

# Buying the *Front Door*

*NVIDIA has paid \$12.93 billion for Hugging Face - the largest deal in its history, and the first time the company that sells the world its compute has bought the place developers go to decide what to run on it.*

NVIDIA already owned the shovels. It has now bought the map room. Whether that makes open-source AI bigger or narrower is the only question that matters - and it will not be answered for years.

# The compute company bought the community.

- **The deal.** On 3 September 2026 NVIDIA agreed to acquire Hugging Face for **\$12,930,300,000** - roughly \$11.9bn to shareholders and up to \$1bn in stock retention for about 750 employees. It is NVIDIA's largest acquisition ever, nearly double the \$6.9bn Mellanox deal. It needs regulatory approval and is expected to close in the first half of 2027.
- **What was bought.** Hugging Face is the default public repository for open AI models - a GitHub for machine learning. NVIDIA's announcement puts it at **18 million developers, 3 million models, 500,000 datasets and over 200,000 companies**. Reported revenue is about \$150m a year, making this an ~86× multiple. The price is not about the revenue.
- **Why NVIDIA wants it.** Every large closed lab - OpenAI, Google, Amazon, Anthropic - is now designing its own silicon. Open models are the counterweight: they spread AI capability to hundreds of thousands of organisations that will never build a chip, and those buyers run commodity GPUs. **Fragmented demand favours the incumbent with the best toolchain.**
- **The pledge.** NVIDIA has committed publicly that the platform stays open - multi-cloud, multi-accelerator, continued AMD, Intel and Google TPU support, brand and founders retained, and **NVIDIA hardware not required** to build or deploy through it.
- **The gap in the pledge.** It covers availability, not prominence. Nothing is said about search ranking, default deployment routing or which model is surfaced first. There is no stated duration, no governance mechanism and no committed spend on rival backends - and those are the levers that actually move developer behaviour.
- **What happens next.** This is a direct acquisition, not the licence-plus-equity structure NVIDIA used for Groq and Poolside, so it triggers mandatory US and EU merger review. For users, nothing changes on day one. The real question is what the defaults look like in 2028.

**\$12.93B**

total deal value,  
largest in NVIDIA's history

**86×**

multiple on reported  
annualised revenue

**18M**

developers on the  
platform acquired

**H1 '27**

expected close,  
subject to approval

# What Actually Happened

Buyer: **NVIDIA** | Target: **Hugging Face** | Announced: **3 Sep 2026** | Status: **pending regulatory review**

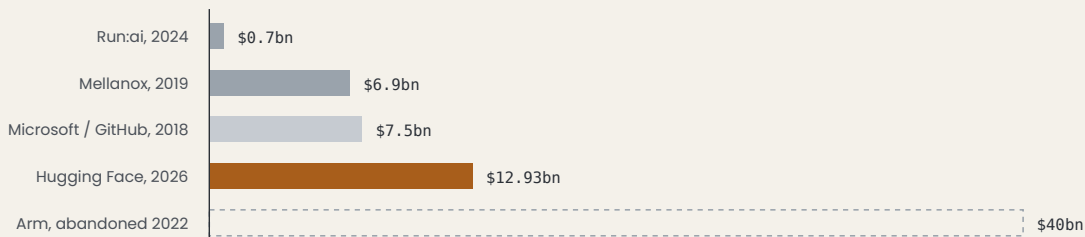
Hugging Face was founded in New York in 2016 by Clément Delangue, Julien Chaumond and Thomas Wolf. It began as a chatbot company, open-sourced its machine-learning tooling, and the tooling ate the company. Its transformers library became the standard way to load a model; its Hub became the standard place to put one. By 2026 it was less a product than a piece of public infrastructure that happened to be privately owned.

NVIDIA is the company whose chips train and serve almost all of that software. In its most recent quarter it reported \$96.2bn of revenue, of which \$89.0bn - about 92% - came from data centre, up 117% year on year. Against that, \$12.93bn is roughly six weeks of sales.

## HOW IT CAME TOGETHER - A SHORT TIMELINE

| WHEN        | WHAT HAPPENED                                                                                                                                                                                      |
|-------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2016        | Hugging Face founded in New York.                                                                                                                                                                  |
| Aug 2023    | The two companies partner on DGX Cloud training. Weeks later Hugging Face raises \$235m led by Salesforce Ventures at a \$4.5bn valuation - with Google, Amazon, IBM and NVIDIA all participating. |
| Late 2025   | <b>Hugging Face rejects</b> a reported \$500m investment from NVIDIA at a \$7bn valuation.                                                                                                         |
| Jul 2026    | Hugging Face discloses a cyberattack carried out by autonomous AI agents - reported as the first documented multi-stage intrusion of its kind.                                                     |
| Summer 2026 | Delangue concludes the platform needs more compute, scale and visibility, and approaches Jensen Huang directly.                                                                                    |
| 26 Aug 2026 | Reports of an agreed \$12.9bn deal surface. NVIDIA stays silent, which is itself unusual.                                                                                                          |
| 3 Sep 2026  | NVIDIA confirms: \$12,930,300,000. Shares rise about 2%.                                                                                                                                           |

## THE SHAPE OF NVIDIA'S M&A - WHAT \$12.93BN LOOKS LIKE



Largest completed acquisition in company history - and the first aimed at developers rather than hardware. GitHub shown for comparison; the Arm bid collapsed under regulatory pressure in 2022.

### THE PRICE IS, LITERALLY, A JOKE

Co-founder Thomas Wolf has noted the figure is not arbitrary: **129,303** is the decimal value of a binary string encoding references to both companies. A small thing, but it tells you the tone of the transaction. This was a negotiated marriage, not a rescue.

# Why a Repository Is Worth \$13 Billion

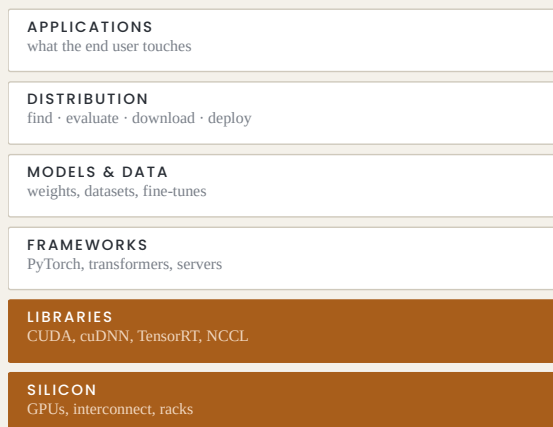
To see what NVIDIA bought, it helps to strip modern AI down to four things. None of this is technical.

A **model** is a large file of numbers - "weights" - produced by training. Feed it text or an image and it produces an output. The file itself is inert. **Compute** is the hardware that makes the file do something: chips that perform the arithmetic, both during training (expensive, occasional) and during inference (cheaper per run, but constant, and therefore where the money eventually lives). **Data** is what the model was trained on. And a **platform** is the connective tissue: the place you find a model, check whether it is any good, download it, and wire it into an application.

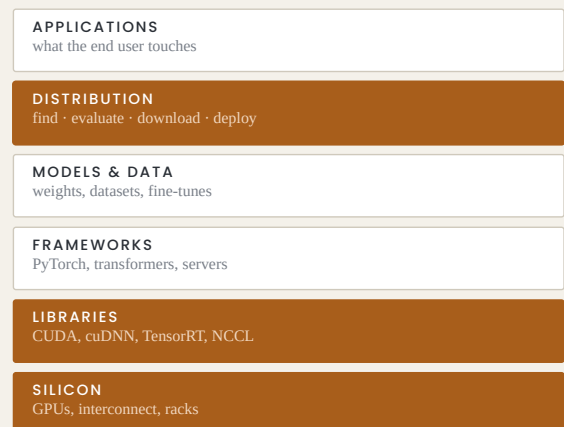
"Open-weight" models are those whose weight files are published for anyone to download. They are not the same as open-source software - training data and code are often withheld - but the practical effect is real: you can run the model on your own hardware, inspect it, modify it, and never send your data to anyone else's API.

## THE STACK, AND WHERE EACH COMPANY SITS

### BEFORE - NVIDIA OWNS THE BOTTOM



### AFTER - AND THE CHOICE LAYER



Ownership shown in orange. The only layer that changed hands is the one where a developer chooses.

NVIDIA's durable advantage has never been the chips alone. It is **CUDA** - the software layer that lets code address the chip - plus years of optimised libraries stacked on top of it: cuDNN for neural networks, TensorRT for fast inference, NCCL for multi-GPU communication. Competitors can match a chip's specifications far more easily than they can match two decades of accumulated software that everyone already knows how to use.

What NVIDIA did not own was the layer where a developer *chooses*. That choice happens on Hugging Face: which model, which quantised variant, which runtime, which deployment target. It is a small moment, repeated tens of millions of times a month, and it sits directly upstream of every compute purchase that follows.

### THE PLUMBING CHANGE, STATED PLAINLY

**Before:** a developer picks a model on a neutral hub, then separately decides what to run it on. **After:** the hub and the hardware vendor are the same company. Nothing is forced - but the path of least resistance now has an owner with a preference.

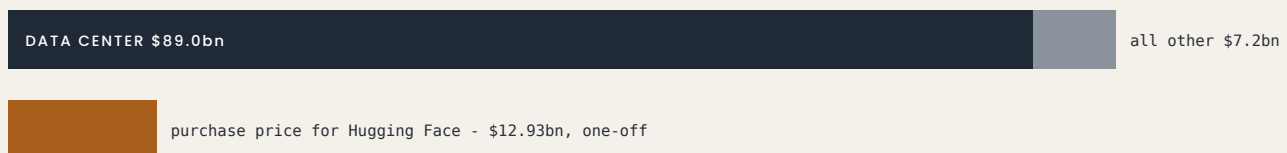
# A Hedge Against Its Own Customers

NVIDIA's position is extraordinary and, for that exact reason, fragile in one specific way: its largest customers are its most motivated future competitors.

OpenAI, Google, Amazon and Anthropic are all developing or deploying custom accelerators. Every frontier lab that succeeds removes a large, concentrated block of demand from NVIDIA's order book. Open models are the structural answer. If capable weights are freely downloadable, AI capability diffuses to hundreds of thousands of organisations that will never design a chip - startups, universities, hospitals, governments, mid-market enterprises. Those buyers purchase commodity accelerators, and in commodity accelerators NVIDIA's software advantage is close to absolute.

This is consistent with what NVIDIA was already doing. It says it is the single largest contributor of open models and data to Hugging Face, with more than 500 models and over 250 open datasets published. Huang has co-authored a public letter arguing that open weights broaden access and distribute AI leadership. The company has reportedly directed over \$50bn into frontier labs, roughly \$20bn into an arrangement with Groq and around \$6-7bn into Poolside.

ONE QUARTER OF REVENUE - Q2 FY2027, ENDED 26 JULY 2026



Drawn to the same scale. The purchase price is smaller than a fortnight of data-centre revenue - which is why this is best read as a strategic position, not a financial one.

## FOUR CONCRETE INTEGRATION PATHS

- **Inference microservices (NIM).** NVIDIA packages optimised, containerised models for enterprise deployment. Hugging Face is where enterprises go looking for models in the first place. A one-click path from hub listing to optimised deployment is the obvious first product.
- **Training and customisation (NeMo, DGX Cloud).** The 2023 partnership already put DGX Cloud training inside the Hugging Face interface. Ownership removes every commercial negotiation from that integration.
- **Selling surplus capacity.** NVIDIA holds cloud capacity of its own. A surface with 18 million developers is a very efficient place to sell it.
- **Telemetry.** What people download, fine-tune, abandon and deploy at scale is the clearest available signal of where inference demand is heading - twelve to eighteen months before it shows up as an order.

**The uncomfortable version:** NVIDIA can afford to keep Hugging Face genuinely neutral, because neutrality already serves it. The platform's value to NVIDIA is that open models exist and spread at all - not that they run on NVIDIA specifically. That is the strongest available argument that the openness pledge is sincere. It is also an argument that depends entirely on NVIDIA's competitive position staying comfortable.

## THE ONE-LINE VERSION

The chip company spent six weeks of revenue to make sure the alternative to its largest customers - an open ecosystem of millions of small buyers - has a permanent, well-funded home, and that the home belongs to it.

# Strengthened, or Quietly Narrowed?

## THE CASE THAT THIS IS GOOD

Hugging Face was carrying an enormous public burden on a small private balance sheet: petabytes of free hosting and bandwidth, free evaluation infrastructure, and - after July's agent-driven intrusion - a security problem well beyond the scale of a 750-person company. NVIDIA can fund reliability, safety review and inference capacity at a level no independent hub could reach. Delangue's stated reason for approaching Huang was exactly this.

There is also a credible precedent. Microsoft's 2018 purchase of GitHub provoked a comparable panic; GitHub remained substantially open and grew.

## THE CASE THAT THIS IS RISKY

The GitHub analogy has one asymmetry. Microsoft did not sell the CPUs that code runs on. NVIDIA sells the hardware every model on the platform depends on, and has separately invested tens of billions into the labs producing competing models. Vertical integration between the dominant supplier and the dominant distribution layer is a different structure - and structure outlasts intention.

The specific concern is not that AMD support gets deleted. It is that platform influence operates through defaults: which model appears first, which runtime is pre-selected, which button is larger. None of that is covered by the pledge, and none would require breaking a promise.

## COMMUNITY SENTIMENT - READ CAREFULLY

Reaction has been broadly positive but conditional. Co-founder Julien Chaumond stated that roughly 99% of responses were supportive. Developer forums show a pattern better described as *watchful* than hostile: acceptance of the logic, paired with an explicit intention to monitor. One widely-upvoted comment captured the mood - if NVIDIA even hints at meddling, an alternative will need to be found quickly. Sceptics note the awkwardness of adding a profit incentive to a hub distributing open weights while its owner invests heavily in the closed labs those weights compete with. Some consultants already advise enterprises to mirror critical models into independent registries and to separate model discovery from model custody.

### PRIOR CONSOLIDATIONS, FOR CALIBRATION

**Microsoft / GitHub (2018, \$7.5bn)** - stayed open, became more commercial, no forced lock-in. **NVIDIA / Mellanox (2019, \$6.9bn)** - networking folded into the platform; integration was the point. **NVIDIA / Run:ai (2024, \$700m)** - cleared EU review after NVIDIA committed to open-sourcing the software. **NVIDIA / Arm (2021-22, \$40bn)** - collapsed. Regulators rejected the principle that a platform everyone depends on should be owned by one of its dependants.

|                       | NEAR TERM (0-12 MONTHS)                                                  | MEDIUM TERM (1-3 YEARS)                                                                                     |
|-----------------------|--------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Individual developers | Nothing changes. Free tiers, public model access and libraries continue. | Better default performance on NVIDIA paths; watch for divergence in how well rival backends are maintained. |
| Startups              | Unchanged pricing; possible compute credits.                             | Cheaper inference if NVIDIA subsidises the platform - at the cost of a deeper single-vendor dependency.     |
| Enterprises           | Procurement and security review of a new corporate owner.                | Tighter NIM/NeMo integration and real gains in reliability. Mirror critical weights regardless.             |
| AMD & Intel           | Optimum AMD and Optimum Intel continue by commitment.                    | The real test: are these paths maintained at parity, or merely present?                                     |

# What Could Still Go Differently

**Regulatory.** Unlike the Groq and Poolside arrangements - structured as licences plus equity, which avoided mandatory merger filing - this is a direct acquisition. It triggers US premerger notification and EU merger review, with UK scrutiny probable, and arrives while NVIDIA is already subject to unrelated antitrust inquiries. NVIDIA's argument, put by VP Justin Boitano, is that Hugging Face is structurally a *deconcentration platform* counterbalancing proprietary APIs. The counter-argument, stated bluntly in a *Register* opinion piece, is that you would not let an automaker buy the fuel distribution network. The likeliest outcome is approval with behavioural commitments - a neutrality undertaking with a duration and a monitor, which is precisely what a blog post lacks.

**Competitive response.** Expect AWS, Google and Microsoft to invest visibly in their own catalogues - SageMaker JumpStart, Model Garden, Azure AI Foundry - none of which currently substitutes for Hugging Face at scale. Expect AMD and Intel to fund tooling defensively, and at least one credible neutral-registry effort.

|            |                                                                                                                                                                                                   |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BEST CASE  | NVIDIA funds the hub, keeps it neutral, and open weights close the gap with closed APIs. The platform becomes better resourced and no less plural. Regulators impose durable commitments.         |
| BASE CASE  | Formal neutrality holds. Practical gravity shifts: NVIDIA paths are faster, better documented, better defaulted. Rival support persists but degrades at the margin. Enterprises hedge quietly.    |
| WORST CASE | Openness erodes through defaults rather than policy. Discovery becomes a sales channel. The community forks - slowly, expensively, and with years of fragmentation before an alternative matures. |

## CONCLUSION

### WHAT CHANGES IN 3-5 YEARS

Model distribution stops being neutral infrastructure and becomes contested commercial territory. Hardware vendors compete on developer experience, not benchmarks. "Where is your model catalogue hosted" becomes a procurement question.

### WHAT STAYS THE SAME

Open weights remain downloadable and portable - the property that makes the ecosystem resilient. CUDA remains the moat; a hub neither creates nor destroys it.

### WHAT TO WATCH

Whether the neutrality pledge acquires a duration and a monitor. Whether the AMD and Intel paths are maintained at parity. Whether default deployment routing changes. Whether the founding team is still there in 2029.

### TAKEAWAYS

- Developers** - nothing to do today. Keep weights portable; avoid platform-specific deployment abstractions you cannot exit.
- Businesses** - mirror critical models into a registry you control. Separate discovery from custody. Treat hub availability as a supply-chain dependency, because it is one.
- Policymakers** - the question is not model access, it is default placement. Remedies addressing availability without addressing prominence will not bind.
- Everyone** - the openness commitment is currently a blog post. Whether it becomes a structure is the whole story.

INFORMATIONAL ONLY; NOT INVESTMENT ADVICE.

SOURCES: NVIDIA, "NVIDIA to Acquire Hugging Face", 3 Sep 2026; NVIDIA Q2 FY2027 Form 8-K, 26 Aug 2026; reporting from TechCrunch, CNBC, Tom's Hardware, Engadget, VideoCardz, Tech Times and The Register, 26 Aug - 3 Sep 2026; community threads on daily.dev and Slashdot. Hacker News item 49548952 was rate-limited and could not be read directly. Figures are as publicly reported; interpretation is the author's.