

Snowflake | Databricks AI Benchmark

Metrics and Findings

Hitachi Solutions Empower Team Study



The Empower Data [Platform](#) is a comprehensive suite of products providing an enterprise-class data platform on Azure Data Services. Developed by Hitachi Solutions, the ecosystem includes a Delta Lake, automated data acquisition, data governance tools, industry data models, machine learning, and deep integration with Dynamics 365 to make insights actionable.

Empower is a subscription-based offering which manages all the repetitious elements of building a data platform, eliminating up-front cost, and allowing our customers and internal teams to focus on adding value – not managing processes that the robots can handle.

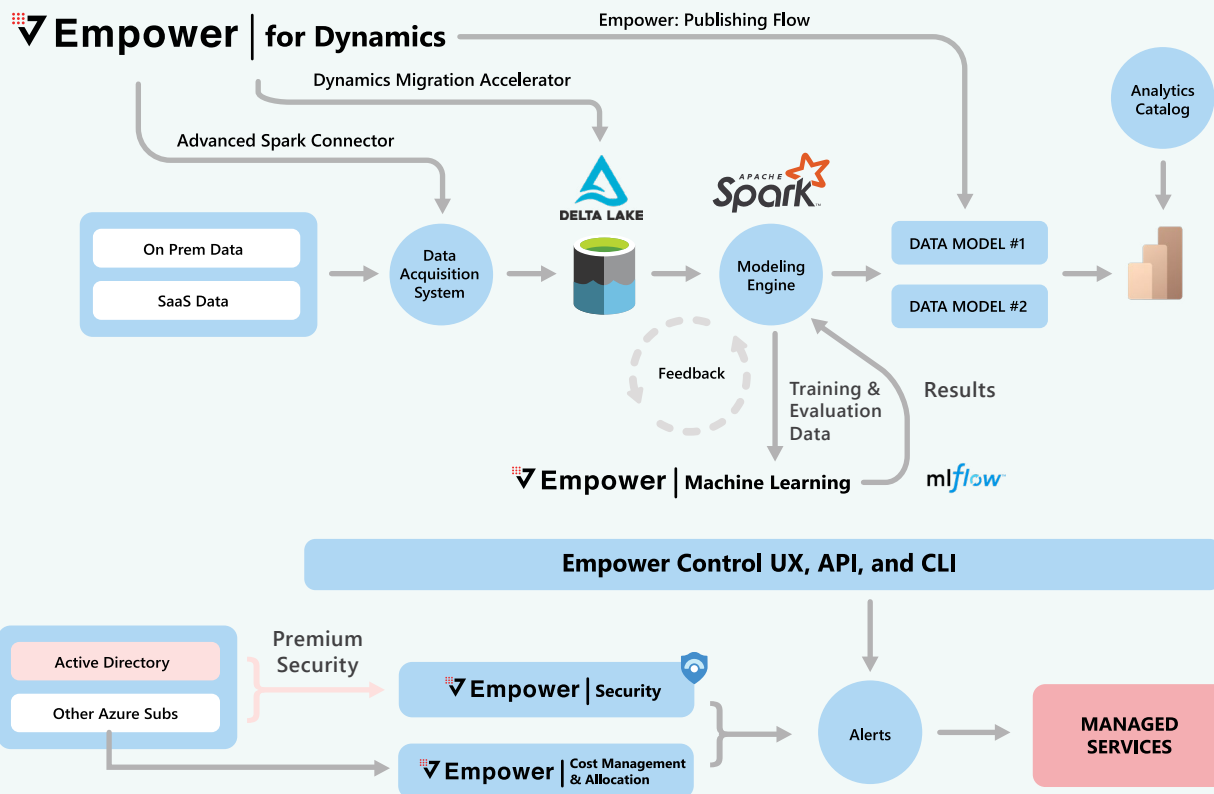
As an industry platform solution, Empower is geared to many different verticals including manufacturing, healthcare, CPG, retail, construction, and financial services. Many organizations are using Empower to manage massive data workloads for analysis to help them make faster, more informed business decisions.

- An entire Empower data solution can be stood up and deployed in a matter of days, reducing the time it takes to get actionable awareness to propel decision making.
- Hitachi Solutions' Empower excels in the orchestration and optimization of **Microsoft Azure** data analytics solutions. As such, we are award-winning [designated Microsoft Partner](#) for Azure Data and AI.
- With Empower, the infrastructure is managed in the customer tenant, and is augmented with incident response, **security monitoring**, and private networking support.

At its core, Empower's analytics and artificial intelligence models are driven by [Delta Lake](#), an optimized layer for storing data and tables in a lakehouse architecture. Delta Lake is driven with [Spark](#) APIs for tight integration and structured streaming in a scalable fashion. Delta Lake and Spark are key components of the [Databricks](#) analytics platform.

As part of our ongoing analytics research and development work at Hitachi Solutions, we recently conducted a benchmark study to compare how Databricks and [Snowflake](#) facilitate artificial intelligence workflows and determine which product performed better in structured testing.

Empower | Ecosystem



The Benchmark

Rooted in data science, a benchmark is a set of conditions against which a product or system is measured using predefined metrics for performance and cost. Hitachi Solutions used [TPCx-AI](#), an end-to-end AI benchmark standard developed by the Transaction Processing Performance Council (TPC). The TPC's standards are used to create benchmarks that can result in higher performing, efficient systems at a lower cost.

Our testing used TPCx-AI use cases that are relevant in current production datacenters and cloud environments. We focused our study on four out of 10 use cases; these use cases were selected for their applicability in common ML scenarios in retail, as well as data type (structured versus unstructured), and dataset size.

Selected Use cases

Audio Transcription of Customer Logs (TPCx-AI Use Case 2)

This natural language processing use case is designed to emulate translating customer audio conversations to text. The core of the model is a recurrent neural network (RNN) trained to ingest speech spectrograms and generate English text transcriptions.

Next-Purchase Prediction (TPCx-AI Use Case 8)

This classical machine learning use case is designed to emulate training a classifier on a data set of shopping transactions, labeled with trip types, to predict a future shopping trip type. The underlying model for this use case is the XGBoost algorithm.

Price Forecasting (TPCx-AI Use Case 5)

This natural language processing use case emulates the data science pipeline to suggest or predict the price of a retail item based on its product description. The underlying model is an RNN.

Customer Identification (TPCx-AI Use Case 9)

This computer vision use case is designed to emulate an end-to-end facial-recognition model. Customer images are aligned based on facial features to all have the same positioning and orientation. The aligned images are used to create an embedding as input to a logistic regression model to recognize the customer.transcriptions.

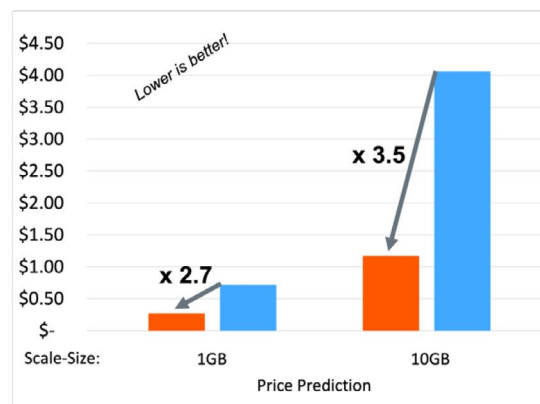
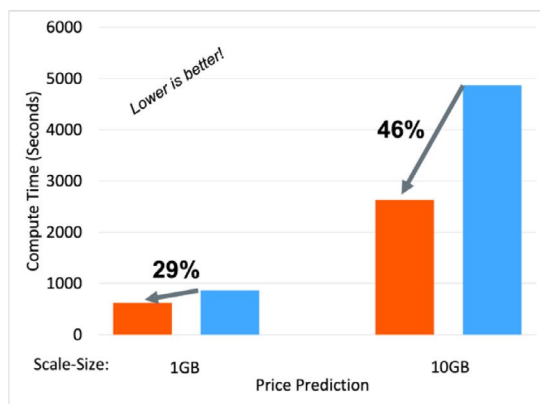
The table below details the number of datapoints generated for different dataset scales, categorized by use case.

Number of datapoints tested per use case (1, 10 and 100 scale dataset sizes)			
Use Case	Scale-1 Dataset	Scale-10 Dataset	Scale-100 Dataset
Audio transcription (Customer Audio Files)	387	1,964	11,764
Natural Language Processing (Product Descriptions)	70,710	358,817	2,152,033
Classical Machine Learning (Shopping Trips)	3,676,784	27,986,550	223,801,916
Computer Vision (Customer Images)	70	1291	46,268

The Results

We ran the benchmark on two different infrastructure configurations: a single-node system to mimic initial R&D for an AI/ML pipeline, and a multi-node cluster to simulate a full production deployment for training and inference. Below are the results for each.

Single Node



The previous graphics illustrate the speed and cost of single-node Snowflake and Databricks solutions for the price prediction use case.

Both Snowflake and Databricks natively handle training AI in single node configurations. Databricks does so using a single-node Spark cluster, while Snowflake uses UDFs/Stored Procedures via Snowpark. On the audio transcription and computer vision use cases we tested, Snowflake is unable to complete the benchmarks on any size dataset due to missing libraries. Snowflake does not allow the user to install either unsupported Conda libraries like the popular OpenCV for computer vision, or Linux packages like `libsndfile`. Snowflake is unable to complete the shopping trips CML use case due to its stored procedure memory limit.

Lastly, in the price prediction use case depicted above, on the 10 GB dataset size, Snowflake stored procedures trigger a timeout exception after 1 hour. The statistics for Snowflake were calculated by training the model on one third of the 10GB dataset and multiplying by three to extrapolate a time estimate. For this use case, Databricks is both faster and cheaper than Snowflake by a significant margin. Although the scale of time and cost are relatively small, they are significant when running hundreds of experiments to test different data representations and model architectures.

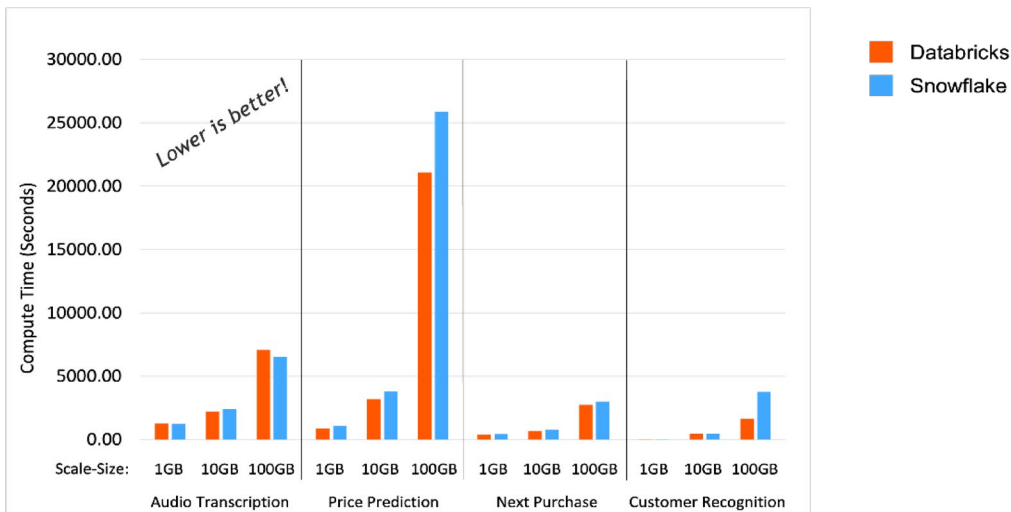
Note: Unlike Databricks, Snowflake does not support GPU training within its SQL engine. On Databricks, you can expect faster performance when using accelerated compute on large datasets and with more complex model architectures.

Multi-Node

Snowflake is unable to support true distributed, multi-node artificial intelligence training.

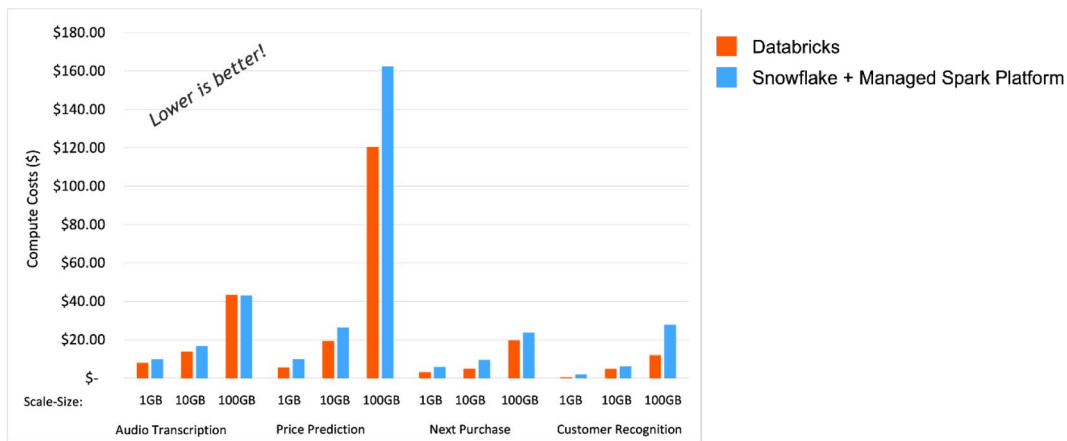
This is because *AI models must be trained iteratively*, and nodes must communicate with each other in multi-node to share gradients in a multi-node configuration. UDFs and stored procedures do not support inter-node communication for gradient sharing, and therefore Snowflake cannot do multi-node distributed training within its data warehouses.

Because of this limitation, we had to set up a third-party compute cluster to train the AI models, using Snowflake only as a data processing and storage solution for all multi-node scenarios. Meanwhile, these benchmarks were easily achievable on Databricks by configuring a multi-node Spark cluster, using Spark dataframes, and importing Horovod versions of the machine learning libraries required for each use case.



Above are the performance results of the two platforms across all use cases and dataset scales.

In most scenarios, Databricks is equivalent or faster to Snowflake while keeping the data all within one platform. Usually, the time differences can be explained by the additional ingress/egress times required when sending data from Snowflake to the secondary compute environment outside of Snowflake to perform multi-node training, as training typically uses an outsized amount of the total compute time.



The above diagram depicts the cost results of the two platforms across all use cases and dataset scales.

As discussed, we needed to leverage an additional compute environment for Snowflake’s multi-node scenarios. The Databricks columns contain a single cost metric (Azure VM spend + DBUs). Snowflake columns contain the costs of Snowflake warehouses, a Spark cluster on Azure (both included in the light blue bars), and license costs for using a managed Spark platform. Additionally, expect ingress/egress fees for moving data outside of Snowflake warehouses into the third-party compute cluster, which are not factored into the costs as displayed.


If Snowflake is used along with a managed Spark service (like Databricks, Dataiku, DataRobot, Sagemaker, and others), **it will be more expensive than Databricks alone**, which can easily handle multi-node workflows. The technical debt introduced by self-managed Spark environments can materially drive-up infrastructure, deployment, and management costs even while negating some of third-party costs illustrated above. We recommend our clients use managed Spark clusters like Databricks for data science workflows rather than trying to self-manage, as we believe the simplicity of managed Spark far outweighs the costs.

Our assessment of Snowflake and Snowpark

Strengths	Weaknesses	Benchmark Insights
Clusters are faster with default settings	Separate storage, development, and execution environments	First-time setup is easy
Clusters will auto suspend and auto scale quickly to save costs	No built-in feature store	Great platform for data analysis and data exploration
Extensive tutorials; light learning curve	No support for open-source Spark	Fast clusters for data manipulations
Fully self-managed platform: little configuration required	No file streaming	Poor environment control means some data science workflows are not possible
API allows users to manipulate SQL with their own language	No built-in Python notebooks	Poor scaling for large datasets or large ML models
Local development environment support	No distributed training support	No support for distributed data handling
Highly scalable for data transformations	Non-collaborative environment	Need to use external clusters for multi-node workflows
	No native GPU support	
	No native support for MLflow	
	Poor support for popular packages and libraries	

Our assessment of Databricks

Strengths	Weaknesses	Benchmark Insights
A unified store, development, and execution workspace	Default cluster settings are slower than Snowflake	All-in-one platform for data engineering, data science, data analysis, and machine learning
Intuitive feature store	Clusters always charged while cluster is running	Dedicated ML persona and runtime for quick environment setup
Full power of Spark	Steep learning curve for all Databricks functionality	Fast and cheap for training small models or training on small datasets
Structured streaming support	So many configuration options can make optimal environment difficult to setup	Easily scales to large models and/or large datasets
Familiar data science workspace	Poor support for cluster auto-scaling; spin-up can feel slow	Data storage, pipeline execution, and code development all in one place
Multi-node support	Limited library support for serverless inferencing	
Delta lake		
Collaboration		
GPU and deep learning training support		
Native model store and experiment tracking		



FINAL THOUGHTS

Hitachi Solutions is a supporter and proponent of the lakehouse and Delta Lake architecture, where data is open source and can be transparently accessed regardless of the compute platform. This open-source philosophy delivers benefits now and into the future as a protection against proprietary lock-in.

Benchmark results such as this solidify our choice of Databricks as the most effective data platform for managing our Empower solution's data management and data science workflows. Overall, it's more cost effective, easier to deploy and scales more effectively than Snowflake.

As a trusted Microsoft Data and AI partner, we continue to test and evaluate new and emerging technologies in database architecture and analysis. This dedicated attention to evolving technologies ensures that Empower will continue to be a best-in-breed product that provides our customers with cost-effective data-driven solutions for optimal and actionable insights.


Working with Hitachi Solutions


We have been implementing data and business system modernization solutions for nearly two decades. Our [data and analytics practice](#) is expertly skilled at designing innovative AI and machine learning solutions that help organizations securely capture, access, and extract knowledge and insights from their data. We give them the visibility and analytics they need to be more efficient, agile, and responsive, and to create elevated, more meaningful insights.

Our deep industry knowledge and experience and Microsoft expertise, combined with our depth of solutions and services portfolio, is how we support our customers and help them gain the most value from all their digital transformation investments.

[Contact Us](#)
to get started!

Ready to chat about how [Empower](#) can extend the value of your Microsoft investments and help your organization work smarter and faster?

Email Us  NA.Marketing@hitachisolutions.com

Visit Us  www.global.hitachi-solutions.com

Follow Us   