Combined Reference

Data Infrastructure Audit Results — Full Provenance

Table of contents

This document reproduces every detail from the Data Infrastructure Audit Results with inline provenance tags. Each fact is marked as one of:

- **SOURCED** — directly from a file on the Cloudcore website, with the file path noted

- **INFERRED** — a reasonable conclusion drawn from sourced material, but not explicitly stated
- **INVENTED** — created for the brief; plausible and non-contradictory, but not in the repo

---

Data Inventory

1. Infrastructure Telemetry

Detail	Status	Source / Reasoning
System: Prometheus and Grafana	SOURCED	`mark_gonzalez_cto.md`
Volume: ~2.1M data points/day	INVENTED	Plausible for ~2,500 VMs with standard exporter intervals
~2,500 VMs, ~200 servers	SOURCED	`david_wilson_cloud_infrastructure_architect.md`, `raj_patel_it_manager.md`
Quality score: 4.5/5	INVENTED	Design principle: infrastructure data should be high quality
Data owner: Martin Nguyen	SOURCED	`cloud_service_operations_manager_martin_nugyen.md`
Sensitivity: INTERNAL	INVENTED	Classification levels sourced from draft policy; assignment to this source is fabricated
Retention: 90 days full, 1-year downsampled	INVENTED	Standard Prometheus retention pattern; not specified in repo
98.5% completeness	INVENTED	Specific percentage fabricated
1.5% gaps during maintenance	INVENTED	Specific percentage fabricated
Metric naming inconsistency (cpu_usage_percent vs node_cpu_utilisation)	INVENTED	Specific example fabricated; naming drift is plausible
8% metrics lack client attribution labels	INVENTED	Specific percentage fabricated

2. Security Event Logs

Detail	Status	Source / Reasoning
System: Splunk SIEM	SOURCED	`mark_gonzalez_cto.md`
Aggregating CrowdStrike, Palo Alto, Auth0, VPN, app logs	SOURCED	Systems confirmed across multiple backstories; aggregation is standard SIEM function
Volume: ~12 GB/day	INVENTED	Plausible for this scale of Splunk deployment
500-800 daily alerts	SOURCED	`docs/policies/` (incident response implementation notes)
Quality score: 4.0/5	INVENTED	Machine-generated but human classification elements reduce score
Data owner: Sophia Martines	SOURCED	`sophia_martines_ciso.md`
Sensitivity: CONFIDENTIAL	INVENTED	Classification assignment fabricated
Retention: 12 months online	INVENTED	Not specified
Retention: 7 years archived	SOURCED	`docs/policies/data_management.qmd` (audit trail retention)
Breach provides documented dataset for model training	SOURCED	`docs/logs/` contains full timestamped entries
Alert classification accuracy 72%	INVENTED	Specific percentage fabricated
28% miscategorised or lack context	INVENTED	Complement of above
Pre-2021 sources use non-standard timestamps	INVENTED	Plausible for legacy systems
False positive ratio 6:1	INVENTED	Specific ratio fabricated

3. Support Ticket Data

Detail	Status	Source / Reasoning
System: internal ticketing system	SOURCED	`customer_support_lead_samantha_wong.md`
Volume: ~45,000 historical	INVENTED	500+ clients, ~1,200/month over 3+ years
~1,200 new tickets/month	INVENTED	Plausible for 500+ clients
Quality score: 3.0/5	INVENTED	Semi-structured human-entered data
Data owner: Samantha Wong	SOURCED	`customer_support_lead_samantha_wong.md`
Reports to Sarah Thompson	SOURCED	`sarah_thompson_coo.md`
Sensitivity: INTERNAL (some CONFIDENTIAL)	INVENTED	Classification fabricated
Retention: indefinite, no formal policy	INVENTED	No retention policy documented for tickets
Category and priority fields 98% complete	INVENTED	Specific percentage fabricated
Resolution times tracked	SOURCED	`sarah_thompson_coo.md` (4.2hr average resolution)
18% miscategorised priority	INVENTED	Specific percentage fabricated
12% descriptions under 20 words	INVENTED	Specific percentage fabricated
Customer identifier inconsistency (22% account numbers, 31% emails)	INVENTED	All percentages fabricated
No structured root cause field	INVENTED	Plausible gap
Pre-mid-2022 tickets lack satisfaction scores	INVENTED	Fabricated historical gap

4. Customer Records (CRM)

Detail	Status	Source / Reasoning
System: HubSpot CRM	SOURCED	`tom_bradley_marketing_manager.md`
Volume: ~85,000 contacts, ~4,200 companies	INVENTED	Plausible for 500+ active clients plus historical data
500+ active client accounts	SOURCED	`marcell_ziemann_ceo.md`
Quality score: 2.0/5	INVENTED	Design principle: customer-facing data should be messy
Data owner: Lisa Chen (marketing), Sales (pipeline)	SOURCED	`lisa_chen_cmo.md`, `tom_bradley_marketing_manager.md`
Sensitivity: CONFIDENTIAL	INVENTED	Classification fabricated
Retention: indefinite, no hygiene schedule	INVENTED	No data hygiene process documented
2022 CRM migration introduced quality problems	INVENTED	Problems exist (sourced); migration as specific cause is fabricated (see Brief 3 companion)
15% confirmed duplicates	INVENTED	Repo confirms duplicates exist; percentage fabricated
8% probable duplicates	INVENTED	Fabricated
26% incomplete industry classification	INVENTED	Fabricated
34% missing job title/role	INVENTED	Fabricated
11.2% email bounce rate	INVENTED	Fabricated
41% missing lead source attribution pre-migration	INVENTED	Fabricated
Sales staff use personal spreadsheets	INVENTED	Plausible; not confirmed
No integration with billing or support	SOURCED	`jamal_al_sayed_data_analyst.md` (data silos confirmed)

5. Billing and Financial Data

Detail	Status	Source / Reasoning
System: internal billing and invoicing	SOURCED	`aisha_rahman_cfo.md`
Volume: ~6,000 invoices/year	INVENTED	500+ clients, monthly billing
Quality score: 3.5/5	INVENTED	Audit requirements enforce discipline; manual processes introduce errors
Data owner: Aisha Rahman	SOURCED	`aisha_rahman_cfo.md`
Sensitivity: RESTRICTED	INVENTED	Contains payment info; RESTRICTED is appropriate per draft policy
Retention: 7 years	SOURCED	`docs/policies/data_management.qmd`
6% monthly billing discrepancies	INVENTED	Fabricated
Product categorisation changed twice	INVENTED	Products are real (sourced from sales CSV); naming changes fabricated
Products: CloudSync, DataVault, SecureLink, Analytics Pro	SOURCED	`data/cloudcore-sales-data.csv`
Contract terms stored as PDFs	INVENTED	Plausible; not described
Revenue attribution by sector relies on manual spreadsheet	INVENTED	Consistent with data silos; not explicitly described

6. Sales Performance Data

Detail	Status	Source / Reasoning
System: HubSpot + regional spreadsheets	INFERRED	HubSpot confirmed for marketing; spreadsheet use inferred from incomplete sales adoption
6 regions: North, South, East, West, Central, Metro	SOURCED	`data/cloudcore-sales-data.csv`
4 product lines	SOURCED	Same file

Detail	Status	Source / Reasoning
Quality score: 2.5/5	INVENTED	Dual systems, no single source of truth
No single data owner	INFERRED	No sales leader identified as data owner
Sensitivity: INTERNAL	INVENTED	Classification fabricated
~2 years structured data	SOURCED	Sales CSV covers Q1 2023 to Q4 2024
No single source of truth for pipeline	INVENTED	Inferred from incomplete CRM adoption
Regional formats not standardised	INVENTED	Fabricated
Segment definitions vary between systems	SOURCED	`data/cloudcore-customer-data.csv` uses Small/Medium/Large; `data/cloudcore-sales-data.csv` uses Enterprise/SME
35% closed-lost missing loss reason	INVENTED	Fabricated
Sales rep attribution clean from 2023 only	INVENTED	Fabricated

7. HR and Access Control Data

Detail	Status	Source / Reasoning
Systems: Auth0, Active Directory, HR system	SOURCED	`docs/policies/access_control.qmd`, `karen_lee_hr_manager.md`
47 active employees	SOURCED	`marcell_ziemann_ceo.md`
~120 historical records	INVENTED	Plausible given 15% turnover over company history
Quality score: 3.0/5	INVENTED	HR data accurate; access data has drift
Data owners: Karen Lee, Raj Patel (shared)	SOURCED	`karen_lee_hr_manager.md`, `raj_patel_it_manager.md`
Sensitivity: RESTRICTED	INVENTED	Contains employee PII; RESTRICTED appropriate
Employee retention: 7 years post-departure	SOURCED	`docs/policies/data_management.qmd`

Detail	Status	Source / Reasoning
Access logs: 12 months	INVENTED	Not specified
~40% over-provisioned	SOURCED	`karen_lee_hr_manager.md`
RBAC definitions incomplete	SOURCED	Same file
Manual onboarding/offboarding coordination	SOURCED	Same file
Termination: policy 24hrs vs HR 2hrs	SOURCED	`docs/policies/access_control.qmd` vs HR process
Auth0 policies reference Okta	SOURCED	`docs/policies/access_control.qmd`

Data Quality Summary Table

Data Source	Score	AI Readiness	Status of Score	Status of Readiness
Infrastructure telemetry	4.5	High	INVENTED	INVENTED
Security event logs	4.0	Medium-High	INVENTED	INVENTED
Support tickets	3.0	Medium	INVENTED	INVENTED
Billing/financial	3.5	Medium	INVENTED	INVENTED
HR/access	3.0	Low	INVENTED	INVENTED
Sales performance	2.5	Low	INVENTED	INVENTED
CRM	2.0	Low	INVENTED	INVENTED

Pattern observation	Status	Reasoning
Infrastructure/operational data is clean; customer-facing data is messy	INVENTED (as a stated finding)	This is the design principle from the handoff document, presented as an audit finding
Reflects mature engineering vs organisational challenges	INFERRED	Consistent with backstory descriptions

---

Data Value Pyramid Assessment

Level	Placement	Status	Reasoning
Descriptive	Partially achieved	INFERRED	Power BI dashboards confirmed (`jamal_al_sayed_data_analyst.md`); weekly/monthly reporting inferred
Diagnostic	Minimal	INFERRED	Manual RCA for incidents described; no automated correlation
Predictive	Not attempted	SOURCED	CTO confirms no predictive models (`mark_gonzalez_cto.md`)
Prescriptive	Not attempted	INFERRED	No automated decision support mentioned anywhere in repo
Cross-system data needed for prediction	SOURCED	`jamal_al_sayed_data_analyst.md` (data silos prevent cross-system analysis)	
Overall framing as "Level 1 with pockets of Level 2"	INVENTED	Editorial characterisation	

Integration Architecture Assessment

Current Approach

Detail	Status	Source / Reasoning
"Point-to-point with manual bridges" characterisation	INVENTED	Describes the pattern found; not a term used in the repo
No integration middleware, ESB, or API gateway	INFERRED	None of these appear anywhere in the repo

Integration Types Table

Type	Examples	Status
Automated point-to-point: Prometheus to Grafana	SOURCED	Both systems confirmed
Automated: CrowdStrike to Splunk	SOURCED	Both systems confirmed
Automated: GitHub Actions to ArgoCD	SOURCED	`michael_thompson_lead_software_develope`
Batch: usage to billing (daily)	INVENTED	Plausible; no details exist
Batch: support metrics to Power BI (weekly)	INVENTED	Manual Power BI imports confirmed; "weekly" cadence fabricated
Manual: CRM to financial reporting	INVENTED	No integration confirmed (sourced); "manual" process fabricated
Manual: support data to customer health	SOURCED	`jamal_al_sayed_data_analyst.md` (manual correlation described)
Manual: sales data consolidation	INVENTED	Dual systems inferred; "consolidation" process fabricated
API: HubSpot lead capture from website	INFERRED	Website lead forms exist; HubSpot integration standard

Type	Examples	Status
API: Auth0 SSO	SOURCED	`docs/policies/access_control.qmd`

Existing ETL Processes

Detail	Status	Source / Reasoning
No formal ETL platform	INFERRED	No ETL tool appears in repo
Scheduled Python scripts for billing aggregation	INVENTED	Python confirmed as dev stack; ETL use assumed
Manual CSV exports to Power BI	SOURCED	`jamal_al_sayed_data_analyst.md` (manual imports confirmed)
Splunk log collection as only formal ETL	INFERRED	Splunk normalisation is standard; "only formal" is editorial
Prometheus federation	INVENTED	Standard capability; assumed in use
Processes are fragile, undocumented, maintained by individuals	INVENTED	Plausible for 2-person team; not stated
Loss of Jamal or junior would create knowledge gaps	INVENTED	Strongly implied by 2-person team size

Missing Capabilities Table

Capability	Status of Gap	Status of Impact Description
Data warehouse	SOURCED gap	INVENTED impact description
Master data management	INFERRED gap (inconsistent definitions confirmed)	INVENTED impact description
Real-time pipelines	SOURCED gap	INVENTED impact description
API gateway	INFERRED gap (not present in repo)	INVENTED impact description

Capability	Status of Gap	Status of Impact Description
Data catalogue	INFERRED gap (tribal knowledge described)	INVENTED impact description

---

Compliance and Data Handling

Compliance Posture Table

Framework	Status	Provenance	Source
ISO 27001	Certified	SOURCED	`sophia_martines_ciso.md`, `cloudcore_company_overview.md`
ISO controls A.12.1.2, A.14.2.2	Referenced	SOURCED	`docs/policies/change_managemer`
SOC 2 Type II	Compliant	SOURCED	`sophia_martines_ciso.md`, `cloudcore_company_overview.md`
AI systems must meet SOC 2 criteria	Requirement	INFERRED	Standard SOC 2 scope extension
Australian Privacy Act	Compliant	SOURCED	`cloudcore_company_overview.md`
APP 6 and APP 11 most relevant	Mapping	INVENTED	Accurate representation of APPs; application to Cloudcore's AI plans is analysis
NDB scheme	Compliant	SOURCED	`security_compliance_officer_sa`
GDPR	Compliant (EU data)	SOURCED	`emily_chen_head_of_compliance.`
Article 22 automated decision-making	Reference	INVENTED	Accurate GDPR provision; application to Cloudcore is analysis
DPIA required for new initiatives	Reference	SOURCED	DPO interview mentions annual DPIAs
HIPAA	Partial (in progress)	SOURCED	`emily_chen_head_of_compliance.`

Framework	Status	Provenance	Source
BAA coverage needed for AI	Requirement	INFERRED	Standard HIPAA requirement for data processing

Healthcare Client Contract Requirements

Detail	Status	Reasoning
Healthcare is ~25% of revenue	INVENTED	See Brief 2 companion
All contract terms listed	INVENTED	Plausible for Australian healthcare cloud contracts; no contract terms in repo
Data residency in Australian DCs	INVENTED	Common requirement
Logged and auditable access	INVENTED	Common requirement
Prior notification for new systems	INVENTED	Plausible clause
Annual security assessments	INVENTED	Common requirement
24-hour breach notification (stricter than NDB 72hr)	INVENTED	Plausible contractual tightening
AI implication about client notification	INVENTED	Logical consequence of invented contract terms

Finance Client Contract Requirements

Detail	Status	Reasoning
Finance is ~20% of revenue	INVENTED	See Brief 2 companion
All contract terms listed	INVENTED	Plausible for Australian finance cloud contracts
Third-party access requires approval	INVENTED	Common finance sector requirement
AI implication about external platforms	INVENTED	Logical consequence of invented terms

Australian Privacy Act (APP Mapping)

Detail	Status	Reasoning
Privacy Act and APPs referenced in backstories	SOURCED	`security_compliance_officer_samuel_torres.md`, `emily_chen_head_of_compliance.md`
APP 1 (open management) mapped to AI	INVENTED	Accurate APP; application is analysis
APP 3 (collection) mapped to training data	INVENTED	Same
APP 6 (use and disclosure) mapped to purpose limitation	INVENTED	Same
APP 11 (security) mapped to model protection	INVENTED	Same
NDB 72-hour reporting for AI systems	SOURCED (NDB scheme) / INFERRED (AI extension)	NDB is confirmed; extension to AI is logical

AI-Specific Data Impact Assessment Gap

Detail	Status	Source / Reasoning
No AI governance framework	SOURCED	`mark_gonzalez_cto.md`, `sophia_martines_ciso.md`
No AI-specific security review process	SOURCED	`sophia_martines_ciso.md`
Existing DPIA process doesn't address AI	INFERRED	DPIAs exist; AI-specific concerns not mentioned
List of unaddressed AI concerns (bias, explainability, etc.)	INVENTED	Comprehensive but fabricated list of standard AI governance concerns
Data classification policy still in draft	SOURCED	`docs/policies/data_classification.qmd` (POL-DATA-001 v1.2 DRAFT)

Infrastructure Cost Benchmarks

Every cost figure in this section was INVENTED based on typical Australian market rates. The only sourced figure is the ML engineer salary.

Component	Range (AUD/yr)	Status	Reasoning
Cloud data warehouse	$36-72K	INVENTED	Snowflake/BigQuery at moderate scale
ETL platform	$18-36K	INVENTED	Fivetran/dbt Cloud mid-tier
Data catalogue	$12-24K	INVENTED	Commercial tooling
MDM	$24-48K	INVENTED	Implementation-heavy
ML platform	$36-96K	INVENTED	SageMaker/Azure ML, highly variable
ML engineer salary	$180-250K	SOURCED	`karen_lee_hr_manager.md`
AI contractor	$2-3.5K/day	INVENTED	Australian specialist rates
MLOps tooling	$6-18K	INVENTED	MLflow as free alternative noted
API gateway	$6-24K	INVENTED	AWS available through partnership
Event streaming	$18-48K	INVENTED	Managed Kafka at modest scale
Integration middleware	$24-60K	INVENTED	Significant implementation cost

Cost Context Analysis

Detail	Status	Source / Reasoning
$250K budget reference	INVENTED	See Brief 2 companion
Data warehouse + ETL consumes $54-108K	INVENTED	Arithmetic from invented figures
Single ML engineer consumes most/all budget	INFERRED	Sourced salary ($180-250K) vs invented budget ($250K)

Detail	Status	Source / Reasoning
Existing partnerships reduce platform costs	INFERRED	AWS/Azure partnerships confirmed; managed AI access is standard
Existing tools have AI features not being used	INFERRED	Splunk, CrowdStrike, HubSpot all have documented AI capabilities; Cloudcore's use of these features is not mentioned

---

Cross-References

Reference	Status
Policies at cloudcore.eduserver.au/docs/policies/	SOURCED — real pages
Logs at cloudcore.eduserver.au/docs/logs/	SOURCED — real pages
Risk frameworks at cloudcore.eduserver.au/docs/support/risk_assessment_frameworks	SOURCED — real page

---

This reference document is for instructor use. It combines sourced facts and invented details into a single annotated view of the Data Infrastructure Audit Results.