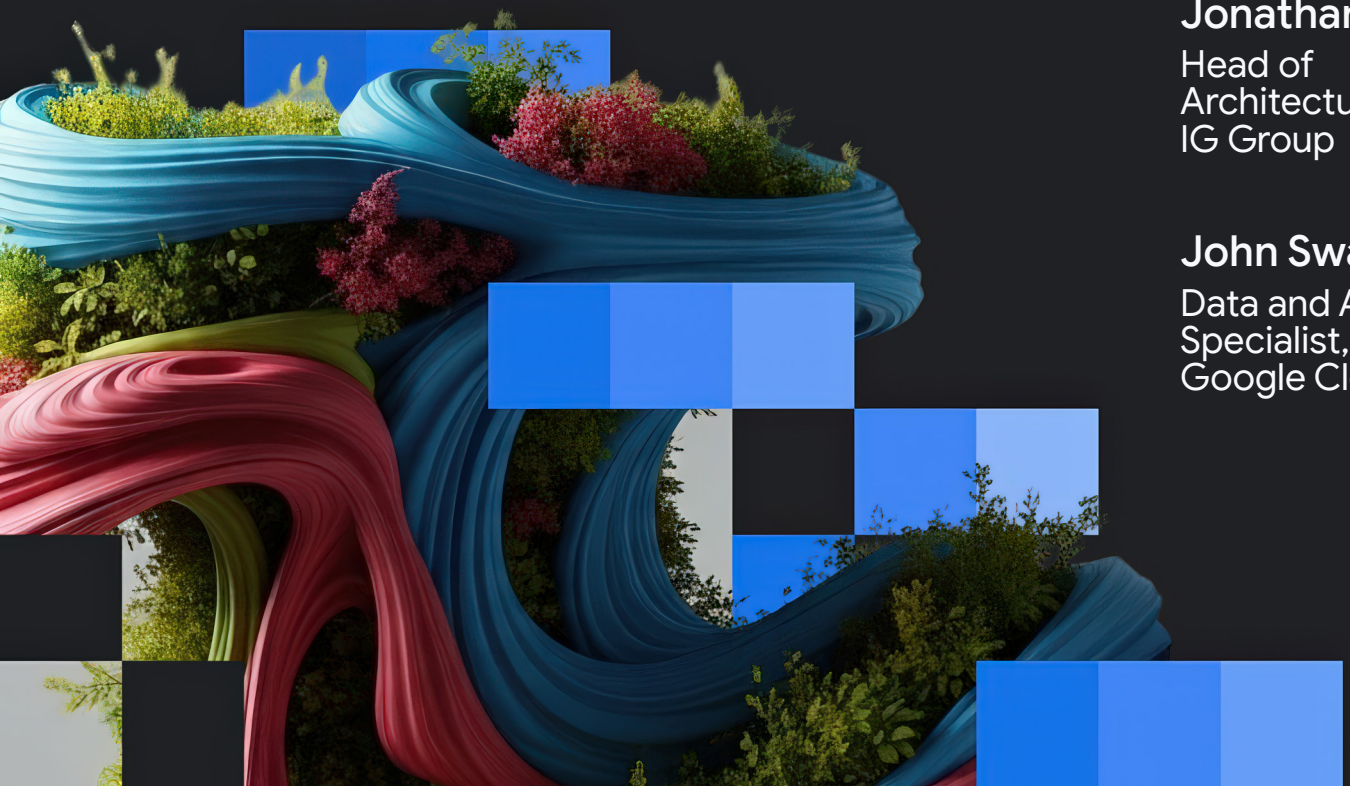
Google Cloud

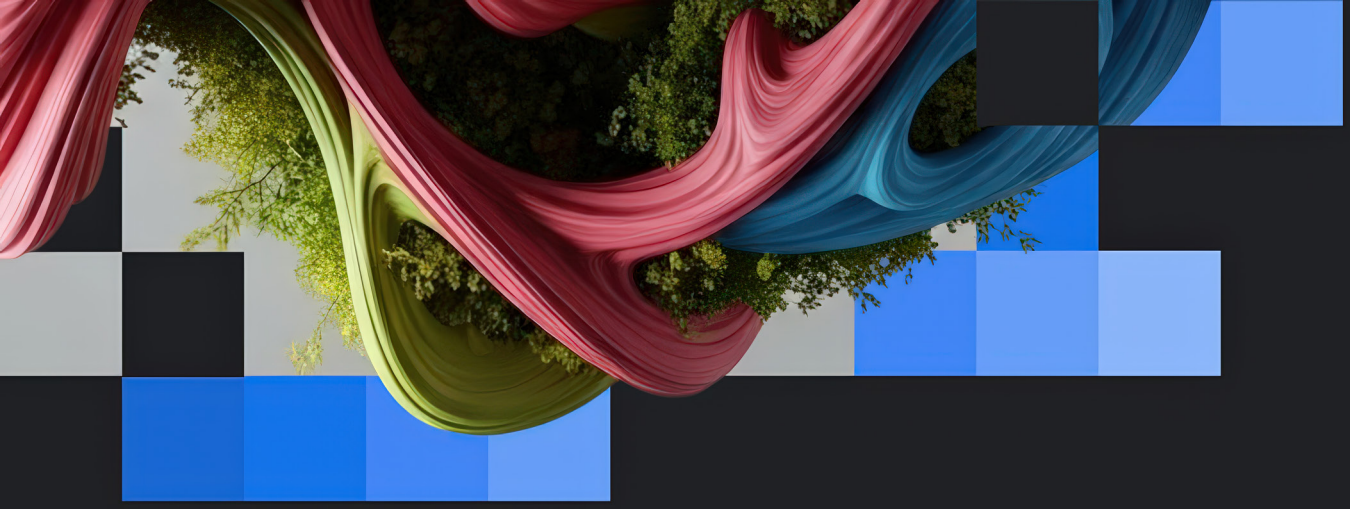
Beyond the data bottleneck

Scaling innovation with the extended medallion architecture.



Jonathan King
Head of Architecture, IG Group

John Swain
Data and Analytics Specialist, Google Cloud



Executive summary

Data teams still spend inordinate amounts of time on prep work, causing backlogs and bottlenecks that force businesses to watch opportunities pass them by.

The root cause of so much prep? Source-system data is complex and lacks conformity. When data is ingested in its raw state, users must navigate cryptic schemas and inconsistent formats. It's exhausting work that creates doubt around the data produced—hence more prep work.

The medallion architecture addresses this challenge through progressive data refinement across bronze, silver, and gold zones. However, traditional implementations create a critical bottleneck: central data engineering teams must own all zones to maintain quality and governance. This means domain teams need to queue for resources, overwhelming the data team and slowing business innovation.

This guide introduces the extended medallion architecture—a refined pattern designed to help organisations break through the central bottleneck while preserving data quality and governance.

By introducing domain-specific gold zone projects, traditional management overheads are removed and domain teams get the ownership and autonomy they need, all while maintaining centralised governance of the foundational data layers. Built on Google Cloud using BigQuery, Cloud Storage, and dbt, the extended architecture delivers 'ready data' that is immediately discoverable, accessible, trustworthy, and usable.

Table of contents

The ‘ready data’ advantage

Modern enterprises generate data at unprecedented volume and velocity.

Yet data teams spend the majority of their time on things like cleansing complex data, preparing datasets, and discovering what data exists and what it means. While these are absolutely necessary, they can and should be automated to free up time for value-adding activities.

This overemphasis on preparatory work is a fundamental failure of data platforms—holding subject matter experts, analysts, and data scientists back from capitalising on business opportunities.

The core problem isn't volume or velocity. It's the complexity of source system data. When platforms ingest data and leave it in its raw state, every consumer must independently understand proprietary schemas, decode cryptic column names, and reconcile inconsistent formats across systems.

The solution is to make understandability the primary output of the data platform. Data must be structured not just for storage efficiency or processing speed, but for effective human comprehension and use. This is where the DATU framework comes in.

Setting the bar for ‘ready data’

The aim of any data architecture is to provide ‘ready data’, which is defined by four key characteristics.

01
Discoverable

Metadata is catalogued and governed. Users can find data easily through search and understand its context through rich documentation and lineage.

02
Accessible

Data is available via standard query languages with secure, performant access. Modern cloud platforms enable zero-copy architectures where users query data directly without moving it.

03
Trustworthy

Data has undergone validation, cleansing, and conforming transformations. Quality is verified, lineage is clear, and users can rely on accuracy for critical decisions.

04
Usable

Data is modelled dimensionally, aggregated where appropriate, and enriched to support business use cases directly. To further enhance usability, technical complexity is abstracted away from consumers.

These characteristics make up the DATU framework, which provides the criteria for evaluating whether a data platform truly delivers value or simply shifts the burden to consumers. And, until now, the traditional medallion architecture—with its structured refinement pipeline—has supported this framework.

The zones of the traditional medallion architecture

The medallion architecture is a data design pattern focused on progressively improving data structure and quality as it flows through three zones—bronze, silver, and gold.

Each zone builds on the previous outputs and, at each boundary, specific quality controls and structural improvements are applied. All data is immutable and read-only to ensure accuracy and maintain a clear source of truth.

Bronze zone: Raw fidelity

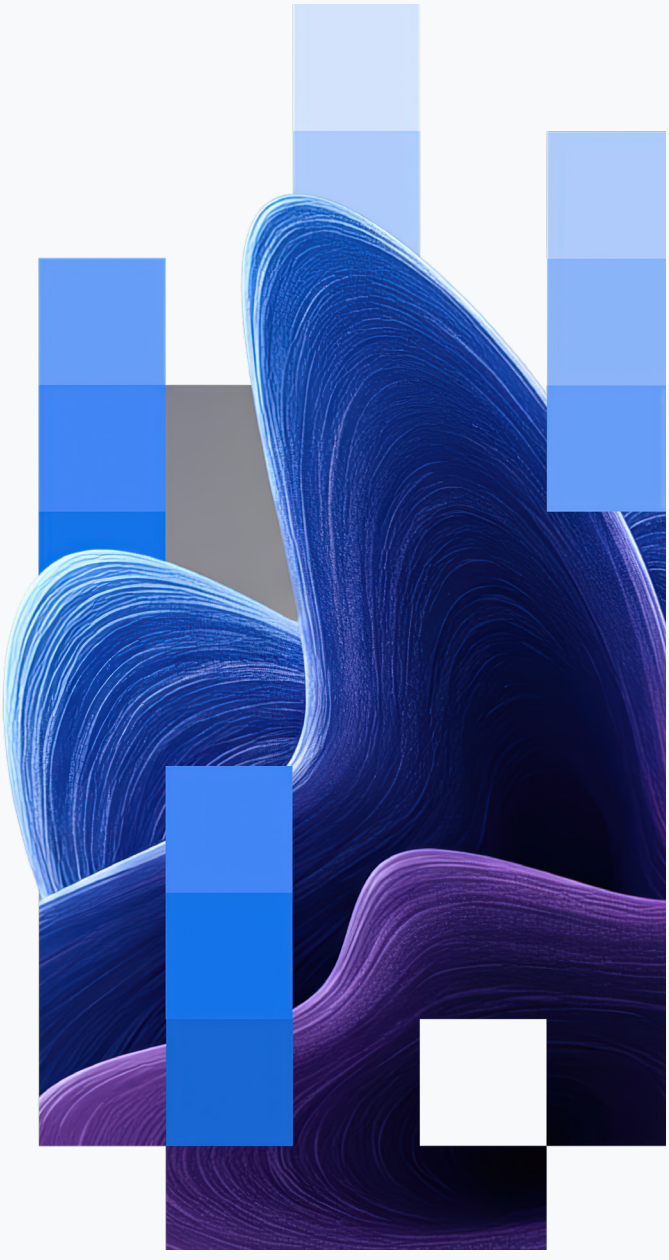
The bronze zone is the landing area for all structured data entering the platform, focusing on faithful ingestion and auditability.

Data structure remains close to the source system model with no business logic or complex transformations applied. The source system can be upstream operational systems (ideally over a messaging broker) or from external partners—the data pipeline from this point is the same.

It serves as the immutable, auditable record of what was received from source systems, enabling data lineage tracing and providing the foundation for reprocessing.

Core functions

- Preserve source system fidelity
- Deduplicate records and add technical metadata (ingestion timestamp, source system identifier, batch ID)
- Perform basic validation checks (field presence, type casting)



Silver zone: Enterprise truth

The silver zone focuses on enterprise-wide standardisation and quality.

It is the most critical zone for achieving understandability and transforming technical, source system-oriented data into a clean, canonical structure that reflects business entities and relationships.

The silver zone acts as an anti-corruption layer, so when a source system changes its data model, only the bronze-to-silver transformation requires adjustment. Multiple source systems representing the same business entity in different ways are harmonised into a single canonical representation, and complex business rules are implemented once rather than duplicated.

Core functions

- Transform data into a fully catalogued canonical data model with business terminology
- Flatten complex nested structures and apply dimensional modelling
- Unify entities across different source system versions or schemas
- Enrich with business metadata (descriptions, ownership, data quality indicators)
- Enforce referential integrity and comprehensive data quality checks

Gold zone: Business value

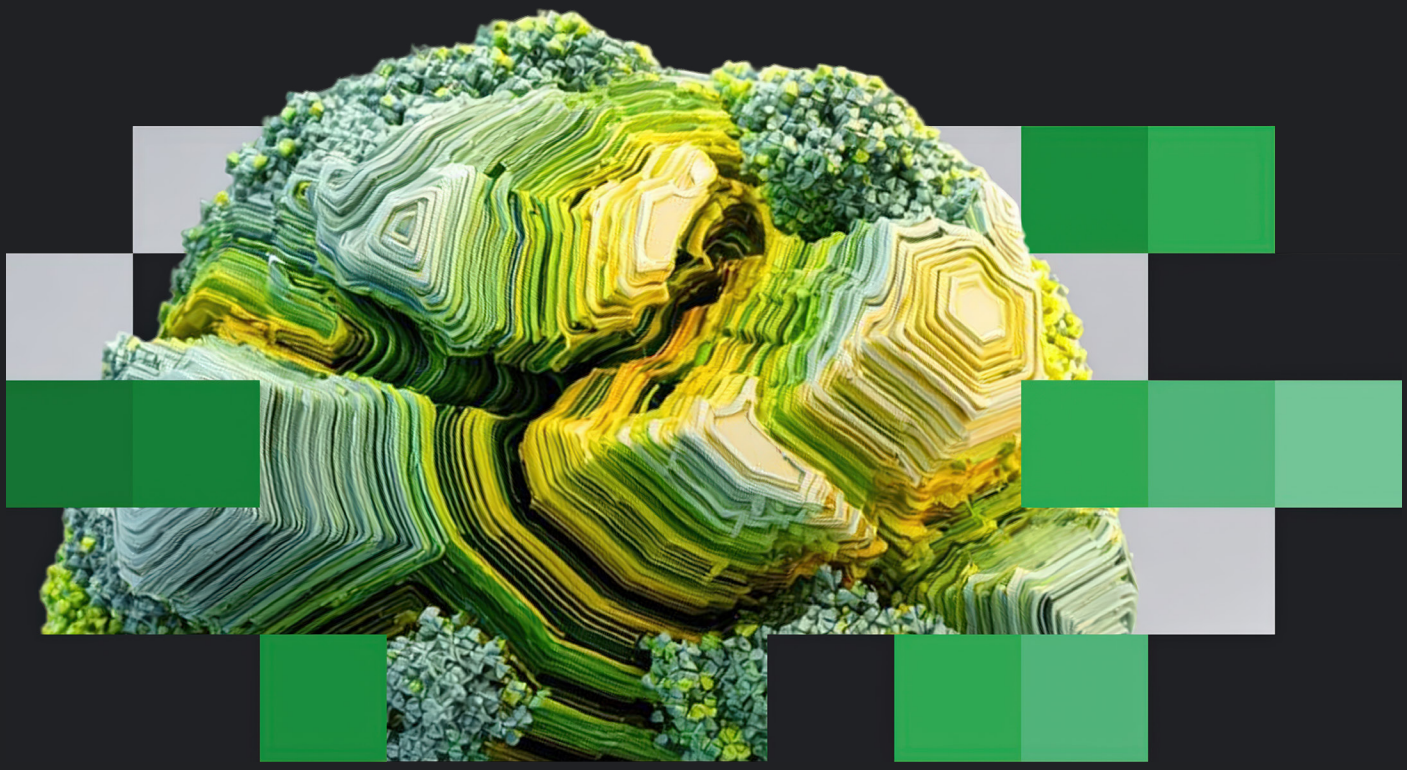
The gold zone contains curated, aggregated datasets aligned to specific business use cases.

Data reaches its highest level of refinement here, optimised for consumption by reporting, business intelligence, and operational dashboards. Multiple gold datasets may exist, each tailored to different consumption patterns. This data can be trusted without the consumer(s) needing to apply their own data quality controls.

Core functions

- Apply business logic to calculate metrics and KPIs
- Aggregate data by key business dimensions (time periods, organisational units, product categories)
- Project-only relevant columns for specific use cases
- Apply performance optimisations (clustering, partitioning, materialisation)

Limitations of the traditional medallion architecture



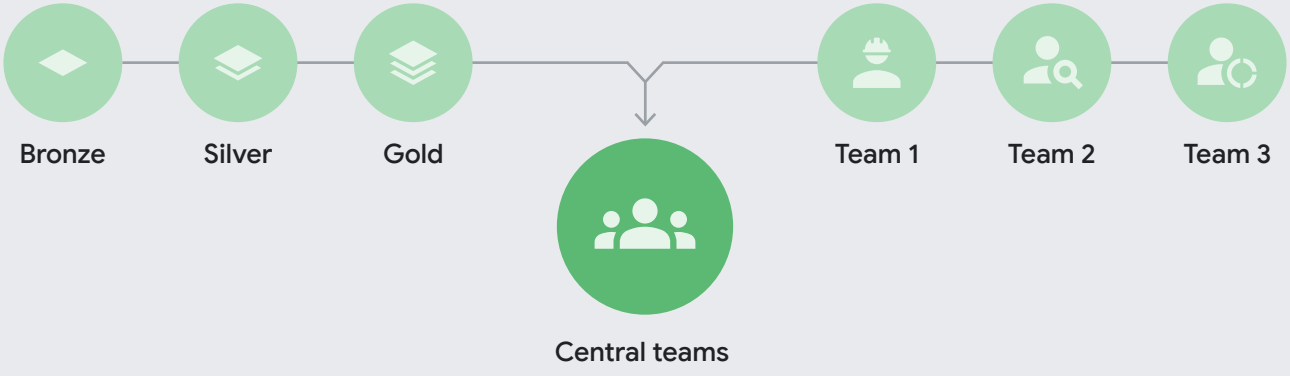
The medallion architecture's strength lies in its ability to progressively improve data quality through bronze, silver, and gold zones.

Each zone in this traditional architecture must be owned and controlled by a central data engineering team in order to maintain accuracy, immutability, and quality. Without this centralisation, critical risks emerge:

- Inconsistent application of data quality rules across zones
- Ungoverned transformations that bypass the canonical model
- Multiple competing versions of "truth" in the silver and gold zones
- Loss of lineage and auditability
- Degradation of the anti-corruption layer's protection

Yet centralisation can cause a problematic bottleneck.

The traditional medallion architecture is centrally controlled



Issues arise when a central team holds full control

With all work funnelled through a single team, there are conflicts between cost and priority. Different domains compete for limited data engineering capacity, and it becomes impossible to see which business units are actually driving value from the data.

Common problems include:

Domain teams queuing for resources

They must submit requests for every domain-specific dataset, aggregation, or transformation—even simple changes.

Underutilisation of domain expertise

The domain team's deep understanding of their business context can't be applied to data transformation. Knowledge is transferred to data engineers, causing delays and potential miscommunication.

Innovation limited to the central team's capacity

The pace of business innovation becomes limited by data engineering bandwidth, and domain teams can't move fast or independently.

Data engineering overwhelm

Central teams are overwhelmed with domain-specific requests that require deep business context—taking focus from their core mission of ingesting and modelling data in bronze and silver zones.

Beyond the bottleneck, other problems arise

The control-vs-agility dilemma

Organisations face an impossible choice between business speed and maintaining central control. Protecting data quality stalls innovation, but allowing domain teams to access and modify the gold zone leads to a wild west of ungoverned data, inconsistent models, and a lack of ownership and accountability. Most companies default to the safety of central control, while those that choose decentralisation often discover the governance costs only after data chaos has taken root.

External frictions

Sharing data with partners or regulators shouldn't mean opening a backdoor to your enterprise. In traditional architectures, external integrations often connect directly to the core lakehouse, creating a security risk where a single partner could access your sensitive canonical data.

Beyond security, organisations are often forced to change internal data models to meet a partner's formatting requirements. Managing multiple data model standards is a complex and risky compromise, making it nearly impossible to scale without putting your core data at risk.

Security and blast radius concerns

When all gold zone datasets reside in a single central project, managing fine-grained access for dozens of teams in one place becomes a juggling act of complex policies. You must account for all possible combinations of users, domains, and data sensitivity within one monolithic project structure.

At the same time, centralised zones make it difficult to contain a crisis. Without hard isolation boundaries between domains, a single security breach or a simple data quality error can have a much larger blast radius and ripple through the entire organisation.

A lack of cost clarity

It is nearly impossible to prove ROI when your compute and storage costs are all piled into one central bill. When domain-specific workloads are bundled with core engineering, finance teams can't allocate costs, report accurately, or determine who's driving value. And without attribution, domain teams aren't incentivised to write efficient queries or manage their own data footprint. Ultimately, organisations rely on estimations rather than actual consumption, making chargeback models or budget planning a matter of guesswork.

Competing for compute resources

When every workload runs on the same shared infrastructure, your most critical data pipelines compete with unrelated ad-hoc queries. A heavy ML training job or a massive, unoptimised search from one department can easily starve your core engineering pipelines of the resources they need to run. This creates a situation of unpredictable performance where data freshness SLAs are difficult to guarantee. Without dedicated compute resources for each domain, you're stuck either over-provisioning capacity and wasting money, or under-provisioning and leaving teams to fight for the same limited processing power.

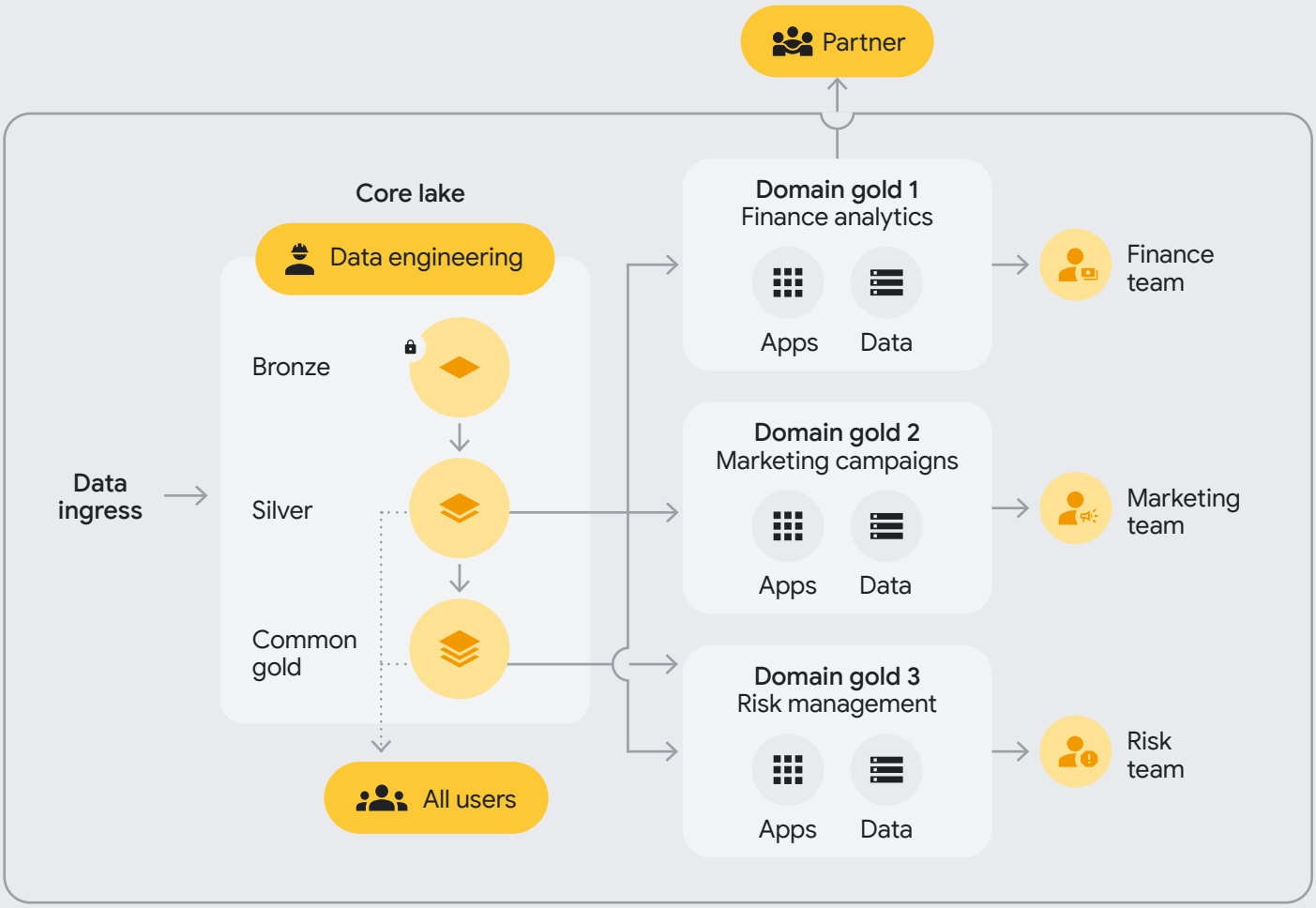
Why an extended medallion architecture is needed

The extended medallion architecture solves the control-versus-agility dilemma through a structural innovation: domain-specific gold zone projects that are separate from the core lakehouse infrastructure.

In this pattern, the standard medallion zones (bronze, silver, and common gold) remain under central data engineering ownership in a core lakehouse project. However, each business domain or division receives its own dedicated gold zone project.

The extended medallion architecture

Teams are aligned to domain-specific gold zones, which are created from refined silver or common gold datasets.



The domain projects in the extended medallion architecture are not replicas. Raw and canonical data live once in the core lakehouse, while domain projects contain either canonical model data (e.g., for partner access) or value-added derivatives, transformations, and enrichments.

In these consumption and transformation environments, domain teams read from the centrally governed silver zone and create domain-specific aggregations, features, and datasets. There are clear lines of responsibility.

As new domains emerge or analytical needs grow, new domain projects are created—scaling without increasing load on central data engineering.

Data team accountability

- Raw data ingestion, deduplication, and technical metadata in the bronze zone
- Canonical model, anti-corruption, and enterprise-wide data quality in the silver zone
- Cross-functional aggregations serving multiple domains in the common gold zone
- Owning and curating the central data catalogue

Domain team accountability

- Owning and closely collaborating on its areas of the canonical model
- Domain-specific transformations and aggregations in the domain gold zone
- Managing data protection and access control for the domain gold zone
- Feature engineering for domain ML models
- Supporting domain analysts and data consumers
- Managing domain-specific costs and performance
- Documenting and cataloguing domain data products

The extended architecture solves traditional challenges

Enabling control and agility

The extended medallion architecture maintains central governance where it matters most, while enabling domain autonomy where it creates the most value.

Data engineering retains full ownership of bronze, silver, and common gold zones. The canonical model remains authoritative and data quality, lineage, and the anti-corruption layer stay intact under central control. This preserves the integrity and governance that make the medallion pattern effective.

Domain teams gain the ability to create their own transformations, aggregations, and data products without queuing for central resources. They decide what to build and how to prioritise, while moving at the speed their business requires.

This separation eliminates the resource bottleneck that slows innovation while preserving the quality and governance of foundational data layers, maintaining a single source of truth. Ultimately, it eliminates the control-vs-agility dilemma.

Simplifying external integrations

APIs, file transfers, partner data sharing, and all other external integrations occur exclusively from domain gold zones, never from the core lakehouse. This isolates the core bronze, silver, and common gold zones from external exposure and partner security incidents.

In domain gold zones, data models can be tailored to each partner's needs without impacting the internal canonical model, and each integration can have its own agreed schema optimised for that use case. With partner-specific transformations isolated to a domain project, changes to one integration don't affect others—or internal consumers. It also leaves the silver zone's canonical model stable and unaffected, and multiple integrations with different schema requirements can coexist without conflict.

Removing security and blast radius concerns

Each domain project can have distinct IAM policies and access controls, with security tailored to its specific requirements. Sensitive domain data can be isolated within its own project with domain-specific security policies, addressing the granular access control challenges of monolithic architectures.

The project-level isolation creates natural containment boundaries where security incidents, data breaches, integration failures, or misconfigurations are contained within a single domain project. This means:

- Compromised credentials or misconfigured policies won't expose data across all business units or the entire core lakehouse
- Data loss is limited to domain-specific datasets, not enterprise canonical data
- Projects can be selectively exposed to external partners without risking ungoverned access to other domains
- Compliance and audit scope is reduced to specific domain projects for external partnerships

Cost attribution and accountability

Domain-specific projects now have their own compute and storage costs, with queries billed to the executing project and storage costs to the domain project. This provides clear cost visibility for each domain, with accurate attribution to the business units using them. Domain teams have transparency into resource consumption and can make informed trade-offs between analytical sophistication and cost.

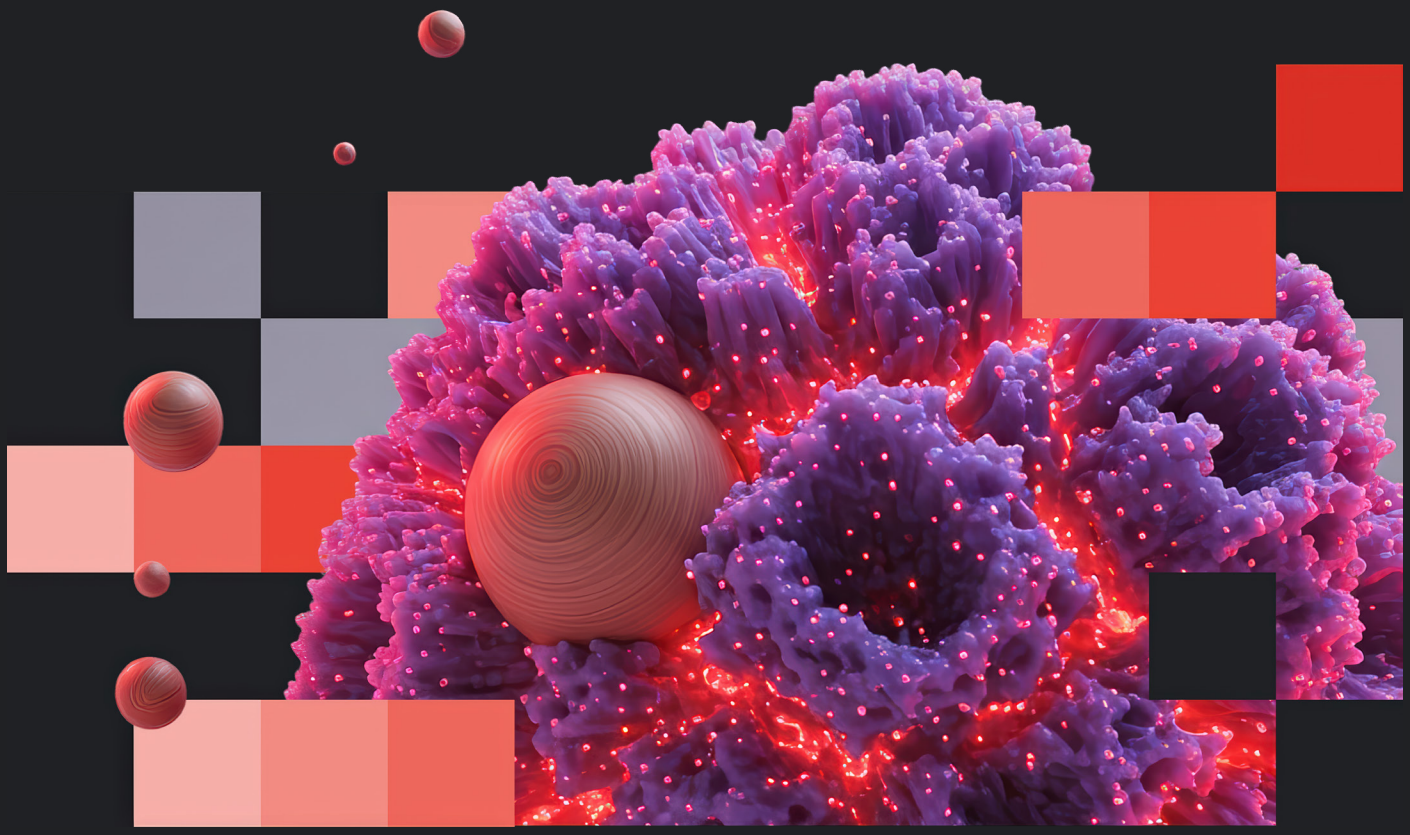
This accountability helps finance teams calculate ROI for domain-specific products, configure budget alerts per project, and establish straightforward chargeback models based on actual consumption metrics.

Eliminating compute resource contention

Domain projects use compute resources (e.g., BigQuery slots, processing capacity) allocated to those projects—not the core lakehouse project—while critical data engineering pipelines remain protected from resource contention.

This means heavy ML training, large aggregations, or exploratory queries in domain projects won't disrupt core data pipelines. Now, the performance of core transformations is predictable, data freshness SLAs can be guaranteed, and capacity planning becomes manageable with independent provisioning based on distinct requirements and priorities.

Unlocking value for every data consumer



Across the organisation, different teams interact with the data platform in distinct ways. The extended medallion architecture delivers targeted value to each group.

When providing lakehouse data to data stakeholders and teams, it is important to distinguish between user access and process access:

> **User access**
>
> This is typically an interested user (e.g., a business analyst, an engineer, or a business stakeholder) running a single query to answer a specific question.

> **Process access**
>
> This is when any kind of automation, application, or scheduled job needs to run a query, which can be anything from a fully fledged enterprise application to an automation running regularly.

This distinction is important. The architecture can only protect the central medallion zones for heavy or frequent loads if they come from a domain gold zone where the query and compute slots can be configured—removing the risk of an impact.

Ad-hoc queries (user access) will come from anywhere in the organisation to the central silver and common gold zones, and could impact their performance. **Any process access queries must come from a domain gold zone.**

Business analysts and data stakeholders

Access type: User

Challenge

Analysts waste time locating the correct datasets, understanding schemas, and validating completeness—often relying on data engineering for simple data access tasks.

Value

With data that already conforms to business rules, their focus shifts from data preparation to insight generation as they query the silver zone's canonical, clean data or the gold zone's pre-aggregated views.

Because the canonical model in the silver zone uses familiar business terminology, comprehensive documentation, and visible lineage, they can confidently build analyses without additional validation.

Data scientists

Access type: Process

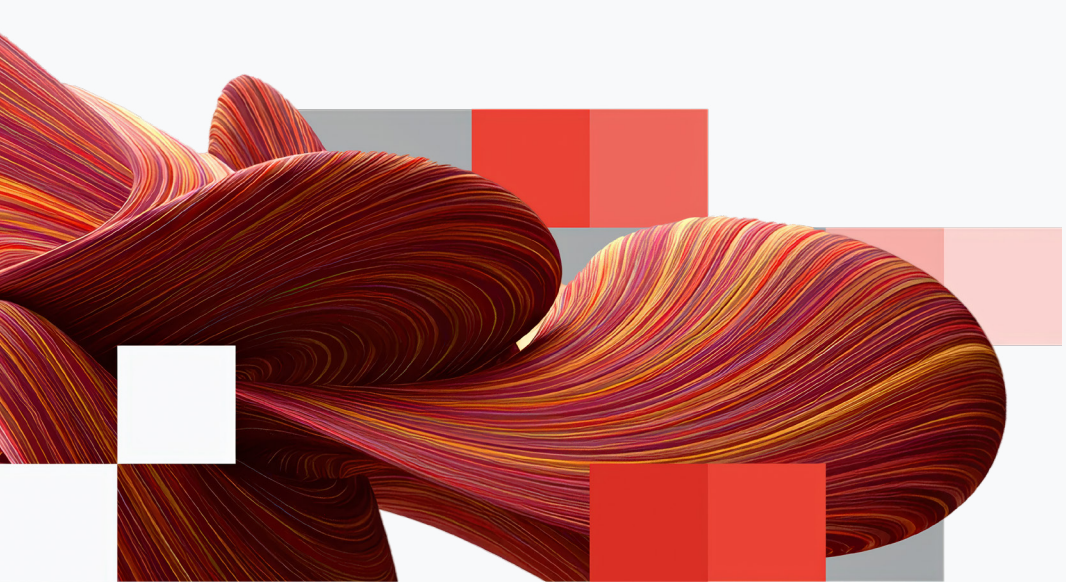
Challenge

Data scientists spend lots of time on feature engineering, cleaning raw data, harmonising source systems, and managing complex lineage for models—ultimately delaying their deployment.

Value

They can now access the silver zone's enterprise-conformed data for consistent, clean features. These are then used in domain-specific gold zones to provide pre-computed, validated features that accelerate model training and deployment.

Consistent features across the organisation allow models to be compared and combined reliably, while the silver zone's anti-corruption layer acts as a safeguard, ensuring feature definitions remain stable regardless of source system changes.



Business intelligence (BI) and reporting

Access type: Process

Challenge

Traditionally, BI dashboards suffer bottlenecks when querying complex source tables, while data inconsistency across reports erodes trust. And because report developers manually implement the business logic, there are often varying, unreliable results.

Value

BI and reporting teams can now use the high-performance common gold zone or domain-specific aggregations to ensure rapid query response times and consistent results across all reports.

The gold zone's optimised structure (clustering, partitioning, materialisation) means dashboards load quickly even with large data volumes. By implementing business logic once upstream, there is only one source of truth. And, crucially, the data is consistent across domains and data products. With self-service access to high-quality data, there's stronger innovation and collaboration across the organisation.

AI and machine learning (ML) applications

Access type: Process

Challenge

ML models are fundamentally dependent on data quality, consistency, and governance. Using raw data leads to unreliable features, hallucinations in generative models, and unsafe automated decisions.

Value

The silver zone now acts as a standardised feature store foundation, while domain-specific gold zones are ideal for domain-specific feature sets—ready for ML services with low-latency access.

For advanced applications like Retrieval-Augmented Generation (RAG), the silver zone's canonical model provides trusted, context-rich data for accurate grounding, and the standardised, well-documented data reduces hallucinations.

Thanks to the gold zone metrics serving as governed tools for AI agents, automated actions are safer and more consistent. This makes sure the tools use approved business logic rather than attempting to recalculate complex metrics from raw data.

Technical analysts

Access type: Process

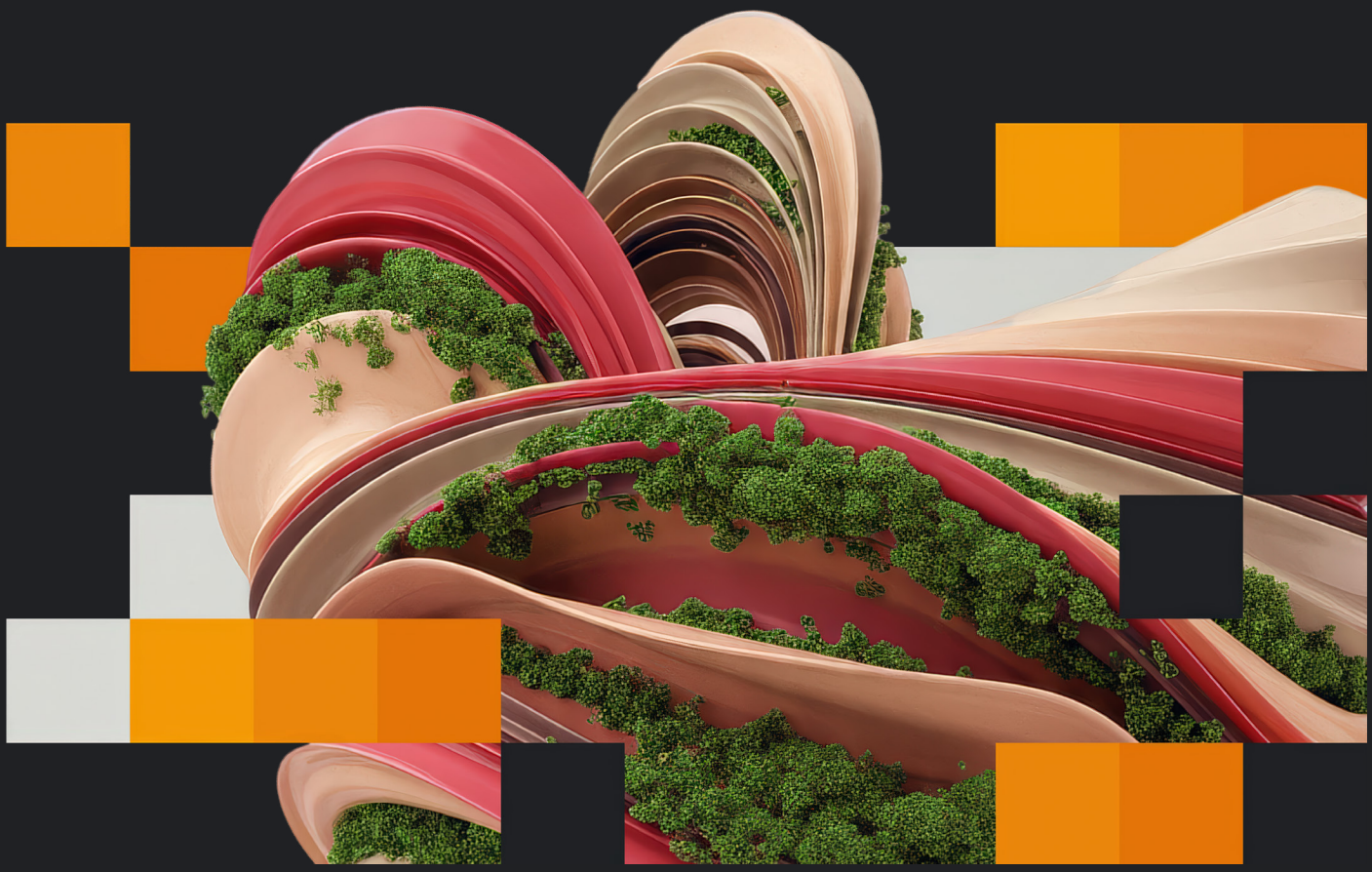
Challenge

Technical analysts spend a lot of effort reverse-engineering source systems and connecting the dots to join disparate tables. Because they lack a reliable middle ground between raw data and final reports, their custom analyses are difficult to maintain and often break when source schemas change.

Value

Instead of starting from scratch with raw data, they now have a stable foundation in the silver zone. This allows them to perform deep analysis using clean, pre-connected data that follows business logic, while still having the flexibility to build new, domain-specific insights in their own gold zone projects.

Google Cloud data foundation



While the extended medallion architecture is by design platform and hyperscaler agnostic, for the purposes of this guide, we'll show how it can be deployed on Google Cloud.

Data processing and transformation

The transformation pipelines are owned by data engineering teams and executed using the dbt framework orchestrated by Cloud Composer (Apache Airflow).

BigQuery serves as the primary data store for all medallion layers. Its serverless nature and decoupled storage and compute architecture enable the rapid, cost-effective, and scalable transformations required to move data between zones. By using BigQuery, organisations can avoid complex, expensive ETL infrastructure.

Then, to achieve high data quality, dbt enforces data quality checks, with unit and schema tests run on every model transformation. What's more, the entire transformation logic is versioned, and dbt automatically maps data lineage from bronze up to gold, fulfilling the traceability objective.

Optimising consumption and access

Consumption follows the principle of strict isolation, using Google Cloud best practices to secure data while maintaining performance.

With consumer-specific workloads hosted in dedicated Google Cloud projects (such as finance or marketing projects), the architecture supports:

> **Cost management (FinOps)**
>
> Compute costs for reading data from the core lakehouse and performing final-mile transformations are attributed to the consuming project—ensuring clear accountability.

> **Compute isolation**
>
> Heavy analytical or ML workloads use the domain project's allocated BigQuery slots. This prevents any contention with the critical data engineering pipelines running in the core lakehouse.

Access is managed through predefined roles and the principle of least privilege, and is granted primarily to federated groups (not individuals) and service accounts for automated workloads. BI tools like Looker connect to the curated gold zones, and technical analysts connect to the silver zone for deeper, ad-hoc querying—with domain-specific user projects bearing the query cost.

Data governance and discoverability

To ensure users know data exists—and trust it—the architecture enforces consistent naming and detailed descriptions across all BigQuery layers. The silver zone's canonical data model is the primary source of business understanding.

Furthermore, a central data cataloguing solution like Dataplex integrates with BigQuery to index and organise the metadata, providing search functionality, lineage visibility, and a unified governance layer across the entire platform.

Data consumption patterns and access models

The 'ready data' philosophy dictates that the consumption layer should be flexible, allowing users to select the layer of refinement (and complexity) that matches their need, while ensuring access is governed by IAM.

Function	Layers	Access	Rationale
BI and analytics (Looker and dashboards)	• Common gold zone • Domain-specific gold zone • Silver zone (ad-hoc)	BI tools (e.g., Looker, Tableau) connect directly to BigQuery datasets.	**Gold zone** Optimised for high-performance, aggregated reporting on common or specific use cases. **Silver zone** Used only for ad-hoc, deep-dive analysis when unaggregated, historical data is required.
AI and ML (ML Ops and Vertex AI)	• Silver zone • Gold zone • Domain gold zones • Unstructured data	Service accounts running Vertex AI models access data via IAM access controls.	**Silver zone** Provides clean, canonical features for model training and deployment.

Function	Layers	Access	Rationale
Reporting (financial or regulatory)	• Domain-specific gold zone (transactional) • Silver zone	Reporting projects use views on silver or gold data to generate reportable datasets. GoAnywhere is used for external file transfer.	**Domain-specific gold zone** Isolates the compute and storage of sensitive, reportable datasets, ensuring compliance with strict external transfer requirements (e.g., regulator CSV export).
API (external partner integration)	• Domain gold zone **only**	Cloud Functions or Cloud Run are used to pull data and expose it via REST services (e.g., feeding external partners like Facebook).	**Gold zone** Used for aggregated, business-ready data required by partners. **Silver zone** Used when entity-level, canonical data is required for external systems.
Data analysis (technical analysts)	• Silver zone • Common gold zone • Domain gold zone	Analysts connect directly to BigQuery via the console, running their queries within their **domain-specific user projects.**	**Silver zone** Provides the clean, understandable canonical data necessary for reliable exploratory analysis without needing to bypass the anti-corruption layer.

Summary

The extended medallion architecture is a pragmatic evolution of the standard medallion pattern—solving the tension between centralised governance and domain agility.

By extending the three-zone model with domain-specific gold zone projects, there are no more data bottlenecks.

While this architecture is designed to be platform agnostic, organisations using this pattern on Google Cloud (as is IG Group) can access native capabilities like BigQuery's zero-copy architecture, IAM's fine-grained access control, and project-based resource isolation—ultimately implementing the pattern with minimal operational overhead.

The result is a data platform that truly serves the business. Central teams maintain the canonical model and enterprise data quality, while domain teams move at the speed of business opportunity. Data becomes 'ready data'—discoverable, accessible, trustworthy, and usable. No longer just for a privileged few, this data now benefits the entire organisation.

With thoughtful design, the extended medallion architecture shows that modern cloud platforms can support both rigorous data governance and organisational agility. It's the new blueprint for the data-driven enterprise.

About IG Group Holdings

IG Group Holdings plc ("IG") is a FTSE 100 financial technology company operating at the intersection of retail trading, technology and capital markets. Through its trusted brands—IG, tastytrade, Freetrade and Independent Reserve—the Group serves over 1.3 million customers worldwide, providing leveraged trading, stock trading and investments, and cryptocurrency trading via its proprietary platforms.

Ready to build your data-driven enterprise?

Contact us