

# DataOS<sup>®</sup>

Data Evolution &  
Modern Data Management

## Executive Summary

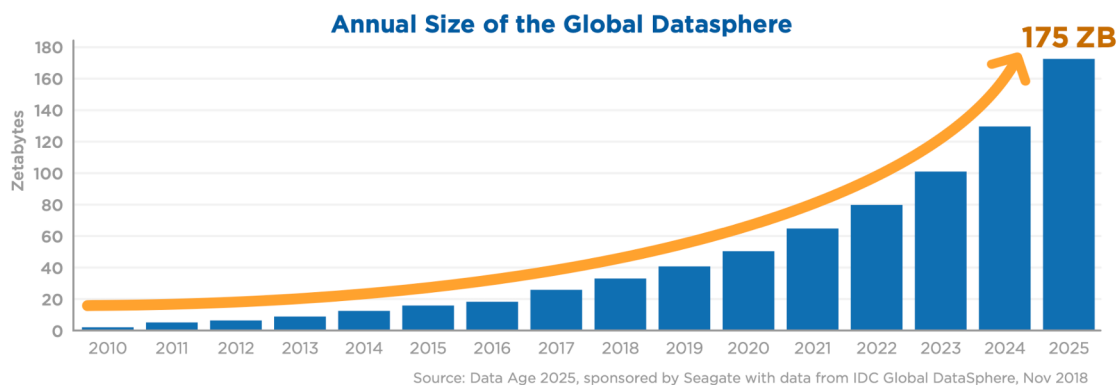
- The growth of the size of data is accelerating as businesses re-oriented to digital first engagement models
- Data silos associated with legacy applications have multiplied as a result of data management strategies to accelerate these systems to real-time
- Democratization requirements of the business has left the traditional enterprise data and analytics infrastructure breathless
- Data fabric empowers organizations and their users to leverage their data without the burden of redesigning and overspending on data management solutions

## Growth Rate of the Size of Data has Blown Up

The world is rapidly shifting to a digital first model for every organization due to world events. The long known expansion of the size of data has suddenly kicked up its pace. Organizations are competing on analytics and data more than ever and their only recourse for capitalizing on this rapidly growing data is to double down on even more advanced analytics with major investments in data and analytics infrastructure, machine learning, and tools for democratizing data and analytics access.

# Growth of Data

The use of data today is transforming the way we live, work, and play. Businesses in industries around the world are using data to transform themselves to become more agile, improve customer experience, introduce new business models, and develop new sources of competitive advantage. Much of today's economy relies on data, and this reliance is accelerating due to current world events around increasing tariffs, Covid-19 restrictions and even social unrest movements. As organizations shift to a digital first world, it is a foregone conclusion that there will be a never-ending expansion in the size of the global datasphere. Estimated to be 33 ZB in 2018, IDC forecasts the Global Datasphere to grow to 175 ZB by 2025.



## Data Silos Multiplying

Organizations continue to experience challenges with data silos created by legacy applications. Matter of fact, the number of data silos have increased as additional layers of data management systems we added on top of older legacy systems to bring their performance up to real-time for use in customer and partner facing portals and applications. As early as 2018, nearly 80% of corporate executives were already worried about the threat of being disrupted by competitors who were demonstrating an ability to compete on analytics (New Vantage Partners, Big Data Executive Survey 2018). Enterprises invested heavily in user and operational analytics adding cloud data warehouses to support users and operational data management systems to support real-time web applications.

## What does “digital-business” mean to organizations?

Enable worker productivity through tools such as mobile, data access and AI-assisted processes

52%

Ability to better manage business performance through data availability and visibility

49%

Meet customer experience expectations

46%

IDC: 2018 State of Digital Business Transformation

# Rise of the Citizen Data Scientist

Coming behind the growth in data and investments in machine learning is the creation of new job functions in the organization- namely data scientists. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from the even expanding data.

Recent surveys of job creation in the data science field reveal:

- 650% job growth since 2012 (source: LinkedIn).
- An estimated 11.5 million new jobs by 2026 (source: U.S. Bureau of Labor Statistics).

The increase in organizational uptake of data isn't only limited to data scientists. Many organizations have embarked on strategic transformations to inculcate a culture of being data driven with the creation of C -level positions for Chief Data officers and Chief Analytics Officers but also driving analytics usage down to the front line worker. The business strategy being data driven has resulted in a dramatic increase in the use of data and analytics throughout an organization as well as creation of a new trend the "citizen data scientist" first coined by Gartner. Both Forrester and Gartner point to new product requirements for self-service in both data integration tools as well as analytics with much of the self-service functionality being bolstered by Machine Learning for task automation. This trend is both business strategy and an evolution of the latest technology to become autonomous and automated serving any level of user in an organization.

- "By 2022, manual data integration tasks (including recognition of performance and optimization issues across multiple environments) will be reduced by 45% through the addition of ML and automated service-level management." ~Gartner Magic Quadrant for Data Integration Tools, August 2019
- "Self-service data integration — Finally, there is an increasing expectation of organizations to allow business users or citizen integrators to be able to integrate "their own data" through data preparation tools and techniques. The notion of data management being able to govern and control the flow in a synergistic manner through the data integration tool is a challenge that data integration tool vendors are expected to solve." ~Gartner Magic Quadrant for Data Integration Tools, August 2019
- The best data fabric solutions focus on data democratization by allowing business users to support easy discovery and navigation of data assets. In addition, vendors now offer zero-code and low-code functionality to accelerate even large and complex fabric deployments. Look for vendors that have expanded AI/ML capabilities to automate data discovery, classification, security, ingestion, transformation." ~Forrester Research, The Forrester Wave: Enterprise Data Fabric, Q22020

- “The tools available to create and execute analytics have evolved massively in recent years. Activities that used to require a lot of custom coding can now be done with just clicks, drags, and drops. Whether it be visual workflows, visualization tools, search-based analytics, or automated modeling tools, creating and executing analytics has never been easier. In today’s world, people without a deep and formal education in math or statistics can now make use of sophisticated analytics. It’s no longer as important to understand the detailed math behind a logistic regression, for example, as to understand when a logistic regression should be used (which machines can even recommend) and how to interpret its results. Citizen data scientists and business analysts can handle this and are starting to do a lot of it. The biggest challenge with this trend is to ensure that guardrails are put in place so that a citizen data scientist can increase his or her productivity while minimizing the risk that an inappropriate process is created due to their lack of deeper technical knowledge.” ~International Institute for Analytics, 2019 ANALYTICS PREDICTIONS AND PRIORITIES

## Data Fabric is the Answer

With today’s world events, many businesses are facing a potential “mass extinction” event if they are unable to rapidly reforecast business plans and reorganize supply chains all of which requires voluminous analysis of an organization’s data in a very short order of time. Business leaders are left working with an extremely complicated and layered data and analytics infrastructure built up over time to accomplish this. Legacy systems and traditional data warehouses where data assets reside are still critical; new cloud data warehouses and operational data stores hold the keys for the business model of the future. Now more than ever organizations are looking for modern data management solutions to bridge this gap. Forrester has coined the term “data fabric” for technology solutions which help to close this gap.

### **Forrester Research defines data fabric as:**

*“Orchestrating disparate data sources intelligently and securely in a self-service manner, leveraging data platforms such as data lakes, Hadoop, Spark, in-memory, and NoSQL to deliver a unified, trusted, and comprehensive view of customer and business data across the enterprise.”*

# DataOS<sup>®</sup> - The Modern Data Fabric

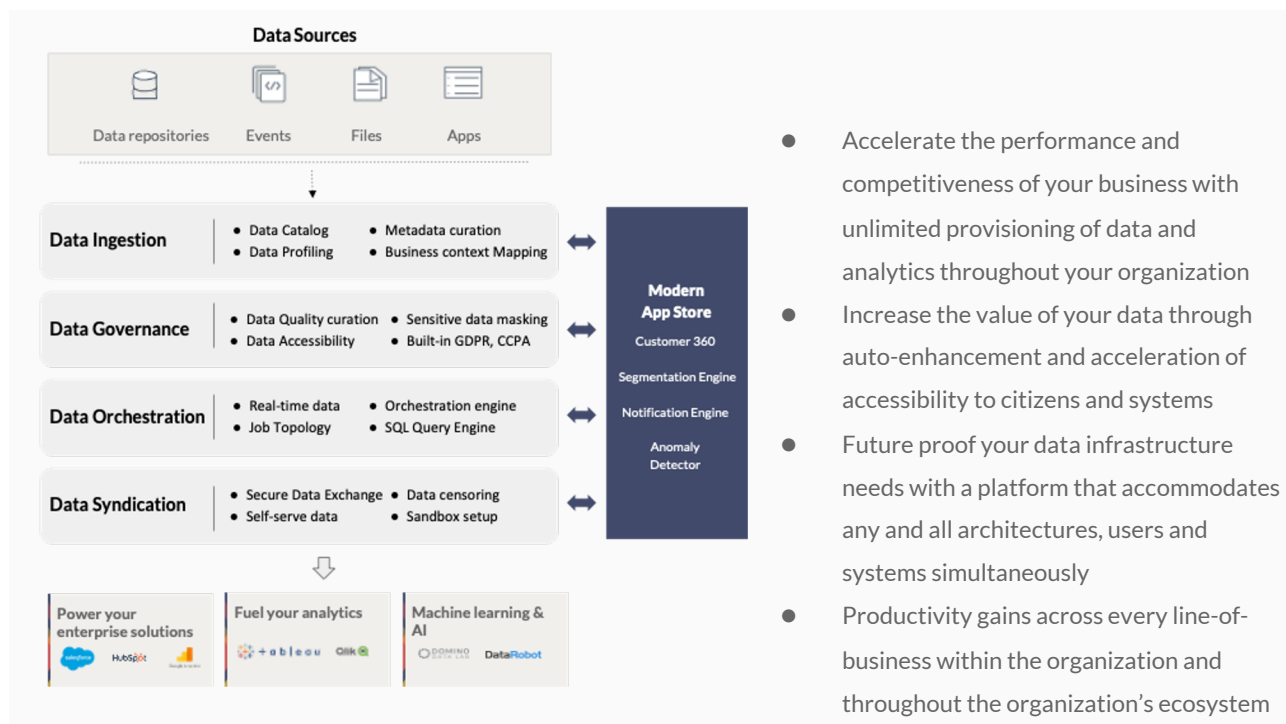
DataOS is a modern data fabric designed specifically for the age of the citizen data scientist amidst the complexity of today's enterprise data and analytics infrastructure. As the number of workers who need to use data in their everyday jobs increases, it no longer makes sense to relegate data to the purview of IT only. DataOS is an organization-wide data management practice that can help more stakeholders drive more value from data.

*By 2022, 90% of corporate strategies will explicitly mention information as a critical enterprise asset, and analytics as an essential competency. ~Gartner “- 10 Ways CDOs Can Succeed in Forging a Data-Driven Organization” May 2019*

*“Unique data can differentiate a firm as much or more than analytics today. It is very hard in today's world to stay too far ahead purely by out analyzing the same data that the competition is also analyzing.” - IA 2019 Predictions*

The Modern DataOS Data Fabric disaggregates your data from its residence in data warehouses, data marts, applications, APIs, services, users, and even external sources and automatically enhances and deploys it strategically throughout your organization. It does this by finding, integrating, cataloging it and then applying Machine Learning logic to auto-analyze and auto-enhance it --increasing its intrinsic value. Additionally, DataOS governs access to the data, and enables its sharing throughout an organization and its ecosystem. This ensures the entirety of the data is protected by enterprise-grade security and provisions the data with high-performance speed to analysts, data scientists, ML systems, BI tools and data warehouses.

## DataOS – The Modern Data Fabric



## Capabilities of the Modern Data OS<sup>®</sup> Data Fabric

The DataOS Data Fabric consists of several data management components and software, all of which must work in tandem to deliver an agile, integrated, secure, and scalable data ecosystem. Regardless of the data's physical location, any application, process, tool, or user can make one simple request using SQL, an API, interactive search, or a BI tool to get an integrated view of business data. The data fabric integrates various data sources in real time based on metadata/data catalog and policies. It minimizes complexity and hides heterogeneity by embodying a coherent model of data that reflects business requirements rather than the details of underlying systems and sources. It leverages several core open source and commercial technologies, including Apache Hadoop, Apache Spark, Apache Kafka, Apache Flume, Apache Hive, data warehouse, NoSQL, replication, and in-memory.

### The DataOS Data Fabric - architecture comprises six core capabilities:

**Data ingestion & Streaming:** DataOS offers a huge library of connectors to support multiple data source systems and users can effortlessly ingest real-time, batch, incremental and one-time historical data - be it structured or unstructured data. Seamlessly ingest any data at any volume and velocity from any source.

**Data Access & Discovery:** This capability automates the discovery of new internal or external data sources and presents them as a new data asset for consumption by business users. Dynamic discovery includes several components, such as data modeling, data preparation, curation, and virtualization, to deliver a flexible big data platform to support any use case.

**Data Management:** A critical capability of the data fabric includes automated metadata curation, data catalog, data dictionary, and data lineage. Data management enables the reliability, security, integrity, integration, and governance of data across the organization by orchestrating data flows, integrating disparate data, and transforming data across various silos.

**Data Processing & orchestration :** Data orchestration capability transforms, integrates, and cleans data to support various use cases in real time or near real time. From data cleansing to advanced transformations, DataOS is designed to simplify even the most complex data tasks into a few simple clicks. DataOS simplifies ETL jobs deployment by writing and executing ETL jobs on top of Apache Spark, Kafka Streams and supports standard Java or Scala based Spark programs.

**Data Governance and Compliance :** Data governance in DataOS is built on three key pillars - data quality, ownership, and data accessibility. With an inbuilt advanced governance engine, data admins can set rules, track datasets and define policies for secure data sharing. Moreover, with increasing regulatory norms around customer data privacy and protection, organizations need data platforms that automate and ensure customer data privacy. DataOS's inbuilt data compliance helps organizations to stay compliant with all existing regulatory norms (GDPR, CCPA etc.) put in place by state regulators.

**Self-serve data and analytics :** Business teams are dependent on IT teams to get hold of data that they want to perform analytics on and draw insights. This results in huge time delay between hypothesis initiation and insight generation. Business teams can utilize the self-serve data applications of DataOS, fetch the data by themselves, analyze, model and share the data without any dependency. These applications are completely code-free and can be utilized by users with limited technical abilities.



# Data OS<sup>®</sup> Data Fabric - Enabling Business Opportunities

Modern Data Fabric, as a modern data management platform, supports many functions in an organization:

- At the corporate level, DataOS enables the Chief Data Officer (CDO) to implement their data strategies aimed at enhancing data quality, reliability and access, enhancing analytical decision making and driving business or product innovation. DataOS does through its intelligent automation of data tasks and provisioning. DataOS through its sole focus on data can even enable CDOs to construct governance strategies to ensure data ethics throughout an organization and its ecosystem.
- DataOS modernizes your data warehouse and corporate data infrastructure by providing governed citizen access to all of our data and analytics, automated integration, analysis and data prep, and high-performance provisioning of data services. DataOS does this without burdening your current data warehouse and infrastructure with increased workloads thereby accelerating your business data and it's intrinsic value.
- The DataOS Data Fabric is a “citizen” data management tool enabling the automated integration, clean-up/prep and analysis of the ever growing and changing body of organizational data. It's unique AI capabilities can be used to create automated analysis/pipeline marcos which monitor operational process data such as demand planning or supply chain systems to alert business users or trigger automated causality reporting.
- The DataOS Data Fabric is also a data science laboratory assisting in the survey and assessment of organizational data through automated discovery and analysis of data mining for insights used in the development of Machine Learning models.
- DataOS also supports the operationalization of data science by pipelining data to data warehouses or data science platforms for consumption in production models. DataOS can also auto-analyze the operational data of a machine learning system to assess its performance in real-time, monitoring for model drift or bias, and even assist in the determination of explainability of a model's performance for further improvement and refinement over the course of its lifetime.