fsinsight powered by fundstrat

Market Data Insight for Actionable Strategy

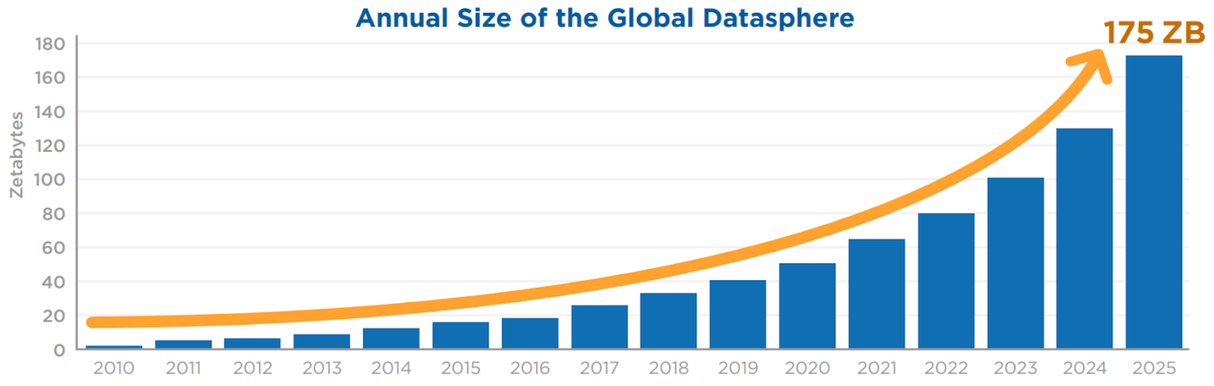
Signal From Noise

October 12, 2023

Big Data, Big World

“There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days.” ~ Eric Schmidt, former CEO of Google

The extent to which data drives our world is increasing. In 2018, IDC estimated the annual size of the global datasphere – the amount of new data created, captured, and copied – at around 33 zettabytes (ZB). At the time, International Data Corporation (IDC) predicted that the number would grow to 175 ZB by 2025. (For reference, a zettabyte is 2^{70} bytes, or roughly one trillion gigabytes.) The IDC’s methodology incorporated estimates for data related to non-entertainment imaging (medical imaging, security footage, etc.) as well as entertainment, productivity, and voice data.



Source: Seagate Technology, “The Digitization of the World from Edge to Core”

The expanding datasphere is the result of an acceleration in disparate trends – from remote work to adoption of IoT devices to machine learning to AI research and implementation. It is also rooted in the growing penetration of fast Internet access around the world. Consequently, the global Big Data market – companies that help enterprises manage and use all that data – is expected to experience annualized growth rates of 11%-13.5% from now until 2030.

All of this constantly generated data has immediate, obvious usefulness, whether it's a selfie shared on social media, a transaction record at a convenience store, or a research note. Yet when aggregated into large sets, data can become a profoundly powerful and versatile tool. Among other things, large datasets are critical to the development, training, and testing of cutting-edge research into generative artificial intelligence.

Perhaps less glamorously, although certainly no less importantly, large datasets are used by companies in a broad range of sectors, including finance, agriculture, insurance, retail, transportation, and health care, as well as in public sector fields such as education and law enforcement, to improve:

- **Risk management.** Large datasets allow businesses, companies, and governments to take a larger view of potential risks, identifying otherwise hidden patterns and trends that can foreshadow risk events. These can include operational losses such as cyberattacks and fraud, leading to operational losses, as well as losses associated with various types of financial risk, and losses related to natural disasters.
- **Operational efficiency.** Businesses can use large data sets to optimize their supply chains, improve or streamline production, allocate resources, and facilitate transportation and delivery. Large datasets help hospitals and other healthcare providers anticipate public health needs and provide better patient care.
- **Research and product development.** Big data improves the quality of machine-learning models, thus facilitating the discovery of new medical treatments, cutting-edge materials, and granular insights into customer needs and preferences. Large datasets can help in the development of more accurate financial models, optimize crop management for farmers, and aid retailers in making better product recommendations.

Data Storage

Of course, before anyone can do anything with all that data, it needs to be stored somewhere. This is typically done in data lakes and data warehouses. Data lakes store massive amounts of raw data that has yet to be organized, tagged, cleaned, validated, and standardized. Often, the data is from different sources – combining, say, video footage of customers talking about a company's product with comments made on social media and via e-mail.

In contrast, data warehousing involves the storage of data that has been cleaned and validated, with its formatting standardized. That is, data warehouses store data that is ready for use and will likely be accessed more frequently.

Companies that help enterprises to store and manage large quantities of data include:

Snowflake ($SNOW)

Snowflake is one of the largest providers of data lake and data warehousing services in the world, with 30 locations in the Americas, Europe and the Middle East, and throughout the Asia-Pacific region.

MongoDB ($MDB)

MongoDB's business is based on free and paid versions of an open-source, non-SQL database platform. The company primarily derives revenues from a variety of subscription-based, software-as-a-service offerings that include stream processing, continuous data validation, cloud-based analytics-optimized data lake solutions, as well as data visualization tools.

In many ways, Snowflake and MongoDB compete with Big Tech companies that also offer cloud-based data storage – think Amazon ($MDB), Alphabet ($GOOGL), and Microsoft ($MSFT). However, unlike their Big Tech competitors, Snowflake and MongoDB are cloud-agnostic.

Data cleaning and validation

Making the most of the advantages provided by large datasets requires the data to be cleaned. This generally includes most or all of the following:

- **Duplicates** must be removed.
- Formats need to be **standardized**
- **Syntax** errors are corrected
- **Incomplete or inaccurate records** need to be rectified or removed
- **Outliers** should be identified and, if deemed appropriate, removed
- **Obsolete** or untimely data needs to be identified and removed

Finally, at the end of this process, the dataset needs to be validated to make sure data cleaning is complete. Some of the leaders in this field include:

Informatica ($INFA)

Informatica helps enterprises integrate data from multiple disparate sources, with an array of tools to clean, categorize, and improve data quality, in part with AI-powered automation. The company's offerings also include data warehousing and real-time data-quality monitoring. Informatica's revenues from a broad range of industry sectors, including Financial Services, Energy, Health Care, Manufacturing, Retail, Telecommunications, and Education.

Workiva ($WK)

Workiva specializes in helping businesses that need to work with large financial datasets. It can help them collect data from disparate sources, clean and standardize that data, and maintain a transparent audit trail for the entire process

Data Mining (process mining software market)

With the right dataset cleaned and validated, data mining can begin. This is the process of identifying otherwise hidden patterns and trends in order to glean insights and spot trends. Data-mining techniques include:

- **Cluster analysis** – searching a dataset for records with similar characteristics, assessing clusters for size and data distribution
- **Regression analysis** – the use of statistical techniques to identify relationships between variables with a dataset's records
- **Association rule-learning** – looking for variables with strong correlations
- **Anomaly detection** – learning to identify records that are somehow unusual, hoping to better understand when and why they occur
- **Sequential pattern mining** – seeking to identify patterns in events over time

The results of this process are assessed and used to build models and visualizations that can help enterprises better manage their ongoing operations and to anticipate, predict, and prepare for what might happen in the future.

Alteryx ($AYX)

Alteryx is a leading provider of data analytics products and solutions. Its Big Data offerings include an integrated product that cleans and validates data, then uses machine learning to mine that data to produce models for enterprise use. Its clients include companies in retail, food services, consumer goods, entertainment, financial services, health care, and financial services.

Teradata ($TDC)

This San Diego-based company provides one of the industry's leading data platforms, which includes multi-cloud compatible data lakes, data warehousing, and cloud-based analytics. Its solutions provide clients with automated tools to take in, store, analyze, and retrieve data of any type or structure quickly and easily.

Elastic ($ESTC)

Elastic's primary product is Elastic Stack, a platform of software products that allows for the virtually real-time ingesting, formatting, analysis, and visualization of data from any source and in any format. Elastic Stack can be implemented on premises and a cloud-based basis; it is cloud-agnostic.

Splunk ($SPLK)/Cisco Systems ($CSCO)

Splunk went public in 2012 and on September 21, 203, agreed to be acquired by Cisco Systems (at $157 per share all-cash, a 31% premium over its September 20 closing price.) Splunk provides products and solutions that help companies store, manage, and analyze data in real-time, particularly machine-generated data such as computer servers, mechanical devices, and industrial control systems. This has a variety of use cases, most prominently in network cybersecurity. Unsurprisingly, therefore, the company also offers a suite of cybersecurity and related analytics solutions. It is estimated that 90% of Fortune 100 companies use Splunk offerings to some extent.

Before a data model can be used, it needs to be tested. Sometimes, this involves the use of synthetic data – data is generated by applying algorithms to a small set of actual data points from the real world in order to create numerous artificial data points. Large sets of synthetic data can thus be used for more effective testing of models, particularly when real data is unavailable due to logistical limitations or because of privacy concerns surrounding the use of personal data.

This is a relatively new field. While there are (as of this writing) no publicly traded companies that specialize in the generation or creation of synthetic data, many of the Big Tech companies, like Amazon ($AMZN), Microsoft, ($MSFT) and Google ($GOOGL) have a presence in this space.

Data infrastructure

Even electronic data requires a physical xx, and companies that work with large datasets require specialized facilities to house their data operations. Many often choose to rely on outsourced solutions to fulfill those needs. Among the companies

Equinix ($EQIX)

Equinix is arguably the largest dedicated provider of data center solutions, enabling companies to easily setup colocation data centers in any of roughly 240 facilities in 71 major metropolitan areas on five continents. In essence, Equinix provides clients with power, floor space, security, and cooling in locations with access to multiple major telecommunications-services providers. It also offers a limited range of data storage solutions. Equinix's primary selling point is in the robust interconnectedness of its facilities.

Digital Realty Trust ($DLR)

DLR is Equinix's biggest competitor, Like its rival, it provides enterprises with colocation data-center solutions. Its geographic reach is not quite as expansive as Equinix's, but Digital Realty is generally regarded as having a stronger focus on sustainability and lower price point.

Iron Mountain ($IRM)

Iron Mountain is one of the largest information storage and management companies, and it claims to have 95% of the world's Fortune 1000 companies as clients. While it has primarily been known for its physical records storage business, it is expanding into data, with solutions that include colocation and wholesale data center space.

Working with large datasets requires deep technical knowledge in a variety of fields, and this piece is not intended to be a guide to the intricacies involved. Rather, it was written to serve as an introduction into some of the component processes involved in the growth of the global datasphere, and some of the companies that might be poised to benefit.

As always, *Signal From Noise* should be seen as a starting point for further investigation. We encourage you to explore our full Signal From Noise library, which includes deep dives on the path to automation and opportunities arising from the ever-increasing global water crisis and related to the impending allergy season. You'll also find a recent discussion on the Magnificent Seven and the coming EV revolution.

Your feedback is welcome and appreciated. What do you want to see more of in this column? Let us know. We read everything our members send and make every effort to write back. Thank you.

Disclosures

This research is for the clients of FS Insight only. FSI Subscription entitles the subscriber to 1 user, research cannot be shared or redistributed. For additional information, please contact your sales representative or FS Insight at fsinsight.com.

Conflicts of Interest

This research contains the views, opinions and recommendations of FS Insight. At the time of publication of this report, FS Insight does not know of, or have reason to know of any material conflicts of interest.

General Disclosures

FS Insight is an independent research company and is not a registered investment advisor and is not acting as a broker dealer under any federal or state securities laws.

FS Insight is a member of IRC Securities' Research Prime Services Platform. IRC Securities is a FINRA registered broker-dealer that is focused on supporting the independent research industry. Certain personnel of FS Insight (i.e. Research Analysts) are registered representatives of IRC Securities, a FINRA member firm registered as a broker-dealer with the Securities and Exchange Commission and certain state securities regulators. As registered representatives and independent contractors of IRC Securities, such personnel may receive commissions paid to or shared with IRC Securities for transactions placed by FS Insight clients directly with IRC Securities or with securities firms that may share commissions with IRC Securities in accordance with applicable SEC and FINRA requirements. IRC Securities does not distribute the research of FS Insight, which is available to select institutional clients that have engaged FS Insight.

As registered representatives of IRC Securities our analysts must follow IRC Securities' Written Supervisory Procedures. Notable compliance policies include (1) prohibition of insider trading or the facilitation thereof, (2) maintaining client confidentiality, (3) archival of electronic communications, and (4) appropriate use of electronic communications, amongst other compliance related policies.

FS Insight does not have the same conflicts that traditional sell-side research organizations have because FS Insight (1) does not conduct any investment banking activities, and (2) does not manage any investment funds.

This communication is issued by FS Insight and/or affiliates of FS Insight. This is not a personal recommendation, nor an offer to buy or sell nor a solicitation to buy or sell any securities, investment products or other financial instruments or services. This material is distributed for general informational and educational purposes only and is not intended to constitute legal, tax, accounting or investment advice. The statements in this document shall not be considered as an objective or independent explanation of the matters. Please note that this document (a) has not been prepared in accordance with legal requirements designed to promote the independence of investment research, and (b) is not subject to any prohibition on dealing ahead of the dissemination or publication of investment research. Intended for recipient only and not for further distribution without the consent of FS Insight.

This research is for the clients of FS Insight only. Additional information is available upon request. Information has been obtained from sources believed to be reliable, but FS Insight does not warrant its completeness or accuracy except with respect to any disclosures relative to FS Insight and the analyst's involvement (if any) with any of the subject companies of the research. All pricing is as of the market close for the securities discussed, unless otherwise stated. Opinions and estimates constitute our judgment as of the date of this material and are subject to change without notice. Past performance is not indicative of future results. This material is not intended as an offer or solicitation for the purchase or sale of any financial instrument. The opinions and recommendations herein do not take into account individual client circumstances, risk tolerance, objectives, or needs and are not intended as recommendations of particular securities, financial instruments or strategies. The recipient of this report must make its own independent decision regarding any securities or financial instruments mentioned herein. Except in circumstances where FS Insight expressly agrees otherwise in writing, FS Insight is not acting as a municipal advisor and the opinions or views contained herein are not intended to be, and do not constitute, advice, including within the meaning of Section 15B of the Securities Exchange Act of 1934. All research reports are disseminated and available to all clients simultaneously through electronic publication to our internal client website, fsinsight.com. Not all research content is redistributed to our clients or made available to third-party aggregators or the media. Please contact your sales representative if you would like to receive any of our research publications.