Companion Notes

Data Infrastructure Audit Results — Source Tracing

Table of contents

This companion document traces every detail in the Data Infrastructure Audit Results to either a specific location on the Cloudcore website or flags it as an assumption invented for the brief.

Part 1: Facts Sourced from the Cloudcore Website

Data Source Identification

All seven data sources are real systems confirmed in the repo:

Data Source	System	Confirmed In
Infrastructure telemetry	Prometheus + Grafana	`chatbots/_backstories/mark_gonzalez_cto.md`
Security event logs	Splunk SIEM	Same file; also `docs/logs/`
Support tickets	Internal ticketing system	`chatbots/_backstories/customer_support_lead`
Customer records	HubSpot CRM	`chatbots/_backstories/tom_bradley_marketing`
Billing and financial	Internal billing system	`chatbots/_backstories/aisha_rahman_cfo.md`
Sales performance	HubSpot + spreadsheets	`tom_bradley_marketing_manager.md`, `data/cloudcore-sales-data.csv`
HR and access control	Auth0 + Active Directory	`docs/policies/access_control.qmd`, `chatbots/_backstories/karen_lee_hr_manager.`

Data Owners

All data owner assignments match the organisational structure in backstory files:

Owner	Role	Source
Martin Nguyen	Cloud Service Operations Manager	`chatbots/_backstories/cloud_service_operati`
Sophia Martines	CISO	`chatbots/_backstories/sophia_martines_ciso.`
Samantha Wong	Customer Support Lead	`chatbots/_backstories/customer_support_lead`
Lisa Chen	CMO	`chatbots/_backstories/lisa_chen_cmo.md`
Aisha Rahman	CFO	`chatbots/_backstories/aisha_rahman_cfo.md`
Karen Lee	HR Manager	`chatbots/_backstories/karen_lee_hr_manager.`
Raj Patel	IT Manager	`chatbots/_backstories/raj_patel_it_manager.`

Data Quality — Sourced Issues

Issue	Source
Data siloed across systems, no unified platform	chatbots/_backstories/jamal_al_sayed_data_analyst.md, mark_gonzalez_cto.md
CTO rates data readiness 2/5	mark_gonzalez_cto.md
Data "acceptable" for operations but not assessed for ML	Same file
No data warehouse or data lake	jamal_al_sayed_data_analyst.md
Basic BI tools only (Power BI, Excel)	Same file
No formal data governance	Same file
Data definitions vary across systems	Same file
3-4 years historical data (completeness varies)	Same file
Data team: 2 people (stretched thin)	Same file
6-12 months data prep needed before ML	Same file; also mark_gonzalez_cto.md
CRM has duplicate records, inconsistent formatting, missing fields	jamal_al_sayed_data_analyst.md (general); CRM migration issues implied
~40% employees have broader access than required	karen_lee_hr_manager.md
RBAC definitions incomplete	Same file
Account termination: policy says 24hrs, HR says 2hrs	docs/policies/access_control.qmd vs HR backstory
Auth0 migration left policies referencing Okta	docs/policies/access_control.qmd
500-800 daily security alerts	docs/policies/ (incident response notes)
Products: CloudSync, DataVault, SecureLink, Analytics Pro	data/cloudcore-sales-data.csv
Sales regions: North, South, East, West, Central, Metro	Same file
Customer segments vary (Small, Medium, Large vs Enterprise, SME)	data/cloudcore-customer-data.csv vs data/cloudcore-sales-data.csv
No AI governance framework	mark_gonzalez_cto.md, sophia_martines_ciso.md
Data classification policy still in DRAFT	docs/policies/data_classification.qmd (POL-DATA-001 v1.2 DRAFT)

Compliance Posture

Framework	Status	Source
ISO 27001 certified	Confirmed	sophia_martines_ciso.md, cloudcore_company_overview.md

Framework	Status	Source
SOC 2 Type II	Confirmed	Same files
Australian Privacy Act compliant	Confirmed	`cloudcore_company_overview.md`
NDB scheme compliant	Confirmed	`chatbots/_backstories/security_compli`
GDPR compliant (EU data)	Confirmed	`emily_chen_head_of_compliance.md`
HIPAA in progress (partial)	Confirmed	Same file
No AI-specific data impact assessment	Confirmed	`mark_gonzalez_cto.md`, `sophia_martines_ciso.md`, `emily_chen_head_of_compliance.md`

Compliance Framework Details

Detail	Source
ISO 27001 controls A.12.1.2 and A.14.2.2	`docs/policies/change_management.qmd`
GDPR 72-hour breach notification	`docs/policies/breach_notification.qmd`
GDPR penalties up to 4% revenue or EUR 20M	`docs/articles/` (risk analysis article)
Privacy Act fines up to $2.2M per violation	`chatbots/_backstories/data_breach_overview.md`
HIPAA penalties $100-$50K per violation	`docs/policies/breach_notification.qmd`
7-year audit trail retention	`docs/policies/data_management.qmd`

Integration Architecture — Confirmed Gaps

Gap	Source
No data warehouse	`jamal_al_sayed_data_analyst.md`, `mark_gonzalez_cto.md`
No unified analytics platform	Same files
No real-time analytics pipeline	Same files
No ML infrastructure	`mark_gonzalez_cto.md`
No GPU instances	`david_wilson_cloud_infrastructure_architect.md`

Breach Incident Data

The reference to the September 2024 breach as a documented dataset for model training is supported by extensive log files in `docs/logs/` including VPN, database, firewall, EDR, SIEM, and application server entries with full timestamps.

Cross-References

All website URLs reference real pages on the Cloudcore site, including the risk assessment frameworks document at `docs/support/risk_assessment_frameworks.md`.

Part 2: Assumptions and Invented Details

All Data Volume Figures

No data volume figures exist anywhere in the repo. Every volume number was invented:

Data Source	Invented Volume	Reasoning
Infrastructure telemetry	~2.1M data points/day	Plausible for Prometheus monitoring ~2,500 VMs with standard exporter intervals
Security logs	~12 GB/day	Plausible for Splunk ingestion across multiple log sources at this scale
Support tickets	~45,000 historical; ~1,200/month	The sample CSV has 100 tickets; 500+ clients generating ~2-3 tickets each per month is plausible
CRM contacts	~85,000 records	500+ active clients plus historical contacts, prospects, and marketing list
CRM companies	~4,200 records	Includes prospects, former clients, and partners
Billing invoices	~6,000/year	500+ clients, monthly billing cycles

All Quality Scores (1-5 Scale)

The quality scores were invented to follow the design principle that infrastructure/operational data should be high quality and customer-facing data should be messy:

Data Source	Score	Design Reasoning
Infrastructure telemetry	4.5	Machine-generated, automated, minimal human intervention
Security logs	4.0	Machine-generated but alert classification has human elements
Billing/financial	3.5	Audit requirements enforce some discipline
Support tickets	3.0	Semi-structured; human-entered data with quality variance
HR/access	3.0	HR data accurate; access control data has known drift
Sales performance	2.5	Dual systems, no single source of truth
CRM	2.0	Migration-damaged, never cleaned, poorly adopted

All Specific Quality Percentages

Every percentage figure describing data quality issues was invented. None appear in the repo:

Infrastructure telemetry:

- 98.5% data completeness — invented
- 1.5% gaps during maintenance — invented
- 8% of metrics lacking standardised client attribution labels — invented
- Inconsistent naming (cpu_usage_percent vs node_cpu_utilisation) — invented example

Security logs:

- 72% alert classification accuracy — invented
- 6:1 false positive to true positive ratio — invented
- Non-standard timestamp formats for pre-2021 sources — invented

Support tickets:

- 18% miscategorised priority levels — invented
- 12% descriptions under 20 words — invented
- 22% use account numbers, 31% use email addresses for customer ID — invented
- Historical tickets pre-mid-2022 lack satisfaction scores — invented

CRM:

- 15% confirmed duplicate contacts — invented
- 8% probable duplicates — invented
- 26% incomplete industry classification — invented
- 34% missing job title/role — invented
- 11.2% email bounce rate — invented
- 41% missing lead source attribution (pre-migration) — invented

Billing:

- 6% monthly billing cycle discrepancies — invented
- Product codes changed twice in three years — invented (products are real)
- Contracts stored as PDFs — invented

Sales:

- 35% closed-lost opportunities missing loss reason — invented
- Regional reporting format inconsistencies — invented

HR/access:

- ~40% over-provisioned — SOURCED (Karen Lee backstory)
- Account termination timing inconsistency — SOURCED (policy conflict)
- Auth0/Okta gap — SOURCED

Data Value Pyramid Assessment

The entire data value pyramid section was composed for the brief. The repo does not contain an analytics maturity assessment. The placement decisions:

- Descriptive: "partially achieved" — based on Power BI and Excel usage confirmed in Jamal's backstory
- Diagnostic: "minimal" — based on manual root cause analysis described in operations backstories
- Predictive: "not attempted" — confirmed by CTO ("no predictive models exist" equivalent in backstory)
- Prescriptive: "not attempted" — no evidence of automated decision support anywhere in repo

Integration Architecture Details

Sourced elements:

- Point-to-point integrations exist (Prometheus to Grafana, CrowdStrike to Splunk, etc.)
- No integration middleware confirmed
- Python scripts for data movement — inferred from dev team tech stack

Invented elements:

- Characterisation as "point-to-point with manual bridges"
- "Scheduled Python scripts for billing data aggregation" — invented (Python is confirmed as the dev stack; its use for ETL is assumed)
- "Splunk built-in log collection and normalisation" as the only formal ETL — framing invented
- "Prometheus federation for metrics consolidation" — standard Prometheus capability, assumed in use
- "Loss of Jamal or junior analyst would create immediate knowledge gaps" — invented but strongly implied by backstory

Missing capabilities table:

- All items listed as missing are confirmed absent in backstory files
- The descriptions of impact (e.g., "every analysis requires manual data assembly") are framing, not direct quotes

Compliance — Healthcare and Finance Contract Requirements

Entirely invented. The repo confirms that healthcare and finance are key client sectors with strict compliance requirements, but no specific contract terms appear anywhere. The invented terms are plausible for Australian healthcare and finance cloud contracts:

- Healthcare: data residency, audit logging, client notification for new systems, annual security assessments, 24-hour breach notification
- Finance: data classification documentation, third-party access approval, pen test sharing, data retention schedules, real-time access monitoring

Australian Privacy Act — AI-Specific Obligations

The Privacy Act and APPs are referenced in compliance backstories. The specific mapping of APP 1, 3, 6, and 11 to AI use cases was composed for the brief. These are accurate representations of the APPs but their application to Cloudcore's AI plans is analysis, not sourced fact.

AI-Specific Data Impact Assessment Gap

The gap itself is sourced (multiple backstories confirm no AI governance framework). The specific list of AI concerns that the gap leaves unaddressed (training data consent, model bias, explainability, re-identification risk, etc.) was composed for the brief.

All Infrastructure Cost Benchmarks

Every cost figure in the benchmarks section was invented based on typical Australian market rates in 2024-2025:

Component	Invented Range	Basis
Cloud data warehouse	$36-72K/yr	Typical Snowflake/BigQuery pricing at moderate scale
ETL platform	$18-36K/yr	Fivetran/dbt Cloud mid-tier pricing
Data catalogue	$12-24K/yr	Commercial tooling; open-source alternative noted
MDM	$24-48K/yr	Implementation-heavy; conservative estimate
ML platform	$36-96K/yr	SageMaker/Azure ML compute costs are highly variable
ML engineer salary	$180-250K/yr	SOURCED from Karen Lee backstory
AI contractor	$2-3.5K/day	Typical Australian specialist consulting rates
MLOps tooling	$6-18K/yr	Commercial options; MLflow as free alternative noted

Component	Invented Range	Basis
API gateway	$6-24K/yr	AWS API Gateway available through existing partnership
Event streaming	$18-48K/yr	Managed Kafka pricing at modest scale
Integration middleware	$24-60K/yr	Significant implementation cost acknowledged

Sensitivity Classifications

The data classification levels (PUBLIC, INTERNAL, CONFIDENTIAL, RESTRICTED) are sourced from the draft data classification policy (POL-DATA-001). The assignment of specific data sources to these levels was invented, though informed by the policy's descriptions of each tier.

Retention Periods

Detail	Status
Security logs: 12 months online, 7 years archived	7-year audit trail retention is sourced; 12-month online is invented
Infrastructure telemetry: 90 days full, 1-year downsampled	Entirely invented
Support tickets: indefinite, no formal policy	Invented; plausible given no data retention schedule is documented for tickets
CRM: indefinite, no hygiene schedule	Invented
Billing: 7 years	Sourced (financial compliance requirement in data management policy)
HR: 7 years post-departure, access logs 12 months	7-year retention sourced; 12-month access log period invented

---

This companion document is for instructor reference. It is not intended for student distribution unless adapted.