Databricks: Breakdowns Research

Research by: Erick Mokaya

Research Sources

- Company's Website, Blog
- Ali Ghodsi, Databricks CEO: A Fortt Knox Conversation

Company History

- **2013:** Databricks is co-founded by Matei Zaharia alongside Ali Ghodsi, Ion Stoica, Patrick Wendell, Reynold Xin, Arsalan Tavakoli-Shiraji, and Andy Konwinski. The seven co-founders initially worked together when they developed Apache Spark in the University of California's Algorithms, Machines & People Lab (AMP Lab) back in 2009. The company also secured its inaugural funding during this year - $14M Series A investment in a round led by VC Andreessen Horowitz.
- **2014:** Databricks Cloud, the company's Apache Spark-based cloud platform is unveiled. That same year the company also raised $33M in Series B funding in a round led by New Enterprise Associates (NEA) with follow-on investment from existing investor VC Andreessen Horowitz.
- **2015:** Databricks partners with Amazon Web Services (AWS) unifying Spark-AWS infrastructures and making it easier for companies to use Databricks on AWS's cloud computing platform.
- **2016:** A similar partnership to the one with AWS, is made with Microsoft's Azure. That same year Ali Ghodsi became CEO and Ion Stoica Executive Chairman. The company also raised $60M in Series C funding in a round led by existing investor New Enterprise Associates (NEA).
- **2017:** Databricks Delta, the company's unified data management system is unveiled with the aim of simplifying large-scale data management. According to Databricks data management was challenging at the time owing to the multiplicity of systems involving streaming systems, data lakes, and data warehouses.
    - That same year the company also raised $140M in Series D funding in a round led by existing investor New Enterprise Associates (NEA)
- **2019:** Databricks secures $250M Series E investment in a round led by Andreessen Horowitz, Coatue Management, and Microsoft (following the 2016 partnership). The company would also later in the year raise $400M at a $6.2B valuation in a Series F round led by Andreessen Horowitz and joined by other investors including BlackRock, T.Rowe Price Associates, and Tiger Global Management.
- **2020:** Databricks acquires Redash - an open-source dashboard and visualization service. The acquisition brings visualization capabilities to Databricks' data lakes.
- **2021:** Fundraising continues as Databricks secures a $1B Series G investment at a USD 28B valuation (post-money) in a round led by Franklin Templeton and joined by new investors; Fidelity and Whale Rock. Seven

months later, the company would also raise $1.6B at a $38B valuation in a Series H round led by Counterpart Global (Morgan Stanley) and joined by other new investors including Baillie Gifford, ClearBridge Investments, and UC Investments.

    - The company also acquired German startup 8080 Labs that year to widen its accessibility to citizen data scientists.

- **2022:** Databricks acquires Data Joy and Cortex Labs. Following the investments secured in 2021, Databricks also expanded its global operations into Italy and Korea with the opening of office spaces and hiring of talent.
- **2023:** Databricks completes the acquisition of MosaicML, Rubicon, and Okera. MosaicML is known for its leading generative AI platform. Databricks has also recently partnered with Salesforce in an integration that combines Salesforce's Data Cloud and Databricks' Lakehouse.
    - The company has also raised $500M in a Series I round valuing the firm at $43B.
    - Launched Dolly, a ChatGPT rival. The Dolly LLM was trained for less than $30 using just three machines. Dolly 2.0 is a small, open-source, generative model that can follow instructions and is licensed for commercial and research use.

Key People

Ali Ghodsi - Co-Founder and CEO

- Ali is part of the team that co-founded Databricks and was named CEO in January 2016. Before taking up the CEO role, he served as VP of Engineering and Product Management for two and a half years.
- He joined UC Berkeley as a visiting scholar in 2009 where he met Matei Zaharia, who was then a Ph.D. student. The two together with five others created an open-source project Apache Spark at the University of California's Algorithms, Machines & People Lab back in 2009. Today, Ali serves as an adjunct professor at the University and is also on the board of the University's RiseLab.
- Ali was born in Iran but grew up in Sweden where his family moved in 1984 after fleeing the Iranian Revolution. He started programming around the age of 9, and by the age of 14 he was helping fix computers for friends and family and earning money from it. He holds an MBA in Logistics & Strategic Marketing and an MSc in Computer Engineering from Mid-Sweden University and a Ph.D. in Distributed Computing from KHT/Royal Institute of Technology in Sweden. Before moving to the US, he was an Associate Professor with the Royal Institute of Technology and co-founded Peerialism AB, a software development firm based in Sweden.

Ion Stoica - Co-Founder and Executive Chairman

- Ion was named Executive Chairman in January 2016. He is a Databricks co-founder and a pioneer developer of Apache Spark at the University of California. He served as Databricks CEO from 2013 to 2015 and had to step away to focus on his responsibilities at UC Berkeley.
- He has served as a Professor in the EECS Department since 2009 and is a co-director of the Algorithms, Machines & People Lab (AMP Lab) at UC Berkeley. He also serves as CTO at Conviva, a firm that commercializes technologies for large-scale video distribution that he co-founded in 2006.
- He holds a Ph.D. in Electrical & Computer Engineering from Carnegie Mellon University, and a Master's in Computer Science & Control Engineering from Polytechnic University, Bucharest.

Matei Zaharia - Co-Founder and CTO

- Matei is a co-founder of Databricks and has held the role of CTO since 2013. He started the Apache Spark research project at the University of California back in 2009 during his Ph.D. program. Besides Apache Spark, he has also worked on Delta Lake, MLFlow, and Dolly. He has been recognized for his research work in combining LLMs with external data sources to improve the efficiency and quality of results. He serves as an Associate Professor of Computer Science at UC Berkeley.
- He holds a Ph.D. in Computer Science from UC Berkeley, and a Bachelor of Mathematics, Honors Computer Science from the University of Waterloo. From 2016-2022, he was an Assistant Professor of Computer Science at

Stanford University, then moved up to Associate professor for about a year before leaving in July 2023.

Business Model

Overview

- Databricks seeks to simplify and democratize data by combining the best features of data warehouses and data lakes. The company is headquartered in San Francisco with offices around the world.
- Databricks' unique "Lakehouse" platform is designed to help customers overcome the common limitations of data warehouses and data lakes. Data warehouses are costly and complex to build and maintain, have a limited scope and variety of data analysis, and require vast data integration. Data lake limitations include data discoverability challenges, lack of ACID support (an industry standard for properties of database transactions), inadequate data maturity, limited support for BI workloads, and sluggish query performance.

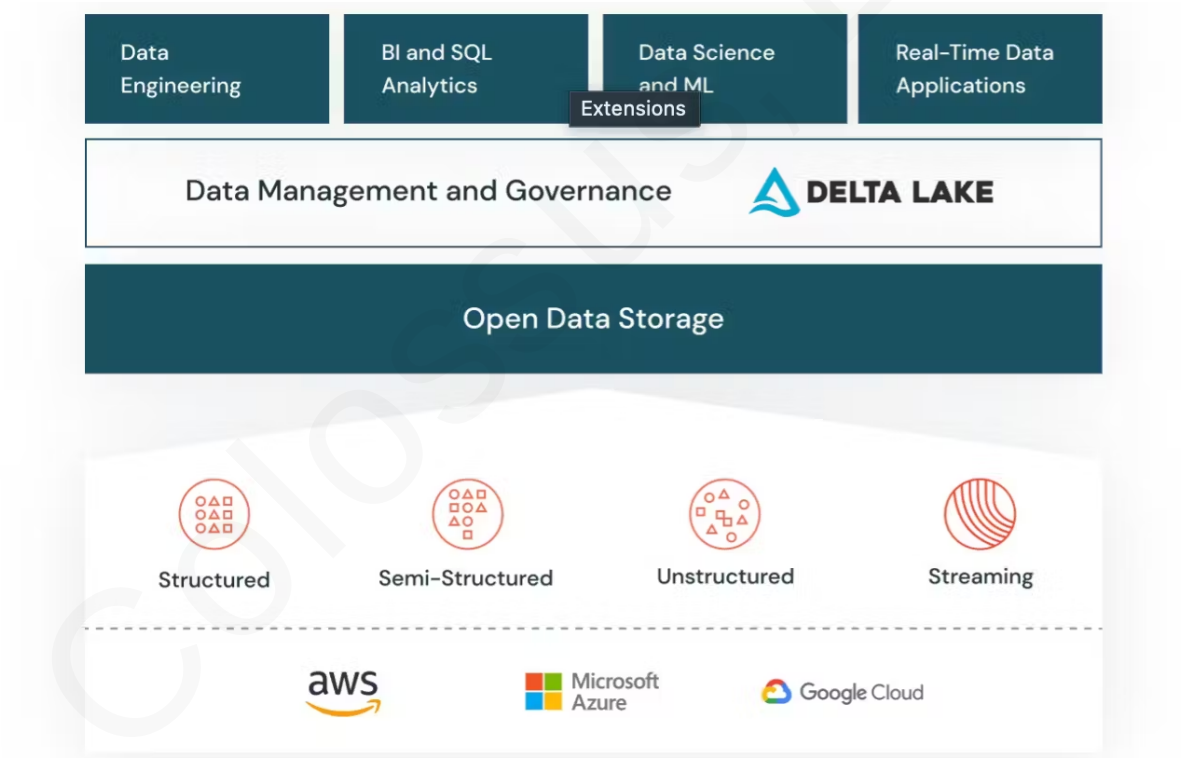
Source

Product

- Databricks Lakehouse Platform is the world's first and only lakehouse in the cloud and runs on every popular public cloud offering. It's a unified, scalable platform built on open source and open standards: Apache Spark, Delta Lake, and MLFlow. Databricks also features support for popular open-source technologies such as TensorFlow, PyTorch, RStudio, Keras, Scikit-Learn, XGBoost, and Terraform.
    - *Apache Spark:* This is a very fast, easy-to-use multi-language engine for large-scale data processing. It's the largest open-source project in data processing with more than 1,000 contributors from more than 250 organizations. It provides support for streaming data, SQL queries, machine learning, and graph processing.

Source

- *Delta Lake:* Open protocol for secure data sharing that offers massive speed, and scale. With Delta Lake, customers can share business data across organizations securely and in a compliant manner.



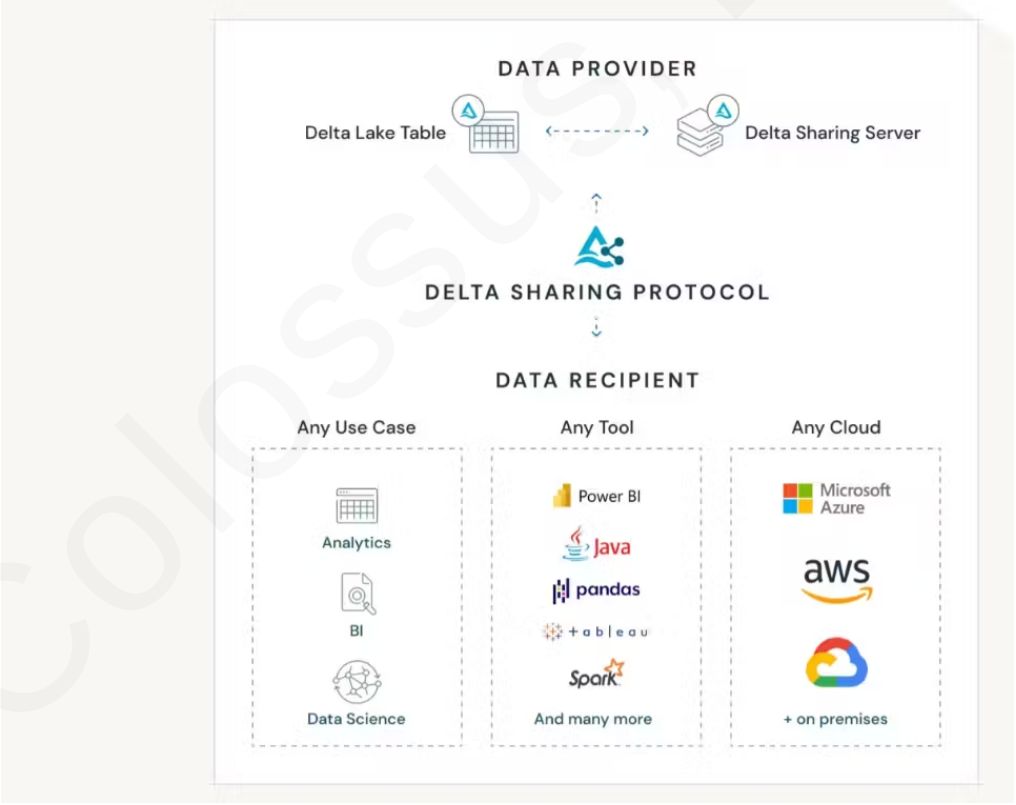
Source

- *MLflow:* Databricks' implementation of the MLflow protocol is used to manage the entire machine learning lifecycle and deploy large language models.

- The Databricks platform provides solutions to both small businesses and large enterprises across a wide range of industries which includes financial services, broadcast & streaming, manufacturing, gaming, healthcare & life sciences, energy & utilities, public sector, technology & software, advertising & marketing, communications, media & entertainment, and retail & consumer goods.

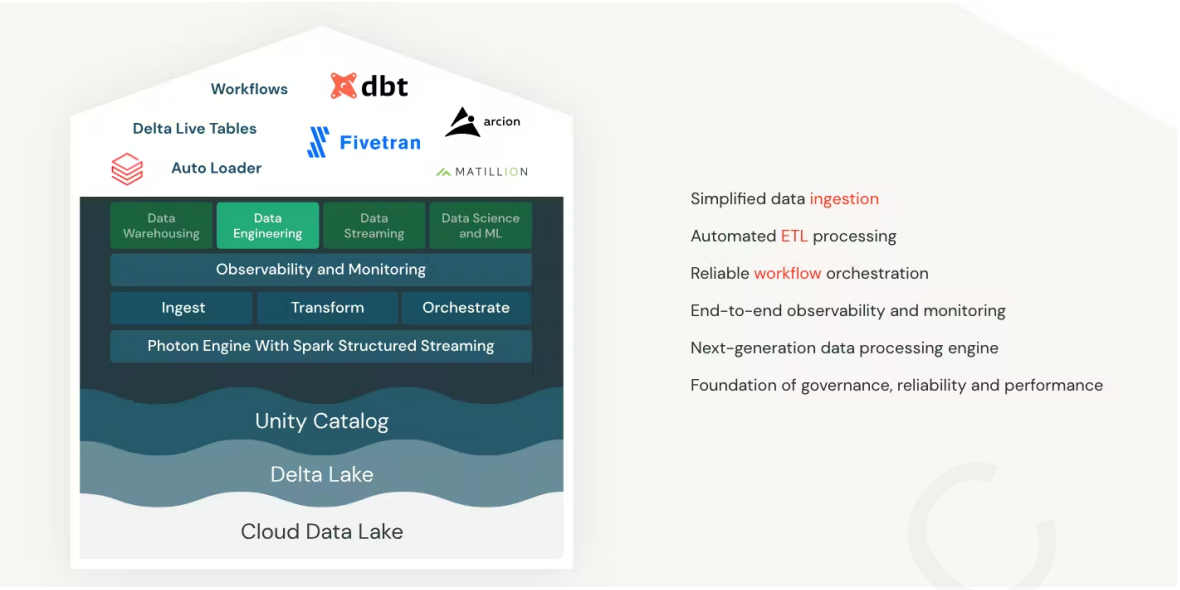
- *There are an infinite number of ways to use Databricks, and we've been blown away by all the cool things our customers are doing. For example, Regeneron uses our ML algorithms to detect the gene in DNA that's responsible for chronic liver disease, and then they were able to develop a drug that targets that particular gene. Or a company like Comcast uses Databricks to make their voice-activated remote controls work. When you talk to the remote control, that voice data goes into the cloud for Databricks to process using machine learning, and it figures out what you said and directs the TV to the right channel. And during the pandemic, hospitals used Databricks to get a real-time picture of how full their ERs were so they could redirect patients in ambulances to different hospitals that had space. Financial services firms are analyzing satellite data to make predictions about which global sectors and companies to invest in. Shell uses Databricks to monitor sensor data from 200 million valves to predict if any are going to break, so they can replace them ahead of time to keep systems running, save money, and ensure employees stay safe. - Ali Ghodsi, CEO & Co-founder*

- The company creates new products through collaboration with developers and heavy investment in research and development. It develops open-source software for data processing and AI applications and offers paid fully managed enterprise versions of the projects which include additional proprietary features. Some of its recent product success stories include:
  - *Data sharing:* **Delta Sharing** was developed by Databricks and the Linux Foundation, and is the industry's first open protocol for simple and secure data access organizations. Allows for the sharing of live data sets, models, dashboards, and notebooks across platforms, clouds, and regions without relying on specific data-sharing services.



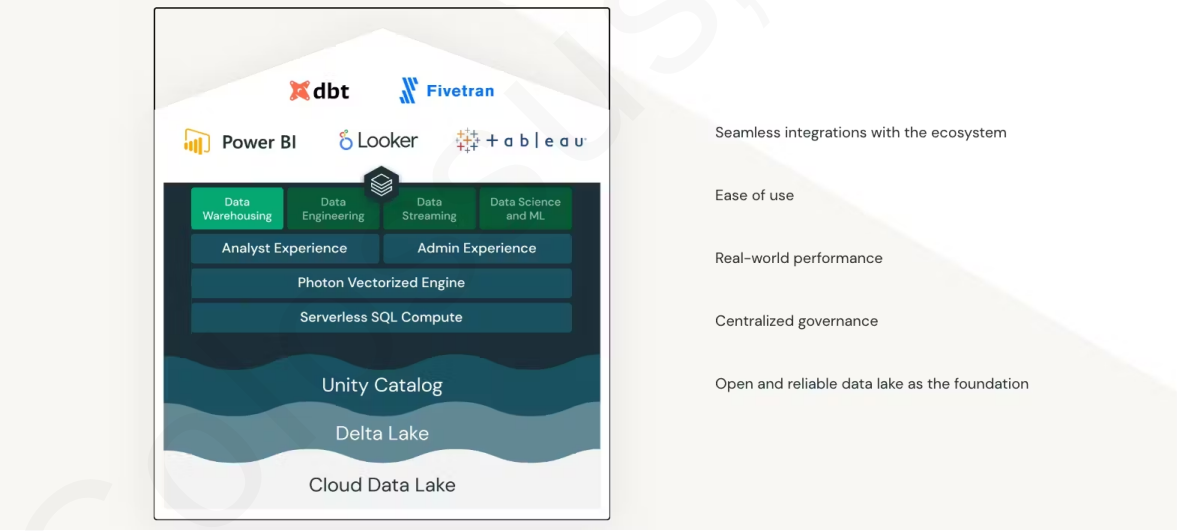
Source

  - *Data Engineering:* This solution is powered by **Photon** (a next-generation engine compatible with Apache Spark APIs) and reduces the time it takes to develop new ideas and increases the availability and accuracy of data.



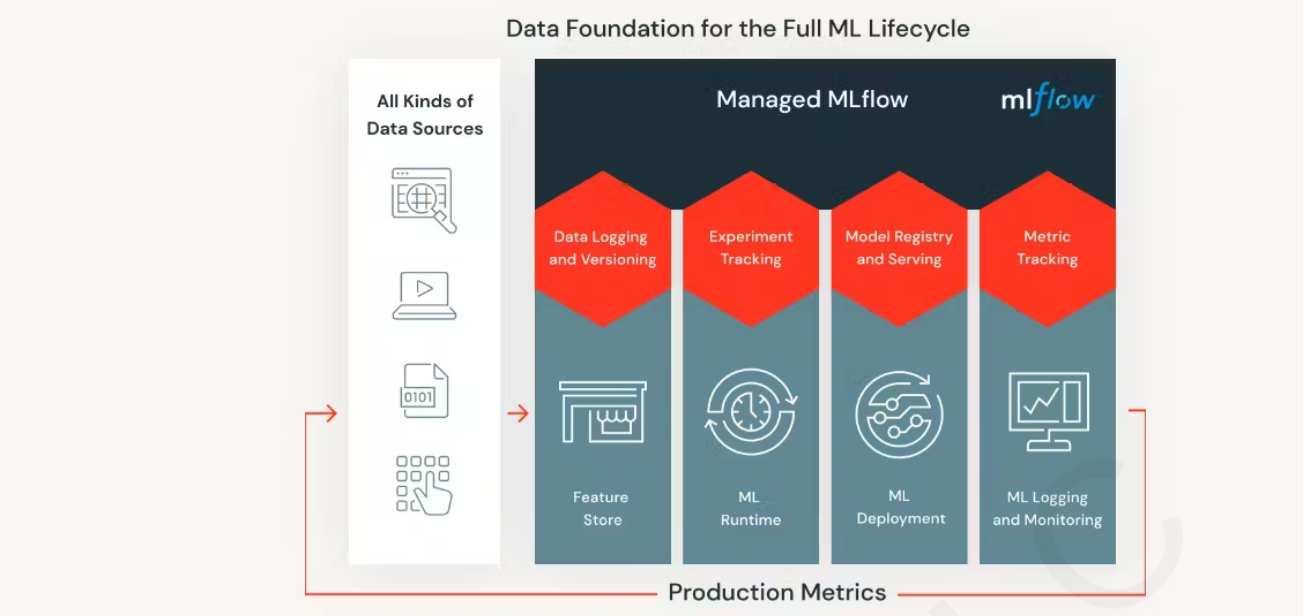
Source

- *Data warehousing:* **Databricks SQL** (DB SQL) is a serverless data warehouse on the Databricks Lakehouse Platform that provides general compute resources for business analytics. DB SQL eliminates the need to manage, configure, or scale cloud infrastructure. It has built-in governance, and modern analytics and BI tools to ingest, transform, and query data, and provides analysts with access to the latest data faster for real-time analytics.



Source

- *Real-time streaming:* Databricks' Lakehouse Platform simplifies data streaming to deliver real-time analytics, machine learning, and applications on a single platform.
- *Machine learning:* AI and ML on Databricks help organizations accelerate innovation, unlock high ROI, and improve customer experience while reducing the total cost of ownership. Features include LLM support, model serving, AI governance, and Lakehouse monitoring. Databricks enables ML teams to prepare any type of data for AI and ML, automate experiment tracking and governance, manage the full model lifecycle from data to production, and deploy models at scale and low latency.

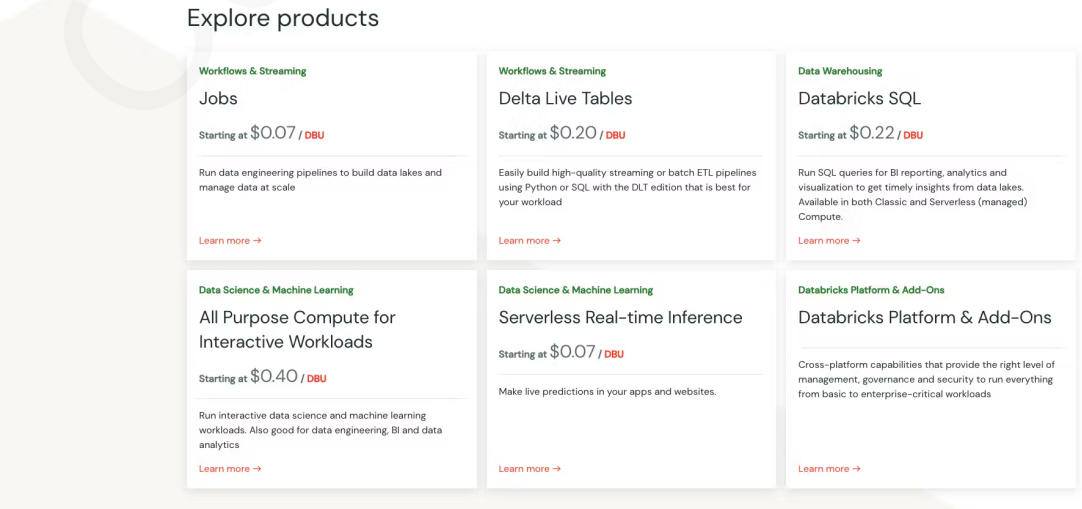


Source

- *Data Science:* Tools available include Databricks Notebooks, IDE integrations, Databrick Workflows, Repos, Databrick solution accelerators, and machine learning. This product streamlines the end-to-end data science workflow from data to modeling to insights in a collaborative environment. Databrick solution accelerators enable companies to deliver data and AI-driven solutions faster.
- *Marketplace:* This is an open marketplace powered by Delta Sharing standards where one can access and share data, analytics, and AI assets.

Monetization

- Databricks *generates revenue* from subscription plans, enterprise solutions, consulting, professional services, and the marketplace. The company monetizes open-source projects by offering a fully managed and optimized version to customers. Subscription fees cover storage, compute resources, security, and customer support while proprietary features cover data integration, monitoring, security, and collaboration features.
- The company offers a *Pay-as-you-go model* with a 14-day free trial. There are no upfront costs, customers are charged depending on the amount of compute they consume. Billing is done monthly based on per-second usage. Databricks Units, a proprietary and normalized measure of processing power consumption, are used to measure and price customers' usage. The company offers discounts to customers who commit to certain levels of usage. Prices may vary depending on geographical region and choice of cloud provider.



Source


Select cloud

AWS Azure Google Cloud

Platform Tiers

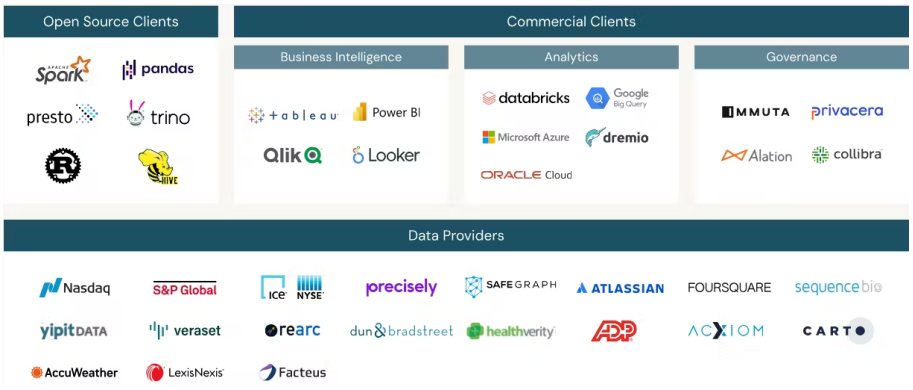
Feature Detail	Standard	Premium	Enterprise
+ Databricks Workspace	Workspace for production jobs, analytics and ML	Workspace for production jobs, analytics and ML	Workspace for production jobs, analytics and ML
Performance	Up to 50x faster than Apache Spark™	Optimized Runtime Engine	
+ Governance and Manageability	Databricks Workspace administration	Audit logs and automated policy controls	Audit logs and automated policy controls
+ Enterprise Security	Secured cloud and network architecture with authentications like single sign-on	Extend your cloud-native security for company-wide adoption	Advanced compliance and security for mission-critical data

Source

- *More than 9,000 organizations globally* rely on Databricks for full-cycle machine learning, business analytics, and massive-scale data engineering. Customers include Burberry, AT&T, Nasdaq, HSBC, Adobe, Absa, Walgreens, US Airforce, ABN-Amro, Sam's Club, Warner Bros Discovery and Air Canada.
- *Distribution:* Databricks can reach its target audience and expand its reach through sales & marketing efforts, partnerships, and community engagement. The most effective strategy for Databricks has been bottom-up developer adoption. The company engages developers to code on open-source projects and also hosts developer-focused presentations. These projects and events have seen massive adoption.

Partnerships & Ecosystem

- Over the years, Databricks has developed *strong partnerships* with cloud providers, consulting partners, and technology & data companies. The company boasts more than 1,200 partners globally who provide data, analytics, and AI solutions. Partners include Amazon Web Services, Capgemini, Tableau, Microsoft, Informatica, Google Cloud, and Booz Allen Hamilton.
    - Through Databricks' integrations with cloud providers, customers can leverage AI services, storage, compute, analytics, and security to meet their needs.
    - Integrations with technology partners provide additional capabilities in data ingestion, machine learning, business intelligence, ETL, and governance.
    - Data Provider Partner Program: This program helps data providers monetize their data assets to a broad, open ecosystem of data consumers from a single platform. Through the Databricks Lakehouse Platform, Data Providers can reach a large number of customers, reduce costs, and provide a better customer experience. The product also gives data providers access to Databricks' Product and R&D team, marketing support, and industry teams to build specific industry solutions.
    - Consulting partners are experts who help organizations build, deploy, scale, design transformation strategies, manage data, set up governance structures, or migrate to the Lakehouse Platform. Organizations can leverage brickbuilder solutions to migrate to the platform. Some of the consulting partners providing these solutions are Deloitte, Capgemini, Celebal, Impetus, Avanade, Wipro, Lovelytics, and Perficient.



Source

- *Acquisition*: Databricks has made at least seven acquisitions, indicating a willingness to pursue growth by buying competitors and established companies where it can integrate complementary offerings. The MosaicML and Okera transactions were completed in 2023, with MosaicML being the most recent acquisition at a price of $1.3B. This is slightly above Databricks' $1.24B of revenue in 2022 and represented 3% of the company's valuation at the time.

Databricks Acquisition Record

Acquired Company	Announced Date	Price	Field
MosaicML	June 26, 2023	$1.3B	Generative AI
Rubicon	June 13, 2023	—	AI storage systems
Okera	May 3, 2023	—	Data governance
DataJoy Inc.	October 13, 2022	—	Revenue analytics
Cortex Labs	April 25, 2022	—	ML operations
8080 Labs	October 6, 2021	—	Low-code/no-code capabilities
Redash	June 24, 2020	—	Visualization and dashboards

Table: @ekmokaya • Source: Databricks • Created with Datawrapper

- Databricks Ventures invests in innovative early and growth-stage companies that are committed to using the Lakehouse architecture to build the next generation of data and AI companies. Companies in the portfolio include Alation, Catalyst, DBT Labs, Hightouch, Hex, Neon, Snowplow, Tecton, Revelate, Matillion, perplexity, Hunters, Lovelytics, Labelbox, Immuta, and Arcion.

Economics

Revenue Composition

- As of the end of Q2 2023, Databricks' customer base surpassed the 10K mark worldwide, representing a growth of 11.1% from Q1 2023. The revenue run rate grew by 50% year-on-year to cross the $1.5B mark on the back of rapid customer growth. At least 300 customers, or 3% of the customer base are generating in excess of $1M per year, or $300M in retained revenue per year in absolute terms. Databricks closed Q2 2023 with Non-GAAP subscription margins at 85%.

Databricks Total Customer Count

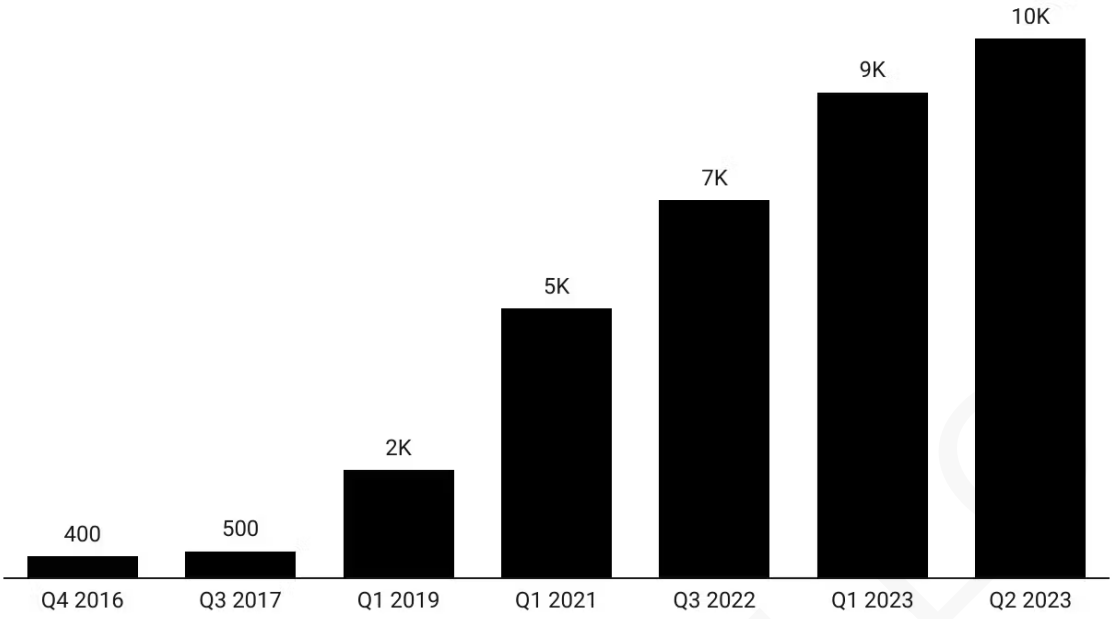


Chart: @ekmokaya • Source: Company Financials • Created with Datawrapper

- In FY 2022, Databricks became a $1B revenue-generating company, with annual recurring revenue growing by 55% year-on-year to reach $1.24B. Out of this, more than 90% were from core products, which mostly include professional data management and processing services.

Databricks Revenue (USD)

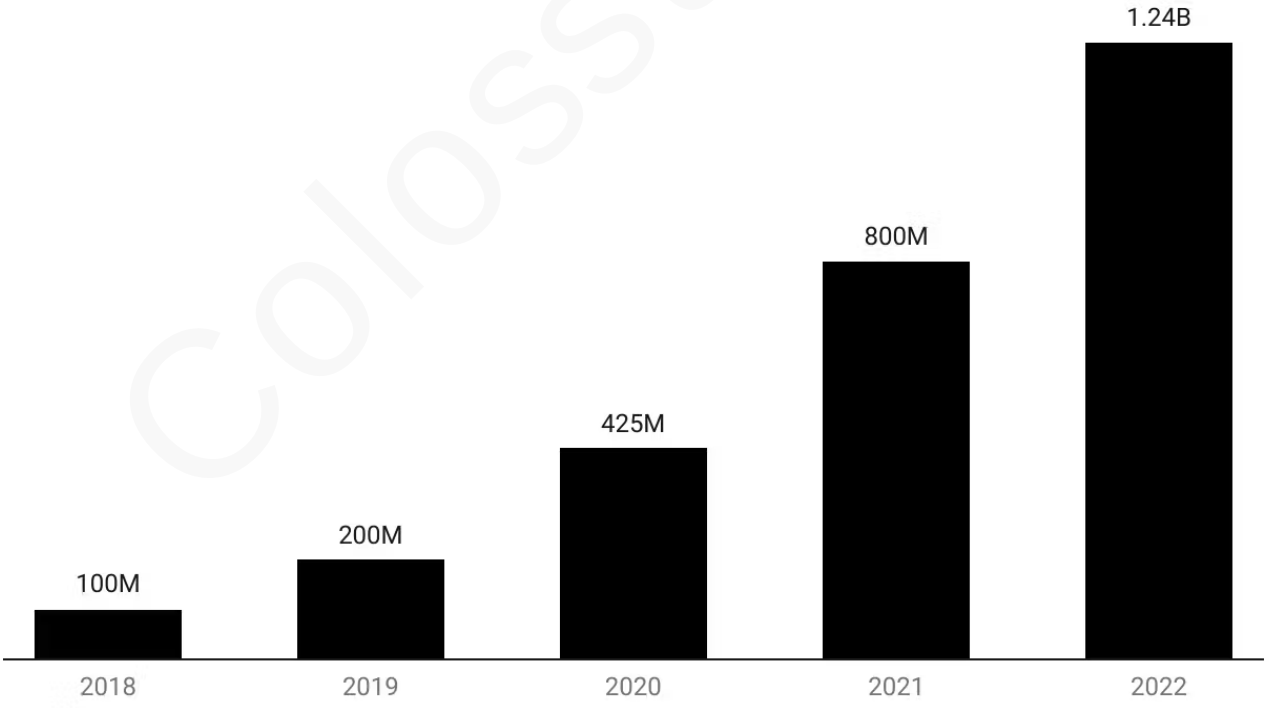


Chart: @ekmokaya • Source: Company Financials • Created with Datawrapper

- Databricks has so far successfully closed 9 fundraising rounds which have yielded an aggregate amount in the region of $4B. The latest round in September 2023 raised $500M at a $43B valuation - and most of this capital

will be expended in enhancing the product base and augmenting efforts around AI and ML given the recent acquisition of AI firm MosaicML and Nvidia's participation in the funding round.

Databricks Fundraising Record

Round	Date	Amount ($M)	Valuation ($M)
Series I	September 14, 2023	500	43,000
Series H	August 31, 2021	1,600	38,000
Series G	February 1, 2021	1,000	28,000
Series F	October 22, 2019	400	6,200
Series E	February 5, 2019	250	2,750
Series D	August 22, 2017	140	900
Series C	December 15, 2016	60	371
Series B	June 30, 2014	33	236
Series A	September 25, 2013	14	43

Table: @ekmokaya • Source: Databricks • Created with Datawrapper

Profitability

- Databricks has never turned a profit since inception in 2013, bringing into focus the net cash burn over the period. For Databricks, breaking even could open up the company to new opportunities, including exploring the idea of going public through an Initial Public Offering. On pursuing the path to profitability, CEO Ali is of the following mindset:

> *"We haven't disclosed a specific timeline to reaching profitability, and we are continuing to invest aggressively to go after the data and AI market."*

Competitive Position

Industry

- In 2022, the global data analytics market size reached $30B and is expected to grow at a CAGR of 29.4% between 2023 and 2032 to reach USD 393.35B by 2032. The global artificial intelligence market was valued at $1.82T in 2022 and is expected to grow at a CAGR of 32.4% between 2023 and 2028 to reach $9.08T by 2028.
- The volume of data available is exploding, and the future of digital business will be determined to a great extent by the ability to effectively capture value from data. The need to extract as much value as possible from this data is driving innovations in analytics, AI, and machine learning. The adoption of AI has skyrocketed, and businesses are increasingly prioritizing data-driven insights with new use cases emerging every so often. It is, therefore, important for organizations to figure out their AI strategies, and for that, they require tools and capabilities. Some of the trends in AI and Machine Learning that are proving to be significant in business are Natural Language Processing (NLP), edge computing, computer vision, and Generative AI.
- AI is a virtuous circle for companies in the data analytics business and those who harness it will undoubtedly benefit. The increased adoption of tools such as ChatGPT means more data being generated worldwide on a daily basis. This additional data in turn fuels demand for data warehousing as well as AI-powered offerings to analyze the data.
- Databricks faces competition from not just other unified data analytics companies like Snowflake but also cloud computing tech giants including Amazon Web Services, Microsoft's Azure, and Google Cloud Platform (Big Query) which also offer data analytics and AI services.

- *Snowflake:* Both Databricks and Snowflake are global leaders serving notable customers across various industries including; Adobe, AT&T, Capital One, and S&P Global among others.
- *Amazon Web Services (AWS):* Databricks and AWS have a partnership with the Databricks platform being offered on the public cloud provided by AWS. The two however are primary competitors as AWS also offers cloud computing services including big data analytics and AI.
- *Google Cloud Platform:* Google's Big Query also offers big data analytics and AI services just like AWS and is also in direct competition with Databricks.
- *Microsoft's Azure:* Databricks has a partnership with Microsoft dubbed Azure Databricks - an analytics platform based on Apache Spark. Microsoft also provides big data analytics and AI services

- While data analytics as an industry has relatively lower technological barriers to entry as compared to more high-tech industries, new entrants into the data and AI market would have a difficult time building out their product and trying to catch up with established companies like Databricks, Snowflake, Amazon Web Services, Microsoft Azure, and Google Cloud Platform. Besides the financial muscle that these companies possess, they do own massive amounts of data and established customer relationships. In this space, data is everything, and whoever can extract the most value from data will win.

Risks to Competitive Position

- **Competition:** As discussed earlier, Databricks operates in a space that has large and well-established data and AI companies. There is also a steady stream of competition from new players entering the market. As the cloud market continues to mature and new technologies emerge, Databricks needs to continuously innovate in response to the rapidly changing environment.
- **Dependency on Open-Source Projects:** These projects offer numerous benefits but carry a huge risk should there be changes in governance or differences in the community that would disrupt the project. Also, with the growth of data and AI tools, more and more open-source projects will come up and that will of course have a significant impact on Databricks' offering.
- **Relying on cloud providers as partners** who also happen to be competitors would pose a significant risk to the business should the companies change their strategies.

Competitive Advantage

- **An open and unified platform for AI:** The company offers a scalable platform that eliminates the silos that complicate data and AI. The platform provides end-to-end capabilities with integration, storage, governance, data sharing, processing, analytics, and AI. Automatic optimization for performance and storage reduces the total cost of ownership of any platform. Data teams/scientists can work seamlessly on the platform, and organizations are able to leverage the tools available to build and deploy models faster.
- **All-in-one product solutions:** DB SQL together with Photon gives 12x better price/performance than other cloud data warehouses.
- **Diversified and large customer base:** Databricks offers solutions to various businesses across diverse industries which provides stability in terms of revenue and mitigates risks associated with industry-specific disruptions. The business is still beating revenue targets despite the slowdown in the economy. Demand for diverse products and services drives innovation and provides valuable market insights.
- **Expertise in open-source frameworks/Long-tenured and experienced management team:** The co-founders and executive team are well-versed in the space and have a wealth of experience in building data and AI tools. The co-founders built the open-source projects that underpin the Lakehouse Platform.
- **Strong partnerships:** Databricks is partnered with major cloud providers including Amazon Web Services and Microsoft's Azure enabling the offering of Databricks platform on the cloud platforms. Databricks has also most recently announced a partnership with Salesforce in an integration that combines Salesforce's Data Cloud and Databricks' Lakehouse.
- **Strong investor backing:** Databricks has raised funds from some of the big names in investment circles like BlackRock, Andreessen Horowitz, Baillie Gifford, and Counterpoint Global (Morgan Stanley) among others.

Microsoft and AWS are also investors in the firm, besides being partners. This has provided the firm with financial resources to pursue new technologies and expand existing ones. The company also benefits from the knowledge and expertise that these firms have gathered over years of investing in different industries.