Data Architecture & Governance  
KOMEZA BANK – Churn Intelligence Platform for Bank of Kigali  
Version: 0.1  
Date: October 2025

**Document Control and Information**

| **Name** | **Title** | **Action** |
| --- | --- | --- |
| Rose NGABIRE | Chief Operations Officer | Approve |
| Nicholas MURIMI | Chief Product Development Officer | Review |
| Samuel MUTINDA | Head, Data & Analytics | Review |
| Steven SHYAKA | Machine Learning Engineer | Compile |

**Revision History**

| **Version** | **Author** | **Summary of Change** |
| --- | --- | --- |
| 0.1 | Steven SHYAKA | Initial draft aligned to KOMEZA BRS & SRS |

**Distribution List**

| **Name** | **Area** | **Action** |
| --- | --- | --- |
| Samuel MUTINDA | Data Science and Analytics | Review |
| Herve TWAHIRWA | Data Science and Analytics | Solution Development |
| Alphonse MUHAYIMANA | Business Applications | CRM Integration |
| Sandra UWADEDE | Business Applications | CRM Integration |
| Francine UMULISA | Customer Experience | Comment/Noting |
| Compliance Team | Legal & Risk | Compliance Review |

**Glossary**

| **Term/Abbreviation** | **Definition** |
| --- | --- |
| PII | Personally Identifiable Information |
| ETL | Extract, Transform, Load |
| CDC | Change Data Capture |
| DQ | Data Quality |
| DLP | Data Loss Prevention |
| CLV | Customer Lifetime Value |
| ROI | Return on Investment |
| Data Catalog | Centralized metadata registry describing datasets |

**1. Purpose**

This document defines the **Data Architecture and Governance** standards for the KOMEZA BANK Churn Intelligence Platform built for Bank of Kigali (BK). It specifies data sources, schemas, lineage, quality rules, anonymization and PII handling, retention and archival policies, roles & responsibilities, and operational controls required to ensure secure, compliant, and trustworthy use of data for predictive analytics.

**2. Principles**

1. **Compliance-first:** All data handling complies with Rwanda’s Law No. 058/2021 on Data Protection and BK internal policies.
2. **Least-privilege access:** Users get only the access necessary for their role.
3. **Auditability:** All data access and transformations are logged and traceable.
4. **Data quality by design:** Validation and monitoring are embedded into ingestion and transformation pipelines.
5. **Privacy-preserving analytics:** PII is anonymized or tokenized prior to model training and non-production environments.
6. **Single source of truth:** Curated datasets live in the data warehouse/feature store and are discoverable in the data catalog.

**3. Data Sources (Inventory)**

| **Source** | **Owner** | **Primary Data** | **Access Method** |
| --- | --- | --- | --- |
| T24 Core Banking | Core Banking Team | Customer profiles, accounts, balances, transactions | JDBC / Secure API |
| CRM | Business Applications | Customer interactions, complaints, opportunities | API / DB Extract |
| Digital App Logs | Digital Banking Team | Mobile app sessions, logins, transactions | Event export (Kafka/S3) / API |
| USSD Platform | Digital Banking Team | USSD transactions and sessions | Batch export / API |
| Call Center | Contact Center | Calls, call outcomes, recordings metadata | API / CSV export |
| Loan Management | Credit Team | Loan applications, approvals, repayment status | DB / API |
| Surveys & NPS | CX Team | NPS scores, survey responses | API / CSV |
| External (e.g., credit bureau) | Risk & Compliance | Credit scores, bureau reports | Secure API (on request) |

**4. Logical Data Architecture**

Layers:

1. **Ingestion Layer:** Connectors (JDBC, API, SFTP, CDC) bring data into staging.
2. **Staging Zone (Secure):** Short-lived raw extracts; PII masked here where possible.
3. **Raw Data Lake:** Immutable object store with time-stamped partitions for auditability.
4. **Curated Data Warehouse / Feature Store:** Cleaned, normalized tables and feature views for ML & reporting.
5. **Model Output Store:** Scored results, SHAP explainability artifacts, and recommendations.
6. **Serving Layer:** APIs and exports to CRM and dashboards.

Data Flow Summary:

* Source → Ingestion → Staging (validation, masking) → Raw Lake → ETL → Curated Warehouse / Feature Store → Model Training → Model Registry → Scoring → Serving → CRM / Dashboard

**5. Data Schema & Sample Mapping**

**5.1 Canonical Customer Table (Curated)**

| **Field Name** | **Description** | **Type** | **PII?** | **Notes** |
| --- | --- | --- | --- | --- |
| customer\_id | Internal anonymized ID | STRING | No | Derived from Customer\_ID hashed with salt |
| external\_customer\_ref | Bank account number (tokenized) | STRING | Yes (tokenized) | Tokenized for reversible lookup by privileged services |
| segment | Retail/SME/Corporate | STRING | No | Source: T24 |
| account\_status | Active/Dormant/Closed | STRING | No | Source: T24 |
| branch\_id | Branch code | STRING | No | Source: T24 |
| account\_open\_date | Date | DATE | No | Source: T24 |
| last\_transaction\_date | Date | DATE | No | Derived field |
| avg\_balance\_3mo | Numeric | FLOAT | No | Derived feature |
| txn\_count\_3mo | Integer | INT | No | Derived feature |
| digital\_login\_freq\_3mo | Integer | INT | No | Derived from app logs |
| ussd\_usage\_3mo | Integer | INT | No | Derived from USSD platform |
| complaints\_count\_6mo | Integer | INT | No | CRM |
| loan\_outstanding\_amount | Numeric | FLOAT | No | Loan system |
| nps\_score | Integer (0-10) | INT | No | Surveys |
| churn\_flag | Binary (1=churned) | INT | No | Label derived for training |

**5.2 Feature Store Table Example**

* feature\_customer\_snapshot\_v{date}: time-stamped table with computed features used for model training/scoring.

**6. Data Lineage & Provenance**

* Every curated dataset includes metadata fields: source\_system, extraction\_time, etl\_job\_id, transform\_version, and data\_owner.
* Lineage captured in data catalog and in ETL orchestration logs (Airflow run IDs) to support traceability.
* Model training records include pointer to feature snapshot versions and dataset hashes.

**7. PII, Anonymization & Tokenization Strategy**

**7.1 PII Classification**

* **Direct PII:** Names, national ID, phone numbers, email, account numbers.
* **Indirect PII / Quasi-identifiers:** Birthdate, address, transaction timestamps, device IDs.

**7.2 Handling Rules**

1. **Minimize PII Use:** Avoid storing direct PII in model training datasets; use anonymized customer\_id instead.
2. **Hashing:** Use strong salted hashing (e.g., HMAC-SHA256 with KMS-managed salt) for irreversible identifiers where reversible mapping is not required.
3. **Tokenization:** For cases requiring reversible mapping (e.g., customer support), use token service with strict RBAC and audit logs.
4. **Masking:** Mask partial fields in dashboards (e.g., show only last 4 digits of account number).
5. **Removal in Non-Prod:** Strip or synthetic-generate PII for development and testing environments.

**8. Data Quality Framework**

**8.1 DQ Dimensions**

* **Completeness:** Required fields must be present (e.g., customer\_id, transaction\_date).
* **Accuracy:** Numeric ranges and pattern checks (e.g., valid account formats).
* **Timeliness:** Data freshness targets (e.g., daily ingestion for transaction data).
* **Uniqueness:** No duplicate customer\_id entries in canonical table.
* **Consistency:** Field value consistency across sources (e.g., segment values aligned).

**8.2 Validation Rules (Examples)**

* customer\_id not null and unique per snapshot.
* last\_transaction\_date <= extraction\_date.
* avg\_balance\_3mo >= 0.
* Null rate alert: raise if nulls > 10% on critical fields.
* Schema drift detection: alert if extra/missing columns are present.

**8.3 Tools & Implementation**

* Implement DQ checks in ETL using Great Expectations or custom validation steps in Airflow.
* DQ metrics written to monitoring dashboard and alerted via Slack/PagerDuty for thresholds exceeded.

**9. Data Catalog & Metadata Management**

* Maintain a central **Data Catalog** (e.g., Amundsen, DataHub, or internal wiki) containing dataset descriptions, owners, freshness, lineage, and sensitivity tags.
* Metadata captured: schema, sample rows, tags (PII, Sensitive), retention policy, contact person.
* Catalog must be searchable by business users and data engineers.

**10. Access Control & Roles**

**10.1 Roles**

* **Data Steward:** Responsible for data quality, metadata, and owner of dataset definitions.
* **Data Engineer:** Builds ETL pipelines and maintains ingestion jobs.
* **Data Scientist:** Uses curated datasets and feature store for model development.
* **BI Analyst:** Builds dashboards and reports.
* **System Admin / DevOps:** Manages infra, secrets, and deployment.
* **Compliance Officer:** Reviews data usage and approves sensitive data access.

**10.2 Access Policies**

* RBAC enforced via BK identity provider (AD/Okta).
* Least privilege by default; access granted through request/approval workflow.
* Time-bound access for sensitive datasets; approvals logged.
* Production data access requires MFA and just-in-time access where feasible.

**11. Data Retention & Archival Policy**

* **Raw Data Lake:** Keep raw extracts for **2 years** in hot storage, then archive to cold storage for **3 additional years** (total 5 years) unless legal holds apply.
* **Curated Data Warehouse / Feature Store:** Keep 3 years of time-stamped feature snapshots online; older snapshots archived for 2 years.
* **Model Artifacts & Training Data Hashes:** Retain indefinitely or per BK policy; at minimum, store last 10 production model artifacts with training snapshot references.
* **PII Retention:** PII retained only as necessary for business operations and legal compliance; purge requests handled according to BK DPO (Data Protection Officer).

**12. Data Security & Encryption**

* **At Rest:** Use AES-256 encryption for object storage and database volumes.
* **In Transit:** TLS 1.2+ for all API and DB connections.
* **Key Management:** Use cloud KMS or HSM for key storage; rotate keys per BK policy.
* **DLP:** Configure DLP rules to flag exports containing PII and enforce review workflows.

**13. Data Sharing & Third-Party Integrations**

* Data sharing outside BK requires a Data Sharing Agreement (DSA) approved by Legal & Compliance.
* External vendors given least privilege and subject to security assessments.
* All third-party transfers must be encrypted and logged.

**14. Monitoring, Observability & Alerts**

* **Data Pipeline Health:** Track ETL success rates, run durations, row counts, and schema changes.
* **DQ Metrics:** Dashboards for completeness, null rates, and drift; alerts for threshold breaches.
* **Model Data Drift:** Monitor feature distribution drift, population stability index (PSI), and prediction distribution shifts.
* **Anomaly Detection:** Automated alerts for unusual spikes/dips in transaction volumes.

**15. Data Governance Processes**

**15.1 Onboarding a New Data Source**

1. Submit Data Onboarding Request to Data Steward.
2. Data Steward performs data profiling and sensitivity classification.
3. Build/approve ingestion connector and DQ rules.
4. Add dataset to Data Catalog with owner and retention policy.

**15.2 Data Access Request**

1. User submits access request via access management portal.
2. Data Steward and Compliance review and approve/reject.
3. If approved, granting is time-boxed and logged.

**15.3 Data Incident Handling**

* Incident response runbook to contain steps for containment, assessment, notification (DPO), and remediation.
* Breaches reported per BK policy and regulatory requirements.

**16. Compliance & Legal Considerations**

* All processing adheres to Rwanda Law No. 058/2021 on Data Protection and Privacy.
* Maintain records of processing activities (RoPA) for audit readiness.
* Support data subject rights (access, rectification, erasure) through DPO and operational workflows.

**17. Feature Engineering Governance**

* Feature definitions must be versioned and stored in the Feature Registry.
* Each feature includes: definition, SQL/formula, owner, freshness schedule, and lineage to raw columns.
* Changes to feature definitions require review and sign-off (Data Steward + Data Scientist).

**18. Model Training Data Governance**

* Training datasets must reference snapshot versions and include dataset hashes for reproducibility.
* Training on PII requires documented justification and Compliance sign-off.
* Feedback from CRM/Telesales stored and used for retraining with provenance.

**19. Audit & Reporting**

* Quarterly data governance reports covering DQ trends, access reviews, incidents, and compliance checks.
* Annual external audit (if required) and internal audits as part of BK IT assurance calendar.

**20. Roadmap & Next Steps**

1. Implement Data Catalog and register all curated datasets.
2. Deploy initial DQ checks in ETL pipelines and dashboard DQ metrics.
3. Implement PII tokenization service and secrets management.
4. Define SLA and runbooks for data incidents and onboarding.
5. Conduct governance training for data consumers and stewards.

**Appendix A: Sample Feature Definitions**

* avg\_balance\_3mo = rolling average of end-of-day balances over last 90 days.
* txn\_count\_3mo = count of customer-initiated transactions in last 90 days.
* days\_since\_last\_txn = difference in days between scoring date and last transaction date.
* digital\_login\_freq\_3mo = count of unique login sessions in last 90 days.

**Appendix B: Contacts**

* **Data Steward:** Samuel MUTINDA (Head, Data & Analytics)
* **Compliance / DPO:** [To be filled]
* **Platform Owner:** Steven SHYAKA (Machine Learning Engineer)

**Conclusion**

This Data Architecture & Governance document establishes the standards and operational controls necessary to make the KOMEZA BANK Churn Intelligence Platform trustworthy, compliant, and production-ready for Bank of Kigali. It is designed to evolve as the platform matures and more channels (BK Agent, mobile money) are onboarded.