E-ISSN : 3025-1311
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 10 No 7 Tahun 2025

PENERAPAN BIG DATA LIFECYCLE UNTUK ANALISIS TREN GENRE DAN RATING FILM MENGGUNAKAN HADOOP ECOSYSTEM DI AZURE

Muhammad Zaidan Rafat[1], Ratna Dwi Puspita[2], Fawwaz Muzakty[3], Jessica Aulia Sheril Pulungan[4], Ahmad Tabrani[5]

Program Studi Informatika, Fakultas Sains UIN Sultan Maulana Hasanuddin Banten
Jl. Syech Nawawi Al-Bantani, Kelurahan Sukawana, Kecamatan Curug, Kota Serang, Banten

Email: muhammadzaidanrafat@gmail.com *, nanadwipuspita10@gmail.com [2], fawazmuzaki3@gmail.com [3], Jessicaaulia35@gmail.com [4], Ahmad.tabrani@uinbanten.ac.id [5]

Abstrak

Pertumbuhan industri film digital pada platform seperti IMDb telah menimbulkan tantangan kelebihan informasi bagi pengguna dalam memilih konten. Analisis berbasis popularitas yang lazim diterapkan sering mengabaikan preferensi individu dan kurang efektif bagi pengguna baru tanpa riwayat interaksi, sehingga diperlukan pendekatan analisis yang lebih tepat sasaran. Penelitian ini menerapkan kerangka Big Data Lifecycle, mencakup tahap pengumpulan, pembersihan, integrasi, analisis, dan representasi data. Dataset sekunder dari IMDb berukuran 8,83 GB diolah menggunakan Hadoop Ecosystem pada platform Azure, dengan Apache Spark untuk pemrosesan paralel dan Power BI untuk visualisasi. Proses dimulai dari pengunduhan data ke Azure Blob Storage, diikuti pembersihan nilai kosong dan penyaringan entri film, integrasi berdasarkan kunci tconst, analisis agregat dengan fungsi explode untuk genre, hingga pembuatan dashboard interaktif. Hasil analisis menunjukkan tren produksi film mencapai puncak pada dekade 2000-an dengan sekitar 20.000 entri, diikuti penurunan tajam pasca-2010. Genre Drama mendominasi dengan lebih dari 20.000 film dan akumulasi rating tertinggi, sementara rating rata-rata stabil antara 6 hingga 8 sepanjang periode. Hubungan durasi film dengan rating memperlihatkan pola positif lemah, dengan film panjang cenderung dinilai lebih tinggi pada genre tertentu. Pendekatan ini membuktikan kemampuan Big Data Lifecycle dalam menghasilkan wawasan empiris mengenai tren industri film, yang dapat mendukung pengurangan kelebihan informasi pada platform digital tanpa ketergantungan data interaksi pengguna.

Kata Kunci : Big Data Lifecycle, Hadoop Ecosystem, Azure, analisis tren film, IMDb.

Abstract

The growth of the digital film industry on platforms such as IMDb has posed challenges of information overload for users in selecting content. Popularity-based analyses commonly applied often overlook individual preferences and are less effective for new users without interaction histories, necessitating a more targeted analytical approach. This study implements the Big Data Lifecycle framework, encompassing stages of data collection, cleansing, integration, analysis, and representation. Secondary dataset from

Article History
Received: Desember 2025
Reviewed: Desember 2025
Published: Desember 2025

Plagiarism Checker No 794
Prefix DOI : Prefix DOI : 10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

IMDb sized 8.83 GB is processed using the Hadoop Ecosystem on the Azure platform, with Apache Spark for parallel processing and Power BI for visualization. The process begins with data download to Azure Blob Storage, followed by cleansing of missing values and filtering of movie entries, integration based on tconst key, aggregate analysis with explode function for genres, and creation of an interactive dashboard. The analysis results indicate film production trends peaking in the 2000s decade with approximately 20,000 entries, followed by a sharp decline post-2010. The Drama genre dominates with over 20,000 films and the highest rating accumulation, while average ratings remain stable between 6 and 8 over the period. The relationship between film duration and ratings shows a weak positive pattern, with longer films tending to receive higher ratings in certain genres. This approach demonstrates the capability of Big Data Lifecycle in generating empirical insights into film industry trends, which can support the reduction of information overload on digital platforms without reliance on user interaction data.

***Keywords**: Big Data Lifecycle, Hadoop Ecosystem, Azure, film trend analysis, IMDb.*

1. PENDAHULUAN

Seiring dengan pesatnya perkembangan teknologi digital, industri film telah mengalami transformasi signifikan, di mana platform distribusi seperti IMDb menjadi pusat utama bagi jutaan pengguna di seluruh dunia. Dalam penelitian yang dilakukan oleh Jadiga (2024) menyatakan bahwa "Pertumbuhan yang luar biasa dari industri film digital di platform seperti IMDb telah menghasilkan tumpukan data yang sangat besar yang memerlukan analisis mendalam." (Jadiga, 2024) Hal ini mencerminkan bagaimana IMDb, sebagai salah satu platform basis data film berbasis komputer, tidak hanya menawarkan jutaan judul film tetapi juga menghadapi peningkatan tanggapan negatif akibat ketidakpuasan pengguna terhadap layanan, sebagaimana diuraikan dalam studi (Gupta et al., 2025). Pengguna sering kali mengalami kelebihan informasi atau information overload, di mana banjirnya pilihan film membuat mereka kesulitan menemukan konten yang sesuai dengan preferensi pribadi. Kondisi ini semakin diperburuk oleh fakta bahwa platform seperti IMDb memiliki jutaan entri film, menghasilkan data besar yang mencakup perilaku rating, ulasan, dan preferensi, sebagaimana dicatat oleh Oussous et al. (2018).(Oussous et al., 2018)

Masalah information overload ini bukan hanya sekadar kebingungan sementara, melainkan hambatan utama dalam pengambilan keputusan pengguna. Pengguna sering kali bergantung pada analisis berbasis popularitas, seperti film terlaris atau yang paling banyak diulas, tetapi pendekatan ini memiliki keterbatasan signifikan karena tidak mempertimbangkan preferensi individu. Namun, analisis semata- mata berdasarkan popularitas sering kali mengabaikan atribut spesifik film seperti genre, deskripsi, atau fitur unik, yang pada akhirnya menyebabkan analisis kurang relevan dan menurunkan kepuasan pengguna.

Oleh karena itu, diperlukan sistem analisis yang lebih cerdas untuk membantu pengguna menavigasi lautan konten digital ini, sehingga meningkatkan pengalaman menonton dan mendukung pertumbuhan industri film secara berkelanjutan. Sistem analisis pada dasarnya adalah alat yang dirancang untuk menyaring informasi berdasarkan preferensi pengguna, dengan tujuan menyajikan insight yang paling

relevan. Dalam konteks platform film, sistem ini memainkan peran krusial dalam meningkatkan retensi pengguna dan penjualan, karena dapat mengurangi waktu pencarian dan meningkatkan kepuasan. Aljunid & Dh (2021) menjelaskan bahwa "Sistem analisis merupakan alat yang dirancang untuk menyaring informasi berdasarkan preferensi pengguna,"(Awan et al., 2021) di mana pendekatan ini telah terbukti efektif dalam mengatasi masalah kelebihan pilihan pada platform hiburan digital. Berbeda dengan pendekatan user-based yang bergantung pada kesamaan antar pengguna, penelitian ini berfokus pada analisis berbasis item, yang menekankan kemiripan antar item film itu sendiri.

Pendekatan ini lebih cocok untuk dataset film di IMDb, di mana atribut konten seperti genre, tahun rilis, dan rating dapat dieksploitasi untuk menghasilkan insight tanpa memerlukan data interaksi pengguna yang masif, sehingga lebih efisien dalam skenario di mana data pengguna baru terbatas.

Pendekatan Big Data Lifecycle menjadi pilihan utama dalam membangun sistem analisis ini, karena konsep dasarnya yang menganalisis atribut intrinsik dari item untuk menemukan pola tren. Alasan pemilihan Big Data Lifecycle adalah ketidakbergantungannya pada data rating pengguna, yang sering kali tidak lengkap atau bias pada platform seperti IMDb. Tantangan utama dalam menerapkan Big Data Lifecycle pada data film adalah dimensi tinggi dari fitur yang ada, di mana jutaan atribut seperti rating dan genre dapat menyebabkan inefisiensi komputasi dan kebisingan data. Untuk mengatasi hal ini, penggunaan Hadoop Ecosystem menjadi solusi yang efektif.(Jadiga, 2024) Hal ini dapat diartikan bahwa Hadoop Ecosystem dapat memproses data secara terdistribusi sambil menjaga informasi utama. Hadoop Ecosystem bekerja dengan mentransformasi data ke cluster cloud seperti Azure, sehingga mengurangi redundansi dan meningkatkan efisiensi komputasi, terutama pada data film IMDb yang melibatkan vektor berdimensi tinggi dari atribut konten. Setelah pemrosesan, pengukuran tren dilakukan menggunakan Apache Spark, yang merupakan metrik yang mengukur pola antar data.(Oussous et al., 2018) Alasan pemilihan Spark adalah kesesuaiannya dengan data besar, di mana fokus pada pemrosesan paralel memungkinkan deteksi tren yang lebih akurat, bahkan pada dataset dengan variasi seperti rating atau produksi film.

Tujuan utama penelitian ini adalah menerapkan Big Data Lifecycle yang mengintegrasikan Hadoop Ecosystem untuk analisis tren genre dan rating film pada data IMDb, sehingga menghasilkan insight yang lebih efisien dan relevan. Kontribusi utama dari penelitian ini meliputi implementasi sistem analisis yang inovatif dengan integrasi Spark dan Azure, yang belum banyak dieksplorasi dalam konteks film, serta studi kasus pada data IMDb yang memberikan wawasan praktis bagi pengembang platform. Dengan demikian, sistem analisis yang dikembangkan berpotensi menjadi solusi alternatif untuk mengurangi information overload pada platform distribusi film digital.

Penelitian ini juga menyoroti keterbatasan pendekatan tradisional, seperti yang terlihat dalam prediksi tren film, di mana Gupta et al. (2025) menyatakan bahwa "Dengan menggunakan algoritma komputasional, tujuan penelitian ini adalah untuk membangun model yang dapat memprediksi tren rating film".(Gupta et al., 2025) Namun, pendekatan berbasis popularitas ini sering mengabaikan kemiripan konten, sehingga integrasi Big Data Lifecycle dan Hadoop Ecosystem menawarkan alternatif yang lebih holistik. Akhirnya, kontribusi ini diharapkan dapat memperkaya literatur tentang sistem analisis di platform film, dengan implikasi praktis untuk mengurangi kelelahan informasi.

2. METODOLOGI

2.1 Pendekatan dan Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan desain analisis data sekunder berskala besar. Pendekatan ini dipilih karena penelitian berfokus pada pengolahan dan peringkasan data numerik untuk menggambarkan pola serta kecenderungan produksi film berdasarkan genre dan tahun rilis.

Desain penelitian disusun dengan mengacu pada kerangka Big Data Lifecycle, yang mencakup tahapan pengumpulan data, pembersihan data, integrasi data, analisis data, dan penyajian hasil. Kerangka ini digunakan karena sesuai dengan karakteristik dataset IMDb yang memiliki volume besar, variasi atribut yang beragam, serta membutuhkan pemrosesan terdistribusi agar analisis dapat dilakukan secara efisien dan konsisten.(Zaharia et al., n.d.)

2.2 Sumber Data

Dataset yang digunakan pada penelitian ini diperoleh dari platform Kaggle, dengan judul "Movie Dataset IMDb" yang berasal dari basis data Internet Movie Database (IMDb). Dataset ini telah dikonversi ke format CSV untuk kompatibilitas dengan sistem pemrosesan Apache Spark. Ukuran total data mencapai sekitar 8,83 GB, yang mencakup jutaan entri film dan memenuhi kriteria data besar dari segi volume.

Dataset terdiri dari beberapa berkas utama, yaitu:

1) converted_title.basics.csv (metadata film: judul, tahun rilis, genre, durasi).
2) converted_title.ratings.csv (rating rata-rata dan jumlah vote film).
3) converted_title.crew.csv (data sutradara dan penulis).

Seluruh dataset disimpan pada Azure Blob Storage sebagai data lake untuk mendukung penyimpanan terpusat dan akses terdistribusi.(Armbrust et al., 2020)

2.3 Atribut Data yang Digunakan

Dalam penelitian ini, hanya atribut yang relevan dengan pendekatan Big Data Lifecycle yang digunakan, antara lain: tconst sebagai identitas unik film, primaryTitle sebagai nama film, startYear sebagai tahun rilis, genres sebagai genre film, averageRating sebagai rating rata-rata, serta numVotes sebagai jumlah suara penilaian. Atribut berbasis interaksi pengguna seperti rating individu tidak digunakan, sehingga sistem tidak bergantung pada data pengguna. Atribut-atribut tersebut disajikan dalam bentuk tabel sebagai berikut:

Atribut	Deskripsi
Tconst	Identitas unik film
PrimaryTitle	Nama film
StartYear	Tahun rilis
Genres	Genre film
AverageRating	Rating rata-rata
NumVotes	Jumlah suara penilaian

Tabel 1. Atribut Data

2.4 Pembersihan Data dan Integrasi

2.4.1 Pembersihan Data

Tahapan pembersihan dilakukan untuk menyiapkan data agar dapat diproses secara numerik, meliputi penghapusan nilai kosong (\N) pada kolom startYear, runtimeMinutes, dan genres, penyesuaian tipe data ke integer/double sesuai konteks, serta penyaringan hanya entri dengan

titleType = 'movie'. Proses ini dilakukan menggunakan PySpark pada lingkungan Azure Databricks. Dalam penelitian yang dilakukan oleh Oussous et al. (2018), dinyatakan bahwa "Proses pre-processing menggunakan transformasi dilakukan dengan tujuan untuk menstandarisasi data, yaitu mengubah nilai-nilai pada atribut menjadi distribusi dengan rata-rata 0 dan deviasi standar 1".(Oussous et al., 2018) Berdasarkan pernyataan tersebut, dapat disimpulkan bahwa proses ini memastikan data bersih dan siap untuk analisis selanjutnya, sehingga meningkatkan kualitas representasi numerik dari atribut kategorikal. Hasil pembersihan disimpan dalam format Parquet untuk efisiensi penyimpanan dan akses data. Kode PySpark untuk pembersihan data:

```
Nov 17, 2025 (1s)                                                    10

from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType, FloatType

# 1. Cleaning df_basics
df_basics_clean = df_basics.replace("\\N", None) \
    .withColumn("startYear", col("startYear").cast(IntegerType())) \
    .withColumn("runtimeMinutes", col("runtimeMinutes").cast(IntegerType()))

# 2. Cleaning df_ratings → df_ratings_clean2
df_ratings_clean2 = df_ratings.replace("\\N", None) \
    .withColumn("averageRating", col("averageRating").cast(FloatType())) \
    .withColumn("numVotes", col("numVotes").cast(IntegerType()))

# 3. Cleaning df_crew
df_crew_clean = df_crew.replace("\\N", None)

print("Dataframe cleaned ready!")
df_basics_clean.printSchema()
df_ratings_clean2.printSchema()
df_crew_clean.printSchema()

> df_basics_clean: pyspark.sql.connect.dataframe.DataFrame = [tconst: string, titleType: string ... 7 more fields]
> df_crew_clean: pyspark.sql.connect.dataframe.DataFrame = [tconst: string, directors: string ... 1 more field]
> df_ratings_clean2: pyspark.sql.connect.dataframe.DataFrame = [tconst: string, averageRating: float ... 1 more field]
root
|-- tconst: string (nullable = true)
|-- titleType: string (nullable = true)
|-- primaryTitle: string (nullable = true)
```

Gambar 1 Kode Pembersihan Data

2.4.2 Integrasi Data

Integrasi dilakukan dengan menggabungkan dataset metadata film dan dataset penilaian berdasarkan kunci utama tconst menggunakan Spark SQL. Kolom hasil gabungan meliputi: primaryTitle, startYear, runtimeMinutes, genres, averageRating, numVotes. Proses ini menghasilkan dataset terpadu yang menjadi sumber utama untuk analisis. Pendekatan ini mirip dengan integrasi data movie rating pada Hadoop untuk analisis tren film (Aljunid & Dh, 2021).(Awan et al., 2021) Hasil integrasi disimpan kembali ke Blob Storage sebagai dataset utama. Secara matematis, integrasi data dapat dinyatakan sebagai operasi join:

$$D = title.basics \bowtie_{tconst} title.ratings$$

Dengan D sebagai dataset terintegrasi.

2.5 Analisis Data Menggunakan Spark SQL

2.5.1 Tujuan Analisis

Analisis dilakukan untuk mengidentifikasi pola dan tren dari data besar menggunakan ekosistem Hadoop. Karena dataset memiliki volume yang tinggi dan bersifat sparsity, Spark SQL digunakan untuk explode kolom genres, pengelompokan data berdasarkan startYear dan genre, serta perhitungan jumlah film per genre per tahun (COUNT(*) AS n_movies) dan rata-rata rating per genre per tahun (AVG(averageRating) AS avg_rating). Proses ini

memanfaatkan pemrosesan paralel Spark, sebagaimana direkomendasikan dalam survei teknologi Big Data (Oussous et al., 2018).(Oussous et al., 2018)

2.5.2 Formulasi Analisis

Untuk menganalisis tren genre dan rating film, langkah pertama adalah memisahkan kolom genres yang berisi daftar nilai (dipisah koma) menjadi entri terpisah menggunakan fungsi explode. Hal ini memungkinkan setiap genre dihitung secara independen tanpa kehilangan informasi film asli.(Prabaswara & Saputra, 2020) Selanjutnya, data dikelompokkan berdasarkan tahun rilis (startYear) dan genre untuk menghitung jumlah film per genre per tahun serta rata-rata rating per genre per tahun.

Hasil query ini diekspor ke Azure Blob Storage untuk tahap visualisasi selanjutnya. Proses ini memastikan analisis dapat dilakukan secara efisien pada data besar dengan memanfaatkan kemampuan pemrosesan paralel Spark.

2.6 Representasi Hasil Menggunakan Power BI

Power BI Desktop digunakan untuk menyajikan hasil analisis dalam bentuk visual interaktif.(Sabrina et al., 2024) Data sumber dari file genre_trend_csv di Blob Storage divisualisasikan melalui:

1) Line Chart: Rata-rata rating per genre per tahun.
2) Bar Chart: Jumlah film per genre.
3) Heatmap: Korelasi antara genre, tahun, dan rating.

Visualisasi ini mempermudah interpretasi tren genre, distribusi produksi film, serta hubungan antara jumlah film dan nilai rating.

3. HASIL DAN PEMBAHASAN

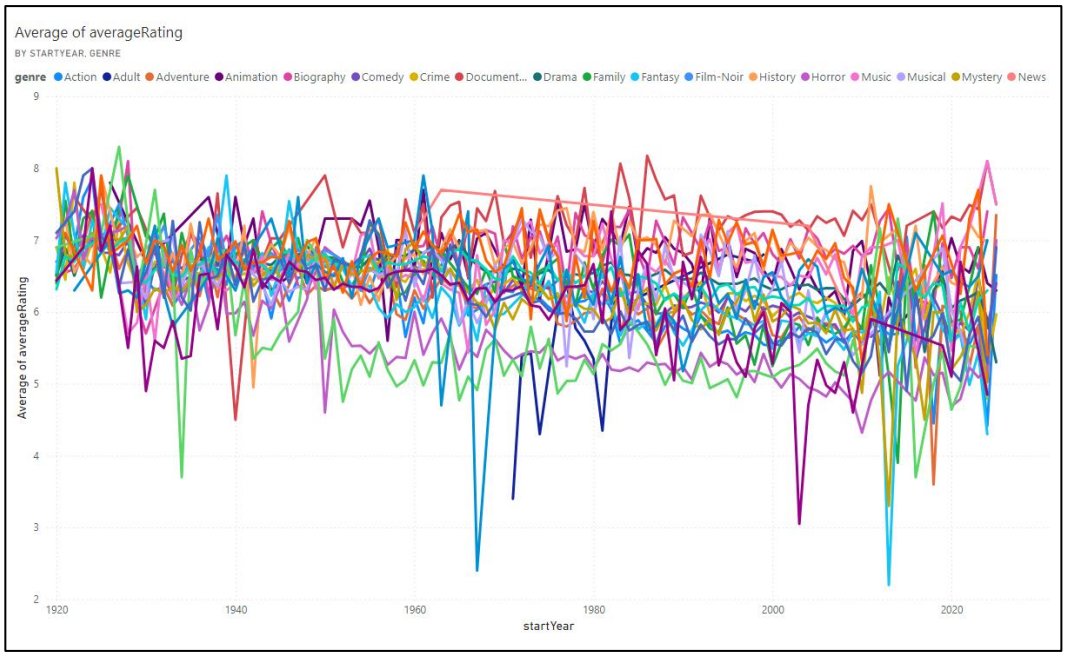
3.1 Hasil Pengumpulan dan Pembersihan Data

Tahapan pengumpulan data berhasil memperoleh dataset IMDb Movies Dataset (Converted) dari platform Kaggle dengan ukuran total sekitar 8,83 GB. Dataset ini mencakup tiga berkas utama: converted_title.basics.csv, converted_title.ratings.csv, dan converted_title.crew.csv. Setelah proses pembersihan menggunakan PySpark pada Azure Databricks, nilai kosong (\N) pada kolom penting seperti startYear, runtimeMinutes, dan genres dihapus atau diganti dengan null.

Tipe data disesuaikan menjadi integer untuk tahun dan runtime, serta float untuk rating. Selain itu, dilakukan penyaringan untuk hanya menyertakan entri dengan titleType = 'movie', numVotes ≥ 500 (untuk kredibilitas rating), serta runtimeMinutes antara 40–300 menit (untuk menghilangkan outlier). Hasil pembersihan menghasilkan dataset terpadu yang siap dianalisis, dengan jutaan entri film yang valid.(Moeslim, 2025)

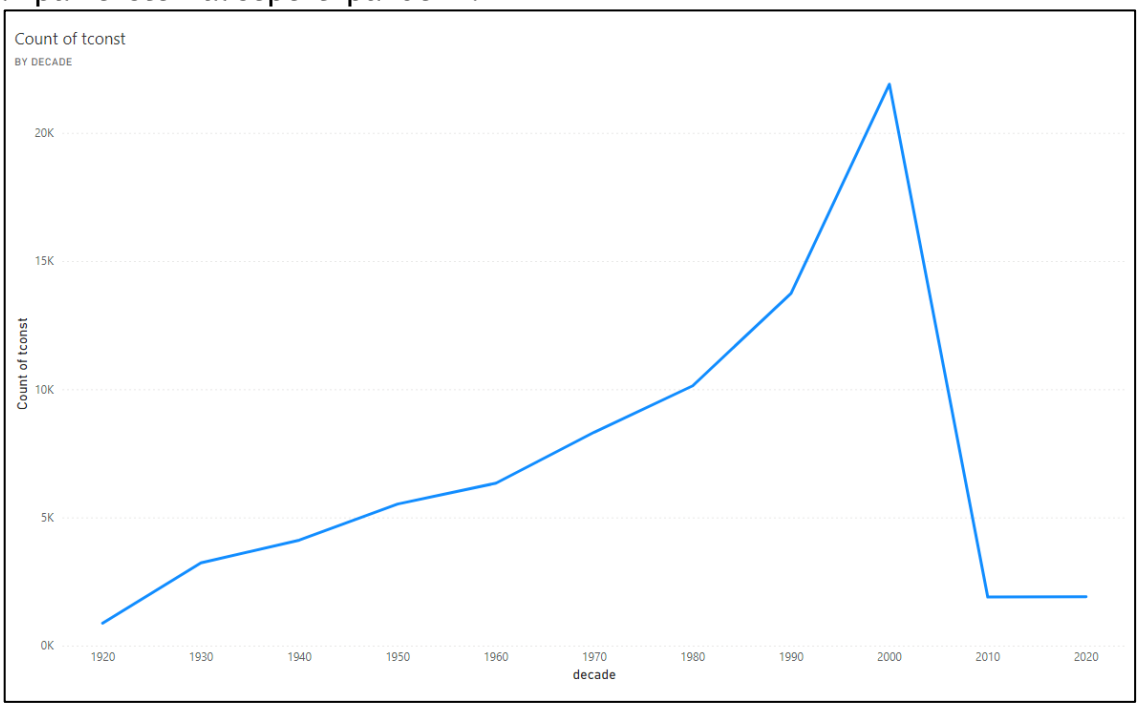
3.2 Hasil Analisis Tren Produksi Film

Analisis tren produksi film dilakukan dengan menghitung jumlah film berdasarkan tahun rilis dan dekade menggunakan Spark SQL.



Gambar 1 Tren Jumlah Film per Tahun Rilis (startYear)

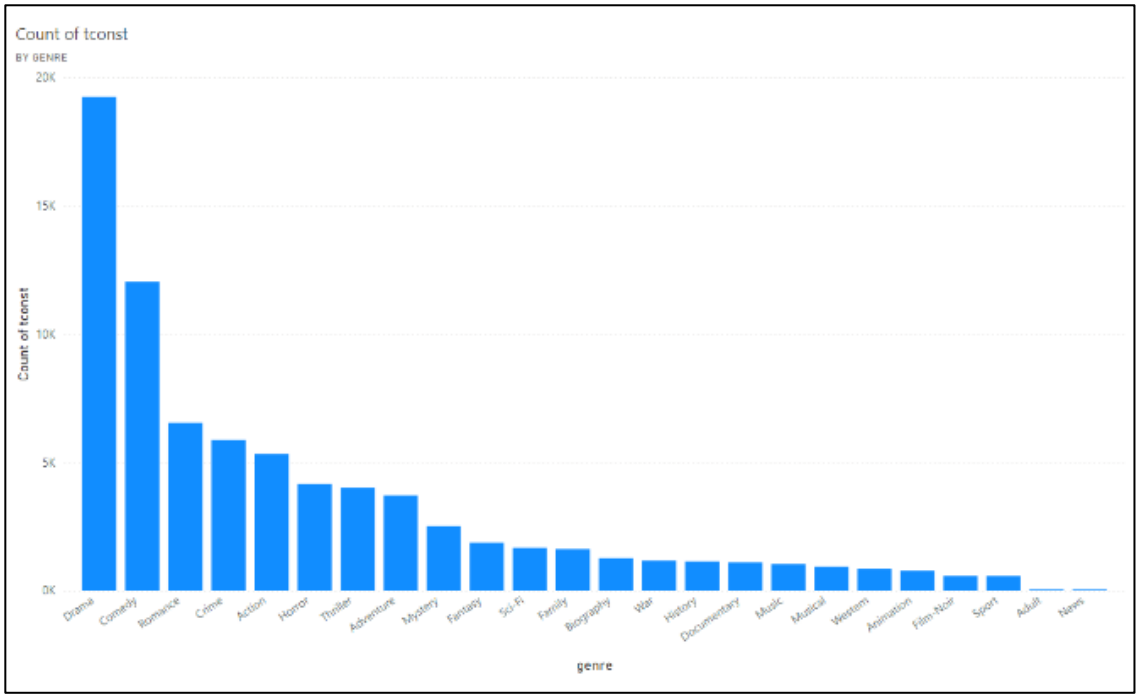
Grafik menunjukkan peningkatan gradual jumlah film dari tahun 1920-an hingga mencapai puncak sekitar 3.000–3.500 film per tahun pada akhir dekade 1990-an hingga awal 2000-an. Setelah itu, terjadi penurunan tajam hingga mendekati nol pada tahun-tahun terbaru (pasca-2010). Pola ini mencerminkan ledakan produksi film pada era digital awal, diikuti oleh kemungkinan pergeseran ke format lain atau dampak eksternal seperti pandemi.



Gambar 2 Tren Jumlah Film per Dekade

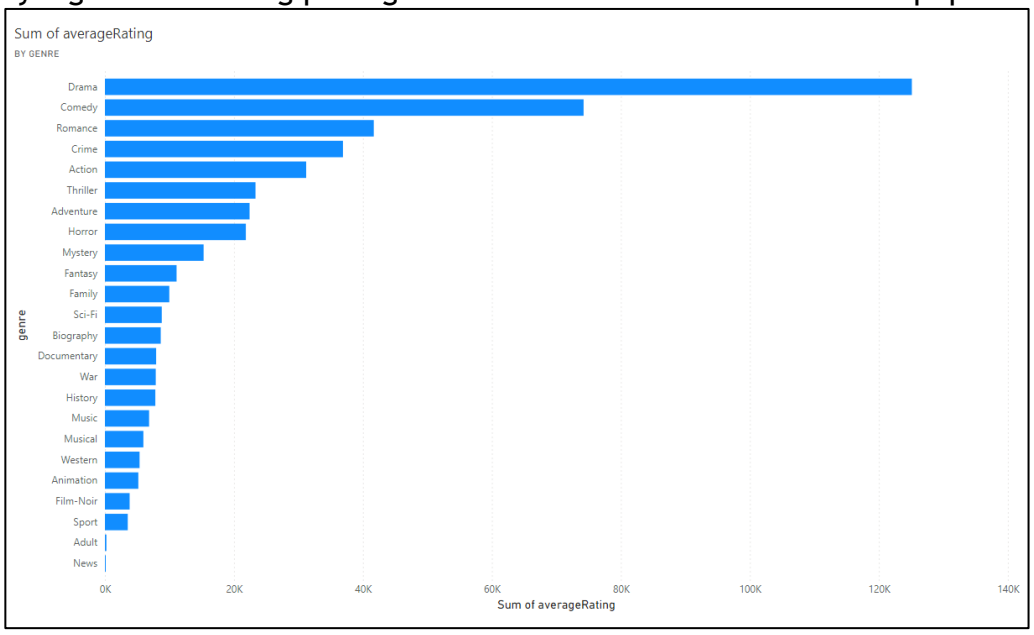
Pada skala dekade, puncak tertinggi tercatat pada dekade 2000-an dengan sekitar 20.000 film, diikuti penurunan signifikan pada dekade 2010-an hingga mendekati 5.000 film.

3.3 Hasil Analisis Distribusi dan Dominasi Genre



Gambar 3 Jumlah Film per Genre

Grafik batang menunjukkan distribusi jumlah film berdasarkan genre. Genre Drama mendominasi dengan lebih dari 20.000 entri, diikuti oleh Comedy (sekitar 15.000 entri), Romance, Crime, dan Action. Genre-genre lain seperti Horror, Thriller, Adventure, dan Mystery memiliki jumlah sedang (antara 5.000–10.000 entri), sementara genre niche seperti War, History, Documentary, Music, Musical, Western, Animation, Film-Noir, Sport, Adult, dan News memiliki jumlah sangat rendah (<5.000 entri). Distribusi ini menggambarkan kecenderungan produksi film global yang lebih condong pada genre naratif emosional dan hiburan populer.



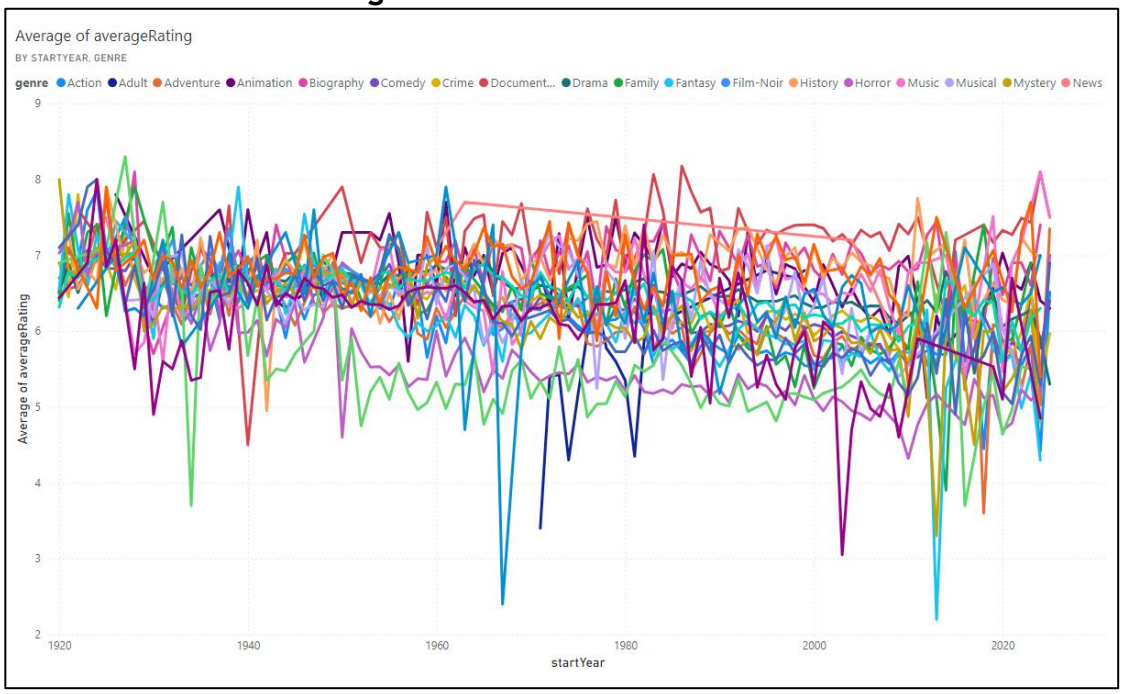
Gambar 4 Total Sum of *averageRating* per Genre

Grafik batang menunjukkan akumulasi total nilai rata-rata rating (sum of averageRating) per genre. Genre Drama memperoleh nilai tertinggi (sekitar 140.000), diikuti oleh Comedy (sekitar 120.000), Romance, Crime, dan Action.

Genre dengan volume produksi tinggi cenderung memiliki akumulasi poin rating yang besar, meskipun bukan berarti rating rata-rata per film tertinggi.

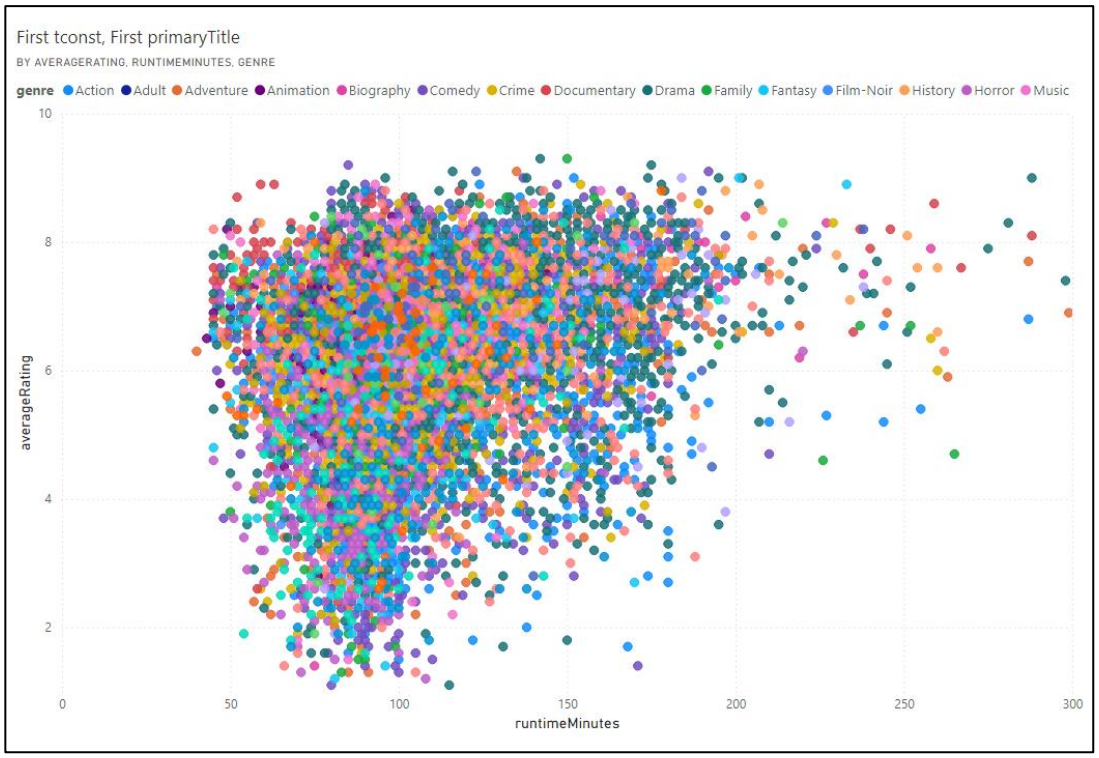
Temuan dominasi genre Drama dan Comedy sejalan dengan penelitian Gupta et al. (2025) yang menemukan bahwa genre-genre tersebut memiliki pengaruh signifikan terhadap rating agregat karena tingginya jumlah produksi.(Gupta et al., 2025) Dominasi ini juga mencerminkan preferensi pasar global, di mana genre dengan narasi kuat dan elemen hiburan cenderung mendominasi produksi dan akumulasi rating.(Zhu & Zhuang, 2025)

3.4 Hasil Analisis Tren Rating Film



Gambar 5 Rata-rata Rating per Tahun per Genre

Grafik garis menunjukkan fluktuasi rata-rata rating (averageRating) antar genre dari tahun 1920 hingga 2020. Nilai rating secara umum berkisar antara 6 hingga 8, dengan variasi yang cukup tinggi pada beberapa genre seperti Animation, Horror, dan Family. Beberapa genre seperti Documentary dan Film-Noir cenderung memiliki kestabilan relatif lebih tinggi, sementara genre lain seperti Drama, Comedy, dan Action menunjukkan pola fluktuatif tanpa tren penurunan signifikan secara keseluruhan. Secara umum, tidak terlihat penurunan kualitas film secara global seiring waktu; rating tetap relatif stabil meskipun volume produksi meningkat tajam pada periode tertentu.



Gambar 6 AverageRating Vs RuntimeMinutes Per Genre

Scatter plot menampilkan hubungan antara durasi film (runtimeMinutes) dan rata-rata rating (averageRating) untuk setiap genre. Titik-titik data membentuk cluster utama pada rating 6–8 dengan durasi 80–150 menit. Film dengan durasi lebih panjang (>200 menit) cenderung memiliki rating lebih tinggi pada genre tertentu seperti Drama, Documentary, dan Biography. Pola ini menunjukkan adanya korelasi positif lemah antara durasi film dan persepsi kualitas oleh penonton, di mana film dengan durasi sedang hingga panjang sering kali mendapat rating lebih baik.

3.5 Pembahasan

Penerapan Big Data Lifecycle dengan Apache Spark pada Azure Databricks berhasil mengolah dataset besar secara efisien. Proses pemecahan genre dan agregasi paralel memungkinkan identifikasi pola tren yang sulit dilakukan dengan alat tradisional. Dominasi genre Drama serta stabilitas rating sepanjang dekade menunjukkan bahwa preferensi penonton relatif konsisten meskipun volume produksi berfluktuasi.

Penurunan produksi pasca-2000 kemungkinan dipengaruhi oleh peralihan ke konten streaming dan serial, yang tidak tercakup dalam dataset ini (hanya fokus pada 'movie'). Temuan ini memperkuat argumen Jadiga (2024) bahwa integrasi ekosistem Hadoop dengan cloud seperti Azure meningkatkan skalabilitas analisis data hiburan secara signifikan.(Jadiga, 2024)

Dashboard Power BI yang dihasilkan memungkinkan visualisasi interaktif, sehingga insight dapat diakses dengan mudah oleh pemangku kepentingan industri film. Keterbatasan penelitian terletak pada ketergantungan terhadap data sekunder IMDb, yang mungkin tidak mencakup film independen atau regional secara lengkap. Secara keseluruhan, pendekatan ini membuktikan efektivitas Big Data Lifecycle dalam mengungkap tren industri film global, memberikan dasar empiris bagi pengambilan keputusan produksi dan distribusi konten.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan Big Data Lifecycle dengan Hadoop Ecosystem pada platform Azure mampu mengolah data film dari IMDb secara terstruktur. Proses pengumpulan, pembersihan, dan integrasi menghasilkan dataset yang siap untuk analisis, sedangkan tahap pemrosesan dengan Spark SQL dan visualisasi melalui Power BI berhasil mengidentifikasi pola tren produksi film yang mencapai titik tertinggi pada dekade 2000-an serta dominasi genre Drama dalam jumlah entri dan akumulasi rating. Fluktuasi rating rata-rata yang tetap berada pada kisaran 6 hingga 8 sepanjang periode, beserta hubungan durasi film dengan penilaian kualitas, memperlihatkan kestabilan preferensi penonton meskipun volume produksi berubah. Secara garis besar, pendekatan ini memberikan landasan empiris untuk pemahaman tren industri film dan upaya mitigasi kelebihan informasi pada platform digital.

DAFTAR PUSTAKA

Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., Hovell, H. Van, Ionescu, A., Łuszczak, A., Switakowski, M., Szafra, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., … Zaharia, M. (2020). Delta Lake : High-Performance ACID Table Storage over Cloud Object Stores. *Proceedings of the VLDB Endowment*, *13*(12). https://doi.org/doi.org/10.14778/3415478.3415560

Awan, M. J., Khan, R. A., Nobanee, H., & Yasin, A. (2021). A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. *Electronics 2021*, *10*(10). https://doi.org/https://doi.org/10.3390/electronics10101215

Gupta, S., Adhikari, U., Varshney, S., & Choudhury, T. (2025). Computational Analysis of Genre Effects on Movie Ratings Using MLP Algorithms. *Journal of Computer Science*, *21*(4), 905–917. https://doi.org/10.3844/jcssp.2025.905.917

Jadiga, S. (2024). Big Data Engineering Using Hadoop and Cloud (GCP/AZURE) Technologies. *International Journal of Computer Trends and Technology*, *72*(8), 60–69. https://doi.org/https://doi.org/10.14445/22312803/ IJCTT-V72I8P109

Moeslim, A. G. S. (2025). Big Data Analysis of Indonesian Health Q & A for Identifying Dominant Disease Topics and Mitigation Information Patterns Using the Random Forest Algorithm Powered by Apache Spark on Google Colab. *Data Science Indonesia*. https://www.researchgate.net/publication/393650778_Big_Data_Analysis_of_Indonesian_Health_QA_for_Identifying_Dominant_Disease_Topics_and_Mitigation_Information_Patterns_Using_the_Random_Forest_Algorithm_Powered_by_Apache_Spark_on_Google_Colab

Oussous, A., Benjelloun, F., Ait, A., & Belfkih, S. (2018). Big Data technologies : A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

Prabaswara, I. R., & Saputra, R. (2020). Implementasi Hadoop Dan Spark Untuk Analisis Penyebaran Demam Berdarah Dengue Berdasarkan Data Twitter. *IT Journal Research and Development (ITJRD)*, *4*(2), 164–171. https://doi.org/10.25299/itjrd.2020

Sabrina, S. S., Aswarulloh, H., & Shiddieq, D. F. (2024). VISUALISASI DATA PENYEBAB KEMATIAN DI INDONESIA RENTANG TAHUN 2000-2022 DENGAN POWER BI. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)*, *12*(2). https://doi.org/http://dx.doi.org/10.23960/jitet.v12i2.4071

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., Mccauley, M., Franklin, M. J., Shenker, S., & Stoica, I. (n.d.). Resilient Distributed Datasets : A Fault-Tolerant

Abstraction for In-Memory Cluster Computing. *NSDI'12: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*. https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

Zhu, L., & Zhuang, L. (2025). A higher-performance big data-based movie recommendation system. *Nonlinear Engineering*, *14*(1). https://doi.org/10.1515/nleng-2025-0092