The Past, Present and Future of the Modern Data Stack

AlbionVC

From the outside it may seem that European VCs find a new sector to fall head over heels for every few years - in 2016 we had the direct-to-consumer brands, 2021 was all about crypto and web3, 2023 sees AI and climate tech as being in vogue.

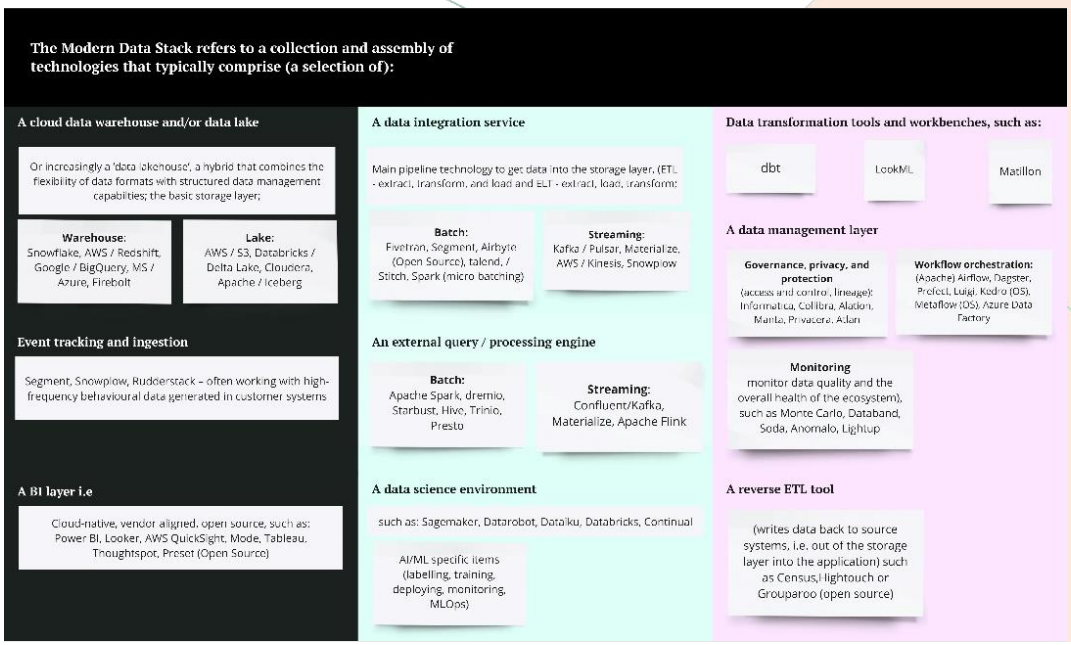
Today the modern data stack is the fundamental layer of startups and corporates, with many leading companies having been built and sold in the space. Circa 50% of all private AI and DataTech companies operating today were founded since 2010 with over $100bn invested into them by VCs. Just this September Databricks announced a $43bn Series I, prompting many to compare it to the public markets darling Snowflake. While it may lead some to conclude that the sector is set to be dominated by existing later stage companies, at AlbionVC we are highly optimistic about the potential and growth of the early-stage European data stack ecosystem and anticipate rapid acceleration in certain segments of the market over the next 2-3 years.

With the recent explosion in AI, data infrastructure has never been more important and business critical. As Frank Slootman, CEO of Snowflake, said, "*Enterprises cannot have an AI strategy without a data strategy.*" In that context we thought it would be helpful to provide an overview of where the enterprise data market is today, and the trends we see coming over the next few years.

1. Defining the modern data stack

Over the past decade, the modern data stack has emerged with three main aims: (1) to democratise data by making it broadly accessible within an organisation, (2) to normalise, organise and aggregate multiple sources and (3) in so doing to utilise it for faster, more elastic, and automated customer and organisational experiences. Usually, the modern data stack's infrastructure is built upon a data warehouse that takes input from different data sources, transforms it within the warehouse, and provides clean and often-reformatted data to support further processes.

The modern data stack refers to a collection and assembly of technologies that typically comprise (a selection of):



- A **cloud data warehouse** and/or **data lake**; or increasingly a 'data lakehouse', a hybrid that combines the flexibility of data formats with structured data management capabilities. Increasingly these hybrid implementations are developed with a focus on being schema-agnostic which reduces the need for long-running data integration pipelines before value can be realised. e.g.
    - Warehouse: Snowflake, AWS / Redshift, Google / BigQuery, MS / Azure, Firebolt
    - Lake: AWS / S3, Databricks / Delta Lake, Cloudera, Apache / Iceberg
- A **data integration** service: main pipeline technology to get data into the storage layer. (ETL - extract, transform, and load and ELT - extract, load, transform; e.g:
    - Batch: Fivetran, Segment, Airbyte (Open Source), talend / Stitch, Spark (micro batching)
    - Streaming: Kafka / Pulsar, Materialize, AWS / Kinesis, Snowplow
- **Data transformation** tools and workbenches, such as dbt, LookML, Matillion
- **Event tracking and ingestion**; e.g. Segment, Snowplow, Rudderstack – often working with high-frequency behavioural data generated in customer systems
- An external **query / processing engine**:
    - Batch: Apache Spark, dremio, Starbust, Hive, Trinio, Presto
    - Streaming: Confluent/Kafka, Materialize, Apache Flink
    - Multi-language engines: Apache Spark, Databricks, Apache Beam
- A **data science environment**, such as: Sagemaker, Datarobot, Dataiku, Databricks, Continual
    - + AI/ML specific items (labelling, training, deploying, monitoring, MLOps)
- A **BI layer i.e.** cloud-native, vendor aligned, open source, such as: Power BI, Looker, AWS QuickSight, Mode, Tableau, Thoughtspot, Preset (Open Source)
- A **data management** layer:
    - Workflow orchestration: (Apache) Airflow, Dagster, Prefect, Luigi, Kedro (OS), Metaflow (OS), Azure Data Factory
    - Governance, privacy, and protection (access and control, lineage): Informatica, Collibra, Alation, Manta, Privacera, Atlan
    - Monitoring (monitor data quality and the overall health of the ecosystem), such as Monte Carlo, Databand, Soda, Anomalo, Lightup
- A **reverse ETL tool** (writes data back to source systems, i.e. out of the storage layer into the application) such as Census, Hightouch or Grouparoo (open source)[1]

The main drivers of the fast adoption of the modern data stack are related to the unique value of this technology:

- It requires **minimal infrastructure setup and software configuration,** reducing the need for specialist in-house engineering resource; set up and **time to value is relatively short;** costs are shifted towards operational expenditure and away from heavy capital investment.
- Working out the box with other elements of the data stack, often automating connections with **plug and play** integrations and limited vendor lock-in.
- It is based on **common languages and standards** (SQL, Python, Parquet), and increasingly low/no-code implementation approaches, democratising access and thus extending user base.
- It is **hyperscalable**: use and pay for what you need, when you need it, separating cost, and compute (at vastly different price points).

[1] ICON Corporate Finance, AI & DataTech Atlas 2023

2. Brief history of the modern data stack

The Modern Data Stack is the state-of-the-art ecosystem of technologies for all data needs, from data-driven business intelligence to the deployment of AI-powered products. In the mid to late 2000s, the Big Data era, businesses started to analyse larger, more unstructured data sets using expensive and complex tools such as Hadoop and MapReduce. Then in 2012 the introduction of Amazon Redshift revolutionised analytical query workloads and response times for vast datasets by offering these services at around a 1/10th of the price. The subsequent releases of Google BigQuery, Apache Spark and Snowflake accelerated this transition.

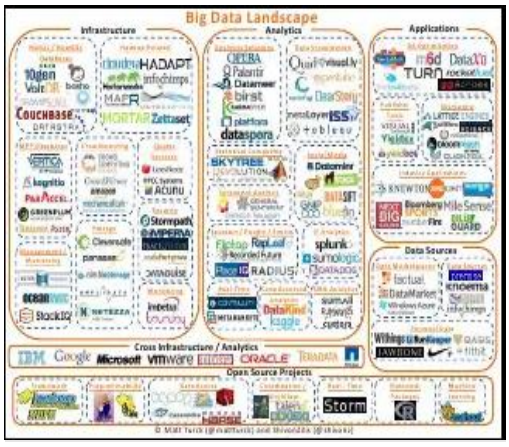
This significant reduction in the price point paved the way for a new era of technological innovation with breakthrough levels of performance, efficiency, and novel behaviours within the data stack. Tristian Handy, in his article, "The Modern Data Stack: Past, Present and, and Future" refers to this as the First Cambrian Explosion. Today, these products have reached a state of maturity in data platforms, which serve as a trusted foundation upon which a new wave of successful innovations is being built. The stage is set for the Second Cambrian Explosion. What types of innovation will it bring?

3. Dynamics of the Data'Tech flywheel

Over the last 10 years, the DataTech landscape has experienced dramatic change driven by what we call the dynamics of the DataTech Flywheel.

Matt Tuck, Partnerat Firstmark, has been publishing a market map of the global DataTech landscape since 2012. The pictures paint a thousand words and are testament to the rapid rate of change and technological innovation in that space.

2012



2022

We believe that the main growth-driving mechanisms inside today's DataTech flywheel are:

Foundational technology shifts, especially the transition to public cloud architecture, which created dramatic new opportunities in cost reduction and operational convenience. The rise of the Modern Data Stack can be seen as the resurgence of the data warehouse as a primary data store for data workloads. in the early 2000s the data warehouse started losing its dominance to data lakes due to a lack of horizontal scalabilty, inflexible heavyweight vendors and expensive hardware. Ultimately, due to a fast adoption of cloud infrastructure in the 2010s data warehouses made a comeback, this time built for the cloud and incorporating data lake approaches that enable the greater flexibility, experimentation, and auditability that modern advanced data science (including AI/ML) approaches require. The modern approach is intrinsically less centralised and less reliant on a single vendor or walled-garden approach.

According to the International Data Corporation, the premier global market intelligence firm, the global public cloud services market is expected to increase from $292 billion to $628 billion, from 2020 to 2024, respectively. Further, approximately only 4% of global data was stored in public cloud environments in 2010 compared to a projected 53% in 2024.

Better access to data. Customer expectations and faster product and market cycles are increasing demand for low-latency data insights, proliferating through the data value chain. These shifts are creating massive growth in new data generation. We observe data in greater volume, with more complexity (unstructured and semi-structured data), from more diverse sources, delivered at much higher speed (real-time and streaming data). In addition, with the central focus on data quality there has been increased appetite for and development of 'self-healing data' which enables the automation of the detection, correction, validation and learning about errors in the data to ensure the highest quality possible in an ongoing manner.

According to Cisco, global annual internet traffic surpassed one zettabyte (1021 bytes)—the equivalent, by one calculation, of 150 million years of high-definition video. It took 40 years to get to this point, but in the next four, data traffic will double. Further, by 2024, 25% of the data generated will be real-time.

Organisational change is happening in response to these technological shifts. Data Tech is moving out of IT into dedicated data departments and other functional business units (such as marketing and HR), leaving some gaps in cross-functional tooling and skills. This creates an ideal environment for the adoption of low/no-code tooling for business users to self-serve and accelerates data democratisation. For both startups and established data product vendors, this has led to major shifts in customer adoption and buying patterns and personas, moving away from the traditional CIO and CTO.

In addition to the increasing volume and complexity of data, the increased use of data across enterprises and regions/divisions with globalisation have led to regulatory complexity, with concerns like GDPR/CCPA and data residency for compliance reasons, not technical ones. This need for additional governance also drives a need for specialist skills and an organisational structure that can provide the needed oversight.

Currently, 60% of Fortune 1000 companies employ a Chief Data Officer (CDO) up from 12% in 2012 (NewVantge) and DataTech talent in general is in huge demand. Linkedin's 3 of top 10 roles in 2020 were data related. DataOps is one of the fastest growing disciplines; 73% of data professionals said they need to hire more in this domain (Nexla).

More data leads to better infrastructure, more roles and adoption, which drives data product innovation and in turn leads to the creation of even more data. The DataTech flywheel propelled by these forces is just beginning to turn. Despite the fast growth in the global cloud computing market, over the past five years, only 17% of the system infrastructure software spend has been made up of spend from public

cloud services (IDC). Only [2]13k companies use Redshift and [3]8.5k companies used Snowflake as of 2023. So, there is still a massive room for growth and further technological innovation. Which trends are going to drive it?

Academic Focus and Talent Development is also undergoing significant changes as universities increasingly offer more practically focused data science degrees and conversion programs for students with non-STEM backgrounds. In addition, there has been a proliferation of alternative resources for learning such as online courses and bootcamps. This has meant that there has been a dramatic increase in data talent which in turn creates fertile ground new ideas and growth in the data tech sector.

4. Trends we see in the space

At AlbionVC, with support from Seedcloud, we have developed a market map of the leading European DataTech startups. By the virtue of performing this exercise – and blending in what Seedcloud are seeing in active use cases across hundreds of engagements in the industry – we were able to reflect on the direction of trends that are currently shaping the Modern Data Stack. We believe that the likely defining features of the current wave of technological innovation in the space, that is the Second Cambrian Explosion, are:

Metadata

Metadata augmented with business context enables self-serve data products and data governance to be owned by various stakeholders but within a central framework, critical to breaking down data silos while maintaining holistic visibility of an entire organisation's data estate. This is fundamental today in pursuing operational efficiency and automation. Furthermore, while implications and applications of enterprise generative AI are still nascent, a full mapping of an organisation's data estate as enabled by advances in metadata management and data lineage become a precursor to achieving business value from LLMs - what datasets are available, to what quality, and how are they interrelated up and down stream become fundamental questions. As active metadata is more widely adopted, breaking down data silos, and connecting to semantic models, data discovery and data analysis cross-enterprise becomes possible while maintaining the integrity of governance and rights management. Solidatus, an award-winning data lineage solution, for example, is providing the next wave of capabilities, enabling enterprises to digitally map policies onto the data estate to enable organisations to understand departures and take remedial action. Virtually connected governance if finally possible, connecting data governance regimes across business functions, legal entities, and geographies without the operational complexity of implementation.

Expanding on this, there's untapped potential for businesses to gain significant operational advantages by integrating functional value chain metadata with industry-specific data ontologies. This aligns with the notion that enterprises can extract considerable value from combining internal data sources, like transaction records, with external data, such as news about their customers and suppliers, to generate deeper insights.

Overall, it is clear that metadata is becoming the magic dust that balances access, integrity, and governance across organisations' increasingly complex data estate.

[2] 6sense.com
[3] Statista.com

Data Observability

A combination of supply side factors (data and infrastructure complexity and diversity) and demand side factors (data quality and limited/no “data downtime”) has made Data Observability (DO) a hot topic. DO tooling is doing for organisation’s data leaders what employee engagement tooling did for organisation’s people leaders. The analogy is apt given it is often argued that data is fast becoming an organisation’s most valuable asset. Key to progressing the DO category is the further development of holistic tooling, providing a single pane of glass over each of the major areas of observability (quality, pipelines, infrastructure, usage, and costs). Siloed tools of course exist today but the real value comes in breaking down these silos and providing a 360-degree view of the data domain and its performance. Acceldata, Monte Carlo, and IBM are across a number of observability areas but still not completely end-to-end. The DO journey is also progressing from a pure detection and monitoring exercise to recommendation of fixes, and into pre-emptive measures. This is still the nirvana and a number of limitations remain to be alleviated (data in motion, infrastructure flexibility/on-prem, and non technical user engagement). Undoubtedly the prominence of DO will increase but for all the investment and insurgent companies will it remain a standalone capability?

Emergence of Data platforms

The Modern Data Stack has a stable core and a vibrant shell. Much of the development we see is in the ecosystem of firms who create applications on top of stable core capabilities, which are increasingly provided by few established players, such as Snowflake and Databricks, who have become an essential part of any data pipeline. The more the data stack resembles a platform, like Windows and Mac OS, the more those players will have a moat at the core of any future developments. In a recent blog post a16z suggests that high valuations for reverse ETL and metrics layer companies can only be justified on the grounds of them having the potential to become a core part of the platform. The rise of data platforms created the conditions for the “Cambrian explosion” of new data functionalities. As long as the platforms remain open, they will enable new firms to innovate in the application layer and find large cohorts of users as first-class part of the stack. The coming year will see further consolidation in the platform layer and creative innovation in the application layer.

An additional, emerging characteristic of leading players operating across multiple domains within the stack is interoperability where companies architect their platforms to allow for decomposition such that components can be combined with ‘replacement parts’ from other companies in the ecosystem.

Real-time data

Increasing the frequency of data updates is the next frontier in the evolution of the core capabilities of the Modern Data Stack. As it becomes increasingly affordable, real-time data enables a host of new use cases as well as improving existing processes. 70% of real-time companies found in our market map were founded in the last 5 years, which indicates this is currently an area of rapid innovation.

Though real-time data is not the main use case of the Modern Data Stack today, a drastic reduction in the pipeline latency can unlock plenty of new use cases for this technology. The analysis of real-time companies featured in our market map can be a good indication of the emerging use cases within the real-time start-ups space.

With TinyBird, Ably and Scramjet, developers can create simple APIs in seconds that allow fast and powerful queries onto real-time data. Quix enables real-time calibration and deployment of ML models for data scientists. Conduktor Platform is an interface and management tool for streaming applications

that sits on top of Kafka. DoubleCloud and Altinity build the real-time analytics applications layer on top of ClickHouse. Elastic detects instances of fraud and security anomalies when it really matters, right as they occur. Snowplow has recently fine-tuned its primary proposition into a real-time “data creation” platform, specialised for customer data platform use cases.

Best of breed startups compete against limited capabilities from existing players. Snowflake, BigQuery, Redshift, and Databricks already offer some features for streaming data. They may not be as powerful or flexible as the specialized tools but have the advantage of integrating streaming and traditional use cases.

Vector databases

Generative AI is accelerating the development and adoption of vector databases. These can store and query large amounts of vector embeddings that have been derived from unstructured data (such as text, images, audio, and video) which can be leveraged by advanced machine learning models. Vector embeddings are low-dimensional but semantically rich representations of data that capture their meaning and context. Vector databases enable fast and accurate similarity search and retrieval of data based on their content and style. They are ideal for generative AI applications that need to access and manipulate large amounts of data in real time, such as chatbots, text generation, and image generation. Vector-databases are already on the rise and will continue to grow as firms turn to AI to make their unstructured data useful.

As our European DataTech market map indicates, we can see important players in that space emerging in Europe. The two most prominent European vector databases providers are Berlin-based Qdrant (founded in 2021, valued at $33-50 ml) and Dutch Weaviate (founded in 2019, valued at $200m). Qdrant is powering the next generation of AI applications with vector similarity search technologies. Weaviate is an open-source vector database for storing data objects and vector embeddings for ML-models.

Data governance

As more data becomes available, data governance products are becoming increasingly indispensable. With more data available, ensuring its quality becomes a necessity to avoid chaos. Data governance consists of many product categories. Our market map distinguishes the following: synthetic data, data catalogue, data quality, data observability, data privacy and governance and security.

The most striking fact about the data governance category in our market map is its size – it is by far the biggest category listed. We were able to identify 76 European companies focused on data governance and we believe this indicates that prime time for this product category is coming. Until now, much of the progress in data governance has been led by the Big Tech and their internal governance tools such as: DataHub (Linkedin), Dataportal (Airbnb,) Metacat (Netflix), Databook (Uber) and others. Many smaller companies were not integrating any governance tools into their data stacks. However, now we see a strong push for innovation in this space and believe that, with time, the technology is going to be more widely adopted by smaller players as well.

AI influence

We expect the latest wave of AI progression to have the most significant influence on the Modern Data Stack over the next 2-3 years. The sudden acceleration of generative AI, as well as the continuing advancement and proliferation of predictive AI approaches, will demand data platforms that can integrate and support these capabilities as first-class, native components.

AI is already influencing data architecture and tooling right across the stack. This includes data pipelines that can handle a previously unseen diversity of data formats and step-changes in data volumes. In fact, much of the value from the proliferation of generative AI and LLMs is related to the transformative ease by which users can now build ELT/data processing pipelines for unstructured data. With 80% of all enterprise data being unstructured this dramatic decrease in complexity is a major driver of the shift towards schema-agnostic data storage approaches and the ability to leverage unstructured data for real-time applications.

Data needs to be stored in both raw and processed formats to enable the constant experimentation, training, adversarial testing, measurement and refinement of models and neural networks. As a result, the data lake pattern will remain embedded in the stack, alongside more traditional data stores and warehouses. Metadata to describe and track complex data assets is increasingly a requirement.

A clear audit trail is increasingly a vital requirement as well, being able to track and audit the lineage of data used in machine learning – both for practical and compliance reasons. This leads to changes in storage (e.g. data lakes to keep source data), modelling (event source models, to track all changes and support point-in-time queries), and supporting products (e.g. Solidatus, Collibra to help track lineage), often for regulatory compliance. It is always important to know why a loan gets denied, insurance policy declined, job hire rejected, or price adjusted – and even more so when these have legal ramifications. **Adding AI over the top where decision is made without human justification means that there is an even stronger need for 'explainability'.**
Specialist tools are being adopted to help handle these responsibilities, due to businesses needing to meet existing GDPR requirements and the inevitability that the EU will drive new and stringent AI-specific regulations across the bloc.

Multiple data stores, from document to relational to vector to multi-model, are needed to handle different workloads most efficiently at different stages in the model lifecycle. Some of these will be provided by specialists (such as vector), adding further complexity and decreasing the reliance on single vendors – at least in the short term, while the major players catch up. All of this places greater demand on storage, processing and monitoring – and the costs of operating the modern data stack.

The adoption of structured MLOps and DataOps approaches and the deeper embedding of data science roles into cross-functional engineering teams will boost the use of tooling that allows AI/ML teams to mirror and join with modern automated DevOps approaches around testing, deployment and operational management.

The proliferation of AI-driven insights and decisions is driving increasingly regulatory and compliance requirements that create more complexity for organisations, as well as the everyday management of data. The need to enforce and audit robust frameworks and governance patterns is leading also to new tools entering the modern stack. For example, specialist components that ensure the reliable anonymisation of personally identifiable information in unstructured data, or attempt to provide greater explainability of black-box models for regulators, by attempting to extract the features that led to outcomes (such as refusing a loan application).

Europe is well positioned to capitalise on these opportunities because of the availability of high-quality technical talent at a relatively affordable price. This allows European AI companies to differentiate their approach from many application focused AI companies which are built as thin wrappers around foundational models and API exposed data retrieval mechanisms. Really understanding the strengths and limitations provided by generative AI from a technical and commercial perspective enables companies to better leverage the existing data tech stack. This is compounded by an increasing outlier focus in European VC. This has been validated by the wave of relatively recent unicorn companies in

Europe and in turn allows European companies the chance to genuinely compete on a more even playing field with US competitors.

Next-generation cloud cost optimisation

Many businesses have taken advantage of the inherent horizontal scalability and operational convenience of public cloud infrastructure when assembling their data platforms. This has naturally enabled more data to be collected, processed and generated than ever before. In many cases, however, cloud costs have scaled linearly with the data volumes, leading to budget pressure and ceilings to be imposed on data storage capacities.

This, as well as demands to get data processed more quickly and into the hands of end-consumers (people and other systems) in near real-time, has led a wave of startup activity in recent years targeting improvements in compute and storage optimisation – much of it around essential monitoring and backup management.

This is a well-known problem, but not yet a fully solved one. Many of the cost-inflating factors in modern data and AI platforms are deeply embedded in the data engines, query processing plans and CPU architectures in use – which simple cost and utilisation monitoring cannot reach. The explosion of AI has also led to significant GPU shortages, which adds further constraints. As a result, some emerging startups are looking at ways to take optimisation approaches to a new level of sophistication through intricate bottom-up technology approaches.

XONAI, for example, provides a drop-in optimisation product that accelerates processing of analytical data in Apache Spark-based platforms such as Databricks and AWS EMR – resulting in significant, directly measurable cost savings.

CAST.AI has built a specialist ML-driven cost optimisation platform for Kubernetes clusters, that learns the workload demands of the applications being run on a particular cluster, then automatically adjusts the resource allocation to find the optimal balance of performance and cost.

As optimisation specialists, such companies believe they can provide a defensible enhancement layer on top of the core features of mainstream platforms. This is because those platforms' business models have been traditionally aligned with driving increased consumption and spend, rather than reducing it. That said, both Databricks and Snowflake have recently released new features that are directly targeted at reducing costs, through targeted performance enhancements.

By being independent, startups in this space also have the ability to enable optimal resource consumption across multiple cloud providers and abstract some of the complexity of multi-cloud architectures.

Final thoughts

Over the past ten years the Modern Data Stack has been transformed from an adjacent part of the IT set up to an integral part of the modern organisation driving decision making, customer experience and ultimately revenue. Having explored the reasons for such drastic shift, we remain bullish about the prospects of the European data ecosystem. While between 2012 and 2022 many categories leading global data companies have been built, we believe that we are posted for further acceleration. The need for metadata, data observability, real-time data, vector databases, data governance and next-generation cloud cost optimisation are all driving the next wave of technological innovation creating a generation of European companies with strong business models, healthy growth margins and unit

economics, alongside a sticky and loyal customer base. Yet it the latest wave of AI progression, in a now maturing European tech ecosystem, that we expect to have the largest influence on the modern data stack over the next 2-3 years.

At AlbionVC we have been partnering with data and AI companies for well over a decade - our current portfolio includes 20 early and growth stage European companies across a range of data stack categories. We've had 9 exits in the sector including to Apple, Meta and Microsoft. If you are building or considering starting a company in the space, we'd love to hear from you.

This project has been developed in partnership with Seedcloud, a leading advisory firm working on specialist technical and product strategy engagements for technology-led companies and their investors.

Appendix

Methodology of Modern Data Stack market map

In H1 2023, we scraped data from Dealroom, Crunchbase, and Beauhurst to generate a long list of almost 6,000 data-related companies based on their description or tagline. We only considered firms founded after 2012 which had HQ in Europe or Israel to reflect Albion's investment criteria. We examined the website of each company and selected only those that were directly relevant to the Modern Data Stack infrastructure as defined in section 1 of the report. Most of the companies we screened out were application layer data tools. This process resulted in selection of 265 Modern Data Stack companies, which we further categorised drawing on experience within the team at AlbionVC. We checked for completeness of our results against an overlapping dataset of c. 4,000 companies from Specter.

Since we compiled our list, the generative AI category has exploded, and as such, our map and analysis likely underrepresent this segment.

See the map on our website HERE.