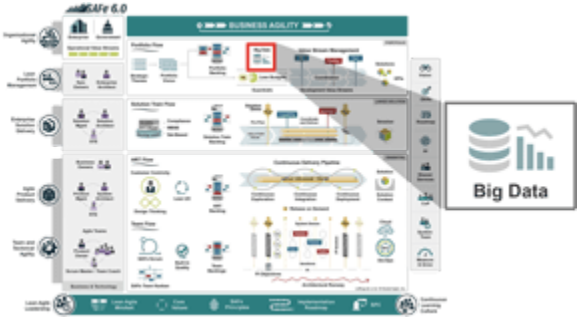
SAFe® STUDIO

Search

+ Framework



> *Just like oil was a natural resource powering the last industrial revolution, data is going to be the natural resource for this industrial revolution.*
>
> —Abhishek Mehta, CEO Tresata [1]

Big Data

Big Data refers to the roles and practices required to collect, manage, normalize and deliver large datasets that help enterprises make more informed, fact-based decisions.

Data has become critically important across the entire enterprise. It influences business decisions, helps create better products, improves product development, and drives operational efficiencies. This article describes data's critical role in the enterprise, the *DataOps* process to manage and deliver extensive volumes of data, and how to apply DataOps in SAFe.

Details

In the digital age, enterprises generate data at an astonishing rate. Each website click, turbine engine rotation, vehicle acceleration, and credit card transaction creates new information about products, consumers, and operating environments. The rapid acceleration of information has led to new practices for storing, managing and serving massive data collections [2]. These Big Data practices deliver purpose-built data products to provide value across the entire enterprise, as Figure 1 illustrates.

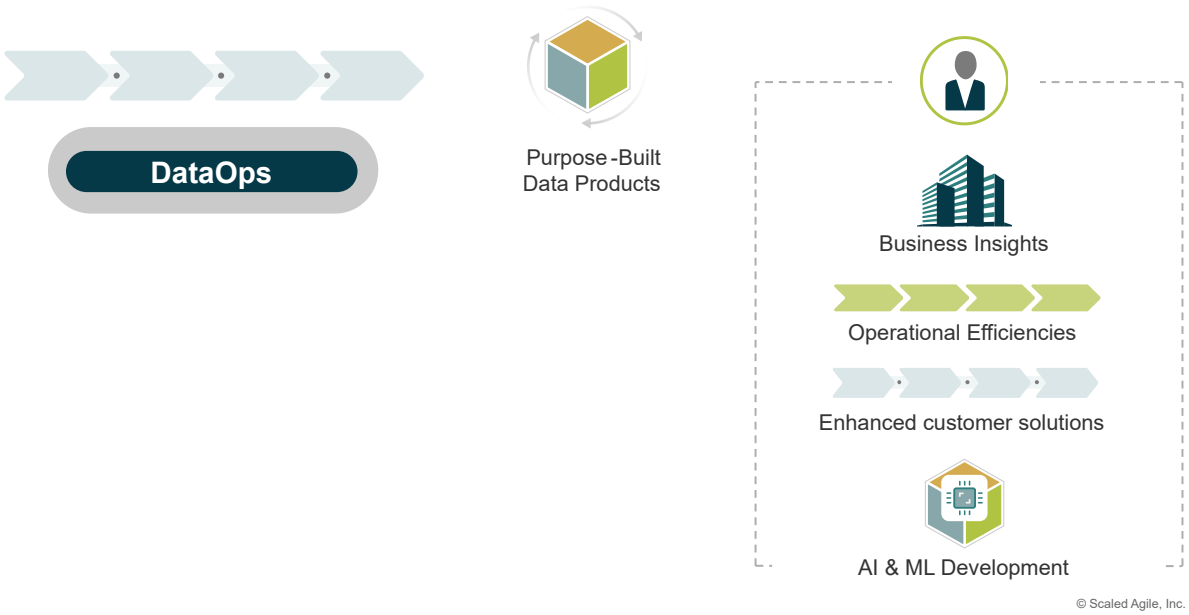


Figure 1. Big Data products support all parts of the enterprise

The Evolving Role of Big Data in the Enterprise

Data accumulation typically begins within organizational silos. A department collects information about its users and systems to enhance products, discover operational improvements, improve marketing and sales, etc. While this localized data is valuable, aggregating large data sets across the entire organization provides exponentially more value than siloed data.

Exploiting Big Data for Competitive Advantage

Every enterprise uses data to improve its products, optimize operations, and better understand its customers and markets. Media and consumer product organizations use big data solutions to build predictive models for new products and services to anticipate customer demand. Manufacturing uses big data solutions for predictive maintenance to anticipate failures. Retail businesses utilize big data solutions to

improve Customer experiences and effectively manage supply chains. Financial organizations use big data solutions to look for patterns in data that indicate potential fraud.

Supporting AI Initiatives

Organizations use Artificial Intelligence (AI) and machine learning (ML) as a competitive advantage to provide better products to their customers, improve operational and development efficiencies, and provide insights that will enhance the business. Artificial Intelligence initiatives focused on machine learning require large sets of rich data to train and validate models. Lack of sufficient data is a common reason for the failure of AI initiatives. To achieve AI goals, an organization must develop an enterprise-wide approach to collecting, managing, and delivering data collected across the organization integrated with external data to fill gaps.

Big Data Challenges

Collecting and aggregating this data poses challenges. The data community characterizes Big Data with the '3 Vs':

- **Volume** – Data insights require a broad spectrum of data collected across the enterprise that can scale to hundreds of petabytes. As an example, Google processes 20 petabytes of web data each day. Big data solutions must collect, aggregate, and deliver massive volumes of data to data consumers.
- **Velocity** – Data-driven decisions require the latest data. Velocity determines how quickly new data is received and refreshed from data sources. For example, a Boeing 737 engine generates 20 terabytes of information every hour. Big data solutions must decide which data to store and for what duration.
- **Variety** – Data originates in many forms across the organization. Traditional data from databases, spreadsheets, and text are easy to store and analyze. Unstructured data from video, images, and sensors presents new challenges. Big data solutions must address all types of data.

More recently, the data community has added Variability, Veracity, Value, Visibility, and other 'Vs' to characterize Big Data further and add to the challenges of storing, managing, and serving it.

Understand DataOps in the Enterprise

To address these challenges, organizations need a unifying approach. The Data Science community recognizes organizations' stages in the *Data Science Hierarchy of Needs* [3] (Figure 2). At the foundation, Data Engineers (likely System or Solution Architects on a Development Value Stream) design solutions that collect and manage data. This data and its storage are optimized for the application without concern for broader use.

As this data becomes more pertinent to the organization, Data Engineers (usually as part of a centralized data function) transform and aggregate the broader sets of application data into a data warehouse to make it available through data products such as marts, cubes, and views. Data Analysts and others use these read-only, purpose-fit data products for statistical analysis and visualizations. Data Scientists use them to develop and train models for AI and ML.

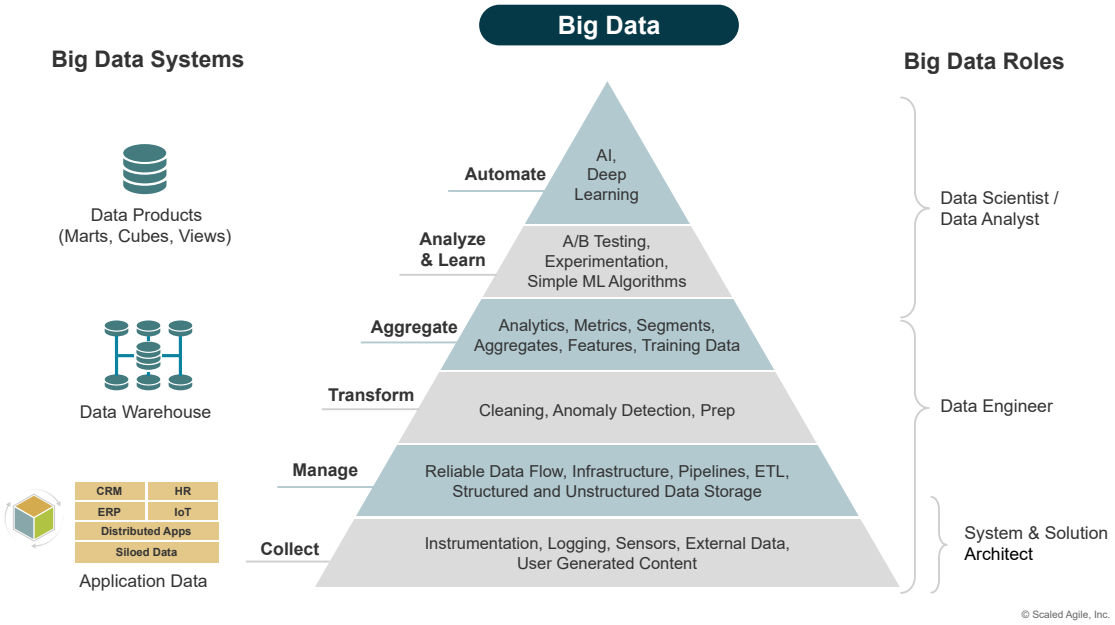


Figure 2. The Data Science Hierarchy of Needs

Individuals often play multiple roles. For example, a Data Analyst creating a dashboard with Data Engineering skills may go back into the data warehouse to transform and re-aggregate data for a new or updated view. However, an organization's governance rules may restrict the ability of individuals to act in multiple parts of the pyramid. While most organizations take a centralized Data Warehouse approach to data, distributed strategies like the Data Mesh [4] are emerging, particularly for large organizations.

The DataOps Lifecycle

The Big Data practices above are performed continuously as part of the *DataOps* lifecycle model shown in Figure 3. DataOps is a collaborative data management activity across Agile Teams, data practitioners, and enterprise stakeholders that leverages Lean-Agile and DevOps mindset, principles, and practices to deliver quality data products predictably and reliably.

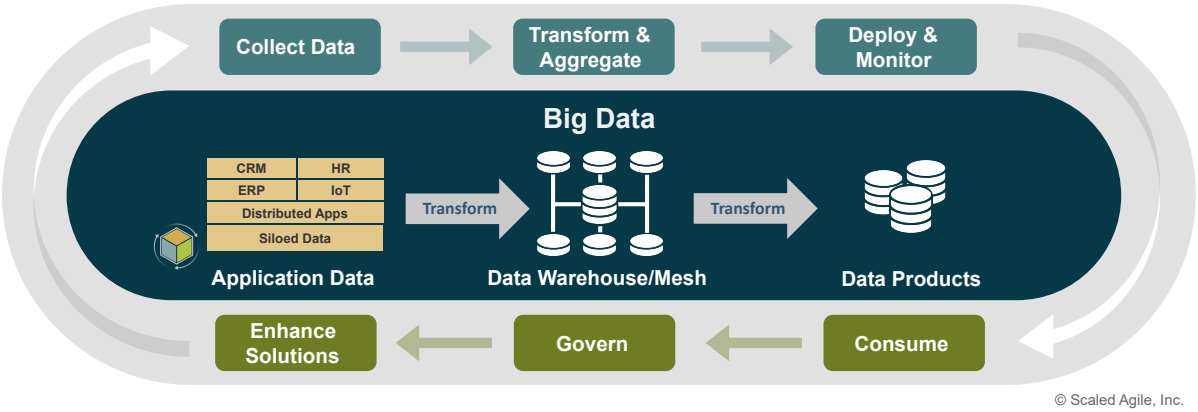


Figure 3. The DataOps lifecycle

The top portion of the lifecycle shows how siloed, application-specific data flows into data products offered to various consumers. The bottom part shows how data is consumed, governed, and used to enhance solutions that generate even more data into the pipeline. The remainder of this section describes each DataOps activity.

Collect Data – System and Solution Architects, Agile teams, and system administrators build telemetry, logging, and monitoring to gather data on system and user behaviors. Product Management prioritizes this work. System Architects ensure the solutions are easy to instrument and architected so the application data can be accessed externally, often through APIs. As data becomes more critical to the broader enterprise, work to collect better data and expose it through APIs may increase in priority and require capacity allocation (see ART Backlog) to ensure proper balance with other backlog items.

Aggregate & Transform – Data Engineers transform and aggregate data from across the enterprise into normalized forms optimized for efficient use by data consumers. A centralized data architecture ensures efficient storage and delivery of consistent data products across the enterprise. Data Engineers must balance the 'Vs' discussed earlier and determine which data to store, how long, tolerable access

times, etc. They view data as a product and apply Design-Thinking and Built-in Quality. Personas and Journey Maps help them empathize with data consumers' pains, gains, and user experiences and determine how to offer better data products.

Deploy & Monitor – Data Engineers deploy data products from the data warehouse into various forms, including data marts, cubes, and views used by data consumers. Like other technology solutions, data solutions leverage Cloud technology and apply Continuous Delivery through a DevOps pipeline designed for data. These practices quickly move data changes through development, Q/A, UAT, and production environments for consumer feedback. Multiple environments ensure dashboards, reports, models, and other artifacts dependent on the data content and formats can evolve with the data.

In the spirit of DevOps, monitoring occurs across all stages of the data pipeline to detect anomalies in the data (row counts, data out of range) and data operations (unexpected timeouts) and sends alerts to the data team.

Consume – Big Data consumers can be categorized into two groups, as shown in Figure 4. Analysts use data products to discover insights and create visualizations for specific data customers. Data Scientists and ML developers use them to develop and train models. Their customers (Data Customers in Figure 4) are enterprise-wide stakeholders seeking business insights for decision-making, improving operations, and enhancing solutions.

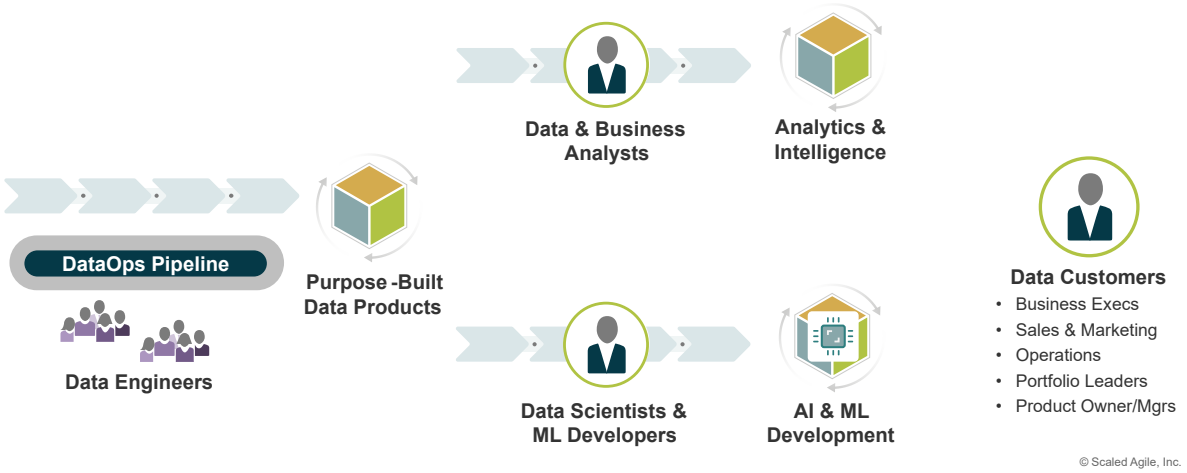


Figure 4. Big Data has many customers across the entire enterprise

Different lines of business, product managers, AI developers, and others have unique data needs. To enable these individuals through a self-serve model, enterprises should invest in the tools, analytics packages, and resources to support ad hoc

access for localized manipulation. This investment reduces the load on members of the centralized data team for reporting and other mundane tasks.

Govern – DataOps must enforce data privacy, confidentiality, residency, sharing, retention, and other legal requirements. The Big Data solution must ensure security through access controls, audits, and monitoring that detect intrusion and data breaches. Like other digital products, it must also provide fault tolerance and disaster recovery through vendors or home-grown solutions.

Enhance Solutions – Portfolio and Agile Release Train (ART) leaders use data to enhance solutions that provide better customer value (better products) and business value (better data). Data analytics can reveal opportunities for new solution offerings and inform feature prioritization for existing solutions. Data gaps also show opportunities to enhance solutions to collect additional data.

Applying DataOps in SAFe

The previous sections describe the importance of a clear and compelling data strategy. This section describes additional guidance SAFe organizations can take to support their Big Data journey.

DataOps is a Portfolio Level Concern

SAFe addresses Big Data concerns at the portfolio level as it requires vision, investment, and governance at the highest levels in the organization (Figure 5). While ARTs create the data, the value comes from data aggregation at the portfolio and enterprise levels. Big Data solutions require strategic investment from the organization and a comprehensive approach that aligns each of the organization's development value streams to common DataOps practices that produce cohesive data sets used across the entire organization. Portfolio leaders use Lean Budgets to invest in Big Data infrastructure and DataOps practices to accomplish this. They also use Portfolio Epics and the Portfolio Backlog to specify and prioritize the infrastructure, technologies, and data needs to support the organization (Figure 5).

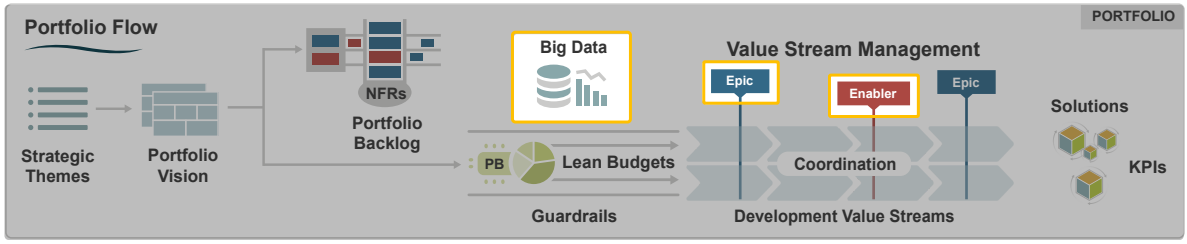


Figure 5. Big Data is a portfolio concern

Initially Centralize the Data Function

SAFe's Principle #10, Organize Around Value, strives to optimize flow by ensuring teams and ARTs have all the skills necessary to deliver value. Unfortunately, most organizations are still growing their data engineering and science functions, resulting in more demand than capacity for these skills. Initial centralization is often helpful for early technology adoption to maximize available skills and create the DataOps infrastructure and practices. Centralization also simplifies governance for privacy and security that would be virtually impossible to safeguard with a siloed approach.

This centralized function can meet most of the organization's data needs through the customer-centric approaches described earlier. Where additional support is needed, SAFe has a known pattern of providing additional services to other parts of the organization through a Shared Service, as shown in Figure 6.

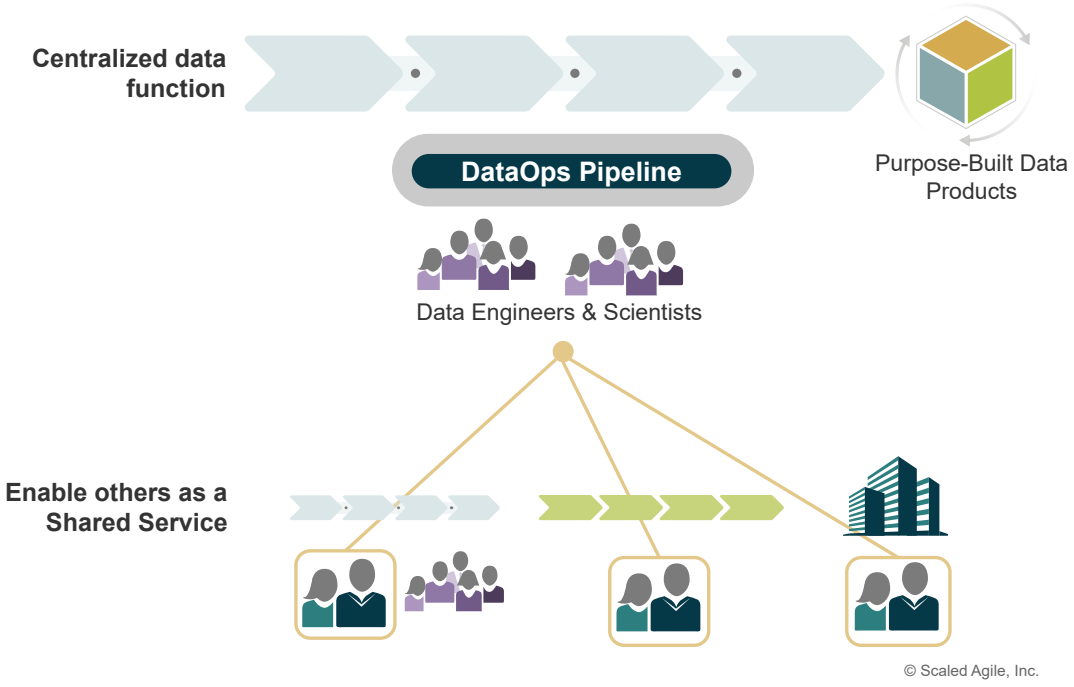


Figure 6. Start with a centralized data function that supports other parts of the organization as a Shared Service

Over time, organizations will grow their data functions to support the broader enterprise and embed individuals in ARTs and Operational Value Streams. However, a 'Big Data' development value stream of Data Engineers providing aggregated enterprise data will likely exist for some time.

Grow Technical Talent

Data Engineering, Analysts, and AI/ML developer skills are in huge demand, and recruiting knowledgeable and skilled data specialists is a significant challenge. Organizations must create a compelling and inspiring data and AI vision to attract and retain this talent. Data specialists want to learn and grow from other data specialists to keep pace with the rapidly evolving technologies and practices.

Applying DataOps to Build Better Solutions

Solutions are informed and augmented by data, and ARTs require Data Science and Data Engineering skills. As described earlier, the data functions can provide some resources to ARTs as a shared service. But they must balance that ART work with their primary responsibility to create and evolve the DataOps practices that include providing data products to the enterprise. When supporting ARTs, they should act like an Enabling Team (see Agile Teams) to grow the technical data competence across the organization. In this capacity, they are not there to do the work but to teach others how to do it.

Learn More

[1] Deutscher, Maria. *IBM's CEO Says Big Data is Like Oil, Enterprises Need Help Extracting the Value.* SiliconANGLE, March 11, 2013. Retrieved October 13, 2023, from https://siliconangle.com/2013/03/11/ibms-ceo-says-big-data-is-like-oil-enterprises-need-help-extracting-the-value/

[2] Kleppman, Martin. *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems.* O'Reilly Media, 2017.

[3] Arsenault, Marc-Olivier. *The data science pyramid.* Medium, September 26, 2018. Retrieved October 13, 2023, from https://towardsdatascience.com/the-data-science-pyramid-8a018013c490

[4] *Data mesh.* Wikipedia. Retrieved October 13, 2023, from https://en.wikipedia.org/wiki/Data_mesh

Last Update: 13 October 2023

Framework

Download SAFe Posters & Graphics

Watch and download SAFe videos and presentations

Blog

Training

Course Calendar

About Certification

Become a Trainer

Content & Trademarks

FAQs on how to use SAFe content and trademarks

Permissions Form

Usage and Permissions

Scaled Agile, Inc

Contact Us

5400 Airport Blvd., Suite 300
Boulder, CO 80301 USA

Business Hours

Weekdays: 9am to 5pm
Weekends: CLOSED

FAQs on how to use SAFe content and trademarks

Permissions Form

Usage and Permissions