pemo

Data Strategy Document - Detailed

by **Mehmet Akif Kucuk**

2025-06-12

Thursday

In a nutshell; **3**

Current Data Structure of pemo **4**

Proposed High-Level Data Structure for pemo **5**

Proposed Detailed Data Structure for pemo **6**

1. Data Governance Foundation **8**

1.1. Data Inventory and Asset Management (Table-Level) - Data Catalog L1 8

1.2. Data Classification and Sensitivity Mapping (Column-Level) Data Catalog L2 8

2. Compliance and Regional Requirements **9**

2.1. Data Residency Policy 9

2.2. Data Retention and Lifecycle Management 9

3. Access Control Security **10**

3.1 Role-Based Access Management 10

4. Data Architecture and Performance **10**

4.1. Database Optimization 10

4.2. Data Pipeline Architecture 11

4.3. Curated Data and User Access Optimization 11

5. Data Visualization **11**

5.1. Data Visualization 11

6. Data Quality and Monitoring **12**

6.1. Data Quality and Monitoring 12

7. External Integrations and Partnerships **12**

7.1. Third-Party Data Sharing Controls 12

8. Data Backup and Disaster Recovery **13**

9. Monitoring Success Metrics and KPIs **14**

In a nutshell;

Risk Mitigation: Eliminates potential millions in UAE Data Regulatory/SAMA fines and prevents data breaches

Revenue Growth: Enables $2-5M additional revenue through better data utilization and faster decision-making

Cost Optimization: Reduces data-related costs by 30-50% through precalculation, automation and efficiency

Curated Data: Maintains curated data by following medallion structure to drive faster, accurate and consistent decisions

Competitive Advantage: Provides real-time insights with <1 hour time-to-insight for critical business questions

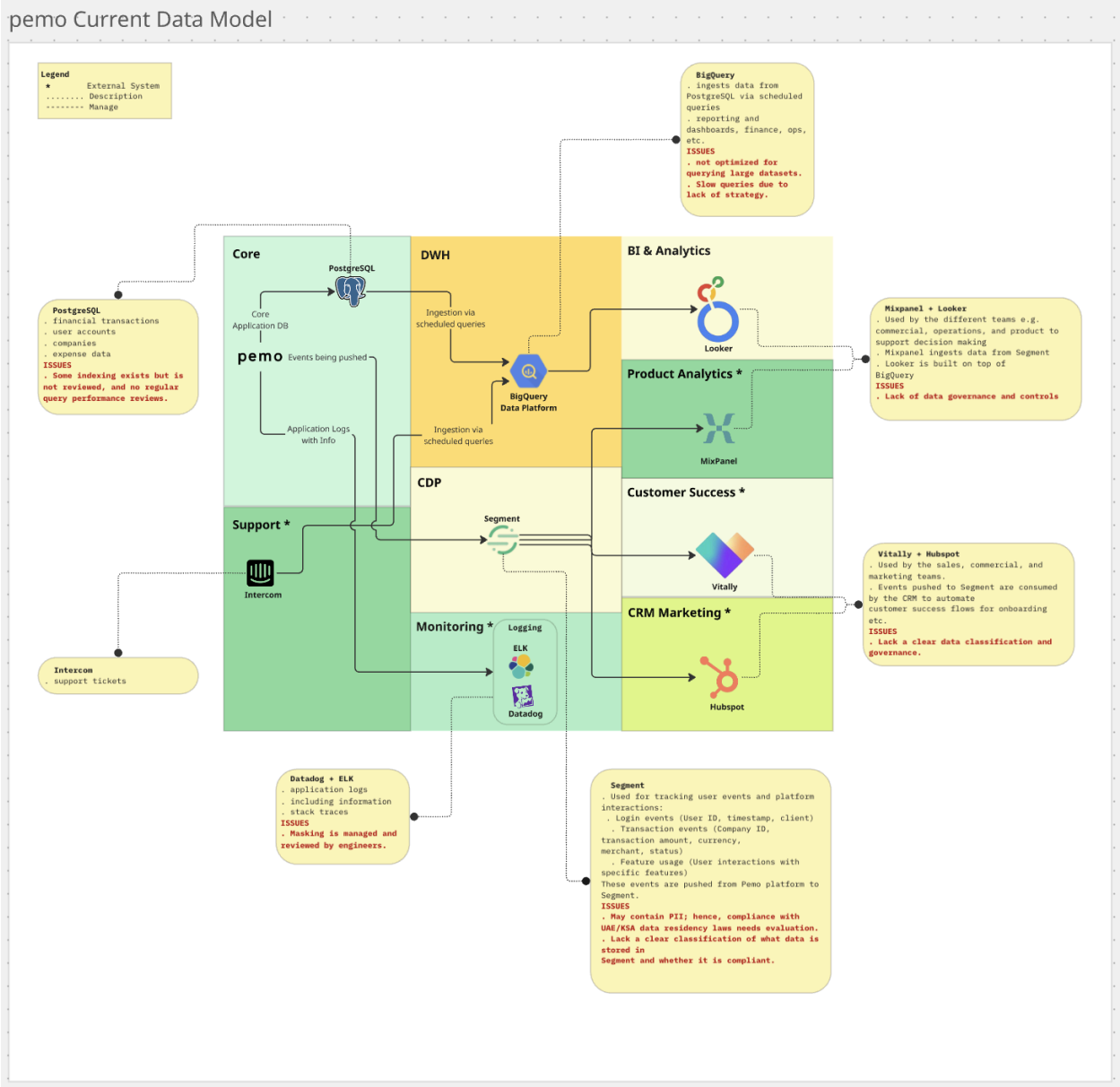
Business Continuity: Ensures <4 hour recovery time during disasters, protecting against extended downtime

Customer Trust: Demonstrates strong data protection, essential for customer retention and growth

Current Data Structure of pemo

pemo Current Data Model Frame

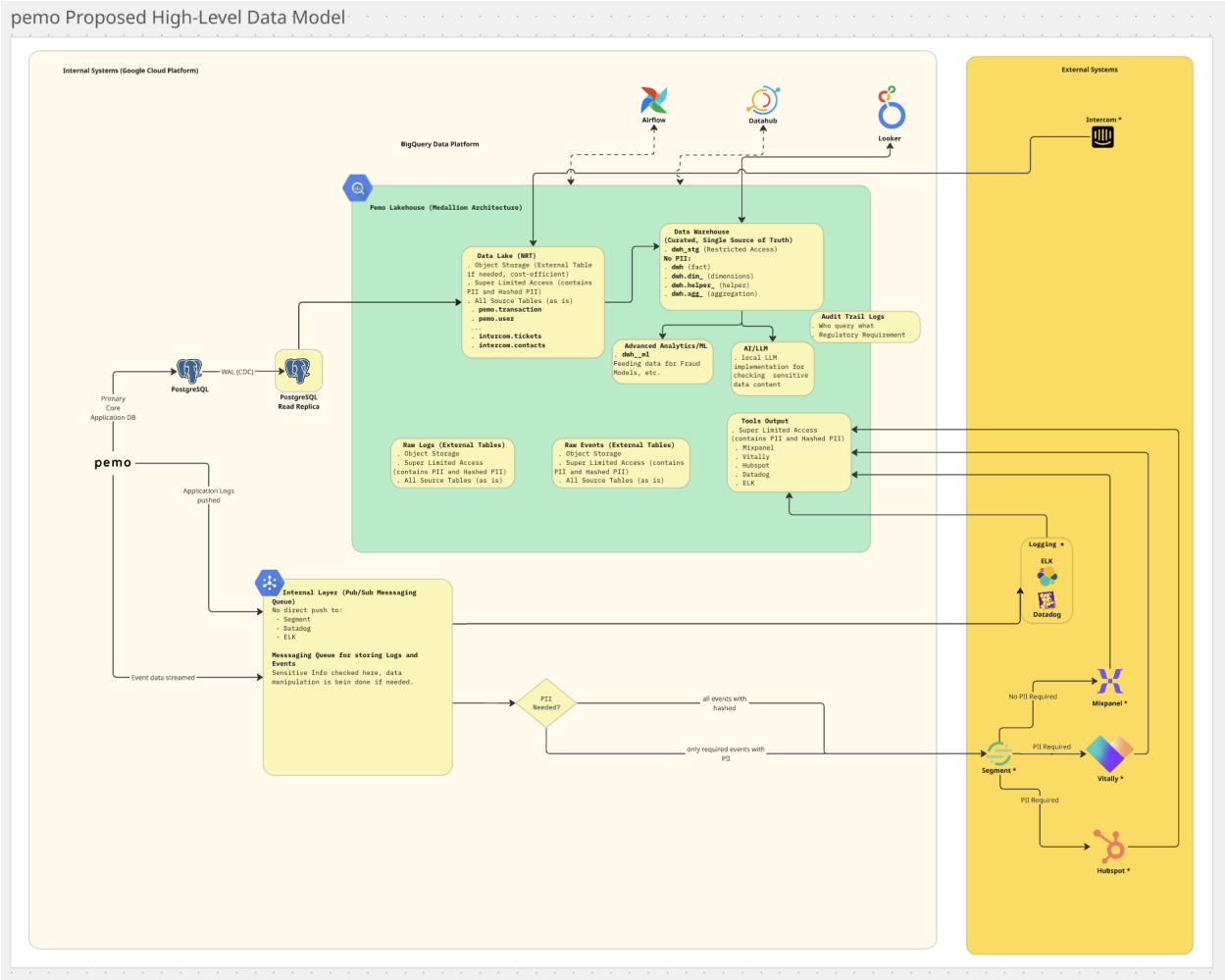
https://miro.com/app/board/uXjVIr6Vhbw=/?share_link_id=264565662366



Proposed High-Level Data Structure for pemo

pemo Proposed High Level Data Model Frame

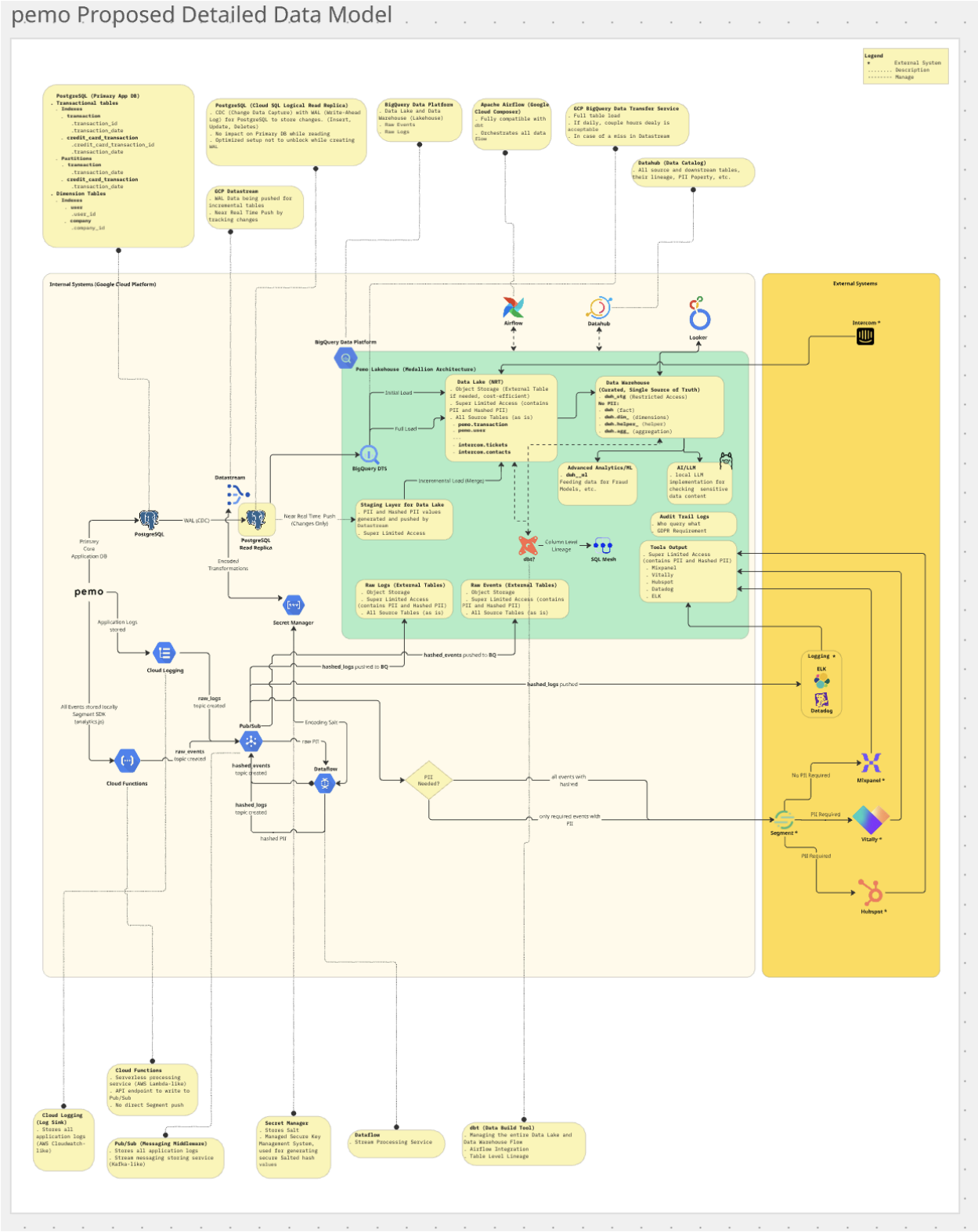
https://miro.com/app/board/uXjVIr6Vhbw=/?share_link_id=264565662366



Proposed Detailed Data Structure for pemo

pemo Proposed Detailed Data Model Frame

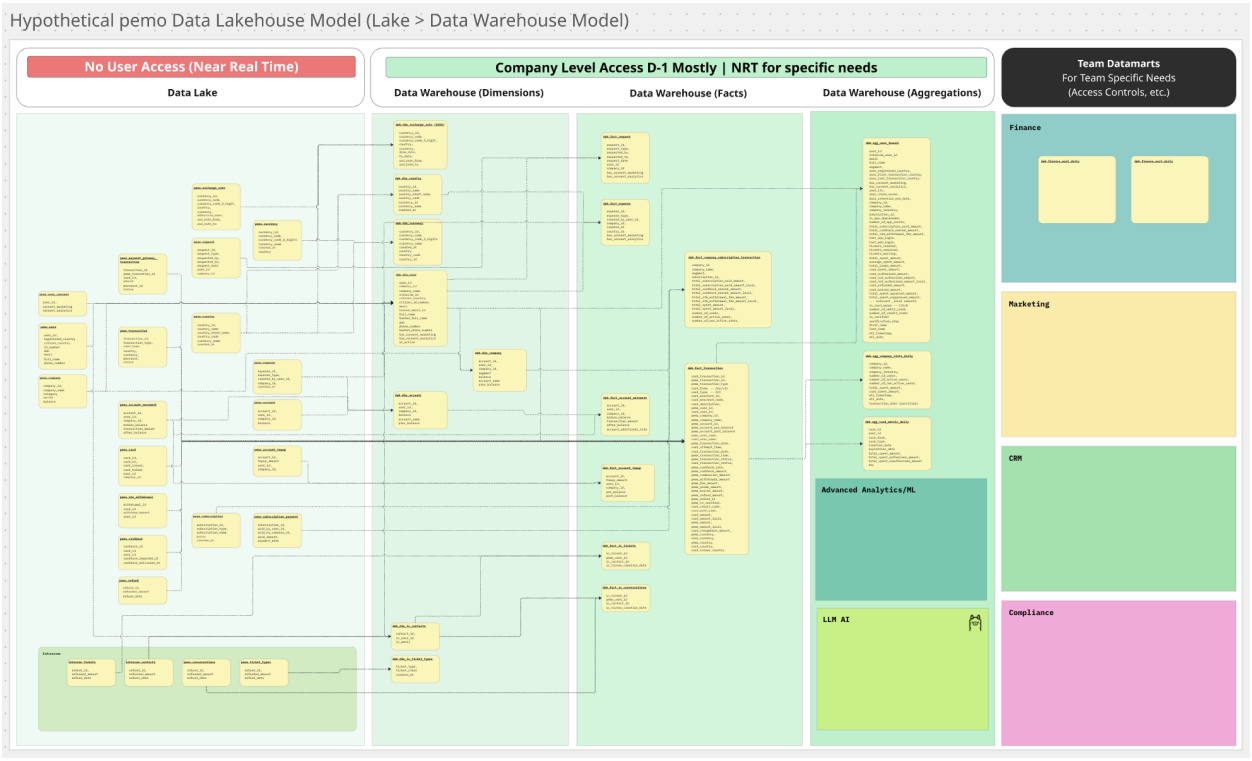
https://miro.com/app/board/uXjVIr6Vhbw=/?share_link_id=264565662366



Hypothetical pemo Data Lakehouse Model

Hypothetical pemo Data Lakehouse Model (Lake > Data Warehouse Model) Frame

https://miro.com/app/board/uXjVIr6Vhbw=/?share_link_id=264565662366



1. Data Governance Foundation

1.1. Data Inventory and Asset Management (Table-Level) - Data Catalog L1

Objective: Provide clear answers to "Where does data come from, where does it go, and how is it transferred?" This enables proactive risk identification.

Source System	Src Table/Stream	Target System	Target Table/Stream	Migration Method	Frequency	Data Residency	Target Retention	Notes
PostgreSQL	users	BigQuery	dataset.users	Scheduled Query	Daily	me-central2	5 years	ID No, email
Event Stream	payments	BigQuery	dataset.payments	Datastream	Saatlik	me-central2	3 years	Card no
Log System	app_logs	Datadog	-	Cloud Function	Real Time	-	90 days	user, ip
DWH	transaction, user company	DWH	agg_company_stats_daily	Transform ETL	NRT every 10m	me-central2	No Retention	Precalculated, very fast result

Current State	No central inventory and catalog exists
Business Impact	- Enables identification of risky data flows across all internal and external systems - Simplifies compliance audits - Protects company from fines and builds customer trust
Cost Impact	- Requires initial time and resource investment - However, potential fines and reputation damage from data loss or breaches are significantly higher
Technical Notes	*Utilize dbt and SQL Mesh for source-target mapping*

1.2. Data Classification and Sensitivity Mapping (Column-Level) Data Catalog L2

Objective: Identify sensitive information within data and ensure adequate protection measures.

Table/Stream	Column Name	Description	Sensitivity Level	DLP Found	Manipulation	Data Residency	Retention	Notes
dataset.users	customer_id	ID Number	PII	Yes	Hash	europe-west1	5 years	Not hashed currently
dataset.events	card_number	Credit Card No	Confidential	Yes	Custom	europe-west1	5 years	Already custom manipulated
app_logs	ip_address	IP Address	Confidential	No	Masking	-	90 days	Already masked to Segment

Current State	No central column-level catalog
Business Impact	- Ensures compliance with UAE & SAMA regulations - Protects brand security - Reduces potential lawsuit and penalty risks
Cost Impact	Late discovery of violations can result in very high penalties

Technical Notes	Google DLP, Tokenization, Masking, Hashing via Dataflow

2. Compliance and Regional Requirements

2.1. Data Residency Policy

Objective: Ensure data remains within required geographical boundaries per local laws.

Maintain and follow the rules defined here:

Region	Data Types	Storage Location	Transfer Rules	Local Requirements
EU	All Personal Data	europe-west1	No US transfer	GDPR Article 44-49
UAE	Customer Data	me-central2	Local processing	UAE Data Protection Law
KSA	Financial Data	Local only	No cloud export	SAMA Cyber Security Framework

Current State	Core applications are aligned with data residency requirements
Business Impact	Prevents business shutdowns due to non-compliance
Cost Impact	Additional regional infrastructure costs
Technical Notes	Explore GCP's multi-region solutions with strong compliance alignment

2.2. Data Retention and Lifecycle Management

Objective: Define how long data should be stored and subsequent actions, considering Compliance as well.

*Check retention period in table and column level mapping **(1.1, 1.2)** and apply deletion:*

Current State	No central data retention system
Business Impact	- Eliminates unnecessary data burden - Ensures compliance (e.g., "emails must be stored for 5 years") - Reduces risk of old data breaches
Cost Impact	- Archived data reduces storage costs - Manual SQL deletion is risky → Automation essential
Technical Notes	- BigQuery TTL (Time-to-Live) configuration - "Soft Delete + Retention Policy" → archive first, permanent delete later

3. Access Control Security

3.1 Role-Based Access Management

Objective: Control who accesses which data at what level, minimizing unnecessary access.

Role	Dataset	Access Level	Notes
Compliance Officer	dwh.raw_events	Full Access	Original PII visible
Data Engineer	dwh.raw_events	Masked	Masked PII visible
Analyst	dwh.raw_events	Anonymized	Unencodable Data Visible
UAE Analyst	dwh.raw_events	UAE-Only	Regional Records Only
Marketing Analyst	segment.event_metrics	Marketing	Role to show who has access to external systems as well

Current State	Poor level of access controls
Business Impact	- Only necessary personnel access data → Enhanced security - Simplified audit processes
Cost Impact	Access confusion leads to data breach risks
Technical Notes	- BigQuery IAM Management with on-the-fly hashing (5-10% CPU cost) - Column and Row Level Access Controls

4. Data Architecture and Performance

4.1. Database Optimization

Objective: Achieve faster, accurate results at lower cost through precalculation strategies.

Current State	Unoptimized databases, queries
Business Impact	Quick, accurate, and consistent data access
Cost Impact	Increased costs due to precalculation processes and storage
Technical Notes	Partitioning, indexing frequently used columns, materialized views, clustering, and merge operations

4.2. Data Pipeline Architecture

Objective: Improve current data transfer methods using Near Real-Time ingestion with Change Data Capture technology.

Near Real Time Data ingestion by using the power of Change Data Capture technology

Current State	No real time ingestion flow
Business Impact	- Current and real-time data → Better decision making - Enhanced performance and reporting quality
Cost Impact	- Near real-time structures are more expensive and complex - However, delayed data also delays decisions
Technical Notes	- CDC → EventStream → BigQuery (external or internal tables) - GCP Dataflow for real-time pipeline setup - Asynchronous flow separation as scale increases

4.3. Curated Data and User Access Optimization

Objective: Enable end users to access data quickly and accurately.

Warehouse layer objects may run frequently based on time criticality (costly but necessary)

No user access to initial ingestion layer; access limited to Data Warehouse layer and beyond

Current State	No data management in Data Warehouse level
Business Impact	- Golden-level curated data is being flagged and accessed - Enhanced performance and reporting quality
Cost Impact	Some near-real-time data causes cost but also brings faster decision making
Technical Notes	dbt → BigQuery Transformation Management

4.4. Data Generation and Push for Advanced Analytics/Machine Learning

Objective: Empower fraud detection systems with rapid and accurate data access for real-time analytics and machine learning model training.

Warehouse layer objects are scheduled to run frequently to support near-real-time fraud detection

Current State	Limited data orchestration at the Data Warehouse level for fraud-specific use cases.
Business Impact	- Curated, high-quality fraud detection data enables faster identification of suspicious activities. - Improved model accuracy enhances fraud prevention, reducing financial losses.

Cost Impact	Investment in optimized pipelines may yield long-term cost efficiencies.
Technical Notes	Data pipelines feed structured data into machine learning models (e.g., anomaly detection algorithms).

4.5. LLM Implementations

Objective: Leverage Large Language Models (LLMs) to securely explore and extract insights from sensitive data while maintaining compliance and data privacy.

Warehouse layer objects are designed to anonymize and preprocess sensitive data before feeding into LLMs, with strict access controls.

Current State	Limited automation for anonymization and preprocessing of sensitive datasets.
Business Impact	Enhanced data privacy measures build trust with stakeholders and ensure regulatory compliance.
Cost Impact	Processing sensitive data with LLMs incurs higher computational costs due to anonymization and encryption requirements.
Technical Notes	Use batch processing for non-time-critical tasks to balance cost and performance.

5. Data Visualization

5.1. Data Visualization

Objective: Ensure reporting and visualization uses curated data

Curated sources will be mapped in Data Catalog to be used by visualization tool, this layer will be clean, heavy aggregations to be tracked

Current State	Querying BigQuery via Looker
Business Impact	Access beautiful insights with a curated data on the back-end
Cost Impact	Cost will be reduced after proper reporting catalog
Technical Notes	Looker might connect other layers, OLAP layer if needed

6. Data Quality and Monitoring

6.1. Data Quality and Monitoring

Objective: Ensure data accuracy, completeness, and consistency across all systems.

Quality Dimension	Measure	Target	Monitoring	Alert Threshold
Completeness	% of non-null values	>95%	Daily	<90%
Timeliness	Data freshness	<15mins	Continuous	<30mins
Consistency	Source Target Match	100%	Daily	>99%

Current State	No data quality checks and monitoring being implemented
Business Impact	Reduce time spent fixing data issues
Cost Impact	Minimal implementation cost for checks
Technical Notes	Python, Spark Framework or SQL with dashboard visualization

7. External Integrations and Partnerships

7.1. Third-Party Data Sharing Controls

Objective: Monitor and control what data goes to external systems.

External System	Data Shared	Sharing ID	Data Content Controlled by	Sharing Approved by	Shared Time
segment	User Events	share_sgm_01	Mehmet Kucuk	Compliance Officer	2025-09-01 12:30
datadog	Application Log	share_dd_01	Akif Kucuk	Engineering Lead	2025-09-01 13:30

- **Additional Internal layer** before sending, check whether the content has PII by Sensitive Data Exploration, content control person recorded, an, no direct push to external systems

Current State	No sensitive data classification checks, especially for logs and events
Business Impact	- Prevents PII data from going to wrong destinations - Provides audit comfort
Cost Impact	- Internal layers can be costly - However, sensitive data exposure creates reputation and financial loss
Technical Notes	- Internal Layer for Logs and Events - Log flow: Cloud Logging → Pub/Sub → Dataflow → Transformed Pub/Sub → Parquet → BigQuery → External Tools - Event flow: Cloud Function → Internal Endpoint → Pub/Sub → Dataflow → Hashed Pub/Sub → BigQuery - Log all external data transfers with timestamps - Use only anonymous IDs (hashed + salted) for external data - Store external data temporarily for audit queries

8. Data Backup and Disaster Recovery

Objective: Ensure critical data is securely backed up and quickly restored during failures, while meeting compliance requirements.

System	Asset	Backup Frequency	Storage Location	Recovery Time Obj (RTO)	Compliance Notes
PostgreSQL	pemo.user	Daily full, hourly incremental	me-central2, europe-west1	<4h	SAMA, GDPR
bigquery	dwh.transaction	Daily snapshot	share_dd_01	<24h	UAE Data Protection, SAMA

Current State	No centralized backup or disaster recovery plan documented
Business Impact	- Ensures business continuity during outages, minimizing downtime and financial loss - Builds customer trust by protecting sensitive data
Cost Impact	- Initial setup and ongoing storage costs - Avoids much higher costs from data loss, regulatory fines, or reputation damage
Technical Notes	Follow GCP best practices including Cloud SQL Automated Backups and BigQuery Auto Snapshots

9. Monitoring Success Metrics and KPIs

Data Access Metrics	- Who accessed which tables - Mostly queried tables - Mostly querying user - Who tried to access restricted tables systems
Operational Metrics	- Data Pipeline Uptime: >99.5% - Query Performance: <3 second average response - Data Quality Score: >95% - Compliance Audit Results: Zero findings
Business Metrics	- Time to Insight: <1 hour for critical business questions - Decision Making Speed: 50% faster data-driven decisions - Revenue Impact: $2-5M additional revenue from better data usage - Cost Optimization: 30-50% reduction in data-related costs
Risk Metrics	- Security Incidents: Zero data breaches - Compliance Violations: Zero regulatory fines - Data Recovery Time: <4 hours for critical data - Vendor SLA Compliance: >99% across all vendors