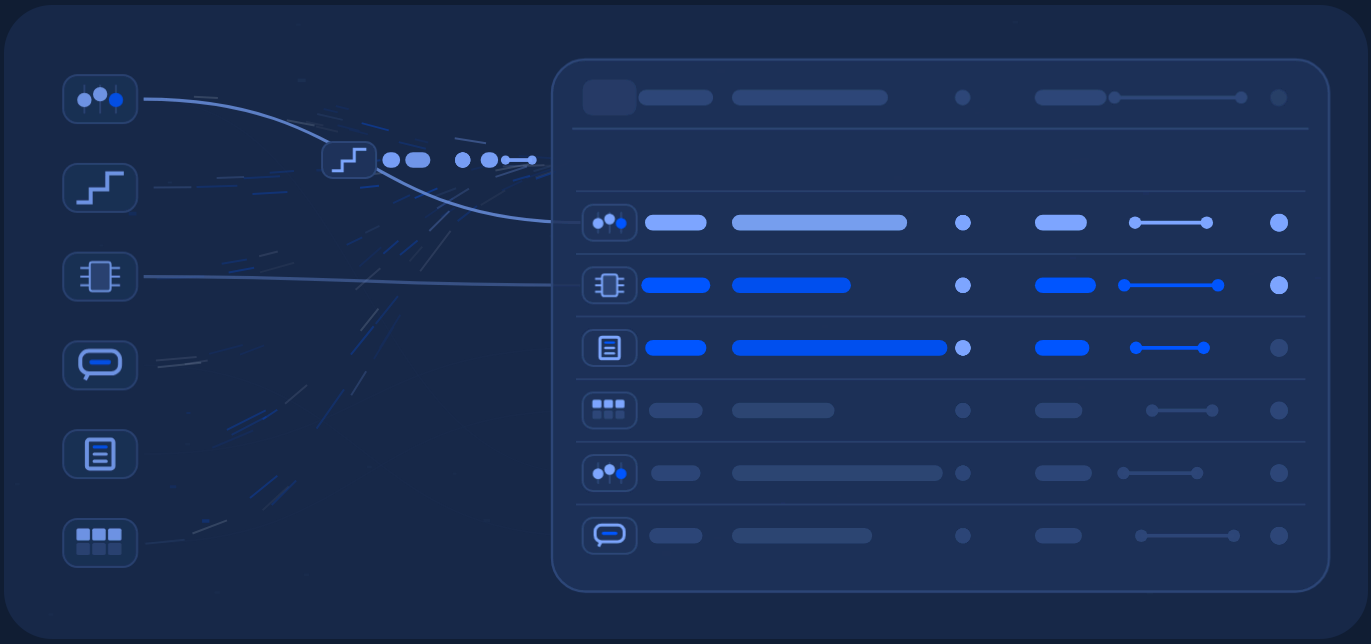
Rillor

EXPLAINER

How data marketplaces for model training work

How a training data marketplace works: listings, samples, prices, licenses and delivery on the large cloud platforms, and the rules for data intermediaries and data brokers.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A training data marketplace lists datasets from providers, sets the terms on which model developers license them and delivers access, and its samples and trials answer a problem Kenneth Arrow described in 1962: a buyer cannot value information before having it [1, 4].

- A data marketplace is a digital platform on which data products are traded, run by an operator that reviews listings, supplies contract terms and bills buyers [2, 6, 7, 8]. A 2022 survey studied 97 of 180 data trading entities in depth and counted 2,015 providers selling in nine marketplaces [3].
- Marketplaces let buyers judge data before paying through metadata, samples and trials. AWS allows up to 10 samples of up to 50 MB per data set, and Snowflake trial listings run for 1 to 90 days [4, 9, 16].
- Prices are often private. In a 2022 survey, 25% of data marketplaces gave no clear public information on pricing, and 57% of the entities that published price information supported fixed prices [3].
- AWS's published listing fee for public offers of data products is 3% of the pre-tax total contract value, and its private offer fees run from 3% down to 1.5% as contract value rises [13].
- AWS's standard Data Subscription Agreement, in its July 2022 template, grants a nonexclusive, nontransferable license, bars distribution to third parties, counts "models" among the Derived Data a subscriber creates and requires removal of the data within 90 days after the term ends [14].
- Delivery takes the form of file copies, APIs or read-only shares in which no data is copied. On AWS, a provider can revoke access to a revision, after which subscribers cannot export or query it [3, 7, 10, 17].
- The EU's Data Governance Act counts data marketplaces among data intermediation services, which are subject to a notification procedure and must run through a separate legal person. The European Commission proposed a voluntary regime on 19 November 2025 [25, 26].
- Sites also license content to crawlers directly. Cloudflare's pay per crawl answers an unpaid request with HTTP 402 and a price, and the Really Simple Licensing (RSL) 1.0 standard of December 2025 lets a site publish machine-readable license and payment terms [31, 33].

Background

What a data marketplace is

Markus Spiekermann's 2019 paper defines a data marketplace as "a digital platform on which data products are traded" [2]. Such a platform, he writes, must act like a neutral intermediary and let anyone, or at least many registered customers, upload and sell data products. It can provide static data or data streams, through file downloads, APIs or web interfaces, and so provides "standardised licensing models, as well as regulations regarding data access and usage" [2]. His taxonomy also sorts marketplaces by openness: closed marketplaces regulate access to the platform, while open ones give up some control over the quality and use of the data [2].

Why data is hard to sell

A 2022 survey notes that data "is non-depletable and hence a non-rivalrous good: selling data to customer A does not prevent the owner from selling it to customer B", so entities trading data aim at making it excludable [3]. A 2019 paper on market design lists four traits that make data hard to sell. It is freely replicable, its value is combinatorial because of its correlation with signal in other data, prediction tasks and the value of accuracy vary widely, and the usefulness of training data is difficult to verify before applying it to a prediction task [5].

Who buys, and what this report covers

A 2023 review of data acquisition names a buyer specific to machine learning: data acquirers, "who need to acquire data for their own ML applications" [4]. It lists varying data file formats as a major challenge in data marketplaces [4].

This report explains how a marketplace for such data works, from the listing to the end of a term, and the rules and crawler licensing schemes around it. What a buyer checks before signing, and what to ask a vendor, are outside its scope. It is written for model developers, data providers and the staff who read their agreements. It is research, never legal advice.

Method

What was examined. Parties and roles, listings and metadata, evaluation before purchase, prices and fees, standard license terms, delivery and revocation, the regulation of intermediaries and data brokers, and licensing schemes for crawlers.

Sources. Thirty-five sources, each opened on 8 November 2026 and read for the passages cited: five economics and research papers, the documentation, policies, fee schedule and standard contract of four cloud data platforms, one open sharing protocol, one dataset hosting hub and one metadata format, EU and California legal texts and agency pages, a U.S. Copyright Office report, three crawler licensing and preference schemes, one acquisition announcement and one annual report.

How. Each mechanism is described in the words of the platform, law or paper that sets it, with the date of the version read. The BigQuery sharing page was read again on 10 November 2026 after its update date changed [22]. Undated platform documentation describes the platform as it stood on 8 November 2026. Statements marked as this report's reading are inferences from the sources cited with them. No platform is ranked.

Weight. The surveys describe marketplaces as their authors found them in 2021 to 2023 [3, 4]. The 2019 market design paper is used only for its account of why data is hard to sell [5]. The Data Subscription Agreement is a template that providers may change [8, 14]. The Digital Omnibus is a proposal [26, 27], the Copyright Office report a pre-publication version [30] and the IETF vocabulary an Internet-Draft [34].

Findings

1. Three parties, and an operator that sets the rules

The 2023 review names the participants: "data providers, who want to sell their data, data acquirers, who need to acquire data for their own ML applications, and sometimes data brokers, who serve as a middleman between data providers and data acquirers" [4]. The cloud platforms give the operator a role of its own. On AWS, a data product is listed in the catalog "after being reviewed by AWS against our guidelines and terms and conditions" [6]. AWS removes any product that breaches those guidelines and may suspend the provider [12]. The guidelines bar information that identifies any person unless it is publicly available or the provider qualifies under the Extended Provider Program, and require sensitive categories to be aggregated or anonymized [12].

Snowflake makes providers "solely responsible for their Products and Listing Information", requires them to "have all the necessary rights to share or sell their Products" and caps each provider at 100 listings [19]. Databricks providers apply through its Data Partner Program, or sign up in the provider console if they share only through private exchanges [20], and its Marketplace carries datasets, models, notebooks, apps and Model Context Protocol (MCP) servers [20]. Listings can be public or private. Snowflake listings can be available "Privately, available only to specific consumers" [16], and Databricks lets a provider share listings only with member consumers, a set-up its documentation calls a "private exchange" [20].

Some operators also screen buyers. Since 8 April 2025, AWS has barred the use of AWS Data Exchange to provide products containing bulk U.S. sensitive personal data or U.S. government-related data to countries of concern or covered persons, as defined in a U.S. Department of Justice final rule [11]. Subscription verification is enabled for every public offer, and the form asks for contact details, country-of-concern status and the "Prospective subscriber's intended use case" [11]. AWS "doesn't review or validate the information provided by a prospective subscriber", and the provider is solely responsible for verifying it [11]. If the provider changes its offer terms after a request, that subscriber keeps the terms as they were at the time of the request [11].

A hosting hub works differently. On Hugging Face, authors can gate a dataset behind access requests that share a user's username and email address with them, approved automatically by default or by hand, with added fields such as a checkbox agreeing to non-commercial use only [23].

The 2022 survey listed 180 data trading entities, studied 97 in depth, including 38 data marketplaces, and counted 2,015 data providers selling in a sample of nine public or semi-private marketplaces [3]. The 2023 review tabulated 14 markets [4]. It set out four properties of an ideal marketplace (budget awareness, price transparency, multiple providers and useful information sharing) and said that "none of the existing data marketplaces satisfy all four properties" [4].

2. Buyers cannot value data before they see it

Arrow's 1962 essay states the paradox: the value of information "for the purchaser is not known until he has the information, but then he has in effect acquired it without cost" [1]. He adds that "if the seller can retain property rights in the use of the information, this would be no problem", and that, given incomplete appropriability, the buyer decides "on less than optimal criteria" [1]. Spiekermann applies the point to data, whose value "cannot be fully disclosed prior to purchase (known as the 'Arrow paradox')" [2].

The 2023 review describes how this plays out. Providers are often unwilling to share their dataset before purchase, and acquirers are unwilling to share their use case or models, for confidentiality. "This creates the biggest challenge in the marketplace, how to evaluate the value of the data with limited information?" [4]. The "most common practice is to share only metadata information about the dataset or a few data samples" [4].

Platforms offer four answers.

- **Samples.** On AWS, a sample is "a small part of the data in your product" meant to show what the whole data set is like. Subscribers can view and download samples before subscribing, providers can upload up to 10 samples of up to 50 MB to each data set, and CSV samples can be previewed [9].
- **Free and outdated data.** The 2022 survey found "most data products in general-purpose DMs are made available for free", and that providers "upload outdated samples of their products so that buyers can play around with them" [3].
- **Trials.** A Snowflake provider can list free access, limited trial access with full access on request, or paid access [16]. A trial offers "a subset of data" or the full product "for a short period of time", for 1 to 90 days, and suits providers who want only certain consumers to purchase "due to licensing agreements, regulatory requirements, or other commercial reasons" [16].
- **Computation that comes to the data.** AWS Clean Rooms ML "allows two or more parties to run machine learning models on their data without the need to share their data with each other" [15]. Its terms name a "Training data provider", the party that contributes the training data [15].

In this report's reading, licenses supply the property rights Arrow named. The 2023 review found the most important usage restriction to be that "the dataset cannot be further sold by the acquirer", because of licensing restrictions [4].

3. Listings carry the metadata a buyer evaluates

In the 2023 review, "each provider gives a description of its own dataset, a pricing mechanism, and potentially a few samples from the dataset" [4]. The cloud platforms set the fields.

- **AWS.** Product details include a name, short and long descriptions, data samples, a logo and support contact information [6]. Listing descriptions must be accurate, give valid contact details and say whether any data was aggregated or anonymized [12].

- **Snowflake.** A listing adds metadata to a share, "such as a title, description, sample SQL queries, and information about the data provider" [16]. Dataset listings must give "accurate descriptions of the update frequency, the geographical scope, region availability, and the completeness of the fields in the tables/views", and the documentation link must give additional information about the product, such as the "sources and methodology utilized for compiling" it [19].
- **Hugging Face.** A dataset card's metadata describes its license, language and size, and a listed license keyword is displayed on the dataset page [23].

Shared formats make such fields readable by machines. MLCommons released Croissant on 6 March 2024 as "a metadata format to help standardize machine learning datasets", supported by Hugging Face, Google Dataset Search, Kaggle and OpenML [24]. Its vocabulary extends schema.org, which MLCommons says is used by over 40 million datasets on the web [24]. The 2023 review names Croissant as an industry answer to varying file formats [4].

4. Prices: fixed, subscription or per use, and often private

Spiekermann's taxonomy names free data, fixed prices or subscriptions, package prices, pay-per-use, which "sets a price for each unit consumed", and progressive prices, which rise "the more customers purchase a license to use the data set" [2]. The 2023 review groups transactions into one-time upfront, query-based and subscription pricing [4].

Observed practice is less transparent. In the 2022 survey, "25% of DMs do not provide any clear explicit public information about how data pricing works on their platform" [3]. Of the entities that published price information, 57% supported fixed prices [3]. Other mechanisms included volume-based and usage-based prices, bids by buyers, free data and customized prices, where "Price is set by the seller case by case, and depends on who the buyer is and what the data is intended to be used for" [3]. The 2023 review found that "in the majority of the markets prices are not advertised publicly and contacting sales is required" [4].

On the cloud platforms the structure of offers is public.

- **AWS.** An offer sets the subscription agreement, price and duration combinations, U.S. sales tax, a refund policy, subscription verification and auto-renewal [8]. Durations run from 1 to 36 months, a public offer can list up to 5 durations, and the only pricing currency is the U.S. dollar [8]. A subscriber chooses one price and duration, agrees to the terms and pays upfront [8]. A provider can extend a custom offer to a select AWS account with its own terms and pricing [8].
- **Snowflake.** Usage-based plans charge per query or a monthly fee, billed in arrears in months where usage occurs, and subscription plans charge for a term, billed upfront [18]. A per-query plan must include a maximum total charge per month, after which further usage, such as queries, is free [18]. Offers are specific to each consumer, with individualized billing, payment terms and contract dates [16].

Operators take a share. Marketplaces "usually calculate a commission for each data transaction", with fees for membership, listing, storage space or data services [2], and the 2022 survey found time-based subscriptions, revenue sharing and one-off connection fees [3]. AWS Marketplace's fee schedule, effective 5 January 2024, is the one this report opened [13].

AWS Marketplace offer	Listing fee	Source
Public offer, AWS Data Exchange product	3%	[13]
Private offer, total contract value under $1M	3%	[13]
Private offer, $1M to under $10M	2%	[13]
Private offer, $10M or more	1.5%	[13]
Renewals	1.5%	[13]

Table 1. Listing fees on the pre-tax total contract value. Regional listing fees apply in addition for buyers in some jurisdictions [13].

5. The license travels with the listing

The Data Subscription Agreement (DSA) "is the standard contract template that AWS Data Exchange offers as the default", and "As a provider, you control the legal terms and usage rights" [8]. The DSA covers use, warranty, indemnification and governing law, and providers may edit it or upload their own [8]. The template dated 14 July 2022 in its file name sets out these terms [14].

- **Grant.** A "nonexclusive, worldwide, nontransferable license to receive, retain, use, and modify the Data and to create Derived Data using the Data", with the provider retaining "all rights not expressly granted" (section 2.1) [14].
- **Restrictions.** No publishing or distributing the data "or any material subset thereof, to any third party", no sublicensing, no acting as "a consultant, service bureau, or application service provider", and, for data that does not identify individuals, no use "to create, generate, or infer any information relating to the identity of an individual" (section 2.3) [14].
- **Derived Data.** Materials "derived from or created by or for Subscriber using the Data, including data analytics, reports, research, analysis, tools, notes, presentations, discussions and/or models, calculations, algorithms or statistical methods, but excluding any unmodified Data" (section 11.7) [14]. The subscriber represents and warrants that it owns the Derived Data, apart from the provider's data within it, and nothing in the agreement gives the provider rights in it (section 3.3) [14].

- **End of term.** The subscriber removes the data, "within ninety (90) calendar days following such termination or expiration", from the AWS infrastructure it used "and any other computer systems operated by or for Subscriber, and if instructed by Provider, destroy all other copies of the Data" (section 8.5.1) [14].
- **Personal data.** The provider warrants that data identifying individuals "has already lawfully been made available to the general public" (section 4.1) [14].
- **Order of precedence.** An amendment agreed by the parties, then the agreement, then the offer (section 1.3) [14].

This report quotes the template and does not interpret how its terms apply to a trained model under any contract.

Snowflake places the terms in the listing, whose information should cover "the applicable costs and Listing Terms, including any use restrictions, license grants, and other terms and conditions covering a Consumer's use of their Products" [19]. A provider that shares anonymized or aggregated data must keep it anonymous "even when combined with additional information" [19]. On Databricks, some products are available as soon as a consumer requests them and agrees to the terms, while others need provider approval and a transaction through the provider's own interfaces [20].

6. Delivery: copies, shares and revocation

Data is often delivered "as a bulk download or through APIs" [3]. The cloud platforms add access in place.

Platform	How access is delivered	Where the data sits	Controls named in the documentation	Sources
AWS	Files, APIs, Amazon Redshift, Amazon S3, AWS Lake Formation (Preview)	Files: a copy for the subscriber. S3 data access: the provider's own objects	Revision access rules, revocation of a revision	[6, 7, 10]
Snowflake	Share, Native App or Connected App	Provider's account, no data copied or transferred	Shared objects read-only	[17, 19]
Google BigQuery sharing	Linked dataset or linked Pub/Sub subscription	Publisher's project, shared without replicating	Egress options block copy, clone, export and snapshot	[22]
Databricks (OpenSharing)	Open protocol across computing platforms	Provider's cloud dataset, read through pre-signed file URLs or temporary credentials	Bearer tokens for recipients	[20, 21]

Table 2. Delivery modes on four platforms, from their documentation viewed 8 November 2026.

On AWS, Redshift products and AWS Lake Formation (Preview) give read-only access, and S3 data access gives subscribers "the same Amazon S3 objects that the data provider maintains" [7].

Snowflake's Secure Data Sharing copies nothing: "no actual data is copied or transferred between accounts", and the consumer pays only for the compute used to query the shared data [17]. "All database objects shared between accounts are read-only" [17]. Google describes BigQuery sharing, formerly Analytics Hub, as a way to share and access data across organizations "without replicating data" [22]. A subscription creates "a read-only BigQuery dataset" in the subscriber's project, and publishers can set egress options under which "Copy, clone, export, and snapshot APIs are unavailable" [22]. Databricks describes OpenSharing as "an open protocol developed by Databricks for secure data sharing with other organizations" that "works regardless of the computing platforms those organizations use" [20]. In its protocol, a recipient is "a principal that has a bearer token to access shared tables", and the server returns pre-signed URLs for data files or issues temporary cloud credentials [21].

Access over time is set per listing. An AWS provider chooses whether a subscriber receives all historical revisions, a fixed number of trailing revisions from 1 to 100, or none, and whether future revisions during the subscription are included, though it cannot choose no historical and no future revisions together [10]. A provider "can revoke subscriber access to a specific revision at any time", typically for compliance reasons. Revoking does not delete the underlying assets, and "Subscribers can't export or query the data within a revoked revision" [10].

In this report's reading, revocation stops access through the platform, while a copy already exported falls under the agreement, which in AWS's template means its end-of-term removal clause [10, 14]. A training pipeline that needs files outside the platform needs a listing whose terms and egress controls allow export [22].

7. Rules for intermediaries and data brokers

The EU's Data Governance Act, Regulation (EU) 2022/868, applies from 24 September 2023 [25]. It defines a data intermediation service as one "which aims to establish commercial relationships for the purposes of data sharing between an undetermined number of data subjects and data holders on the one hand and data users on the other, through technical, legal or other means" (Article 2(11)) [25]. Recital 28 gives "data marketplaces on which undertakings could make data available to others" as an example [25].

Two exclusions matter for training data. The definition leaves out "services that obtain data from data holders and aggregate, enrich or transform the data for the purpose of adding substantial value to it and license the use of the resulting data to data users, without establishing a commercial relationship between data holders and data users", and "services that focus on the intermediation of copyright-protected content" [25].

Listed services are subject to a notification procedure and must comply with Article 12 (Article 10) [25]. Among its conditions:

- The provider may not use the intermediated data for purposes other than putting it at the disposal of data users, and must provide the service "through a separate legal person" [25].
- Pricing may not depend on whether a data holder or user uses the provider's other services [25].
- Access must be "fair, transparent and non-discriminatory", including on prices and terms of service [25].
- The provider must keep a log record of the intermediation activity [25].

The Commission keeps a public register (Article 11) [25]. Its page describes intermediaries as "neutral third parties that connect individuals and companies with data users", which may charge for facilitating data sharing but "cannot directly use the data that they intermediate for financial profit" [28]. On 8 November 2026, the register listed 40 providers in 10 member states, by this report's count, on a page last updated on 3 November 2026 [28].

On 19 November 2025 the Commission proposed, in its Digital Omnibus, to repeal Regulation (EU) 2022/868 and to transform the mandatory regime for data intermediation services "into a voluntary, trust-enhancing regime" in the Data Act [26]. Its recital 8 says the compulsory regime "has created unnecessary compliance costs" and proposes a functional separation in place of the legal one [26]. On 8 November 2026, the European Parliament's procedure file showed the stage reached as "Awaiting committee decision" [27].

In California, a business is a data broker if it knowingly collects and sells to "third parties the personal information of a consumer with whom the business does not have a direct relationship" [29]. Data brokers register each year, for a fee the agency's page gave on 8 November 2026 as $6,000 plus a third-party processing fee [29]. Beginning 1 August 2026, they must access the state's "accessible deletion mechanism" at least once every 45 days and process consumer deletion requests [29]. This report describes what the definition says and does not say who meets it.

8. Licensing content for crawlers, and direct deals

Licensing for training also happens outside dataset marketplaces. The Copyright Office's May 2025 pre-publication report records licensing across fields, "as well as various content aggregators offering or facilitating collective licensing of training materials" [30]. It concludes that "at this point in time, the Office recommends allowing the licensing market to continue to develop without government intervention" [30].

Three schemes put licensing at the point of crawl.

- **Pay per crawl.** Each time a crawler requests content, it either presents payment intent in request headers for an HTTP 200 response, or receives an HTTP 402 Payment Required response with pricing, and Cloudflare "acts as the Merchant of Record" [31]. Site owners can set a flat per-request price and choose, per crawler, to allow, charge or block [31]. The documentation, last updated 28 July 2026, says pay per crawl "is currently in closed beta" [31].

- **Really Simple Licensing.** RSL 1.0, published on 10 December 2025 with the status "Recommendation", is "an open, XML-based standard" for machine-readable usage, licensing and legal terms, discovered through robots.txt, HTTP headers, RSS feeds and HTML elements [33]. Its usage tokens include `ai-train`, for training or fine-tuning models [33]. Its payment types include a one-time purchase, a subscription, a payment for each training use, a payment per crawl and a payment each time content contributes to an output [33].
- **Usage preferences.** The IETF Internet-Draft draft-ietf-aipref-vocab-08, dated 14 September 2026 and intended for the Standards Track, defines "a vocabulary for expressing preferences regarding how digital assets are used by automated processing systems", with a training category labelled `train-ai` [34]. Where legal agreements explicitly permit a use, "those arrangements likely apply despite the existence of machine-readable statements of preference" [34].

On 15 January 2026 Cloudflare announced it had acquired Human Native, a data marketplace founded in 2024 that connects creators with developers [32].

Direct licensing runs beside all of these. Shutterstock's annual report for 2025, filed on 17 February 2026, says the company has "seen increased demand for access to our metadata for machine learning and generative artificial intelligence model training" [35]. Its Data, Distribution, and Services revenues rose 16% to $203.3 million in 2025, mainly from the sale and delivery of metadata and from growth in distribution and services [35]. The offering also includes Giphy distribution and advertising and Shutterstock Studios services, so the figure is not a licensing total [35].

Implications

For model developers licensing data. Prices are often private and samples small [3, 4, 9], so plan early for a sales conversation, a trial or a sample request. Fit the delivery mode to the pipeline: a read-only share with export switched off suits work inside the platform, while training on files outside it needs terms and controls that allow export [17, 22]. Read what the agreement counts as derived work and what happens at the end of the term [14]. On AWS, expect a verification form that asks for the intended use case [11].

For data providers. The operator sets a listing's fields: descriptions, samples, update frequency, scope, completeness and sources [6, 12, 19]. Revision access rules decide what history and updates a subscriber gets [10]. The standard contract is a starting point the provider controls [8], and on AWS the provider carries the duty to check what subscribers say [11].

For legal and procurement staff. On AWS, a subscriber's terms freeze at the time of its verification request [11], and an amendment outranks the agreement and the offer in the template [14]. Whether a marketplace is a data intermediation service turns on its role between data holders and users, with aggregators that license enriched data outside the definition [25], and the regime is subject to the November 2025 proposal [26, 27]. California's definition turns on selling personal information of consumers the business has no direct relationship with [29].

Limits of this analysis

- This report is research on how marketplaces work, never legal advice.
- Most platform documentation is undated and describes the platforms as viewed on 8 November 2026.
- Fees come from AWS's schedule only, and the only standard contract read is AWS's July 2022 template [13, 14].
- The survey counts come from 2021 to 2023 [3, 4]. No source opened gives current counts of listings, providers or buyers, or a size for the training data market.
- The register count is this report's count of a page updated on 3 November 2026 [28].
- The Digital Omnibus was a proposal awaiting a committee decision on 8 November 2026 [26, 27].
- Adoption of pay per crawl, RSL and the IETF vocabulary was not measured [31, 33, 34].
- Buyer screening was examined on AWS only [11].

Sources

1. Economic Welfare and the Allocation of Resources for Invention. Kenneth J. Arrow, in The Rate and Direction of Inventive Activity: Economic and Social Factors, Princeton University Press for the National Bureau of Economic Research, 1962, pp. 609 to 626, viewed 8 November 2026. https://www.nber.org/books-and-chapters/rate-and-direction-inventive-activity-economic-and-social-factors/economic-welfare-and-allocation-resources-invention

2. Data Marketplaces: Trends and Monetisation of Data Goods. Markus Spiekermann, Intereconomics, Vol. 54, No. 4, 2019, pp. 208 to 216, viewed 8 November 2026. https://www.intereconomics.eu/contents/year/2019/number/4/article/data-marketplaces-trends-and-monetisation-of-data-goods.html

3. A Survey of Data Marketplaces and Their Business Models. Santiago Andrés Azcoitia and Nikolaos Laoutaris, SIGMOD Record, Vol. 51, No. 3, September 2022, arXiv:2201.04561, 11 January 2022, viewed 8 November 2026. https://arxiv.org/abs/2201.04561

4. Data Acquisition: A New Frontier in Data-centric AI. Lingjiao Chen, Bilge Acun, Newsha Ardalani and others, arXiv:2311.13712, 22 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2311.13712

5. A Marketplace for Data: An Algorithmic Solution. Anish Agarwal, Munther Dahleh and Tuhin Sarkar, ACM Conference on Economics and Computation (EC '19), arXiv:1805.08125, version 4, 12 May 2019, viewed 8 November 2026. https://arxiv.org/abs/1805.08125

6. What is AWS Data Exchange? AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/what-is.html

7. Providing AWS Data Exchange data products on AWS Marketplace. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/providing-data-sets.html

8. Creating an offer for AWS Data Exchange products. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/prepare-offers.html

9. Sample data in AWS Data Exchange. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/samples-pro.html

10. Revision access rules in AWS Data Exchange, and Revoking access to revisions in AWS Data Exchange. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/best-practices-revisions.html and https://docs.aws.amazon.com/data-exchange/latest/userguide/revoking-revisions.html

11. Subscription verification for providers in AWS Data Exchange. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/subscription-verification-pro.html

12. Publishing guidelines for AWS Data Exchange. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/publishing-guidelines.html

13. Understanding listing fees for AWS Marketplace sellers. AWS Marketplace Seller Guide, Amazon Web Services, fees effective 5 January 2024, viewed 8 November 2026. https://docs.aws.amazon.com/marketplace/latest/userguide/listing-fees.html

14. Data Subscription Agreement for AWS Marketplace. Amazon Web Services, version of 14 July 2022, viewed 8 November 2026. https://aws-mp-standard-contracts.s3.amazonaws.com/Data-Subscription-Agreement-for-AWS-Marketplace-2022-07-14.pdf

15. AWS Clean Rooms ML. AWS Clean Rooms User Guide, Amazon Web Services, undated, viewed 8 November 2026. https://docs.aws.amazon.com/clean-rooms/latest/userguide/machine-learning.html

16. About listings. Snowflake Documentation, Snowflake Inc., undated, viewed 8 November 2026. https://docs.snowflake.com/en/collaboration/collaboration-listings-about

17. About Secure Data Sharing. Snowflake Documentation, Snowflake Inc., undated, viewed 8 November 2026. https://docs.snowflake.com/en/user-guide/data-sharing-intro

18. Paid listings pricing models. Snowflake Documentation, Snowflake Inc., undated, viewed 8 November 2026. https://other-docs.snowflake.com/en/collaboration/provider-listings-pricing-model

19. Provider and consumer policies. Snowflake Inc., undated, viewed 8 November 2026. https://www.snowflake.com/provider-and-consumer-policies/

20. What is Databricks Marketplace? and What is OpenSharing? Databricks Documentation, Databricks, last updated 11 September 2026 and 27 October 2026, viewed 8 November 2026. https://docs.databricks.com/aws/en/marketplace/ and https://docs.databricks.com/aws/en/delta-sharing/

21. Delta Sharing Protocol (PROTOCOL.md). Delta Lake project, GitHub repository delta-io/delta-sharing, last changed 5 August 2026, viewed 8 November 2026. https://github.com/delta-io/delta-sharing/blob/main/PROTOCOL.md

22. Introduction to BigQuery sharing. Google Cloud Documentation, Google, last updated 7 November 2026, viewed 10 November 2026. https://cloud.google.com/bigquery/docs/analytics-hub-introduction

23. Gated datasets, and Dataset Cards. Hugging Face Hub documentation, Hugging Face, gated datasets page last changed 23 October 2026, viewed 8 November 2026. https://huggingface.co/docs/hub/datasets-gated and https://huggingface.co/docs/hub/datasets-cards

24. New Croissant Metadata Format helps Standardize ML Datasets. MLCommons, 6 March 2024, viewed 8 November 2026. https://mlcommons.org/2024/03/croissant_metadata_announce/

25. Regulation (EU) 2022/868 (Data Governance Act). European Parliament and Council of the European Union, 30 May 2022, Official Journal L 152, 3 June 2022, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R0868

26. Proposal for a Regulation (Digital Omnibus), COM(2025) 837 final, 2025/0360(COD). European Commission, 19 November 2025, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52025PC0837

27. Procedure file 2025/0360(COD), Digital Omnibus. European Parliament, Legislative Observatory, undated procedure page, viewed 8 November 2026. https://oeil.secure.europarl.europa.eu/oeil/en/procedure-file?reference=2025/0360(COD)

28. EU register of data intermediation services. European Commission, Shaping Europe's digital future, last updated 3 November 2026, viewed 8 November 2026. https://digital-strategy.ec.europa.eu/en/policies/data-intermediary-services

29. Information for Data Brokers. California Privacy Protection Agency, undated web page, viewed 8 November 2026. https://cppa.ca.gov/data_brokers/

30. Copyright and Artificial Intelligence, Part 3: Generative AI Training, pre-publication version. U.S. Copyright Office, May 2025, viewed 8 November 2026. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

31. What is Pay Per Crawl? Cloudflare documentation, last updated 28 July 2026, and Introducing pay per crawl, Cloudflare blog, 1 July 2025, viewed 8 November 2026. https://developers.cloudflare.com/ai-crawl-control/features/pay-per-crawl/what-is-pay-per-crawl/ and https://blog.cloudflare.com/introducing-pay-per-crawl/

32. Cloudflare Strengthens Content Offering to AI Companies with Acquisition of Human Native. Cloudflare, press release, 15 January 2026, viewed 8 November 2026. https://www.cloudflare.com/press/press-releases/2026/cloudflare-strengthens-content-offering-to-ai-companies-with-acquisition-of-human-native/

33. Really Simple Licensing (RSL) 1.0 Specification, RSL-SPEC-1.0. RSL Technical Steering Committee, 10 December 2025, viewed 8 November 2026. https://rslstandard.org/rsl

34. A Vocabulary For Expressing AI Usage Preferences, draft-ietf-aipref-vocab-08. P. Keller and M. Thomson (Ed.), IETF AI Preferences Working Group, Internet-Draft, 14 September 2026, viewed 8 November 2026. https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/

35. Shutterstock, Inc. Form 10-K for the fiscal year ended 31 December 2025. U.S. Securities and Exchange Commission (EDGAR), filed 17 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1549346/000154934626000008/sstk-20251231.htm

About this report

Published by Rillor on 11 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How data marketplaces for model training work. 11 November 2026. https://rillor.com/insights/how-data-marketplaces-for-model-training-work

Work with Rillor

Send the datasets you need, the coverage period, the delivery format and the intended use. Contact us at rillor.com/contact?type=market-data or contact@rillor.com. More on data marketplace: rillor.com/data/marketplace