Cloudcore Networks

Data Infrastructure Audit Results

Table of contents

Purpose of This Document

This document presents the results of an internal audit of Cloudcore's data infrastructure, conducted to assess readiness for AI and advanced analytics initiatives. The audit examined data quality, integration architecture, compliance posture, and infrastructure cost context. Findings are intended to support infrastructure planning decisions.

Data Inventory

The audit identified seven major data sources across Cloudcore's environment. Each was assessed for volume, quality, ownership, sensitivity, and specific issues.

1. Infrastructure Telemetry

Dimension	Assessment
System	Prometheus and Grafana
Volume	~2.1 million metric data points per day across ~2,500 VMs and ~200 physical/virtual servers
Quality score	4.5 / 5
Data owner	Martin Nguyen, Cloud Service Operations Manager
Sensitivity	INTERNAL
Retention	90 days at full resolution; downsampled to 1-year rolling archive

Quality justification: Infrastructure telemetry is machine-generated, consistently structured, and timestamped. Collection is automated via Prometheus exporters with minimal human intervention. Data completeness is estimated at 98.5%; the remaining 1.5% represents brief gaps during maintenance windows or agent restarts. This is Cloudcore's cleanest and most AI-ready data source.

Specific issues: Metric naming conventions are inconsistent across older and newer deployments (e.g., `cpu_usage_percent` vs. `node_cpu_utilisation`). Approximately 8% of metrics lack standardised labels for client attribution, making per-client analysis difficult without cross-referencing provisioning records.

---

2. Security Event Logs

Dimension	Assessment
System	Splunk SIEM (aggregating CrowdStrike EDR, Palo Alto firewall, Auth0, VPN, application logs)
Volume	~12 GB per day; 500 to 800 alerts generated daily
Quality score	4.0 / 5
Data owner	Sophia Martines, CISO
Sensitivity	CONFIDENTIAL
Retention	12 months online; 7 years archived (compliance requirement)

Quality justification: Log data is machine-generated and well-structured. Splunk normalises formats across sources. The September 2024 breach provided a fully documented incident with timestamped entries across VPN, database, firewall, EDR, SIEM, and application logs, offering a validated dataset for anomaly detection model training.

Specific issues: Alert classification accuracy is estimated at 72%; the remaining 28% are miscategorised or lack sufficient context for automated triage. Some legacy log sources (pre-2021 systems) use non-standard timestamp formats requiring manual parsing. The ratio of false positives to true positives in daily alerts is approximately 6:1, contributing to alert fatigue.

---

3. Support Ticket Data

Dimension	Assessment
System	Internal ticketing system
Volume	~45,000 tickets historically (3.5 years); ~1,200 new tickets per month
Quality score	3.0 / 5
Data owner	Samantha Wong, Customer Support Lead (reports to Sarah Thompson, COO)
Sensitivity	INTERNAL (some tickets contain CONFIDENTIAL client details)
Retention	Indefinite; no formal retention policy applied

Quality justification: Ticket data is semi-structured. Category and priority fields are reliably populated (98% complete), and resolution times are tracked. However, free-text descriptions vary significantly in detail and consistency.

Specific issues:

- 18% of tickets have miscategorised priority levels (identified by comparing resolution urgency against assigned priority)
- 12% of ticket descriptions contain fewer than 20 words, providing insufficient detail for text-based analysis or automated routing
- Customer identifiers are inconsistently formatted: 22% use account numbers, 31% use email addresses, and the remainder use company names with variable spelling
- No structured field for root cause; resolution notes are free-text with no controlled vocabulary
- Historical tickets prior to mid-2022 lack satisfaction scores (field was added later)

---

4. Customer Records (CRM)

Dimension	Assessment
System	HubSpot CRM
Volume	~85,000 contact records; ~4,200 company records; ~500+ active client accounts
Quality score	2.0 / 5
Data owner	Lisa Chen, CMO (marketing data); Sales team (pipeline data)
Sensitivity	CONFIDENTIAL
Retention	Indefinite; no data hygiene schedule

Quality justification: The 2022 CRM migration from a legacy contact management system introduced significant data quality problems that have never been systematically addressed.

Specific issues:

- 15% of contact records are confirmed duplicates (same person, different records created by different teams)
- An additional estimated 8% are probable duplicates requiring manual review
- 26% of company records have incomplete industry classification
- 34% of contact records are missing job title or role information
- Email bounce rate on the full database is 11.2%, indicating substantial stale data

- Lead source attribution is missing for 41% of records created before the HubSpot migration
- Sales pipeline data is unreliable; many sales staff continue to track opportunities in personal spreadsheets rather than HubSpot
- No integration with billing or support systems; customer health must be assessed manually by cross-referencing exports

---

5. Billing and Financial Data

Dimension	Assessment
System	Internal billing and invoicing system
Volume	~6,000 invoices per year; monthly usage records for 500+ clients
Quality score	3.5 / 5
Data owner	Aisha Rahman, CFO
Sensitivity	RESTRICTED (contains payment information)
Retention	7 years (financial compliance)

Quality justification: Financial data is relatively well-maintained due to audit requirements and regulatory obligations. Invoice records are complete and reconciled monthly.

Specific issues:

- Usage metering data requires manual validation against service records; discrepancies found in approximately 6% of monthly billing cycles
- Product categorisation has changed twice in the past three years (CloudSync, DataVault, SecureLink, Analytics Pro are current names); historical data uses legacy product codes that are not consistently mapped
- Client contract terms are stored as PDF attachments rather than structured data, making automated analysis of contract value, renewal dates, and SLA terms impossible without manual extraction
- Revenue attribution by sector relies on manually maintained spreadsheet classifications, not CRM data

---

6. Sales Performance Data

Dimension	Assessment
System	Combination of HubSpot CRM and regional spreadsheets
Volume	Quarterly records across 6 regions (North, South, East, West, Central, Metro) and 4 product lines
Quality score	2.5 / 5
Data owner	Sales team; no single owner
Sensitivity	INTERNAL
Retention	~2 years structured; earlier data in inconsistent formats

Quality justification: Sales data exists in two parallel systems. Marketing tracks leads and top-of-funnel metrics in HubSpot. Individual sales representatives maintain pipeline and closed-deal data in personal or regional spreadsheets.

Specific issues:

- No single source of truth for sales pipeline; HubSpot and spreadsheet figures often conflict
- Regional reporting formats are not standardised; Metro and North regions use different column structures
- Customer segment definitions (Small, Medium, Large, Enterprise, SME) vary between sales reports and CRM records
- Win/loss data is incomplete; approximately 35% of closed-lost opportunities have no recorded reason for loss
- Sales representative attribution is clean for 2023 onward but unreliable for earlier periods

---

7. HR and Access Control Data

Dimension	Assessment
System	Auth0 (identity), Active Directory (on-premise), HR management system
Volume	47 active employee records; ~120 historical records; access permissions across all systems
Quality score	3.0 / 5
Data owner	Karen Lee, HR Manager; Raj Patel, IT Manager (shared)
Sensitivity	RESTRICTED

Dimension	Assessment
Retention	Employee records: 7 years post-departure; access logs: 12 months

Quality justification: Employee master data is maintained by HR and is generally accurate. However, access control data has known integrity issues.

Specific issues:

- ~40% of employees have broader system access than their role requires (identified in quarterly access review)
- Role-based access control (RBAC) definitions are incomplete; actual permissions often diverge from documented role templates
- Onboarding and offboarding access changes involve manual coordination between HR, IT, and department managers, with no automated workflow
- Account termination timing is inconsistent: policy states 24 hours, HR procedure states 2 hours, and actual practice varies
- Auth0 migration from Okta (December 2023) left some access policies referencing the old identity provider

---

Data Quality Summary

Data Source	Quality Score	AI Readiness	Key Barrier
Infrastructure telemetry	4.5 / 5	High	Metric naming inconsistency
Security event logs	4.0 / 5	Medium-High	Alert classification accuracy; false positive ratio
Support tickets	3.0 / 5	Medium	Inconsistent customer identifiers; sparse descriptions
Billing and financial	3.5 / 5	Medium	Contract data unstructured; product code mapping
HR and access control	3.0 / 5	Low	Access permission drift; manual processes

Data Source	Quality Score	AI Readiness	Key Barrier
Sales performance	2.5 / 5	Low	Dual systems; no single source of truth
Customer records (CRM)	2.0 / 5	Low	Duplicates, missing fields, no integration

Pattern: Cloudcore's infrastructure and operational data is relatively clean and well-structured, reflecting mature engineering practices. Customer-facing and commercial data is significantly messier, reflecting the organisational challenges of the CRM migration, rapid growth, and siloed teams. This contrast is the central data quality challenge for any AI initiative targeting customer experience or commercial outcomes.

Data Value Pyramid Assessment

The data value pyramid maps an organisation's analytics maturity from descriptive (what happened) through diagnostic (why), predictive (what will happen), and prescriptive (what should we do).

Level	Status	Evidence
Descriptive (what happened)	Partially achieved	Power BI dashboards exist for operational metrics. Support metrics (resolution time, satisfaction, ticket volume) are reported weekly. Financial reporting is monthly. However, cross-system views require manual assembly by the data team.
Diagnostic (why it happened)	Minimal	Root cause analysis is performed manually for major incidents. No automated correlation between data sources. Jamal Al-Sayed's team can investigate specific questions but there is no self-service diagnostic capability.

Level	Status	Evidence
Predictive (what will happen)	Not attempted	No predictive models exist. Capacity planning uses historical trend extrapolation in spreadsheets. Churn risk is identified reactively (after customers raise concerns), not proactively.
Prescriptive (what should we do)	Not attempted	No automated decision support. Resource allocation, staffing, and investment decisions are based on experience and judgment, not data-driven recommendations.

Assessment: Cloudcore is operating primarily at Level 1 (descriptive) with pockets of Level 2 (diagnostic) for security incidents and major operational issues. Moving to predictive analytics would require solving the data integration challenge first, as no single system currently holds the cross-functional data needed for meaningful prediction.

Integration Architecture Assessment

Current Approach

Cloudcore's integration architecture is best described as **point-to-point with manual bridges**. There is no integration middleware, enterprise service bus, or API gateway connecting internal systems.

Integration Type	Examples	Status
Automated point-to-point	Prometheus to Grafana; CrowdStrike to Splunk; GitHub Actions to ArgoCD	Working well within functional silos
Batch file transfer	Service usage data to billing (daily batch); support metrics to Power BI (weekly export)	Functional but error-prone; manual validation required

Integration Type	Examples	Status
Manual data transfer	CRM data to financial reporting; support data to customer health assessment; sales data consolidation	Labour-intensive; relies on the 2-person data team
API integration	HubSpot lead capture from website; Auth0 SSO across applications	Limited to a few well-defined use cases

Existing ETL Processes

Cloudcore has no formal ETL platform. Data movement between systems relies on:

- Scheduled Python scripts (maintained by the development team) for billing data aggregation
- Manual CSV exports from individual systems into Power BI
- Splunk's built-in log collection and normalisation (security data only)
- Prometheus federation for infrastructure metrics consolidation

These processes are fragile, undocumented, and maintained by individuals rather than teams. The data team has flagged that the loss of either Jamal Al-Sayed or his junior analyst would create immediate knowledge gaps in how data is extracted and transformed.

What Is Missing

Capability	Current State	Impact
Data warehouse	Does not exist	No single source of truth for cross-functional analytics; every analysis requires manual data assembly
Master data management (MDM)	Does not exist	Customer identifiers, product codes, and segment definitions are inconsistent across systems

Capability	Current State	Impact
Real-time data pipelines	Does not exist	All cross-system data movement is batch or manual; minimum latency is daily
API gateway	Does not exist	No centralised API management, rate limiting, or access control for internal integrations
Data catalogue	Does not exist	No inventory of available datasets, their definitions, or their lineage; tribal knowledge only

---

Compliance and Data Handling

Current Compliance Posture

Framework	Status	Relevance to AI
ISO 27001	Certified (achieved ~18 months ago)	Requires documented risk assessment for new technology initiatives including AI; controls A.12.1.2 (change management) and A.14.2.2 (system change control) apply
SOC 2 Type II	Compliant (renewed annually)	AI systems processing customer data must meet SOC 2 trust service criteria for security, availability, and confidentiality
Australian Privacy Act (APPs)	Compliant	AI systems using personal information must comply with Australian Privacy Principles; APP 6 (use and disclosure) and APP 11 (security) are most relevant
Notifiable Data Breaches (NDB)	Compliant	Any AI system with access to personal information falls under NDB reporting obligations if compromised

Framework	Status	Relevance to AI
GDPR	Compliant (EU customer data)	AI decisions affecting EU data subjects may trigger Article 22 (automated decision-making) requirements; Data Protection Impact Assessments required
HIPAA	Partially compliant (in progress)	Healthcare client data used for AI training would require Business Associate Agreement coverage and additional safeguards

Healthcare Client Contract Requirements

Cloudcore's healthcare clients (representing approximately 25% of revenue) operate under contractual terms that include:

- All patient-adjacent data must remain within Australian data centres
- Data access must be logged and auditable
- Any new system processing healthcare data requires prior written notification to the client
- Annual security assessments must be provided to the client
- Breach notification within 24 hours (stricter than the statutory 72-hour NDB requirement)

AI implication: Any AI system trained on or processing healthcare client data would require individual client notification and potentially contract amendments. Using healthcare data for model training (even anonymised) may require explicit consent depending on contract terms.

Finance Client Contract Requirements

Finance sector clients (representing approximately 20% of revenue) have similarly strict requirements:

- Data classification and handling procedures must be documented and provided
- Third-party access to client data (including AI vendor platforms) requires prior approval
- Regular penetration testing results must be shared
- Data retention and deletion must follow agreed schedules
- Real-time monitoring of access to financial data is required

AI implication: Sending finance client data to external AI platforms (e.g., cloud-hosted ML services) may breach third-party access clauses unless explicitly approved. On-premise or private-cloud AI deployment may be necessary for finance workloads.

Australian Privacy Act Obligations Relevant to AI

The Australian Privacy Act and Australian Privacy Principles create several obligations relevant to AI deployment:

- **APP 1 (Open and transparent management):** Organisations must have a clearly expressed privacy policy covering how AI uses personal information
- **APP 3 (Collection):** Personal information should only be collected where reasonably necessary; AI training data collection must be justified
- **APP 6 (Use and disclosure):** Personal information collected for one purpose cannot be used for a materially different purpose (e.g., support ticket data collected for service improvement cannot be repurposed for marketing AI without consent)
- **APP 11 (Security):** Organisations must take reasonable steps to protect personal information from misuse, interference, and loss; this extends to AI model security and training data protection
- **Notifiable Data Breaches scheme:** Any eligible data breach involving AI systems must be reported to the OAIC within 72 hours

Gap: No AI-Specific Data Impact Assessment Process

Cloudcore currently has no process for assessing the data protection implications of AI initiatives. The existing Data Protection Impact Assessment (DPIA) process covers new systems and data handling changes but does not address AI-specific concerns including:

- Training data sourcing and consent
- Model bias and fairness assessment
- Automated decision-making transparency
- Model output explainability
- Training data retention and deletion
- Re-identification risk from anonymised datasets
- Cross-border data transfer for cloud AI processing

The data classification policy (POL-DATA-001) remains in draft status and has not been formally approved, further complicating the governance foundation for AI data handling.

Infrastructure Cost Benchmarks

The following cost ranges are based on Australian market rates for organisations at Cloudcore's scale (~500 clients, ~47 employees, two data centres). All figures are annual unless noted.

Data Platform Costs

Component	Estimated Annual Cost (AUD)	Notes
Cloud data warehouse (e.g., Snowflake, BigQuery, Redshift)	$36,000 to $72,000	Based on moderate query volume and ~5 TB storage; scales with usage
ETL/data integration platform (e.g., Fivetran, dbt Cloud)	$18,000 to $36,000	Depends on number of connectors and data volume
Data catalogue and governance tooling	$12,000 to $24,000	Could start with open-source alternatives to reduce cost
Master data management	$24,000 to $48,000	Significant implementation effort beyond licensing

AI/ML Platform Costs

Component	Estimated Annual Cost (AUD)	Notes
Managed ML platform (e.g., SageMaker, Azure ML)	$36,000 to $96,000	Highly variable; depends on compute usage and model training frequency

Component	Estimated Annual Cost (AUD)	Notes
ML engineer salary	$180,000 to $250,000	Market rate for Perth/Sydney; scarce talent pool
AI/ML contractor or consulting engagement	$2,000 to $3,500 per day	For specialist advisory or implementation support
MLOps tooling (experiment tracking, model registry)	$6,000 to $18,000	Could start with open-source (MLflow) at minimal cost

Integration Infrastructure Costs

Component	Estimated Annual Cost (AUD)	Notes
API gateway (e.g., Kong, AWS API Gateway)	$6,000 to $24,000	AWS API Gateway available through existing partnership
Event streaming platform (e.g., Kafka, managed equivalent)	$18,000 to $48,000	Only needed if real-time pipelines are required
Integration middleware	$24,000 to $60,000	Significant implementation cost beyond licensing

Cost Context

Against the proposed $250,000 AI investment envelope, these benchmarks illustrate the trade-offs:

- A data warehouse plus basic ETL tooling would consume $54,000 to $108,000 annually, leaving limited room for AI-specific investment
- A single ML engineer at market rate ($180,000 to $250,000) would consume most or all of the budget alone
- Leveraging existing AWS or Azure partnerships for managed AI services could reduce platform costs but still requires skilled staff to build and maintain models
- The most cost-effective path may involve using AI features already embedded in existing tools (Splunk ML analytics, HubSpot predictive lead scoring, CrowdStrike AI threat detection) while building foundational data infrastructure

---

Cross-References

For additional context, the following resources are available on the Cloudcore Networks website:

- **Security policies and data handling:** The data classification policy (draft), data protection policy, and access control policy are available at cloudcore.eduserver.au/docs/policies/
- **Breach incident documentation:** Detailed security logs from the September 2024 breach, including database query logs, firewall alerts, and SIEM correlation events, are available at cloudcore.eduserver.au/docs/logs/
- **Risk assessment frameworks:** ISO 27005 and NIST SP 800-30 templates used by Cloudcore are documented at cloudcore.eduserver.au/docs/support/risk_assessment_frameworks

---

Cloudcore Networks is a fictional company created for educational purposes. Any resemblance to real organisations is coincidental.