

WEB DATA COLLECTION

OVERVIEW

Titan delivers managed data acquisition (a fully custom pipeline Titan builds and maintains), a dataset marketplace (ready-to-use datasets for immediate access), and real-time data feeds (ongoing for fast-changing sources), and delivers through API, S3, or other cloud storage.

EVALUATION PATH

1	Technical consultation Define sources, fields, refresh frequency, geography, and delivery format with Titan engineers.
2	10 TB evaluation agreement Secure the evaluation window and set up cloud delivery permissions (S3/GCS/Azure). Agreement to launch in 3-5 days.
3	Data delivery Receive the validated dataset with QA report and full technical support before scaling to a recurring feed or production program.

STAKEHOLDER REVIEW

Engineering

The question: *Will the data arrive clean, schema-conformant, and maintained, so I am not rebuilding scrapers?*

Titan's answer: Titan runs the full pipeline: source discovery, collection, anti-bot access, deduplication, validation, enrichment, and normalization against your defined field schema. You receive finished, structured data, not raw HTML to clean.

Proof: Multi-step QA on every dataset so completeness and accuracy are verifiable before ingestion.

The question: *Does it deliver into our stack, at the refresh cadence our sources demand?*

Titan's answer: Datasets are delivered as JSON, CSV, or Parquet by API, S3/GCS/Azure, webhook, or direct ingestion into Snowflake or BigQuery. Engagements scale from a one-time dataset to a recurring feed or real-time firehose on a cadence you define.

Proof: Delivery aligned to your schema feeds on a defined cadence; three product modes: managed acquisition, dataset marketplace, and real-time feeds.

Data

The question: *Does buying data beat managed data collection?*

Titan's answer: Off-the-shelf marketplace datasets cover known needs immediately while managed acquisition handles specific or ongoing requirements. Either way the team works from decision-ready inputs instead of maintaining scrapers.

Proof: Categories span AI/LLM and multimodal training, e-commerce, travel, financial, news, company/B2B, real estate, jobs, and institutional data, validated in production by two of the world's top-10 AI companies.

Finance

The question: *What is the cost model, and how do we contain it?*

Titan's answer: Pricing varies by product type, volume, and delivery frequency, on a consultative model that typically starts with a pilot before a production contract. High-volume record-based datasets carry a low published per-record floor.

Proof: E-commerce product datasets and managed feeds from \$0.005 per record; custom pipelines and large AI datasets scoped per engagement, with a 10 TB evaluation on-ramp. Figures quoted as ranges, confirmed per deal in writing.

Product Management

The question: *Is the data current, complete, and geographically accurate enough to act on?*

Titan's answer: Collection cadence is matched to how fast the source changes, with regional collection for geo-specific accuracy, then deduplicated and enriched before delivery.

Proof: Recurring and real-time feed options; QA report and inventory file for completeness; regional collection across 195 routable countries.

The question: *Is Titan a viable vendor, to add to our vendor pool and how do we de-risk the commitment?*

Titan's answer: Engagements start with a marketplace sample or a low-commitment 10 TB evaluation that validates pipeline quality before any production terms, then scale to a production contract.

Proof: 90%+ proof-of-concept-to-paying conversion; QA report and inventory file on every large delivery; consultative pricing with a pilot period before production.

Privacy & Security

The question: *Where does the data flow, and does it touch our environment unsafely?*

Titan's answer: Collection runs on Titan infrastructure; validated output is pushed one-way into your own cloud bucket. There is no inbound access to your systems, and contributor devices carry no customer data.

Proof: One-directional delivery to customer-controlled S3/GCS/Azure. Node layer routes bandwidth only; no browsing data, files, or credentials traverse it.

The question: *Is the collection lawful, and how is personal data handled?*

Titan's answer: Titan collects only publicly available data and does not collect or process personal user data. Residential IPs are opt-in and consent-based; no personal data is transmitted through the collection layer.

Proof: GDPR-safe posture; public-data-only scope; compliance documentation available to enterprise partners on request; the team works directly with legal and procurement.

BUILD VS BUY

Most enterprise teams already run more than one data provider. The question is which provider earns the next workload. Set against an in-house build and against other vendors, this is where Titan lands.

Dimension	In-house build	Other providers	Titan
Infrastructure	Costly DIY server and pipeline management	Provider-run, generic tooling	Fully managed, elastic scale
IP resources	Fragmented pools, high ban rates	Recycled or limited residential pools	40M+ residential global pool
Long-video reliability	Unstable, partial downloads, corruption	Breaks on long-form media	99.9% completion guarantee
Data quality	Raw, messy HTML formats	Raw or semi-structured output	AI-ready structured JSON
Team focus	2-3 engineers on ops and maintenance	Integration and cleanup still on you	100% focused on ML training
Cost per TB	Headcount plus maintenance	\$2,500 to \$8,000/TB (DIY residential proxy)	Under \$20/TB managed

Pricing is quoted as ranges and confirmed per deal in writing. Published rate scales from an evaluation tier down to under \$20/TB at scale; volumes above 1 PB are custom.

Titan Web Services · titannet.io/web-services/web-data/scrapers/video-scraper · Book a technical consultation. © 2026 Titan Labs LLC. Figures quoted as ranges; final terms confirmed per deal in writing.