Engineering the Global Identity Data Pipeline

The Foundation for High-Fidelity Identity Graphs and Intelligence Systems

WHITE PAPER | JUNE 2026

1datapipe®

Contents

03 EXECUTIVE SUMMARY

04 THE GLOBAL IDENTITY DATA CHALLENGE

05 MEDALLION ARCHITECTURE

06 MEDALLION ARCHITECTURE: AT A GLANCE

07 WHY MEDALLION ARCHITECTURE

08 ARCHITECTURE COMPONENTS OVERVIEW

09 HOW DATA MOVES THROUGH THE PIPELINE

10 WHAT THE PIPELINE PRODUCES

11 BUILT FOR INTEGRATION

12 ACTIVATING IDENTITY INTELLIGENCE

13 INTELLIGENCE DOMAINS

14 ARCHITECTURAL DISCIPLINE AT GLOBAL SCALE

15 IDENTITY INFRASTRUCTURE AS A STRATEGIC ASSET

16 ABOUT 1DATAPIPE®

Executive Summary

In today's data-driven global economy, financial institutions and enterprises face an urgent challenge: how to unify massive volumes of identity data scattered across multiple countries, formats, and regulatory environments.

Living Identity® is 1datapipe's persistent identity graph — a unified, continuously evolving representation of individuals built by connecting identifiers, attributes, and events across sources, geographies, and time.

The 1datapipe® global pipeline exists to feed, maintain, and govern this graph— enforcing data quality, normalization, lineage, and verification at scale to ensure every identity is accurate, explainable, and continuously up to date—producing compliant, interoperable, high-fidelity identity data that enables trusted resolution, regulatory confidence, and advanced analytics across financial services and AI-driven use cases.

This white paper outlines the engineering principles, architecture choices, and technical frameworks that enable 1datapipe® to process nearly two billion profiles with precision, compliance, and operational efficiency at global scale.

Architected for scale, auditability, and regulatory alignment.

KEY METRICS

What This Paper Covers

- The global challenge of fragmented identity data
- How Medallion Architecture ensures scale, quality, compliance
- Bronze → Silver → Gold pipeline framework
- Country-specific ID logic and validation rules
- Technical components enabling automation and speed
- The value of the Living Identity® output
- Use cases: fraud, scoring, inclusion

Architectural Impact

Modern financial ecosystems depend on the integrity, accuracy, and consistency of identity data. By enforcing standards, lineage, and verification across multi-country identity datasets, 1datapipe® provides the foundation for trusted identity intelligence—supporting safer onboarding, stronger fraud defenses, and reliable AI-driven decisioning at global scale.

Architecture, not aggregation.

26 MARKETS COVERED

LATAM

Brazil · Chile · Colombia
Ecuador · Mexico · Venezuela

SOUTH EAST ASIA

Bangladesh · Indonesia
Malaysia · Myanmar · Pakistan
Philippines · Sri Lanka
Thailand · Vietnam

MIDDLE EAST & NORTH AFRICA

Algeria · Egypt · Morocco
Qatar · Saudi Arabia · Turkey
UAE

AFRICA

Cameroon · Kenya · Nigeria
South Africa

The Global Identity Data Challenge At Graph Scale

Modern enterprises face a fundamental barrier when working with identity data across multiple countries: fragmentation—preventing the formation of persistent, unified identity graphs at scale. Every market, every dataset, and every source system introduces inconsistencies that make identity verification, analytics, and AI-driven decisions unreliable or impossible to scale.

Across 26 global markets, 1datapipe® evaluated the core issues preventing organizations from achieving a unified identity foundation:

The Five Critical Challenges

01

Country-Specific ID Formats

Each country enforces unique government ID structures, validation rules, and embedded attributes — creating incompatible datasets.

Impact: Validation failures, mismatched profiles, country-level silos.

02

Data Quality and Duplication

Government ID datasets contain missing values, formatting inconsistencies, and duplicate identities from multiple sources.

Impact: Inflated risk scores, poor matching, synthetic identities.

03

Scalability Requirements

Processing billions of identity records requires distributed, cloud-scale architecture. Legacy pipelines bottleneck.

Impact: Slow refresh cycles, stale data, limited AI workload readiness.

04

PII and Compliance Mandates

Each country enforces unique privacy, residency, and age validation rules — requiring masking, filtering, selective processing.

Impact: Regulatory risk, fines, no unified cross-border models.

05

Identity Continuity Over Time

Identities evolve — names change, addresses update, new identifiers are issued. Without persistent linkage, continuity is lost.

Impact: Broken histories, weaker fraud signals, no behavior tracking.

Together, these challenges fragment identity at every layer.

Solving them requires architecture, not aggregation.

Medallion Architecture

To unify identity data across 26 countries, 1datapipe® required a pipeline design capable of handling extreme variability in data formats, regulatory environments, and source quality. The solution is a Medallion Architecture–based system designed to construct and maintain the Living Identity® graph— engineered for global scalability, high performance, and regulatory-aligned compliance controls. This architecture ensures that fragmented identity data is continuously transformed into persistent, graph-ready entities.

The Medallion approach introduces a layered model — Bronze → Silver → Gold — where each stage progressively improves data quality, structure, and readiness for downstream identity intelligence applications.

This layered design builds reliability into the pipeline at every step, ensuring that raw government ID data becomes standardized, deduplicated, enriched, and ready for real-time decisioning systems.

Built for persistent identity intelligence, not single-use checks.

Why Medallion Architecture Works for Global Identity Data

01

Traceability at Scale

Because each layer builds upon the previous one, data lineage is preserved end-to-end — essential for regulated industries and audit requirements.

02

Configurability Through YAML Modules

Country-specific logic is defined through configuration, not hard-coded logic. This enables rapid onboarding of new markets and updates without disruptive redevelopment cycles.

03

Distributed Processing for Large-Scale Datasets

Built on PySpark and Delta Lake, the pipeline is designed to process datasets at national scale with ACID guarantees and parallelized execution.

04

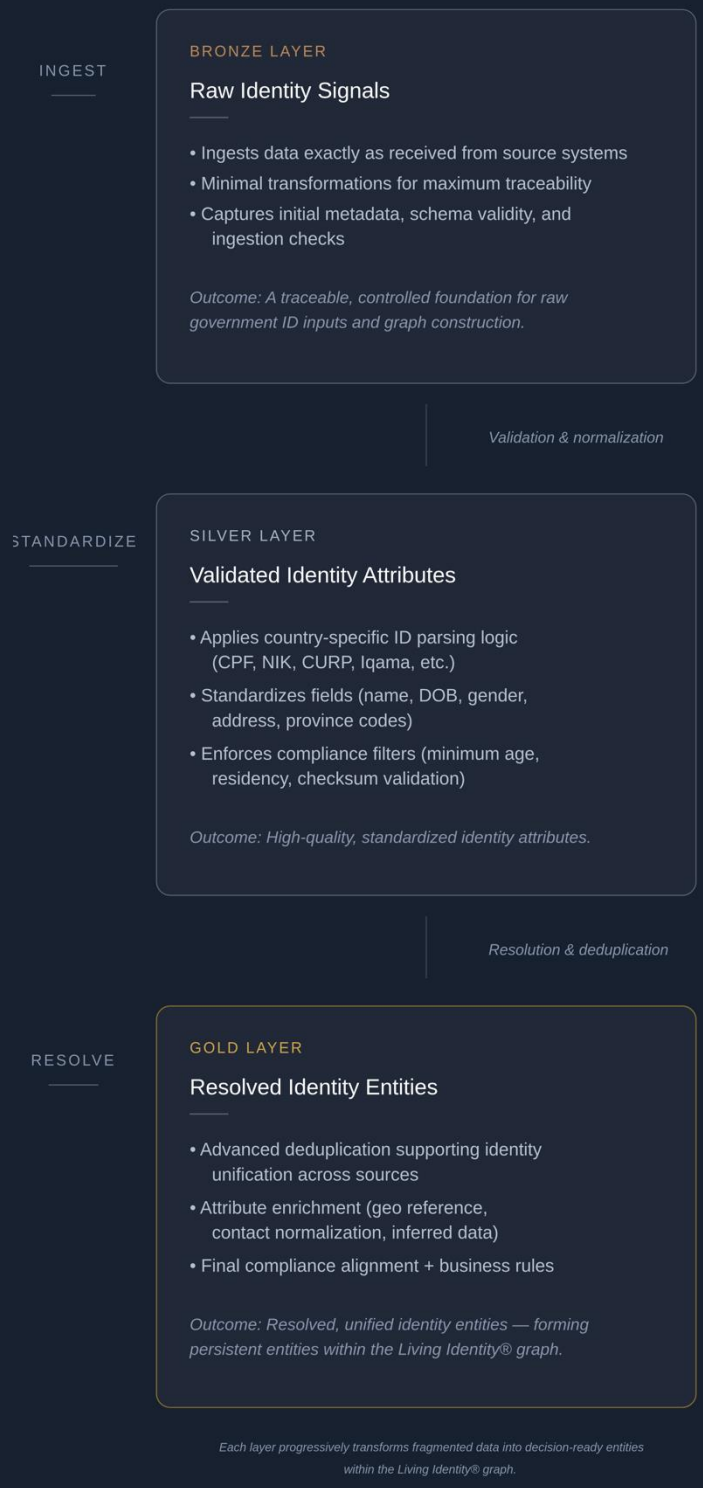
Compliance by Design

Requirements for data residency, PII handling, masking, and validation are embedded directly into each stage of the architecture.

Medallion Architecture: At A Glance

Each layer progressively prepares and transforms identity data for inclusion in the living identity® graph. This layered design is powered by country-specific validation, normalization, and deduplication logic — configured through modular YAML-based rules and executed at scale.

Three layers.
One foundation.



Each layer progressively transforms fragmented data into decision-ready entities within the Living Identity® graph.

This layered design is powered by country-specific validation, normalization, and deduplication logic — configured through modular YAML-based rules and executed at scale.

Why Identity Intelligence Requires a Medallion Architecture

Identity intelligence places fundamentally different demands on data systems than traditional analytics. Identity records must remain accurate over time, explainable across decisions, and resilient across fragmented, regulated, and evolving data environments.

At global scale, these requirements cannot be satisfied through ad hoc pipelines or downstream controls. Identity intelligence depends on a layered, lineage-preserving architecture to function reliably.

Identity intelligence only works when scale, governance, and explainability are structural.

Explainability Across the Identity Lifecycle

Identity decisions must remain explainable long after they are made. For regulatory review, fraud investigation, or dispute resolution, organizations must be able to trace how an identity record was formed, modified, and used over time.

A lineage-preserving architecture enables end-to-end auditability and correction without compromising system integrity, performance, or regulatory compliance.

Consistency Across Jurisdictions

Identity logic varies by country, regulator, and ID system, but identity behavior must remain consistent globally. Without architectural controls, localized rules fragment identity logic, creating inconsistent outcomes and governance risk.

Configuration-driven approaches apply country-specific requirements without hard-coded logic, preserving global consistency while enabling controlled regional adaptation.

Population-scale Identity Resolution

Identity intelligence must operate at population scale—not just for point-in-time checks. Government-issued identity systems and fraud platforms require consistent resolution across millions or billions of records, often with incomplete or variable data.

A layered, lineage-preserving architecture enforces standardization and governance at each stage of processing, maintaining accuracy, auditability, and operational stability as datasets expand across markets.

Architecture Components Overview

Each layer progressively prepares and transforms identity data for inclusion in the Living Identity® graph.

Configurable by market.
Consistent by design.

Country-Specific Modules

Extensible identity logic, defined in YAML

ID format rules (length, regex, encoding)
Checksum and authenticity logic
Province, region, and state mappings
Localization (accents, transliteration)
Age and residency validations
Local privacy, masking, and PII rules

Configurable per market

Utility Modules

Reusable shared logic across all markets

Schema Validator
Quality Evaluator
Deduplication Engine
Enrichment Services
Error Handling and Logging

Base Framework — Core Engine

PySpark + Delta Lake + YAML-Defined Logic

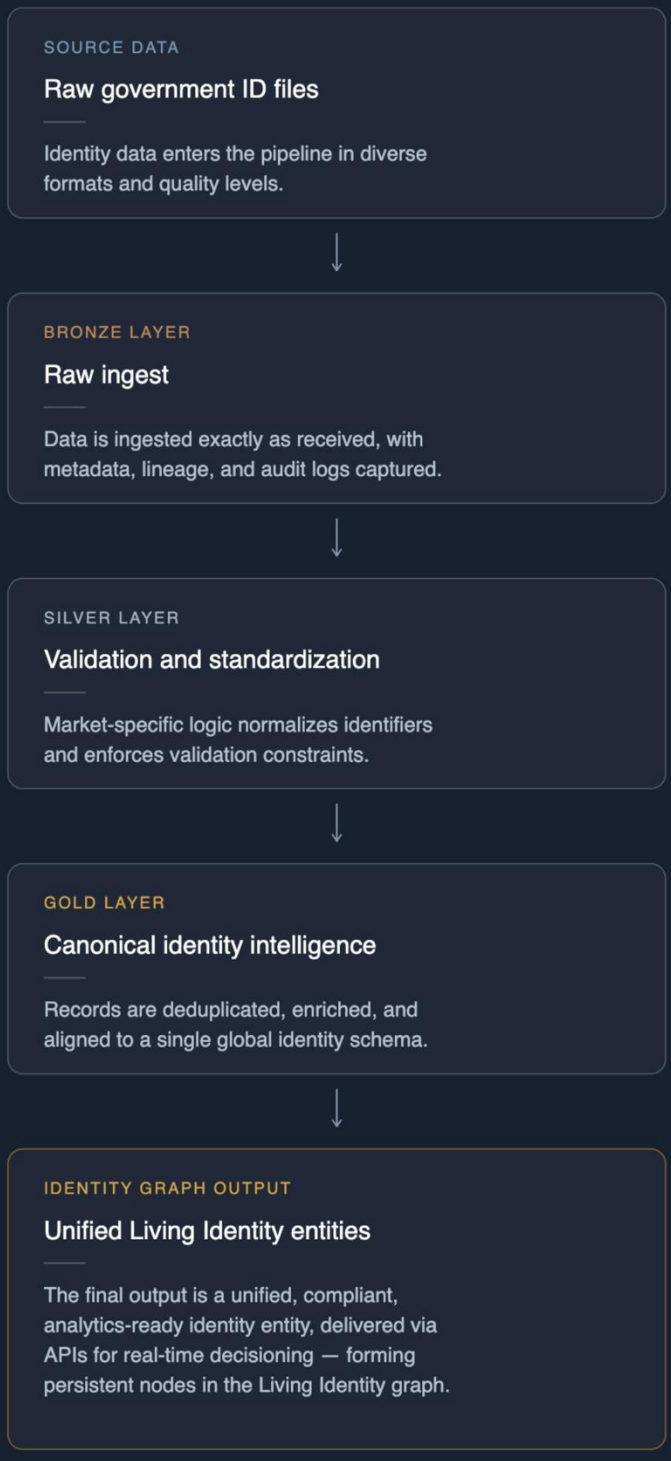
Distributed execution
ACID storage and transactions
End-to-end data lineage
Incremental processing
Versioning and reproducibility

Ensures stability, repeatability, and compliance at scale

How Data Moves Through the Pipeline

Each stage progressively improves data quality, structural consistency, and readiness for identity decisioning—transforming fragmented inputs into persistent identity entities within the Living Identity® graph.

From Raw Signals to Resolved Entities



Each stage progressively transforms fragmented data into decision-ready entities.

What the Pipeline Produces

The output of the pipeline is not a file or a one-time report — it is a continuously maintained set of identity entities within the Living Identity® graph. Each entity is built to be reused across decisioning systems, refreshed as new data arrives, and governed by the same compliance rules applied during processing.

Continuously refreshed, versioned, and governed.

Characteristics of the Living Identity® Output

Unified Schema

Each profile conforms to a single global identity model, enabling consistent use across regions.

High-Completeness Profiles

Core attributes are consistently populated, supporting high-confidence verification and modeling.

Continuously Updated

Profiles are refreshed on a regular cadence to reflect the most current available data.

Compliance-Aligned by Design

Output respects local privacy, residency, and regulatory constraints embedded upstream.

Analytics-Ready

Data is delivered in a structure optimized for scoring, matching, and downstream decisioning systems.

Fully Traceable

Every attribute carries full lineage, supporting auditability across regulatory and enterprise review.

What this enables

Faster, High-Confidence Onboarding

Reduced friction with high-confidence identity verification across markets.

Stronger Risk & Fraud Detection

Consistent identity signals for risk assessment and cross-border decisioning.

Reusable Identity Intelligence

A standardized identity layer for onboarding, risk, analytics, and AI workloads.

Built for Integration

Living Identity® is delivered as standardized profiles via APIs and secure data interfaces, allowing enterprises to integrate identity intelligence directly into onboarding flows, fraud systems, and analytics platforms—enabling access to persistent identity entities within the Living Identity® graph without custom per-market logic.

Continuously refreshed, versioned, and governed.

Integration modes

API-Based Access

RESTful endpoints for real-time identity lookups, integrated directly into onboarding flows and decisioning systems.

Standardized request and response formats across all markets, with consistent identity entity structure regardless of source country.

Real-Time & Batch

Continuous refresh for live decisioning, or scheduled batch delivery for analytics and modeling workloads.

Same governed identity foundation, accessed at the cadence your business operations require — from real-time decisions to overnight workloads.

Enterprise Integration

Compatible with major data platforms and cloud warehouses, supporting existing data pipelines without rework.

Designed to plug into the systems enterprises already run, with no per-market customization required.

One governed identity foundation. Multiple ways to consume it.

A single identity layer, reused across every decisioning system in the enterprise.

Activating Identity Intelligence

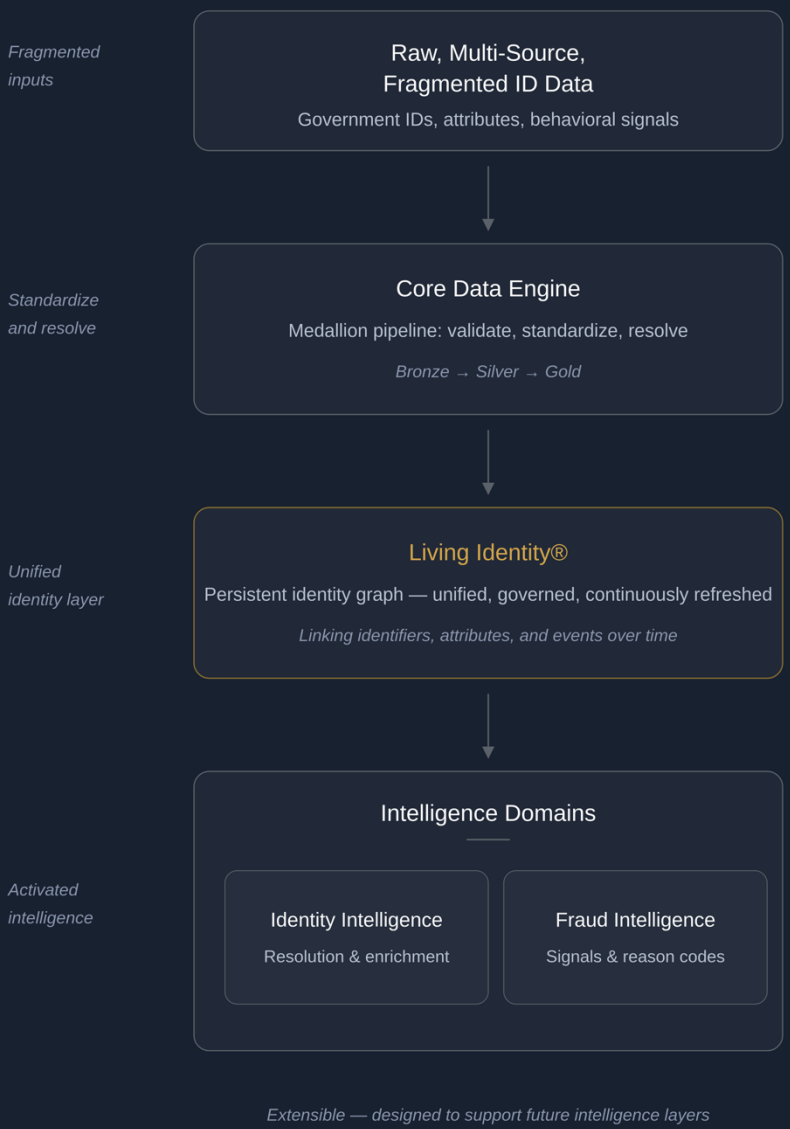
Living Identity® is not an endpoint—it is a persistent identity graph and reusable intelligence layer designed to power multiple decisioning systems across the enterprise. Once identity data has been unified, verified, deduplicated, and enriched, it becomes a trusted input for analytics, models, and workflows that operate across markets and use cases.

This transformation is enabled by the underlying identity graph, where persistent entities and their relationships provide the foundation for consistent, cross-market intelligence. It allows organizations to move beyond point-in-time identity checks toward continuous, population-scale intelligence applied across onboarding, fraud prevention, risk assessment, and customer engagement.

One identity foundation. Multiple intelligence outcomes.

HOW IDENTITY BECOMES INTELLIGENCE

The flow from raw identity data to decision-ready intelligence can be summarized as:



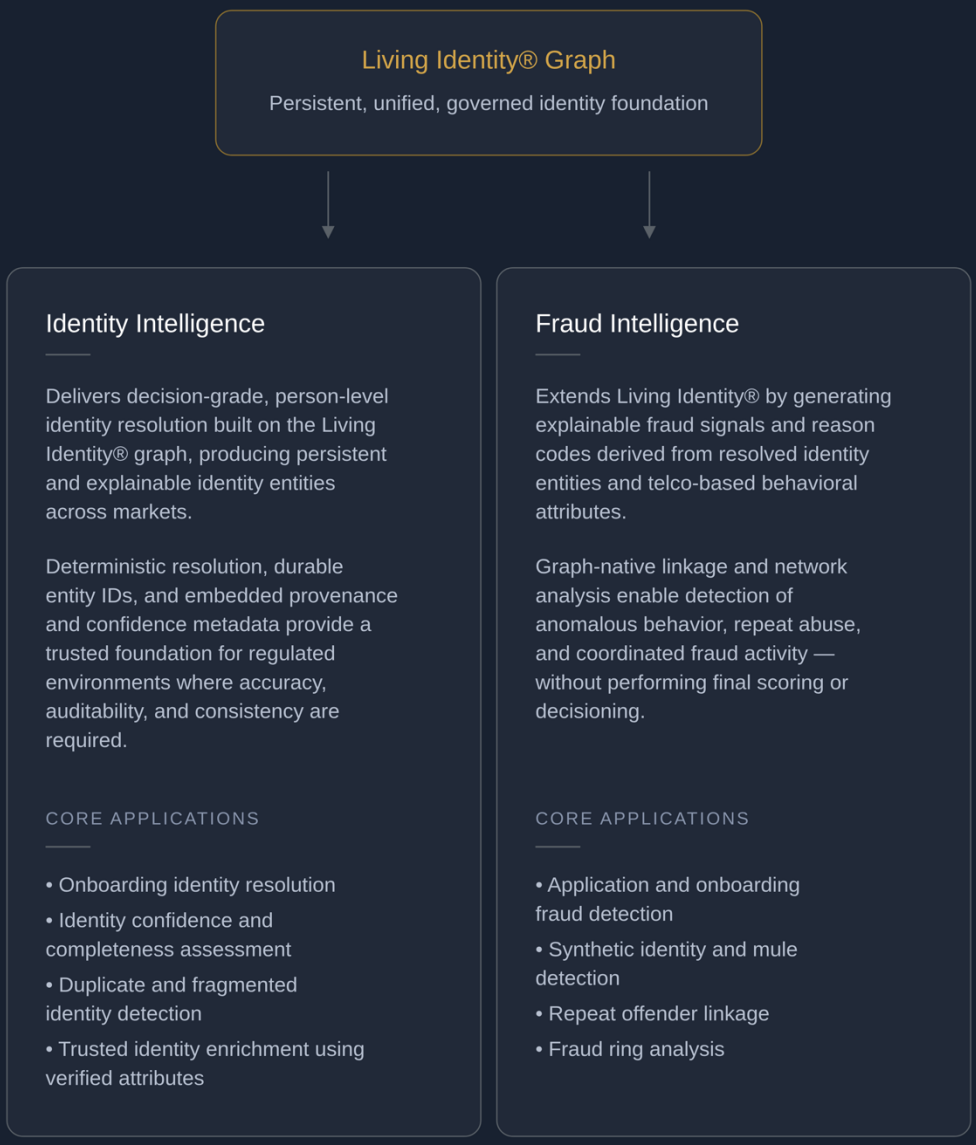
Extensible — designed to support future intelligence layers

INTELLIGENCE DOMAINS

The Living Identity® graph is not a single product — it is a foundation that powers multiple intelligence layers across the enterprise. Once identity data is unified, verified, and resolved, it becomes a reusable input for analytics, models, and decisioning systems that operate across markets and use cases.

Today, two intelligence domains operate directly on the graph. The architecture is designed to support additional domains as identity ecosystems evolve.

Different domains.
Same governed graph.



Both domains operate directly on the same governed identity graph.

One identity foundation. Multiple intelligence outcomes.

Architectural Discipline at Global Scale

Global identity intelligence—and the integrity of the underlying identity graph—cannot be delivered through ad hoc pipelines or market-specific implementations. It requires an architecture designed from the outset to operate across heterogeneous data environments, regulatory frameworks, and quality baselines — without sacrificing consistency or trust.

The Medallion Architecture underpins the Living Identity® graph by enforcing discipline at every stage of the data lifecycle, ensuring identity intelligence remains reliable, auditable, and scalable across all markets.

Architectural controls enforced across every layer.

ARCHITECTURE-ENFORCED GUARANTEES

These architectural guarantees ensure the stability, explainability, and long-term integrity of the Living Identity® graph at global scale.

End-to-End Traceability

Every identity record maintains full lineage from ingestion through output. Transformations are version-controlled, auditable, and reproducible to support regulatory review and enterprise governance.

Configuration Without Fragmentation

Country-specific logic is applied through configuration rather than custom code, enabling rapid market extensions and consistent behavior across regions without fragmenting the core pipeline.

Population-Scale Determinism

Distributed processing and ACID-compliant storage support deterministic performance at national and multi-national scale, even under heavy operational workloads.

Compliance Embedded by Design

Privacy, residency, masking, and validation rules are enforced directly within the pipeline, making compliance structural to how identity data is processed and produced.

ONE IDENTITY LAYER, GLOBALLY CONSISTENT

By standardizing identity outputs across all markets, the architecture enables a single identity representation to be reused consistently across onboarding, fraud, risk, and AI-driven decisioning systems — eliminating the need for market-specific adaptations.

ENTERPRISE IMPACT

This architectural approach transforms fragmented identity data into a durable, governed identity graph—supporting scale, regulatory confidence, and consistent decisioning across markets through a single, unified identity layer.

Identity Infrastructure as a Strategic Asset

Government-issued identity data is inherently fragmented—across formats, jurisdictions, data owners, and regulatory regimes—making it difficult to establish a persistent, unified identity graph at scale. This fragmentation limits the ability of institutions to build scalable, trusted identity systems. Solving this problem requires more than aggregation or point solutions; it requires purpose-built identity infrastructure designed for global and regulatory complexity from day one.

1datapipe® addresses this challenge through a Medallion Architecture–based identity pipeline that constructs and maintains a persistent, governed identity graph from fragmented source data. Traceability, provenance, configurability, performance, and governance are embedded directly into the identity lifecycle—ensuring consistency and trust across markets, rather than attempting to impose them downstream.

The result is Living Identity®—a persistent identity graph that maintains entity continuity across sources, attributes, and time, operating at population scale across regulated environments.

As infrastructure, the Living Identity® graph enables multiple intelligence domains—foundational identity intelligence and derived fraud intelligence—while remaining signals-first, auditable, and partner-compatible. This foundation is intentionally designed to support future intelligence layers, including deeper graph-based reasoning and AI-driven analytics, without compromising governance, explainability, or regulatory rigor.

At global scale, trust is not asserted—it is engineered into identity graph infrastructure designed to persist, govern, and scale.

1datapipe® operates a governed identity graph covering 24 markets and nearly 2 billion verified profiles.

Built on the architectural principles in this paper, the graph supports enterprise-scale onboarding, fraud prevention, and AI-driven decisioning — and is designed to extend into deeper graph-based intelligence as identity ecosystems evolve.