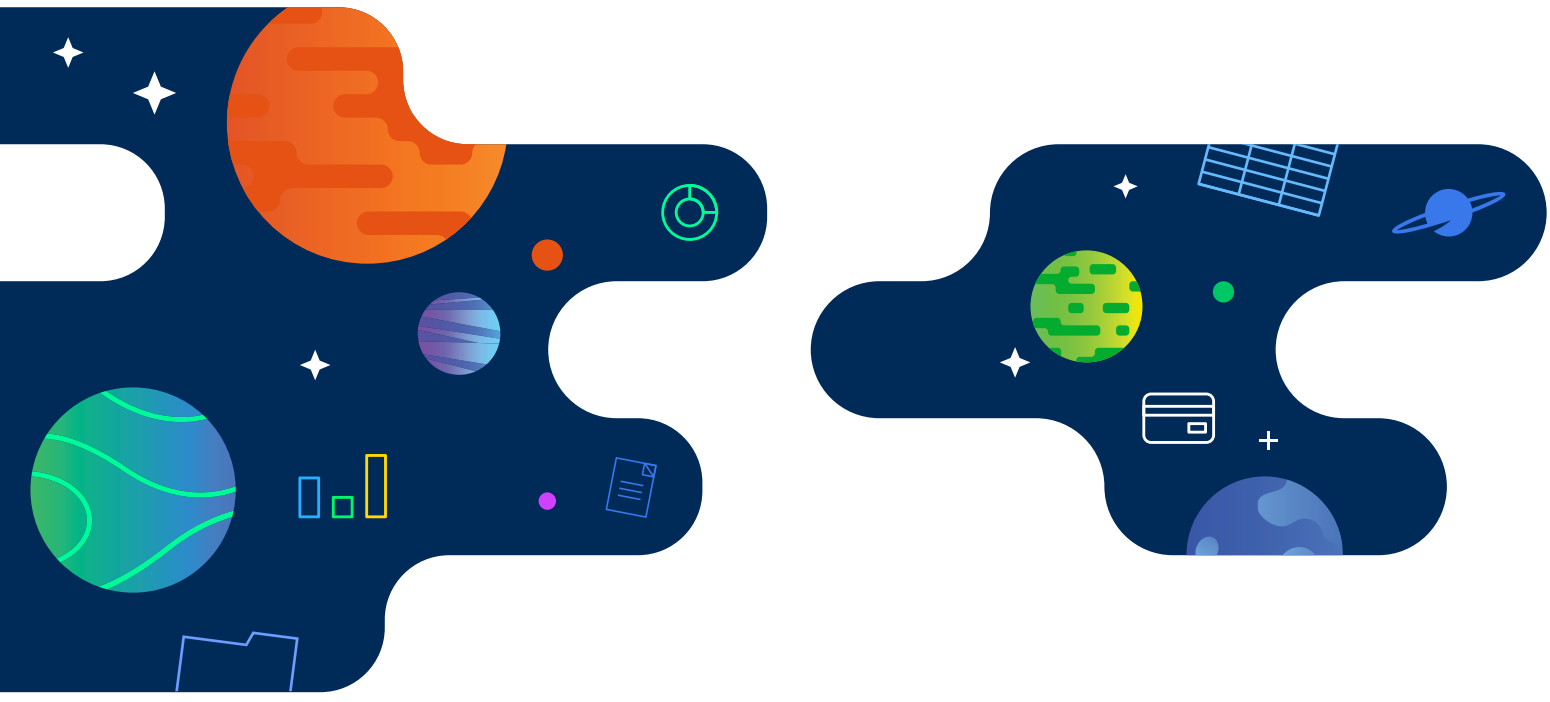
The CIO's Guide to Shattering Data Silos

Why standardizing on an open stack is the key to unlocking your data's value



DataStax

Table of Contents

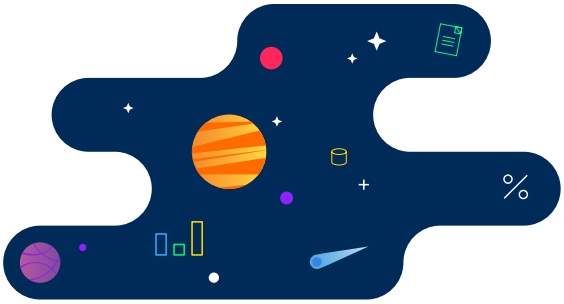
"If 60% of organizations have embarked on digital transformation, yet on average, only 38% of data is leveraged, there is a gap between intent and results in many cases."

IDC, December 2021[1]

Enterprises are awash in data, and the incoming flood isn't about to abate any time soon. Analyst firm IDC reports that 64.2 zettabytes of data were created in 2020, and predicts the amount to increase by a compound annual growth rate of 23% through 2025.[2]

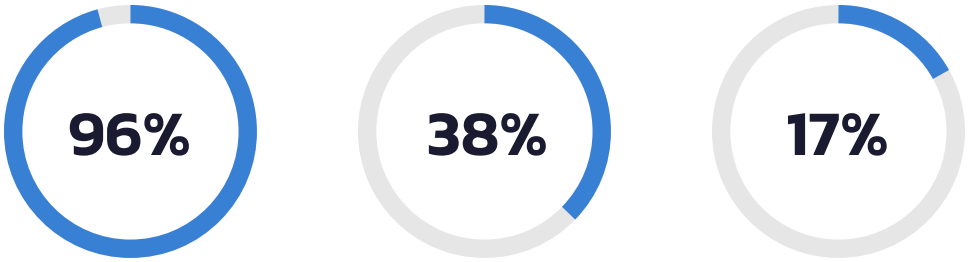
Most technology leaders understand that data is at the heart of how their organizations create value for their customers. Data growth should provide enterprises with fuel for transformation and growth acceleration, but often, it only increases cost, complexity, and frustration. Data is often stored in ways that make it very difficult to put to work. The question comes up: Why can't I take full advantage of this asset that I know is so critical to winning?

The challenges of complexity



A survey of over 500 technology executives and practitioners revealed that 96% claimed to have a data strategy, but only 38% agreed strongly that their data was creating value for customers, and even fewer (17%) said data and analytics drove more than 20% of revenue.

Universal ambition, uneven progress



Have a data strategy

Excel at using data to deliver value to customers

Attribute more than 20% of revenue to data and analytics

There are several reasons for this disconnect. In an attempt to harness the value of their data, many enterprises, over time, invested in a variety of point technology solutions. While this might have worked for one team or one project or one application, the end result of this effort was to lock data in a variety of silos across the organization.

Digital transformation, somewhat ironically, has contributed to this complexity. As services and applications became increasingly decoupled and finer-grained, they multiplied. And so have the operational databases and abstraction layers that support them. Most large enterprises support dozens of mostly proprietary and highly dislocated NoSQL and SQL databases. That translates into a large number of costly licenses and a lot of operational maintenance resources.

These silos make it hard for developers to be agile, because the data they need to build new applications is scattered about the organization and difficult to access behind disparate interfaces and rigid database technologies that absorb a lot of developers' attention. Likewise, they prevent enterprises from getting the big picture about and context regarding the intent of their customers. When data is fragmented in silos, innovation gets stymied—we refer to it as the "innovation stalemate." Without modern, cloud-native technologies, bringing new innovations to market quickly is almost impossible.

Cost becomes a major challenge, too. With all these different technologies and data silos, enterprises have to maintain too many products and too many disparate skills; the costs of scaling data begin to spiral out of control.

"The immense potential of the data—as a real-time and contextual revenue driver—remains unrealized. It's just too hard to capitalize on the opportunity.

But what if we could simplify these data environments and unlock the value of this data trove, by increasing data availability and speed, and what if we could significantly reduce the license and operating costs?"

Cody Sanford, former CIO, T-Mobile

Data shouldn't be a problem

How does an enterprise do away with the fragmented array of data silos and disparate technologies that hamper data availability and speed? How can it reduce the licensing and operating costs that come along with managing so many data stores—and unify security practices and data governance? In other words, how can an organization unlock the power of its data and free its developers to innovate? The answer is to store and manage your data in a consistent way that enables you to use it in real time.

"Enterprises accomplish data standardization by unifying operational data stores into a single, cloud-native architecture to abstract, stream, combine, and make it readily available."

This isn't to suggest that organizations do away with the multitude of different data types, use cases, and data sources. Rather, enterprises need to unify all of their operational data stores into a single, cloud-native architecture to abstract, stream, combine, and make data—regardless of the data type—readily available to all the high-scale applications (and the developers that build them) across the organization.

In a non-standardized environment, data can become a burden—regardless of its potential value. Introducing new data causes a series of side effects, not the least of which is the need to decide where to store that data, and whether a data store is scalable enough to handle it. There are, of course, strict controls on how new data enters the organization.

This isn't surprising, as security and compliance requirements are always critical to protecting valuable (and often sensitive) data. But in a fragmented approach, when there are multiple data stores and other point technology solutions across the organization, those protective requirements must be applied multiple times, in multiple ways.

"CIOs would be greedy when it comes to data. If there isn't a scalable way to absorb whatever data comes an organization's way, modern CIOs aren't doing their jobs."

This is antithesis to the way a modern enterprise should approach data. A data stack that provides end-to-end security helps enterprises manage risks by standardizing an organization's security posture.

And it enables CIOs to be greedy when it comes to data; they should prioritize a scalable capability for ingesting it, because it's absolutely critical to meeting customer expectations, building new products, and driving revenue. Essentially, if there isn't a scalable way to absorb whatever data comes an organization's way, modern CIOs aren't doing their jobs.

Apache Cassandra: The NoSQL standard

A way that more and more organizations are unifying their data architectures is to "rationalize and consolidate the overall DBMS portfolio," as analyst firm Gartner puts it.

"Reducing the number of supported platforms can simplify operations, make the most of skills, reduce spending, and streamline data integration and governance overhead," Gartner said in a 2021 report.[3]

Apache Cassandra®, the open source, distributed NoSQL database, is **a logical choice due to its global scale and high availability**.

NoSQL (or non-relational) databases are optimized specifically for modern data applications that require large data volume, low latency, and flexible data models. This is achieved by relaxing some of the data consistency restrictions of other databases. Cassandra's appeal is broadening quickly, driven by a rush to harness today's "data deluge" with apps that are globally distributed and always-on.

"Companies across industries have standardized the way they manage their data with Cassandra—and staked their business on its reliability and scalability."

Cassandra was developed by Facebook in 2007 to store and access huge amounts of data for Messenger, which was growing very fast. Cassandra scaled quickly, and accessed huge amounts of data within strict SLAs—in a way that relational databases and SQL, which had long been the standard ways to access and manipulate data, couldn't. As it became clear that this technology was suitable for other use cases, Facebook handed Cassandra to the Apache Software Foundation, where it became an open source project (it was voted into a top-level project in 2010).

Since then, companies across industries have standardized the way they manage their data with Cassandra—and staked their business on its reliability and scalability.

Best Buy, the world's largest multichannel consumer electronics retailer, describes Cassandra as "flawless" in how it handles massive spikes in holiday purchasing traffic. Bloomberg News has relied on Cassandra since 2016 because it's easy to use, easy to scale, and always available; the financial news service serves 20 billion requests per day on nearly a petabyte of data (that's the rough equivalent of over 4,000 digital pictures a day—for every day of an average person's life).

Today, Cassandra **has matured** into the battle-tested, standard choice for organizations that demand reliability and scalability. But it isn't just for big, established sector leaders like Best Buy or Bloomberg. **Ankeri**, an Icelandic startup that operates a platform to help cargo shipping operators manage real-time vessel data, chose Cassandra—delivered through DataStax's database-as-a-service **Astra DB**—in part because of its ability to scale as the company gathers an increasing amount of data from disparate systems across a growing number of ships. It wanted a data platform that wouldn't make data a problem, and wouldn't get in the way of its success.

DataStax Astra DB: Knocking down obstacles to Cassandra

Standardizing on Cassandra (and using other databases, for that matter) can come with a host of operational, maintenance, provisioning, and cost challenges.

Cassandra's robustness and complexity—the very characteristics that have made it the choice for many of the most challenging database needs—have created development barriers and operational challenges, slowing its adoption as a more mainstream data solution for every all developers.

One of the biggest obstacles to using Cassandra has been operational overhead. Deploying the database can take weeks, and operating it can pose a host of challenges that fly in the face of standardization.

"Cassandra's robustness and complexity—the very characteristics that have made it the choice for many of the most challenging database needs—have created development barriers and operational challenges."

For one, individual teams across an organization can end up implementing the database on an ad hoc basis. As the deployment scales and multiplies across the organization, the need for support services increases. In many cases, organizations end up with a patchwork quilt of support and services from a variety of different sources: some in-house resources, the open source community, and third-party agencies. All of these come with varying levels of Cassandra expertise and response time.

On top of this, implementing and configuring Cassandra can require a significant learning curve. Most companies find out that it's difficult and costly to hire in-house expertise due to a limited supply of talent. Employees usually end up educating themselves on Cassandra, using a combination of open source documentation, help from the community, and trial and error. This slows down adoption and puts a significant administrative burden on IT.

One of the most challenging tasks when it comes to managing Cassandra (and other modern databases, for that matter) has been provisioning. With cloud computing services, scaling, capacity planning, and cost management are all automated, resulting in software that's easy to maintain, and cost effective—it's described as "serverless."

But because modern databases store data by partitioning it across nodes of a database cluster, they've proved challenging to make serverless. Doing so requires rebalancing data across nodes when more are added, in order to balance storage and computing capabilities.

Because of this, enterprises have been required to guess what their peak usage will be—and pay for that level, even if they aren't using that capacity.

DataStax Astra DB, a database-as-a-service (DBaaS) built on Cassandra, tackles the challenges posed by this powerful database head-on. Because it's a fully managed cloud DBaaS, architects and developers no longer have to grapple with challenges associated with operating and managing Cassandra. Deploying new applications on Cassandra takes minutes rather than weeks, and software patches and updates are no longer a burden carried by the enterprise. Additionally, Astra DB can be deployed in multiple regions on any cloud (AWS, Azure, and Google).

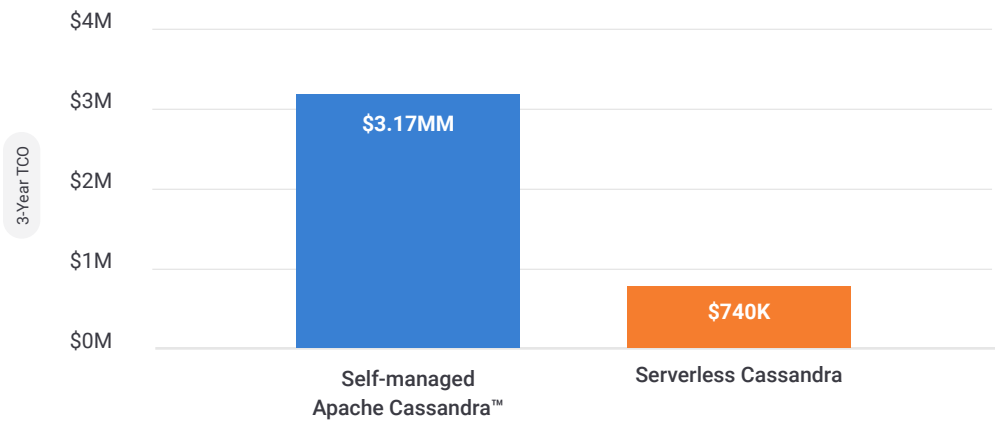
To standardize development on Cassandra even further, Astra DB also includes **Stargate**, which provides developers with an API gateway that eliminates the need to code endpoints to access data via JSON, GraphQL, gRPC, and REST approaches.

"We already use Cassandra for the reliability and resilience to handle big workloads with zero downtime. The promise of Astra DB is a true Database-as-a-Service with no ops. This will help developers work more efficiently so they can spend more time innovating."

Dipak Chandan, Venmo senior engineering manager

Perhaps the most significant recent advance is the **availability** of Astra DB as a serverless, pay-as-you-go service. According to **research** by analyst firm GigaOm, the serverless Astra DB can deliver significant cost savings over self-managed Cassandra—over 75% overall savings, in fact. And developers only pay for what they use, no matter how many database clusters they create and deploy.

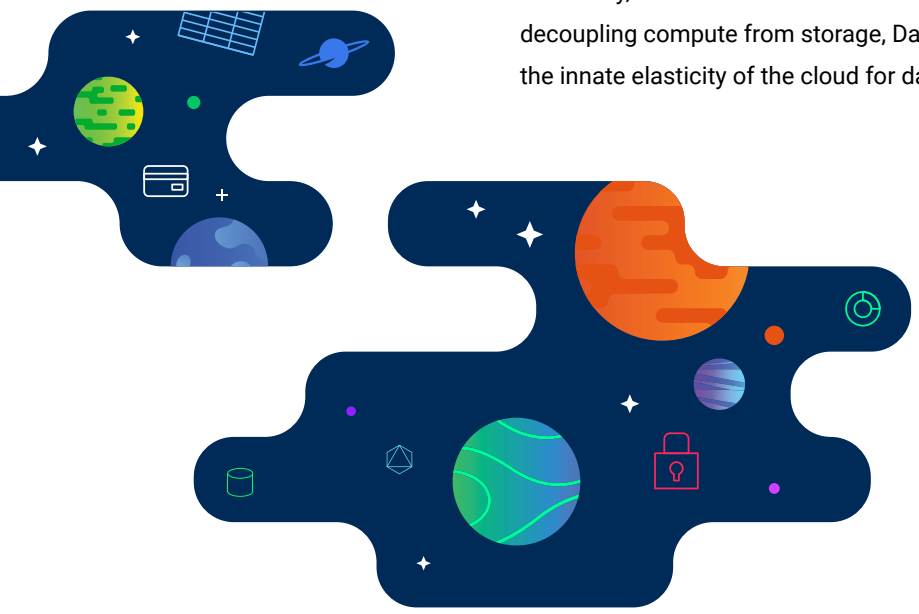
Serverless Cassandra Reduces TCO by 76%



*Based on findings fo the GigaOm: Cassandra Total Cost of Ownership Study

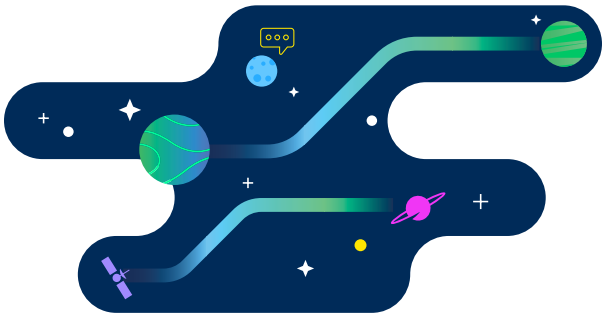
GIGAOM

Carl Olofson, research vice president at IDC, notes: “A core benefit of the cloud is dynamic scalability, but this has been more difficult to achieve for storage than with compute. By decoupling compute from storage, DataStax’s Astra DB service lets users take advantage of the innate elasticity of the cloud for data with a cloud agnostic database.”



DS

Streaming toward unified data



Standardizing on Cassandra is one key part of unifying an organization's data architecture. It's how organizations develop a clear view of all of their "data at rest"—the kind of information that's often captured in databases, data warehouses, and even data lakes, to be used immediately (in "real time") or to fuel applications and analysis later.

But "data in motion," which is driven by activities, actions, and events that happen in real-time across an organization, is becoming increasingly valuable. While it might eventually be captured in a database or other store, data is often worth more to a bank, for example, when it's still in flight, because it enables the detection and prevention of fraud in real time.

Streaming is a key part of an event-driven architecture, which is a software architecture or programming approach built around the capture, communication, processing, and persistence of events—mouse clicks and sensor outputs, for instance. It's also how events are passed along from one service to another within an architecture, creating a standard way for data to be connected and accessible across an organization.

Some enterprises have recognized that they need to derive value from their data in motion and have assembled their own event-driven architectures from a patchwork of legacy technologies that weren't built to meet the scale and performance needs of a real-time, event-driven architecture. These technologies might include message-oriented middleware systems like Java messaging service (JMS) or message queue (MQ) platforms.

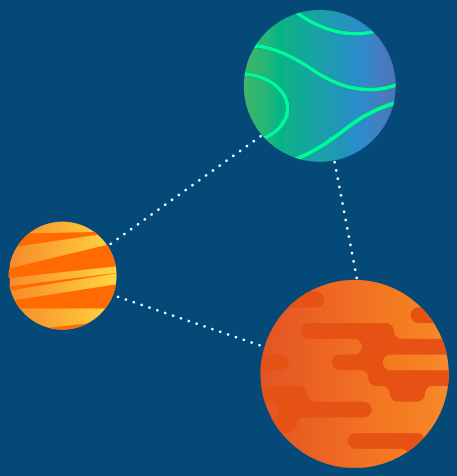
But, similar to how siloed data architectures can't keep up with the data demands of modern enterprises, MQ platforms have become outdated and expensive to maintain. They were built on the premise that the value of data they processed was transient, and should be immediately discarded one each message is delivered. This essentially throws away a highly valuable asset: data that's identifiable as arriving at a particular point in time. Time-series information is critical for applications that involve asynchronous analysis, like machine learning.

"Apache Pulsar is an advanced, open-source, unified streaming and message queuing platform for handling real-time data. It was built for the high throughput and scalability that many data-driven applications require."

A modern streaming system needs to not only pass events along from one service to another, but also store them in a way that retains their value or usage later. The system also needs to be able to scale to manage terabytes of data and millions of messages per second. The old MQ systems are not designed to do either of these.

Apache Pulsar™ is an advanced, open-source, unified streaming and message queuing platform for handling real-time data. It was built for the high throughput and scalability that many data-driven applications require. It's a key component of the ultimate way to standardize a data architecture: the open data stack.

Mobilizing real-time data with an open stack



So how do you bring this all together—data persistence, support for a wide range of standard APIs to connect your developers with data, and the capability to get the most out of data in motion–in a way that **supports real-time applications**?

This requires an open and integrated data stack that can handle all of an organization's real-time data, all in one open, standardized stack. It's the polar opposite of the data complexity and fragmentation that gets so many enterprises stuck in an innovation stalemate.

As mentioned above, Astra DB makes the infinitely scalable Cassandra easy to use, build on, and afford. This instantly available "data at rest" is critical to many use cases (customer profile, session information, etc.) as a system of record.

But the world operates in real time, and streaming data in motion captures changes on the fly. Pulsar enables real-time data to be acted upon as it's generated (like when FedEx sends a notification to the buyer that their package is being delivered).

It's a key reason DataStax chose Pulsar as the foundation for DataStax Astra Streaming. Only a stack that unifies both real-time data at rest and in motion can help solve for data complexity and deliver a new level of digital excellence for our customers.

Because it includes the open-source, data API layer Stargate, developers and their applications have a standard way to connect to Cassandra. This does away with the need to install and update drivers, reduces your team's learning curve, and completely eliminates the need to maintain your own abstraction layer.

Taken together, the DataStax Astra cloud service—with the world's most scalable database, cutting-edge streaming technology, and a set of powerful, easy-to-use APIs—delivers a unified stack that helps enterprises build high-scale, high-impact applications. In other words, it enables organizations to standardize the way they store, access, and build on the data that matters most.

Additional resources

- **Learn more about Astra DB**
- **Try Astra DB with up to 80 GB of free monthly usage**
- **Request a demo**

Endnotes

1 IDC, Establishing Data Logistics to Support DataOps and Better Data Value, Doc # US46402421, December 2021

2 IDC, Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All?, Doc # US46410421, March 2021

3 Gartner, Cost Optimization Is Crucial for Modern Data Management Programs, Ankush Jain, Nina Showell, Dec.31, 2021

ABOUT DATASTAX

DataStax is the real-time data company. With DataStax, any enterprise can mobilize real-time data and quickly build the smart, high-scale applications required to become a true data-driven business. The **DataStax Astra** cloud service uniquely delivers the power of Apache Cassandra™—the world's most scalable database—with the advanced Apache Pulsar™ streaming technology in a unified stack, available on any cloud. Hundreds of the world's leading enterprises, including Verizon, Audi, FedEx, ESL Gaming and many more rely on DataStax to unleash the power of real-time data to create the in-the-moment digital experiences that can win new markets and change industries. Learn more at **DataStax.com**.