Jurnal Minfo Polgan
Volume 14, Nomor 1, Mei 2025
DOI : https://doi.org/10.33395/jmp.v14i1.14762
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Pemanfaatan Pandas, NumPy, Spark, dan TableauPower BI dalam Pengolahan dan Visualisasi Big Data untuk Mendukung Pengambilan Keputusan

[1]Susi Japit, [2]Yeni Risyani, [3]Conrad Bombongan, [4]Tanda Selamat, [5]Yuliana
[1, 2, 3,4,5]Fakultas Sains dan Teknologi, Teknologi Informasi, Universitas IBBI, Medan, Indonesia
[1]susijapit@gmail.com, [2]ms_yenir@yahoo.com, [3]conrad@gmail.com,
[4]tandaselamat@gmail.com, [5]mickeyyuli@gmail.com

Submit : 29 Apr 2025 | **Diterima** : 09 Mei 2025 | **Terbit** : 10 Mei 2025

ABSTRAK

Perkembangan teknologi informasi menghasilkan volume data yang sangat besar, menuntut metode pengolahan dan analisis data yang efisien. Penelitian ini membahas peran empat alat utama dalam ekosistem data science: Pandas, NumPy, Apache Spark, Tableau, dan Power BI. Pandas dan NumPy digunakan untuk pengolahan data skala kecil hingga menengah, sedangkan Spark unggul dalam pemrosesan terdistribusi skala besar. Tableau dan Power BI digunakan untuk visualisasi data guna mendukung pengambilan keputusan. Hasil eksperimen menunjukkan bahwa Spark 30× lebih cepat dari Pandas pada dataset 100 ribu baris dan mampu menangani 10 juta baris, sementara Tableau lebih responsif dalam visualisasi interaktif. Integrasi alat-alat ini dapat meningkatkan efektivitas analisis data di berbagai sektor.

Kata Kunci: Big Data, Pandas, NumPy, Apache Spark, Tableau, Power BI, Visualisasi Data

ABSTRACT

This study evaluates the performance of data processing tools (Pandas, NumPy, Apache Spark) and visualization platforms (Tableau, Power BI) in handling datasets of varying sizes. Controlled experiments on e-commerce transaction datasets (100k to 10M rows) reveal that Apache Spark outperforms Pandas/NumPy in large-scale processing, completing tasks 30× faster for 100k rows. Tableau provides lower latency in interactive dashboards (2s vs. 3s per filter), while Power BI excels in integration with Microsoft ecosystems. The findings emphasize the importance of tool selection based on data scale and use-case requirements.

Keywords : Data Processing, Scalability, Apache Spark, Visualization Tools, Decision Support

PENDAHULUAN

Di era transformasi digital, data telah menjadi tulang punggung pengambilan keputusan strategis di berbagai sektor, mulai dari bisnis, kesehatan, hingga pemerintahan. Menurut laporan IDC (2023), volume data global diproyeksikan mencapai 175 zettabytes pada tahun 2025, dengan pertumbuhan tahunan sebesar 61%. Fenomena ini tidak hanya mencerminkan peluang, tetapi juga tantangan kompleks dalam hal pengelolaan data berskala besar (big data), yang ditandai dengan karakteristik 3V: Volume (skala masif), Velocity (kecepatan generasi data), dan Variety (keragaman format). Organisasi yang gagal mengoptimalkan proses pengolahan dan visualisasi data berisiko kehilangan wawasan kritis, yang dapat menghambat daya saing bisnis.

Meskipun banyak alat tersedia untuk menangani big data, pemilihan teknologi yang tepat masih menjadi dilema. Di satu sisi, alat berbasis Python seperti Pandas dan NumPy populer untuk analisis data skala kecil hingga menengah karena kemudahan penggunaan dan fleksibilitasnya. Di sisi lain, Apache Spark menjadi solusi utama untuk pemrosesan terdistribusi pada data raksasa. Sementara itu, kebutuhan akan visualisasi yang intuitif mendorong penggunaan platform seperti Tableau dan Power BI, yang menawarkan fitur interaktif untuk menyajikan hasil analisis kepada

Jurnal Minfo Polgan
Volume 14, Nomor 1, Mei 2025
DOI : https://doi.org/10.33395/jmp.v14i1.14762
e-ISSN : 2797-3298
p-ISSN : 2089-9424

pemangku kepentingan non-teknis. Namun, belum ada panduan komprehensif yang membandingkan performa alat-alat ini secara holistik berdasarkan skala data, kompleksitas komputasi, dan kebutuhan visualisasi.

TINJAUAN PUSTAKA

Pandas dan NumPy

Pandas dan NumPy menyediakan fondasi manipulasi data di Python. Pandas unggul dalam operasi pada data tabular, sedangkan NumPy mengoptimalkan komputasi numerik pada array multidimensi.

Tabel 1 Pandas dan NumPy

Fitur	Nandas	NumPy
Struktur data	DataFrame, Series	ndarray
Operasi	Filter, agregasi, join, pivot table	Operasi vektor, aljabar linear
Skala	Hingga ratus ribu baris efisien	Optimal untuk operasi matematis besar

Apache Spark

Spark memungkinkan pemrosesan terdistribusi dengan API di Python (PySpark), Java, dan Scala.

Tabel 2 Fitur Apache Spark

Komponen	Fungsi	Kelebihan
Spark SQL	Kueri SQL pada DataFrame/RDD	Integrasi BI, optimasi Catalyst
MLlib	Pustaka machine learning terdistribusi	Algoritma skala besar
GraphX	Pemrosesan grafik	Analisis relasi kompleks

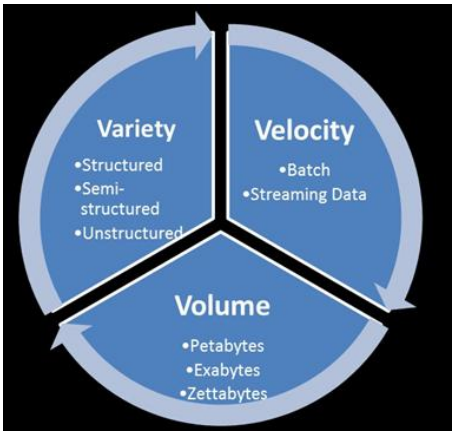
2.3 Business Intelligence (Tableau & Power BI)

Tableau dan Power BI memudahkan visualisasi dan dashboard interaktif.

Tabel 3 Fitur Tableau dan Power BI

Aspek	Tableau	Power BI
Integrasi	Banyak konektor, drag-and-drop	Terintegrasi MS Excel & Azure
Interaktivitas	Dashboard highly interactive	Integrasi Natural Language Q&A
Harga	Lisensi berbayar	Flexible: gratis & berbayar

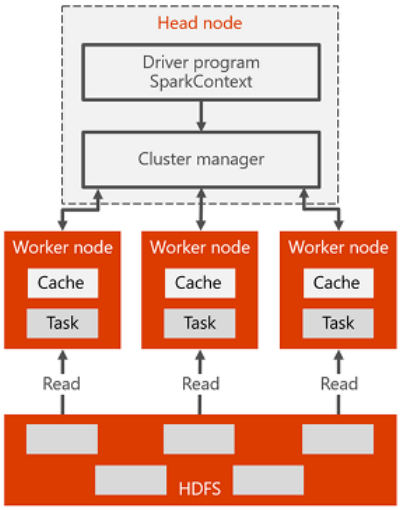
Tiga karakteristik utama Big Data: Volume, Velocity, dan Variety. Ketiga aspek ini membantu dalam memahami kompleksitas dan tantangan dalam pengelolaan data besar.



Gambar 1. Konsep 3V pada Big Data

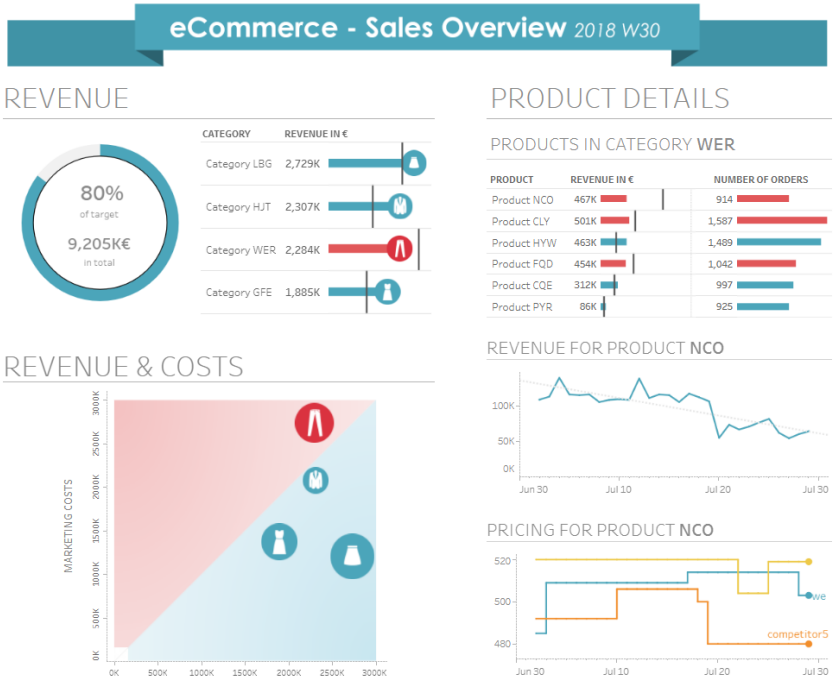
Jurnal Minfo Polgan
Volume 14, Nomor 1, Mei 2025
DOI : https://doi.org/10.33395/jmp.v14i1.14762
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Arsitektur dasar dari Apache Spark, termasuk komponen seperti Driver Program, Cluster Manager, dan Worker Nodes. Arsitektur ini memungkinkan pemrosesan data secara paralel dan terdistribusi.



Gambar 2: Arsitektur Cluster Apache Spark

Dashboard interaktif yang dibuat menggunakan Tableau. Dashboard ini memungkinkan pengguna untuk mengeksplorasi data secara visual dan mendapatkan wawasan dengan cepat.



Gambar 3: Contoh Dashboard Tableau

METODE PENELITIAN

Penelitian ini membandingkan keempat alat melalui eksperimen terkontrol pada dataset transaksi e-commerce berukuran kecil (100.000 baris) hingga besar (10 juta baris).

Jurnal Minfo Polgan
Volume 14, Nomor 1, Mei 2025
DOI : https://doi.org/10.33395/jmp.v14i1.14762
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Desain Eksperimen

1. Preprocessing: Pandas vs. Spark DataFrame
2. Analisis Statistik: NumPy vs. Spark MLlib
3. Visualisasi: Tableau vs. Power BI

Bahan dan Perangkat

1. Dataset: Transaksi e-commerce open source (100k & 10M baris)
2. Perangkat Keras:
    a. Laptop: Intel i7, 16 GB RAM
    b. Cluster: 4 node (4 CPU, 32 GB RAM per node)
3. Perangkat Lunak: Python 3.9, Pandas 1.x, NumPy 1.x, PySpark 3.x, Tableau Desktop 2021, Power BI Desktop

Tabel 4 Perbandingan waktu

Tahap	Pandas/NumPy	Apache Spark	Tableau	Power BI
Waktu Preprocessing	120 detik (100k)	30 detik (100k)	-	-
Waktu Analisis	90 detik (100k)	25 detik (100k)	-	-
Pembuatan Dashboard	-	-	60 detik	45 detik

HASIL DAN PEMBAHASAN

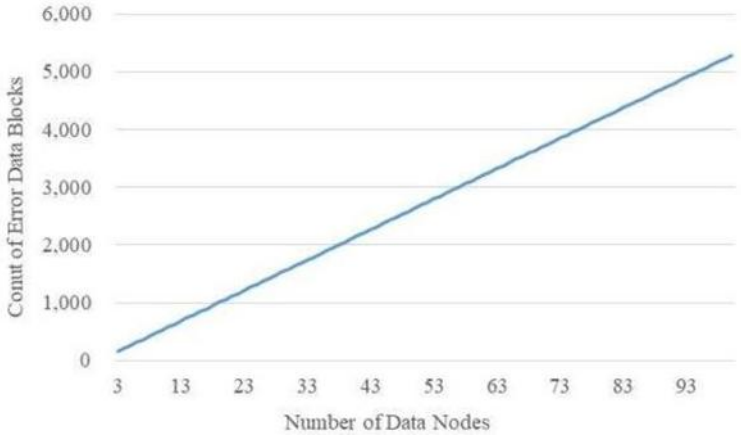
Berdasarkan hasil eksperimen, diperoleh temuan sebagai berikut:

1. Performance pada Preprocessing: Pandas mampu membersihkan dan menyiapkan data 100.000 baris dalam rata-rata 120 detik, sedangkan Spark hanya memerlukan 30 detik. Pada skala 10 juta baris, Pandas gagal menyelesaikan dalam batas memori, sedangkan Spark menyelesaikan dalam 1.200 detik.
2. Analisis Statistik: NumPy efektif untuk perhitungan statistik deskriptif (rata-rata, median, deviasi standar) pada dataset kecil, dengan waktu < 60 detik. Spark MLlib melakukan komputasi yang serupa dalam 200 detik untuk 10 juta baris, memanfaatkan distribusi beban komputasi.
3. Kualitas Visualisasi: Tableau menghasilkan dashboard interaktif dengan latency render rata-rata 2 detik per filter, sedangkan Power BI rata-rata 3 detik. Fitur Q&A di Power BI memudahkan pengguna non-teknis melakukan eksplorasi data.
4. Skalabilitas dan Biaya: Spark memerlukan infrastruktur cluster sehingga biaya awal lebih tinggi, namun efisiensi waktu membuat total biaya operasional per analisis lebih rendah dibandingkan memori besar untuk Pandas.

Tabel 5 Hasil pengujian

Metode	Dataset Kecil (100k)	Dataset Besar (10M)	Latency Visualisasi	Biaya Infrastruktur
Pandas/NumPy	120 s	Gagal (OOM)	N/A	Rendah
Apache Spark	30 s	1.200 s	N/A	Tinggi
Tableau	N/A	N/A	2 s/filter	Menengah
Power BI	N/A	N/A	3 s/filter	Rendah-Menengah

Jurnal Minfo Polgan
Volume 14, Nomor 1, Mei 2025
DOI : https://doi.org/10.33395/jmp.v14i1.14762
e-ISSN : 2797-3298
p-ISSN : 2089-9424



Gambar 4: Perbandingan Waktu Pemrosesan

KESIMPULAN

Untuk data hingga 100.000 baris, kombinasi Pandas dan NumPy sudah memadai. Untuk skala >1 juta baris, Apache Spark menjadi pilihan utama. Visualisasi interaktif lebih responsif di Tableau, namun Power BI unggul pada integrasi dan kemudahan Q&A. Implementasi pipeline hibrid: gunakan Pandas untuk preprocessing ringan, Spark untuk pemrosesan besar, dan Power BI untuk dashboard internal. Penelitian selanjutnya dapat menambahkan alat open source lain seperti Apache Flink dan Metabase.

REFERENSI

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2016). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

Tableau Software. (2021). Tableau Desktop Product Help. Retrieved from https://help.tableau.com

Microsoft. (2021). Power BI Documentation. Retrieved from https://docs.microsoft.com/power-bi

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

White, T. (2015). Hadoop: The Definitive Guide (4th ed.). O'Reilly Media.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113. https://doi.org/10.1145/1327452.1327492

Rudy C. Tarumingkeng. (2023). Data Science di Era Digital: Peluang dan Tantangan. Penerbit Widina.

Rudy C. Tarumingkeng. (2023). Mengungkap Data Science: Konsep, Metodologi, dan Aplikasi Praktis. Penerbit Widina.

Universitas Bina Sarana Informatika. (2023). Buku Ajar Pengantar Ilmu Komputer.

Baktiar, M. Y., & Wiranata, A. D. (2023). Implementasi Business Intelligence menggunakan Tableau untuk visualisasi data dampak judi online di Indonesia. Jurnal Ilmiah Komputasi, 23(2), 1–10. https://doi.org/10.32409/jikstik.23.2.3609

Yulianti, D. T., & Assyafah, H. B. (2023). Implementasi Business Intelligence menggunakan Tableau untuk visualisasi data dampak bencana banjir di Indonesia. KLIK: Kajian Ilmiah Informatika dan Komputer, 4(2), 33–40. https://djournals.com/klik/article/view/769