam MongoDB, data disimpan dalam dokumen yang terstruktur dalam format BSON (Binary JSON), yang memungkinkan fleksibilitas dan skalabilitas yang tinggi.

2.1.4 Python

Python adalah bahasa pemrograman tingkat tinggi yang populer dan serbaguna. Dikembangkan oleh Guido van Rossum pada tahun 1991, Python dirancang dengan filosofi yang mengutamakan keterbacaan kode (readability) dan sintaks yang bersih, sehingga membuatnya mudah dipahami dan ditulis oleh para pengembang.

2.2 Penelitian Terkait

1. **Erwin Eko Wahyudi (2022)** dengan judul *“Akuisisi Data Prediksi Curah Hujan Secara Periodik Menggunakan Apache Airflow”.*

Terdapat beberapa faktor yang mempengaruhi terjadinya tanah longsor, seperti curah hujan. Ketika hujan semakin deras, curah hujan akan semakin tinggi semakin tinggi curah hujan, maka peluang terjadinya tanah longsor akan semakin besar. Infiltrasi air hujan dapat memicu longsor karena tanah menjadi jenuh dan meningkatkan tekanan air pori dalam tanah*.* jurnal penelitian ini akan memproses pemodelan, seperti pembuatan model untuk memprediksi potensi terjadinya tanah longsor. Data prediksi curah hujan yang disediakan oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dapat digunakan untuk pemodelan tersebut. Data akan disimpan di komputer lokal dengan menggunakan alat atau aplikasi otomasi yang bernama Apache Airflow. Hasil dari penelitian ini adalah terdapat pipeline pada aplikasi otomasi Apache Airflow untuk membantu proses akuisisi data secara periodik.

1. **Dion Ricky Saputra (2022)** dengan judul *“SISTEM PROPAGASI ANOTASI PADA METADATA LINEAGE UNTUK MANAJEMEN DATA WAREHOUSE”.*

Data yang berada di data warehouse berasal dari beragam sumber data baik didalam maupun diluar organisasi. Data yang pada mulanya berada di sumber data akan dipindahkan melalui proses ekstraksi, transformasi, dan loading atau disebut juga dengan proses ETL. Sulitnya melakukan propagasi anotasi dengan tepat diakibatkan karena tidak adanya informasi jalur perpindahan data. Pengujian sistem dilakukan menggunakan teknik black box dan user acceptance testing bersama pengguna akhir. Pengujian blackbox menghasilkan 30 kasus uji yang hasilnya valid. Hasil evaluasi user acceptance testing menunjukkan bahwa rata-rata pengguna menyatakan sangat setuju dengan sistem yang dikembangkan.

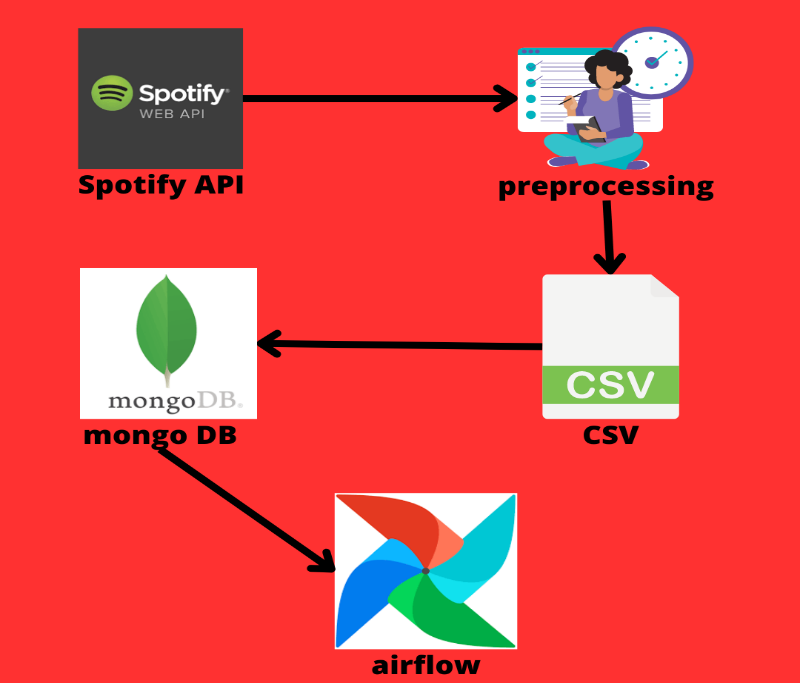
1. **I Nyoman Adnyana Putra (2019)** dengan judul *“OPTIMASI PROSES ETL DENGAN METODE HEURISTIK UNTUK MEMBANGUN DATA WAREHOUSE”.*

Perkembangan global teknologi informasi saat ini berperan penting dalam dunia bisnis baik swasta maupun pemerintahan. Dalam dunia pemerintahan, Pemerintah Provinsi Bali yang saat ini masih melakukan proses pelaporan realisasi pengadaan secara manual, dapat memanfaatkan teknologi informasi untuk mendigitalisasi laporan tersebut. Data warehouse merupakan salah satu konsep pengolahan data yang baik dengan sumber data yang dapat berasal dari banyak sistem. Menurut beberapa pakar, sekitar 60%-80% upaya dalam pembangunan data warehouse dihabiskan untuk proses ETL. Proses ETL yang menangani jumlah data yang banyak serta kerja yang rumit, sehingga memerlukan waktu penyelesaian yang lama dan membutuhkan memori yang besar. Dalam penelitian ini akan dilakukan modifikasi untuk mengurangi jumlah baris data yang diproses dengan mengatur ulang transformasi pada proses ETL dengan menggunakan metode heuristik. Hasil yang didapatkan dari penelitian ini adalah optimasi ETL menggunakan metode heuristik dapat memaksimalkan waktu dan beban memory saat proses ETL. Dari tiga sumer data yang digunakan dengan 17.024 jumlah data yang diolah, selisih waktu antara ETL sebelum optimasi dengan yang di optimasi mencapai 103 menit.

**BAB III : METODOLOGI PENELITIAN**

3.1 Desain dan Rancangan Penelitian

Untuk memahami alur dari penelitian, akan disajikan flowchart penelitian kami dibawah ini.



**Gambar 1**

Untuk melakukan penelitian ini, diperlukan beberapa tahapan yang harus dilaksanakan dengan tujuan penelitian berjalan seperti yang telah diharapkan dan direncanakan seperti:

3.1.1 Crawling Data

Tahap ini dilakukan untuk mengumpulkan data dari sumber yang dapat diakses melalui API Spotify, seperti situs web musik, forum, atau sumber data lainnya. Dengan ini kita dapat memperoleh informasi tambahan yang berguna untuk penelitian.

3.1.2 Pra-processing

Pra-processing merupakan tahap untuk pembersihan dari data yang tidak relevan atau tidak konsisten agar data menjadi lebih spesifik dan lebih mudah untuk diproses.

3.1.3. Data Transform

Tahap ini merupakan proses merubah data yang telah diambil menjadi CSV setelah menetapkan variabel, struktur, dan lain-lain agar dapat digunakan dalam penelitian ini.

3.1.4.  Database

Pada tahap ini database yang digunakan adalah mongo DB. Database digunakan untuk menyimpan data sebelum dieksekusi dengan Airflow.

3.1.5.  Real-time Data

Tahap ini digunakan Airflow untuk menjadwalkan dan menjalankan alur kerja ETL yang kompleks. Pada penelitian ini Airflow berintegrasi dengan mongoDB sehingga memudahkan untuk menghubungkan alur kerja ETL Spotify dengan mongoDB.

# **BAB IV : HASIL DAN PEMBAHASAN**

4.1 Authenticate spotify API



Proses pertama yang dilakukan adalah import modul yang digunakan, seperti spotify, pandas, dotenv, dan os. pada proses ini dilakukan autentikasi spotify API.

4.2 Crawling Data

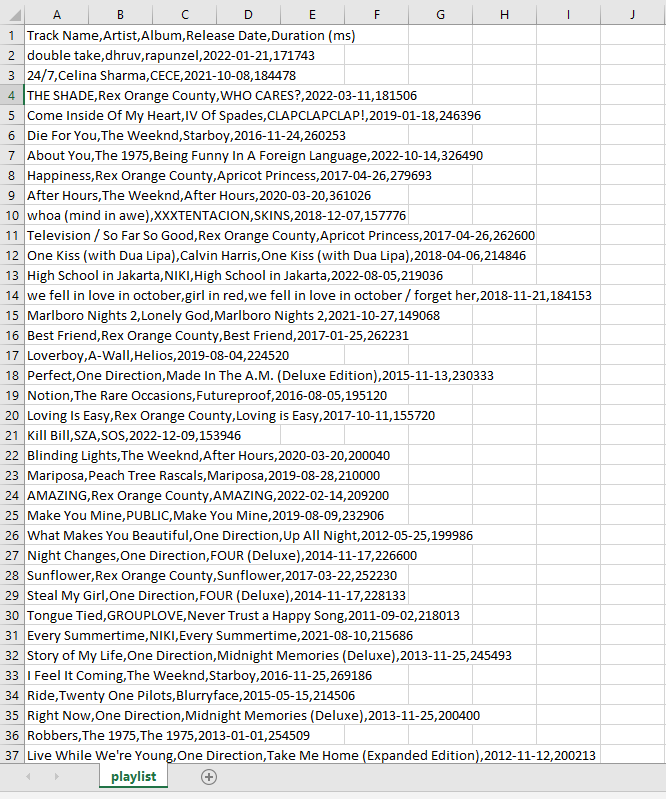




Pada proses ini dilakukan crawling data dari spotify yang berupa tracks dari sebuah playlist. Sekaligus melakukan tranformasi data.

4.3 Ekstrak data menjadi CSV

Pada proses ini data yang diambil dari spotify di extrak menjadi dataset berupa file csv. Output dari code script tersebut berupa file csv seperti gambar dibawah ini.





4.4 Load data ke dalam database MongoDB



Code di atas merupakan contoh implementasi pengambilan data lagu dari playlist Spotify, dan penyimpanan data ke dalam MongoDB serta ekspor data ke dalam file CSV. Berikut adalah penjelasan untuk setiap bagian code tersebut:

* + mongo\_url = 'mongodb://localhost:27017/', mongo\_db = 'spootif-etl-db', dan mongo\_collection = 'tracks': Menentukan URL MongoDB, nama basis data, dan nama koleksi yang akan digunakan untuk menyimpan data.
  + Fungsi Load: def load\_tracks\_to\_mongodb(transformed\_tracks): Fungsi ini melakukan koneksi ke MongoDB menggunakan URI yang telah ditentukan dan memasukkan data yang telah ditransformasi ke dalam koleksi yang ditentukan. Sebelum memasukkan data baru, fungsi ini juga menghapus data yang sudah ada dalam koleksi.
  + Fungsi Export: def export\_data\_to\_csv(): Fungsi ini melakukan ekspor data dari MongoDB ke dalam file CSV. Data dari koleksi MongoDB diambil menggunakan find() dan disimpan dalam DataFrame menggunakan Pandas. DataFrame tersebut kemudian diekspor ke dalam file CSV dengan menggunakan to\_csv

4.5 Connect Data to Airflow

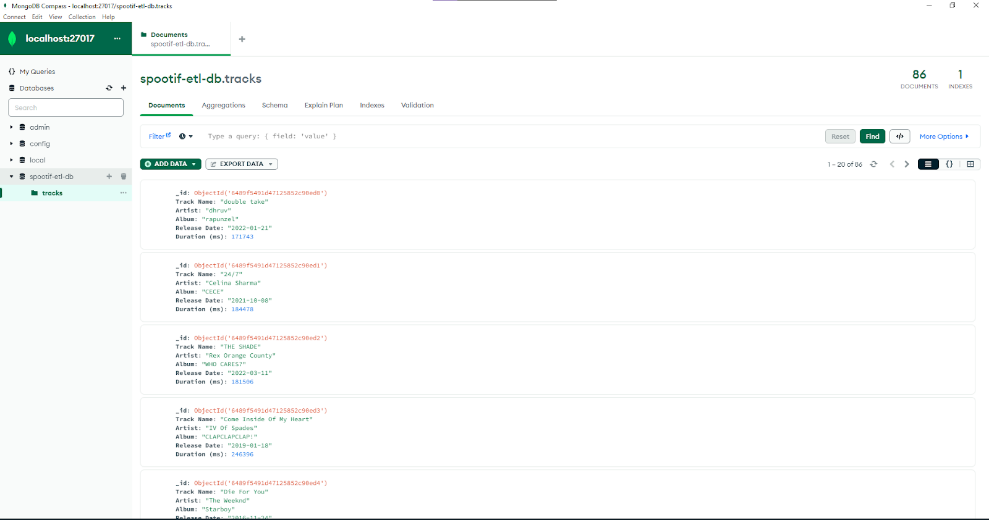






* from airflow import DAG dan from airflow.operators.python import PythonOperator: Mengimpor modul dan kelas yang diperlukan dari library Apache Airflow untuk membuat DAG dan operator Python.
* Membuat DAG (Directed Acyclic Graph): with DAG('spotify\_etl\_dag', default\_args=default\_args, schedule\_interval='@daily') as dag: Membuat DAG dengan nama 'spotify\_etl\_dag' dan menggunakan default\_args serta jadwal interval harian.
* default\_args: Mendefinisikan argumen default untuk DAG seperti pemilik (owner), tanggal mulai, retries, dan retry delay.
* def run\_spotify\_etl(): Mendefinisikan fungsi yang akan dijalankan. namun pada bagian ini code script yang telah dibuat belum bisa connect dengan DAG dari apache Airflow.

Dengan menggunakan Airflow, proses ETL ini dapat diatur untuk berjalan secara otomatis sesuai jadwal yang telah ditentukan. Hal ini memungkinkan pembaruan data playlist lagu secara otomatis di MongoDB dan file CSV saat playlist di Spotify diperbarui.



# **BAB V : KESIMPULAN**

Pada project ini kami yang berjudul Spotify ETL Menggunakan Apache Airflow dan MongoDB mempunyai tujuan menggambarkan data dari spotify dapat diambil di transformasikan dimuat di dalam database dan atur dengan bantuan alur kerja Airflow untuk menjalankan proses ETL secara terjadwal dan terkontrol

Sedangkan untuk melakukan penelitian ini diperlukan beberapa tahap yang harus dilaksanakan. Yang pertama ada proses crawling data yang pada tahap ini kita mengambil data dari spotify untuk diproses. Tahap selanjutnya ada pra-processing data yang merupakan tahap untuk membersihkan dari data yang tidak relevan. Kemudian, yang ketiga yaitu tahap data transform proses merubah data yang telah diambil. Lalu tahap selanjutnya yaitu ada database data yang digunakan ntuk menyimpan data sebelum di eksekusi. Dan yang terakhir, adalah tahap real-time data yang digunakan untuk menjadwalkan dan menjalankan alur kerja ETL yang kompleks.

Hasil yang diperoleh dari project ini, dengan menggunakan Airflow, proses ETL ini dapat diatur untuk berjalan secara otomatis sesuai jadwal yang telah ditentukan. Hal ini memungkinkan pembaruan data playlist lagu secara otomatis di MongoDB saat playlist di Spotify diperbarui.

Keseluruhan hasil ini mencerminkan berhasilnya implementasi proses ETL dari Spotify API ke MongoDB dengan bantuan Airflow. Data playlist lagu dapat diolah, disimpan, dan diekspor dengan menggunakan teknologi yang relevan, sehingga memungkinkan analisis lebih lanjut atau penggunaan data tersebut sesuai kebutuhan.

**DAFTAR PUSTAKA**

Abanto, J. (2004, 12). *Airflow modelling in a computer room*. Retrieved from sciencedirect: https://www.sciencedirect.com/science/article/abs/pii/S0360132304001167

Anderson, A. (2020, april). *Conference: WWW '20: The Web Conference 2020*. Retrieved from ResearchGate : https://www.researchgate.net/publication/341126150\_Algorithmic\_Effects\_on\_the\_Diversity\_of\_Consumption\_on\_Spotify#:~:text=Recent%20research%20commissioned%20by%20Spotify%20%28Anderson%20et%20al.,away%20from%20such%20recommendations%2C%20their%20consumptio

Hardi, N. (Volume 10, Nomor 1, Januari 2021). Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier. *SISTEMASI: Jurnal Sistem Informasi*, 131-138.

Mahawardana, P. P. ( Vol. 3, No. 1 April 2022). Analisis Sentimen Berdasarkan Opini dari Media Sosial Twitter terhadap “Figure Pemimpin” Menggunakan Python. *JITTER- Jurnal Ilmiah Teknologi dan Komputer*.

Pratama, A. Y. (Volume 5 Nomor 2, September 2021). Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja). *Jurnal Sains Komputer & Informatika (J-SAKTI)* , pp. 897-910.

*Programming Python, 4th Edition*. (2010, Desember). Retrieved from Publisher(s): O'Reilly Media, Inc.: https://www.oreilly.com/library/view/programming-python-4th/9781449398712/

Renaldi. (2020, 2). *JII: Jurnal Inovasi Informatika Universitas Pradita*. Retrieved from media.neliti: https://media.neliti.com/media/publications/465239-none-c7412118.pdf

Yoga, I. K. (Vol 1 No. 3 , 2018). IMPLEMENTASI KONSEP CREATING SHARED VALUE (CSV) SEBAGAI PROGRAM CORPORATE SOCIAL RESPONSIBILITY (CSR) DALAM PENINGKATAN KESEJAHTERAAN STAKEHOLDER. *Pactum Law Journal*, ISSN: 2615-7837. Retrieved from http://repository.lppm.unila.ac.id/9091/1/IMPLEMENTASI%20KONSEP%20CREATING%20SHARED%20VALUE%20%28CSV%29%20SEBAGAI%20PROGRAM%20CORPORATE%20SOCILA%20RESPONBILLITY%20%28CSR%29%20DALAM%20PENINGKATAN%20KESEJAHTERAAN%20STAKEHOLDER.pdf

**LAMPIRAN**

Tautan source code python :

<https://colab.research.google.com/drive/16sUBwYTSuZIEUBBQNEKyVVmR1k2o8voC?usp=sharing>

Tautan vidio youtube : <https://youtu.be/eHr8-lXmOWg>