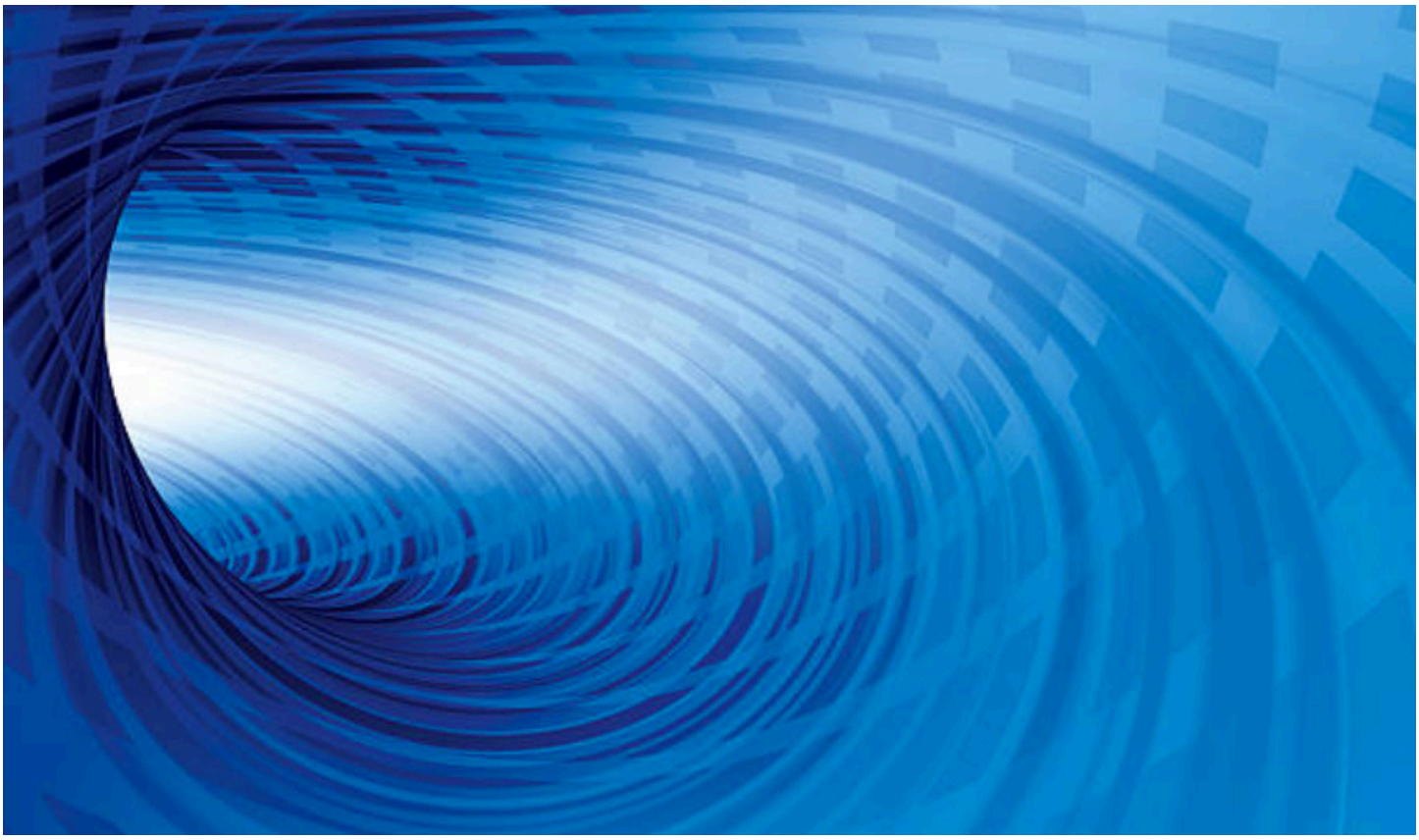
October 2018

MIT Technology Review

Custom Research

Fueling Decisions and Actions With Data Pipelining



Need to respond instantly to fast-changing markets? Automated, repeatable data orchestration speeds insights for the intelligent organization.

Sponsored by:

SAP®

EXECUTIVE SUMMARY

Digital businesses need to act quickly on their data, yet traditional data tools and processes are slow. By the time these tools can find, move, process and analyze critical information, the moment for gaining a competitive advantage may have passed.

A new approach can help. Known as "data pipelining," it uses automated, repeatable processes and new technologies to find, cleanse, enhance and analyze any form of data at its source, eliminating the cost and delay of moving the data. Data pipelining is one component of data orchestration, which delivers real-time awareness of customer needs, market trends and operational issues. These deliverables can drive benefits, including:

- **Excellence in customer-facing activities.** By using intelligent data, organizations can respond quickly to changes in customer behavior and markets. These responses can include revised pricing, churn prevention, cross- and up-sell offers, and promotion optimization.

- **Lower risk and cost,** the result of using comprehensive, up-to-date data to optimize internal operations in areas such as predictive maintenance, supply chain optimization and fraud prevention.

- **Expanded digital offerings,** enabling new business models; sales of customer, market and product data; and new analytics-as-a-service offerings.

Data pipelining can also help enterprises scale their big data programs without adding expensive staff, instead using automated and reusable processes and best practices. And by applying machine learning to data management processes, a data pipeline can become faster and more effective with each passing day.

CHANGING NEEDS, CHANGING PROCESSES

The raw materials of industrial businesses in the last century were oil, sweat and steel. For today's digital businesses, it's data. A modern retailer can gain a competitive advantage by knowing the shopping habits of individual consumers; a manufacturer, by knowing the wear patterns of a drill; and a pharmaceuticals supplier, by detecting quality issues in its supplies. They share the pressing need to obtain and act on data quickly and easily.

Today's digital businesses need to make sense, quickly and cost-effectively, of the mass of information available to them. This data comes not just from PCs, tablets and smartphones, but also from new internet of things (IoT) devices and sensors. With so many devices generating so many kinds of data that could drive competitive advantage, an organization's ability to quickly analyze and act on that data is a new business requirement.

Many such projects fall under the heading of big data. These require an organization to quickly discover, understand, gather, cleanse and analyze all business-relevant information about a product, customer or market. This might include online information, such as the location of a customer, the vibration level in a turbine or the moisture level in a farm field. It could also include off-line or historical information, such as a customer's past purchases, previous links between vibration and turbine failure, and the relationship of moisture levels to crop yields in a specific farm field.

The sheer volume, variety and velocity of such data — plus the fact that it's often stored in multiple locations — makes much big data analysis difficult. Nine in 10 organizations have not yet reached a "transformational level of maturity" in data and analytics, finds a recent global survey by Gartner, despite this being a top investment priority for CIOs.[1]

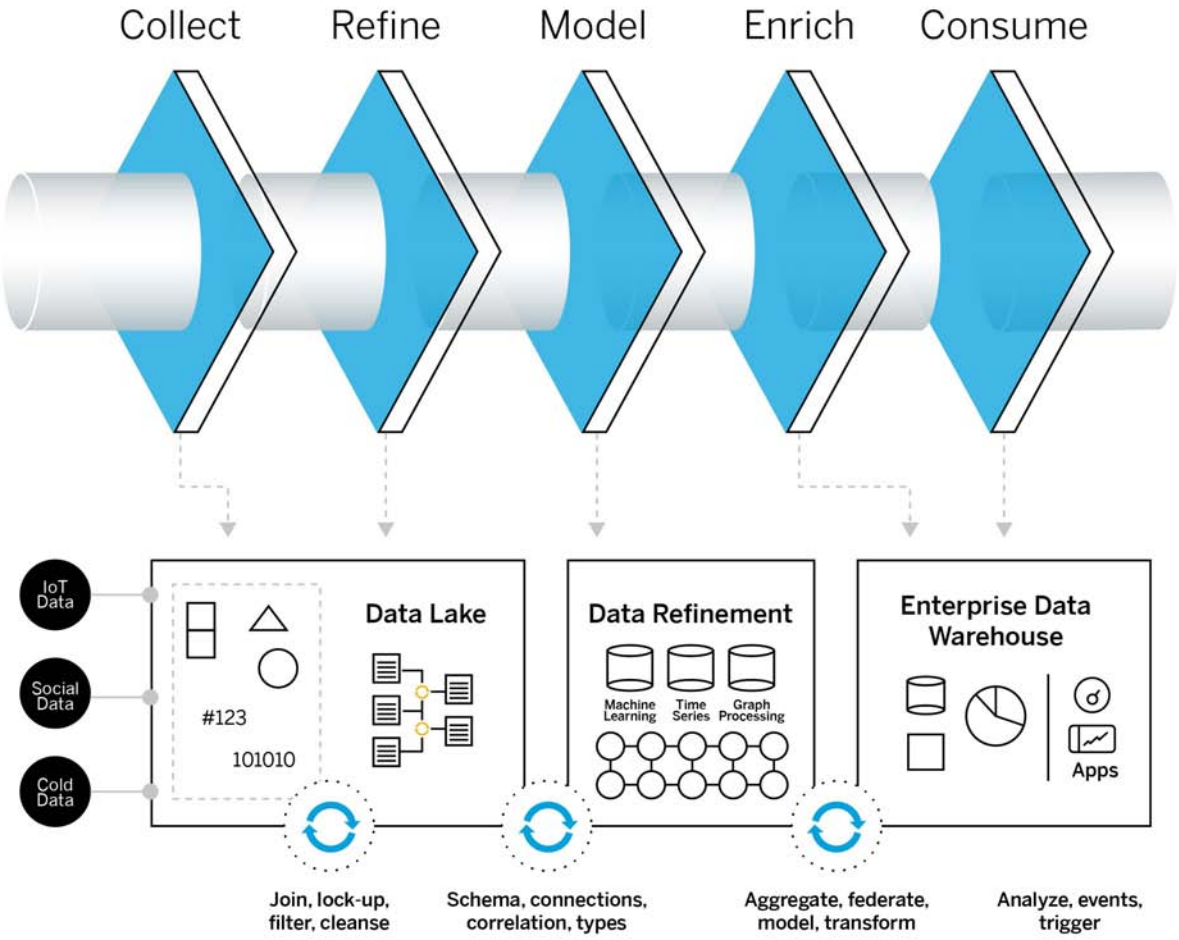
Why? In part, because a great deal of data remains in silos. Analyzing this data requires manual processes, such as loading transaction data into data warehouses, as well as cleansing and enhancing the data. Only then can the data be analyzed and acted on. For organizations seeking to drive the frontiers of excellence with higher levels of speed and innovation, this process is hazardously slow. For example, by the time a retailer knows a customer is interested in a product, the customer may have already moved on, either to another store aisle or another website. Similarly, by the time a plant manager knows a drill has gone dull, the tool may have already been used to create defective parts. And by the time a pharmaceuticals supplier discovers a raw material is contaminated, thousands of pills may have already been created, all needing to be recalled.

A new approach known as "data pipelining" can help by giving business decision-makers new tools for uncovering valuable insights from their masses of data. Data pipelining is the automated and reusable process of data ingestion, cleansing, enrichment, refinement and sharing among all relevant business stakeholders. It uses automation, processing and analyzing data close to where

1. "Survey Shows Organizations Are Slow to Advance in Data and Analytics." Gartner press release, February 5, 2018. https://www.gartner.com/newsroom/id/3851963

Figure 1

Inside the End-to-End Data Pipeline



it was generated (see sidebar, "What Moves Through the Data Pipeline?"). Data pipelining also helps achieve the data orchestration required for real-time analytics even when dealing with the data volumes generated by IoT deployments or required for big data analytics.

Such automated, reusable data orchestration is both much faster than traditional manual processes and much less expensive. It also provides a higher degree of governance and metadata — that is, helpful data about the data — enabling even nontechnical users to quickly find and evaluate the information they need.

Data pipelining achieves these goals by combining technologies and processes to provide repeatable, factory-like workflows that meet the multiple challenges that can slow or even stop traditional approaches. That speed is vital, as organizations in every industry now struggle to respond to fast-changing user needs and market conditions. "Things that used to be analyzed in batch time now need to be analyzed in real time to provide relevant feedback to the business," says Robin Bloor, a principal at IT research firm Bloor Group.

Data pipelining offers a flow of insights so rapid and sophisticated that organizations can move from merely descriptive analytics to diagnostic, predictive and, ultimately, prescriptive insights. "Whoever is on the front line with customers should be looking at how to take advantage of data in a smart way," says Toph Whitmore, an independent analyst and researcher, and the author of a recent report on DataOps (another term for data pipelining).[2]

2. Toph Whitmore. "DataOps: The Collaborative Framework for Enterprise Data-Flow Orchestration." Blue Hill Research, January 2017.
https://www.datakitchen.io/content/RT-A0287-DataOpsDefined-TW-DataKitchen-Final.pdf

Figure 2

The Data Pipeline Difference

Digital Challenge	How Data Pipelining Delivers
Complex data landscape limits agility	Automated, centralized orchestration of data
Having confidence in data quality and meeting compliance regulations	A single, cross-landscape data control center to manage governance policies
Inability to get the most out of data	Efficiently processes data from all sources to unlock new use cases
High data costs	Minimizes data movement and duplication

Data pipelining enables organizations to "operationalize" new analytics capabilities, such as machine learning and artificial intelligence (AI), and leverage data from new sources, such as IoT deployments. Rather than requiring a new, expensive data preparation process for each new analytical request, data pipelines provide automated and even self-service access to data.

One powerful application is anomaly detection. Machine learning systems can first observe and learn the baseline behavior of a system or users, and then watch for exceptions. For example, if an employee who normally works 9-to-5 suddenly attempts to log in at 3 a.m., the system, detecting anomalous behavior, would trigger an alert, spurring a deeper investigation and possible preventive action. If the login is valid, because the user was on a business trip, the system would take that into account for the future, generating fewer warnings for after-hours logins for this person. If the attack was genuine, the system could raise its future alerts to even higher levels. Either way, the system "learns" and improves its own performance over time.

Unstructured data (mentioned above) is another important application for data pipelining. Traditional database systems were designed to handle data that's highly structured, such as customer accounts, parts inventories and financial records. But today's organizations must also deal with a large and fast-growing body of semi- and unstructured data, much of it coming from videos, social media, text messages and industrial IoT sensors. Any organization that can correlate both types of data — structured and unstructured — can enjoy a major competitive advantage. For example, consider a retailer able to identify which of its SKU numbers describes the "cool red handbag" now trending on social media. The retailer could then increase the bag's production and distribution, capitalizing quickly on the new demand.

Further, data pipelining can help organizations identify, access, manage and analyze valuable data that may be stored almost anywhere, including on-premises data centers, the public cloud and massive repositories of raw data known as data lakes. With traditional approaches, most data must be moved to a central repository before it can be analyzed. But with data pipelining, the data can be analyzed wherever it resides. The results can then be quickly shared with the right users. What's more, the ability of data pipelining to manage and present metadata can help users find, identify and evaluate the information that's most useful to them — all major challenges of the digital age. Storing data is increasingly inexpensive, but getting value out of that data is trickier, says industry watcher Whitmore. "I hear customers complaining they have a massive data lake, but don't know what to do with it."

SEVEN DATA PIPELINING ESSENTIALS

To meet today's needs for faster, more accessible analytics, the data hub that enables data pipelining should include several key capabilities:

1. **Real-time or near-real-time data management** to speed analytics. This requires careful prepositioning of the required data, the tools to manage it, and the methodologies to properly prepare and then analyze the data. Organizations need to monitor not only the time required for each step in the data management process, but also the overall time needed for all steps. That's the only way to ensure a delay in one process, such as data cleansing, doesn't interfere with other data preparation steps necessary for the timely execution of critical analytics.

2. The ability to **analyze the data in place,** near where it was created. This eliminates the cost and delay of moving data for processing. Researcher Bloor says doing this effectively requires a central querying capability that runs distributed queries and then aggregates the results. The larger the number of nodes, he adds, the more sophisticated the software needs to be.

3. The ability to **discover, process and analyze both structured enterprise data and unstructured data** from, for example, social media and video. This is needed for the most comprehensive analysis of customer, market and business issues.

4. An **open design** that enables the use of multiple machine learning technologies, such as the TensorFlow software library for data flow programming and Apache Spark's MLlib scalable machine learning library, to automatically improve data orchestration over time. The orchestration platform must also allow the use of popular data management, distributed computing, analytics and query engines, such as Apache Hadoop, Apache Hive, Presto and Impala.

> "We're seeing self-service driving more adoption of analytics, with data preparation and management done by those end users."
>
> —Matt Aslett, research director, 451 Research

5. **Self-service capabilities** that speed user access to both data access tools and the required analytics infrastructure. "Historically, data management products were sold to data administrators and data engineers in IT who create data management pipelines, reports and analytics for end users," says Matt Aslett, a research director at IT analyst firm 451 Research. Today, he adds, "we're seeing self-service driving more adoption of analytics, with data preparation and management done by those end users." To help users perform more of that work themselves, Aslett suggests, organizations can look for tools with features such as an easy-to-use interface and collaboration capabilities.

6. **Metadata management** facilitates the discovery, management and sharing of metadata to help users find and understand data, and to help applications and big data tools refine and enrich it. For example, metadata can help sales agents evaluate the usefulness of raw sales data, showing them when the data was generated, who else has used it and how useful these other users found it.

7. **Support for data orchestration** — the automated, reusable processes required to ingest, process, share and analyze all the data business users need to make better decisions — and the use of AI to improve those processes. Along with data pipelining, orchestration comprises two other components: landscape management (discovering, understanding, managing and securing the ever-changing sources of enterprise data) and data governance (assuring the provenance and integrity of the data with policy-based management and the use of metadata catalogs to understand the state of the data and how it can be shared). Using these capabilities, data orchestration can reduce the time and cost required for data management.

NEW CHALLENGES, NEW TOOLS

Data pipelining also requires organizations to adopt new technologies that enable both real-time analytics and the underlying required data management. These move data management beyond the conventional functions of extract/transform/load, data cleansing and master data management to true real-time and in-place data enrichment, refinement and analysis. Such technologies include:

- **In-memory cluster computing and distributed computing solutions** such as Apache Spark, which can execute high-performance analytics on data from multiple sources.

- **Machine learning.** A subset of AI, this technology continuously uses experience to improve the analysis of data and, eventually, the data management functions needed to enable that analysis.

- **Cloud platforms.** Today's highly extensible cloud platforms provide not only storage, compute and networking but also big data tools. And they do so more flexibly and at lower cost than has ever been possible before.

- **Data hubs.** These enterprise orchestration platforms enable data pipelining by providing consistent, centralized policies and processes for everything from data cleansing to data governance. In this way, enterprises can quickly and cost-effectively manage and share data from any source, including data lakes, data warehouses and legacy applications running either on-premises or in the cloud.

- **A new generation of query and reporting tools** deliver advanced analytics capabilities. They're also far easier to use than earlier versions.

8 STOPS ON THE DATA PIPELINE

Data pipelining isn't a product or even a combination of products. Instead, it's a set of standardized, repeatable and automated data operations that enable the orchestration of data business users need. It empowers a broad range of users and applications to analyze a broad range of data in place, all without the cost and delay of moving that data to a central location.

Organizations that want to take a comprehensive approach to data pipelining should make these stops:

1. **Ingest the data.** This is the ongoing process of identifying and understanding all the internal and external sources of data required for a complete picture of customer needs, operational trends and market conditions. This includes understanding when new data becomes available, where it's stored and how it must be processed to become useful.

2. **Understand the data.** Correlations and relationships in the data can be made visible by offering metadata to users. By describing the source of a dataset, the type of information it contains or previous ways the data has been used, metadata can give business users new and different types of information for solving business problems. For example, if the metadata for a collection of social media comments includes the term 'customer satisfaction,' business users looking to increase customer retention might know to include that data in their analysis, and thus identify users at risk of defecting to a competitor through their posts. Easy-to-use metadata catalogs can also help nontechnical users find, understand and use this information.

> "It's difficult to know when to inject data cleansing without some analysis of the system."
>
> —Robin Bloor, principal, Bloor Group

3. **Refine the data.** Data quality management and data cleansing become increasingly important in the context of real-time analytics and new data sources, such as IoT deployments. "People sometimes assume that data is good if it comes out of a sensor, because it didn't go through a process with humans involved. But a lot of data is bad because of program or sensor error," says researcher Bloor. "It's difficult to know when to inject data cleansing without some analysis of the system."

4. **Enrich the data.** This involves correlating data from the enterprise with data from other sources, such as social media. It provides accurate, real-time insights into which products or services a customer is looking for, commenting on or seeking to purchase, ultimately helping the organization increase both sales and customer retention. For example, if a customer complained about a store's customer service on social media, finding his order details in a customer relationship management system could enable the retailer to proactively contact that customer with detailed help that might prevent him moving to a competitor.

5. **Analyze the data in place.** This eliminates the delay and cost of moving data to a central location for analysis, a key advantage of data pipelining. In-place analysis also helps an organization understand activities that take place at the edge of the network — such as a production facility or store — and then take needed actions quickly enough to prevent an interruption in production or a lost sale.

What Moves Through the Data Pipeline?

Despite its name, a data pipeline doesn't move data. It moves automated, proven and reusable instructions for how to quickly and efficiently ingest, cleanse, manage, share and analyze the data from a central repository to every platform that manages the data.

Also, each process moves automatically through the pipeline from one step to the next. This automates processes such as eliminating redundant or misleading information without requiring highly paid staff to manually trigger new processes and applications.

By enabling data orchestration, a data pipeline minimizes the cost, delay and effort of moving large amounts of data from the point of creation to where that data is cleansed, refined, enriched and analyzed. A data pipeline also assures that the most current data is used. It does this by eliminating the need to capture and move a data snapshot that rapidly falls out of date.

6. **Automate processes.** Automating the execution of these steps reduces both the time and the cost of performing business-critical data analyses. It also improves the consistency of an organization's data management, which helps increase security and efficiency.

7. **Continuously improve.** Organizations that track metrics — such as the cost and speed of data movement, curation and transformation, as well as workflow speed — can more easily improve them. Researcher Whitmore expects AI will eventually be able to improve process management itself, but he adds that "eventually" could still be five years off.

8. **Assure compliance.** Security, privacy and regulatory compliance can be automated with centralized, orchestrated processes such as encryption, identity management, access control and data usage audits. Together, these measures can dramatically reduce both human error and criminal theft.

THE FUTURE OF DATA PIPELINING

Even if most organizations aren't yet thinking in terms of data pipelining, many are implementing some of its capabilities. Researcher Aslett says these include self-service analytics, self-service data preparation, collaborative data governance, and modernizing both data governance and data management processes. In the future, organizations may trade open-source pipelines for common business needs in online marketplaces. They may also create multithreaded pipelines that can provide different data orchestration processes to different types of workers.

Many organizations seeking to adopt data pipelining will need new technology platforms to support the centralized orchestration of data management, as well as automate complex data operations such as refinement and enrichment. These platforms will also need to work with multiple on-premises platforms, cloud platforms and data management execution engines.

With the right platform and processes, the data orchestration enabled by data pipelining can help organizations in a wide range of industries drive faster insights from their data. This, in turn, can help them achieve competitive advantages that are both serious and sustainable. ■

ABOUT MIT TECHNOLOGY REVIEW CUSTOM RESEARCH

MIT Technology Review Custom Research is the strategic research and thought leadership publishing practice of MIT Technology Review. The mission of MIT Technology Review is to equip its audiences with the intelligence to understand a world shaped by technology.

The author of this report, Robert L. Scheier, has written extensively on business technology for more than 15 years.

COPYRIGHT AND DISCLAIMER

MIT Technology Review

Custom Research

Orchestrate Your Data for an Intelligent Enterprise

Corporate data landscapes are growing increasingly diverse and distributed. Data volume is exploding with unstructured data from internet of things (IoT) deployments and social sites. Moreover, data is stored in multiple locations — on-premises, in the cloud, in data warehouses and data marts, and on edge devices. Uncontrolled data consumption and insufficient security and governance across the distributed data landscape make it difficult for companies to share intelligent data and insights. Furthermore, existing systems and processing methods, which were built primarily for managing structured transactional data, are typically point-to-point and highly manual. The result: Enterprises are collecting vast amounts of data, but for most, it's a treasure trove of data they can't unlock.

> SAP Data Hub ingests, processes, shares and analyzes diverse data with data pipelines. It orchestrates the pipelines with data workflows, provides effective data governance with metadata discovery and profiling capabilities, and helps provide complete visibility and control across the data landscape with an open and flexible architecture.

To become an intelligent enterprise, businesses need to make sense of this growing volume and diversity of data. For example, new technologies such as machine learning need scalable techniques to gather, prepare and manage massive quantities of data. SAP has addressed this need with SAP Data Hub, an all-in-one data orchestration solution that covers distributed landscapes across multiple cloud providers and on-premises data centers and streamlines innovation projects across different data sources and data processing paradigms. SAP Data Hub ingests, processes, shares and analyzes diverse data with data pipelines. It orchestrates the pipelines with data workflows, provides effective data governance with metadata discovery and profiling capabilities, and helps provide complete visibility and control across the data landscape with an open and flexible architecture.

"The ability to quickly create powerful, scalable data pipelines that refine and enrich all the raw data — while also ensuring data governance — dramatically speeds the delivery of the intelligent data businesses can use," says Amit Satoor, Senior Director, Digital Platform Product Marketing at SAP.

By orchestrating and governing any type, variety and volume of data across an entire distributed data landscape, SAP Data Hub rapidly delivers enriched, trustworthy, intelligent data to the right users with the right context at the right time.

For more information, please visit www.sap.com/datahub

PRESENTED BY: