

TUGAS 2
TRANSFORMASI
MATA KULIAH BIG DATA PLATFORM

Dosen Pengampu:

Dr. Yudi Wibisono, M.T.



Disusun oleh:

Kelompok 1

Faisal Nur Qolbi	2311399
Hafidz Tantowi	2308817
Muhammad Farhan	2309323
Muhammad Helmi Rahmadi	2311574
Sifa Imania Nurul Hidayah	2312084
Yazid Madarizel	2305328

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA

2025

A. Identifikasi Sumber Data

Sumber Data:

- YFinance: <https://finance.yahoo.com>

The screenshot shows the Yahoo Finance homepage with a dark theme. On the left, there's a sidebar with categories like Overview, World Indices, Futures, Bonds, Currencies, Options, Sectors, Stocks, and Crypto. The main content area has a banner for 'Private Company Data' and a 'Stocks' section with tabs for Most Active, Trending Now, Top Gainers, Top Losers, 52 Week Gainers, and 52 Week Losers. Below this is a table view of stock data for companies like HTZ, NFLX, PLD, AREC, CGC, 9618.HK, and VIST. To the right, there's a 'Quote Lookup' section showing U.S. markets close in 6h 26m with data for S&P Futures, Dow Futures, Nasdaq Futures, Russell 2000, Crude Oil, and Gold. At the bottom, there's a 'My portfolios' section with a 'Sign in to access your portfolio' button.

Gambar 1. Halaman Web YFinance

- IDX (Indonesia Stock Exchange):

<https://www.idx.co.id/id/perusahaan-tercatat/laporan-keuangan-dan-tahunan>

The screenshot shows the IDX website with a header including the IDX logo, a search bar, and buttons for MASUK, DAFTAR, EN, and ID. Below the header is a navigation menu with links to DATA PASAR, PRODUK & LAYANAN, PERUSAHAAN TERCATAT, IDX SYARIAH, ANGGOTA BURSA & PARTISIPAN, BERITA, PERATURAN, INVESTOR, and TENTANG BEI. A breadcrumb trail shows the user is at Perusahaan Tercatat > Laporan Keuangan Dan Tahunan. The main content title is 'Laporan Keuangan dan Tahunan'. Below it is a search bar with dropdowns for 'Search Company Cr' (set to 12), 'A-Z' (sorted by name), and a 'Filter' button. There are four filter sections: 'Jenis Laporan' (radio buttons for Laporan Keuangan and Laporan Tahunan), 'Jenis Efek' (radio buttons for Saham and Obligasi), 'Tahun' (radio buttons for 2025, 2024, 2023, 2022, 2021), and 'Periode' (radio buttons for Triwulan 1, Triwulan 2, Triwulan 3, and Tahunan). At the bottom are 'RESET' and 'Terapkan' buttons.

Gambar 2. Halaman Web IDX

- IQplus: <http://www.iqplus.info/index.php>



Gambar 3 Halaman Web IQPLUS

1. ISI

1) YFinance

YFinance menyajikan berbagai data pasar saham dan keuangan perusahaan dalam format tabular yang terstruktur. Data yang tersedia meliputi:

- **Date** : Tanggal pencatatan harga saham.
- **Open** : Harga pembukaan saham pada hari tersebut.
- **High** : Harga tertinggi saham yang dicapai dalam satu hari perdagangan.
- **Low** : Harga terendah saham dalam satu hari perdagangan.
- **Close** : Harga penutupan saham pada hari tersebut.
- **Volume** : Jumlah saham yang diperdagangkan pada hari tersebut.
- **Dividens** : Dividen yang dibayarkan per saham pada tanggal tersebut
- **Stock Splits** : Informasi mengenai pemecahan saham (stock split) yang terjadi.

Data dapat diunduh dalam format CSV atau JSON melalui **library yfinance** di python yang memudahkan pengolahan lebih lanjut.

2) IDX (Indonesia Stock Exchange)

IDX menyediakan laporan keuangan perusahaan yang terdaftar di Bursa Efek Indonesia dalam zip bernama instance.zip dimana didalamnya ada file Taxonomy.xsd dan instance.xrbl. Data mencakup laporan laba rugi, neraca, laporan arus kas, dan catatan atas laporan keuangan. Laporan Keuangan yang kami ambil

berada di rentang laporan tahunan saja untuk sekarang dari 2021-2024 khusus jenis efek saham. Dengan format XML maka membutuhkan parsing lebih lanjut untuk mengekstrak informasi yang dibutuhkan.

3) IQPLUS

IQPLUS menyediakan artikel berita yang berfokus pada keuangan, analisis pasar, dan informasi ekonomi di Indonesia. Konten disajikan dalam bentuk teks bebas dengan tag atau kategori untuk mempermudah pencarian, namun memerlukan pengolahan lebih lanjut untuk mendapatkan struktur yang konsisten.

2. ACCESSIBILITY

1) YFinance

Data YFinance dapat diakses secara bebas untuk penggunaan dasar, namun terdapat batasan untuk penggunaan skala besar. Penggunaan dalam aplikasi komersial mungkin memerlukan izin lebih lanjut dari penyedia data asli. API YFinance menyediakan cara mudah untuk mengakses data secara terprogram.

2) IDX

Data laporan keuangan IDX tersedia untuk umum, namun akses melalui API membutuhkan otorisasi. Tidak ada lisensi khusus yang membatasi penggunaan data untuk keperluan analisis atau publikasi, selama sumber data disebutkan dengan benar.

3) IQPLUS

Sebagian besar berita IQPLUS dapat diakses secara gratis, tetapi terdapat artikel premium yang memerlukan langganan. Semua konten memiliki hak cipta, sehingga penggunaan untuk keperluan bisnis atau publikasi ulang memerlukan izin dari penerbit

3. USABILITY

1) YFinance

Data YFinance sudah terstruktur dengan baik dan mudah diekstrak menggunakan API YFinance. Proses pengolahan minimal diperlukan karena format data sudah konsisten dan siap untuk analisis lebih lanjut.

2) IDX

Data IDX dalam format XML memerlukan parsing lebih kompleks untuk mengekstrak informasi yang relevan. Diperlukan praproses data untuk membersihkan dan menstandarisasi format laporan keuangan dari berbagai perusahaan.

3) IQPLUS

Artikel IQPLUS berbentuk teks bebas yang memerlukan pengolahan lebih lanjut untuk mendapatkan struktur yang konsisten. Teknik natural language processing (NLP) mungkin diperlukan untuk mengekstrak informasi penting dari artikel.

4. COVERAGE

1) YFinance

YFinance menyediakan cakupan global untuk data saham dan indeks utama, namun mungkin memiliki keterbatasan untuk pasar tertentu atau data yang sangat spesifik. Cakupan terbaik adalah untuk pasar saham AS dan indeks global utama.

2) IDX

IDX menyediakan data komprehensif untuk seluruh perusahaan publik yang terdaftar di BEI, tetapi tidak mencakup perusahaan swasta atau data non-finansial.

Fokus utama adalah pada data keuangan formal yang dilaporkan.

3) IQPLUS

IQPLUS fokus pada berita dan analisis ekonomi-keuangan Indonesia, dengan cakupan yang lebih terbatas untuk pasar global atau sektor-sektor khusus. Lebih baik digunakan untuk memahami sentimen pasar Indonesia dan berita ekonomi lokal.

5. RELIABILITY

1) YFinance

YFinance cukup terpercaya dengan data yang diperbarui secara teratur, meskipun terkadang terdapat keterlambatan kecil atau masalah teknis. Untuk analisis penting, verifikasi dengan sumber resmi mungkin diperlukan.

2) IDX

Data IDX sangat terpercaya karena bersumber dari laporan yang diaudit oleh otoritas resmi. Namun, kualitas data tetap bergantung pada ketepatan dan kepatuhan pelaporan masing-masing perusahaan

3) IQPLUS

Berita IQPLUS umumnya dapat dipercaya karena berasal dari sumber yang kredibel dan diperbarui secara rutin. Namun, seperti semua sumber berita, verifikasi tambahan direkomendasikan untuk memastikan keakuratan informasi sebelum pengambilan keputusan.

6. VOLUME

1) YFinance

Yahoo Finance Menjadi sumber data karena cukup dikenal dan banyak dipakai oleh investor maupun analis. Data yang tersedia cukup konsisten dan bisa diandalkan, terutama untuk analisis keuangan tingkat dasar sampai menengah. Selama proses pengambilan data, terkumpul total 1.833.217 dokumen yang disimpan di MongoDB. Data ini mencakup 951 saham, sesuai dengan target yang sudah direncanakan sejak awal (100% tercapai). Rata-rata tiap saham punya sekitar 1.928 dokumen, meskipun jumlah pastinya bisa berbeda tergantung seberapa panjang riwayat perdagangan saham tersebut dan ketersediaan datanya. Volume data yang dikumpulkan menunjukkan skala pengambilan yang besar dan cukup lengkap, jadi bisa digunakan untuk berbagai keperluan analisis historis secara mendalam.

2) IDX

Data IDX sangat kredibel karena berasal dari laporan keuangan resmi perusahaan yang telah diaudit. Struktur data terstandarisasi sesuai dengan peraturan pelaporan yang berlaku di Indonesia. Dari rentang 2021-2024 kami melakukan scraping dan berhasil mengumpulkan total 1844 dokumen.

3) IQPLUS

Kualitas berita IQPLUS bervariasi tergantung pada jurnalis atau penulis artikel. Sebagai sumber berita, IQPLUS menyediakan analisis dan pembaruan pasar yang bermanfaat, tetapi dengan tingkat kedalaman yang bervariasi. Publikasi artikel sekitar 100-150 artikel per hari, kami scraping total 9462 artikel dari kategori market dan stock dimana dalam prosesnya membutuhkan waktu kisaran 3 jam 19 menit.

7. VELOCITY

1) YFinance

Data dari Yahoo Finance umumnya diperbarui secara real-time untuk harga saham dan indeks, sementara laporan keuangan dan informasi fundamental perusahaan mengikuti siklus pelaporan yang lebih jarang diperbarui. Dalam proses pengambilan data yang dilakukan, waktu total yang dibutuhkan adalah sekitar 367,69 detik (atau sekitar 6 menit 8 detik). Rata-rata kecepatan pengambilan adalah sekitar 0,39 detik per saham, dengan throughput mencapai sekitar 4.986 dokumen per detik. Proses dilakukan secara sekuensial, artinya data untuk satu saham diambil dan diproses hingga selesai sebelum berpindah ke saham berikutnya. Meskipun belum menggunakan metode paralel atau multithread, kecepatan yang dicapai sudah cukup efisien untuk pengumpulan data historis dalam jumlah besar

2) IDX

Data IDX diperbarui sesuai dengan siklus pelaporan perusahaan, umumnya kuartalan atau tahunan. Terdapat periode intensif publikasi laporan setelah tenggat waktu pelaporan resmi. Pembaruan laporan keuangan biasanya dilakukan dalam rentang kuartalan/tahunan.

Kami memerlukan waktu selama 3 jam 54 menit untuk scraping 1844 total dokumen, artinya kami bisa scraping dengan kecepatan kisaran 473 dokumen/jam.

3) IQPLUS

IQPLUS memperbarui berita cukup cepat, dengan publikasi artikel baru sepanjang hari terutama pada jam kerja. Volume publikasi meningkat signifikan saat terjadi peristiwa ekonomi atau bisnis penting. Publikasi artikel sekitar 5-10 artikel per jam jadi tidak bisa disebut real-time juga sih, tapi cukup cepat untuk media pemberitaan. Saat scraping program kami mendapatkan rata-rata kecepatan 2853 artikel per jam, tapi scraping tentu banyak hal yang memengaruhi mulai dari koneksi internet sampai keefisienan program.

B. Ingest

1) YFINANCE

The screenshot shows the MongoDB Compass interface with the 'yfinance' database selected. The left sidebar lists connections, and the main area displays a table of collections with their storage sizes, document counts, average document sizes, index counts, and total index sizes. The collections listed are:

Collection Name	Storage size	Documents	Avg. document size	Indexes	Total index size
ABM Investama Tbk	122.88 kB	0	0 B	1	45.04 kB
Aset Indonesia Tbk	125.17 kB	0	0 B	1	45.04 kB
Adhi Karya (Persero) Tbk	140.46 kB	0	0 B	1	45.04 kB
Adi Sarana Armada Tbk	151.55 kB	0	0 B	1	45.04 kB
Adira Dinamika Multi Finance T	114.69 kB	0	0 B	1	45.04 kB
Agung Podomoro Land Tbk	122.88 kB	0	0 B	1	45.04 kB
Akash Wira International Tbk	104.69 kB	0	0 B	1	45.04 kB
AKR Corporindo Tbk	114.69 kB	0	0 B	1	45.04 kB

The screenshot shows the Compass MongoDB interface. The left sidebar lists connections, including 'localhost:27017' which contains databases like 'BertiniPosr', 'IXData', 'StockNews', 'ABM Investama Tbk', and 'yfinance'. The main area shows the 'yfinance' database with the 'ABM Investama Tbk' collection selected. A table displays five documents, each representing a stock price entry for ABM Investama Tbk. The columns include '_id', 'Date', 'Open', 'High', 'Low', 'Close', 'Volume', 'Dividende', and 'Stock Split'. The dates range from 2014-01-01 to 2014-01-02.

_id	Date	Open	High	Low	Close	Volume	Dividende	Stock Split
<code>ObjectID('0f7f73e86331d027056aa31')</code>	2014-01-01T00:00:00+07:00	2191.185571542	2172.307555938	2172.307555938	2172.307555938	470000	0	0
<code>ObjectID('0f7f73e86331d027056aa37')</code>	2014-01-01T00:00:00+07:00	2172.307555938	2172.307555938	2172.307555938	2172.307555938	3975585	0	0
<code>ObjectID('0f7f73e86331d027056aa39')</code>	2014-01-01T00:00:00+07:00	2169.175985979	2169.175985979	2169.175985979	2169.175985979	3975585	0	0
<code>ObjectID('0f7f73e86331d027056aa3b')</code>	2014-01-01T00:00:00+07:00	2136.788532772	2136.788532772	2136.788532772	2136.788532772	354288944	0	0
<code>ObjectID('0f7f73e86331d027056aa3c')</code>	2014-01-01T00:00:00+07:00	2176.433941566	2176.433941566	2176.433941566	2176.433941566	354288944	0	0

Gambar 5. Ingest MongoDB YFinance

Proses ingestion data menggunakan yfinance diawali dengan membaca daftar kode saham yang tersimpan dalam file CSV. File ini berisi daftar simbol saham yang akan digunakan sebagai parameter untuk mengambil data historis dari Yahoo Finance. Daftar saham ini kemudian diolah satu per satu untuk mendapatkan informasi pasar dari masing-masing kode saham.

Setelah mendapatkan daftar saham, yfinance digunakan untuk mengambil historical data dengan rentang waktu 2014 hingga 2025. Data ini mencakup harga pembukaan (open), harga penutupan (close), harga tertinggi (high), harga terendah (low), volume transaksi, serta adjusted close price. Data yang diperoleh dalam format tabular ini kemudian dikonversi ke dalam format JSON, yang lebih fleksibel untuk penyimpanan dan analisis lebih lanjut.

Format JSON dipilih karena kompatibel dengan berbagai sistem basis data, termasuk MongoDB, yang sering digunakan dalam proyek Big Data. Dengan format ini, data dapat dengan mudah diolah, dianalisis, atau diintegrasikan ke dalam pipeline pemrosesan data lainnya. Selain itu, format JSON memudahkan transfer data antar sistem serta dapat digunakan untuk membangun aplikasi analitik berbasis data.

2) IDX(Indonesia Stock Exchange)

The screenshot shows the Compass MongoDB interface. On the left, the 'CONNECTIONS' sidebar lists 'localhost:27017', 'BeritaPasar', and 'IDXData'. The 'IDXData' connection is expanded, showing four collections: 'Financial_2021', 'Financial_2022', 'Financial_2023', and 'Financial_2024'. Each collection card provides storage size, document count, average document size, index count, and total index size. For example, 'Financial_2021' has 737 documents, 4.23 MB storage, and 19.51 kB average document size.

Collection	Storage size	Documents	Avg. document size	Indexes	Total index size
Financial_2021	4.23 MB	737	19.51 kB	1	28.67 kB
Financial_2022	19.36 MB	815	54.76 kB	1	32.77 kB
Financial_2023	2.86 MB	149	46.18 kB	1	20.48 kB
Financial_2024	3.47 MB	143	56.66 kB	1	20.48 kB

Gambar 6. Ingest IDX

Dalam proyek ini, kami mengembangkan program untuk mengambil laporan keuangan perusahaan yang terdaftar di Bursa Efek Indonesia (IDX) dari situs resmi IDX secara otomatis. Tujuan utama dari proyek ini adalah mengumpulkan data keuangan tanpa harus mengunduhnya secara manual, sehingga bisa digunakan untuk analisis lebih lanjut.

Untuk melakukan scraping, kami menggunakan Selenium sebagai alat otomatisasi browser. Selenium digunakan untuk membuka halaman web IDX, memilih tahun dan periode laporan, serta mengunduh file laporan keuangan instance.zip. Setelah diunduh, file ZIP diekstrak untuk mendapatkan file XML berisi Taxonomy.xsd dan instance.xrbl, yang kemudian dikonversi menjadi JSON sebelum disimpan ke dalam database MongoDB.

Langkah pertama dalam scraping adalah membuka halaman laporan keuangan di IDX dan memilih tahun laporan yang diinginkan. Setelah itu, program mencari daftar perusahaan yang memiliki laporan keuangan yang tersedia untuk diunduh. Setiap perusahaan dalam daftar akan diproses satu per satu, dan program akan mengunduh file laporan keuangan dalam format ZIP.

Setelah file ZIP berhasil diunduh, program mengekstrak isi file untuk mendapatkan file XML yang berisi data keuangan perusahaan. File XML ini kemudian dikonversi menjadi format JSON agar lebih mudah diolah dan dianalisis. JSON hasil konversi kemudian disimpan ke dalam database MongoDB dengan

struktur yang mencakup kode perusahaan, tahun, periode laporan, serta data keuangan yang diekstrak dari XML.

Agar proses berjalan lancar, kami juga menambahkan beberapa mekanisme *error handling*. Jika terjadi error saat mengunduh file, program akan mencatatnya dan melanjutkan ke perusahaan berikutnya tanpa menghentikan proses secara keseluruhan. Selain itu, setelah file ZIP diproses, file tersebut akan dihapus untuk menghemat ruang penyimpanan.

3) IQPLUS

The screenshot shows the Compass MongoDB interface. On the left, the 'CONNECTIONS' sidebar lists 'localhost:27017' and 'BeritaPasar'. The 'BeritaPasar' connection is expanded, showing collections like MarketNews, MarketNewsSummary, RingkasanStockNews, and StockNews, each with a list of documents. The main area displays four collections:

Collection	Storage size	Documents	Avg. document size	Indexes	Total index size
MarketNews	6.41 MB	5.2 K	2.16 kB	1	10.59 kB
MarketNewsSummary	1.68 MB	5.2 K	436.00 B	1	217.09 kB
RingkasanStockNews	28.67 kB	12	2.16 kB	1	36.86 kB
StockNews	4.97 MB	4.3 K	2.14 kB	1	94.21 kB

Gambar 7. Isi Collection dari Database BeritaPasar

The screenshot shows the Compass MongoDB interface with the 'MarketNews' collection selected. The left sidebar lists various database connections and collections. The main panel displays the 'MarketNews' collection with 5.2K documents. Each document entry includes a preview of the news article's title, date, link, publication date, and content.

Gambar 8. Isi dari Collection Stock dan Market

Dalam proyek ini, kami membuat sistem otomatis untuk mengambil berita pasar dari situs IQPlus dan menyimpannya ke dalam database MongoDB. Tujuan utamanya adalah agar data berita bisa dikumpulkan tanpa harus mengambilnya secara manual, sehingga bisa lebih mudah dianalisis nantinya.

Untuk melakukan scraping, kami menggunakan dua library utama: Selenium dan BeautifulSoup. Selenium digunakan untuk membuka dan menjelajahi halaman web secara otomatis, sedangkan BeautifulSoup bertugas membaca dan mengambil informasi dari kode HTML halaman tersebut. Hasil akhirnya disimpan dalam MongoDB agar bisa diakses dan dianalisis lebih lanjut.

Pertama, program mencari tahu berapa jumlah halaman berita yang tersedia di situs IQPlus. Setelah itu, program mulai mengunjungi setiap halaman satu per satu untuk mengambil daftar berita yang ada. Setiap berita yang ditemukan akan diambil informasinya, seperti judul, waktu publikasi, dan link ke artikel lengkapnya. Setelah mendapatkan daftar berita, program kemudian membuka setiap link berita tersebut untuk mengambil tanggal artikel dan isi beritanya. Agar lebih bersih, elemen-elemen

yang tidak relevan seperti tombol zoom dihapus. Data yang sudah dikumpulkan lalu disimpan ke dalam MongoDB, dengan pengecekan agar tidak ada data yang tersimpan dua kali.

Format data yang tersimpan mencakup beberapa elemen penting: judul berita, waktu publikasi di halaman utama, link artikel, tanggal yang tertera di dalam artikel, serta isi lengkap beritanya. Dengan format ini, data bisa digunakan untuk berbagai keperluan, seperti analisis tren pasar atau riset lebih lanjut.

Setelah semua halaman diproses, seluruh berita pasar dari IQPlus berhasil dikumpulkan dan tersimpan di database. Dengan sistem otomatis ini, pengambilan berita jadi lebih cepat dan efisien tanpa perlu membuka situsnya secara manual setiap hari. Nantinya, data ini bisa dipakai untuk berbagai analisis, seperti mencari pola dalam berita atau memahami sentimen pasar dari isi berita yang dikumpulkan.

C. Simpan Ke MongoDB

1) YFINANCE

The screenshot shows the MongoDB Compass interface connected to the 'localhost:27017/Yfinance.ABM Investama Tbk' database. The 'Yfinance' collection is selected. The interface displays a list of documents with their IDs and some fields like Date, Open, High, Low, Close, Volume, Dividends, and Stock Splits. The first document's details are expanded:

```
_id: ObjectId('68014c96228357e5eaaaf9145')
Date : "2014-01-02T09:00:00+07:00"
Open : 2191.1861177572
High : 2191.1861177572
Low : 2172.9978027344
Close : 2172.9978027344
Volume : 478000
Dividends : 0
Stock Splits : 0
```

Below it, another document is partially visible:

```
_id: ObjectId('68014c96228357e5eaaaf9146')
Date : "2014-01-03T09:00:00+07:00"
Open : 2172.9978027344
High : 2172.9978027344
Low : 2172.9978027344
Close : 2172.9978027344
Volume : 0
Dividends : 0
Stock Splits : 0
```

At the bottom, a third document is shown:

```
_id: ObjectId('68014c96228357e5eaaaf9147')
Date : "2014-01-06T00:00:00+07:00"
Open : 2169.3759885979
High : 2169.3759885979
Low : 2169.3759885979
Close : 2169.3759885979
Volume : 18400
```

Gambar 9. Isi Collection dari Yfinance

Setelah berhasil mengambil data harga saham dari Yahoo Finance menggunakan pustaka yfinance, langkah berikutnya adalah menyimpannya ke dalam MongoDB. Data yang diperoleh mencakup riwayat harga saham harian dari tahun 2014 hingga

2025 untuk setiap perusahaan yang terdaftar dalam file Daftar_Saham.csv. Kami menyimpan data ini dalam sebuah database bernama Yfinance, di mana setiap perusahaan memiliki koleksi tersendiri yang dinamai sesuai dengan nama perusahaannya.

MongoDB dipilih sebagai sistem penyimpanan karena strukturnya yang fleksibel dan mendukung format JSON, sehingga mempermudah proses pengolahan dan analisis data di tahap selanjutnya. Proses penyimpanan dimulai dengan membaca daftar kode saham dari file CSV, kemudian setiap kode saham digunakan untuk mengambil data riwayat harga menggunakan fungsi Ticker.history() dari pustaka yfinance. Data yang diperoleh dikonversi ke dalam format JSON dan disimpan ke dalam koleksi MongoDB menggunakan pustaka PyMongo.

Sebelum menyimpan data sesuai dengan tanggal yang diinginkan, dilakukan pengecekan terlebih dahulu apakah data yang diperoleh tidak kosong, untuk menghindari penyimpanan dokumen yang tidak memiliki nilai. Jika data tersedia, maka insert_many() digunakan untuk menyimpan seluruh riwayat harga saham ke dalam database.

Dalam proses penyimpanan ini, kami menemukan bahwa durasi pengambilan dan penyimpanan data sangat bergantung pada jumlah saham yang diproses dan koneksi ke server Yahoo Finance. Jika respons dari server lambat atau permintaan diblokir sementara karena terlalu banyak permintaan dalam waktu singkat, maka proses pengambilan data bisa menjadi lebih lama. Oleh karena itu, untuk meningkatkan efisiensi, optimasi dapat dilakukan dengan menambahkan mekanisme sleep antara permintaan atau menggunakan teknik paralelisasi untuk mempercepat pengambilan data dalam jumlah besar.

2) IDX(Indonesia Stock Exchange)

The screenshot shows the Compass MongoDB interface. On the left, the 'Connections' sidebar lists 'localhost:27017' and 'BeritaPoser' (with sub-collections like MarketNews, MarketNewsSummary, RingkasanStockNews, and StockNews). Below it is the 'IDXData' database with its own sub-collections: 'Financial_2021', 'Financial_2022', 'Financial_2023', 'Financial_2024', 'Yfinance_Final', 'admin', 'config', 'idx_final', 'local', and 'yfinance'. The main area shows the 'Financial_2021' collection with 737 documents. A search bar at the top says 'Type a query: { field: 'value' } or generate_query'. Below the search bar are buttons for 'ADD DATA', 'EXPORT DATA', 'UPDATE', and 'DELETE'. The results list shows several documents, each with a unique _id, company_code (e.g., 'ABDA', 'AALI', 'ABMH', 'ACES', 'ADCH', 'ACST'), year ('2021'), period ('audit'), and a data object containing audit details.

Gambar 10. Isi Collection dari IDXData

Secara Otomatis dalam program sudah dibuat script untuk melakukan pengunduhan file **instance.zip** yang mana berisi [instance.xbrl](#) dan [Taxonomy.xsd](#), kedua file ini merupakan format xml sehingga program kami melakukan ekstraksi atau pengambilan kedua file tersebut lalu diubah ke format JSON.

Detail data yang kami scraping sebanyak total **1844 dokumen** dari rentang 2021-2024, tapi sebenarnya bisa lebih banyak sehingga untuk projek kedepannya kami berencana melakukan scraping ulang agar dapat data yang lebih banyak dan lengkap.

Setelah data dalam format JSON, program kami secara otomatis mengirimkan dan menyimpan data tersebut di MongoDB melalui MongoClient dari library Pymongo. Dalam MongoDB kami menyimpannya dalam direktori utama yaitu “**IDX_Financial_Data**” dan memisahkannya lagi berdasarkan Tahun data tersebut.

Durasi proses penyimpanan selama **3 jam 54 menit**. Selenium digunakan untuk scraping, tetapi proses rendering halaman membuatnya lebih lambat dibandingkan

BeautifulSoup. Kestabilan jaringan dan respons server IDX juga memengaruhi waktu penyimpanan.

3) IQPLUS

```

_id: ObjectId('67d6397d05f6d8ac5c011881')
judul : "EKONOMI: GUBERNUR OPTIMISTIS JATIM TETAP PRODUSEN BERAS TERTINGGI NASIONAL"
waktu : "14/03/25 - 04:44"
link : "http://www.iqplus.info/news/market_news/ekonomi-gubernur-optimistis-jatim_"
tanggal_artikel : "Friday 14/Mar/2025 at 16:44"
konten : "IQPlus, (14/3) – Gubernur Jawa Timur Khofifah Indar Parawansa optimistis...""

_id: ObjectId('67d6397d05f6d8ac5c011880')
judul : "IFG DAN STIH ADHYAKSA GELAR SEMINAR NASIONAL"
waktu : "14/03/25 - 04:41"
link : "http://www.iqplus.info/news/market_news/ekon-ifg-dan-stih-adhyaksa-gelar_"
tanggal_artikel : "Friday 14/Mar/2025 at 17:05"
konten : "IQPlus, (14/3) – Penahaman generasi muda terhadap literasi dan tata ke...""

_id: ObjectId('67d6397d05f6d8ac5c011885')
judul : "EKONOMI: BI PENGUNJUNGAN QRS TAP DIHARAPKAN SEMAKIN TINGKATKAN INKLUSI KE..."
waktu : "14/03/25 - 04:41"
link : "http://www.iqplus.info/news/market_news/ekon-bi-pengunjungan-qrs-tap-d...""
tanggal_artikel : "Friday 14/Mar/2025 at 16:41"
konten : "IQPlus, (14/3) – Bank Indonesia (BI) menyampaikan bahwa implementasi L...""

_id: ObjectId('67d6397d05f6d8ac5c011884')
judul : "EKONOMI: IHSG SESI III DITUTUP MELEAH 131 POIN KE LEVEL 6.515"
waktu : "14/03/25 - 04:23"
link : "http://www.iqplus.info/news/market_news/ekon-ihsg-sesi-iii-ditutup-meleah_"
tanggal_artikel : "Friday 14/Mar/2025 at 16:23"
konten : "IQPlus, (14/3) – Indeks Harga Saham Gabungan (IHSG) ditutup meleah 13...""

_id: ObjectId('67d6397d05f6d8ac5c011883')
judul : "PEGBRI: PEMBAGIAN BAYAR KUPON OBLIGASI TAHUN 2025"
waktu : "14/03/25 - 04:33"
link : "http://www.iqplus.info/news/market_news/ekon-pegbri-pembagian-bayar-kupon-obl...""
tanggal_artikel : "Friday 14/Mar/2025 at 16:33"
konten : "IQPlus, (14/3) – PT Pegadaian melakukan pembayaran kupon Obligasi Berik...""

```

Gambar 11. Isi Collection dari BeritaPasar (IQPlus)

Setelah berhasil mengambil data berita dari situs IQPlus, langkah berikutnya adalah menyimpannya ke dalam MongoDB. Kami menyimpan berita ini dalam dua koleksi, yaitu iqplus_stock_news yang berisi berita terkait saham, dan iqplus_market_news yang berisi berita tentang kondisi pasar secara umum.

Kami memilih MongoDB karena format penyimpanannya yang fleksibel, mirip dengan JSON, sehingga lebih mudah dalam pengolahan data. Database yang kami gunakan bernama `scraping_db`, di mana setiap berita yang tersimpan memiliki informasi lengkap, seperti judul berita, waktu publikasi di halaman utama, link berita, tanggal dalam artikel, serta isi lengkap berita.

Sebelum berita disimpan, kami melakukan pengecekan terlebih dahulu untuk memastikan bahwa berita tersebut belum ada di database. Ini penting untuk menghindari data ganda yang bisa membuat ukuran database semakin besar dan sulit

dikelola. Jika berita belum tersimpan sebelumnya, maka data tersebut akan dimasukkan ke koleksi yang sesuai menggunakan PyMongo.

Dalam proses penyimpanan data hasil scraping, kami menemukan bahwa waktu yang dibutuhkan cukup lama. Salah satu faktor utama yang mempengaruhi durasi ini adalah jumlah berita yang sangat banyak. Semakin banyak berita yang harus diambil dan disimpan, semakin lama pula keseluruhan proses berlangsung.

Selain itu, kami menggunakan Selenium untuk mengambil data dari situs IQPlus. Selenium memerlukan proses rendering halaman secara langsung, yang berarti setiap kali mengambil data, halaman web harus dibuka dan dimuat sepenuhnya. Hal ini berbeda dengan metode BeautifulSoup, yang hanya membaca struktur HTML tanpa perlu merender tampilan, sehingga dapat bekerja lebih cepat.

Kami juga menemukan bahwa kecepatan proses ini sangat bergantung pada respons dari server IQPlus dan kestabilan jaringan internet. Jika server merespons dengan lambat atau jaringan tidak stabil, waktu yang dibutuhkan untuk mengambil dan menyimpan berita menjadi lebih lama. Dalam beberapa kasus, proses ini dapat memakan waktu dari beberapa menit hingga berjam-jam, tergantung pada jumlah berita yang diproses.

D. Tranformasi Data

1) YFINANCE

- Arsitektur Proses Transformasi Data

Database Sumber: Yfinance

Database Tujuan: Yfinance_Final

Sumber Data: CSV "Daftar_Saham.csv" berisi daftar nama perusahaan

Alat Transformasi: Apache Spark dengan MongoDB Connector

- Alur Proses Transformasi

Proses ini dimulai dengan membaca daftar nama perusahaan dari file CSV bernama "Daftar_Saham.csv". Setiap nama perusahaan dalam file tersebut mewakili satu koleksi data saham di MongoDB. Sistem akan membuat sesi kerja terpisah menggunakan Apache Spark, yang bertugas untuk mengambil dan memproses data.

Setelah data saham berhasil dimuat dari MongoDB, langkah pertama adalah memastikan bahwa kolom tanggal (Date) sudah dalam format waktu (timestamp) yang tepat. Ini penting agar data bisa diolah lebih lanjut berdasarkan periode waktu.

Selanjutnya, sistem membuat kolom baru bernama period_key, yaitu hasil dari memformat tanggal ke tiga jenis format:

- Harian: "yyyy-MM-dd"
- Bulanan: "yyyy-MM"
- Tahunan: "yyyy"

```
_id: ObjectId('68014c96228357e5eaaf9145')
Close : 2172.9978027344
Date : 2014-01-01T17:00:00.000+00:00
Dividends : 0
High : 2191.1061177572
Low : 2172.9978027344
Open : 2191.1061177572
Stock Splits : 0
Volume : 478000
period_key : "2014-01-02"
```

Gambar 11. Contoh hasil dari data saham yang telah di transformasi

Untuk data harian, tidak dilakukan agregasi khusus karena data yang tersedia memang sudah harian (per tanggal), bukan per jam atau lebih rinci. Jadi, data ini hanya dipastikan sudah rapi dan langsung ditulis ulang ke koleksi baru di MongoDB.

```
_id: ObjectId('6801c145cf6fe1723ef159d')
period_key : "2024-01"
avg_open : 47.64142660346818
avg_high : 48.10478264507273
avg_low : 46.7568377454546
avg_close : 47.64142660661364
avg_volume : 737913.6363636364
avg_dividends : 0
avg_stock_splits : 0
sum_open : 1048.1113852763
sum_high : 1058.3052181916
sum_low : 1028.65043104
sum_close : 1048.1113853455001
sum_volume : 16234100
sum_dividends : 0
sum_stock_splits : 0
max_open : 48.1890297684
max_high : 49.115737915
max_low : 47.2623176575
max_close : 48.1890296936
max_volume : 7473300
max_dividends : 0
max_stock_splits : 0
min_open : 46.3356018066
min_high : 46.3356018066
min_low : 46.3356018066
min_close : 46.3356018066
min_volume : 0
min_dividends : 0
min_stock_splits : 0
std_open : 0.6803368828677623
std_high : 0.4877856586123009
std_low : 0.4722966902338275
std_close : 0.6803371042079132
std_volume : 1535526.8418969912
std_dividends : 0
std_stock_splits : 0
row_count : 22
month_number : 28
agg_type : "month"
```

Gambar 12. Hasil data bulanan

```

_id: ObjectId('6801b638cfc6fe1723ed7165')
period_key : "2014"
avg_open : 2044.4767285005182
avg_high : 2061.483242980596
avg_low : 2022.5960767129131
avg_close : 2056.6980795938534
avg_volume : 115685.24590163934
avg_dividends : 0.020409836065573773
avg_stock_splits : 0
sum_open : 498852.32175412646
sum_high : 503001.9112872655
sum_low : 493513.4427179508
sum_close : 501834.33142090024
sum_volume : 28227200
sum_dividends : 4.98
sum_stock_splits : 0
max_open : 2317.8642723881
max_high : 2426.5141601562
max_low : 2281.6474609375
max_close : 2426.5141601562
max_volume : 3215000
max_dividends : 4.98
max_stock_splits : 0
min_open : 1665.0179443359
min_high : 1675.9004145604
min_low : 1581.5857421875
min_close : 1665.0179443359
min_volume : 0
min_dividends : 0
min_stock_splits : 0
std_open : 119.340019770163
std_high : 118.3808572685263
std_low : 135.45902779717005
std_close : 120.85036090112281
std_volume : 378978.00313423213
std_dividends : 0.318811831032911
std_stock_splits : 0
row_count : 244
year_number : 1
agg_type : "year"

```

Gambar 13. Hasil data tahunan

Namun, untuk data bulanan dan tahunan, prosesnya berbeda. Sistem akan mengelompokkan data berdasarkan period_key (misalnya, semua data bulan Januari 2023 atau semua data tahun 2022), lalu menghitung agregasi menggunakan berbagai tipe agregasi seperti rata-rata (AVG), total (SUM), nilai tertinggi (MAX), nilai terendah (MIN), deviasi standar (STD). Agregasi ini dilakukan untuk metrik seperti open, high, low, close, volume, dividend, dan stock splits dalam setiap periode yang ditentukan, baik bulanan maupun tahunan.

Meskipun seluruh proses ini dijalankan satu per satu untuk tiap perusahaan (belum paralel), waktu pemrosesan tetap cepat berkat optimasi dari Spark. Di akhir proses, sistem juga memberikan ringkasan statistik, seperti:

- Jumlah saham yang berhasil diproses: 951 dari 951 (100%)
- Total dokumen yang disimpan ke MongoDB: 1.833.324
- Durasi total proses: sekitar 56 menit 45 detik (3.405,41 detik)
- Rata-rata waktu pemrosesan per saham: 3,58 detik
- Kecepatan pemrosesan: sekitar 538 dokumen per detik

- Penjelasan Atribut

Atribut agregasi perhari :

Atribut	Penjelasan
_id	ID unik dokumen di MongoDB
Close	Harga penutupan saham
Date	Tanggal transaksi
Dividends	Nilai dividen
High	Harga tertinggi saham
Low	Harga terendah saham
Open	Harga pembukaan saham
Stock Splits	Pemecahan saham
Volume	Volume perdagangan
period_key	Kunci periode dalam format "yyyy-mm-dd"

Tabel 1. Atribut data agregasi harian

Atribut agregasi untuk bulanan dan tahunan :

Atribut	Penjelasan
_id	ID unik dokumen yang dibuat otomatis oleh MongoDB.
period_key	Kunci waktu agregasi: - Bulanan: format yyyy-MM - Tahunan: format yyyy
month_number	(Hanya untuk bulanan) Nomor urut bulan berdasarkan total bulan sejak data dimulai.
year_number	(Hanya untuk tahunan) Nomor urut tahun berdasarkan total tahun sejak data dimulai.
row_count	Jumlah data harian (baris) yang dihitung dalam periode tersebut.
agg_type	Jenis agregasi: "month" untuk bulanan, "year" untuk tahunan.

Tabel 2. Hasil atribut data agregasi bulanan

Atribut Baru dari Close, Dividends,High,Low,Open,Stock Splits,Volume.

AVG	SUM	MAX	MIN	STD
avg_open	sum_open	max_open	min_open	std_open
avg_high	sum_high	max_high	min_high	std_high
avg_low	sum_low	max_low	min_low	std_low
avg_close	sum_close	max_close	min_close	std_close
avg_volume	sum_volume	max_volume	min_volume	std_volume
avg_dividends	sum_dividends	max_dividends	min_dividends	std_dividends

avg_stock_splits	sum_stock_splits	max_stock_splits	min_stock_splits	std_stock_splits

Tabel 3. Hasil atribut data agregasi tahunan

- Rata-rata harga pembukaan, penutupan, tertinggi, terendah, volume transaksi, dividen, dan stock split.
- Jumlah total (sum) dari masing-masing metrik tersebut.
- Nilai maksimum dan minimum untuk mengukur batas tertinggi dan terendah.
- Standar deviasi (std) untuk melihat variasi/fluktuasi data dalam periode tersebut.
- Fitur Lain

Program dilengkapi dengan fitur konfigurasi logging menggunakan modul logging Python dengan format timestamp, level, dan pesan, serta level log INFO dan ERROR. Penanganan kesalahan diterapkan pada pembacaan file CSV, validasi kolom "Date", penggunaan try-except block, dan pencatatan kesalahan dalam log. Untuk optimasi performa, program menggunakan Apache Spark untuk pemrosesan data berskala besar, konfigurasi memori 4GB untuk driver dan executor, serta window functions untuk penomoran urut dalam agregasi. Program juga menyediakan ringkasan proses berupa pencatatan total saham yang berhasil diproses, total data yang berhasil disimpan, dan pengukuran waktu eksekusi.

- Keluaran/Output

```

2025-04-18 10:13:24,536 - INFO - ✓ Agregasi year selesai untuk Canaya Bintang Medan Tbk
2025-04-18 10:13:24,457 - INFO - Membaca data dari MongoDB koleksi: Aesler Grup Internasional Tbk
2025-04-18 10:13:24,849 - INFO - ✓ Data harian disimpan untuk Aesler Grup Internasional Tbk
2025-04-18 10:13:26,383 - INFO - ✓ Agregasi month selesai untuk Aesler Grup Internasional Tbk
2025-04-18 10:13:27,189 - INFO - ✓ Agregasi year selesai untuk Aesler Grup Internasional Tbk
2025-04-18 10:13:28,146 - INFO - Spark session dihentikan

 Ringkasan Pengambilan Data:
✖ Total saham yang berhasil ditransformasi: 951 dari 951 saham
📦 Total data yang berhasil disimpan di MongoDB: 1833324 dokumen
⌚ Waktu eksekusi: 3405.41 detik
✓ Proses selesai!
2025-04-18 10:13:28,203 - INFO - Closing down clientserver connection

```

Gambar 14. Log dari transformasi yfinance

Setelah proses transformasi selesai, data disimpan dalam format JSON di database Yfinance_Final, dengan setiap koleksi saham berisi data harian lengkap serta agregasi bulanan dan tahunan yang dilengkapi dengan metrik relevan.. Selama proses pengambilan data, sebanyak 951 saham berhasil diambil dari target 951 saham, yang berarti 100% berhasil tercapai. Total volume data yang terkumpul adalah 1.833.324 dokumen yang disimpan di MongoDB. Proses pengambilan data ini memakan waktu 3.405,41 detik atau sekitar 56 menit 45 detik, dengan rata-rata waktu per saham sekitar 3,58 detik. Kecepatan pemrosesan mencapai sekitar 538 dokumen per detik. Proses pengambilan data dilakukan secara sekuensial, artinya data untuk setiap saham diproses satu per satu sebelum berlanjut ke saham berikutnya. Meskipun belum menggunakan paralelisasi, kecepatan yang tercapai sudah menunjukkan efisiensi yang cukup baik untuk mengelola data dalam jumlah besar dan mendukung analisis yang lebih mendalam.

2) IDX (Indonesia Stock Exchange)

- Arsitektur Proses Transformasi Data

Sumber Data : Hasil scraping instance.zip di IDX yang sudah diproses dan disimpan di MongoDB

Nama Database MongoDB: IDXData

Collection yang Diproses: Financial_2021, Financial_2022, Financial_2023, Financial_2024

- Alur Proses Transformasi

Proses transformasi data dimulai dengan mengatur dua jenis koneksi. Yang pertama menggunakan Apache Spark sebagai metode utama, dan yang kedua menggunakan PyMongo sebagai cadangan jika Spark mengalami kendala. Koneksi ke database MongoDB menggunakan alamat mongodb://localhost:27017 (lokal). Setelah koneksi siap, data dibaca dari koleksi sumber yang ada di dalam database bernama *IDXData*. Dari situ, sistem akan mengenali daftar emiten yang ada di tiap koleksi berdasarkan kode emiten, lalu memproses data per emiten.

Selanjutnya, data akan melalui tahap transformasi. Di tahap ini, sistem akan mengambil data-data keuangan penting dari tiap emiten, mengubah tipe data yang diperlukan, dan menghitung beberapa metrik tambahan. Metrik ini termasuk rasio keuangan seperti current ratio dan asset to equity ratio, margin seperti gross margin, operating margin, dan net margin, serta perhitungan laba-rugi dan analisis arus kas. Untuk detailnya diawal setiap emiten memiliki atribut kisaran 300-500 atribut tergantung isi laporan emiten-nya, dan setelah dipertimbangkan ada 32 atribut yang akan dipertahankan dan diproses, yaitu:

No	Atribut	Penjelasan
Informasi Umum		
1.	company_code	Kode unik perusahaan/emiten yang terdaftar di bursa
2.	year	Tahun pelaporan laporan keuangan
3.	period	Periode pelaporan (misalnya audit, interim, dll.)
4.	company_name	Nama lengkap perusahaan/emiten
5.	sector	Sektor industri tempat perusahaan beroperasi
6.	subsector	Subsektor industri yang lebih spesifik dari sektor utama
Metrik Keuangan		
7.	revenue	Total pendapatan atau penjualan perusahaan selama periode pelaporan
8.	gross_profit	Laba kotor, yaitu pendapatan dikurangi biaya produksi langsung (HPP)

9.	operating_profit	Laba operasional. Dihitung sebagai laba sebelum pajak (ProfitLossBeforeIncomeTax) - biaya keuangan (FinanceCosts)
10.	net_profit	Laba bersih setelah memperhitungkan semua pendapatan, biaya, dan pajak
11.	ebitda	Earnings Before Interest, Taxes, Depreciation, and Amortization. Dihitung sebagai laba sebelum pajak (ProfitLossBeforeIncomeTax) + biaya keuangan (FinanceCosts)
12.	basic_eps	Earnings Per Share, laba bersih dibagi jumlah saham beredar
Metrik Neraca		
13.	cash	Jumlah kas dan setara kas perusahaan
14.	total_assets	Total aset perusahaan, mencakup aset lancar dan tidak lancar
15.	short_term_borrowing	Pinjaman jangka pendek, termasuk pinjaman bank jangka pendek atau cicilan pinjaman jangka panjang yang jatuh tempo
16.	long_term_borrowing	Pinjaman jangka panjang dari bank
17.	total_equity	Total ekuitas pemegang saham
18.	total_liabilities	Total kewajiban perusahaan, mencakup

		kewajiban lancar dan tidak lancar
19.	current_assets	Total aset lancar, seperti kas, piutang, dan persediaan
20.	current_liabilities	Total kewajiban lancar, seperti utang dagang dan utang pajak
Metrik Arus Kas		
21.	cash_from_operations	Arus kas bersih dari aktivitas operasi
22.	cash_from_investing	Arus kas bersih dari aktivitas investasi
23.	cash_from_financing	Arus kas bersih dari aktivitas pendanaan
Biaya Operasional		
24.	selling_expenses	Biaya penjualan, termasuk biaya pemasaran dan distribusi
25.	g_and_a_expenses	Biaya umum dan administrasi, seperti gaji manajemen dan biaya kantor
26.	operating_expenses	Total biaya operasional. Dihitung dari selling_expenses + g_and_a_expenses diatas
Rasio Keuangan		
27.	current_ratio	Rasio lancar untuk mengukur kemampuan perusahaan memenuhi kewajiban jangka pendek.

		Dihitung dari current_assets / current_liabilities.
28.	asset_to_equity_ratio	Rasio aset terhadap ekuitas untuk mengukur leverage perusahaan. Dihitung dari total_assets / total_equity, dengan penanganan null
29.	debt_to_equity_ratio	Rasio utang terhadap ekuitas untuk mengukur tingkat utang perusahaan relatif terhadap modalnya. Dihitung dari total_liabilities / total_equity, dengan penanganan null
Margin Keuangan		
30.	gross_margin_pct	Margin laba kotor. Dihitung dari gross_profit / revenue * 100, dengan penanganan null
31.	operating_margin_pct	Margin laba operasional. Dihitung dari operating_profit / revenue * 100, dengan penanganan null
32.	net_margin_pct	Margin laba bersih untuk menunjukkan profitabilitas keseluruhan Dihitung dari net_profit * 100

Tabel 4. Hasil atribut dari transformasi IDX

Setelah proses selesai, data yang sudah diolah akan disimpan ke dalam database tujuan bernama *idx_final*. Nama koleksi hasil akan mengikuti nama koleksi aslinya, hanya ditambahkan "*_final*" di belakangnya. Misalnya, *Financial_2021* akan menjadi *Financial_2021_final*, dan *Financial_2022* menjadi *Financial_2022_final*.

- Fitur Lain

Program transformasi data ini memiliki beberapa fitur/proses lain, yaitu penanganan nilai *null* dan *error handling*. Nilai-nilai *null* secara otomatis diubah menjadi nol agar tidak mengganggu perhitungan, dan jika terjadi error, proses akan mencatatnya dalam log yang akan dicetak di terminal jadi memudahkan proses debugging dan lebih tergambar jelas alur programnya.

- Keluaran/Output

```
2025-04-17 18:12:27,298 - INFO - Selesai memproses Financial_2024 dalam 0:01:17.797326
2025-04-17 18:12:27,326 - INFO - === LAPORAN PERFORMA KESELURUHAN ===
2025-04-17 18:12:27,326 - INFO - Total waktu eksekusi: 0:16:52.263318
2025-04-17 18:12:27,326 - INFO - Total dokumen diproses: 1844
2025-04-17 18:12:27,326 - INFO - Total perusahaan diproses: 1844
2025-04-17 18:12:27,326 - INFO - Rata-rata dokumen per detik: 1.82
2025-04-17 18:12:27,326 - INFO - Rata-rata penggunaan CPU: 48.49%
2025-04-17 18:12:27,326 - INFO - Rata-rata penggunaan Memory: 76.87%
2025-04-17 18:12:27,326 - INFO - Total error: 0
2025-04-17 18:12:27,326 - INFO - Proses selesai pada: 2025-04-17 18:12:27
2025-04-17 18:12:27,326 - INFO - Total waktu eksekusi: 0:16:56.222896
2025-04-17 18:12:27,354 - INFO - Database views berhasil dibuat
2025-04-17 18:12:27,833 - INFO - Closing down clientserver connection
```

Gambar 15. Log dari hasil transformasi IDX

Hasil akhir dari proses transformasi ini adalah kumpulan data JSON yang disimpan dalam koleksi final pada database *idx_final* berisi atribut-atribut yang sudah diseleksi tapi banyaknya atribut tergantung pada data awal emiten, jadi tidak selalu 32 atribut, jika atribut tidak ada nanti akan ditampilkan 0 atau tidak ada/dilaporkan. Berdasarkan laporan kinerja, proses ini memakan waktu 16 menit 52 detik dan berhasil memproses 1.844 dokumen dari 1.844 perusahaan. Rata-rata kecepatan pemrosesan adalah 1,82 dokumen per detik.

Selama proses berlangsung, sistem menggunakan CPU sebesar 48,49% dan memori sebesar 76,87%. Yang paling penting, seluruh proses berjalan lancar

tanpa adanya kesalahan. Pemrosesan dimulai pukul 18:12:27 dan selesai pada hari yang sama. Total waktu yang dicatat, termasuk semua tahapan, adalah 16 menit 56 detik.

3) IQPLUS

- Arsitektur Proses Transformasi Data

Sumber Data : Hasil scraping scraping situs *IQPlus*, yang berisi informasi dan artikel keuangan perusahaan publik di Indonesia. sudah diproses dan disimpan di MongoDB

Nama Database MongoDB: BeritaPasar

Collection yang Diproses: MarketNews dan StockNews

- Alur Transformasi

Proses peringkasan berita dimulai dengan menghubungkan sistem ke database MongoDB dan mengambil seluruh dokumen berita untuk diproses satu per satu dalam bentuk list. Berbeda dengan proses transformasi data Yfinance dan IDX yang memanfaatkan Apache Spark untuk pemrosesan skala besar, peringkasan berita dilakukan secara sekvensial menggunakan Python standar dan library PyMongo. Pendekatan ini dipilih karena peringkasan teks membutuhkan pemrosesan mendalam per dokumen, bukan pemrosesan paralel.

Untuk setiap dokumen, sistem menghitung jumlah token menggunakan tokenizer dari model BART (facebook/bart-large-cnn). Jika jumlah token tidak melebihi 1024 (batas maksimum model), dokumen langsung diringkas. Namun, jika melebihi, digunakan algoritma split-merge untuk membagi teks berdasarkan kalimat menjadi bagian-bagian kecil yang masing-masing berisi maksimal 1024 token. Kalimat ditambahkan satu per satu hingga batas token tercapai, lalu bagian tersebut disimpan dan bagian baru dibuat.

Setiap bagian kemudian diringkas menggunakan model BART dengan parameter `max_length=512` dan `min_length=30`. Ringkasan tiap bagian dikumpulkan dalam list, lalu digabungkan dan diringkas ulang untuk menghasilkan ringkasan final yang lebih padat (dengan `max_length=250` dan `min_length=50`).

Untuk menjaga sistem, setiap langkah dilengkapi mekanisme try-except agar error saat memproses dokumen tertentu tidak menghentikan seluruh proses. Kesalahan dicatat dalam log, dan dokumen berikutnya tetap diproses. Sistem juga memiliki fitur monitoring dan logging yang mencatat progres setiap 10 dokumen dan total waktu eksekusi di akhir proses. Dokumen hasil ringkasan berita memiliki struktur dengan atribut sebagai berikut:

No	Atribut	Penjelasan
1.	_id	ID unik dokumen di MongoDB
2.	original_id	ID dokumen asli dari koleksi sumber, disimpan sebagai string untuk memudahkan referensi silang
3.	judul	Judul berita yang dipertahankan dari dokumen asli
4.	konten	Konten berita asli secara utuh, dipreservasi untuk perbandingan dengan hasil ringkasan
5.	summary	Hasil ringkasan berita yang dihasilkan oleh model BART menggunakan algoritma split-merge
6.	timestamp	Keperluan log tracing waktu dan kronologi proses

Tabel 5. Atribut dari hasil transformasi database BeritaPasar(IQplus)

• Keluaran/Output

```
PS C:\Users\HP\Downloads\tugas2> & C:/Users/HP/AppData/Local/Programs/Python/Python313/python.exe c:/Users/HP/Downloads/tugas2/ringkas_4.py
Device set to use cpu
 Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.
 Your max_length is set to 311, but your input_length is only 110. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=75)
 Your max_length is set to 250, but your input_length is only 145. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=72)
2025-04-18 15:13:28,702 - INFO - BERMASIL: Dokumen index 1 diproses dalam 18.62 detik
2025-04-18 15:13:35,977 - INFO - BERMASIL: Dokumen index 2 diproses dalam 7.27 detik
2025-04-18 15:13:39,938 - INFO - BERMASIL: Dokumen index 3 diproses dalam 3.96 detik
Your max_length is set to 250, but your input_length is only 184. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=100)
2025-04-18 15:13:44,051 - INFO - BERMASIL: Dokumen index 4 diproses dalam 5.87 detik
2025-04-18 15:13:51,186 - INFO - BERMASIL: Dokumen index 5 diproses dalam 5.38 detik
2025-04-18 15:13:51,844 - INFO - BERMASIL: Dokumen index 6 diproses dalam 6.66 detik
2025-04-18 15:14:06,347 - INFO - BERMASIL: Dokumen index 7 diproses dalam 8.50 detik
2025-04-18 15:14:15,518 - INFO - BERMASIL: Dokumen index 8 diproses dalam 9.17 detik
Your max_length is set to 250, but your input_length is only 207. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=103)
2025-04-18 15:14:21,651 - INFO - BERMASIL: Dokumen index 9 diproses dalam 22.04 detik
2025-04-18 15:14:47,257 - INFO - BERMASIL: Dokumen index 10 diproses dalam 6.80 detik
2025-04-18 15:15:39,342 - INFO - BERMASIL: Dokumen index 11 diproses dalam 7.31 detik
2025-04-18 15:15:45,998 - INFO - BERMASIL: Dokumen index 12 diproses dalam 4.51 detik
2025-04-18 15:15:56,741 - INFO - BERMASIL: Dokumen index 13 diproses dalam 7.64 detik
2025-04-18 15:15:56,741 - INFO - BERMASIL: Dokumen index 14 diproses dalam 5.90 detik
2025-04-18 15:15:57,854 - INFO - BERMASIL: Dokumen index 15 diproses dalam 5.21 detik
2025-04-18 15:15:52,489 - INFO - BERMASIL: Dokumen index 16 diproses dalam 7.03 detik
2025-04-18 15:15:52,489 - INFO - BERMASIL: Dokumen index 17 diproses dalam 8.03 detik
2025-04-18 15:15:52,489 - INFO - BERMASIL: Dokumen index 18 diproses dalam 6.46 detik
2025-04-18 15:15:48,616 - INFO - BERMASIL: Dokumen index 19 diproses dalam 9.27 detik
2025-04-18 15:15:56,885 - INFO - BERMASIL: Dokumen index 20 diproses dalam 6.27 detik
2025-04-18 15:16:08,814 - INFO - BERMASIL: Dokumen index 21 diproses dalam 11.93 detik
2025-04-18 15:16:14,827 - INFO - BERMASIL: Dokumen index 22 diproses dalam 6.01 detik
2025-04-18 15:16:24,104 - INFO - BERMASIL: Dokumen index 23 diproses dalam 9.28 detik
2025-04-18 15:16:30,772 - INFO - BERMASIL: Dokumen index 24 diproses dalam 6.67 detik
2025-04-18 15:16:44,331 - INFO - BERMASIL: Dokumen index 25 diproses dalam 8.58 detik
2025-04-18 15:16:44,331 - INFO - BERMASIL: Dokumen index 26 diproses dalam 6.33 detik
2025-04-18 15:16:51,563 - INFO - BERMASIL: Dokumen index 27 diproses dalam 6.71 detik
2025-04-18 15:16:57,418 - INFO - BERMASIL: Dokumen index 28 diproses dalam 5.85 detik
2025-04-18 15:17:02,235 - INFO - BERMASIL: Dokumen index 29 diproses dalam 4.82 detik
```

Gambar 16. Log Hasil Transformasi MarketNews

```
Last login: Fri Apr 18 14:03:38 on ttys013
helnizahmed@MacBook-Pro-Mini: ~ % cd /Users/helnizahmed/Kuliah/Semester 4/Big Data/Tugas1
helnizahmed@MacBook-Pro-Mini: Tugas 2$ python3 loplusplit.py
/Users/helnizahmed/Library/Python/3.9/lib/python/site-packages/urllib3/_int...py:36: NotOpenSSLWarning: urllib3 v2 only supports OpenSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. See: https://github.com/urllib3/urllib3/pull/2529
  warnings.warn(
Device set to use CPU
 Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.
2025-04-18 15:36:40,920 - INFO - BERMASIL: Dokumen index 1 diproses dalam 18.62 detik
2025-04-18 15:36:47,454 - INFO - BERMASIL: Dokumen index 2 diproses dalam 8.03 detik
2025-04-18 15:36:59,152 - INFO - BERMASIL: Dokumen index 3 diproses dalam 11.73 detik
2025-04-18 15:36:59,152 - INFO - BERMASIL: Dokumen index 4 diproses dalam 7.03 detik
Your max_length is set to 312, but your input_length is only 349. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=174)
Your max_length is set to 250, but your input_length is only 380. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=93)
2025-04-18 15:37:39,627 - INFO - BERMASIL: Dokumen index 5 diproses dalam 28.99 detik
2025-04-18 15:37:52,852 - INFO - BERMASIL: Dokumen index 6 diproses dalam 13.42 detik
2025-04-18 15:38:09,672 - INFO - BERMASIL: Dokumen index 7 diproses dalam 8.62 detik
2025-04-18 15:38:16,152 - INFO - BERMASIL: Dokumen index 8 diproses dalam 8.03 detik
2025-04-18 15:38:15,698 - INFO - BERMASIL: Dokumen index 9 diproses dalam 6.86 detik
2025-04-18 15:38:15,698 - INFO - BERMASIL: Dokumen index 10 diproses dalam 8.03 detik
2025-04-18 15:38:33,989 - INFO - BERMASIL: Dokumen index 11 diproses dalam 10.26 detik
Your max_length is set to 312, but your input_length is only 211. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=174)
Your max_length is set to 250, but your input_length is only 151. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=93)
2025-04-18 15:39:01,878 - INFO - BERMASIL: Dokumen index 12 diproses dalam 27.00 detik
2025-04-18 15:39:15,941 - INFO - BERMASIL: Dokumen index 13 diproses dalam 13.97 detik
2025-04-18 15:39:22,154 - INFO - BERMASIL: Dokumen index 14 diproses dalam 8.03 detik
2025-04-18 15:39:35,414 - INFO - BERMASIL: Dokumen index 15 diproses dalam 6.68 detik
2025-04-18 15:39:35,414 - INFO - BERMASIL: Dokumen index 16 diproses dalam 8.03 detik
2025-04-18 15:39:52,329 - INFO - BERMASIL: Dokumen index 17 diproses dalam 8.12 detik
2025-04-18 15:40:01,452 - INFO - BERMASIL: Dokumen index 18 diproses dalam 9.12 detik
2025-04-18 15:40:01,452 - INFO - BERMASIL: Dokumen index 19 diproses dalam 8.03 detik
2025-04-18 15:40:19,658 - INFO - BERMASIL: Dokumen index 20 diproses dalam 8.24 detik
Your max_length is set to 312, but your input_length is only 426. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=174)
Your max_length is set to 250, but your input_length is only 172. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=93)
2025-04-18 15:40:49,984 - INFO - BERMASIL: Dokumen index 21 diproses dalam 39.24 detik
2025-04-18 15:40:57,212 - INFO - BERMASIL: Dokumen index 22 diproses dalam 7.03 detik
2025-04-18 15:40:57,212 - INFO - BERMASIL: Dokumen index 23 diproses dalam 7.90 detik
2025-04-18 15:41:13,776 - INFO - BERMASIL: Dokumen index 24 diproses dalam 7.68 detik
2025-04-18 15:41:13,776 - INFO - BERMASIL: Dokumen index 25 diproses dalam 7.03 detik
2025-04-18 15:41:31,238 - INFO - BERMASIL: Dokumen index 26 diproses dalam 8.37 detik
2025-04-18 15:41:49,569 - INFO - BERMASIL: Dokumen index 27 diproses dalam 8.03 detik
2025-04-18 15:41:49,569 - INFO - BERMASIL: Dokumen index 28 diproses dalam 6.00 detik
2025-04-18 15:41:55,965 - INFO - BERMASIL: Dokumen index 29 diproses dalam 6.00 detik
2025-04-18 15:41:55,965 - INFO - BERMASIL: Dokumen index 30 diproses dalam 7.03 detik
2025-04-18 15:42:20,978 - INFO - BERMASIL: Dokumen index 31 diproses dalam 15.00 detik
2025-04-18 15:42:36,947 - INFO - BERMASIL: Dokumen index 32 diproses dalam 9.70 detik
2025-04-18 15:42:43,689 - INFO - BERMASIL: Dokumen index 33 diproses dalam 7.90 detik
2025-04-18 15:43:04,399 - INFO - BERMASIL: Dokumen index 34 diproses dalam 7.16 detik
2025-04-18 15:43:04,399 - INFO - BERMASIL: Dokumen index 35 diproses dalam 7.03 detik
2025-04-18 15:43:29,732 - INFO - BERMASIL: Dokumen index 36 diproses dalam 12.27 detik
2025-04-18 15:43:29,732 - INFO - BERMASIL: Dokumen index 37 diproses dalam 7.90 detik
2025-04-18 15:43:37,735 - INFO - BERMASIL: Dokumen index 38 diproses dalam 9.83 detik
2025-04-18 15:43:37,735 - INFO - BERMASIL: Dokumen index 39 diproses dalam 11.08 detik
2025-04-18 15:43:59,722 - INFO - BERMASIL: Dokumen index 40 diproses dalam 10.15 detik
2025-04-18 15:43:59,722 - INFO - BERMASIL: Dokumen index 41 diproses dalam 9.49 detik
2025-04-18 15:44:07,668 - INFO - BERMASIL: Dokumen index 42 diproses dalam 7.63 detik
2025-04-18 15:44:07,668 - INFO - BERMASIL: Dokumen index 43 diproses dalam 7.03 detik
Your max_length is set to 312, but your input_length is only 294. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=174)
Your max_length is set to 250, but your input_length is only 189. Since this is a summarization task, where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually, e.g. summarizer(..., max_length=93)
```

Gambar 17. Log Hasil Transformasi MarketNews

Hasil ringkasan final dari setiap dokumen disimpan ke dalam koleksi target RingkasanMarketNews dan RingkasanStockNews bersama dengan informasi penting dari dokumen asli seperti ID dokumen, judul berita, konten asli (utuh), dan timestamp berita. Pendekatan ini memastikan bahwa ringkasan yang dihasilkan tetap terkait dengan dokumen aslinya.

Berdasarkan data pengambilan hasil transformasi di atas dapat dilihat bahwa rata-rata pengambilan per 30 data adalah 8.768 detik Dalam satu jam, maka sekitar 410 data yang diambil. Untuk data 5183 MarketNews, waktu yang dibutuhkan adalah sekitar 12 jam 37 menit 24 detik. Sedangkan untuk 4279 data StockNews, waktu yang dibutuhkan adalah sekitar 10 jam 25 menit 18 detik.

Kode lengkap untuk Transformasi ini sudah kami unggah ke GitHub, dan bisa diakses di sini:

<https://github.com/salsilsulselsol/Tugas-2-Big-Data.git>