# A Blueprint for Secure Banking Migration

An enterprise-grade data platform engineered for zero-trust security, regulatory compliance, and operational excellence.

## Medallion Architecture

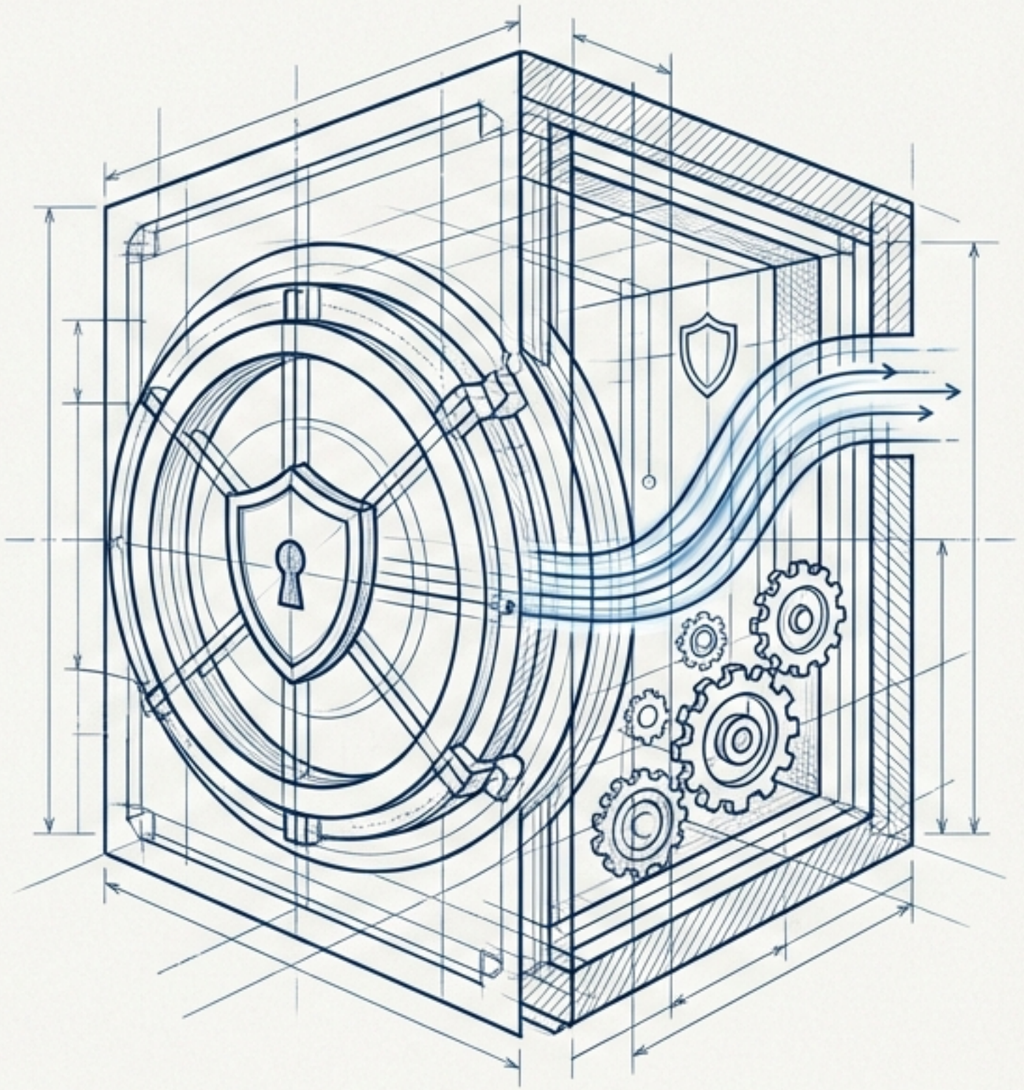
A disciplined, multi-layered approach to data integrity.

## Security & Governance by Design

Cryptographic controls and compliance are embedded, not bolted on.

## Cloud-Native & Automated

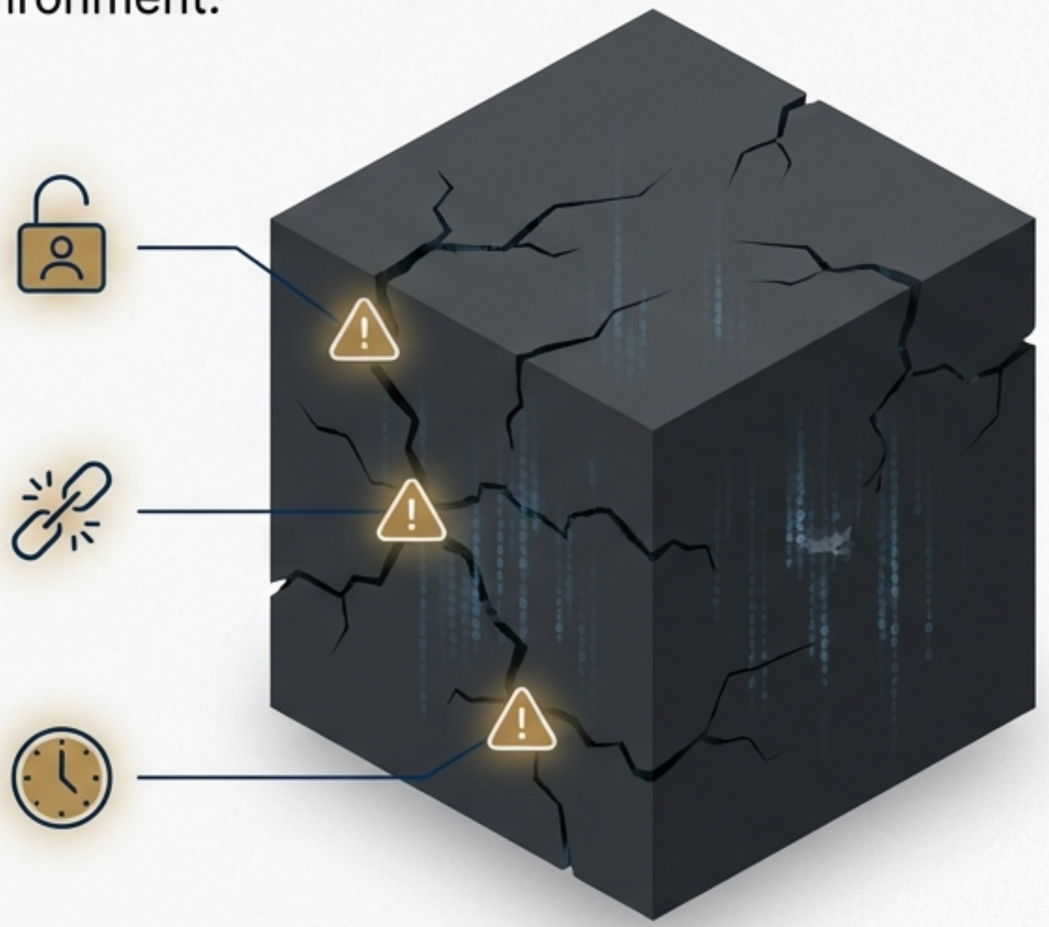
Built for scalability and operational efficiency using Infrastructure as Code.



# The High Cost of Legacy Data Systems

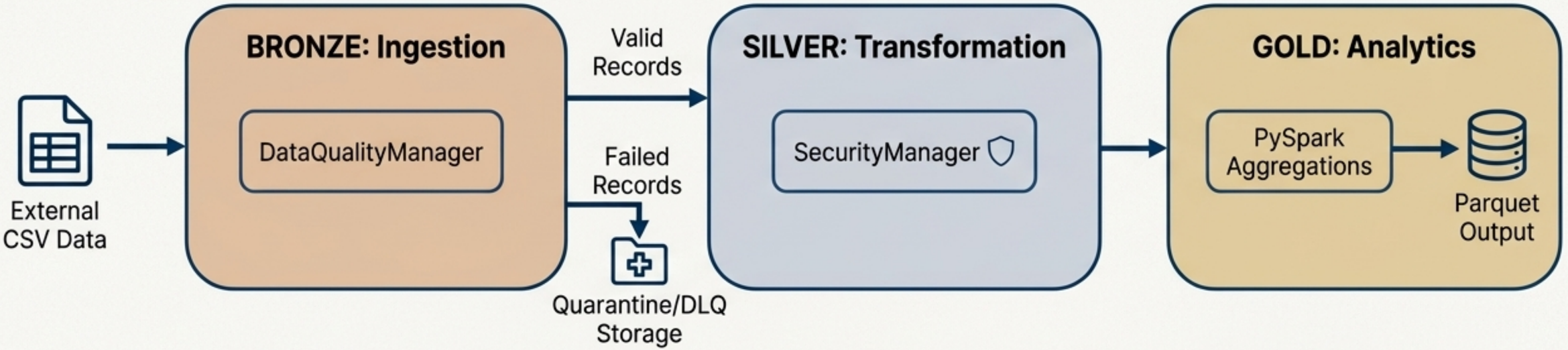
Traditional banking infrastructure presents unacceptable risks and performance bottlenecks in a real-time, high-threat environment.

- **Critical PII Exposure:** Sensitive customer data in plain-text format creates a constant threat of breaches and severe non-compliance penalties under regulations like **GDPR** and **PCI-DSS**.
- **Data Silos & Integrity Issues:** Information is trapped in monolithic systems, preventing effective aggregation for risk modeling and business intelligence. **Data quality** is often **unverified**, leading to unreliable reporting.
- **High-Latency Batch Processing:** Outdated processing models introduce significant delays, impairing the ability to **perform timely fraud detection** and meet the rigorous reporting demands of standards like **BCBS 239**.



# The Medallion Blueprint: A Disciplined Flow for Data Integrity

Our three-tiered architecture progressively refines, secures, and prepares data for **mission-critical analytics**, ensuring quality and control at every stage.




### Bronze (Ingestion)

Raw data lands and is immediately validated by the `DataQualityManager`. Failed records are automatically automatically quarantined into a Dead-Letter Queue (DLQ), ensuring 100% pipeline uptime and preventing data corruption downstream.

### Silver (Transformation)

Cleansed data undergoes security hardening. The `SecurityManager` applies irreversible hashing and strong encryption to all sensitive fields.
**No plain-text PII proceeds past this point.**

### Gold (Analytics)

Fully secured, high-quality data is aggregated by PySpark into optimized Parquet files, ready for high-performance risk modeling and regulatory reporting.

# Security by Design: A Zero-Trust Approach to Data

We implement non-negotiable cryptographic controls at the core of the transformation layer, eliminating plain-text risk entirely and satisfying PCI-DSS requirements.

SSN → 0x...

## PII Anonymization (Hashing)

Sensitive identifiers are pseudonymized using one-way **SHA-256 hashing**. This allows for data linkage without exposing the original identity, making reverse-engineering computationally impossible.


## Full-Field Encryption (AES-256)

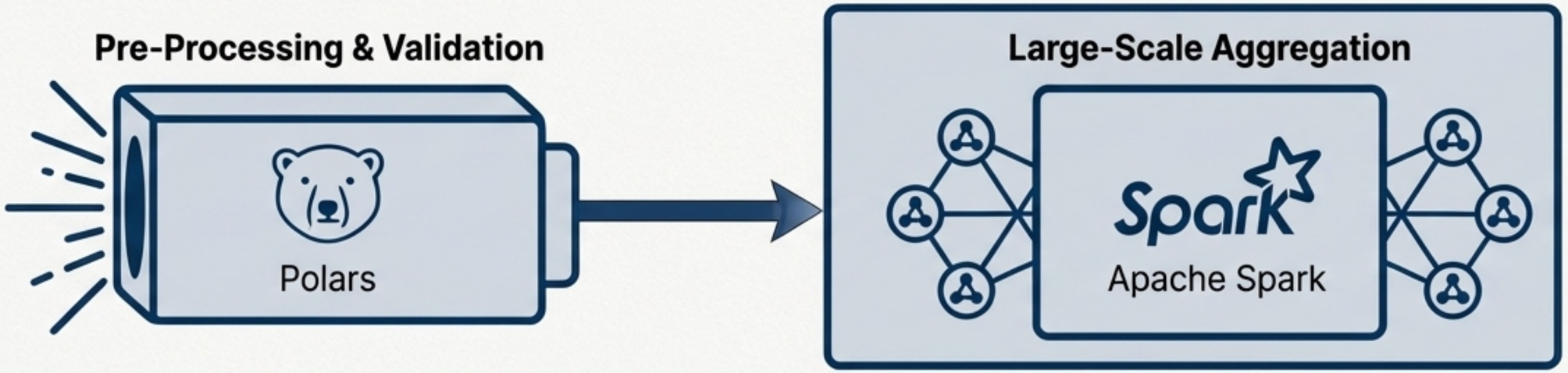
All other sensitive columns are encrypted using a high-performance, **vectorized Fernet (AES-256)** implementation. This ensures data is protected both at rest and in memory during transformation.

## Isolated & Verifiable Cryptographic Core

All security logic is encapsulated in a dedicated module. This core is rigorously validated with its own suite of `pytest` unit tests, providing a verifiable chain of custody for all cryptographic operations.

# Engineered for Efficiency: The Dual-Engine Advantage

We leverage the right tool for each job, maximizing processing speed and **dramatically reducing cloud infrastructure costs** by avoiding over-provisioning.




- **Pre-Processing with Polars:** For initial schema validation and transformations in the Bronze-to-Silver layer, we use the lightning-fast **Polars** engine. This avoids the high startup and orchestration overhead of a distributed cluster for tasks that don't require it.
- **Aggregation with Apache Spark:** For business logic and large-scale aggregations on the secured Silver data, we scale out with **PySpark's** distributed computing engine. Performance is further boosted by using **Apache Arrow-powered Vectorized UDFs**, delivering up to **10x processing speed** over standard implementations.
- **The Economic Impact:** This 'right-sizing' of compute power leads to a significant reduction in TCO by lowering cloud spend and shortening job completion times.

# Automated Governance & Bulletproof Compliance

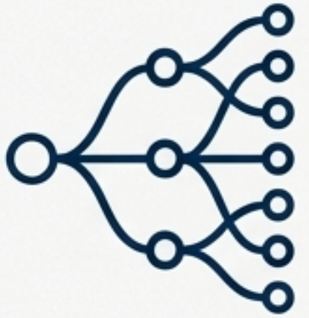
Embedding data quality contracts and regulatory principles directly into the pipeline's DNA for proactive, auditable control.


## Proactive Data Quality with Great Expectations

We define "what good data looks like" as executable code. **Great Expectations** automatically enforces these data contracts, preventing schema drift and quality degradation before they impact business users.

## Audit-Ready for BCBS 239

The architecture is explicitly designed to meet the rigorous principles of **BCBS 239** for risk data aggregation and reporting, ensuring data integrity, timeliness, and accuracy.



## End-to-End Traceability with OpenLineage

Integrated support for **OpenLineage** provides a clear, machine-readable audit trail of the data's entire journey, satisfying regulatory demands for transparency and verifiable lineage.

# Professional Infrastructure & Operations

A cloud-native foundation built with Terraform and GCP for repeatable, scalable, and secure deployment from day one.


- **Infrastructure as Code (IaC):** The entire cloud environment is provisioned using Terraform and Google Cloud Platform (GCP). This eliminates manual configuration errors and ensures a consistent, secure, and easily replicable setup.
- **Integrated Secret Management:** Leverages GCP Secret Manager to securely handle all credentials and sensitive configuration, removing them from code and adhering to security best practices.
- **Flexible Deployment Models:** The platform is designed for professional workflows, supporting local development (Windows portable), containerized execution with Docker for consistency, and seamless deployment to GCP for production workloads.

# The Business Impact: A Future-Proof Data Foundation

Transforming the bank's data capability from a technical liability into a strategic, revenue-enabling asset.

## Fundamentally Secure

Eliminates the risk of plain-text PII breaches, demonstrably meets stringent **PCI-DSS** and **GDPR** standards, and protects the bank's brand and reputation.

## Infinitely Scalable

The cloud-native design and distributed engines ensure the platform can handle exponential data growth without costly re-architecting.

## Completely Auditable

Every step—from quality checks to cryptographic transforms—is logged and traceable via **OpenLineage**, providing unparalleled transparency for both internal audit and external regulators.

## Economically Efficient

Delivers faster, more reliable insights to the business at a lower total cost of ownership through intelligent engineering and automation.