



• PROPOSAL TEKNIS • KLUSTER PELAYANAN

Data Modernization & Integrated Reporting Platform

Modernisasi data mart pelayanan menjadi arsitektur **Lakehouse** modern —
ELT, Medallion, CDC, dan pelaporan terintegrasi dengan **Teknologi Big Data**
untuk BPJS Kesehatan.

Hadoop · Spark

Apache Iceberg

Airflow / Talend

Trino · Hive

CDC · Streaming

Voltage CLS

~42

TABEL SUMBER (ELT)

3+1

BULAN DEV + MAINTENANCE

3

LAYER MEDALLION

8

TENAGA AHLI



01 / 30



• RINGKASAN EKSEKUTIF

Dari Data Mart Statis menuju Lakehouse Analytics

BPJS Kesehatan memerlukan **modernisasi data mart pelayanan** menjadi platform data terpadu yang **lakehouse-ready** — mendukung transaksi ACID, evolusi skema, beban **batch & streaming**, serta pelaporan dan analitik yang konsisten dan siap pakai.

Venturo Pro mengusulkan arsitektur **ELT → Medallion (Bronze/Silver/Gold)** di atas Apache Iceberg, dengan Data Quality di setiap lapisan, keamanan kolom (CLS), metadata & lineage, serta empat dashboard tematik.

ACID

Iceberg transactions + schema evolution

100%

DQ check di setiap layer (end-to-end)

Real-time

CDC + Kafka untuk data streaming

Col-level

Proteksi data Voltage (masking/tokenization)

≈ 12
bln

Estimasi **payback period** melalui efisiensi pipeline, otomasi DQ, dan percepatan pelaporan eksekutif & deteksi fraud.

Lima Gelombang yang Membentuk Healthcare Data Platform



Lakehouse Convergence

Data lake + warehouse menyatu via open table format (Iceberg/Delta/Hudi). ACID di atas object/HDFS storage.



Real-time & CDC

Pergeseran dari batch harian ke change-data-capture & streaming untuk SLA pelaporan yang lebih ketat.



Data Governance & Privacy

CLS/RLS, masking, tokenization & lineage menjadi wajib pada data kesehatan yang sensitif (PII/PHI).



Self-Service Analytics

Semantic layer & dashboard tematik mendekatkan data ke pengambil keputusan tanpa rekayasa ulang.



AI/ML & Fraud Analytics

Gold/feature layer siap pakai untuk deteksi anomali klaim & fraud, prediksi utilisasi layanan.

Gartner

“By 2027, ≥ 60% organisasi data besar akan mengadopsi **open lakehouse** sebagai fondasi analitik enterprise.” — proyeksi industri (ilustratif).

• TANTANGAN SAAT INI

Hambatan pada Data Mart Pelayanan Eksisting

01

Arsitektur Belum ACID

Data lake/warehouse eksisting belum mendukung transaksi ACID & evolusi skema secara native — sulit menangani update/delete dan late-arriving data.

02

Pipeline ETL Rapuh

Penanganan error, retry, logging & alerting belum end-to-end. Beban di sistem sumber tinggi saat penarikan ±42 tabel.

03

Kualitas Data Tidak Terjamin

Belum ada DQ otomatis berbasis dimensi DMBOK di tiap lapisan; integritas referensial & format tidak konsisten.

04

Skalabilitas & Kinerja

Strategi partisi & kompaksi small-files belum optimal; kueri analitik melambat seiring pertumbuhan volume & konkurensi.

05

Keamanan Kolom Terbatas

Proteksi data tingkat kolom (PII/PHI) belum diterapkan menyeluruh; risiko paparan data sensitif peserta.

06

Metadata & Lineage Minim

Metadata teknis/bisnis/operasional & data lineage belum terdokumentasi — menyulitkan audit dan keberlanjutan.

Kriteria Evaluasi Sistem Data Eksisting

DIMENSI EVALUASI (PER TOR)	FOKUS	STATUS
Efektivitas Landing Zone	Ingestion raw → data lake	Sebagian
Organisasi Data & Format File	Partisi · kolumnar · kompaksi	Kurang
Dukungan Transaksi ACID	Hudi/Delta/Iceberg	Belum
Prinsip Desain Warehousing	Kepatuhan pola arsitektur	Sebagian
Skalabilitas & Kinerja	Volume · konkurensi · kueri	Sebagian
Mekanisme Ingestion & Pemuatan	Efisiensi · keandalan · SLA	Sebagian
Model Konseptual/Logis/Fisik	3-layer modelling	Kurang
Normalisasi vs Denormalisasi	Redundansi · kinerja kueri	Sebagian
Desain Skema Datamart	Star/snowflake compliance	Sebagian
Mekanisme Pembaruan	Refresh · latensi · dampak sumber	Kurang
Penanganan Error Pipeline	Logging · alerting · retry	Kurang
Kualitas Data (DMBOK)	6 dimensi kualitas	Belum

SKOR KESIAPAN KESELURUHAN

2.3 / 5

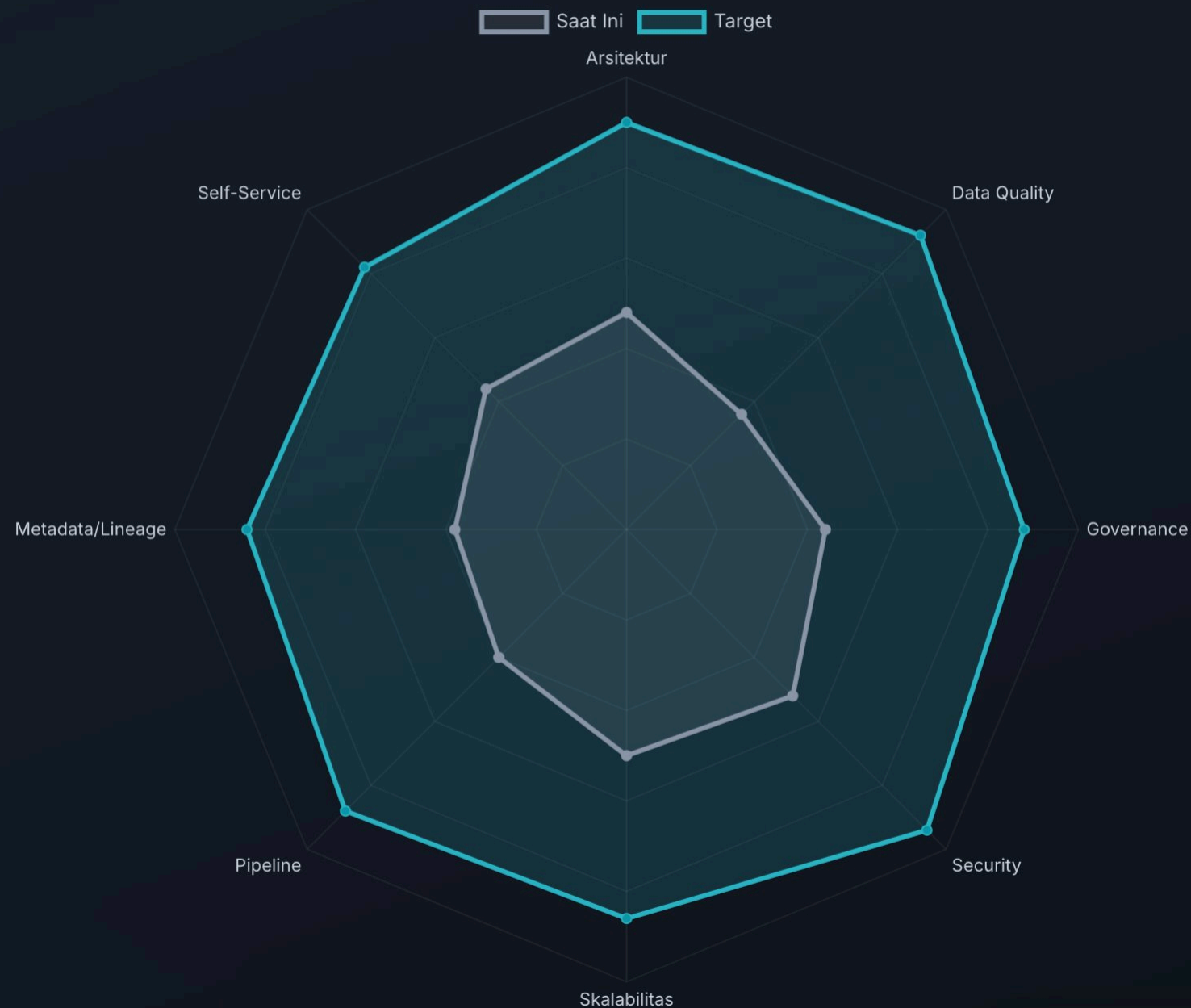
Tahap "Reactive → Managed". Fondasi ada, namun belum lakehouse-ready.

PRIORITAS PERBAIKAN

ACID & Open Table Format	P0
Data Quality Otomatis	P0
Keamanan Kolom (Voltage)	P1
Metadata & Lineage	P1

• DATA MATURITY ASSESSMENT

Current vs Target — 8 Dimensi Kematangan Data



METODOLOGI

Penilaian skala 1–5 mengacu pada **DMBOK** & maturity model lakehouse, mencakup arsitektur, kualitas, governance, dan operasional.

2.3

SKOR SAAT INI

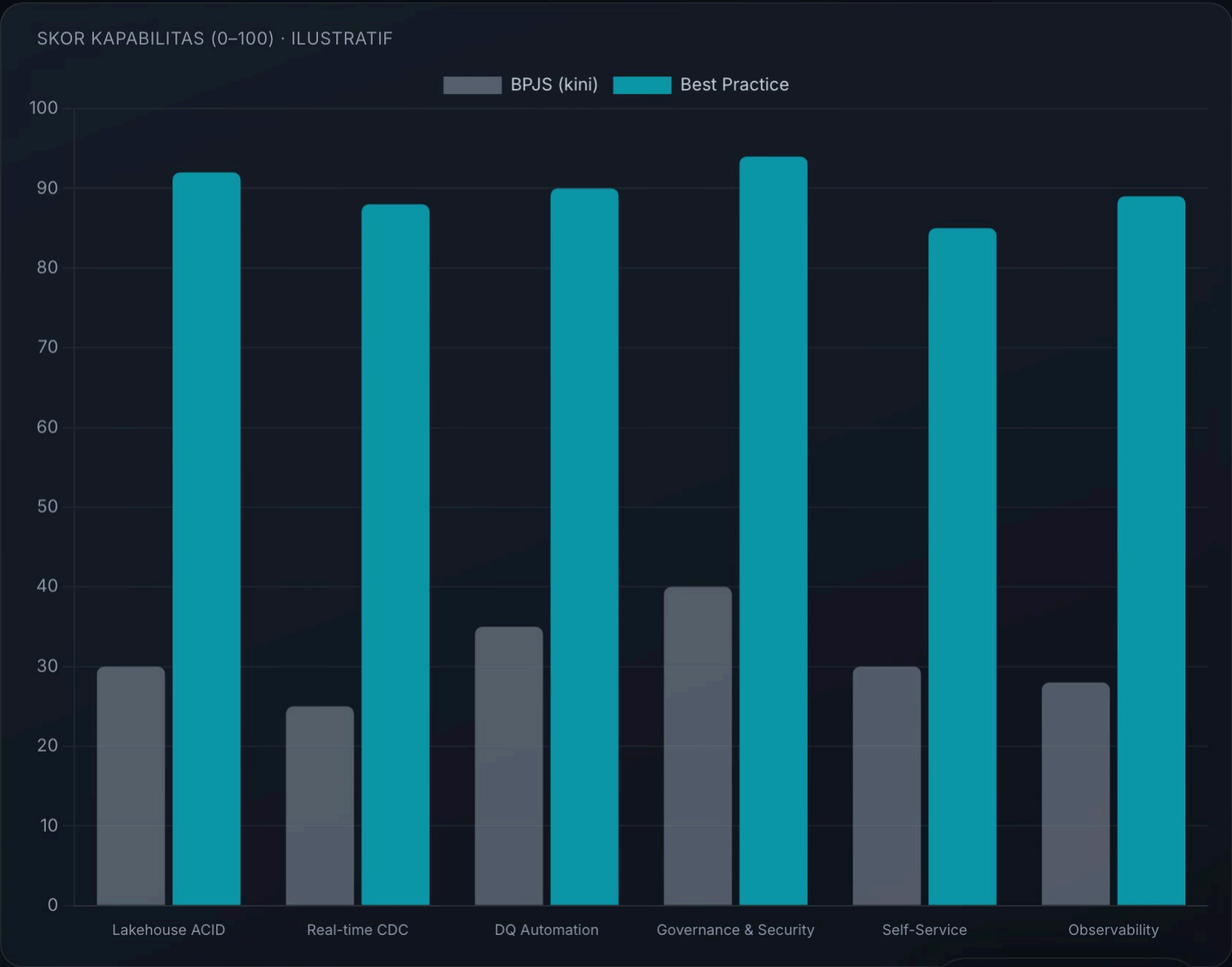
4.4

SKOR TARGET

+91%

PENINGKATAN KEMATANGAN PASCA-IMPLEMENTASI

Healthcare Data Platform — Posisi vs Best Practice



KAPABILITAS	BPJS (KINI)	BEST PRACTICE
Open Lakehouse (ACID)	30	92
Real-time / CDC	25	88
Data Quality Automation	35	90
Governance & Security	40	94
Self-Service / Semantic	30	85
Observability	28	89

Referensi: pola NHS Data Platform, CMS (US), Singapore HealthHub, dan rujukan vendor lakehouse — disetarakan ke konteks BPJS.

• TEMUAN UTAMA

Empat Temuan yang Mengarahkan Desain Solusi

A Open Table Format adalah keharusan

Tanpa Iceberg, kebutuhan ACID, schema evolution, time-travel, dan kompaksi small-files tidak dapat dipenuhi pada kluster Hadoop eksisting.

B ELT + Medallion menyederhanakan

Memuat raw langsung ke landing/Bronze mengurangi beban sistem sumber; transformasi terjadi di Spark pada Silver/Gold.

C DQ harus inline & otomatis

Pemeriksaan kualitas (DMBOK) di setiap layer mencegah propagasi data buruk hingga data mart & dashboard.

D Keamanan & lineage sejak desain

CLS via Voltage + metadata/lineage menjadikan platform auditable dan tahan terhadap kebutuhan kepatuhan masa depan.

Satu Platform Data Terpadu, Tepercaya, Siap Pakai

"Setiap keputusan pelayanan didukung data yang konsisten, aman, dan real-time."

North-Star untuk Kluster Pelayanan BPJS Kesehatan.

Lakehouse-Ready

ACID, schema evolution, pemisahan compute/storage di atas infrastruktur eksisting.

Trusted Data

DQ end-to-end (DMBOK) + integritas referensial pada tiap tahap ELT.

Secure by Design

CLS, masking, tokenization, enkripsi & RLS untuk data PII/PHI peserta.

Real-time Capable

CDC + streaming berdampingan dengan batch tradisional sesuai kebutuhan.

Self-Service Reporting

Semantic layer + dashboard tematik (eksekutif, operasional, fraud, governance).

Sustainable & Documented

Naming convention, kamus data, metadata & knowledge transfer untuk keberlanjutan.

• TARGET OPERATING MODEL

People · Process · Technology · Governance



People

- PM / Team Leader
- Business Analyst
- Data Designer
- 3× Data Engineer (CDC/Talend)
- QC & Admin



Process

- ELT (raw→landing)
- Medallion B/S/G
- DQ gates per layer
- CI/CD pipeline jobs
- Scheduling & alerting



Technology

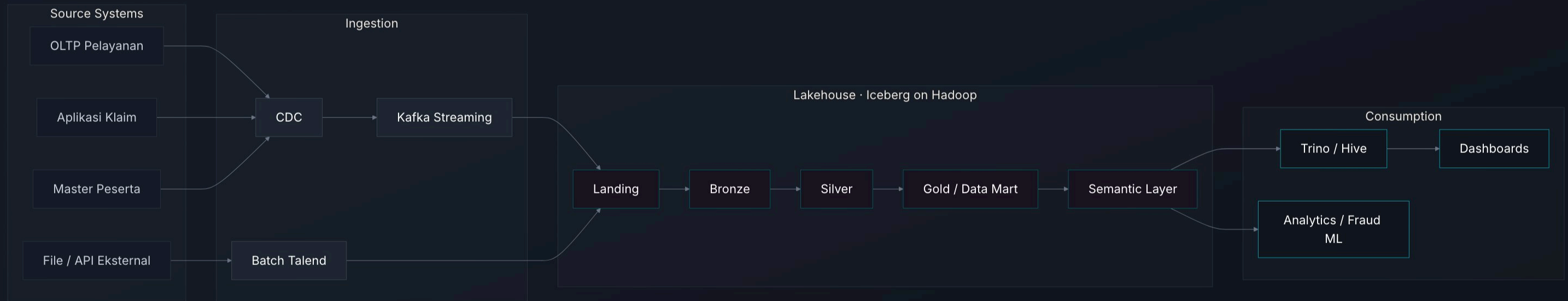
- Hadoop · Spark
- Iceberg · Parquet/ORC
- Kafka · Airflow/Talend
- Trino · Hive
- Voltage (CLS)



Governance

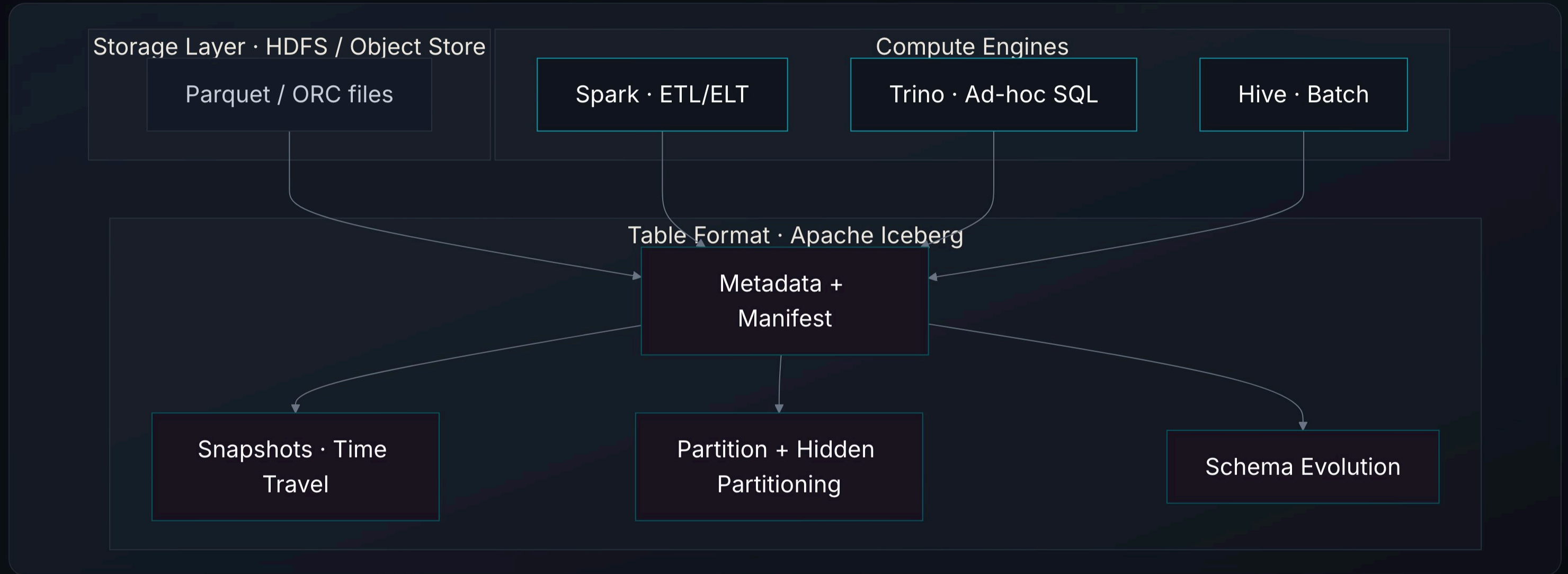
- Metadata teknis/bisnis/ops
- Data lineage
- Naming convention
- Kamus data
- Audit & access control

Arsitektur Data Terpadu — End-to-End



• LAKEHOUSE ARCHITECTURE

Iceberg Lakehouse — Compute & Storage Terpisah



ACID

Transaksi atomik untuk insert/update/delete & merge (UPSERT).

Time Travel

Snapshot & rollback untuk audit dan reproducibility.

Compaction

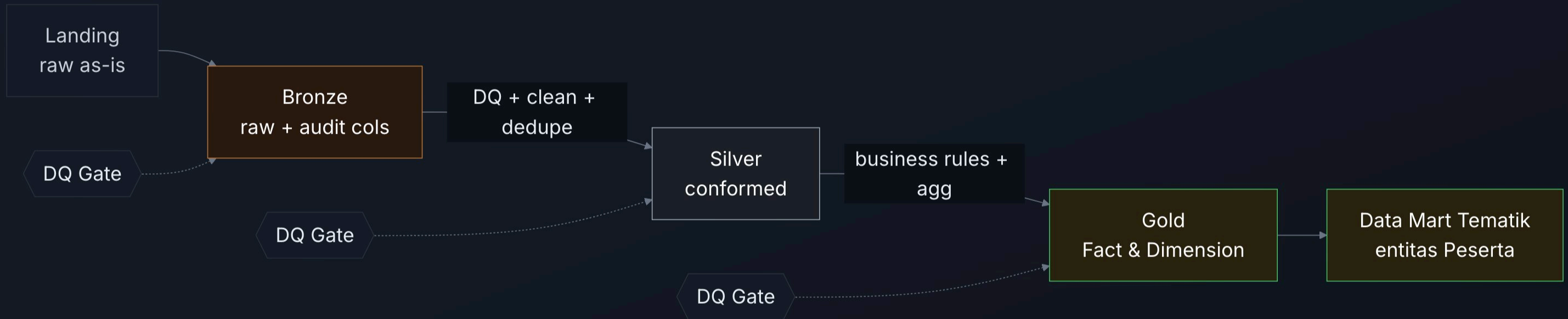
Pemadatan small-files otomatis → kinerja kueri stabil.

Open

Satu copy data, banyak engine (Spark/Trino/Hive).

• MEDALLION ARCHITECTURE

Bronze → Silver → Gold dengan DQ di Setiap Lapisan



Bronze

Raw immutable + kolom audit (ingest_ts, source, op). Schema-on-read, append-only.

Silver

Cleansed, deduplicated, conformed; integritas referensial & tipe data tervalidasi.

Gold

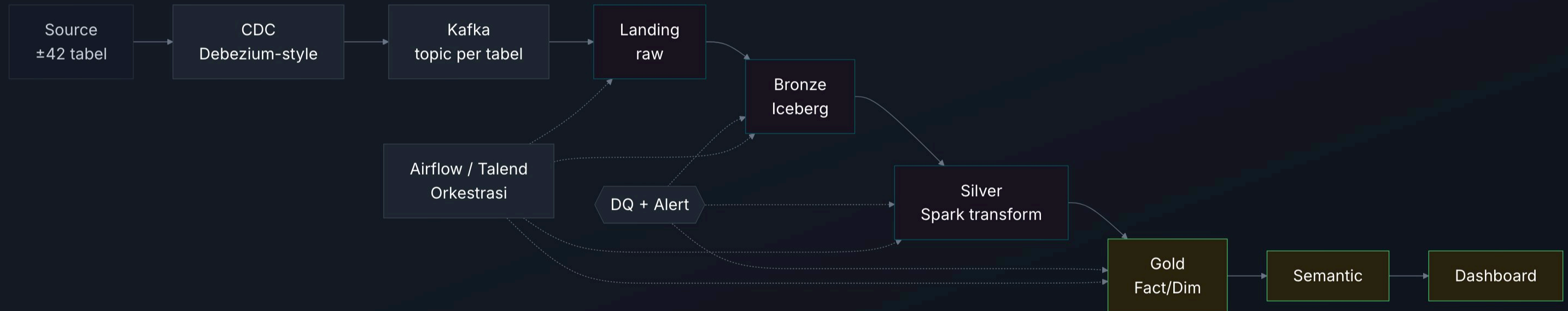
Fact & dimension, agregasi bisnis, data mart tematik peserta — siap analitik & pelaporan.

SOURCE SYSTEM LANDSCAPE

±42 Tabel Sumber Kluster Pelayanan

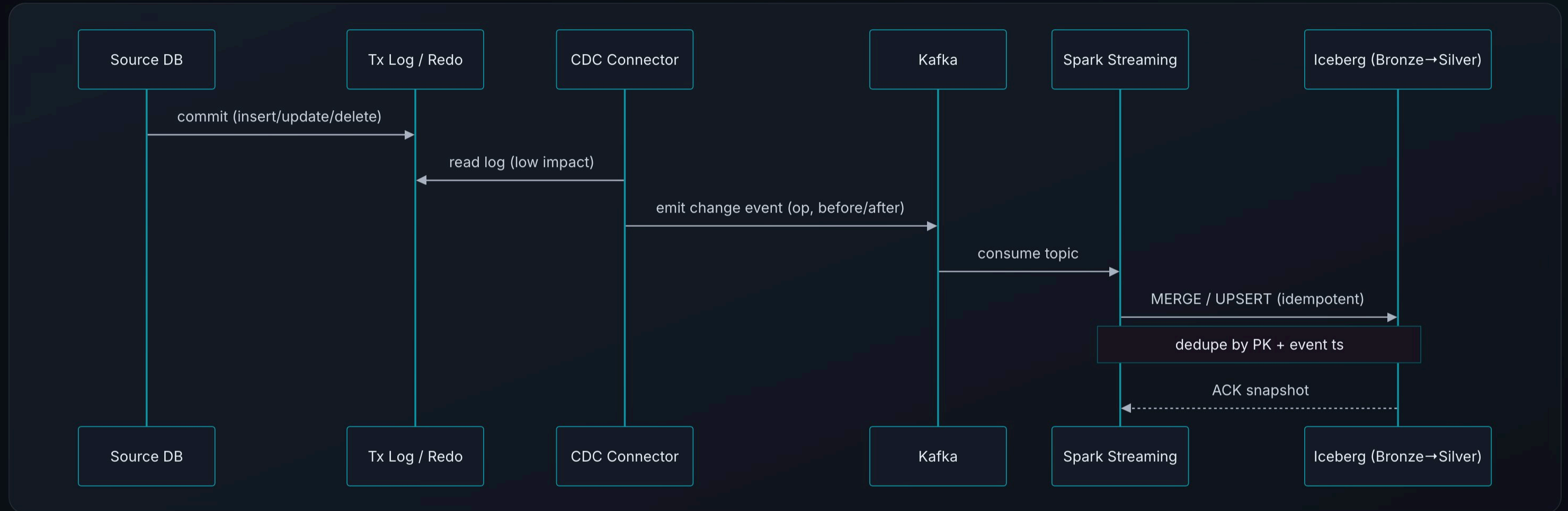
OLTP Pelayanan Transaksi kunjungan, rujukan, FKTP/FKRTL. Mode: CDC	Klaim & Billing Pengajuan & verifikasi klaim. Mode: CDC + batch	Master Peserta Data kepesertaan, segmentasi. Mode: CDC	~42 TABEL (DAPAT BERUBAH PASCA-EVALUASI)
Master Faskes Provider, kapitasi, kontrak. Mode: batch	Referensi / Kode ICD, INA-CBG, wilayah. Mode: batch	File / API Eksternal Data semi/unstructured. Mode: batch/stream	PRINSIP INGESTION Tanpa dampak signifikan ke sistem sumber Incremental load (CDC) diutamakan Konfigurasi koneksi format global Struktur, semi-struktur & unstruktur didukung

Alur End-to-End: Source → CDC → Kafka → ... → Dashboard



• CDC STRATEGY

Change Data Capture — Low-Impact, Near Real-time



Mengapa Log-based CDC?

- Tanpa beban query polling ke sumber
- Menangkap insert/update/delete
- Latensi rendah → mendekati real-time
- Mendukung incremental load

Idempotency & Ordering

MERGE INTO Iceberg berbasis primary key + event timestamp menjamin exactly-once semantics & late-data handling.

• METADATA & LINEAGE

Source-to-Target Data Lineage



Tiga Jenis Metadata

Teknis — struktur & tipe data

Bisnis — definisi elemen data

Operasional — data lineage & run-time

Manfaat

Impact analysis, audit trail, root-cause DQ, dan pemenuhan kamus data + naming convention per TOR.

Enam Dimensi DMBOK — Inline di Setiap Layer

99.8%

Completeness
tidak ada null wajib

99.9%

Accuracy
sesuai sumber benar

100%

Validity
format & range

99.7%

Consistency
lintas tabel

100%

Uniqueness
no duplikat PK

<15m

Timeliness
freshness SLA

Validate

Aturan kualitas dijalankan otomatis di Bronze/Silver/Gold; pelanggaran → quarantine + alert.

Cleanse & Transform

Standardisasi format, dedupe, perbaikan referensial otomatis bila memungkinkan.

Report & Score

DQ scorecard per tabel/layer → Data Governance Dashboard, dengan tren & threshold alerting.

Proteksi Berlapis untuk Data PII / PHI Peserta



Column-Level Security

Kontrol akses per kolom sensitif (NIK, diagnosa) sesuai peran — implementasi **Voltage**.



Row-Level Security

Filter baris berbasis wilayah/peran agar pengguna hanya melihat data yang berwenang.



Masking

Dynamic data masking pada display (mis. NIK 32.....1234) untuk non-privileged.



Tokenization

Format-preserving tokenization (FPE) menjaga referential join tanpa membuka nilai asli.



Encryption

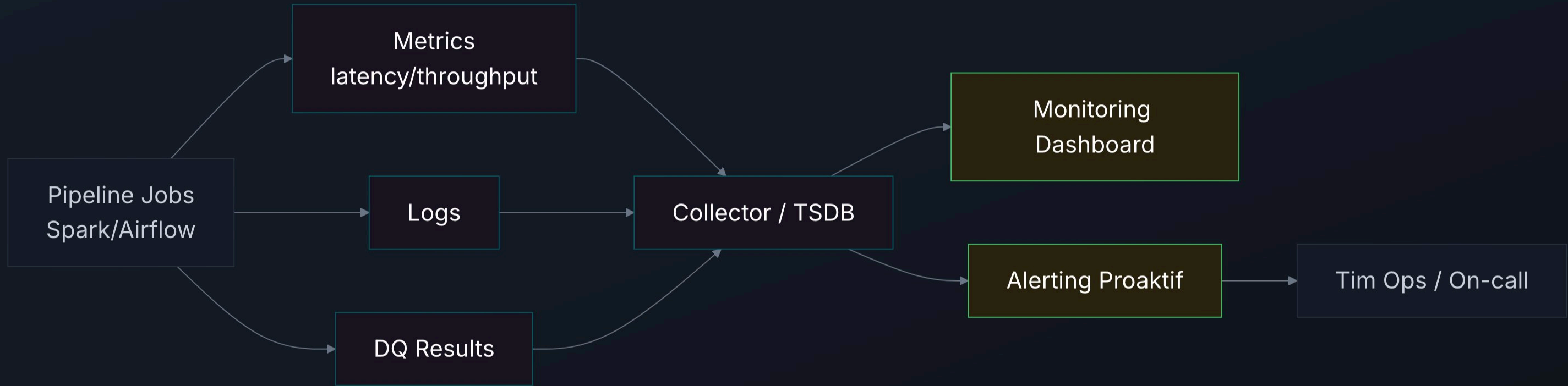
Enkripsi at-rest (storage) & in-transit (TLS); kunci dikelola terpusat (KMS).

Defense in Depth

Akses terbatas & terkontrol, hanya pada hari/jam kerja sesuai ketentuan operasional BPJS Kesehatan (on-site untuk akses data).

• MONITORING ARCHITECTURE

Observability End-to-End — Latensi · Traffic · Error



Latency

durasi per stage & DAG

Throughput

volume/unit waktu

Error Rate

frekuensi kegagalan

Freshness

data lag vs SLA

Alert proaktif → notifikasi sebelum SLA terlewat; runbook troubleshooting terdokumentasi (fase pemeliharaan).

Empat Persona, Satu Semantic Layer

C-Level

Executive

KPI strategis pelayanan, utilisasi, biaya klaim, tren nasional.

Ops

Operational

Kesehatan pipeline, SLA, antrian, throughput faskes harian.

SIU / Audit

Fraud

Deteksi anomali klaim, skor risiko, kasus prioritas.

Data Office

Governance

DQ scorecard, lineage coverage, akses & kepatuhan.

Semantic Layer

Satu definisi metrik terpercaya (single source of truth) di Gold → konsisten lintas seluruh dashboard, melayani via Trino.

Executive Dashboard

KUNJUNGAN (BLN)

8.4jt

▲ 4.2%

BIAYA KLAIM

Rp 12.7T

▲ 2.1%

RASIO KLAIM

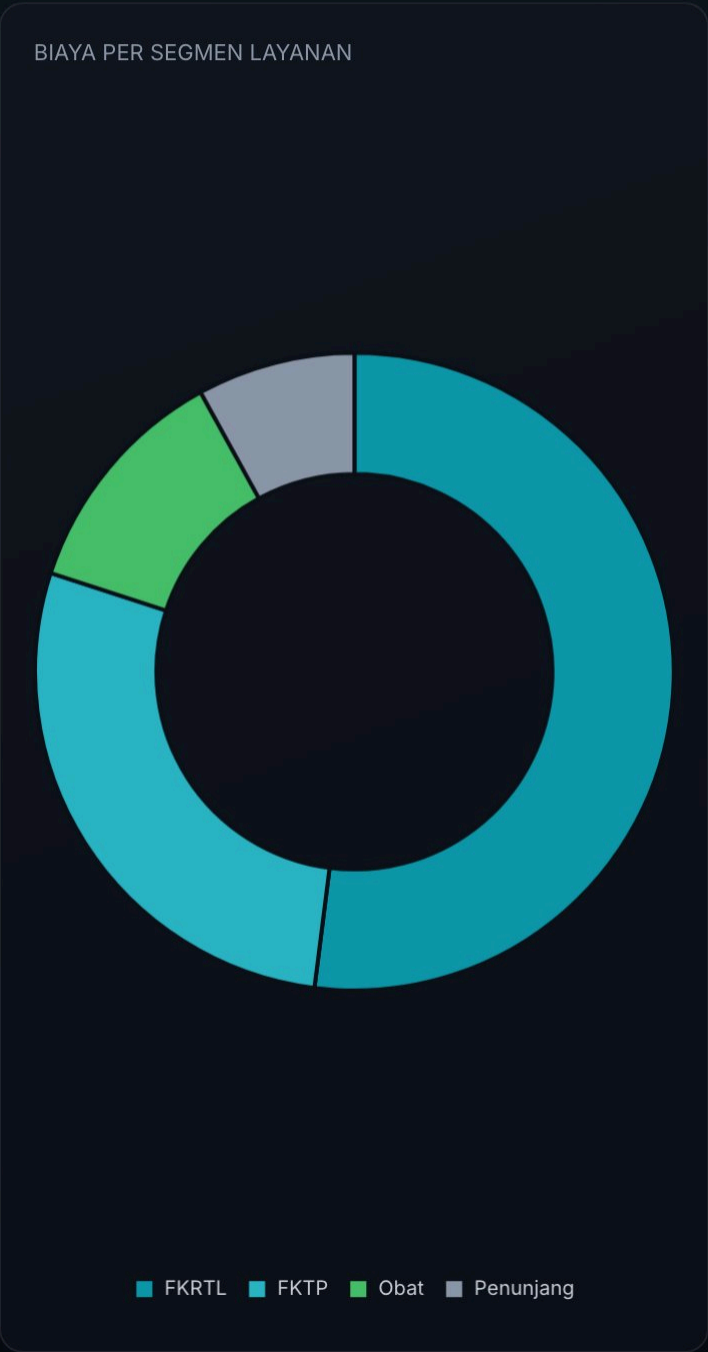
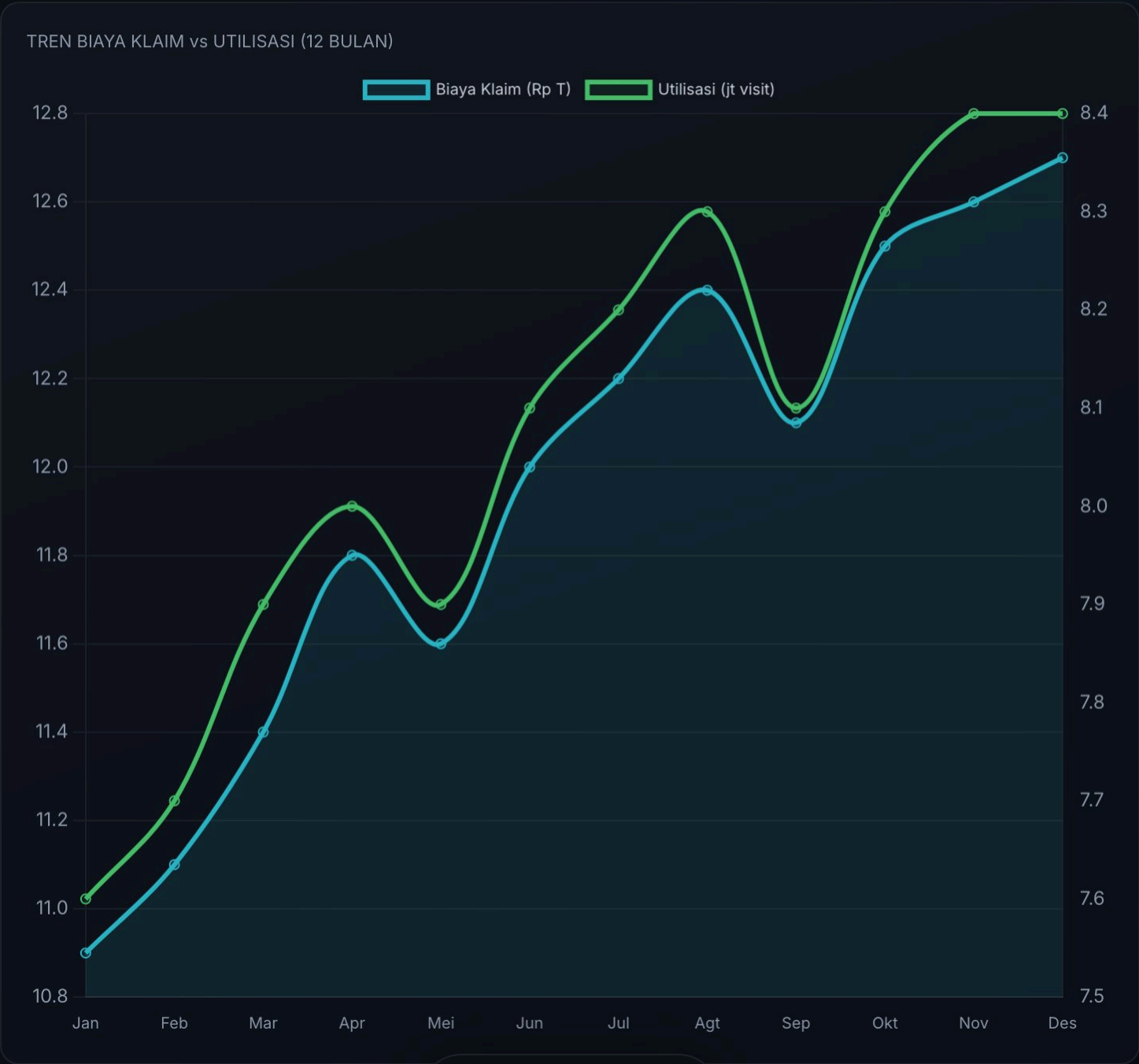
98.6%

▼ 0.8%

PESERTA AKTIF

277jt

▲ 0.6%



• MOCKUP • OPERATIONAL

Operational Dashboard

STATUS PIPELINE HARIAN · 42 JOB

38

SUCCESS

3

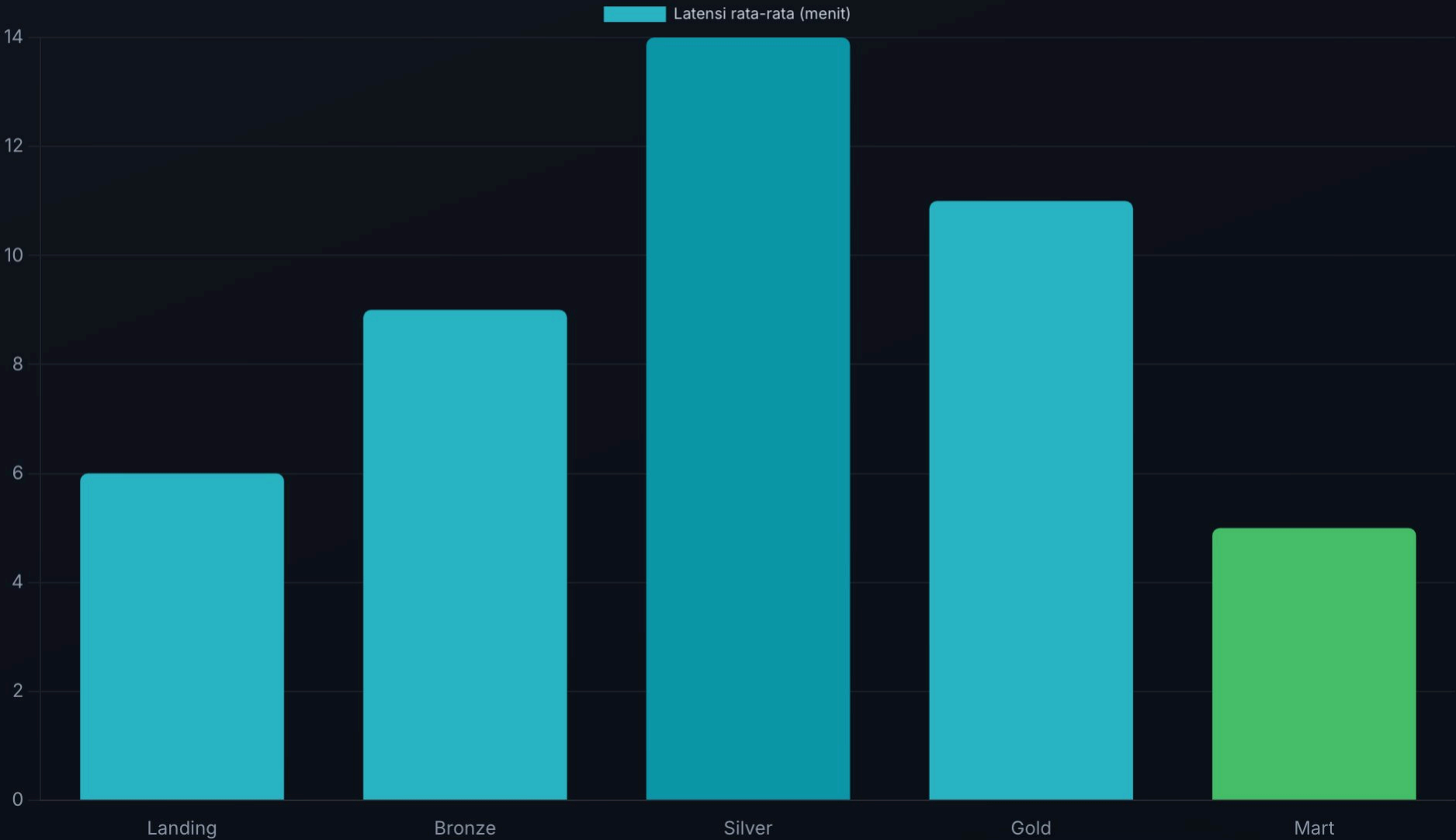
RUNNING

1

FAILED

99.4%

SLA



EVENT TERBARU

OK	bronze.klaim	04:12
OK	silver.visit_clean	04:31
FAIL	silver.peserta_conf	04:33
RUN	gold.fact_klaim	04:40
OK	mart.peserta_360	05:01

Fraud Analytics Dashboard

ALERT AKTIF

214

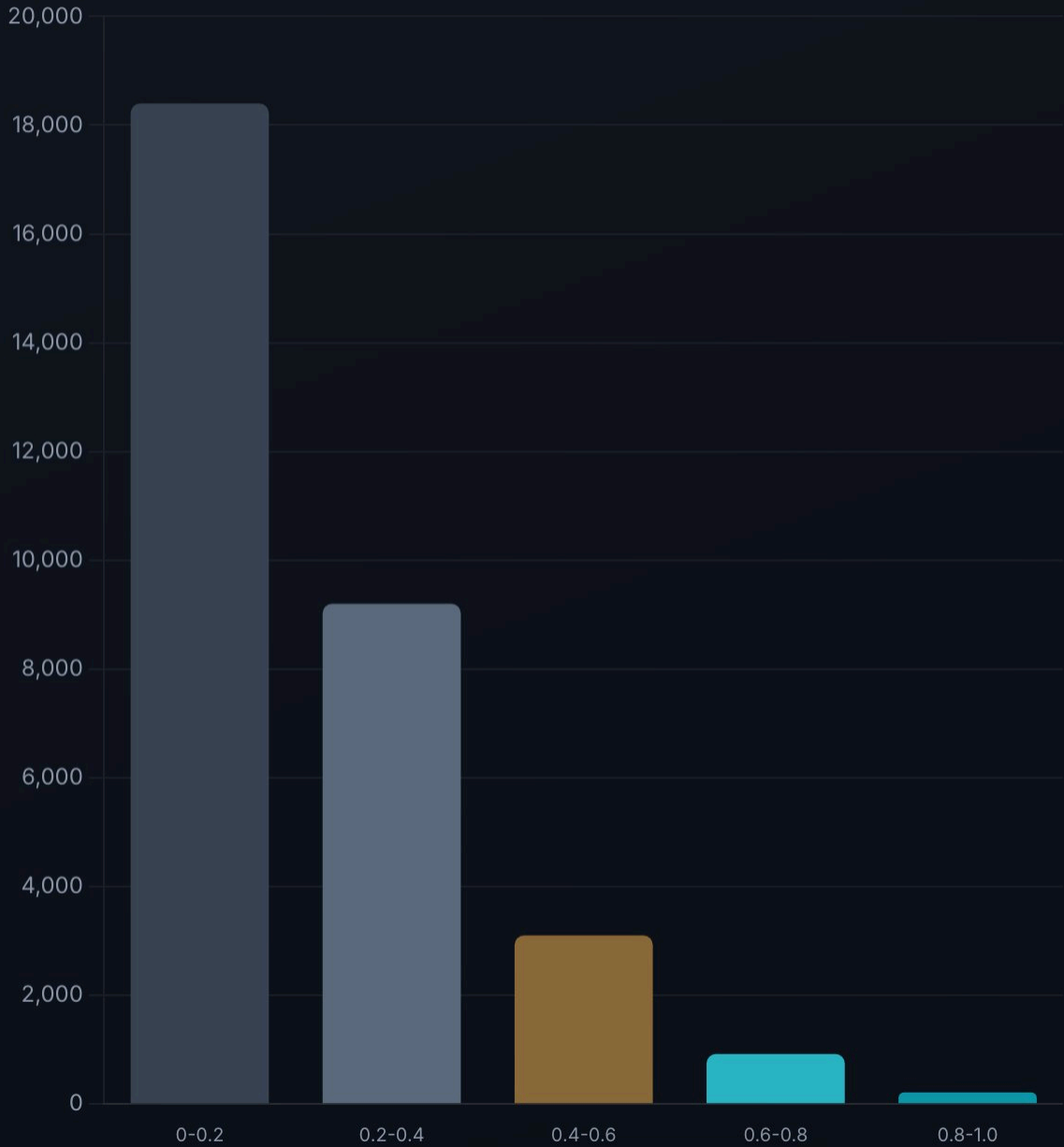
POTENSI SAVING

Rp 86M

PRECISION MODEL

91%

DISTRIBUSI SKOR RISIKO KLAIM



KASUS PRIORITAS

FASKES	POLA	SKOR
FKRTL-0192	Upcoding	0.97
FKTP-3320	Phantom visit	0.95
FKRTL-0871	Readmit loop	0.88
FKTP-1204	Duplikat klaim	0.84
FKRTL-0455	Unbundling	0.81

Data Governance Dashboard

DQ SCORE GLOBAL

98.9%

LINEAGE COVERAGE

100%

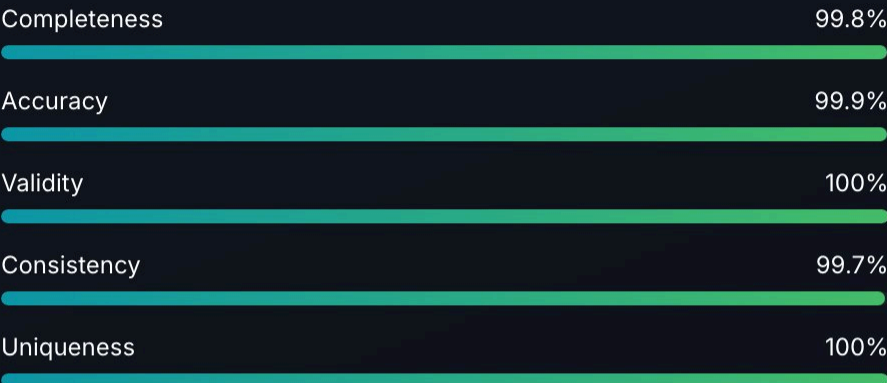
KOLOM TERPROTEKSI

128

DQ SCORE PER LAYER

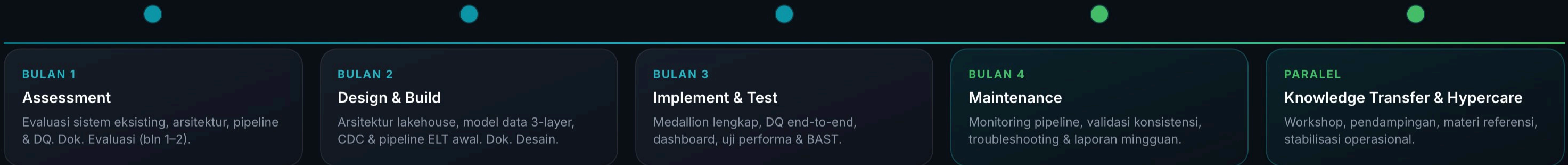


DQ DIMENSI (BULAN INI)



• ROADMAP

3 Bulan Pengembangan + 1 Bulan Pemeliharaan



• RESOURCE PLAN

Tim Inti — 8 Tenaga Ahli

PERAN	KUALIFIKASI	PENGALAMAN	JML	DURASI
Project Manager / Team Leader	S2 TI/terkait	≥ 5 thn PM data	1	32 hari
Business Analyst (Muda)	S1 TI/terkait	≥ 3 thn data	1	4 bln
Data Designer (Muda)	S1 TI/terkait	≥ 3 thn data	1	4 bln
Data Engineer (Muda)	S1 TI · ≥1 CDC · ≥1 Talend	≥ 3 thn data	3	3 bln
Data Engineer (maintenance)	S1 TI/terkait	≥ 3 thn data	1	1 bln
Quality Control	S1 TI/terkait	≥ 3 thn QC	1	4 bln
Tenaga Administrasi	D3	≥ 1 thn admin	1	4 bln

8

TOTAL PERSONEL

3

DATA ENGINEER

CDC

≥1 ENGINEER WAJIB

Talend

≥1 ENGINEER WAJIB

Onsite di kantor pusat BPJS untuk akses data; online untuk pekerjaan non-akses-data.

RISK MITIGATION

Risiko Utama & Strategi Mitigasi

RISIKO	DAMPAK	PROB.	MITIGASI
Dampak penarikan ke sistem sumber	Tinggi	Sedang	Log-based CDC + incremental load + window jam kerja terkontrol
Jumlah tabel berubah pasca-evaluasi	Sedang	Tinggi	Pipeline metadata-driven & template reusable per tabel
Kualitas data sumber buruk	Tinggi	Sedang	DQ gate + quarantine + scorecard governance
Paparan data PII/PHI	Tinggi	Rendah	CLS/RLS Voltage, masking, tokenization, enkripsi, audit akses
Bottleneck kinerja kueri	Sedang	Sedang	Partisi + kompaksi Iceberg + tuning Spark/Trino
Ketergantungan knowledge	Sedang	Sedang	Dokumentasi lengkap + workshop + materi referensi (KT)
Jadwal ketat (3 bulan)	Tinggi	Sedang	Delivery iteratif per bulan + laporan bulanan + scope guard

• EXPECTED BENEFITS

Nilai Bisnis & Payback

PROYEKSI MANFAAT KUMULATIF vs INVESTASI (ILUSTRATIF)



≈12 bln

PAYBACK PERIOD

-60%

WAKTU PELAPORAN

+99%

DATA QUALITY SCORE

-45%

BEBAN SISTEM SUMBER

Rp 86M+

POTENSI SAVING DARI DETEKSI FRAUD / THN

• WHY VENTURO

Mengapa Venturo Pro



Spesialis Lakehouse

Penguasaan mendalam Iceberg, Spark, Trino, Airflow/Talend & CDC pada kluster Hadoop enterprise.



Security & Compliance

Pengalaman menerapkan CLS/RLS, masking & tokenization (Voltage) untuk data sensitif sektor publik.



DMBOK-Grade Quality

Kerangka DQ & governance terstruktur — metadata, lineage, kamus data & naming convention.



Tim Tersertifikasi

8 tenaga ahli memenuhi seluruh kualifikasi TOR, termasuk keahlian CDC & Talend.



Delivery Tepat Waktu

Metodologi iteratif & laporan bulanan menjaga 3 bulan dev + 1 bulan pemeliharaan tetap on-track.

Mitra Keberlanjutan

Knowledge transfer, dokumentasi & hypercare memastikan BPJS Kesehatan mandiri pasca-proyek.