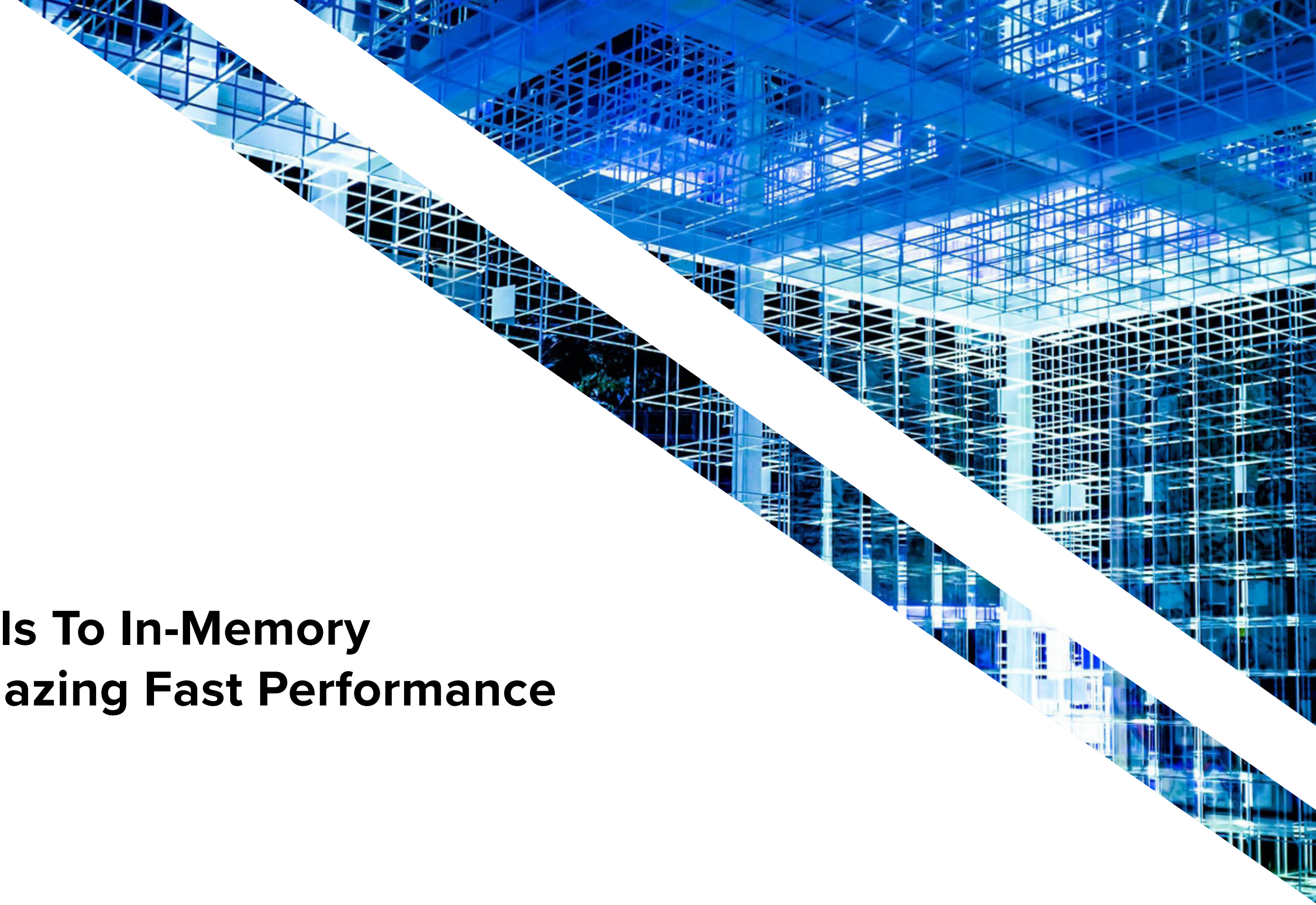
FORRESTER®



Deploy ML Models To In-Memory Databases For Blazing Fast Performance

Get started ⟶

FORRESTER OPPORTUNITY SNAPSHOT: A CUSTOM STUDY COMMISSIONED BY REDIS LABS | JUNE 2021

Real-Time, In-Memory Database ML Is Foundational For Intelligent Digital Transformations

No longer being a nice-to-have feature, AI is a core capability for enterprise digital transformations. To implement AI, firms must train machine learning (ML) models on key business data and then use those models to conduct inference in production applications.

The problem? Deploying ML models in production based on real-time data is challenging because too often such models are offline and slow, and the data feeding them is inaccurate. Additionally, model efficacy can degrade over time due to model and data drift necessitating that features are fresh, and models are retrained on new data inputs.

The solution? Deploying ML models to in-memory databases, where both the transactional and reference data resides, can dramatically reduce latency of both model training and inferencing.

Key Findings

Decision-makers are going all in with ML to create AI apps, but critical hurdles keep them from their desired transformation. Over 40% agree their architecture is not good enough for the future.

The high demand for real-time model inferencing (using ML models in production) exposes major challenges with accuracy, latency, and reliability in current architectures.

Running ML model inferencing in-database where data is stored solves some critical challenges.

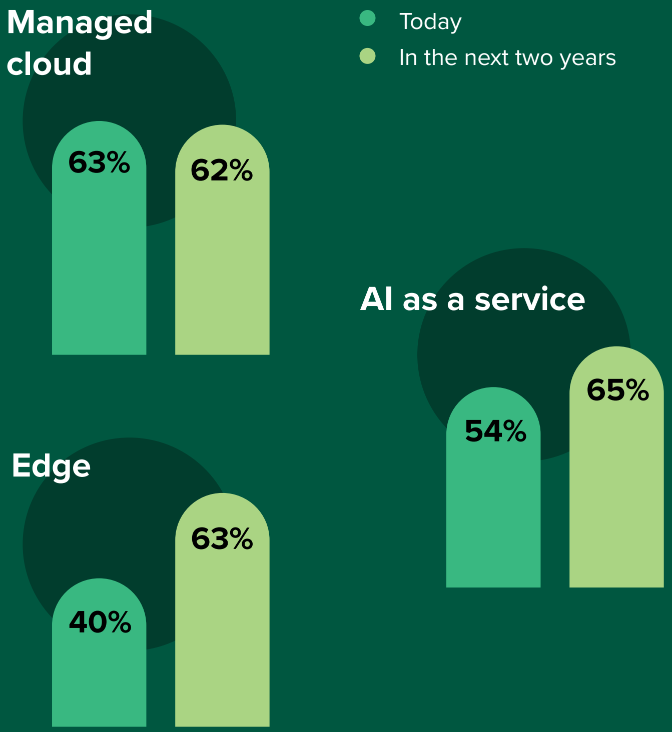
Overview
Current State
Challenges
Opportunity
Conclusion

Cloud Is Powering AI Inferencing; Edge Is Gaining Traction

The use of AI/ML is growing rapidly to keep up with changes in the digital landscape. In fact, 88% of AI/ML decision-makers expect their AI/ML use cases to increase in the next one to two years. Today, model inference mostly happens on managed clouds (63%); however, a shift is coming. In the next two years, AI/ML decision-makers expect to increase their usage of edge and AI as a service (AIaaS) to run their models, building on their existing plans for managed cloud.

Inferencing puts these ML models to work for smarter edge applications that require low latency. As emerging ML techniques like reinforcement learning and federated learning proliferate, edge application intelligence will blossom to accelerate digital transformation, especially in industries that must bridge the physical and digital worlds in real time.[1]

"Where is your AI model inference/production running today, and where do you expect it to be running in the next two years?"



Base: 106 IT manager+ decision-makers in North America responsible for ML/AI operations strategy
Source: A commissioned study conducted by Forrester Consulting on behalf of Redis, December 2020

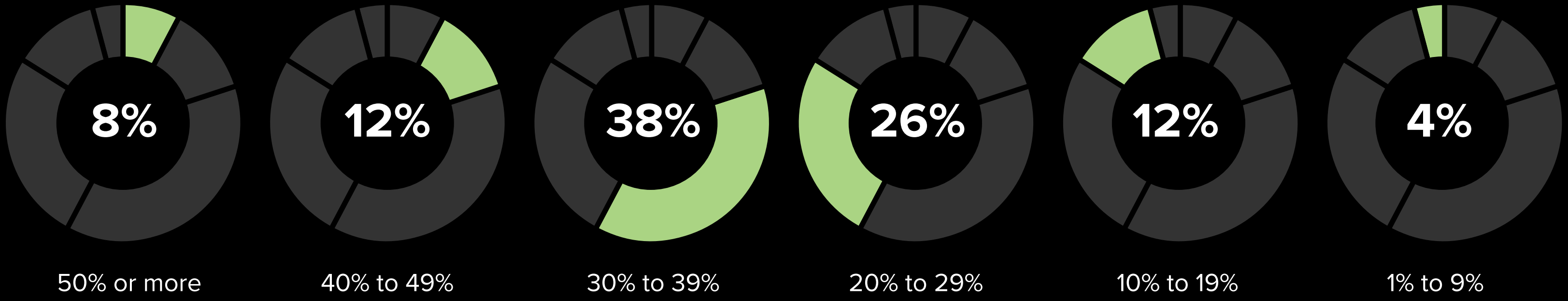
Overview
Current State
Challenges
Opportunity
Conclusion

Real-Time Demands Require Accuracy And Scalability

Today, companies are developing more models based on real-time data. Most decision-makers (64%) say their firms are developing between 20% to 39% of their models on real-time data coming from data streams and connected devices. As teams develop more models on real-time data, the need for accuracy and scalability is becoming increasingly critical. Real-time performance will help prevent models from slowing down applications as they hop to a service and/or microservice for an application to use the model.

38% of leaders say their firms are developing roughly a third of models on the real-time spectrum.

“What percentage of the models you are developing are on the real-time spectrum?”



Base: 106 IT manager+ decision-makers in North America responsible for ML/AI operations strategy
Source: A commissioned study conducted by Forrester Consulting on behalf of Redis, December 2020

Overview
Current State
Challenges
Opportunity
Conclusion

Databases Fail To Meet Reliability And Performance Demands

With the rise in continuous training, companies need accurate, reliable databases for these models. Most decision-makers (94%) claim to be confident that they can deploy a model in production, but firms are settling for AI/ML databases that cannot meet their demands. Nearly half of decision-makers cite reliability (48%) and performance (44%) as their top challenges for getting models deployed with their current databases.

This false confidence is particularly alarming since nearly half (41%) think their databases cannot meet data security and compliance requirements. How can companies be willing to put so much data at risk?

“What are the biggest challenges for your database when it comes to AI/ML?”

48%
Reliability

44%
Performance

41%
Inability to meet data security and compliance requirements

40%
Lack of well-curated data to train an AI system

36%
Lack of integration into the AI ecosystem

25%
Scale

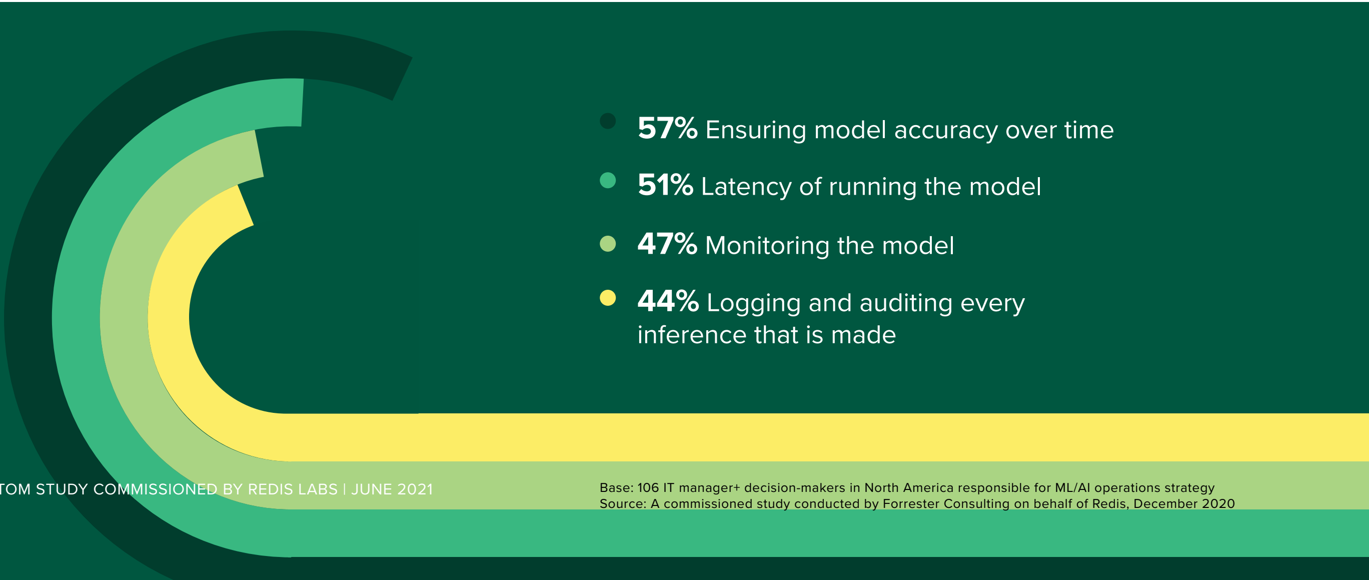
Base: 106 IT manager+ decision-makers in North America responsible for ML/AI operations strategy
Source: A commissioned study conducted by Forrester Consulting on behalf of Redis, December 2020

Latency Challenges Slow Model Inferencing

Companies are also plagued with model inference challenges that slow their AI/ML efforts. Ensuring model accuracy over time (57%) and struggling with the latency of running the model (51%) top the list of inference challenges. As companies move to scale their models and deploy in real time, they must address high latency and low accuracy. As the data representing business and customer operations changes over time, models may become less effective — or worse, they could start to have a negative impact on the business. Training and inferencing ML models in in-memory databases can alleviate this issue by enabling more frequent retraining of models.

Over 40% admit their current data layers/architecture are not good enough for the future.

"What are the biggest inference challenges for your organization?"
(Top 4 shown)



- ● **57%** Ensuring model accuracy over time
- ● **51%** Latency of running the model
- ● **47%** Monitoring the model
- ● **44%** Logging and auditing every inference that is made

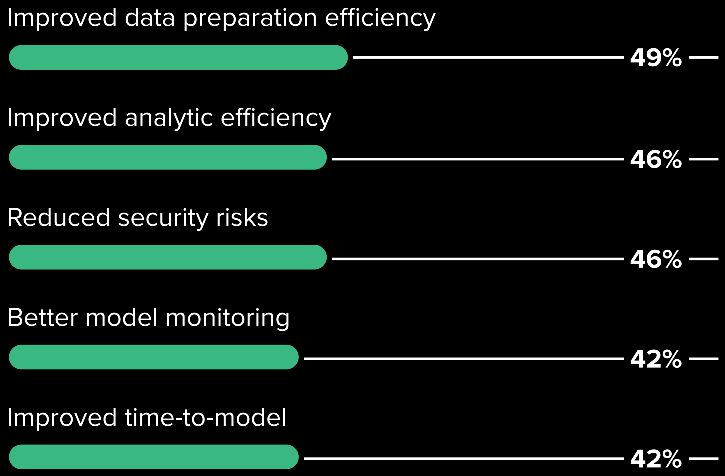
Base: 106 IT manager+ decision-makers in North America responsible for ML/AI operations strategy
Source: A commissioned study conducted by Forrester Consulting on behalf of Redis, December 2020

Overview
Current State
Challenges
Opportunity
Conclusion

Deploy Models Where Data Is Stored And Transacted To Solve Critical Issues

To accomplish the growth in AI/ML that companies have planned for the future, decision-makers agree that locating models in an in-memory database would solve key hurdles currently standing in their way. This tactic would allow firms to prepare data more efficiently (49%), improve analytics efficiency (46%), and keep data safer (46%). Because these are the critical elements that decision-makers cited as challenges, companies should move models into in-memory databases for online predictions to accelerate their AI/ML adoption. Creating more efficient and accurate models helps firms meet business objectives to better serve customers, transform products, empower employees, and optimize operations.

“What benefits have you experienced/would you expect if you could run models where the data is stored?”



Base: 106 IT manager+ decision-makers in North America responsible for ML/AI operations strategy
Source: A commissioned study conducted by Forrester Consulting on behalf of Redis, December 2020

Conclusion

AI powered by ML models mustn't slow down applications by necessitating a network hop to a service and/or microservice for an application to use an ML model and/or get reference data. Most applications, especially transactional applications, can't afford those precious milliseconds while meeting service-level agreements (SLAs). This is a key challenge facing organizations that wish to deploy ML models to either make existing applications smarter or new AI applications altogether. The challenge is exacerbated even further when organizations have globally distributed teams deploying multiple models. Technology leaders can deploy ML models to in-memory databases that support common ML model formats to dramatically reduce variability in the model-building process. Based on the principle of data locality, bringing the model inferencing capability closer to the data reduces latency, which further enables AI/ML to be real-time.

Project Director:

Sarah Brinks,

Senior Market Impact Consultant

Contributing Research:

Forrester's Application Development & Delivery research group

Overview
Current State
Challenges
Opportunity
Conclusion

Methodology

This Opportunity Snapshot was commissioned by Redis Labs. To create this profile, Forrester Consulting surveyed 106 IT decision-makers in North America with insights into AI/ML. The custom survey began and was completed in December 2020.

ENDNOTES

[1] Source: "Predictions 2021: Edge Computing," Forrester Research, Inc., October 26, 2020.

ABOUT FORRESTER CONSULTING

Forrester Consulting provides independent and objective research-based consulting to help leaders succeed in their organizations. Ranging in scope from a short strategy session to custom projects, Forrester's Consulting services connect you directly with research analysts who apply expert insight to your specific business challenges. For more information, visit forrester.com/consulting.

Demographics

COUNTRY

- US: 82%
- Canada: 18%

RESPONDENT ROLE

- C-level executive: 29%
- Vice president: 15%
- Director: 42%
- Manager 13%

TOP 4 INDUSTRIES

- Tech/tech services: 19%
- Manufacturing/materials: 18%
- FinServ/insurance: 18%
- Transportation: 17%

NUMBER OF EMPLOYEES

- 20,000+: 11%
- 5,000 to 19,999: 15%
- 1,000 to 4,999: 43%
- 100 to 999: 30%

Note: Percentages may not total 100 because of rounding.

FORRESTER®