William Blair

Equity Research
Technology, Media, Communications

July 1, 2024

Jason Ader, CFA +1 617 235 7519
Arjun Bhatia, CPA +1 312 364 5696
Jonathan Ho +1 312 364 8276
Jake Roberge +1 312 364 8568
Sebastien Naji +1 212 245 6508

On the Ground and In the Cloud

A Developer Technology Quarterly: *Data Analytics Edition*

Contents

Introduction

On the Ground and In the Cloud is a quarterly publication produced by the William Blair technology team that delves into trends impacting developer technologies across a wide scope of topics that includes software development, DevOps, database, analytics, and observability. Over the past decade, developers have become increasingly important influencers across all organizations, as software applications and digital transformation become critical to business operations, customer interaction, and competitive advantage. More recently, this trend has been accentuated by black swan events like the COVID-19 pandemic and a slew of software supply chain attacks. Developers not only represent the early adopters that will determine the success of a particular software product or project, but also have become primary decisionmakers for making software purchases. As a result, we believe it is essential to examine the key technological and cultural dynamics impacting this all-important cohort of workers.

In this Data Analytics edition of *On the Ground and In the Cloud*, we offer insights into important technology developments and trends in the analytics market, provide results from our recent proprietary survey of developers/practitioners, and review private funding activity in the space. We also provide a recap of key announcements at recent developer conferences, including Snowflake Data Cloud Summit, Databricks Data + AI Summit, and Microsoft Build. Lastly, we offer key takeaways from our recent expert call with analytics consultant and Snowflake Elite partner Paul Corning.

Executive Summary

Rising Demand Should Lift Most Analytics Boats. Given the sheer volume of data generation and retention, unyielding demand for data-driven insights, migration of data to cloud, and the sizable impact of generative AI, we expect strong growth in the data analytics market over the next several years. In its July 2023 Big Data and Analytics (BDA) Software report, IDC Research forecasts a six-year CAGR of 19.3% through 2027 (reaching $253 billion at the end of the forecast period). A key driver of growth is expected to be AI software platforms and related tools, which IDC predicts will grow at a 36% five-year compound annual rate, increasing from 18% of total BDA spend in 2022 to 35% of spend by the end of the forecast period. Analytic data management and integration platforms (which encompasses data warehouses, data lakes, streaming, and ETL tools) is also expected to grow at a healthy 18% over the forecast period, which speaks to the breadth of demand for core analytical systems. While emerging structural shifts in the analytics market could disrupt certain subsegments and vendors over time, we see room for multiple winners and expect the overall data analytics pie to grow larger as new technologies allow customers to generate data-driven insights across more areas of the IT landscape.

Traditional Moats Are Shrinking. Historically, analytics vendors required data to be physically ingested into a centralized repository (data warehouse or DW) and stored in a proprietary format. This is changing with emergence of new technologies that create greater standardization, reduce vendor lock-in, and lower entry barriers for analytics systems. These technologies include a) open tables (e.g., Iceberg, Delta), which allow customers to store data outside the DW in standardized format for analytics (decoupling storage from compute); b) data lakehouses, which obviate the need to maintain separate DWs for structured data and data lakes (DLs) for unstructured/semi-structured data; c) data fabrics, which allow for centralized querying and management of virtualized data across physically distinct silos; and d) zero ETL (extract, transform, load), which enables tighter integration between source data and its analytical destination/sink, reducing the cost and complexity of data transformation.

Gen AI Is a Game Changer for the Analytics Stack. GenAI opens the usage of analytics to a broader range of personas by creating a simple natural language interface for asking questions about data. GenAI also grows the pie of data available for analysis by allowing NLP interfaces to be applied to more datasets and silos. This is particularly impactful for vast amounts of unstructured data (files, images, logs, etc.) within enterprises that historically was burdensome and expensive to analyze. We believe unstructured data specialists like Databricks are best positioned to capitalize on the GenAI opportunity but expect virtually all analytics vendors to grab a piece of the growing pie.

Analytics Increasingly Moving to the Data. Application vendors have historically handed off their data to analytics vendors as running analytics within source application platforms was not possible or practical. This led to creation of a separate analytical data estate with its own software stack—an expensive undertaking that lacked the ability to produce real-time insights (due to latency involved in ETL). This is beginning to change as GenAI democratizes the analytics market, entry barriers to run analytics are now lower (due in part to lakehouses and open table formats), app vendors look for new sources of revenue, and customers demand real-time analysis. The result is that analytics functions are beginning to be applied at the sources of data (the apps themselves) as opposed to the 30-year trend of data needing to move to the analytics estate (analytics moving to the data versus data moving to the analytics).

Salesforce Data Cloud (SFDC): Partner or Competitor to Analytics Systems? SFDC unifies current islands of Salesforce data (Sales Cloud, Marketing Cloud, Service Cloud, etc.) and integrates with external data to create a unified customer view within the CRM app. By adding GenAI (Einstein) to the mix, SFDC creates a powerful querying/analytics engine that can help the company capture more analytics/AI revenue (critical to long-term growth). Over time, SFDC customers should be able to reduce the amount of data moved to external analytics systems like Snowflake, Redshift (AWS), BigQuery (Google) or Databricks—especially as it is no secret that a healthy percentage of data within DWs/DLs comes straight from Salesforce. SFDC's "ace in the hole" is tight integration of holistic customer data with GenAI directly within the CRM app and grounded in the CRM user experience—this could put analytics vendors at a long-term structural disadvantage as they do not own the business apps.

Casualties of the Analytics Wars. Areas of potential disruption and/or disintermediation from new technologies include ETL tools (there will be less need to physically move and transform data due to open tables), traditional/on-prem DWs (due to market shift from BI to AI), and BI/data visualization tools (traditional BI reports/charts could eventually be replaced by vastly simpler natural-language-prompting engines trained on enterprise datasets). Notwithstanding these potential disruptions, we believe analytics expertise has scarcity value, and as such, we expect an uptick in M&A in the analytics space in coming quarters as both analytics and app vendors look to round out their platforms.

For more detail on these and other topics, please see our recent white paper: Data Analytics in the AI Age: Everyone in the Pool!

Proprietary Survey of Developers

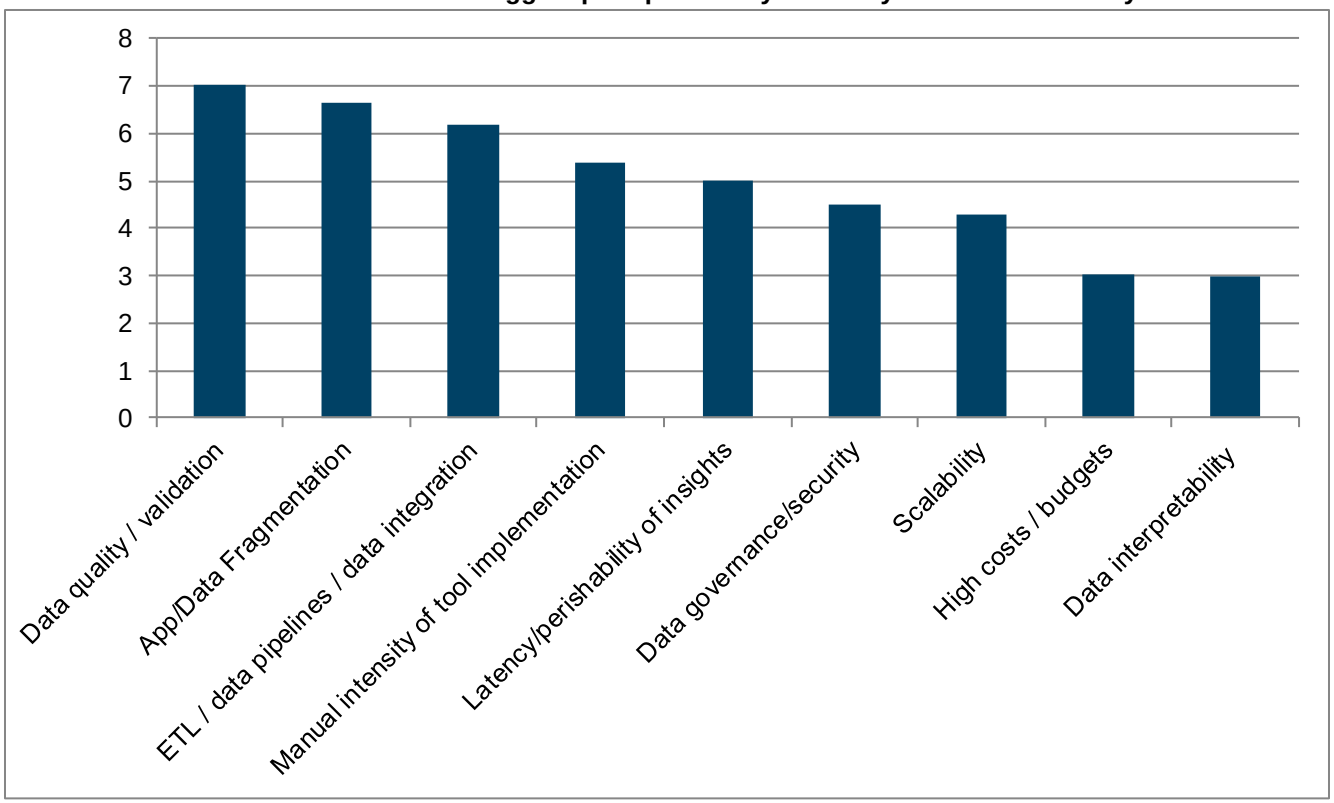
In May 2024, the William Blair technology team performed a proprietary survey of developers, data scientists, data engineers, and data analysts. Our goal with this survey was to better understand how developers and data practitioners view emerging trends in the analytics space, what tools they rely on most today, and how they see the analytics space evolving over the next five years, particularly in view of the rising adoption of generative AI. Much of our research to date has centered on speaking with vendors and analyzing the market from a technological perspective. With this survey, we aim to glean perspectives from the everyday practitioners of these data analytics technologies and hope to parse out any potential disconnects between perceived market trends (proffered by the vendor community) and what is happening on the ground with users.

We surveyed a total of 67 developers/practitioners in the data analytics space, having screened for respondents that engage in data analytics as part of their everyday job functions. Unsurprisingly, survey respondents identified the top three pain points within the data analytics workflow as 1) data quality and validation, 2) app and data fragmentation, and 3) managing ETL, data pipelines, and data integration (see exhibit 1).

Exhibit 1

On the Ground and in the Cloud; A Developer Technology Quarterly

Please rank order the biggest pain points in your analytics workflow today.



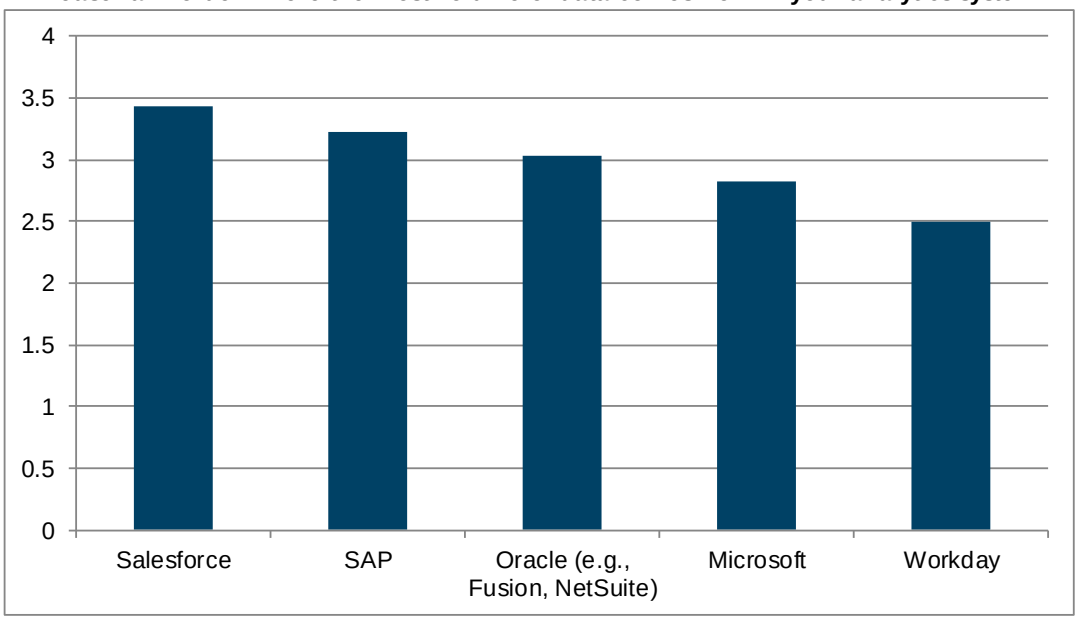
n = 67

Source: William Blair Equity Research

One of the main themes in analytics has been the fragmentation of data, particularly driven by the rise in SaaS applications, which have become a primary source of business data. Across most of the major application platforms (e.g., Salesforce, SAP, Oracle, Workday, and Microsoft), respondents noted differing levels of importance and data volumes. Of the top five, salesforce was consistently the source of the most data used in analytics (31% of responses), followed by Microsoft (with 25%) and SAP (with 21%), as seen in exhibit 2. Importantly, despite concentration of data

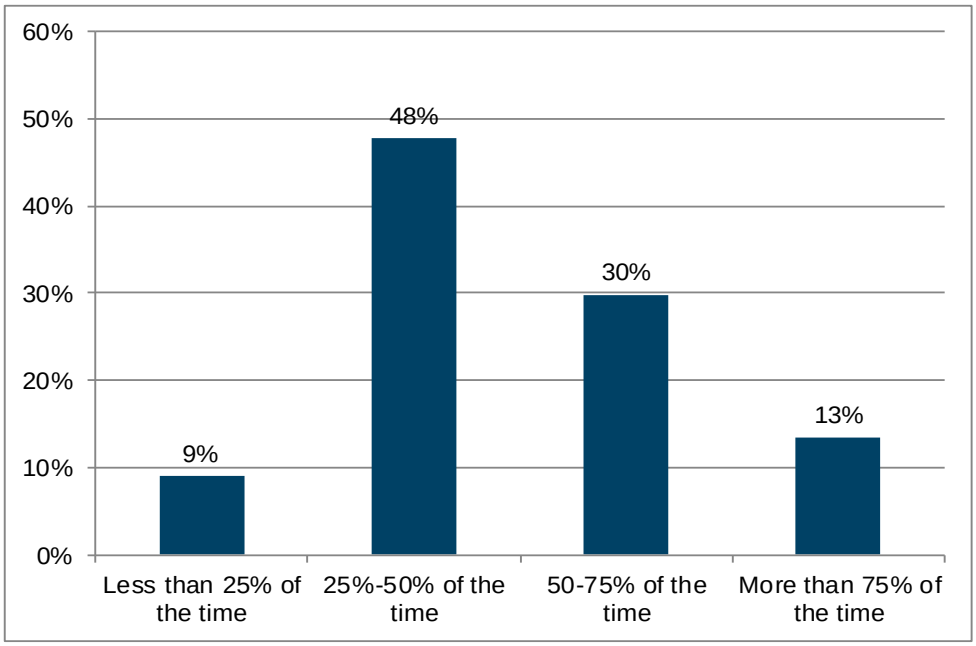
volumes in specific application ecosystems, a notable percentage of companies are also performing at least some analytics on joined data (e.g., data that has been consolidated from different sources and apps). As seen in exhibit 3, 43% of those surveyed are joining datasets for at least half of their analytics jobs.

Exhibit 2
On the Ground and in the Cloud; A Developer Technology Quarterly
Please rank order where the most volume of data comes from in your analytics system.



n = 67
Source: William Blari Equity Research

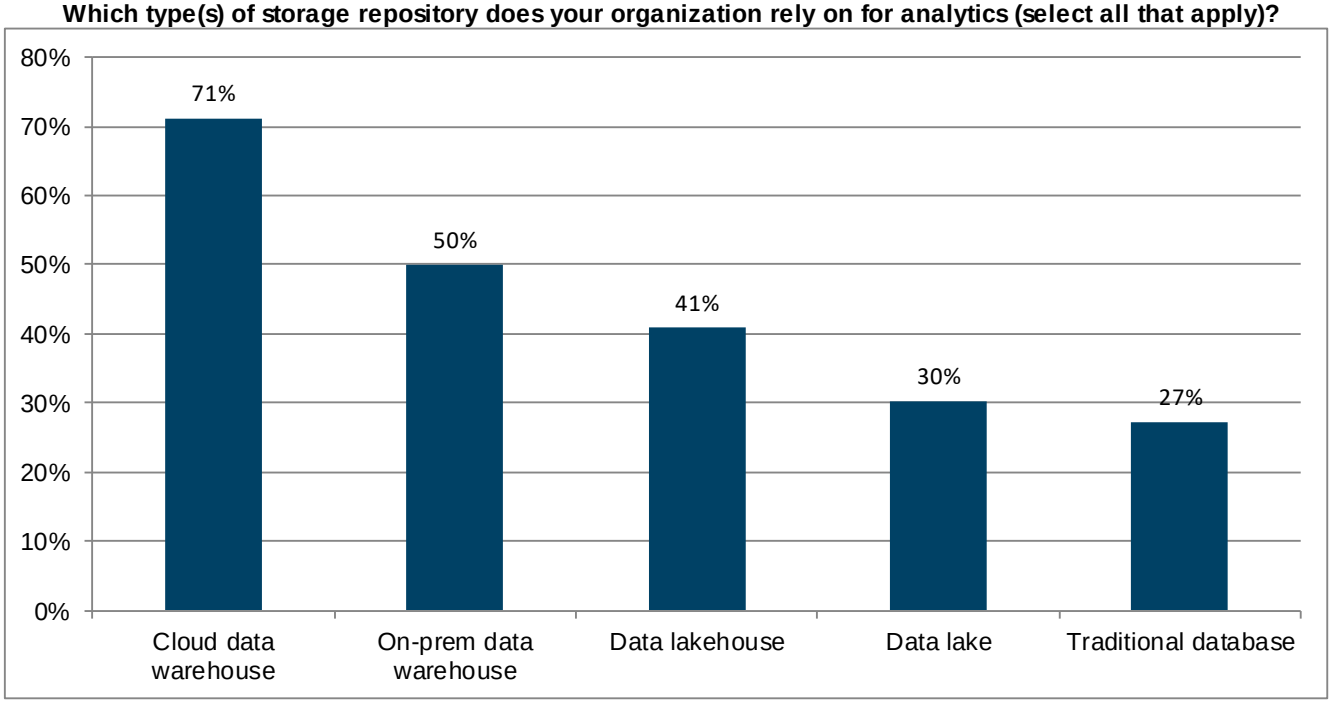
Exhibit 3
On the Ground and in the Cloud; A Developer Technology Quarterly
How often are you joining different data sets across different application and data sources to perform analytics?



n = 67
William Blair Equity Research

The need to consolidate at least some data into a central repository has driven the rapid growth of data warehousing, and most recently cloud data warehouses as the database architecture of choice for analytics. While a significant portion of enterprises retain on-prem data warehouses (half of respondents still use them), cloud data warehouses are the most ubiquitous technology, with 71% of respondents using them. Nonetheless, newer solutions like data lakes and data lakehouses are also seeing significant traction—likely to be further accelerated by the demand for unstructured data in training LLMs for AI use-cases.

Exhibit 4
On the Ground and in the Cloud; A Developer Technology Quarterly
Which type(s) of storage repository does your organization rely on for analytics (select all that apply)?

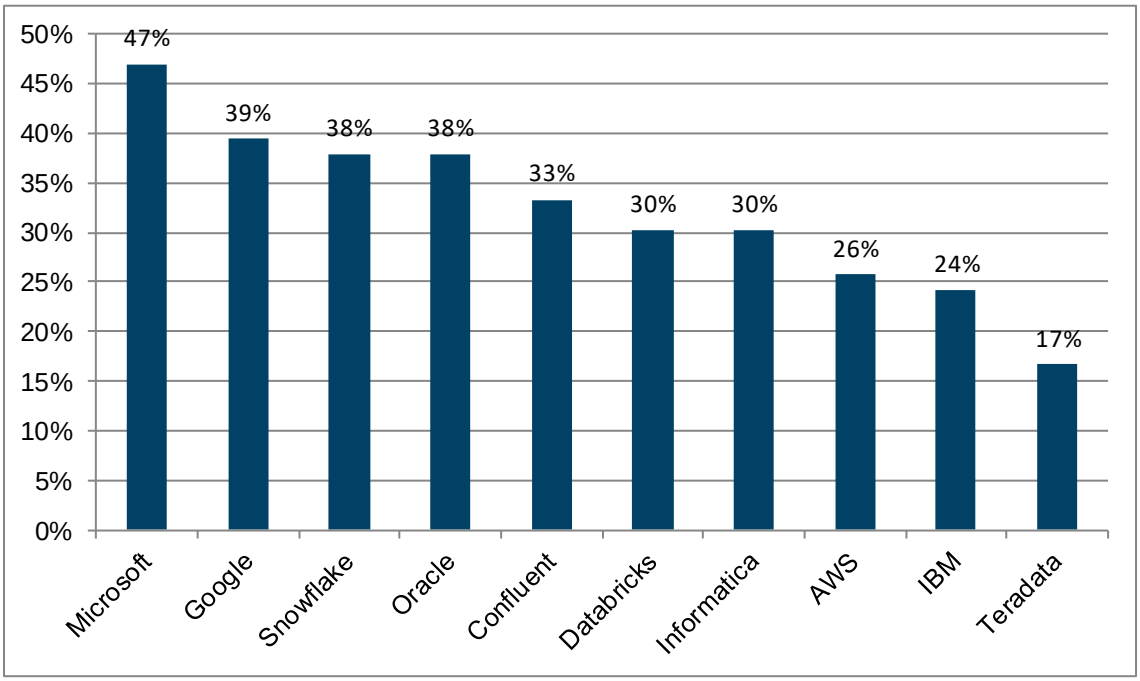


n = 66

Source: William Blair Equity Research

In terms of analytics technologies and vendors used by customers, there was a broad distribution of responses, with many enterprises leveraging multiple technologies across different applications and lines of business. At the top of the list was Microsoft, with 47% of responses highlighting its broad set of data technologies (across Azure Synapse, Azure Data Lake, Azure Data Factory, and Microsoft Fabric). In addition, Microsoft has benefited greatly from its "one hand to shake" approach, bundling together data services, cloud infrastructure, and apps. Behind Microsoft were Google (largely thanks to BigQuery), Snowflake, Oracle, Databricks, Confluent, and AWS.

Exhibit 5
On the Ground and in the Cloud; A Developer Technology Quarterly
Who are your primary analytics vendors (select all that apply)?



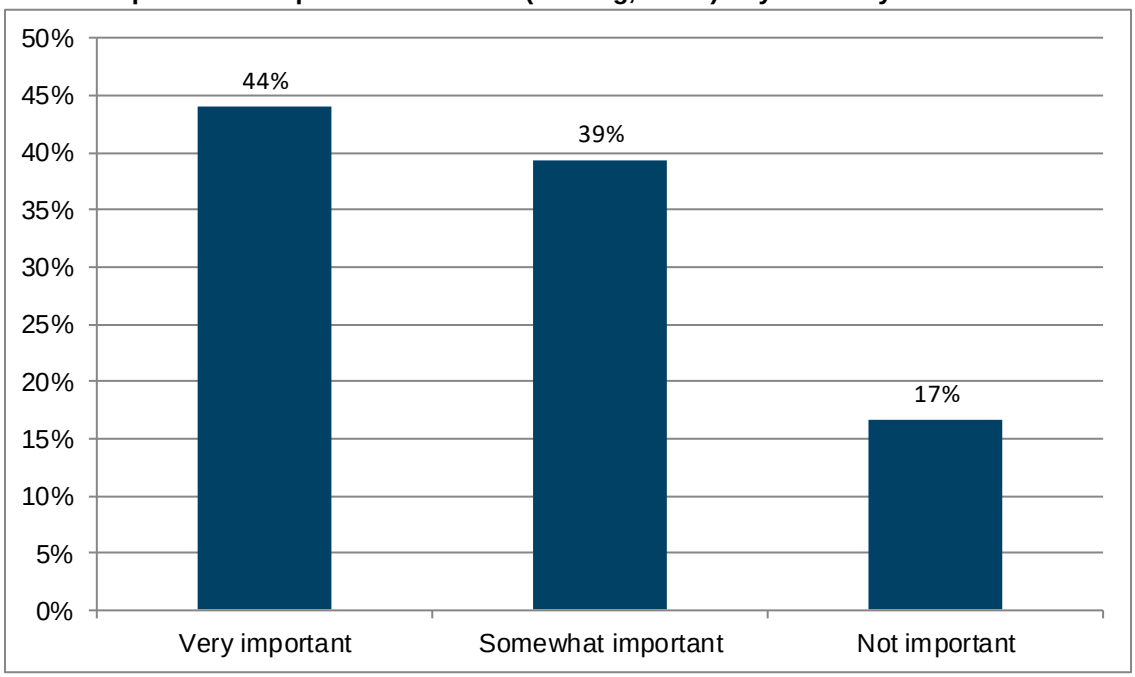
n = 66

Source: William Blair Equity Research

The advent of the new analytics technologies, including lakehouses, open tables, generative AI, and data streaming, has driven enterprises to rethink their data strategies and shift priorities toward supporting applications that are real time in nature, leverage GenAI capabilities, and limit the need to move data or create complex data pipelines that increase the chances that something breaks.

In particular, the data warehousing space has been undergoing an evolution driven by the rise of open table formats (which we discuss further in the trends section on page 11), which allow data to be stored in a consistent fashion and queried by multiple compute engines. Over the last few years, the rising popularity of these open table formats (like Iceberg and Delta) has driven changes at companies like Snowflake, which now supports storage in external Iceberg Tables, in addition to its traditional proprietary storage format. The increased importance of open table formats was corroborated by our survey results, with more than 83% of respondents indicating they were at least somewhat important technologies in their analytics stack.

Exhibit 6
On the Ground and in the Cloud; A Developer Technology Quarterly
How important are open table formats (Iceberg, Delta) to your analytics architecture?

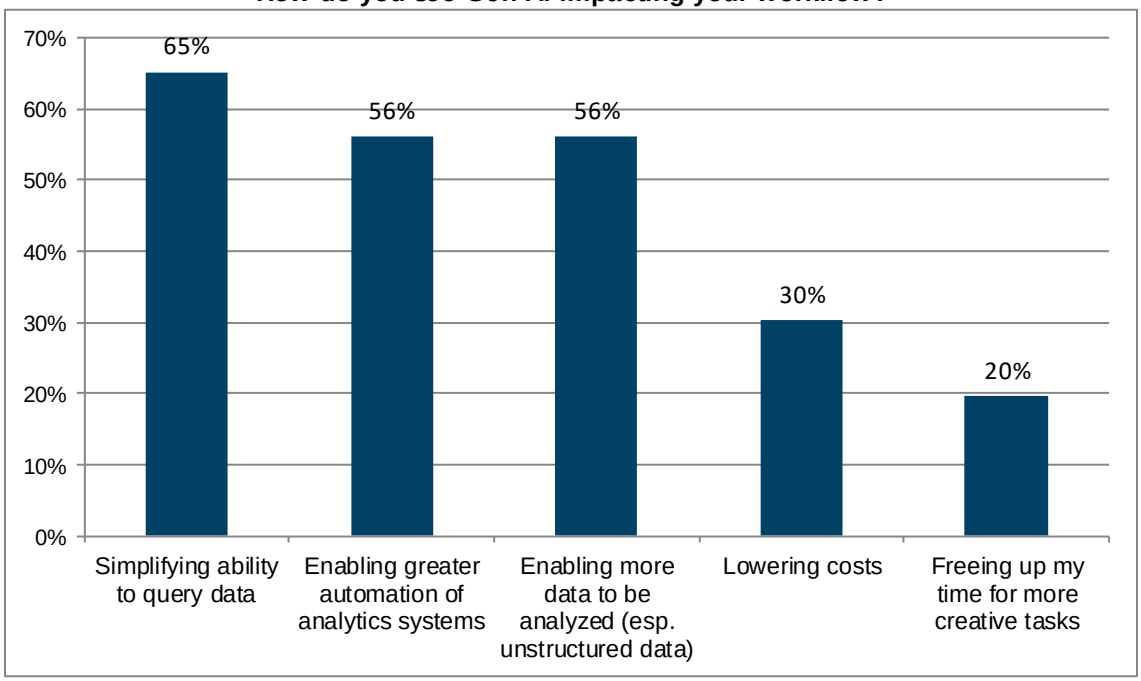


n = 66

Source: William Blair Equity Research

Looking at generative AI, we expected to see a number of potential impacts on the analytics space. The top impact of GenAI on data analytics workflows (with 65% of responses) is simplifying the ability to query data, followed by enabling greater automation, and enabling more data to be analyzed. Notably these responses are aligned with some of the early use-cases that have emerged for generative AI, including the ability to turn natural language prompts into code—in the case of analytics, these are typically SQL queries—which have the effect of allowing a broader set of personas to ask questions about company and customer data. The rising demand for data lakes and lakehouses also highlights the increasing ability to run analytics across large sets of unstructured data—historically, much of this data was ignored because it was deemed impractical or unfeasible to analyze.

Exhibit 7
On the Ground and in the Cloud; A Developer Technology Quarterly
How do you see Gen AI impacting your workflow?

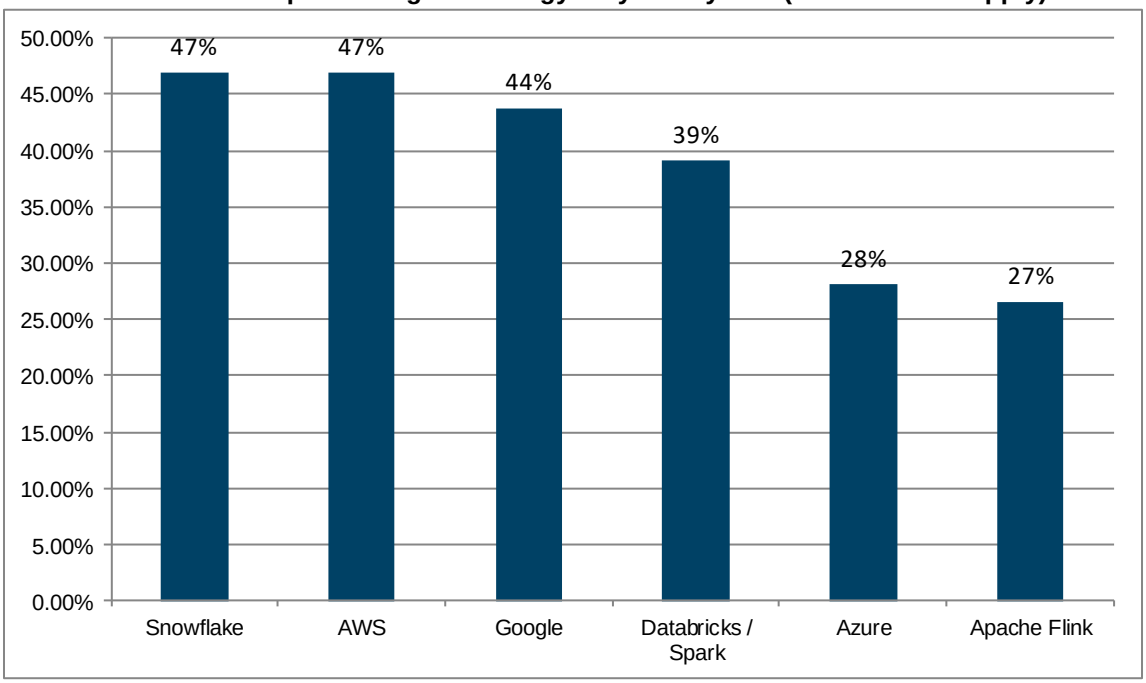


n = 66

Source: William Blair Equtiy Research

When it comes to leveraging real-time data for faster insights, more companies are adopting data streaming and stream processing technologies. While integrating data across data siloes remains a key use-case for data streaming technology (with 46% of respondents), the bigger use-case is driving real-time insights. The importance of processing real-time data streams has led to a number of announcements from database companies about their ability to process streams of data, typically built on top of the Kakfa open-source streaming engine. In our survey, the most popular stream processing technologies were Snowflake and AWS (with MSK), followed by Databricks/Spark (commercialized by Databricks) and Apache Flink (commercialized by Confluent).

Exhibit 8
On the Ground and in the Cloud; A Developer Technology Quarterly
Which stream processing technology do you rely on? (select all that apply)



n = 64

Source: William Blair Equity Research

Top Technology Trends in Data Analytics

We recently published six reports on the analytics market, which delve into the history, competitive dynamics, and disruptive technologies that are shaping the future of the space. These are: 1) Data Analytics in the AI Age: Everyone in the Pool! 2) Snowflake Deep Dive, 3) Snowflake Summit, 4) Databricks Summit, 5) Microsoft Deep Dive, and 6) Takeaways From Expert Call on Data Analytics.

For the purposes of this quarterly, we zone in on what we believe are some of the most disruptive of these technical trends, namely GenAI, data lakehouses, open table formats, metadata catalogs, and zero ETL.

- **Generative AI.** The analytics space has long been an area that requires specific expertise in data science and data engineering. To perform analytics, users developed skills on how to aggregate, transform, query, and analyze different datasets. The advent of GenAI opens the usage of analytics to a broader range of personas by creating a simple natural language interface for asking questions about data and returning an easily understood response. This allows business analysts to tackle higher-value tasks beyond simply retrieving and visualizing data (where today they still spend much of their time). GenAI also massively grows the pool of data available for analysis by allowing NLP (natural language processing) interfaces/prompt boxes to be applied to more datasets and silos.

  For structured data (e.g., SQL-based database records or point-of-sale data), GenAI involves translating natural language queries into traditional SQL queries, which return an answer in SQL that is then converted back into natural language. This is more of an evolutionary shift that should democratize the analytics function in an organization, simplifying responses and reducing the need to rely on traditional BI reports and charts (the traditional purview of tools like Tableau and the bane of many business analysts).

  For unstructured data, the impact of GenAI is more revolutionary. Traditional approaches to processing and analyzing unstructured data have been burdensome and expensive, generally requiring the unstructured data to be converted in some way to structured data. In the GenAI realm, to prepare data for LLM training and prompt engineering, organizations need to employ specialized tools like Snorkel AI or Scale AI. By organizing vast amounts of unstructured

data (e.g., images, audio, files, which IDC estimates accounts for 80%-90% of total enterprise data and is growing much faster than structured data) and making it accessible to model training and inference systems, end-users can ask questions of unstructured data in a way that was not possible previously.

Ultimately, we view analytics as one of the prime beneficiaries from the advent of AI/ML technology, as it will simplify query-ability of enterprise data at scale through NLP—making analytics vastly more accessible and user-friendly. The democratization of analytics through GenAI should allow individuals within departments to explore and analyze data independently, leading to faster insights and more informed decision-making. Put simply, anyone can be a data analyst now!

- **Data Lakehouses**. The analytics industry is coalescing around the concept of a data lakehouse, which combines the best features of a data lake and a data warehouse in a single platform. The benefit here for customers is the ability to consolidate all types of data in a single analytics repository and streamline experiences across different personas in the organization. By uniting two separate analytics systems—a data warehouse for structured data and a data lake for unstructured data—a lakehouse eliminates the need to move data between systems and enables querying across all data types and sets. As companies seek to leverage the benefits of AI, a data lakehouse can also offer AI models a single source of truth and a more comprehensive view of training data.

  At a technical level, lakehouses separate storage and compute processes, allowing each to scale independently. This allows customers to store as much data as needed without worrying about compute resources and scale up compute power as needed without paying for additional storage. Another key feature of lakehouses is the ability to execute ACID transactions; that is, when multiple users are reading and writing data at the same time, lakehouses guarantee data consistency. Lastly, lakehouses offer advanced data management features, including fine-grained access control and auditing to ensure strict data governance and security, and snapshots and "time travel" that are important for production use-cases (enabling rollback of changes).

  The rise of lakehouses is blurring traditional analytics swim lanes and putting traditionally noncompetitive vendors like Snowflake and Databricks on a collision course. We note that Databricks recently disclosed that its SQL data warehousing product grew more than 100% to a run-rate of more than $400 million as of June 2024. Elsewhere, Snowflake expects that its Snowpark product (which competes directly against Spark-based systems like Databricks) will represent 3% of product revenue in the current fiscal year (or about $100 million in revenue).

- **Open Table Formats**. Open table formats allow users to store large collections of analytical data in static files that are well structured, fast to query, modifiable, and fully standardized. Historically, most analytical data was stored in DWs in closed proprietary formats. DLs adopted a more open approach, using open file formats such as CSV, JSON, and Parquet. Open table formats such as Iceberg, Hudi, and Delta—which all came out of distinct open-source projects—have built on this openness and added extra capabilities. Specifically, these formats use a tabular structure and schema, which enables files to be treated like tables with rows and columns. This allows users to perform inserts, updates, deletes, and merge operations, as they would with any relational database table, all while guaranteeing transactional consistency (meaning that multiple users/applications can be inserting, updating, and reading from the same tables concurrently).

  By bringing the advantages of data warehousing (in the form of structured tables) to DLs, open table formats were a key enabler for the creation of data lakehouses. Because these tables are effectively raw data files, this avoids vendor lock-in, makes data migration between tools and platforms significantly easier, and allows customers to leverage low-cost cloud object storage systems like AWS S3. This effectively drives a decoupling of the compute (query) and storage layers and enables any analytics tool to access the data stored within them for both reading and writing.

  The result is that data can theoretically live anywhere, which should fundamentally reduce the traditional competitive moat (as well as the revenue opportunity) for a centralized DW and allow better integration across heterogeneous data sources. Plus, with continual improvements in query performance (due to hardware and network enhancements) and high levels of innovation across open-source query engines, the query function is increasingly becoming democratized. In addition, open table formats support both batch processing and streaming operations, making them a good fit for a world where data increasingly needs to be processed as it is generated.

Put simply, open tables allow customers to store analytical data only once—eliminating compliance and security issues around having data copies in multiple places—and query that data through any number of Iceberg-compliant compute engines. While the benefits for end-users are compelling, the growing popularity of open table formats like Iceberg and Delta has the potential to devalue both the storage layer in a traditional data warehouse (as data can be stored external to the warehouse) and the compute layer (as the need for data ingestion and ETL goes away and the decoupling of storage and compute reduces vendor pricing power).

The most popular open table formats are:

- *Apache Iceberg*: Originally created at Netflix and now a community-run open-source project, the Iceberg ecosystem is growing rapidly, with robust tooling and support from compute engines such as Apache Spark, Snowflake, Amazon Athena, Dremio, Trino (Starburst), Apache Druid, and many others. Our research suggests that Iceberg has the most momentum and user adoption for large-scale datasets stored in lakehouses.

- *Delta Lake*: An open table format that was developed by Databricks and is built on top of Apache Spark. Delta tables provide transactional consistency, schema evolution, and ACID semantics to data. Delta is the default table format in the data lakehouses of both Databricks and Microsoft Fabric (although Microsoft recently announced support for Iceberg as well).

- *Apache Hudi*: An open table format that is designed for efficient incremental data processing and streaming analytics.

As noted earlier, Snowflake recently announced general availability of Snowflake Iceberg Tables, a major shift for the company that built its business on the concept of integrating storage and compute in a cloud data warehouse. Although connectivity to externally stored data will create near-term headwinds to Snowflake's storage and compute revenue in fiscal 2025, management believes that Iceberg Tables are a clear net positive in the medium- to long-term as they will massively increase the pool of data (100 to 1000x increase in top customer cases, according to Snowflake management) available for querying on the Snowflake platform. Despite management's upbeat attitude surrounding Iceberg Tables, the lingering question is whether higher query volume could be offset by lower price per query—especially as management acknowledged that Iceberg levels the playing field for competing query engines like Trino, Presto, Spark, and Flink.

Meanwhile, Databricks' recent acquisition of Iceberg specialist Tabular for total considerations that reportedly exceed $1 billion (founders of Tabular are the original creators of the Iceberg open-source project) appears to be an admission that Iceberg is likely to be the winning format at the end of the day over its Delta Lake format. Databricks today offers Delta Lake UniForm, which aims to provide interoperability across Iceberg, Delta Lake, and Apache Hudi tables. Delta Lake UniForm supports Iceberg's REST-based interface for how compute engines talk to the metadata catalog (see below), so customers can use the analytics engines and tools they are already familiar with, across all their data. This is an ongoing effort, and Databricks (with Tabular) will work closely with the Delta Lake and Iceberg communities to develop lakehouse format compatibility. In the short term, this will be provided inside Delta Lake UniForm, and in the long term, the company aims to develop a single, open, and common standard of interoperability for lakehouse tables.

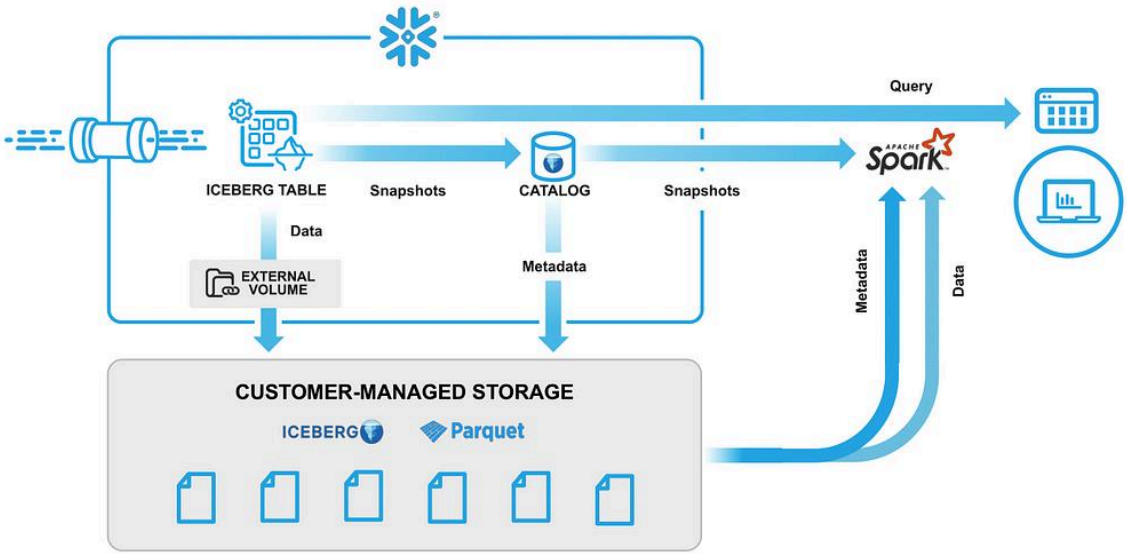
Metadata Catalogs. A critical but underappreciated piece of the data lakehouse is the metadata catalog, which governs how users and compute engines can access data and serves as a trusted inventory/directory of data, helping constituents like data scientists find the data they need for analytics purposes. The metadata from the underlying data sources is fed into the catalog, allowing it to serve as the central directory or hub for the organization's data.

Enterprises need metadata catalogs to help them understand what data assets they have and how this information has transformed over time. Metadata catalogs also are important for data governance and security, since they control the level of access that processing engines (and therefore users) gain to the underlying data. Catalogs manage a collection of tables that are usually grouped into namespaces, and tasks like creating, dropping, and renaming tables are the responsibility of the catalog.

As noted above, to drive greater openness and standardization, the open-source Iceberg community recently launched a REST-based interface for how compute engines talk to the metadata catalog. Vendors have begun to adopt this open interface for metadata catalogs, including Dremio (Project Nessie) and Snowflake (Polaris). For Snowflake customers,

this means that Polaris can be used to manage Iceberg data that is queried by non-Snowflake compute engines, including Spark, Flink, Presto, Trino, and Dremio. From Snowflake's perspective, this will significantly expand the volume of data that can be managed by the Snowflake platform, but further reduces vendor lock-in for customers. In an effort to foster broad adoption and establish Polaris as the de facto standard for Iceberg metadata catalogs, Snowflake will soon be donating the Polaris project to the Apache Software Foundation, effectively ending the metadata catalog's opportunity to be another lock-in element for customers.

Exhibit 9
On the Ground and in the Cloud; A Developer Technology Quarterly
How Iceberg Tables Work



Source: Snowflake, William Blair Equity Research

These moves by Snowflake and others have led Databricks to open source its Unity Catalog, which is the metadata catalog that the company developed for use with Delta Lake, and now will support Iceberg as well. According to Databricks CEO Ali Ghodsi, harmonizing Iceberg with Delta and offering open catalogs will create the "USB" for data access. That is, customers will now be able to access just one copy of analytical data in a standardized, USB-like format under the customer's ownership and governance. This should expand the analytics market, reduce lock-in, and fuel innovation.

Zero ETL. ETL is a data engineering and integration methodology that extracts raw data from source applications, transforms the data (standardizing dates, Booleans, etc.) so that it can be used by data scientists and analysts, and then loads the data into a target database or data warehouse. Traditional ETL processes are time-consuming and complex to develop, maintain, and scale, requiring often brittle point-to-point data pipelines.

As the name suggests, zero ETL eliminates or minimizes the need to build physical ETL data pipelines, often enabling querying across data silos without the need for data movement. Open table formats enable a shift toward zero ETL by creating a uniform data structure between source and sink that affords simpler, lightweight integration versus more traditional ETL solutions that often require complex transformations.

As an example, with zero ETL (aka "zero copy" in Salesforce parlance), Salesforce Data Cloud allows businesses to keep their data in its original location while integrating the metadata in a central catalog. Today, SFDC enables zero ETL integration with data inside Amazon Web Services (AWS), Databricks, Google Cloud, Microsoft, and Snowflake. While some level of data replication is still involved with this approach, zero ETL lets SFDC customers access data where it lives—either through queries or by virtually accessing the file. When source data changes, such as on a vehicle record that displays the mileage and maintenance history of a connected car, it is immediately updated in the SFDC lake that houses the data.

Data Fabrics. To address the challenge of data fragmentation as characterized by relentless growth of isolated data silos and multicloud environments, we are seeing vendors push the concept of a data fabric. This concept aims to centralize data management, governance, and security across various tools, experiences, and personas—using a single dashboard in some cases—but does not require the data itself to be physically centralized in a single data store (it can be virtualized across different locations/clouds using technologies like connectors, APIs, and Iceberg/Delta tables). Here, open table formats and zero ETL become key enablers, because they create standardized data architectures that can improve the ability to integrate and ingest different data into a unified platform. Fabrics do this by helping organizations better manage and gain visibility into their highly distributed data estates through a common metadata or semantic layer that facilitates easy access to those data assets in a self-service manner.

We mentioned above how SFDC leverages this fabric concept. Meanwhile, Microsoft recently launched its Fabric platform, which consolidates numerous standalone Microsoft products into a unified analytics platform that purports to reduce costs, streamline analytics workflow, and enable more efficient collaboration among data users. Other large providers pushing a similar vision include Google (Dataplex), IBM (watsonx.data), HPE (Ezmeral Data Fabric), Informatica (Intelligent Data Management Cloud), and Oracle (Oracle Analytics).

In addition, analytics specialists like Starburst and Dremio have been key proponents of data decentralization and fabric-like models. With Starburst, for example, end-users do not need to know or care if a data table lives in one system or another—Starburst just points connectors at the storage silo. Under the hood, Starburst can join two tables from different data sources by virtualizing the data layer and abstracting the query above it. This is key for use-cases like risk analytics, customer 360, and fraud detection, which all require data that resides across multiple sources. For example, Starburst customer Comcast joins viewer behavior stored in a data lake with billing data stored in a data warehouse to correlate TV shows with spending and personalize customer promotions. To do this traditionally, Comcast had to move all of that data into a DW first. With Starburst, Comcast can simply connect to the different data sources and create a virtual data view that joins data across the different systems.

Microsoft Build Takeaways

Microsoft hosted its annual developer Build conference in mid-May, where the company announced a slew of products, services, and expanded partnerships that demonstrate the company's leadership in enterprise-grade AI.

CEO Satya Nadella set the tone of the event early, emphasizing the pivotal role of AI in shaping our digital landscape as we enter what he referred to as a "golden age of systems," an era characterized by the convergence of infrastructure, data, tooling, and applications. In his keynote, Nadella highlighted three major platform developments in the last year: 1) Microsoft Copilot, the everyday AI companion, 2) Copilot Stack, the platform to build copilots and develop AI solutions, and 3) Copilot+PCs, a new category of PCs, designed to allow developers to deliver differentiated AI experiences locally on their machines.

Copilot: Since Microsoft 365 Copilot's launch in November 2023, Microsoft has already released more than 150 updates and disclosed that nearly 60% of the *Fortune* 500 are customers. At Build, the company unveiled Team Copilot, an expansion of Copilot to act as a collaborator and project manager, enabling agenda management, note-taking, chat moderation, and contextual question-answering. An extensive rollout of these features is scheduled for late 2024, at which point, users will be able to invoke Team Copilot from wherever they collaborate, whether it be Teams, Loop, or Planner.

Doubling down on AI-integrated services, Microsoft also announced *Copilot Connectors* in public preview, allowing developers to ground Copilot with data from a wide range of sources, i.e., the user, their teams, Microsoft Fabric, and third-party connectors from Atlassian, Snowflake, Databricks, and more.

In a rapidly shifting technological landscape, Microsoft is positioning Copilot as an additional layer of abstraction that adjusts to and incorporates model improvements, enabling developers to focus on building features for end-users. Copilot is integrated across the Microsoft ecosystem, leveraging memory and knowledge for context regardless of the user environment and learning based on user feedback.

Copilot Stack: Microsoft announced a set of new capabilities to Copilot Studio for the development of customized

copilots. Copilot can now be built with AI agent capabilities to proactively respond to data and events tailored to specific tasks and functions. In addition, copilot extensions, which include plugins and connectors, allow customers to enhance their copilots by connecting them to new data sources and applications.

Copilot + PC: Perhaps the most unexpected new announcement from Build was the introduction of Copilot+PC, a Windows-based hardware solution aimed at enhancing the PC experience for developers. Management referred to this new category of PCs as the fastest, most AI-ready PCs ever built. The lineup includes the Surface Pro and Surface Laptop, with alternatives from OEM partners ASUS, Dell, Samsung, Acer, Lenovo, and HP on the way. The new line of AI-powered PCs is powered by Qualcomm's Snapdragon X Series chips that boast a turbocharged neural processing unit (NPU)—a specialized chip for AI-intensive processes like real-time translation and image generation—that can perform more than 40 trillion operations per second.

One standout feature of Copilot+PCs is Recall, a snapshot-based and AI-powered photographic memory that captures everything users see or do on the device. Recall effectively acts as an AI explorer, presenting an explorable timeline to give users a snapshot of a period and provide context of the memory. While the feature presents obvious privacy concerns, Microsoft has emphasized that Recall's processing occurs locally on-device and that users can control what the machine stores through the Windows Semantic Index tool. We note that the company has changed course since the Build conference, announcing that the default setting for Recall will be off—i.e., customers will need to opt in to Recall upon setting up their Copilot+PC experience.

GitHub Copilot: While GitHub has its own user conference, GitHub Universe, which will be held in October 2024, Microsoft did introduce GitHub Copilot Extensions. This allows developers and organizations to customize their GitHub Copilot experience with their preferred services such as Azure, Docker, Sentry, and more. Currently, GitHub Copilot stands as the most widely adopted AI developer tool, with 1.8 million paid subscribers across 50,000 organizations at the end of its April 2024 quarter.

New Models and Multimodal Capabilities in Azure AI: While the various Copilots today leverage closed source models from OpenAI, Microsoft is heavily leaning into the extensibility and customization of its experiences with open small language models (SLMs) to address enterprise-grade AI use-cases as well as consumer-centric ones. On that front, Microsoft unveiled Phi-3-vision, a new multimodal model in the Phi-3 family of SLMs. Phi-3 models are cost-effective and optimized for low-computing intensive tasks on personal devices, with Phi-3-vision's multimodality extending to text and images, though it cannot generate images itself yet. Developers can experiment with these models in Azure AI Studio along with GPT-4o, OpenAI's newest flagship model (also available as an API).

Microsoft Fabric: At Build, Microsoft outlined the four key pillars that Fabric delivers on: 1) a complete SaaS analytics platform, encompassing data ingestion, data storage, data engineering and data science, and real-time business intelligence; 2) an open, data lake-centric architecture, reducing data silos while fostering collaboration; 3) empowering business users with real-time analytics and Power BI tools integrated across the stack; and 4) an AI-powered platform with Copilot capabilities integrated in every user experience.

Microsoft unveiled a number of updates to Fabric, including support for open-format Apache Iceberg tables in OneLake (which comes in addition to the default Delta Lake format) and bidirectional data access between Snowflake and Fabric (data written by either platform will be available in both data table formats and natively stored in OneLake). In addition, Microsoft unveiled Real-Time Intelligence for Fabric, a SaaS solution that enables customers to act on large volumes of time-sensitive, highly granular data. Already in preview, Real-Time Intelligence for Fabric allows user to access simple low-code/no-code experiences through a Real-Time Hub that acts as a single place to ingest, process, and route events into Fabric.

For more detail, please see our recent deep dive note on Microsoft Fabric here.

Partnerships: In addition to Microsoft's expanded partnership with Snowflake, the company announced a set of new partnerships and extended collaborations. First and foremost, Microsoft referred to OpenAI as its most strategic and most important partnership shortly before Sam Altman, CEO of OpenAI, made a special appearance during the keynote. Microsoft is also partnering with Meta to bring Windows Volumetric Apps to Meta Quest headsets in an effort to extend Windows apps into the 3D space. On the AI front, the company extended its collaboration with Hugging Face to integrate its LLMs into Azure AI Studio. Lastly, perhaps the most interesting partnership announced was with Khan

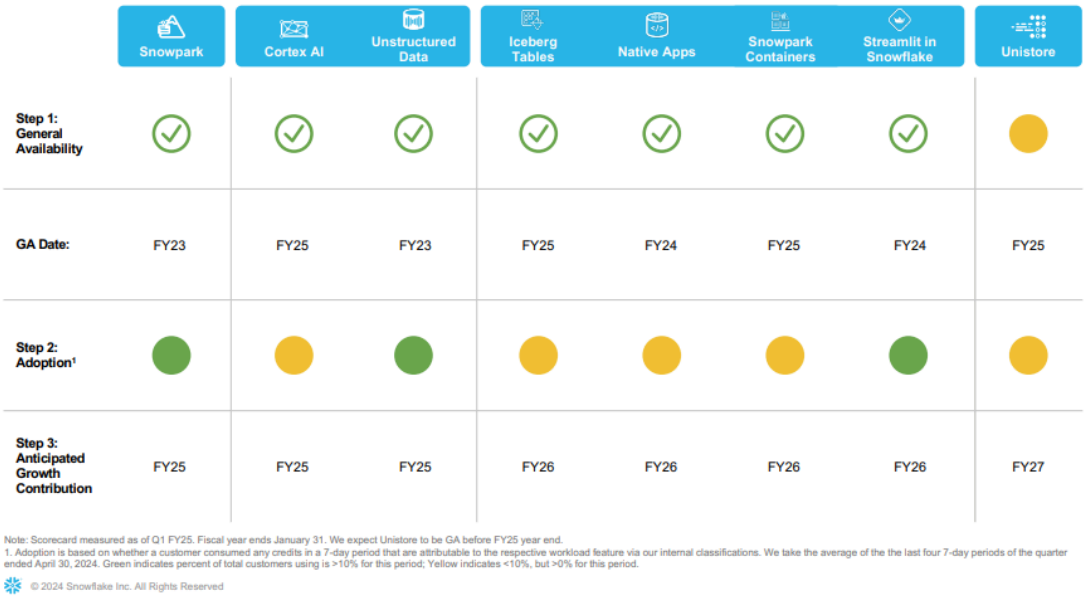
Academy. Microsoft will enable Khan to offer all K-12 U.S.-based educators free access to Khanmigo for Teachers, an AI-powered teaching assistant offered through the Azure OpenAI service.

Snowflake Summit Takeaways

In early June, Snowflake hosted its annual Data Cloud Summit and investor day, where management laid out its vision of expanding the Snowflake platform beyond the core data warehousing (DW) function to become a one-stop shop for all things data. To achieve this vision, management pointed to accelerated product innovation and delivery amid technical changes in the analytics market and the GenAI platform shift.

While management did not announce specific revenue targets for its new offerings (beyond Snowpark's $100 million target in fiscal 2025), it presented a rough timeline of likely customer adoption and revenue impact across the various new products (see exhibit below). More specifically, the company does not expect significant material contribution from most of these new offerings until fiscal 2026 at the earliest, with the currently aggressive R&D and S&M investment posture pressuring margins in the near term.

Exhibit 10
On the Ground and in the Cloud; A Developer Technology Quarterly
Availability and Expected Impact of Snowflake New Products



Source: Snowflake, William Blair Equity Research

Detailed Takeaways

- **Core Business Remains Healthy.** Despite rising concerns around the changing dynamics in the analytics market, Snowflake management spoke to the resilience of its core DW business, built on ease-of-use and best-in-class query performance. Management expects the core business TAM to double from $152 billion in calendar 2023 to $342 billion in calendar 2028, highlighting unyielding enterprise appetite for data analytics, migration to cloud, and data gravity. Snowflake also cited still low penetration of global 2000 (G2K) customers (currently around 7%), which currently hold a higher net retention rate than the overall business despite being less than 10% of the total customer base.

- **New CEO Vision: Building Unified Platform with Expanded Interoperability.** Three months into his tenure, CEO Sridhar Ramaswamy is focused on capitalizing on Snowflake's growth opportunity through 1) accelerated product delivery, made possible by building on the company's strong foundational infrastructure, and 2) driving greater customer adoption, with recent go-to-market changes to consumption-based sales incentives aimed at landing high-quality logos, "heavy" workloads, and ultimately increased consumption.

  At a high level, Snowflake strives to become the central nervous system in the cloud for an enterprise's data needs by creating a unified platform with built-in data interoperability. This vision builds upon the foundation of its core data query engine, with the company looking to expand its reach with 1) enterprise AI, 2) data sharing and collaboration (~one-fourth of customers sharing data today), and 3) refined methods for application deployment and distribution. Based on its unified, cloud-agnostic platform approach combined with its ease-of-use value proposition, management is confident in its ability to gain wallet share in the broader public cloud services market (projected TAM of $2.23 trillion by fiscal 2033).

- **Open Sesame.** Snowflake announced the general availability of Snowflake Iceberg Tables, which enables customers to store data externally outside of Snowflake in low-cost storage services like AWS S3 while still having access to Snowflake's best-in-class compute engine. Initially unveiled at Snowflake Summit 2022, management has touted Iceberg Tables one of the many drivers for appealing to a larger data audience amid the intensifying competition within the core DW market. Although connectivity to externally stored data will create a near-term headwind to storage and compute revenue in fiscal 2025, management believes that Iceberg Tables are a clear net positive in the medium to long term as they will massively increase the pool of data (100 to 1000 times increase in top customer cases) available for querying on the Snowflake platform. Despite management's upbeat attitude surrounding Iceberg Tables, the lingering question is whether higher query volume could be offset by lower price per query—especially as management acknowledged that Iceberg levels the playing field for competing query engines.

- **Capitalizing on AI Opportunity.** GenAI is the No. 1 focus for Snowflake as it aims to enhance the capabilities of its existing products and expand into new areas like machine learning. Despite a late start, Snowflake has significantly advanced its GenAI capabilities over the past year, with Snowflake Cortex, Snowpark Container Services, and Snowflake Arctic all either announced or released as generally available in fiscal 2025. From a near-term impact standpoint, management believes that Cortex could begin generating revenue this fiscal year, though nothing has thus far been baked into guidance. Revenue contribution from the broader AI portfolio should be more meaningful starting next year as customers still ponder their AI deployments and Snowflake's product set matures.

  Another major announcement at the conference was the availability in public preview of Snowflake Notebooks, which management conceded was a major hole in its AI/ML portfolio that it needed to fill. ML notebooks provide an interactive computational environment for developing data science applications, combining software code, computational output, explanatory text, and content in a single document. With Streamlit as the key underlying user interface, Snowflake Notebooks allow programmers to document and demonstrate coding workflows or simply experiment with code. Management expects Notebooks to be a driver of Snowpark adoption, more specifically by attracting data scientists to build their ML workloads on the Snowflake platform and utilize the company's superior data processing engine.

  Elsewhere, Snowflake furthered its partnership with NVIDIA by adopting the NVIDIA AI Enterprise platform into Snowflake Cortex, which provides access to key NVIDIA enterprise AI software such as NeMo Retriever microservices or NVIDIA Triton Inference. In addition, Snowflake Arctic is now available as a NVIDIA Inference Microservice (NIM), allowing it to run on top of any configuration of NVIDIA infrastructure, and supported with NVIDIA TensorRT-LLM software, which helps optimize model performance. The expanded collaboration with NVIDIA highlights the aggressive posture Snowflake is taking to bolster its AI capabilities.

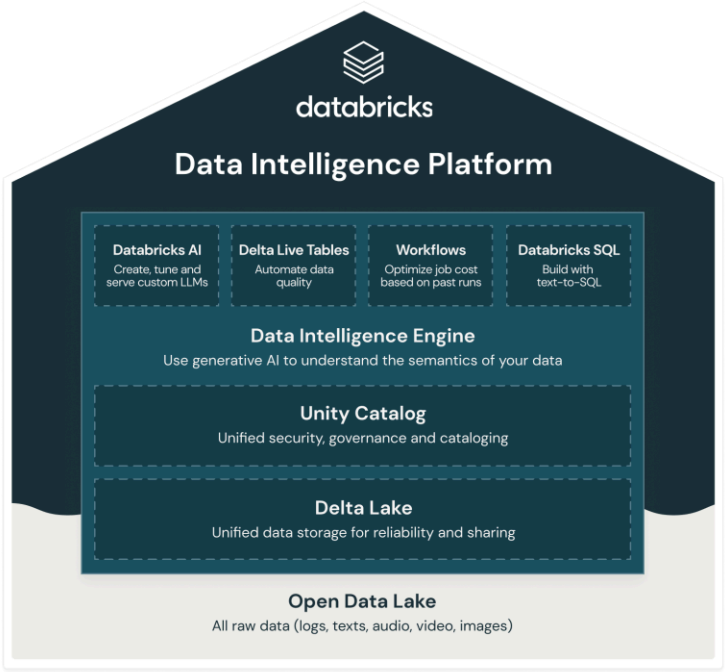
  For more detail, please see our recent deep dive note on Snowflake's new products here.

Databricks Data + AI Summit Takeaways

Not to be outdone, a week following Snowflake's conference in mid-June, Databricks hosted its annual conference and investor event. Similar to Snowflake, Databricks laid out its goal to become the central data platform for both analytics and AI use-cases and unveiled a number of new products that improve the ease-of-use of the Databricks platform.

Notably, management relayed that customers are prioritizing investments in their data estates and AI strategies, which has allowed Databricks to successfully sidestep budget headwinds that have afflicted other vendors. This propelled an acceleration in the company's first-quarter revenue growth to mid-60% compared to 50% growth at the end of fiscal 2024 (January year-end). At of the end of July 2024, Databricks expects to have a roughly $2.4 billion revenue run-rate and NRR above 140%. This strong growth is in part a function of Databrick's expanding portfolio, with the addition of traditional data warehousing capabilities (SQL Warehouse) that is today a $400 million business, as well as new AI capabilities (MosaicML, Notebooks, and the newly announced AI/BI and LakeFlow products).

Exhibit 11
On the Ground and in the Cloud; A Developer Technology Quarterly
Databricks Data Intelligence Platform



Source: Databricks

Detailed Takeaways

- **Consolidating on the Lakehouse.** While Databricks was first to espouse the concept of a lakehouse, we have seen growing adoption of this consolidation approach across all vendors in the analytics space (e.g., Snowflake, Oracle, Microsoft). In particular, Databricks is able to address different data types (structured and unstructured data), data formats (Delta Lake and Iceberg), and workload types (traditional BI workloads as well as AI data pipelines). By centralizing data, not only do customers reduce the costs of managing different technologies and stacks, but also benefit from a single security and governance framework and the flexibility to choose the infrastructure that best supports their lakehouse. During the conference Databricks noted that customers like National Australia Bank that have decided to consolidate their multiple analytics stores into the Databricks lakehouse.

- **Tabular Acquisition and Iceberg.** Databricks announced its planned acquisition of Tabular for a total cash outflow of a "few hundred million" dollars, with the remaining portion of the ~$1 billion price tag tied to performance metrics for the team. Both Snowflake and Confluent were also rumored to be in talks with the company. Tabular, which was founded by the creators of the open-source Iceberg table format and is the most significant contributor to the project,

adds not only a critical team of developers to the Databricks bench, but also helps Databricks unlock a large portion of the market that has preferred the Iceberg open table format versus Databrick's own open Delta Lake format.

- **Robust Growth in Traditional Data Warehousing.** As Databricks and Snowflake have expanded their purview and have moved into each other's swim lanes, our sense is that it is Databricks that has had more measurable success. Databricks' cloud data warehouse offering has revenue of $400 million business today versus Snowpark from Snowflake (the company's data lake offering) expected to generate only $100 million in revenue in 2024. While the two vendors' approaches are largely the same—i.e., building a singular query engine that can be performant across structured and unstructured data—our sense is that Databricks has had an easier time moving into Snowflake's swim lane than vice versa. Under the hood, the SQL data warehouse is a subsegment of a broader lakehouse, where data has already been transformed and cleaned up to give it the structure needed for SQL queries. Nonetheless, to accomplish this, Databricks had to rebuild its Spark-based query engine into a proprietary query engine called Photon (though it remains compatible with Spark APIs).

- **The Rise of Compound AI Models.** As the two companies increasingly move into each other's swim lanes, Databricks has had more measurable success, with its cloud data warehouse offering a $400 million business today versus Snowpark from Snowflake (the company's data lake offering) expected to generate only $100 million in revenue in 2024. While the two vendors' approaches are largely the same—i.e., building a singular query engine that can be performant across structured and unstructured data—our sense is that Databricks has had an easier time moving into Snowflake's swim lane than vice versa. Under the hood, the SQL data warehouse is a subsegment of a broader lakehouse, where data has already been transformed and cleaned up to give it the structure needed for SQL queries. Nonetheless, to accomplish this, Databricks had to rebuild its Spark-based query engine into a proprietary query engine called Photon (though it remains compatible with Spark APIs).

- **Watch Out ETL, LakeFlow Is Here.** Databricks unveiled a new capability called LakeFlow, the feature that management appears most excited about in terms of the revenue opportunity going forward. LakeFlow allows enterprises to create productionized data pipelines (a key element of companies' AI data strategies, as discussed in our recent white paper on enterprise GenAI, that connects data from various sources into Databricks). The company is starting with connectors, built on top of the Arcion technology acquired in November 2023, to connect Databricks to operational databases like Amazon RDS, Azure SQL, Oracle, Postgres, and MySQL, as well as enterprise applications like Salesforce, Microsoft Dynamics/Sharepoint, Workday, and Oracle NetSuite. These connectors are integrated into the Unity Catalog, enabling unified governance and management of different data types. LakeFlow Pipelines help replace traditional ETL solutions and can perform both stream and batch data transformations, competing with other real-time data processing solutions like Apache Flink (Confluent) and Snowflake Streams. Databricks reinforced the view that a real-time data strategy and the ability to process this data rapidly is becoming an important pillar of all data strategies.

  For more details, please see our recent note on Databricks conference and summit here.

Data Analytics Expert Call

Our expert call with analytics consultant and Snowflake Elite partner Paul Corning on June 27 reinforced our view that the analytics market is in the midst of significant upheaval, driven by new technologies like generative AI, data lakehouses, and open table formats.

Detailed Takeaways:

- **From BI to AI**. The analytics market has historically been 80% focused on structured data, with only 20% focused on unstructured data and data science use-cases. With advent of GenAI, Corning expects the market will shift over the next 3-5 years to being 90% focused on unstructured data and AI. Corning believes that organizations have faced a value ceiling with BI and traditional data warehousing, given the lack of predictive insights and largely historical view that BI provides. From a vendor perspective, Databricks has the most market momentum given its expertise in machine learning/data science and unstructured use-cases.

- **Analytics 2.0**. After soft demand in 2022/2023, GenAI has re-energized the analytics market and driven higher customer engagement. Corning believes we have moved into the "Analytics 2.0" phase of the market, characterized by enterprise focus on AI, which should create a burst of new analytics workloads as organizations aim to tap into their large troves of unstructured data. To date, production AI apps have been more vertically oriented, though we have seen some early horizontal use-cases such as anomaly detection, document and video generation, and low-code NLP solutions. But for most companies, AI is still in the planning and experimentation phase—especially given the hot mess that exists inside most data estates.

- **Rise of the Metadata Catalog**. To address the hot mess and prep their data for this new phase of the market, customers need to architecturally rethink data estates, and specifically how data can be organized and trusted within their organizations for analytics/AI purposes. This has massively elevated the importance of metadata catalogs, which enable centralized data governance and security, providing customers with a single point of control to track where data lives (and where it comes from). Once an organization's data is consistently managed and governed, it becomes much easier to build and deploy AI/analytics workloads.

- **Open Table Formats: "The Best Pipeline Is No Pipeline."** While still early in the adoption curve, the rise of open table formats like Iceberg and Delta provides customers with greater flexibility in choosing both storage and query engines for their analytical workloads. Corning believes that most midsize companies prefer data stored in one location due to simplicity, while most large companies will demand some sort of physical data federation due to the impracticality and expense of centralizing data in one physical location (but with centralized governance a must). Despite fears that Snowflake's support for Iceberg tables could be a revenue headwind, Corning expects the company to clearly benefit long term from the ability of customers to apply analytics to data wherever it lives.

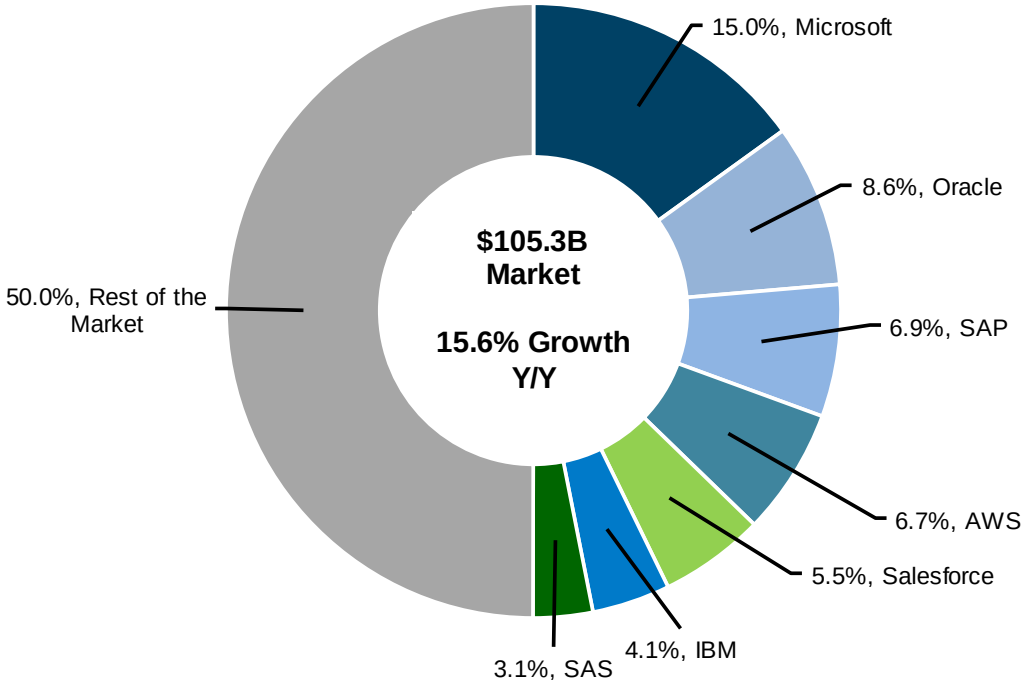
- **Room for Multiple Winners.** While investors intensely focus on the Snowflake-Databricks rivalry—and customers are clearly looking to reduce the number of analytics vendors in their stacks—Corning believes that most large customers are not looking to choose a single vendor for their analytics needs. The reality is that customers are likely to retain multiple analytics vendors in their environments, depending on use-case, prior investments, data incumbency, and product-market fit.

Data Analytics Market Size and Share

According to the latest report from market researcher IDC, the Big Data and Analytics (BDA) software market saw total revenues of $104.9 billion in 2022, or year-over-year growth of 15.7%. Microsoft is the current leader in the BDA software market with $15.8 billion of revenue in 2022, or market share of 15%, followed by Oracle, SAP, AWS, and Salesforce rounding out the top 5. While database incumbents like Microsoft and Oracle hold substantial market share, the growth in cloud-based analytics and in-app analytics has helped drive market share for companies like SAP, AWS, and Salesforce. Importantly, with Snowflake crossing $3.5 billion in revenue in 2023, it will move up into the top 8 vendors on this list.

Despite a handful of leaders in the BDA market, the space remains highly fragmented with over 50% of the market split across a long tail of vendors with 2% market share or less each. Exhibit 12 below shows the market share leaders in the BDA software market from 2022 (the latest available data).

Exhibit 12
On the Ground and in the Cloud; A Developer Technology Quarterly
Big Data and Analytics Market Share, 2022



Source: William Blair Equity Research analysis based on IDC estimates, July 2023

Looking ahead—in part driven by the advent of gen AI technology—IDC projects the aggregate BDA market to reach $253 billion by 2027, a five-year CAGR of 19.3% (see exhibit 14 below). IDC classifies the BDA software market into five primary segments:

1. ***Business intelligence (BI) and analytic tools and platforms*** – 21.1% of the aggregate market in 2022, expected to grow at a five-year compound annual rate of 10.0% and decline to 14.1% of the aggregate market by 2027. This segment can be further broken out into a) BI software and b) advanced and predictive analytics (APA) software. BI software covers multidimensional analysis tools to assist users with data visualization (creating dashboards or visual reports) and creating reports through queries. APA software covers analytics tools used to create statistical models, which include data mining, mathematical optimization, graph analytics, forecasting, and prediction capabilities. Tools like Tableau (acquired by Salesforce), Power BI (Microsoft), Alteryx, Qlik, SAP, SAS, IBM, FICO, Tibco, and Looker (acquired by Google) are included in this segment.

2. ***Enterprise performance management and analytic applications*** – 25% of the aggregate market in 2022, expected to grow at a five-year compound annual rate of 11.9% and decline to 18.2% of the aggregate market by 2027. This segment encompasses prepackaged analytics applications that can be used cross-industry as enterprise performance management applications, or as general-purpose business analytics applications including customer relationship analytic applications, production planning applications, services operations analytic applications, supply chain and product analytic applications, and workforce analytic applications. Vendors in this space include Salesforce, SAP, IBM, Workday, Anaplan, OneStream, and AWS.

3. ***Analytical data management and integration platforms*** – 34% of the aggregate market in 2022, expected to grow at a five-year compound annual rate of 17.8% and decline to 31.9% of the aggregate market by 2027. This segment covers analytic data integration and integrity tools, continuous analytic tools, nonrelational analytic data stores, and

relational DWs. Vendors in this segment include Oracle, Snowflake, Databricks, Confluent, Informatica, Teradata, Cloudera, Microsoft, AWS, Google, Precisely, SAP, Talend, SAS, Alteryx, Denodo, Starburst, Alation, dbt Labs, Fivetran, and IBM.

4. ***Artificial intelligence (AI) software platforms*** – 17.9% of the aggregate market in 2022, expected to grow at a five-year compound annual rate of 36% and approximately double in size to 34.6% of the aggregate BDA market by 2027. AI software platforms includes AI lifecycle software, AI software services, and search and knowledge discovery software. Technologies within this segment contribute to forming AI models and applications as users look to draw the utmost value and insights from their data. Vendors in this segment include Microsoft, Google, IBM, AWS, Palantir, SAS, Dataiku, Amelia, MathWorks, and C3.ai.

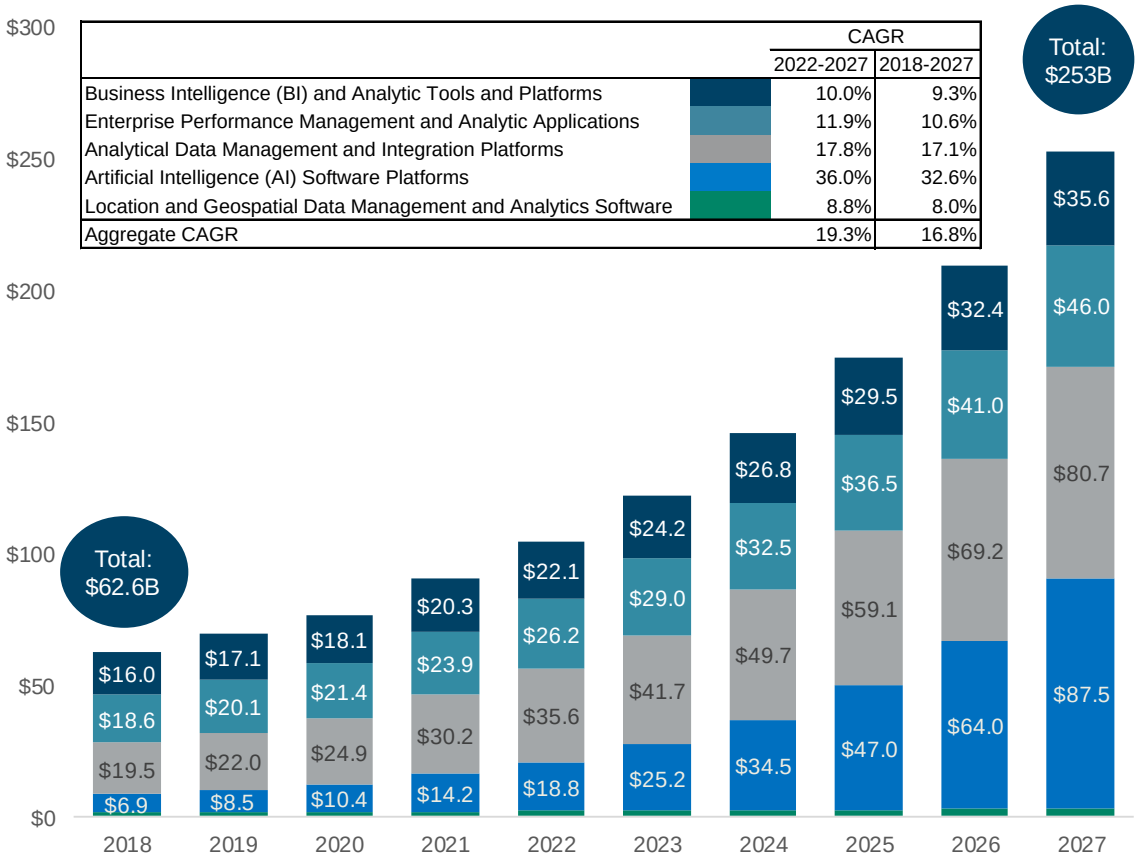
5. ***Location and geospatial data management and analytics software*** – 2% of the aggregate market in 2022, expected to grow at a five-year compound annual rate of 8.8% and decline to 1.3% of the aggregate market by 2027. Software within this segment can be broken into three separate categories: location and geospatial analytics; visualization, mapping, and navigation capabilities; and location-enabling developer tools and platforms. All three subsegments integrate a location and/or geospatial component when analyzing or visualizing data. While location and geospatial data management and analytics software is often integrated within the platforms of the major CSPs and analytics players, a few prominent standalone location and geospatial analytics vendors include Esri, Galigeo, and Carto.

Exhibit 13

On the Ground and in the Cloud; A Developer Technology Quarterly

IDC: Big Data and Analytics (BDA) Forecast 2018-2027

	CAGR 2022-2027	CAGR 2018-2027
Business Intelligence (BI) and Analytic Tools and Platforms	10.0%	9.3%
Enterprise Performance Management and Analytic Applications	11.9%	10.6%
Analytical Data Management and Integration Platforms	17.8%	17.1%
Artificial Intelligence (AI) Software Platforms	36.0%	32.6%
Location and Geospatial Data Management and Analytics Software	8.8%	8.0%
Aggregate CAGR	19.3%	16.8%



Note: All numbers in billions unless otherwise specified

Source: Based on data from IDC; Worldwide Big Data and Analytics Forecast, 2022-2027, Doc #50117823, July 2023

Private Funding Activity in Data Analytics

In this section, we analyze private funding activity for data analytics companies, looking at trends over the past few years to gauge the health of private investments in the big data and analytics market. In addition, we highlight several interesting vendors that we think investors should know more about.

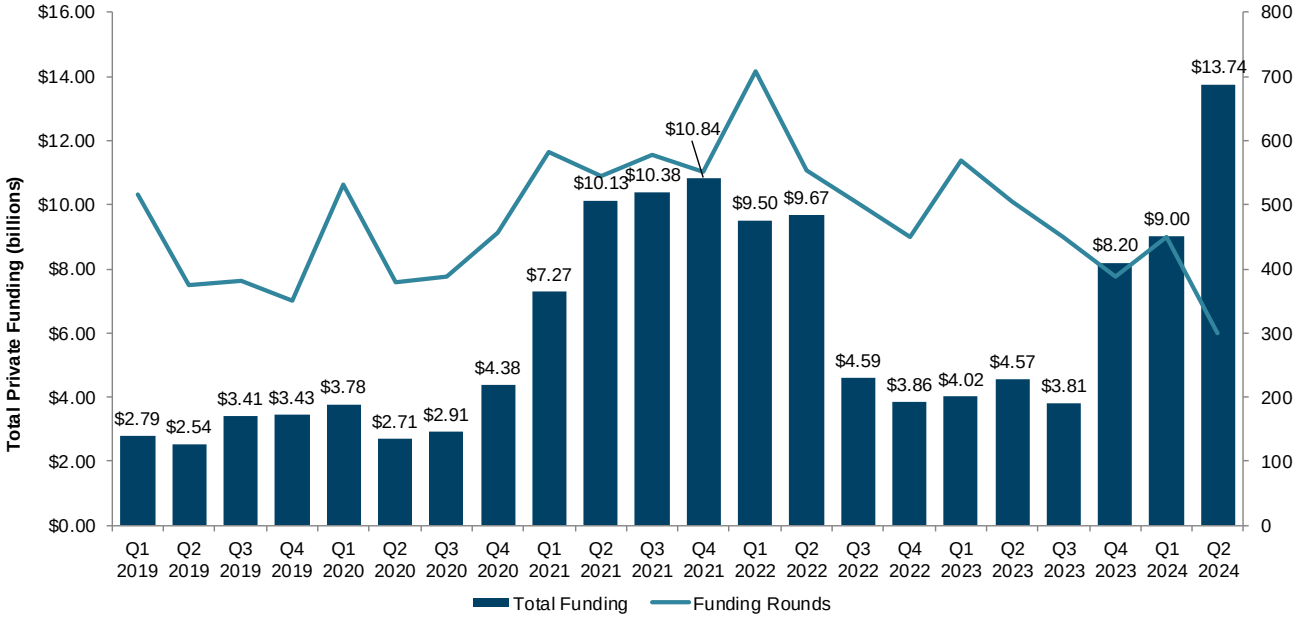
Private Funding Improving from Post-COVID Hangover

In the exhibit below, we detail private market monthly funding for big data and analytic companies since 2019. In particular, our query focuses on companies based in the United States that are characterized as either data analytics, data engineering, or big data companies. Total funding for these companies was $12.2 billion in 2019 and $13.8 billion in 2020. Given the burst of private capital following the COVID slowdown, total funding nearly tripled to $38.6 billion in 2021 (up 180% year-over-year).

While 2022 (down 28% year-over-year) and 2023 (down 25% year-over-year) have seen funding above pre-COVID levels, total private funding has been affected by elongated macro uncertainty and sustained high interest rates. Although we continue to see private markets affected from the continuation of these factors, private funding in 2024 has gotten off to a strong start due to an increased focus in data technologies as it relates to GenAI as well as new market trends like open table formats and the emergence of lakehouses.

The recent strength has been illustrated with the second quarter of 2024 (up to June 15) seeing a record-high total funding raised in a quarter of $13.7 billion. Private funding in the first half of 2024 will grow upward of 165% year-over-year as a result.

Exhibit 14
On the Ground and in the Cloud; A Developer Technology Quarterly
Quarterly Private Funding for Data Companies



*Q2 2024 data up until June 15th, 2024

Source: PitchBook Data, Inc., and William Blair Equity Research

Select BDA companies Receiving Funding Since 2021

We have discussed a host of private database and data streaming companies in past research reports. Below we highlight private BDA companies (in alphabetical order) that have raised sizable rounds over the last few years that we believe deserve airtime. We note that many of these companies raised money under much more favorable valuation conditions than exist today.

- Aerospike: raised $109 million in April 2024
- Airbyte: raised $150 million at a $1.5 billion valuation in December 2021

- Alation: raised $123 million at a $1.7 billion valuation in November 2022
- Atlan: raised $105 million at a $750 million valuation in May 2024
- Census: raised over $60 million at a $630 million valuation in February 2022
- ClickHouse: raised $250 million at a $2 billion valuation in October 2021
- Cloudera: went private in a $5.3 billion transaction in October 2021
- Collibra: raised $250 million at a $5.25 billion valuation in November 2021
- Databricks: raised $500 million at a $43 billion valuation in September 2023
- Data.world: raised $50 million in April 2022
- dbt Labs: raised $222 million at a $4.2 billion valuation in February 2022
- Dremio: raised $160 million at a $2 billion valuation in January 2022
- Firebolt: raised $100 million at a $1.4 billion valuation in January 2022
- Fivetran: raised $565 million at a $5.6 billion valuation in September 2021
- Hightouch: raised $38 million in July 2023
- Imply: raised $100 million at a $1.1 billion valuation in May 2022
- InfluxData: raised $81 million in February 2023
- MotherDuck: raised $52.5 million at a $400 million valuation in September 2023
- Onehouse: raised $25 million in February 2023
- Rivery: raised $30 million in May 2022
- SingleStore: raised $146 million at a valuation of over $1 billion in October 2022
- Starburst: raised $250 million in February 2022
- StarRocks: raised $40 million in February 2021
- StarTree: raised $47 million at a $300 million valuation in August 2022
- ThoughtSpot: raised $100 million at a $4.2 billion valuation in November 2021
- Yellowbrick Data: raised $75 million in November 2021

The prices of the common stock of public companies mentioned in this report follow:

Alphabet Inc. (aka Google) (Outperform)	$183.42
Amazon.com, Inc. (Outperform)	$193.25
C3.ai, Inc.	$28.95
Confluent, Inc. (Outperform)	$29.53
Dell Technologies, Inc.	$137.87
Fair Isaac Corporation	$1,488.39
Hewlett Packard Enterprise Co.	$21.17
Informatica, Inc.	$30.88
International Business Machines Corporation	$173.01
Meta Platforms, Inc. (Outperform)	$75.36
Microsoft Corporation (Outperform)	$446.95
Netflix, Inc. (Outperform)	$674.88
NVIDIA Corporation	$123.54
Oracle Corporation (Outperform)	$141.14
Palantir Technologies Inc. (Underperform)	$25.32
Qualcomm Incorporated	$199.18
Salesforce, Inc. (Outperform)	$257.10
SAP SE	$201.70
Snowflake, Inc. (Outperform)	$135.11
Teradata Corporation	$34.55
Workday, Inc. (Outperform)	$223.56

IMPORTANT DISCLOSURES

This report is available in electronic form to registered users via R*Docs™ at https://williamblairlibrary.bluematrix.com or www.williamblair.com.

Please contact us at +1 800 621 0687 or consult https://www.williamblair.com/equity-research/coverage for all disclosures.

Jason Ader, Jonathan Ho, Jake Roberge and Arjun Bhatia attests that 1) all of the views expressed in this research report accurately reflect his/her personal views about any and all of the securities and companies covered by this report, and 2) no part of his/her compensation was, is, or will be related, directly or indirectly, to the specific recommendations or views expressed by him/her in this report. We seek to update our research as appropriate. Other than certain periodical industry reports, the majority of reports are published at irregular intervals as deemed appropriate by the research analyst.

DOW JONES: 39118.90
S&P 500: 5460.48
NASDAQ: 17858.70

Additional information is available upon request.

Current Rating Distribution (as of June 30, 2024):

Coverage Universe	Percent	Inv. Banking Relationships *	Percent
Outperform (Buy)	72	Outperform (Buy)	8
Market Perform (Hold)	28	Market Perform (Hold)	1
Underperform (Sell)	1	Underperform (Sell)	0

*Percentage of companies in each rating category that are investment banking clients, defined as companies for which William Blair has received compensation for investment banking services within the past 12 months.

The compensation of the research analyst is based on a variety of factors, including performance of his or her stock recommendations; contributions to all of the firm's departments, including asset management, corporate finance, institutional sales, and retail brokerage; firm profitability; and competitive factors.

OTHER IMPORTANT DISCLOSURES

Stock ratings and valuation methodologies: William Blair & Company, L.L.C. uses a three-point system to rate stocks. Individual ratings reflect the expected performance of the stock relative to the broader market (generally the S&P 500, unless otherwise indicated) over the next 12 months. The assessment of expected performance is a function of near-, intermediate-, and long-term company fundamentals, industry outlook, confidence in earnings estimates, valuation (and our valuation methodology), and other factors. Outperform (O) - stock expected to outperform the broader market over the next 12 months; Market Perform (M) - stock expected to perform approximately in line with the broader market over the next 12 months; Underperform (U) - stock expected to underperform the broader market over the next 12 months; not rated (NR) - the stock is not currently rated. The valuation methodologies include (but are not limited to) price-to-earnings multiple (P/E), relative P/E (compared with the relevant market), P/E-to-growth-rate (PEG) ratio, market capitalization/revenue multiple, enterprise value/EBITDA ratio, discounted cash flow, and others. Stock ratings and valuation methodologies should not be used or relied upon as investment advice. Past performance is not necessarily a guide to future performance.

The ratings and valuation methodologies reflect the opinion of the individual analyst and are subject to change at any time.

Our salespeople, traders, and other professionals may provide oral or written market commentary, short-term trade ideas, or trading strategies-to our clients, prospective clients, and our trading desks-that are contrary to opinions expressed in this research report. Certain outstanding research reports may contain discussions or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Always refer to the most recent report on a company or issuer. Our asset management and trading desks may make investment decisions that are inconsistent with recommendations or views expressed in this report. We will from time to time have long or short positions in, act as principal in, and buy or sell the securities referred to in this report. Our research is disseminated primarily electronically, and in some instances in printed form. Research is simultaneously available to all clients. This research report is for our clients only. No part of this material may be copied or duplicated in any form by any means or redistributed without the prior written consent of William Blair & Company, L.L.C.

This is not in any sense an offer or solicitation for the purchase or sale of a security or financial instrument. The factual statements herein have been taken from sources we believe to be reliable, but such statements are made without any representation as to accuracy or completeness or otherwise, except with respect to any disclosures relative to William Blair or its research analysts. Opinions expressed are our own unless otherwise stated and are subject to change without notice. Prices shown are approximate. This report or any portion hereof may not be copied, reprinted, sold, or redistributed or disclosed by the recipient to any third party, by content scraping or extraction, automated processing, or any other form or means, without the prior written consent of William Blair. Any unauthorized use is prohibited.

If the recipient received this research report pursuant to terms of service for, or a contract with William Blair for, the provision of research services for a separate fee, and in connection with the delivery of such research services we may be deemed to be acting as an investment adviser, then such investment adviser status relates, if at all, only to the recipient with whom we have contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing). If such recipient uses these research services in connection with the sale or purchase of a security referred to herein, William Blair may act as principal for our own account or as riskless principal or agent for another party. William Blair is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities referred to herein.

For important disclosures, please visit our website at williamblair.com.

This material is distributed in the United Kingdom and the European Economic Area (EEA) by William Blair International, Ltd., authorised and regulated by the Financial Conduct Authority (FCA). William Blair International, Limited is a limited liability company registered in England and Wales with company number 03619027. This material is only directed and issued to persons regarded as Professional investors or equivalent in their home jurisdiction, or persons falling within articles 19 (5), 38, 47, and 49 of the Financial Services and Markets Act of 2000 (Financial Promotion) Order 2005 (all such persons being referred to as "relevant persons"). This document must not be acted on or relied on by persons who are not "relevant persons."

Any statements in this report that are attributable to IDC Research, Inc. (“IDC”) represent William Blair’s interpretation of data, research opinion or viewpoints published as part of a syndicated subscription service by IDC and have not been reviewed by IDC. IDC’s research is current as of the date IDC published it, not the date that William Blair’s reports are published. Further, IDC’s research contains IDC’s opinion, not representations of fact, and are subject to change without notice.

William Blair & Company, L.L.C. licenses and applies the SASB Materiality Map® and SICSTM in our work.