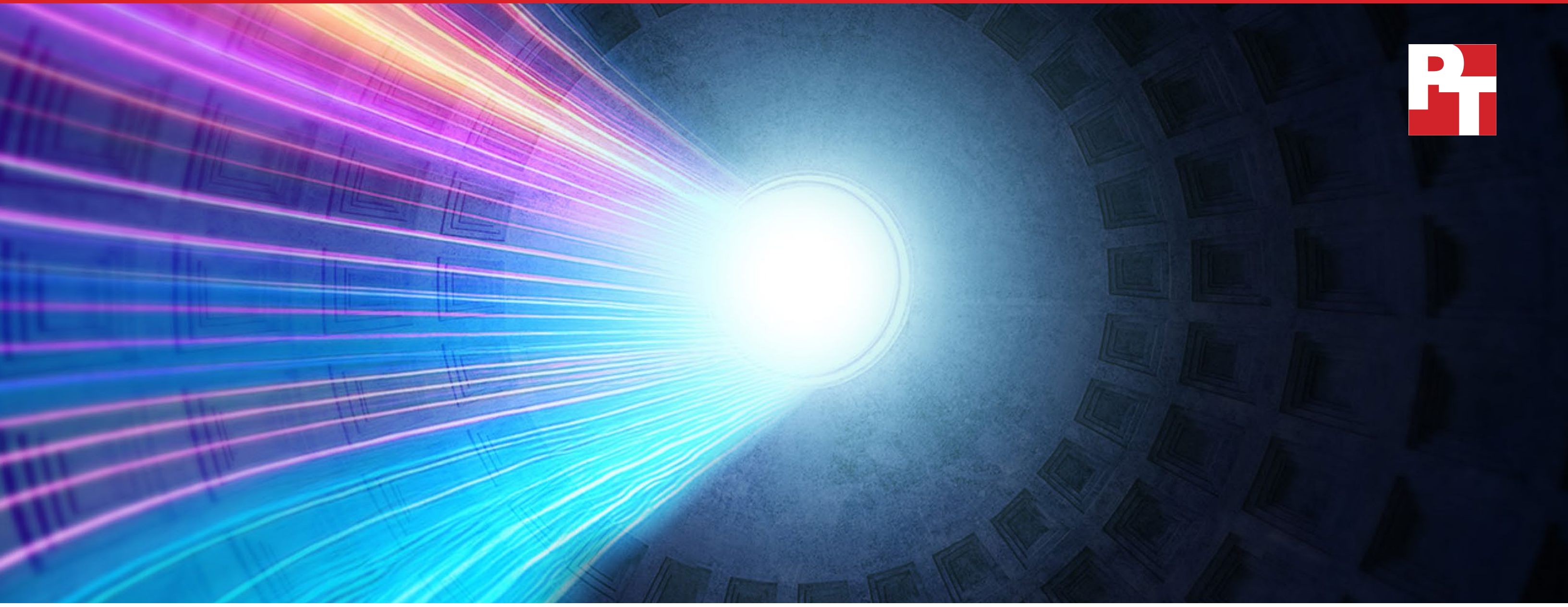


Faster query stream processing

Up to 21.1% less time to complete a stream of queries

Run concurrent query streams more quickly

Saved over 9 minutes while running four concurrent streams

Unlock faster insights with Azure Databricks

On decision support system (DSS) workloads, an Azure Databricks cluster outperformed a Databricks cluster on Amazon Web Services (AWS)

Databricks, with its unified lakehouse architecture, can process vast amounts of structured, semi-structured, and unstructured data. The open-source analytics platform offers distinct integrations with major cloud service providers (CSPs) to align with native services. While Databricks offerings across CSPs share many similarities, notable differences exist in areas such as performance and scalability.

Our analysis aimed to measure Databricks DSS workload performance of two Databricks software-as-a-service (SaaS) solutions: Azure Databricks, the only first-party Databricks service, and Databricks on AWS™, a third-party service. Azure Databricks, running in Microsoft Azure VMs, processed queries faster than a Databricks cluster running in AWS instances, completing lone and concurrent query streams in less time. Both environments used the same Databricks solution but differed in underlying cloud infrastructure, integration, and support models. Choosing the higher-performing Azure Databricks service can lead to faster decision making, improved operational efficiency, and more relevant customer experiences.

Databricks in data-driven organizations

Databricks can streamline the entire data lifecycle—spanning data engineering; extract, transform, and load (ETL) workloads; data science; machine learning (ML); artificial intelligence (AI); and business intelligence (BI)—by leveraging its unified lakehouse architecture as a central data repository that enables seamless transitions across these tasks.

For data-driven organizations, factors such as processing speed, system reliability, and the ability to handle high-volume data workloads help ensure that data pipelines run efficiently. These organizations must also consider how well their CSP integrates with Databricks and vice versa. Optimized storage solutions, high-performance computing resources, and low-latency networking all affect how well Databricks performs.

Why choose Azure Databricks?[1]

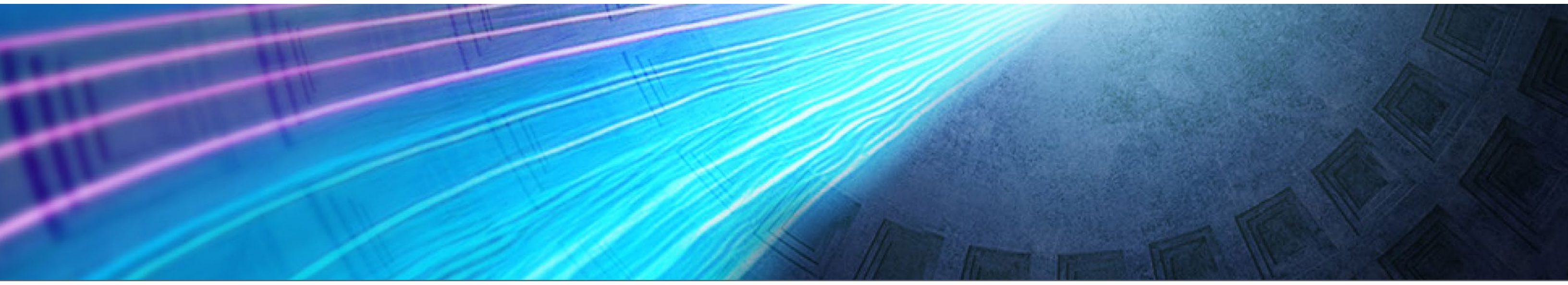
The all-in-one, open analytics platform Azure Databricks helps you build, deploy, share, and maintain critical, scalable, enterprise-grade solutions for data, analytics, AI, and more. The solution integrates the Databricks Data Intelligence Platform with your Azure storage and security, managing and deploying cloud infrastructure for you.

Azure is a first-party Databricks service, meaning that Microsoft and Databricks work together to deliver a unified, cloud-native data platform that integrates seamlessly with the Microsoft Intelligent Data Platform. By integrating the platform with the flexibility and scalability of Azure infrastructure, Azure Databricks helps organizations harness their data more effectively and unlock deeper insights across diverse applications through a single pane of glass.

Defining features of Azure Databricks:

- **Co-engineered:** Microsoft and Databricks have invested significant resources into co-developing solutions that offer tight integration and strong performance. The collaboration enables customers to operate a unified analytics platform that supports various workloads, including BI, ML, and AI.
- **One location, dual support teams:** Azure Databricks is available only in the Azure portal, which can simplify deployment and management of the service. Microsoft manages Azure Databricks and provides support coverage under Microsoft support contracts, which are subject to the same SLAs, security policies, and terms as other Azure services. Both Microsoft and Databricks teams resolve support tickets as needed, with Microsoft support functioning as the primary responder.
- **Centralized billing:** Customers pay a streamlined bill through Azure, gaining smooth cost management and transparency.
- **Seamless integration:** Azure Databricks slots in nicely with other key Azure services, such as Microsoft Entra ID for identity and access management, Azure Data Lake Storage optimizations for efficient and scalable data storage, and Azure Monitor with Log Analytics for comprehensive monitoring and diagnostics.

See this blog post to learn more about the potential benefits of Azure Databricks.



How we tested

We created instance clusters on Azure and AWS. For Azure, we used a Standard_E16ds_v5 instance for the driver and 20 Standard_E8ds_v5 instances as the workers with Azure Data Lake Storage Gen2. For AWS, we used an r6id.4xlarge instance as the driver and 20 r6id.2xlarge instances as the workers with AWS S3 storage. Both Databricks offerings used the Databricks Runtime 15.4 LTS engine (with Photon enabled) powered by Apache® Spark™ 3.5.0. We configured the solutions as comparably as possible in terms of resources.

We used a performance testing framework for Spark SQL in Apache Spark 2.2+ called spark-sql-perf. The framework is a series of Databricks-hosted test scripts and notebooks on GitHub and is based on industry-standard TPC-DS-derived workloads. We executed the TPC-DS-3.2-like benchmark to test the decision support capabilities of each Databricks product using the tpcds_datagen notebook.

We tested configurations with and without AutoScale, which automatically allocates cluster resources based on workload volume. We anticipate many organizations enable AutoScale, but cluster resizing tasks can briefly affect performance, so we tested without AutoScale for those organizations that must maintain consistent performance or don't need to plan for spikes in usage. In the Databricks solution of each CSP, we configured all-purpose compute clusters. We used a 10TB database for all testing.

We ran power and throughput tests from the testing framework. In power tests, a single query session runs a single stream of queries. Throughput tests process multiple concurrent query sessions, each running a single stream of queries. We chose four users for our throughput tests. In both power and throughput tests, a query stream consisted of 99 queries. To learn more, see the science behind the report.

Many cloud platforms offer Databricks, but we focused this study on just two of the largest cloud platforms, Azure and AWS.

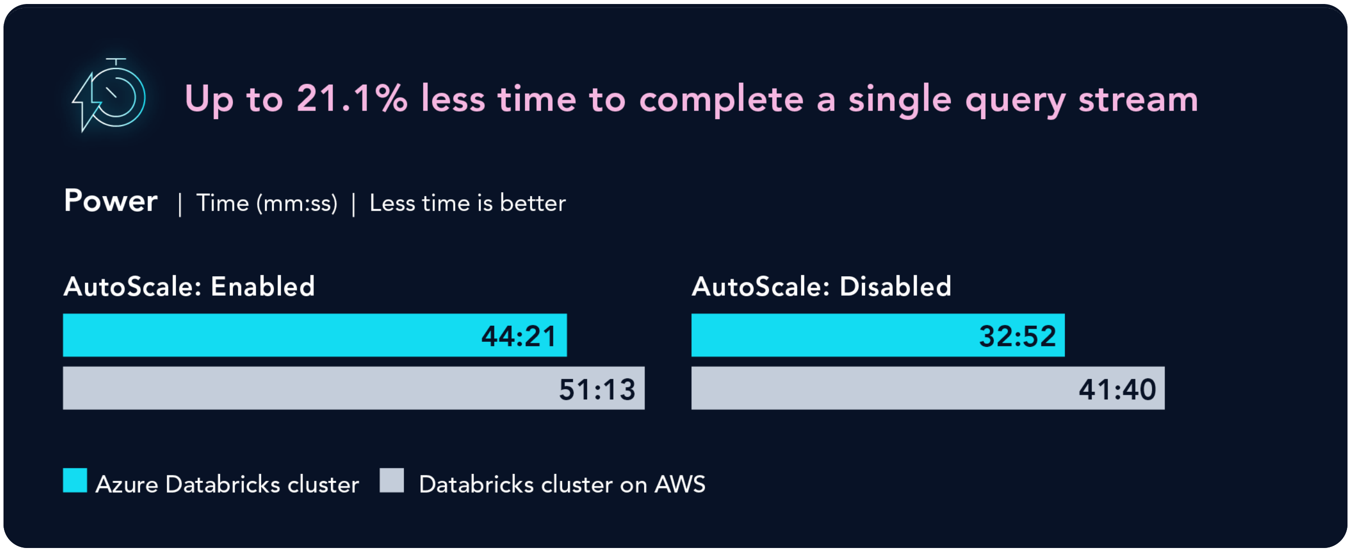
Note: The graphs in this report use different scales to keep a consistent size.
Please be mindful of each graph's data range as you compare.

Get critical business insights sooner with Azure Databricks

Single query session (power test)

This test measured how well each solution can optimize individual query execution without interference from concurrent workloads. Faster execution in this test demonstrates the kind of experience a single user could have. For example, a BI analyst could save time when running multiple detailed reports sequentially because they have tasked the system to handle heavy analytical queries one by one without competing queries from other users.

As Figure 1 shows, the Azure Databricks cluster completed the power test in 13.4 and 21.1 percent less time than the Databricks cluster on AWS. Comparing the configurations with AutoScale disabled, the Azure solution saved nearly 9 minutes.



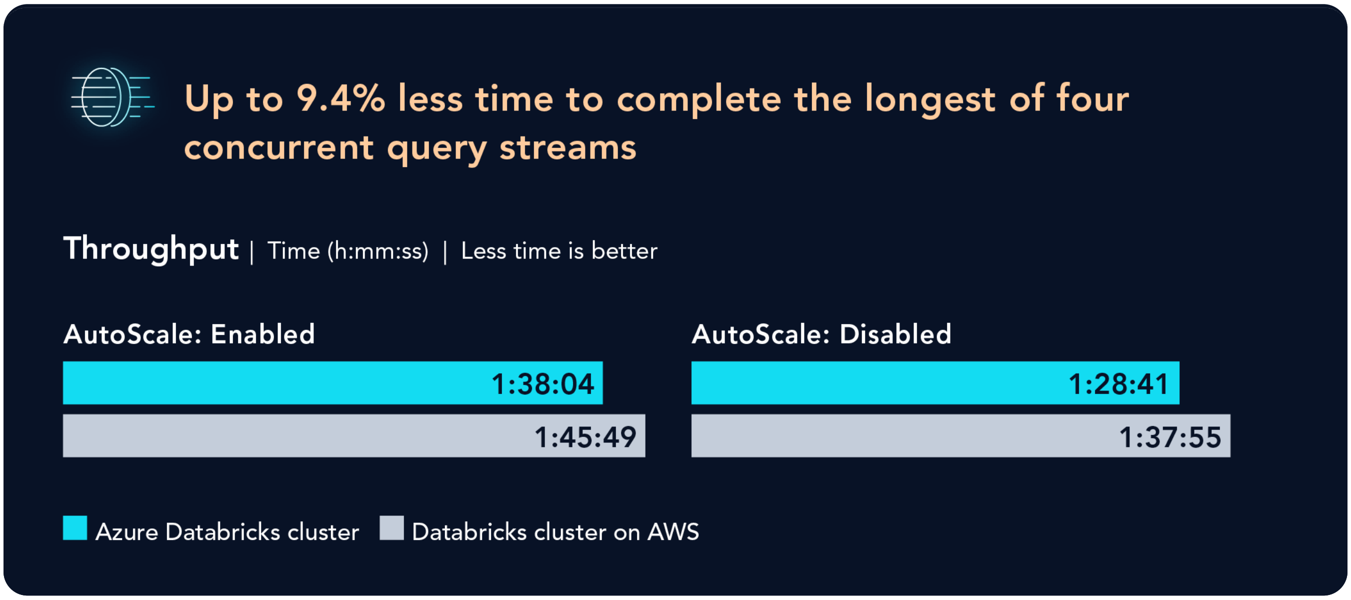
Up to 21.1% less time to complete a single query stream

Figure 1: Time for both solutions in AutoScale enabled and disabled configurations to complete a single query stream. Source: PT.

Four concurrent query sessions (throughput test)

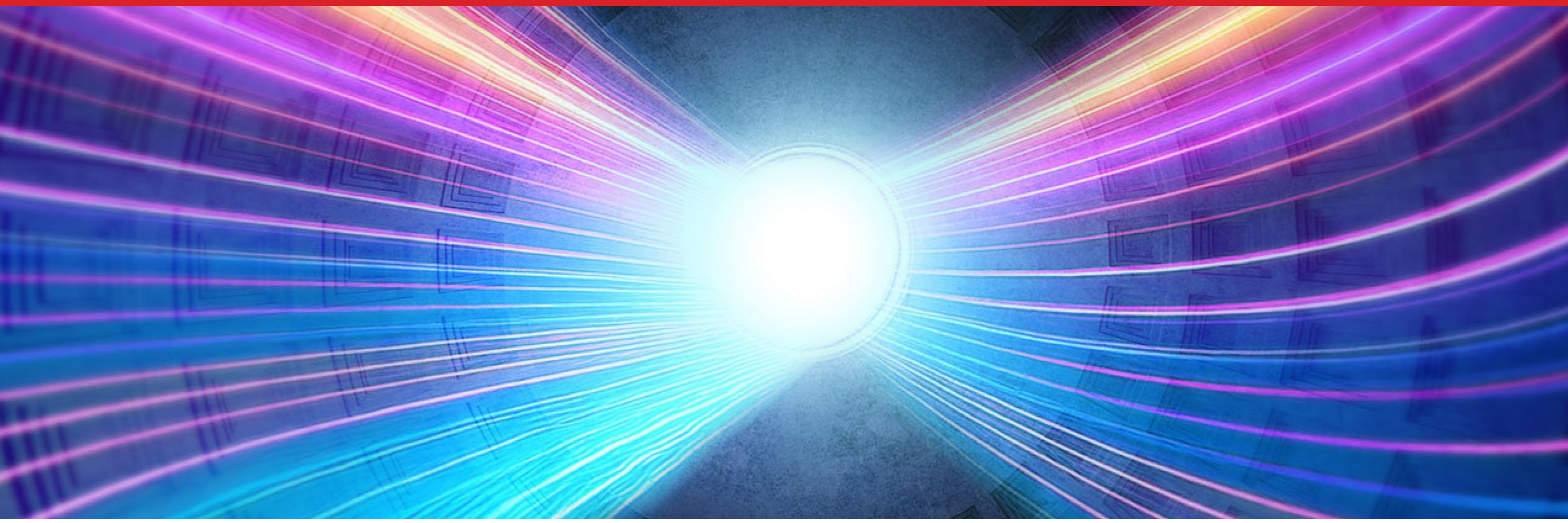
Unlike the power test that runs queries sequentially from a single stream, the throughput test executes concurrent queries from multiple streams. This test could demonstrate the experience a user has while running analysis at the same time as others. For example, an analyst from one department could save time when running reports or dashboards simultaneously with analysts from other departments, sharing cluster resources.

As Figure 2 shows, the Azure Databricks cluster completed the throughput test in 7.3 and 9.4 percent less time than the Databricks cluster on AWS. Comparing the configurations with AutoScale disabled, the Azure solution saved 9 minutes and 14 seconds.



Up to 9.4% less time to complete the longest of four concurrent query streams

Figure 2: Time for both solutions in AutoScale enabled and disabled configurations to complete the longest of four concurrent query streams. Source: PT.



Conclusion

In today's data-driven landscape, organizations rely on robust analytics platforms to transform vast and varied data into actionable insights quickly and reliably. Databricks, with its unified lakehouse architecture, plays a critical role in enabling seamless data workflows, such as data engineering. When integrated with cloud infrastructure, Databricks helps ensure scalable, secure, and efficient access to computing resources.

However, not all Databricks cloud solutions are equal in terms of performance. Our analysis showed that an Azure Databricks cluster processed queries more rapidly than a Databricks cluster on AWS, completing both single query stream and concurrent query stream workloads in less time. These time savings can empower users to accelerate their analytical workloads, improve operational efficiency, and make better-informed decisions faster. For organizations seeking a high-performance integrated analytics solution, Azure Databricks offers compelling advantages.

[1] Jason Pereira and Lindsay Allen, "Azure Databricks: Differentiated synergy," accessed May 28, 2025, https://azure.microsoft.com/en-us/blog/azure-databricks-differentiated-synergy/.

Read the science behind this report at https://facts.pt/KIPWSB9 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Microsoft.