

DATA MINING & PROCESSING • PHILIPPINES

The pipeline-reliability standard: the executive economics of data mining and processing outsourcing to the Philippines.

An analysis of why records processed is a volume vanity metric, how pipeline reliability and output usability — never processing throughput — decide the true cost of a data-pipeline operation once broken jobs, silent corruption, unvalidated output, and re-runs are counted, and the vendor-selection discipline that produces data the business can actually use.

AUTHORS

John Maczynski — Chief Executive Officer
Ralf Ellspermann — Chief Strategy Officer

JULY 2026
MANILA • ESTABLISHED 2001

Contents

–	About the authors	03
01	Executive summary	04
02	The volume mirage: records processed versus reliably processed	05
03	The pipeline contract: build it robust, validate the output, monitor the run	06
04	Cost and performance: cost-per-reliably-processed-record economics and benchmarks	07
05	Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to data pipelines	08
06	Case study: a 44-seat data-pipeline operation re-based on pipeline reliability	09
	The executive playbook: governance for the first 180 days	10
–	Methodology & notes	11

ABOUT THIS SERIES – VOLUME 46

Every data-pipeline proposal is priced per record or per job, because volume is easy to count. This volume is about the only pipeline outcome the business actually depends on: the record processed through a pipeline that runs reliably, with output validated fit-for-use, so no downstream report, model, or decision is quietly built on corrupted data. The gap between processing a record and processing it reliably is where broken jobs and silent corruption live. The full series lives at piton-global.com.

About the authors



John Maczynski

Chief Executive Officer

John has bought data-processing capacity for three decades, and learned that the fast pipeline that processes every record and corrupts a fraction silently is the most expensive one a data team can hire — because a broken job re-runs at a premium, and silent corruption poisons every report and model downstream before anyone notices. His standing question ignores the throughput dashboard: of the records you processed, how many came through a pipeline that ran reliably, with output validated fit-for-use? Applied at PITON-Global, that question is what clients get on every data-pipeline sourcing decision — personally, with no account layer in between.

j.maczynski@piton-global.com
+1 402-598-8740



Ralf Ellspermann

Chief Strategy Officer

Ralf has spent twenty-five years running Philippine data-engineering floors, and he judges a pipeline team by its reliability and output-validation record, not its records-per-hour: whether jobs ran clean and output was validated fit-for-use, or a run broke silently and poisoned everything downstream. The pipeline contract in Section 3 codifies what he learned rebuilding teams that processed records fast and shipped corruption. He now reads data-pipeline providers the way he once ran them — from the job-failure log and the data-validation rate backward.

r.ellspermann@piton-global.com
Manila, Philippines



ABOUT PITON-GLOBAL

PITON-Global is a buyer-side advisory practice, Manila-based since 2001, that has never operated a delivery floor or taken a placement commission. Its stock-in-trade is knowing the Philippine provider market at the operating level — 1,000+ mapped operations, re-verified continuously, including each data-pipeline team's real reliability and output-validation record rather than its records-per-hour headline —

and converting that knowledge into a shortlist of six to ten genuinely qualified partners per engagement. Buyers from Tripadvisor, Garmin, TSYS, HealthPartners, Yetipay, Workrise, and Norman Disney & Young have run their searches through the practice.

SECTION 01

Executive summary

-60%

Fully-loaded cost per reliably-processed record vs. US onshore

99.9%+

Pipeline job-success rate on vetted teams, up from 92-97%

-67%

Reduction in re-runs and downstream data incidents

6.4×

First-year ROI, reconstructed case (Section 6)

A data-pipeline engagement is bought on the price per record or job and paid for on the runs it breaks and the output it corrupts. The vendor reports billions of records processed at a fraction of the onshore cost; the data team finds the jobs that failed silently and were re-run at a premium, the schema drift that slipped through unvalidated, the corrupted output that poisoned a dashboard and a model before anyone noticed, and the engineers firefighting pipelines instead of building them. The operation cleared its records-per-hour and the business inherited unreliable data, a re-run bill, and decisions quietly built on bad numbers. The gap between processing a record and processing it reliably is where the real cost of cheap data processing hides, because a record run through a broken or unvalidated pipeline is not throughput delivered — it is corruption shipped downstream.

The Philippines — with a deep, English-fluent data-engineering talent pool experienced on the modern pipeline stack — is where data mining and processing is delivered with the most pipeline-reliability discipline, because that pool can run the robust engineering, output validation, and monitoring that reliability demands. Five findings:

- 1 Records processed is the industry's most flattering number.** Any pipeline can move records at volume. Vetted teams are measured on outcome: pipeline reliability, output validation, and downstream usability. Weak providers quote a low per-record price the business repays in broken jobs, re-runs, and corrupted data. Section 2 shows the gap.

- 2 The reliable, validated pipeline, not the fast run, is the product.** Data the business can use comes from a pipeline built robust, output validated fit-for-use, and runs monitored — not from moving records at speed. Teams that engineer reliably deliver data you can trust; teams that process loosely ship corruption and re-runs that cost more than the records.
- 3 Silent corruption is worse than a loud failure.** A job that crashes gets fixed; a job that completes with corrupted output poisons every report, model, and decision downstream, undetected, until the damage is done. Vetted teams validate output precisely because the corruption you cannot see is the one that costs the most.
- 4 The re-run and the downstream incident, not the record, are what cost.** A record run through a bad pipeline costs the re-run, the incident response, and the wrong decisions made on corrupted data — many times the per-record saving. Teams that engineer reliably cut re-runs and protect downstream systems, and that protected reliability — not the offshore rate alone — is where the case study's return is built.
- 5 Contract on the reliably-processed record, not the processed one.** The contracting failure of data pipelines is paying per record and hoping the run was clean and the output valid. Vetted deals price against job-success floors, output-validation targets, and re-run ceilings, making pipeline reliability the team's problem, which is where it belongs.

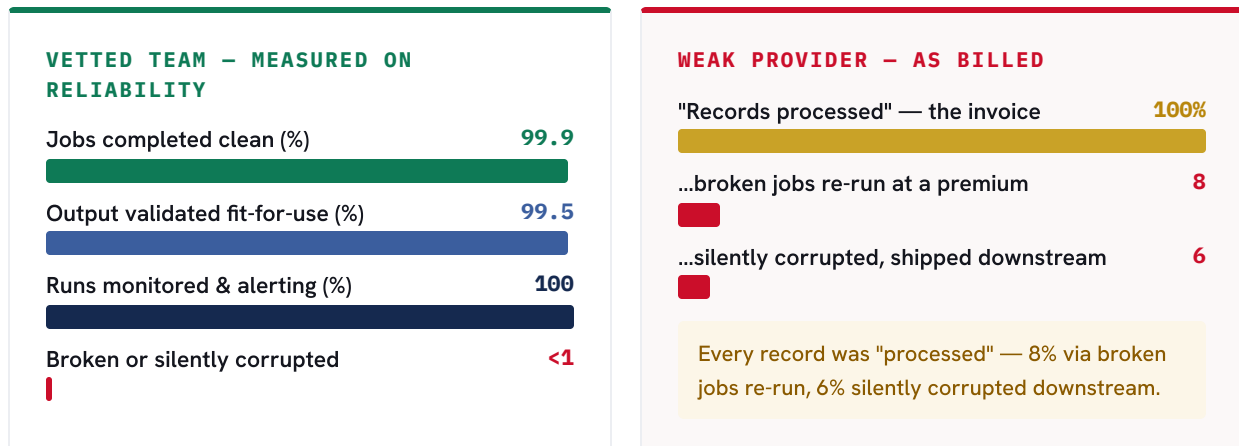
The intended reader is a head of data, VP of engineering, or COO buying data mining and processing per record and paying for it in broken jobs, re-runs, and corrupted downstream data. Sections 2–4 separate the teams that engineer reliably from the teams that move records; Sections 5–7, the method and the proof.

SECTION 02

The volume mirage: records processed versus reliably processed

Figure 1 shows one month of a data team's pipeline runs two ways: as the records-processed report a weak vendor bills, and as the pipeline-reliability reality the validation checks and downstream systems experience. The records-processed number is identical in spirit to a chatbot's containment rate — a figure engineered to look like output while broken jobs, silent corruption, and re-runs accumulate underneath.

FIG. 1 — ONE MONTH OF RUNS: PIPELINE-RELIABILITY QUALITY VS. REPORTED VOLUME



Left: PITON-Global engagement instrumentation, 2025-26, output validated. Right: composite of pre-engagement vendor reporting reviewed in diligence.

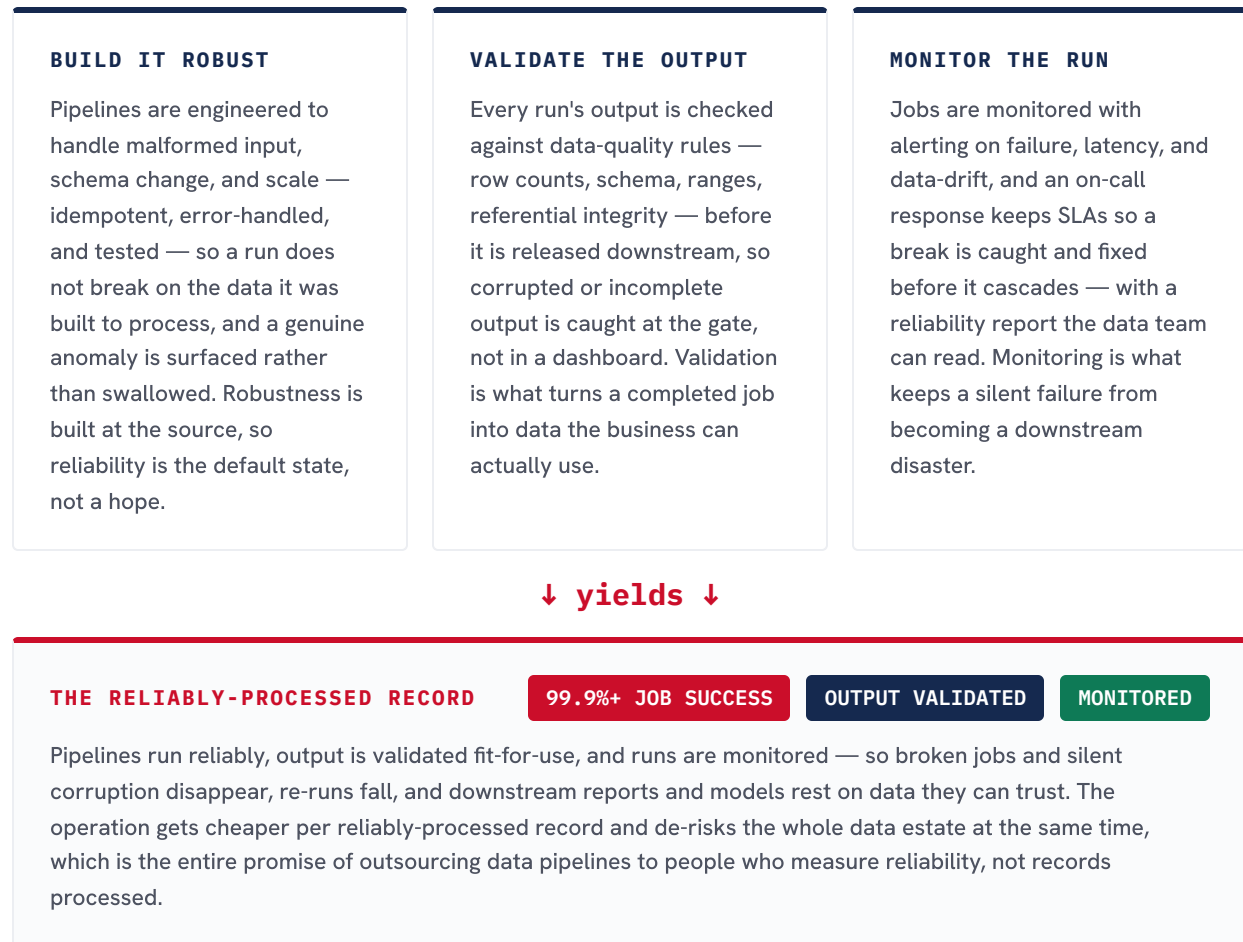
The economics follow the pipeline reliability, not the records count. On the left, jobs complete clean and output is validated fit-for-use, so downstream systems get data they can trust; on the right, a cheap per-record price breaks eight percent of jobs into premium re-runs and ships six percent silently corrupted, poisoning reports and models before anyone notices — paying in re-runs, incident response, and wrong decisions at once. A low per-record price that breaks jobs and corrupts output is not a saving; it is unreliable data and downstream damage, delayed and compounding. The pipeline-reliability audit — job-success rate, output validation, monitoring — is the cheapest due-diligence instrument in this paper.

SECTION 03

The pipeline contract: build it robust, validate the output, monitor the run

Figure 2 lays out the three clauses of a working data-pipeline contract — the machinery a floor read verifies before any per-record price is believed.

FIG. 2 – THE PIPELINE CONTRACT, AS OPERATED ON VETTED TEAMS



Operating parameters from PITON-Global data-engineering engagements, 2025–26.

In diligence, the contract is tested from the output's side: run a representative dataset through and validate the output against data-quality rules to measure true reliability; read the job-failure and re-run logs; audit the monitoring and alerting coverage. Perhaps one team in ten survives it cleanly. Steps 3 and 5 of the framework exist to find that one before the contract does — because everything in Table 3 of Section 6 depends on pipeline reliability, not records processed.

SECTION 04

Cost and performance: cost-per-reliably-

processed-record economics and benchmarks

TABLE 1 – FULLY-LOADED COST PER RELIABLY-PROCESSED RECORD (PER MILLION): THREE OPERATING MODELS (2026)

OPERATING MODEL	COST / MILLION	VS. US BASELINE
US onshore in-house data-engineering	\$78	baseline
Low-cost offshore processing mill (per-record)	\$53	-32%
Vetted Manila team — robust, validated	\$31	-60%

Basis: PITON-Global engagement pricing 2025-26; cost per reliably-processed record = fully-loaded operating cost divided by records processed through clean, validated runs (broken-job and corrupted output excluded); the processing mill looks cheap per record but re-runs, incident response, and downstream corruption lift its true cost per reliable record; US comparator fully-loaded in-house cost. Compute/infrastructure is separately metered; figures reflect the human-engineering operating layer.

TABLE 2 – DATA-PIPELINE BENCHMARKS, VETTED VS. BASELINE (2025-26)

METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
Pipeline job-success rate	99.9%+	92-97%	The only pipeline number trust follows
Output validation pass rate	99.5%+	88-95%	Fit-for-use, or corruption downstream
Re-runs / data incidents vs. baseline	-67%	baseline	"Build it robust" — held or not
Mean time to detect a failure	<5 min	hours-days	Caught at the gate, or in a report
Engineer time on firefighting vs. baseline	-58%	baseline	Building, or firefighting

Source: PITON-Global data-engineering engagement data 2025-26; baseline = typical pre-engagement operations billed on records processed.

The strategic read: the processing mill's -32% per record is a false economy — re-runs, incident response, and downstream corruption lift its true cost per reliable record far above the per-record gap, and a decision made on silently corrupted data is a cost no rate recovers. The vetted Manila team's -60% is real because it is measured where the value is: the record processed through a pipeline that runs reliably, with

output validated fit-for-use. Buy on records processed and you buy the right column of Figure 1; buy on cost-per-reliably-processed-record and you buy pipeline reliability — the same discipline every volume in this series keeps arriving at from a different door.

SECTION 05

Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to data pipelines

Data-pipeline diligence adds one hard requirement to every step: a per-record price is real only when the pipeline reliability behind it survives a validated test run against data-quality rules. The funnel below is representative of a 30–70-seat data-engineering search.

FIG. 3 – THE 7-STEP FRAMEWORK AS A DATA-PIPELINE FUNNEL

STEP	WHAT IT ELIMINATES IN A DATA-PIPELINE SEARCH	FIELD REMAINING
1 Requirements definition	Throughput-first scoping: fixes the job-success floor, the output-validation target, the re-run ceiling, and the monitoring standard the operation must meet by pipeline type	1,000+
2 Market mapping	Processing mills: operations that bill record volume and treat broken jobs and corruption as the client's problem, and generalist floors with no data-engineering or pipeline-tooling depth	~44
3 Capability screening	Teams without the machinery: skilled data engineers, audited reliability history, an output-validation and monitoring workflow, and modern-stack (Spark/Airflow/dbt-class) proficiency across your pipelines	~12
4 Compliance & security audit	Weak data governance: PII protection, access controls on the data stack, lineage and audit trail, breach discipline, SOC 2 scope covering the pipeline environment	~7
5 Operational diligence	The validated test run: run a representative dataset and validate output against data-quality rules for true reliability; read the job-failure and re-run logs; interview on how they handle malformed input and schema drift	4–6
6 Competitive RFP & negotiation	Per-record pricing: finalists bid against job-success floors, output-validation targets, and re-run ceilings — all carrying credits for broken jobs and corrupted output	2–3
7 Transition & launch governance	Cold start: the team rebuilds against your pipelines, data contracts, and stack in a staging environment before it owns production, with the job-success floor certified before cutover, PITON-Global embedded buyer-side	1

Field-remaining counts represent a 30–70-seat data-engineering search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no data floors, runs no pipelines, and takes no placement economics, so the shortlist reflects audited reliability performance, not inventory; **right-sizing** — teams for whom your operation is a flagship account. Six to ten genuinely qualified

providers, delivered within 48–72 hours, free to the buyer, competing through a fully transparent, fair, comprehensive, and highly competitive RFP.

SECTION 06 • ANONYMIZED CASE STUDY

Re-basing the operation on pipeline reliability: a 44-seat data-pipeline rebuild, reconstructed

Engagement DM-060 is a US data-driven enterprise — processing roughly 40 billion records a year across ingestion, transformation, and enrichment pipelines — that had offshored data mining and processing to a per-record mill. The record price was excellent; the reliability was not: the job-success rate sat near 94%, output was going out unvalidated, silent corruption had reached a customer-facing dashboard, and the internal engineers were firefighting instead of building. The head of data wanted the cost without the broken jobs and the corruption. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–8, VALIDATED TEST RUNS)

Requirements fixed a 99.9% job-success floor, a 99.5% output-validation target, and a re-run ceiling, with mandatory monitoring and alerting. Mapping produced 44 claimants; the machinery screen — skilled data engineers, audited reliability, validation-and-monitoring workflow, modern-stack proficiency — left 12; the data-governance-and-security audit left 7. Validated test runs against data-quality rules shortlisted 5. A Manila team won priced against job success and output validation, with a validated re-run written into the reporting spec.

TRANSITION (MONTHS 2–6, REBUILT FIRST)

The team rebuilt against the enterprise's pipelines, data contracts, and stack in staging before owning production; the job-success floor was certified in month 3, and output validation and monitoring stood up from day one. The validation-and-monitoring loop caught failures at the gate. By month 6 the job-success rate reached 99.9%, output validation hit 99.5%, re-runs and data incidents fell 67%, and engineer firefighting dropped 58%. PITON-Global chaired the pipeline-reliability review board to steady state.

TABLE 3 – DM-060 FIRST-YEAR VALUE BUILD-UP (US\$)

VALUE COMPONENT	YEAR 1	BASIS
Processing-cost arbitrage on the 44-seat operation	\$1.58M	Onshore vs. vetted-offshore engineering cost
Re-runs & incident response eliminated	\$0.74M	Premium re-runs and incident cost removed
Downstream-corruption exposure avoided	\$0.40M	Bad-data decisions and rework prevented
Internal-engineer capacity reclaimed	\$0.22M	Engineer hours returned from firefighting
Gross value created	\$2.94M	
Less: engagement & transition costs	(\$0.46M)	Rebuild, validation/monitoring build, staging migration; advisory fee \$0 (provider-paid model)
Net value ÷ cost = first-year ROI	6.4x	\$2.94M ÷ \$0.46M

Figures rounded; reliability, validation, and re-run effects audit-verified. Method in Methodology & notes.

Fifty-four percent arbitrage, forty-six percent from eliminated re-runs, avoided corruption exposure, and reclaimed engineer capacity — a typical split for the series, earned because the operation was re-based on pipeline reliability rather than record volume. Every line traces to a diligence step: the job-success floor to Step 1, the reliability-based pricing to Step 6, the validation-and-monitoring workflow to Step 3's contract. The head of data's line at the year-one review: "The record price barely moved. What moved is that the pipelines just run now — and nothing's silently corrupting a dashboard anymore."

SECTION 07

The executive playbook: governance for the first 180 days

A data-pipeline transition succeeds when the team is measured and paid on pipeline reliability and output validation — client in command throughout.

DAYS 0-30

Define reliability, contract on it

Set the job-success floor, the output-validation target, and the re-run and monitoring thresholds with data leadership before launch, and document the data-quality rules and data contracts by pipeline. Contract against job success and

output validation — never per record alone — with floors and ceilings carrying credits. Stand up the validated test run as the program's arbiter of truth.

DAYS 30–90

Rebuild in staging, prove the floor

Rebuild against your pipelines, data contracts, and stack in staging before the team owns production; certify the job-success floor and the validation-and-monitoring loop. Switch on alerting and on-call from day one. Publish the pipeline dashboard: job-success rate, output-validation pass rate, re-run/incident count, time-to-detect, firefighting hours.

DAYS 90–180

Own production, certify steady state

Cut production over to the team only after it holds the job-success floor and output-validation target. Review the reliability board monthly with the re-run and incident trends on the table. Certify steady state on two months at floor including a peak-load and schema-change cycle; hand over on documented criteria; keep the quarterly validated re-run standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying per record while jobs break silently · moving records instead of engineering robust pipelines · releasing output without validating it · running pipelines with no monitoring or alerting · owning production before reliability is proven. All five are settled at selection and contract.

The data-pipeline argument, in one sentence: a processed record is activity and a reliably-processed one is data the business can trust, and a vetted Philippine team delivers –60% cost per reliably-processed record with a 99.9%+ job-success rate, output validated at 99.5%+, and re-runs cut by two-thirds — for buyers who validate the reliability instead of counting records processed, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can process a record. We make sure you buy the team whose pipelines you can trust.

METHODOLOGY & NOTES

How the numbers were built

Reliability and validation figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global data-engineering engagements, 2025–26. Job-success rate and output validation are measured by running representative datasets and validating output against data-quality rules; re-runs, time-to-detect, and firefighting hours are measured against the client's pre-engagement baseline. The weak-provider column is a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

Cost-per-reliably-processed-record figures (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, pipeline tooling and management/QA overhead including validation and monitoring specialists, attrition-driven recruiting — divided by records processed through clean, validated runs, with broken-job or corrupted output excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. Compute and cloud infrastructure are separately metered and excluded. A reliably-processed record runs through a clean, validated, monitored pipeline.

The case study (DM-060) is anonymized under NDA; volumes and dates are generalized, figures rounded. Reliability, validation, and re-run effects are audit-verified; the re-runs-eliminated line reflects removed premium re-run and incident cost, and the corruption-avoided line is a conservative estimate. $ROI = \text{net first-year value} \div \text{all engagement-related client costs}$. PITON-Global's advisory fee is provider-paid and appears at \$0 in the client cost base; neutrality safeguards are documented at piton-global.com.

Market context draws on IBPAP reporting and PITON-Global's proprietary provider mapping (1,000+ operations; pipeline reliability and output-validation performance re-verified May 2026). Estimates and projections are PITON-Global analysis and should be cited as such. The full series is available at piton-global.com.



Would your provider's record price survive a validated test run?

Thirty minutes with John will tell you. Free to buyers, strictly vendor-neutral — a reliability-verified shortlist within two to three business days.

j.maczynski@piton-global.com
+1 402-598-8740

© 2026 PITON-GLOBAL, INC. · EXECUTIVE WHITE PAPER SERIES WP-46 · DATA MINING & PROCESSING · PHILIPPINES. This publication is provided for general information. Figures identified as estimates or models are PITON-Global analysis. Reproduction with attribution is permitted.