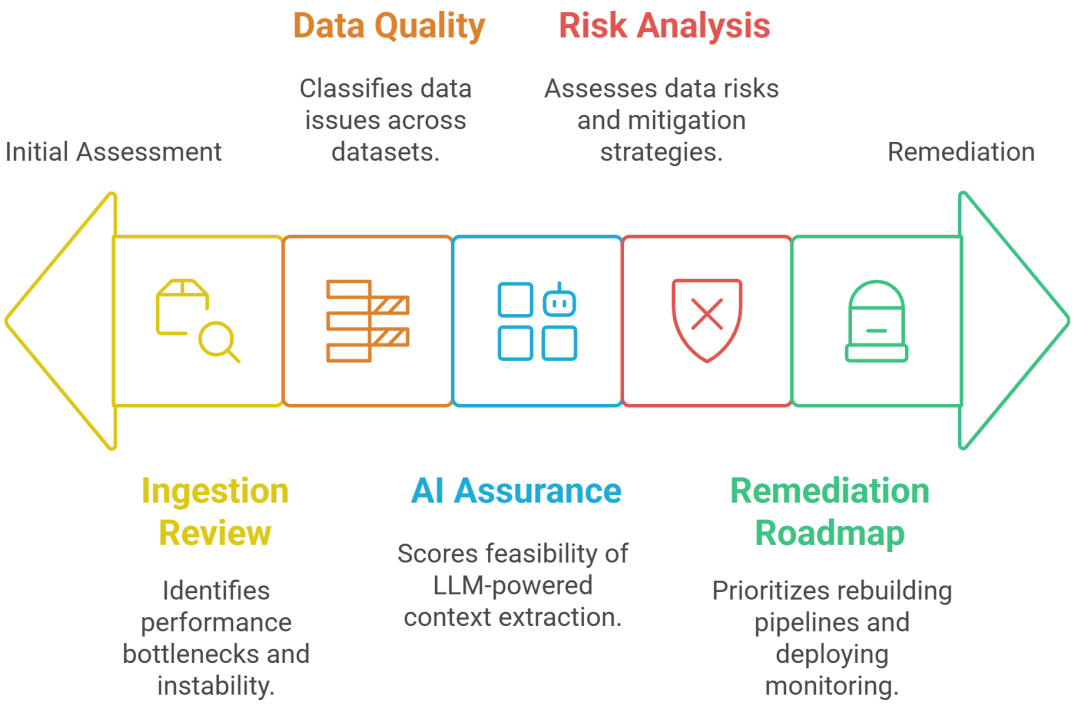
Data Quality & AI Assurance Report

Prepared for: **Stonecliff Partners**
Prepared by: **Eliot Kwan**
Date: 19 Oct 2024

1. Overview & Objectives

Stonecliff Partners engaged us to conduct a comprehensive assessment of its raw data ingestion, quality controls and AI-readiness. This report delivers:



- **Ingestion Review:** A deep dive into your existing Azure SQL–based pipelines (Salesforce, HubSpot, Preqin, PitchBook and internal systems) to surface performance bottlenecks, schema drift and operational instability.
- **Data-Quality Assessment:** Identification and classification of high-impact issues—duplicates, invalid formats, missing values and inconsistency—across your silver- and gold-layer datasets, evaluated against completeness, validity, uniqueness, consistency and timeliness dimensions.
- **AI-Assurance Evaluation:** Scoring the feasibility of LLM-powered context extraction and rule-based augmentation for semi-structured fields, including next-action recommendations and sentiment analysis.
- **Risk Analysis:** Assessment of data risks—unauthorized access, data loss/corruption, schema changes, latency and regulatory non-detection—and corresponding mitigation strategies.
- **Remediation Roadmap:** A prioritized, phase-by-phase plan for rebuilding ingestion and quality pipelines, deploying monitoring tools, integrating APIs, and delivering engagement widgets and executive reports to turn raw feeds into trusted, analytics-ready assets.

2. Data Sources

All of the following are deployed as Azure SQL Databases and represent your **raw ingestion layer**, where overlapping records are expected before deduplication and the creation of silver/gold tables:

Database	Description
`sfdc_core_raw_db`	Raw ingest of the core Salesforce Production org (50+ tables: Accounts, Opportunities, Contacts, etc.)
`sfdc_alpha_partners_raw_db`	Raw ingest from the “Alpha Partners” affiliate Salesforce org (mirrors core schema for reconciliation)
`sfdc_global_ventures_raw_db`	Raw ingest from the “Global Ventures” affiliate Salesforce org (historical snapshot for lineage)
`sfdc_legacy_raw_db`	Historical dump of an older Salesforce org for audit and migration purposes
`hs_core_raw_db`	Raw ingest of HubSpot Core CRM (Contacts, Companies, Deals, Tickets, Lists, Workflows)
`hs_marketing_raw_db`	Raw ingest of HubSpot Marketing module (Emails, Forms, LandingPages, CampaignAnalytics)

`preqin_feed_raw_db`	Raw API ingest of Preqin fund data (FundProfiles, PerformanceMetrics, LPCommitments, NAVHistory)
`pitchbook_feed_raw_db`	Raw API ingest of PitchBook deal data (CompanySnapshots, DealFlow, Valuations, InvestorProfiles)
`int_sales_raw_db`	Internal SQL OLTP for sales transactions (Orders, Invoices, PriceLists, DiscountSchedules)
`int_customer_raw_db`	Internal SQL master for customer records (CustomerAccounts, AddressBook, ContactPreferences)

3. Known Issues

Stonecliff Partners identified the following **known problems** related to their data ingestion and data quality processes:

3.1. Data Ingestion

- **High Subscription Costs:** Consistently high fees with limited options for reducing expenses.
- **Operational Instability:** Frequent errors (over 10 daily) that interrupt regular operations.
- **High Customization Effort:** Significant modifications are needed to accommodate different client requirements, increasing implementation expenses.
- **Slow Vendor Support:** Bug fixes are delayed due to the vendor's lengthy ticketing system.

3.2. Data Quality

- **Numerous Duplicate Records:** Frequent existence of redundant information across datasets.
- **Inefficient Data Design:** Storing all data as nvarchar(max) leads to performance issues and inefficient storage utilization.
- **Inconsistent Data:** Varying formats and classifications across datasets.
- **Complex Manual Mappings:** Extensive and inefficient manual mappings within SQL views.

4. Performance Metrics

Based on our review of the Azure SQL raw data sources and planned stakeholder interviews.

Metric	Definition	Target Threshold
Ingestion Throughput	Number of records ingested per minute	≥ 8,000 records/min

ETL Processing Time	Total runtime for the daily batch job	≤ 3 hours
Query Latency	95th-percentile response time for key analytical queries	≤ 5 seconds
Data Freshness	Time lag from source update to availability in the lake	≤ 2 hours
Job Success Rate	Percentage of ETL jobs completing without errors	≥ 95 %

5. Data Quality Findings

The following table summarizes high-impact data quality findings. For a comprehensive list of all data quality issues, including those with lower impact, please see the Appendix: Detailed Data Analysis.

Metric	Column(s)	Database	Description
Completeness	`email`	`hs_core_raw_db`	15 % of HubSpot contacts lack any email address, preventing outreach and matching.
Validity	`email`	`hs_core_raw_db`	8 % of emails fail the standard regex (^[^@]+@[^@]+\.[^@]+$), causing bounce-backs.
Uniqueness	`deal_id`	`hs_core_raw_db`	6 % duplicate deal_id values indicate overlapping ingestion runs or upstream feeds.
Validity	`currency_code`	`hs_core_raw_db`	3 % of records use non-ISO or typo currency codes (e.g. “US$” vs. “USD”), breaking FX joins.
Completeness	`street_address`	`int_customer_raw_db`	9 % of customer records have missing street address, hampering geospatial analytics.
Consistency	`customer_id`	`int_sales_raw_db`	12 % of orders reference a customer_id not found in the Customer Master table.
Accuracy	`invoice_amount`	`int_sales_raw_db`	~2 % of invoices have negative or zero amounts, violating business rules.
Validity	`created_at`	`int_sales_raw_db`	4 % of order timestamps are non-ISO or malformed, breaking ingestion jobs.
Consistency	`shipped_at vs. created_at`	`int_sales_raw_db`	1 % of shipments have shipped_at < created_at, violating workflow logic.
Uniqueness	`account_id`	`sfdc_core_raw_db`	>10 % of records share the same account_id, risking duplicate account consolidation.

Completeness	`contact_id`	`sfdc_core_raw_db`	~5 % of rows have NULL in the primary contact key, blocking contact-to-account joins.
Timeliness	`nav_date`	`sfdc_core_raw_db`	20 % of NAV entries are older than 90 days, undermining fund performance analyses.
Completeness	`performance_metric`	`sfdc_core_raw_db`	10 % of performance metrics are NULL, leaving gaps in quarterly reporting.
Completeness	`valuation_amount`	`sfdc_global_ventures_raw_db`	7 % of deals are missing their valuation, skewing portfolio aggregation.
Validity	`status`	`sfdc_global_ventures_raw_db`	5 % of status fields contain placeholders (“N/A”, “Unknown”), obscuring true record state.

6. AI Assurance

During our initial stakeholder workshops and requirements-gathering sessions, we shortlisted a set of high-value ML-driven capabilities aligned with your strategic goals—cross-selling, risk scoring, pipeline forecasting and churn prediction. Each use case was then profiled against your remediated silver- and gold-layer datasets to produce a **feasibility score** reflecting feature availability, label completeness, update latency and operational stability. Scores range from 1 (low feasibility) to 10 (ready to launch), with accompanying notes highlighting key data limitations and remediation recommendations.

AI Solution	Feasibility Score	Notes
Cross-Sell Opportunities	3/10	Limited cross-entity joins; only ~20 % of Accounts have linked Opportunity histories in **sfdc_core_raw_db**, which constrains co-occurrence signals.
Predictive Account Risk Scoring	2/10	Sparse interaction logs; < 30 % of Activities contain both `EventDate` and `Description` fields, hampering churn-risk feature extraction.
Fundraising Pipeline Success Predictor	6/10	Moderate label coverage; Deal closures (`DealFlow.CloseDate`) exist for ~55 % of historical deals in **pitchbook_feed_raw_db**, but timestamp lags (often 48–72 h) necessitate careful windowing.
Churn Prediction & Lead Scoring	5/10	Engagement metrics are present, but only ~40 % of Contacts in **hs_core_raw_db** persist past a six-month lifecycle, requiring synthetic upsampling or fallback heuristics.

7. Enrichment & Augmentation

7.1. Rule-Based Tagging

We recommend defining a suite of deterministic, configurable rules to flag and tier records across key dimensions—geography, industry, client profile, compliance, and activity recency. These rules can be exposed via intuitive admin screens, and their outputs should feed both downstream analytics and real-time workflow alerts.

Feature	Data Source(s)	Table(s) / Column(s)	Notes
Geographic Risk Tiering	sfdc_core_raw_db, int_customer_raw_db	`Accounts.Country`, `CustomerAccounts.Address`	Map ISO country codes to Low/Med/High risk buckets based on external country list.
Industry Risk Tiering	sfdc_core_raw_db	`Accounts.IndustryCode`, `Opportunities.IndustrySector`	Use NAICS/SIC → risk tier mappings.
Client-Profile Risk Tiering	int_customer_raw_db, hs_core_raw_db	`CustomerAccounts.Segmentation`, `Contacts.LifecycleStage`	Tier clients by size / engagement metrics (e.g., AUM, deal count, last activity).
PEP / Sanction Flags	ext_pep_list_db (new), ext_sanctions_db (new)	`Watchlist.EntityName`, `Watchlist.IDNumber`	Cross-reference against external PEP and sanctions feeds; flag matches.
Internal Watchlist Flags	int_customer_raw_db, int_sales_raw_db	`ContactPreferences.Flagged`, `Orders.CustomerID`	Mark any customer/contact previously flagged in our internal compliance tracker.
Dormant Account Marker	int_sales_raw_db	`Orders.OrderDate`, `Invoices.InvoiceDate`	Identify accounts with no orders or invoices in the last X days (configurable).
Transaction Recency Marker	int_sales_raw_db	`Orders.OrderDate`, `Transactions.LastModified`	Compute “age” buckets (e.g. <30 days, 30–90 days, >90 days).

7.2. LLM-Powered Context Extraction

We propose leveraging custom-trained large language models to enrich unstructured and semi-structured text fields—surfacing investor sentiment, generating next-action recommendations, and tagging interaction summaries. Prompt templates should be carefully engineered for financial contexts, and a human-in-loop validation process should be established to maintain accuracy and consistency.

Feature	Data Source(s)	Table(s) / Column(s)	Model & Validation
Investor Sentiment Analysis	sfdc_core_raw_db, sfdc_alpha_partners_raw_db	`InvestorProfiles.CommunicationsLog`, `DealFlow.Notes`	Custom GPT-based model; prompt templates tuned for financial tone; sampled results reviewed weekly.
Next-Action Recommendations	sfdc_core_raw_db	`Tasks.Subject`, `Activities.Description`, `Opportunities.StageHistory`	Prompts engineered to suggest follow-ups (e.g. “Schedule call,” “Send proposal”); 10% human QA.
Interaction Summary Tags	hs_core_raw_db	`Tickets.ConversationThread`, `Workflows.History`	Extract key topics (e.g. “pricing issue,” “compliance question”) and tag accordingly.
Custom Note Auto-Classification	int_sales_raw_db, int_customer_raw_db	`Invoices.Comments`, `CustomerAccounts.Notes`	Auto-classify notes into categories (e.g. “billing,” “service request”) to streamline routing.

7.3. Third-Party Data Augmentation

No additional implementation is recommended at this stage, as the ingestion layer already incorporates two robust third-party enrichment sources.

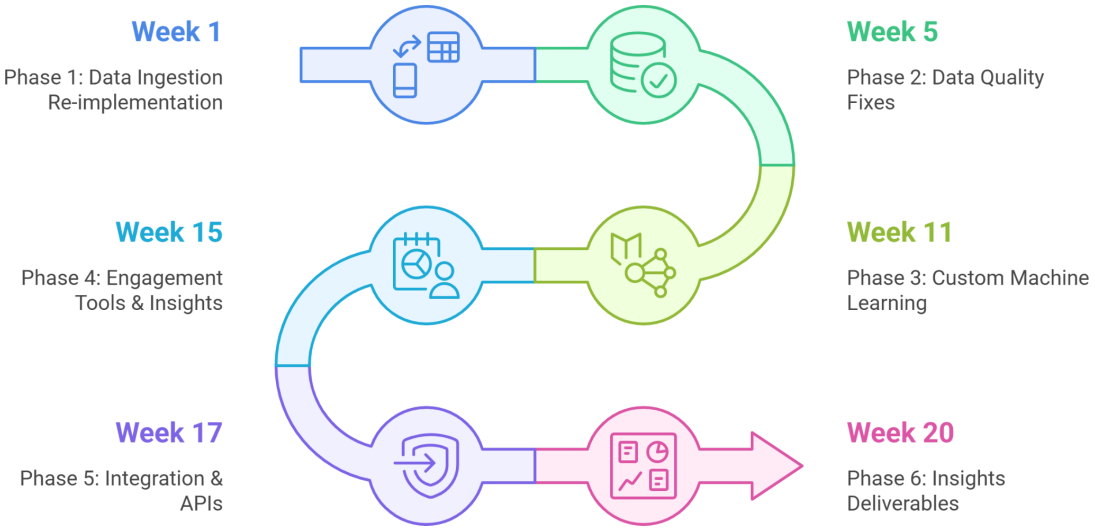
8. Data Risk Assessment

Based on our review of the Azure SQL raw data sources and planned stakeholder interviews.

Risk	Business Impact	Potential Mitigations
Unauthorized Access Unapproved users or roles accessing sensitive tables during ingestion or review.	5/5	• Enforce least-privilege RBAC • Enable Azure AD authentication + MFA • Audit-level logging and alerting
Data Loss or Corruption Errors or failures during ETL can result in missing or altered records.	4/5	• Automated, point-in-time backups • Source-to-target checksum validation • Transactional integrity checks
Inconsistent Definitions Disparate naming conventions or business rules across sources causing misalignment.	3/5	• Maintain central metadata catalog • Data steward–approved glossaries • Pre-ingestion schema validation
Regulatory Non-Detection Failure to identify PII/PHI or other regulated fields within raw tables.	5/5	• Automated data-scanning for sensitive patterns • Quarterly compliance reviews • Inline masking for high-risk columns

Latency & Availability Downtime or slow performance impeding downstream analytics and reporting.	3/5	• Geo-redundant failover groups • Performance monitoring with alert thresholds • Query-level timeouts and retries
Untracked Schema Changes Unexpected alterations to table structures breaking pipelines.	4/5	• Git-backed DDL management • Automated schema drift detection • CI/CD gating for database migrations

9. Remediation Roadmap



Phase 1: Data Ingestion Re-implementation (4 weeks)
We recommend fully re-implementing your ingestion pipeline to resolve the bottlenecks and schema drift uncovered in Section 2 of the Data Quality & AI Assurance Report. This work includes refactoring connectors, standardizing metadata captures, and adding robust error handling.

Phase 2: Data Quality Fixes (6 weeks)
We will rebuild your data quality pipelines end-to-end—implementing advanced deduplication algorithms, automated validity checks and completeness validations as detailed in Section 4. In parallel, we'll deliver a lightweight web app that sits "man-in-the-middle," allowing your team to monitor data health in real time, review flagged records and intervene when necessary. This ensures your silver and gold layers consistently meet the reliability standards required for downstream analytics.

Phase 3: Custom Machine Learning (4 weeks)
Once data quality is assured, we can build bespoke ML models on request—complete with train/validation pipelines, explainability dashboards and handover documentation.

Phase 4: Engagement Tools & Insights (3 weeks)
Develop one engagement widget for Salesforce and one for HubSpot, plus a suite of five executive-grade reports that leverage your cleansed data to drive personalized outreach and strategic decision-making.

Phase 5: Integration & APIs (2 weeks)
Build secure endpoints and webhook integrations to serve cleansed data into both ML workflows and reporting systems—ensuring all downstream tools stay in sync.

Phase 6: Insights Deliverables (3 weeks)
Iterate on and refine dashboard reports, KPI trackers and ad hoc analyses to surface the most impactful insights for your stakeholders.