DOKUMEN SPESIFIKASI PROYEK *BIG DATA*

Pemodelan Harga Emas di Indonesia Menggunakan Regresi Linier Berganda pada Sistem *Big Data* dengan *Batch Processing*

Kelompok 10

Muhammad Bayu Syuhada	122450007
Eksanty F Sugma Islamiaty	122450001
Eli Dwi Putra Berema	122450064
Syalaisha Andini Putriansyah	122450121

PROGRAM STUDI SAINS DATA

FAKULTAS SAINS

INSTITUT TEKNOLOGI SUMATERA

Mei 2025

DAFTAR ISI

1. Pendahuluan

1.1 Tujuan Dokumen

Dokumen ini disusun untuk menjelaskan kebutuhan sistem dalam proyek Pemodelan Harga Emas di Indonesia Menggunakan Regresi Linier Berganda pada Sistem *Big Data* dengan *Batch Processing*. Proyek ini bertujuan untuk membangun model prediktif harga emas berdasarkan variabel ekonomi makro seperti BI Rate, inflasi, dan nilai tukar rupiah terhadap harga emas di indonesia. Dengan pendekatan ini, diharapkan sistem dapat memberikan gambaran yang lebih akurat terhadap faktor-faktor yang mempengaruhi fluktuasi harga emas di Indonesia.

Sistem akan dikembangkan menggunakan pendekatan *Big Data* berbasis *batch processing* untuk menangani volume data ekonomi yang besar dan historis. Arsitektur sistem akan menerapkan pola *Medallion Architecture* yang membagi data ke dalam tiga lapisan yaitu Bronze, Silver, dan Gold guna meningkatkan kualitas dan efisiensi pemrosesan. Teknologi seperti *Hadoop Distributed File System* (HDFS), Apache Spark, dan Apache Hive akan dimanfaatkan untuk menyimpan, mengolah, dan menganalisis data secara terdistribusi. Dengan sistem ini, hasil analisis dapat dimanfaatkan oleh analis dan pembuat kebijakan dalam mendukung keputusan berbasis data.

1.2 Lingkup Sistem

Sistem yang dirancang akan menjalankan proses pengumpulan, pengolahan, dan analisis data ekonomi yang berkaitan dengan harga emas di Indonesia. Pendekatan yang digunakan adalah data *lakehouse* karena mendukung proses analisis data secara mendalam menggunakan regresi linier berganda yang dapat dijalankan langsung melalui engine seperti Apache Spark dan Hive SQL dan pengolahan data dilakukan dengan berbasis *batch processing*, pertama data akan disimpan dan diolah secara bertahap melalui tiga lapisan yaitu *bronze* (data awal mentah), *silver* (data yang telah dibersihkan dan dinormalisasi), dan *gold* (data yang siap digunakan untuk analisis lebih lanjut). Hasil akhir akan berupa model prediksi dan informasi analitik yang bisa diakses melalui tools visualisasi atau diekstrak dengan query langsung.

2. Deskripsi Umum

2.1 Gambaran Umum Sistem

Sistem yang dikembangkan dalam proyek ini bertujuan untuk memodelkan harga emas di Indonesia berdasarkan data makro-ekonomi, yaitu BI Rate, inflasi, dan nilai tukar rupiah terhadap harga emas. Dengan memanfaatkan pendekatan *big data* dan *batch processing*, sistem ini memungkinkan pengolahan data dalam jumlah besar secara efisien dan terstruktur. Pemodelan data dalam hal ini yang digunakan adalah regresi linier berganda, yang mampu mengidentifikasi dan mengukur pengaruh masing-masing variabel independen terhadap pergerakan harga emas sebagai variabel dependen. Hasil dari penelitian ini akan membantu menyediakan informasi yang berguna bagi pengambil kebijakan, analis ekonomi, dan pihak lain yang berkepentingan terhadap tren harga emas. Variabel penelitian dapat dituliskan sebagai berikut:

Tabel 1. Atribut yang digunakan dalam penelitian

Variabel	Keterangan
X_1	BI Rate
X_2	Inflasi

X_3	Nilai Tukar Rupiah
Y	Fluktuasi Harga Emas

2.2 Perspektif Sistem

Sistem yang dikembangkan merupakan sistem *Big Data* berbasis *batch processing* yang berjalan dalam lingkungan Docker multi-container. Sistem ini mengintegrasikan berbagai komponen utama dari ekosistem Apache seperti *Hadoop Distributed File System* (HDFS) untuk penyimpanan data terdistribusi, Apache Spark untuk pemrosesan data *batch*, Apache Hive untuk menjalankan query SQL terhadap data skala besar, serta Apache Ambari sebagai alat monitoring dan manajemen cluster. Arsitektur sistem mengikuti pendekatan *Medallion Architecture*, yang memisahkan alur data ke dalam tiga lapisan yaitu Bronze, Silver, dan Gold. Pendekatan ini digunakan untuk memastikan kualitas data, mempermudah pelacakan transformasi, serta meningkatkan efisiensi dalam proses analisis.

2.3 Fungsi Sistem Utama

Fungsi utama dari sistem ini terbagi dalam empat tahapan utama berikut:

- Mengambil dan menyimpan data ekonomi mentah (Bronze):

  Sistem mengimpor data historis dari berbagai sumber seperti BI Rate, inflasi, kurs rupiah, dan harga emas. Data disimpan dalam format mentah pada HDFS.

- Membersihkan dan mentransformasi data (Silver):

  Data yang telah disimpan akan diproses untuk menghapus duplikasi, menangani missing value, dan melakukan normalisasi menggunakan metode Z-score agar siap digunakan dalam model analitik.

  $$Xstd = \frac{X - \mu}{\sigma}$$

  Persamaan di atas digunakan untuk proses normalisasi data menggunakan metode Z-score, Dalam rumus ini, dengan X adalah nilai asli, μ adalah rata-rata seluruh data, dan σ adalah standar deviasi. Proses ini penting untuk menyamakan skala antar variabel sebelum digunakan dalam analisis seperti regresi linier berganda, agar setiap variabel memiliki kontribusi yang seimbang dan tidak mendominasi model hanya karena perbedaan satuan atau skala. Normalisasi juga membantu mempercepat konvergensi dalam proses pelatihan model serta meningkatkan akurasi prediksi.

- Menggabungkan dan mengagregasi data untuk analisis (Gold):

  Pada tahap ini, data dari berbagai sumber yang sudah dibersihkan akan digabungkan dan digunakan untuk membangun model regresi linier berganda, serta menghasilkan dataset siap pakai untuk analisis lanjutan. Berikut adalah model umum regresi linier berganda:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta k X k + \varepsilon$$

  Keterangan:

  Y = Variabel terikat

  β_0 = Intercept

$\beta_1, \beta_2, \ldots, \beta k$ = Koefisien regresi yang menunjukkan pengaruh masing-masing variabel independen terhadap Y

$X_1, X_2, \ldots, Xk$ = Variabel bebas

ε = Error atau Galat

Model ini digunakan untuk menganalisis hubungan linear antara satu variabel dependen dengan beberapa variabel independen sekaligus.

- Menyediakan akses ke data dan hasil model (Hive dan poster):

  Data yang sudah diolah dapat diakses melalui query SQL menggunakan Hive, dan divisualisasikan dalam bentuk poster analitik untuk mendukung interpretasi hasil model dan pengambilan keputusan.

2.4 Karakteristik Pengguna

Sistem ini dirancang untuk melayani tiga jenis pengguna utama, yaitu *Data Analys*t, *Data Engineer*, analis ekonomi. Masing-masing pengguna memiliki peran dan tanggung jawab dalam mendukung proses analisis dan prediksi harga emas sebagai berikut:

- *Data Analys*t akan menggunakan Hive untuk menjalankan query dan mengeksplorasi data ekonomi seperti BI Rate, inflasi, kurs, dan harga emas guna keperluan analisis statistik dan untuk pembentukan model regresi.
- *Data Engineer* bertanggung jawab terhadap kelangsungan dan stabilitas pipeline ETL (*Extract-Transform-Load*), termasuk proses pembersihan, normalisasi data, serta pengelolaan penyimpanan data di setiap lapisan *Medallion Architecture* (Bronze, Silver, Gold).
- Analis Ekonomi atau Pembuat Kebijakan akan memanfaatkan hasil akhir berupa visualisasi pada dashboard untuk mendukung pengambilan keputusan dan penyusunan kebijakan ekonomi berbasis data makroekonomi dan fluktuasi harga emas secara akurat dan terpercaya.

2.5 Asumsi dan Batasan Sistem

Beberapa asumsi dan batasan sistem dalam penelitian ini diantaranya:

- Sistem ini hanya memproses data historis dalam skema batch, bukan real-time.
- Sumber data terbatas pada instansi resmi seperti Bank Indonesia, BPS, dan situs harga emas.
- Model regresi linier berganda hanya memprediksi berdasarkan tiga variabel ekonomi, tanpa mempertimbangkan faktor lain seperti kondisi global, geopolitik, atau harga komoditas lain.
- Infrastruktur big data diasumsikan tersedia dan dapat digunakan untuk pemrosesan data dalam skala besar.

3. Spesifikasi Proyek

3.1 Ringkasan

Proyek Big Data ini dirancang untuk memodelkan dan memprediksi harga emas di Indonesia dengan memanfaatkan pendekatan *Lakehouse Architecture* serta pemrosesan data secara *batch* menggunakan ekosistem Apache Spark dan Apache Hadoop. Sistem dibangun menggunakan kerangka kerja *Medallion Architecture (Bronze–Silver–Gold)*, yang memungkinkan pengelolaan dan pemurnian data secara bertahap, dari data mentah hingga data siap pakai untuk analisis statistik dan pemodelan prediktif.

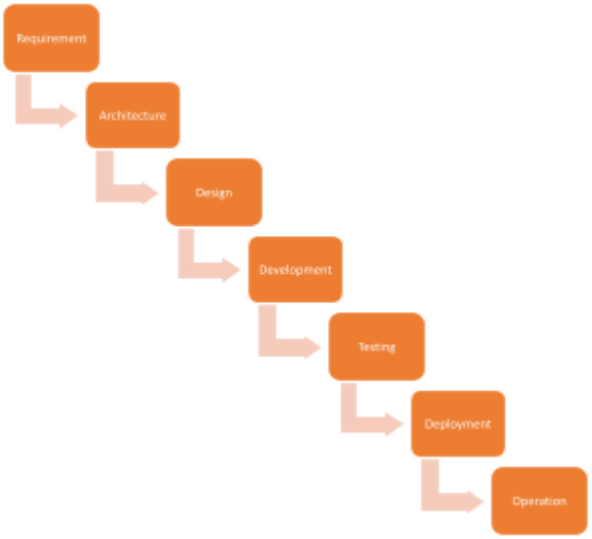
Data utama yang digunakan berasal dari berbagai sumber ekonomi makro, seperti data inflasi, BI *Rate*, nilai tukar rupiah terhadap dolar, dan harga emas harian. Data mentah dikumpulkan dalam format CSV dan JSON dari sumber resmi seperti Bank Indonesia, Badan Pusat Statistik, dan pasar

komoditas global, kemudian disimpan di HDFS pada lapisan Bronze. Proses pembersihan, normalisasi, dan transformasi data dilakukan menggunakan Apache Spark SQL pada lapisan Silver. Setelah data terstruktur dan siap dianalisis, data disimpan dalam format yang efisien seperti *Parquet* pada lapisan *Gold*. Untuk analisis, model Regresi Linier Berganda digunakan untuk mengidentifikasi dan mengukur pengaruh variabel ekonomi seperti inflasi dan nilai tukar terhadap harga emas. Proses ini dilakukan menggunakan Spark MLlib, dengan eksplorasi awal dan pengelolaan metadata dilakukan melalui Apache Hive.

Output sistem berupa model prediksi harga emas, grafik tren historis, serta analisis kontribusi masing-masing variabel terhadap harga emas yang dapat digunakan oleh analis ekonomi dan pembuat kebijakan. Tujuan utama dari proyek ini adalah membangun sistem analitik berbasis Big Data yang mampu menangani volume data makroekonomi dalam skala besar, menyediakan insight berbasis model statistik, serta mendukung pengambilan keputusan strategis di bidang keuangan dan ekonomi nasional.

3.2 Metode Proyek

Metodologi pengembangan proyek ini menggunakan pendekatan *Waterfall,* yaitu model pengembangan sistem yang berjalan secara linear dan berurutan, dengan setiap tahapan harus diselesaikan sebelum masuk ke tahap berikutnya. Proses dimulai dari tahap Analisis Kebutuhan, di mana sistem dirancang untuk mengolah data makroekonomi seperti inflasi, BI Rate, kurs, dan harga emas untuk keperluan analisis prediktif. Proyek dilanjutkan ke tahap Perancangan Sistem, yang mencakup pemodelan arsitektur *Lakehouse* dengan pendekatan *Medallion (Bronze–Silver–Gold)*, perancangan *pipeline batch processing*, serta pemilihan teknologi seperti Apache Spark, Hive, dan HDFS. Pada tahap Implementasi, dilakukan pembangunan *pipeline* ETL, pembersihan data, penyimpanan terstruktur di setiap layer, serta integrasi dengan model regresi linier berganda menggunakan Spark MLlib. Setelah itu, sistem diluncurkan pada tahap *Deployment* di lingkungan lokal atau *cloud* berbasis *cluster*. Data pipeline dijalankan secara periodik dalam mode batch processing serta visualisasi dalam bentuk poster disajikan untuk digunakan oleh pengguna akhir seperti analis ekonomi dan pembuat kebijakan. Kemudian, sistem memasuki tahap Testing, di mana sistem dievaluasi berdasarkan akurasi model prediksi, efisiensi proses batch, dan konsistensi antar lapisan data. Selanjutnya, sistem digunakan untuk menjalankan analisis prediktif secara berkala dan menyajikan hasilnya dalam visualisasi yang berguna bagi pembuat kebijakan ekonomi. Pendekatan *Waterfall* dipilih karena sesuai dengan karakteristik proyek yang memiliki kebutuhan dan alur data yang sudah terdefinisi sejak awal.



Gambar 1. Metode Waterfall

3.3 Studi Kasus

Studi kasus dalam proyek ini berfokus pada pemanfaatan teknologi Big Data untuk menganalisis faktor-faktor ekonomi yang mempengaruhi harga emas di Indonesia sebagai upaya mendukung kebijakan ekonomi. Harga emas memiliki keterkaitan yang erat dengan berbagai indikator makro-ekonomi seperti inflasi, BI Rate, dan nilai tukar rupiah terhadap dolar AS. Namun, hingga kini belum banyak sistem analitik yang mampu mengintegrasikan seluruh indikator tersebut dalam satu ekosistem data yang utuh, fleksibel, dan terukur. Untuk menjawab tantangan ini, studi kasus ini merancang dan mengimplementasikan sistem analitik prediktif berbasis *Big Data Architecture*, dengan mengadopsi pendekatan *Lakehouse* dan struktur *Medallion (Bronze–Silver–Gold)* dengan pendekatan *batch preprocessing*.

Sistem ini menggabungkan beberapa sumber data ekonomi dari berbagai sumber seperti Bank Indonesia, Badan Pusat Statistik, dan pasar komoditas. Data mentah tersebut disimpan di lapisan *Bronze* menggunakan Hadoop Distributed File System (HDFS) dalam format CSV atau JSON, kemudian diproses dan dibersihkan di lapisan *Silver*, hingga akhirnya dianalisis secara prediktif di lapisan *Gold* dengan model regresi linier berganda menggunakan pustaka *MLlib* dari Apache Spark untuk mengidentifikasi hubungan statistik antara harga emas dan variabel-variabel makroekonomi. Output akhir dari sistem ini berupa model prediksi hubungan antar variabel serta tren harga emas ke depan, yang dapat digunakan oleh analis keuangan, pembuat kebijakan ekonomi, dan lembaga keuangan untuk menyusun strategi berbasis data. Dalam studi kasus ini, sistem Big Data tidak hanya digunakan untuk menyimpan dan mengelola data berskala besar, tetapi juga berfungsi sebagai infrastruktur analitik strategis yang mampu mendorong pengambilan keputusan ekonomi yang lebih presisi dan adaptif terhadap perubahan pasar global.

4. Metode Umum

4.1 Analisis Kebutuhan (*Requirement Analysis)*

Tujuan:

1. Memahami masalah utama: Harga emas di Indonesia dipengaruhi oleh berbagai faktor ekonomi, seperti inflasi, BI rate, dan nilai tukar mata uang, yang berdampak pada keputusan investasi dan kebijakan ekonomi.
2. Menentukan stakeholder: Bank Indonesia (BI), Kementerian Keuangan Republik Indonesia (Kemenkeu RI), Analis pasar emas, Investor dan Trader.
3. Mendefinisikan output utama: Model prediksi harga emas berbasis Regresi Linier Berganda, laporan visualisasi tren harga emas, dan analisis hubungan variabel ekonomi.
4. Identifikasi data: BI Rate, Kurs IDR, Inflasi, Harga Emas Indonesia

Tabel 2. Kebutuhan Fungsional

No	Kebutuhan	Prioritas
1	Mengumpulkan data BI Rate, Kurs IDR, Inflasi, dan Harga Emas Indonesia dari sumber terpercaya	Tinggi
2	Menyimpan dan mengelola data dalam sistem Big Data untuk pemrosesan dalam skala besar	Tinggi
3	Preprocessing data: format, missing values, normalisasi untuk regresi	Tinggi
4	Pengembangan pipeline batch processing untuk pemrosesan berkala	Tinggi
5	Analisis korelasi variabel ekonomi terhadap harga emas	Sedang
6	Pelatihan model Regresi Linier Berganda dan tuning parameter	Tinggi
7	Evaluasi model dengan metrik RMSE, R^2 untuk mengukur akurasi	Tinggi
8	Visualisasi tren harga emas dan pengaruh variabel ekonomi	Sedang
9	Integrasi dengan sistem analitik dan query untuk akses mudah	Sedang
10	Monitoring batch processing dan deteksi anomali data	Rendah

11	Update otomatis model berdasarkan data terbaru	Sedang

Tabel 3. Kebutuhan Non Fungsional

No.	Kebutuhan	Prioritas
1	Skalabilitas: Sistem harus mampu menangani pertumbuhan volume data.	Tinggi
2	Keamanan Data: Enkripsi dan kontrol akses ketat untuk mencegah penyalahgunaan.	Rendah
3	Ketersediaan: Menjamin uptime tinggi untuk proses batch yang stabil.	Sedang
4	Performance: Proses batch harus cepat dan efisien dalam analisis data	Sedang
5	Reliabilitas: Mekanisme recovery otomatis jika terjadi kegagalan.	Sedang
6	Maintainability: Dokumentasi sistem harus jelas untuk kemudahan perbaikan dan pengembangan.	Tinggi
7	Portabilitas: Sistem harus fleksibel untuk digunakan di berbagai lingkungan.	Tinggi

4.2 Perancangan Sistem: Arsitektur dan Desain (*System Design*)

Arsitektur Data

Dalam pemodelan harga emas, data seperti BI Rate, kurs IDR, inflasi, dan harga emas perlu diproses secara konsisten dari sumber resmi. Oleh karena itu, diperlukan sebuah arsitektur yang dapat menangani data mentah dalam jumlah besar, melakukan transformasi secara terstruktur, dan menghasilkan output yang siap untuk dianalisis. Arsitektur *Medallion* dengan data *lakehouse* mendukung pengelolaan big data melalui transformasi bertahap dan batch processing yang efisien.

Tabel 4. Arsitektur Medallion – Batch Processing dengan Data Lake

Lapisan	Deskripsi	Komponen Utama (Tools)	Format Data	Tujuan
Bronze	Menyimpan data mentah yang diperoleh dari sumber-sumber resmi: BI Rate, kurs IDR, inflasi, dan harga emas Indonesia tanpa transformasi.	HDFS, Apache Kafka	CSV	Arsip permanen data mentah dari sumber eksternal tanpa perubahan.
Silver	Melakukan pembersihan, transformasi, dan integrasi data (penyesuaian format, penggabungan berdasarkan tanggal, normalisasi) sehingga siap untuk analisis	Apache Spark, Apache Hive	Parquet	Menyediakan data terstruktur dan konsisten yang optimal untuk pemodelan regresi dan analisis korelasi.
Gold	Menyajikan agregasi data, output dari model prediksi (misalnya, tren harga emas dan pengaruh variabel ekonomi) serta hasil visualisasi interaktif.	Apache Spark, PowerBI	Parquet	Menyediakan insight bisnis dan laporan analitik yang mudah diakses oleh stakeholder untuk pengambilan keputusan.

Desain Infrastruktur

Dalam mewujudkan sistem prediksi harga emas berbasis Big Data yang andal dan *scalable,* dibutuhkan infrastruktur terdistribusi yang mampu menangani volume data ekonomi makro secara batch dan mendukung pemodelan prediktif dengan efisien. Desain infrastruktur dalam proyek ini mencakup pengaturan cluster lokal berbasis Hadoop serta pemilihan dan integrasi berbagai teknologi open-source dari ekosistem Apache. Selain itu, orkestrasi pipeline dilakukan secara sistematis guna menjamin kelancaran alur data dari ingestion hingga visualisasi. Rancangan ini disusun agar dapat mendukung pendekatan *Lakehouse Architecture* dengan struktur *Medallion (Bronze, Silver, Gold)* untuk memastikan kebersihan, keterlacakan, dan kegunaan data dalam analisis lanjutan.

Tabel 5. Arsitektur Cluster Lokal (Hadoop Cluster)

Komponen	Jumlah Node	Peran dan Deskripsi
Hadoop Namenode	1	Master node untuk HDFS, mengelola metadata file.
Hadoop Datanode	2	Menyimpan data dalam blok. Direplikasi untuk toleransi kesalahan.
ResourceManager	1	Komponen YARN yang menjadwalkan job Spark/Hive.
NodeManager	2	Mengelola eksekusi task pada masing-masing worker node.
Apache HiveServer2	1	Layanan query engine untuk Hive SQL yang dapat diakses melalui Beeline/BI tools.
Apache Spark Master	1	Koordinator eksekusi job Spark batch.
Apache Spark Worker	2	Mengeksekusi job Spark (ETL, transformasi, agregasi).
Apache Hive Metastore	1	Metadata store untuk skema dan tabel Hive.
Ambari Server	1	Monitoring dan manajemen layanan Hadoop ecosystem.
Superset	1	Untuk visualisasi data gold layer.

Tabel 6. Daftar Teknologi Apache Projects yang digunakan

No	Teknologi	Kategori	Fungsi Utama
1	Hadoop HDFS	Storage	Menyimpan data dan hasil pemrosesan secara terdistribusi
2	Apache Spark	Processing	Memproses data batch untuk transformasi dan pemodelan regresi.
3	YARN	Resource Management	Mengelola sumber daya cluster dan menjalankan aplikasi secara paralel.
4	Apache Hive	Query Engine	Menyediakan antarmuka SQL untuk akses dan analisis data.
5	Apache Oozie	Workflow Scheduling	Menjadwalkan pipeline batch untuk pemrosesan data berkala.
6	Apache Kafka	Data Ingestion	Mengalirkan dan mengumpulkan data secara batch dari berbagai sumber.
7	Ambari	Monitoring & Management	Memantau dan mengelola cluster Hadoop.

Orkestrasi Alur Kegiatan Sistem

Tabel 7. Alur kegiatan sistem

Urutan	Aktivitas	Layer	Tools
1	Ambil file dari FTP/API dan unggah ke HDFS	Bronze	Bash, curl, HDFS CLI
2	Simpan data mentah tanpa ubahan	Bronze	HDFS
3	Validasi schema dan lakukan pembersihan data	Silver	Apache Spark, Hive Metastore
4	Ubah format ke Parquet dan simpan di lokasi refined	Silver	Spark + HDFS
5	Agregasi dan perhitungan metrik ekspor berdasarkan waktu dan kategori	Gold	Spark SQL / Hive
6	Simpan hasil agregasi akhir dalam bentuk tabel analitik	Gold	HDFS, Hive
7	Visualisasi dan eksplorasi data	Gold	Superset / Jupyter
8	Monitoring dan manajemen cluster	Seluruh layer	Apache Ambari

4.3 Implementasi (Implementation)

Mewujudkan arsitektur dan perancangan sistem menjadi sistem nyata (real deployment) pada lingkungan lokal (Docker Desktop VM) berbasis Linux, menggunakan tools utama dari Hadoop ecosystem dan Apache Foundation.

Tabel 8. Lingkungan implementasi

Komponen	Spesifikasi
OS	Ubuntu Server 22.04 (via Docker VM) – OS komputer: Windows 11
Cluster	Pseudo-distributed Hadoop Cluster dengan 2 Node
Orkestrasi	Shell Script + Crontab
Core Tools	Hadoop, HDFS, Hive, Spark, Ambari, Superset, Airflow
Penyimpanan	HDFS sebagai distributed storage
Format Data	CSV (bronze), Parquet (silver & gold)

Proses Batch dan Penyimpanan

Proses ETL berjalan secara batch terjadwal (misalnya harian atau bulanan) menggunakan penjadwalan sederhana seperti cron job atau Oozie/Airflow. Data mentah (CSV) dari sumber resmi (BI Rate, kurs, inflasi, harga emas) dimuat ke HDFS (lapisan Bronze) setiap siklus batch. File CSV yang baru ditempatkan dapat berada di direktori sinkronisasi, lalu dibaca oleh job Spark berikutnya. Setiap tahap disimpan dalam format yang sesuai:

- Bronze: data mentah tetap sebagai CSV di HDFS (arsip permanen)
- Silver: hasil transformasi dan integrasi ditulis dalam format kolumnar efisien seperti Parquet (lebih cepat untuk query analitik}
- Gold: output akhir model dan agregasi disimpan juga sebagai Parquet atau tabel Hive agar mudah di-query dan dianalisis lebih lanjut.

Siklus batch ini ulang secara periodik. Misalnya, setiap malam Spark job dijalankan (spark-submit) untuk melakukan ETL dari Bronze ke Silver, lalu model regresi dilatih, dan hasilnya dioutput ke Gold. Dengan format Parquet dan tabel Hive, proses baca/tulis lebih cepat dan hemat penyimpanan, serta mendukung pemrosesan berskala besar.

a. Persiapan Cluster Lokal

1. Install Docker Desktop + WSL2 di Windows 11

2. Jalankan container berbasis image Hadoop
3. Gunakan docker-compose untuk spin-up:
    - ✔ 2 NameNode
    - ✔ 2 DataNode
    - ✔ 1 Spark Master
    - ✔ 1 Spark Worker
    - ✔ 1 Hive Metastore + HiveServer2
    - ✔ Ambari
    - ✔ Superset
    - ✔ Airflow

b. Struktur Docker-Compose

buat file docker-compose.yml yang memuat baris struktur instruksi docker.

Tabel 9. Struktur Folder HDFS

Layer	Path HDFS	Format
Bronze	`/data/bronze/`	CSV
Silver	`/data/silver/`	Parquet
Gold	`/data/gold/`	Parquet
Temp (untuk temporary)	`/data/tmp/`	-

Implementasi Pipeline Batch

1. **Bronze Layer (Ingestion)**

✔ Jalankan bash script via crontab setiap jam/hari (contoh):

```
curl -o ekspor.csv https://data.example/api/ekspor
hdfs dfs -put -f ekspor.csv /data/bronze/ekspor_$(date +%F).csv
```

curl untuk scrapping atau unduh, hdfs dfs -put untuk upload berkas data.

2. **Silver Layer (Cleansing & Normalization)**

✔ Gunakan Spark Job:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("CleanEkspor").getOrCreate()
df = spark.read.csv("/data/bronze/*.csv", header=True)
```

```
df_clean = df.dropDuplicates().na.fill("N/A")  # Cleaning
df_clean.write.parquet("/data/silver/ekspor_clean.parquet", mode="overwrite")
```

Simpan code pyspark dalam file .ipynb atau .py

3. **Gold Layer (Aggregation & Analytics)**

✔ Agregasi OLAP:

```
df = spark.read.parquet("/data/silver/ekspor_clean.parquet")
agg = df.groupBy("negara", "bulan").agg({"nilai": "sum"})
agg.write.parquet("/data/gold/ekspor_summary.parquet", mode="overwrite")
```

Simpan code pyspark dalam file .ipynb atau .py

Implementasi Query dan Visualisasi

✔ Hive Table (DDL)

```
CREATE EXTERNAL TABLE ekspor_summary (
  negara STRING,
  bulan STRING,
  total_nilai DOUBLE
)
STORED AS PARQUET
LOCATION '/data/gold/ekspor_summary.parquet';
```

Gunakan HiveQL untuk menjalankan kueri.

Pengembangan ETL

✔ Upload file mentah → HDFS (bronze)

✔ ETL Spark → silver layer (cleansing, parsing)

✔ Agregasi Spark → gold layer (komoditas, negara, volume)

✔ Pembuatan Hive Table dari gold layer

Dashboard

Import ke **Apache Superset** atau **Jupyter + PyHive** untuk eksplorasi.

Monitoring dan Logging

✔ **Log Spark**: Tersedia di UI http://localhost:8080

✔ **Ambari UI**: Monitoring cluster, resource, dan service

✔ **Log File**: Tersimpan otomatis di folder /logs/ di container jika dikonfigurasi

Tabel 10. Output implementasi

Output	Lokasi
Data mentah ekspor	`/data/bronze/`
Data bersih (normalized)	`/data/silver/`
Ringkasan nilai ekspor (OLAP)	`/data/gold/`
Tabel Hive (ekspor_summary)	Hive Metastore
Visualisasi ekspor	Superset / Jupyter Dashboard

4.4 Penerapan (*Deployment*)

Mengimplementasikan sistem ke lingkungan lokal berbasis Docker VM yang mensimulasikan cluster Hadoop untuk kebutuhan pemodelan harga emas dengan batch processing.

Tabel 11. Strategi deployment

Tahapan	Deskripsi
1. Setup Environment	Instalasi Docker Desktop + Ubuntu VM, setup docker-compose untuk layanan Hadoop, Spark, Hive, dan Superset.
2. Build & Configure	Konfigurasi cluster Hadoop pseudo-distributed dengan HDFS, Spark, Hive,
3. Data Ingestion	Mengambil data mentah (harga emas, BI Rate, kurs, inflasi) dari API/CSV ke Bronze Layer menggunakan cron job.
4. Spark Transformation	Deploy job Spark untuk pembersihan data, standarisasi (MinMaxScaler), dan transformasi ke Silver Layer.
5. Hive Integration	Membuat tabel Hive dari Gold Layer untuk analisis regresi linier berganda.
6. Visualisasi	Menampilkan tren harga emas dan hasil prediksi model.

Tabel 12. Teknologi Deployment

Tools	Peran
Docker	Virtualisasi container untuk layanan Hadoop, Spark, Hive, dan docker.
Docker Compose	Manajemen multi-container.
Bash + Crontab	Menjadwalkan pengambilan data bulanan ke HDFS.
Spark Submit	Menjalankan job ETL batch untuk preprocessing dan pemodelan.
Hive Metastore (MySQL/PostgreSQL)	Menyimpan metadata tabel Gold Layer untuk analisis regresi.
PowerBI	Visualisasi hasil analisis

Struktur File di Server

```
/opt/gold_price_model/
├── docker-compose.yml
├── data/
│   ├── bronze/ (data mentah: CSV/JSON)
│   ├── silver/ (data bersih: Parquet)
│   └── gold/ (data agregat & hasil model: ORC)
├── scripts/
│   ├── ingest_data.sh
│   ├── spark_clean.py
│   └── spark_model.py
└── logs/
```

4.5 Pengujian (*Testing*)

Memastikan seluruh alur pipeline dari pengambilan data hingga prediksi model berjalan dengan akurat.

Tabel 13. Jenis pengujian

Jenis Pengujian	Tujuan
Unit Test	Memastikan script Spark (pembersihan data, standarisasi, regresi) berjalan tanpa error.
Integration Test	Validasi integrasi Bronze → Silver → Gold Layer dan koneksi Hive.
Data Quality Test	Cek konsistensi data (misal: nilai BI Rate tidak negatif, tanggal valid).
Performance Test	Evaluasi akurasi model regresi linier menggunakan RMSE dan R-squared.
End-to-End Test	Simulasi lengkap: data masuk → preprocessing → pemodelan → visualisasi.

Tabel 14. Kasus uji

No	Kasus Uji	Deskripsi	Status
1	Data mentah terunggah ke Bronze	File CSV harga emas tersimpan di /data/bronze/.	✅
2	Spark membersihkan data	Silver Layer tidak mengandung missing values.	✅
3	Normalisasi data	Nilai fitur (BI Rate, kurs) berada di rentang 0-1.	✅
4	Model regresi menghasilkan prediksi	Output model tersimpan di Gold Layer.	✅

Tabel 15. Tools testing

Tools	Fungsi
Jupyter Notebook	Validasi manual output Spark dan Hive
Bash + Log File	Monitoring job Spark dan log eksekusi.
Hive	Uji query tabel Gold Layer untuk analisis regresi.

4.6 Analitik (*Analytics*)

Melakukan pemodelan harga emas menggunakan regresi linier berganda dengan fitur BI Rate, kurs, dan inflasi.

Tujuan Analitik

- ✔ Memprediksi harga emas bulanan berdasarkan variabel makroekonomi.
- ✔ Menilai pengaruh masing-masing variabel (koefisien regresi).

Data Input untuk ML

1. Sumber: Tabel Gold Layer di Hive (data terstandarisasi).
2. Format: Parquet.
3. Kolom:
    - ✔ tanggal (bulan)
    - ✔ harga_emas (dalam IDR/gram)

✔ bi_rate (% suku bunga BI)

✔ kurs_usd_idr (nilai tukar)

✔ inflasi (% perubahan bulanan)

Tabel 16. Metode analitik yang digunakan

No	Analisis	Algoritma Spark MLlib	Tujuan
1	Prediksi Harga Emas	Linear Regression	Memprediksi harga emas bulanan berdasarkan BI Rate, kurs, dan inflasi.

Tahapan Analisis

1. **Preprocessing:**
   -Standarisasi fitur menggunakan MinMaxScaler.
   -Split data: 80% training (2009-2023), 20% testing (2024-2025).
2. **Pemodelan:**
   -Latih model regresi linier berganda
3. **Evaluasi:**
   -Hitung RMSE dan R-squared untuk mengukur akurasi model.
4. **Visualisasi:**
   -Bandingkan harga aktual vs prediksi di Superset.
   -Tampilkan koefisien regresi sebagai indikator pengaruh variabel.

4.7 Dataset

Data dari berbagai sumber di kumpulkan menjadi satu dalam data lake melalui bacth processing.

Dataset 1: Harga emas idr.csv

Tanggal	Terakhir	Pembukaa	Tertinggi	Terendah	Perubahan%
01/01/2025	1.465.571	1.357.613	1.477.193	1.356.073	7,95%
01/12/2024	1.357.613	1.351.504	1.397.846	1.331.355	0,48%
01/11/2024	1.351.090	1.388.445	1.400.799	1.117.565	-2,70%
01/10/2024	1.388.652	1.288.335	1.410.355	1.286.046	7,78%
01/09/2024	1.288.365	1.250.141	1.306.246	1.154.047	3,06%
01/08/2024	1.250.141	1.278.391	1.289.000	1.145.583	-2,24%
01/07/2024	1.278.844	1.222.114	1.288.723	1.219.059	4,63%
01/06/2024	1.222.195	1.216.433	1.254.386	1.197.154	0,49%
01/05/2024	1.216.236	1.199.616	1.258.255	1.033.294	1,46%
01/04/2024	1.198.769	1.144.636	1.266.303	1.141.716	4,75%
01/03/2024	1.144.430	1.032.577	1.145.381	306.901	10,82%
01/02/2024	1.032.718	1.034.637	1.044.355	994.360	-0,18%
01/01/2024	1.034.609	1.020.702	1.044.271	1.006.608	1,36%
01/12/2023	1.020.702	1.020.516	1.063.470	886.453	0,02%
01/11/2023	1.020.548	1.010.786	1.063.628	925.826	0,96%

Dataset 2: data_inflasi.csv

Data inflasi di indonesia dari 2009 hingga 2025

No	Periode	Data Inflasi
1	Januari 2025	0.76 %
2	Desember 2024	1.57 %
3	November 2024	1.55 %
4	Oktober 2024	1.71 %
5	September 2024	1.84 %
6	Agustus 2024	2.12 %
7	Juli 2024	2.13 %
8	Juni 2024	2.51 %
9	Mei 2024	2.84 %
10	April 2024	3 %
11	Maret 2024	3.05 %
12	Februari 2024	2.75 %
13	Januari 2024	2.57 %

Dataset 3: kurs_transaksi.csv

Data kurs nilai tukar rupiah terhadap dolar dari tahun 2002 hingga 2025

NO	Nilai	Kurs Jual	Kurs Beli	Tanggal
1	1	16650,84	16485,16	5/15/2025 12:00:00 AM
2	1	16614,66	16449,34	5/14/2025 12:00:00 AM
3	1	16579,49	16414,51	5/9/2025 12:00:00 AM
4	1	16615,67	16450,33	5/8/2025 12:00:00 AM
5	1	16554,36	16389,64	5/7/2025 12:00:00 AM
6	1	16503,1	16338,9	5/6/2025 12:00:00 AM
7	1	16575,47	16410,53	5/5/2025 12:00:00 AM
8	1	16762,4	16595,6	5/2/2025 12:00:00 AM
9	1	16870,94	16703,06	4/30/2025 12:00:00 AM
10	1	16946,31	16777,69	4/29/2025 12:00:00 AM
11	1	16913,15	16744,85	4/28/2025 12:00:00 AM
12	1	16968,42	16799,58	4/25/2025 12:00:00 AM
13	1	16964,4	16795,6	4/24/2025 12:00:00 AM

Dataset 4: BI-Rate.csv

Data penetapan suku bunga di indonesia dari tahun 2009 hingga 2025

	BI Rate												
													2009
Variabel	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	Septembe	Oktober	Novembe	Desember	Tahunan
BI Rate	8,75	8,25	7,75	7,5	7,25	7	6,75	6,5	6,5	6,5	6,5	6,5	6,5

Format Penyimpanan

1. **Bronze Layer**: Lokasi: /data/bronze/harga_emas.csv (format CSV).
2. **Silver Layer**: Data terstandarisasi (MinMaxScaler) disimpan di /data/silver/harga_emas.parquet.
3. **Gold Layer**:Tabel Hive prediksi_harga_emas dengan kolom:

   -tanggal

   -harga_aktual

   -harga_prediksi

   -residual (selisih aktual vs prediksi).

4.8 Aliran Data (*Pipeline*) (buat dalam bagan visual bukan seperti ini)

[1] Data Source (CSV, JSON, XML, API, FTP, DB)
- ekspor_bahan_pangan.csv
- cuaca_ekspor.csv
- biaya_logistik.csv
- kurs_mata_uang.csv
- permintaan_global.csv

[2] Bronze Layer (HDFS - Raw Zone)
- Simpan file mentah ke HDFS (/datalake/bronze/)
- Format tetap CSV/JSON (belum diproses)
- Metadata tracking dicatat (timestamp, sumber, ukuran) via Hive Metastore

[3] Ingestion & Staging (Apache Spark + Hive) – Bronze ke Silver
- Gunakan Spark untuk baca data dari HDFS (bronze)
- Lakukan parsing, validasi skema, handling missing values
- Standardisasi waktu, format tanggal, dan satuan
- Tulis ke Silver Layer dalam format kolumnar (Parquet/ORC)

[4] Silver Layer (HDFS - Clean Zone)
- Path: /datalake/silver/
- Data sudah bersih dan distandardisasi
- Join antar dataset (ekspor + cuaca + biaya + kurs)
- Simpan hasil ke format Parquet untuk efisiensi query

[5] Enrichment & Transformation (Apache Spark) – Silver ke Gold
- Hitung agregasi bulanan, nilai ekspor (dalam USD/IDR)
- Integrasi permintaan global
- Siapkan data untuk analitik & ML

[6] Gold Layer (HDFS - Curated Zone)
- Path: /datalake/gold/
- Dataset siap pakai untuk analitik dan machine learning
- Format: Parquet atau Hive Tables
- Gunakan Hive untuk query BI, dan Spark MLlib untuk model prediksi

[7] Analytics Layer
- Apache Hive (OLAP + BI)
- Apache Superset atau Grafana (dashboard)
- Apache Spark MLlib (klasifikasi/prediksi permintaan/volume ekspor)

[8] Monitoring & Orchestration
- Apache Airflow → Workflow terjadwal
- Apache Ambari → Monitoring cluster & resource

5. Lampiran

5.1 Repositori Portofolio

Dokumentasi proyek Pemodelan Harga Emas Menggunakan Regresi Linier Berganda pada Sistem Big Data dengan Batch Processing dapat diakses melalui tautan berikut:. Repository

5.2 Lampiran Code

Analisis Harga Emas di Indonesia