E-BOOK

A little more talk, a little more action

Build a data infrastructure for conversational AI

NetApp

Contents

Tired of small talk? Think bigger.

NLP: Also known as natural language processing. Also known as conversational AI. Also known as *talking robots.*

Whatever you call it, a conversational AI system talks like a human, understands context, and offers intelligent responses—all thanks to massive improvements in deep learning, which makes AI systems more natural and less transactional.

Not only does deep learning make AI more user friendly, it also eliminates the need for deep human knowledge of linguistics and rule-based techniques on the back end. Deep learning opens the door to industries that use more complex, specific language (such as financial services, healthcare and life sciences, government, automotive and manufacturing, and retail) to adopt NLP solutions.

Data is the key to a better conversation

These AI models can be massive and highly complex. They require lots of data moving around at the speed of thought. To be successful, an NLP infrastructure must be able to:

1. Turn up the volume
2. Clear the pipes
3. Respond in a heartbeat

NLP: Not just for chatbots anymore

From smart assistants to search engines and predictive text, NLP is the new global language. It's all around us—sometimes in the places you least expect it.

Creditworthiness assessment

NLP can be used to generate credit scores based on data like geolocation, social media activity, browsing behavior, peer networks, and more.

Clinical trial matching

Getting patient participation for clinical trials can be difficult, mostly because people don’t know that trials are available. With NLP, researchers and manufacturers can automatically match patients to clinical trials.

Law enforcement

Police departments use NLP to identify motive for crimes so they can keep people safe and reduce violence while making policing more understanding and responsive.

Vehicle maintenance

NLP makes it easier for drivers to keep their vehicles in top shape. Instead of flipping through a thick user manual, owners simply need to ask their vehicle: “Which warning light is showing?” “How do I change a fuse?”

Aircraft repair

NLP helps mechanics synthesize information from huge service manuals to increase their understanding of problems reported by pilots.

1. Turn up the volume

Doing NLP right requires an insane amount of data. Think about every word ever spoken and you're almost there.

NLP has to be able to process, understand, and reference speech input against an immense library of data to create an intelligible response in milliseconds.

This requirement is especially difficult given the complexity of human language, which is full of rules and exceptions and becomes even more difficult when you consider the nuances of idioms, sarcasm, and humor. Industry-specific models can also require specific information about a particular domain, company, or products.

That's why the size of conversational AI models has grown to millions or billions of parameters. Typically, the more data, the more accurate the model. Training models of this size can take weeks of compute time and require the best-of-the-best machine learning and deep learning frameworks.

Google Translate

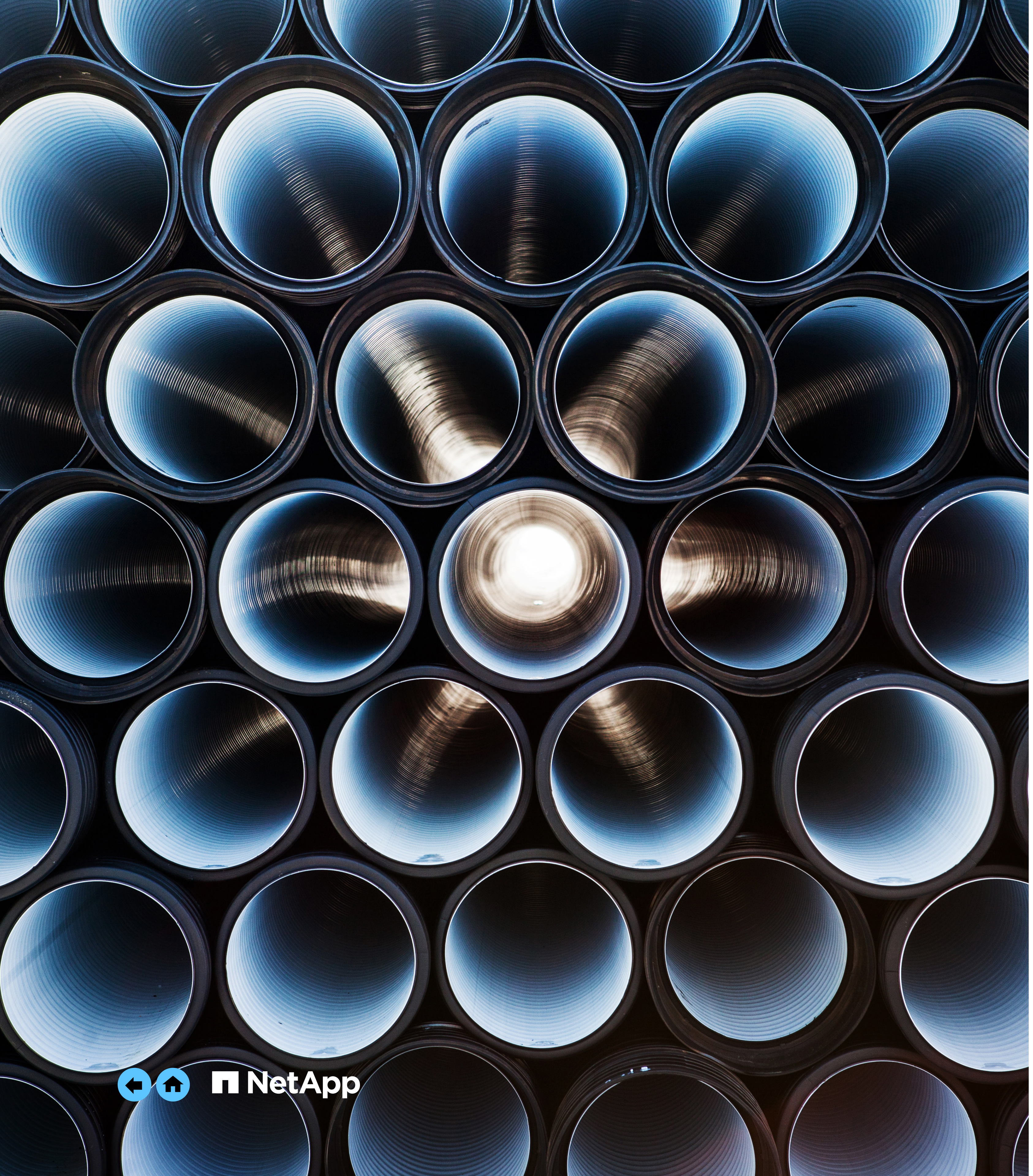
Google Translate supports over 100 languages and is crowdsourcing to help test and improve translation and model training for languages with limited training corpora (fancy words for a source dataset). Google Translate processes 140 billion words every day. That's the work of 70 million human translators. *Every day*.

Google BERT

Google BERT is a popular NLP model that has 340 million parameters. BERT represents a breakthrough in NLP because it goes beyond transactional voice interfaces, such as phone tree algorithms, to become truly conversational. It can read texts and answer questions with extremely high accuracy.

BioMegatron

BioMegatron is the largest biomedical transformer-based language model ever trained. It has up to 1.2 billion parameter variants. It was trained on 6.1 billion words from PubMed, a repository of abstracts and full-text journal articles on biomedical topics.

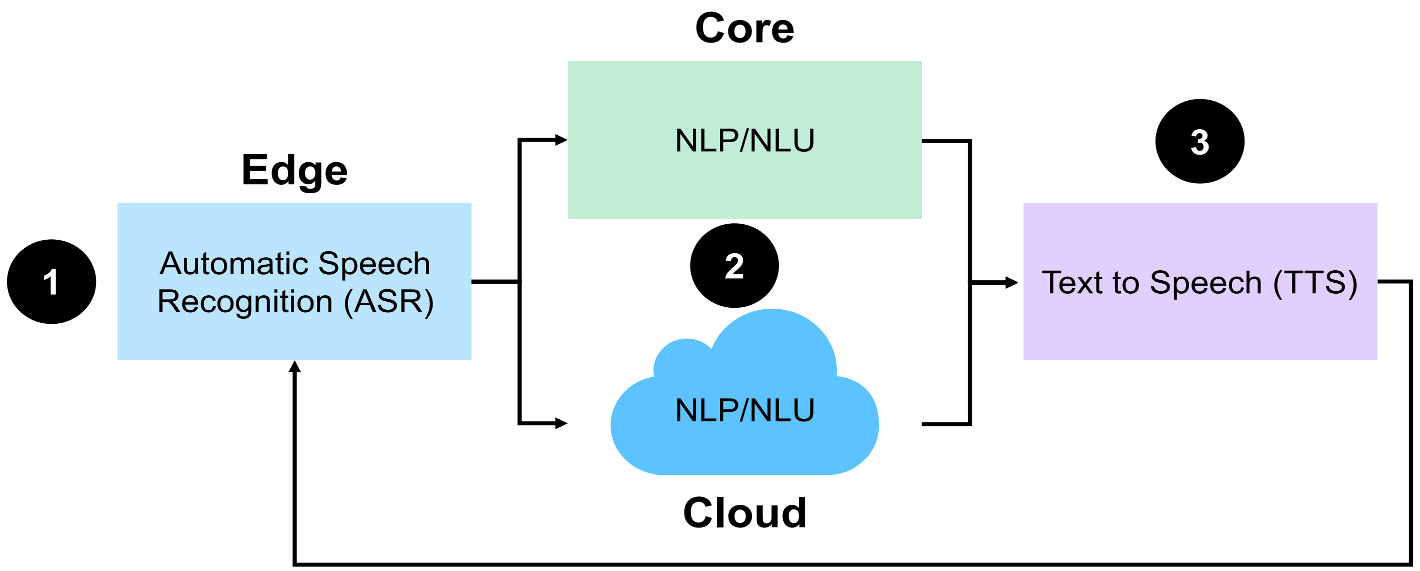


NetApp

2. Clear the pipes

Fast, effective NLP requires a data pipeline that spans the entire ecosystem, from ingest and recognition all the way to voice synthesis. Data must flow quickly and freely throughout each step of the pipeline to drive real-time language processing.

A typical NLP pipeline consists of 3 stages:



In a modern NLP infrastructure, thousands of edge locations gather terabytes of data each day. When access to this data is limited by a siloed infrastructure, deep learning only scratches the surface.

3. Respond in a heartbeat

For AI to replicate human speech, it needs to operate at the speed of a human brain—or even faster. The larger a model is, the longer the lag between a user's question and the AI's response. To sound natural, all the computation must take place in a 300-millisecond window.

That process takes several steps:

1. Convert the user's speech into text
2. Understand the meaning of the text
3. Search for the best response in context
4. Provide the response back in speech

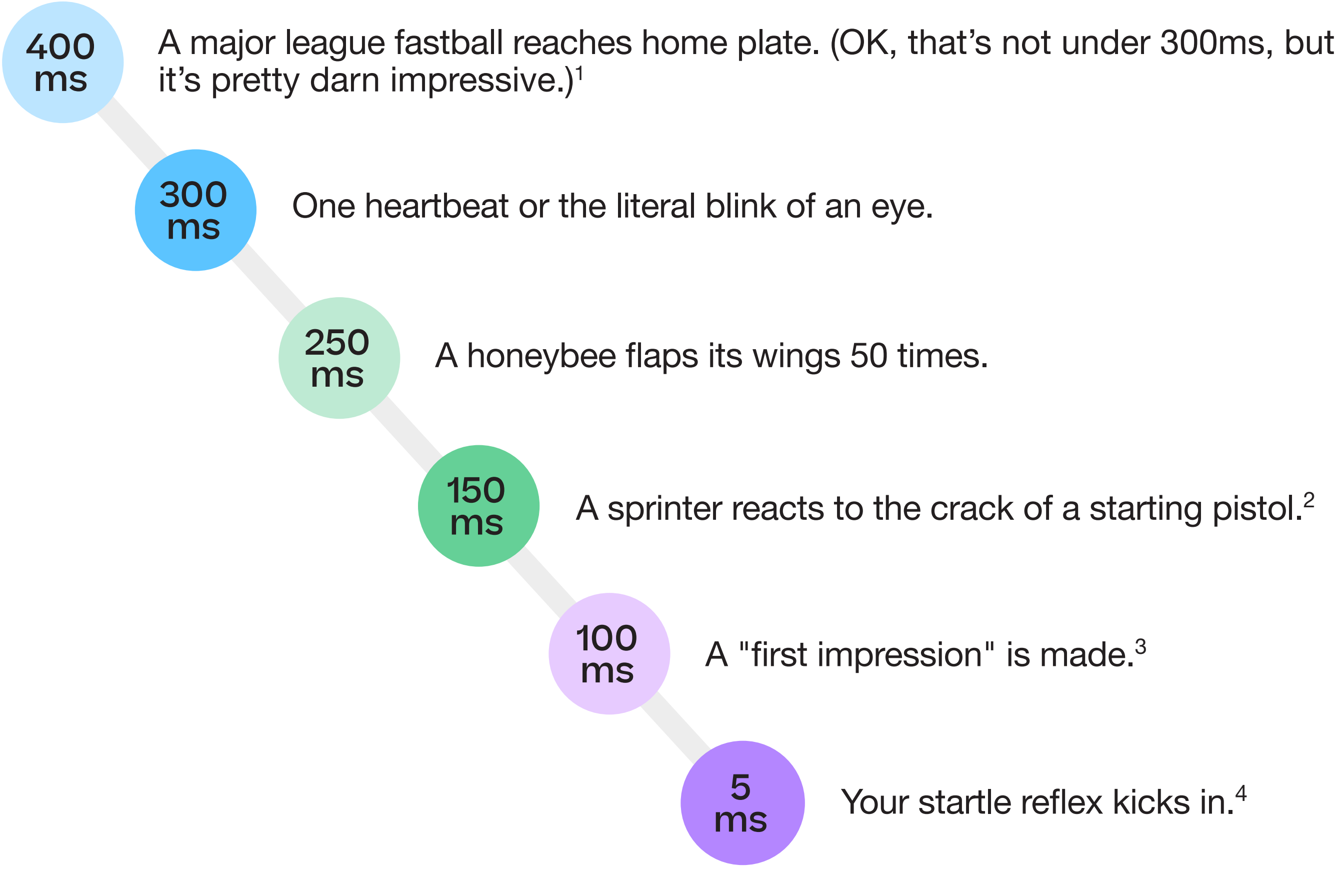
With such tight latency requirements developers of conversational AI often have to make tradeoffs. A high-quality, complex model could take longer than a less bulky language-processing model that delivers results quickly but lacks nuanced responses.

Like a nervous job applicant, a voice assistant may stall for time during conversations by saying, "Let me look that up for you," or making some bleeps and bloops to fill the awkward silence. But the ideal conversational AI—the Holy Grail of NLP—is sophisticated enough to accurately understand a person's queries, and fast enough to respond quickly in seamless natural language.

How fast are we talking?

NLP typically requires less than 300 milliseconds (ms)—that's 0.3 seconds—latency to create a real-time response. How fast is that? Very fast.

Things that happen in 300ms or less:



400 ms — A major league fastball reaches home plate. (OK, that's not under 300ms, but it's pretty darn impressive.)[1]

300 ms — One heartbeat or the literal blink of an eye.

250 ms — A honeybee flaps its wings 50 times.

150 ms — A sprinter reacts to the crack of a starting pistol.[2]

100 ms — A "first impression" is made.[3]

5 ms — Your startle reflex kicks in.[4]

NetApp speaks your language

With NetApp® ONTAP® AI, powered by NVIDIA DGX systems and NetApp cloud-connected all-flash storage systems, massive, state-of-the-art language models can be trained and optimized to run inference rapidly. A data fabric powered by NetApp simplifies data management across the AI data pipeline, from edge to core to cloud.

- **NetApp AI solutions** remove bottlenecks to enable more efficient data collection, accelerated AI workloads, and smoother cloud integration.
- **NetApp unified data management solutions** support seamless, cost-effective data movement across a hybrid multicloud environment.
- **NetApp's world-class partner ecosystem** provides full technical integrations with AI leaders, channel partners and systems integrators, software and hardware providers, and cloud partners. They put together smart, powerful, trusted AI solutions that help achieve your business goals.
- **NetApp Professional Services** provides the specialized expertise that you need to reduce complexity and to expand your AI opportunities and success.

And, by the way, NetApp is positioned as a leader in the IDC MarketScape for worldwide scale-out file-based storage.[5] Which is important because natural language processing workloads are—yep, you guessed it—scale-out and file-based.

Make your data scientists happy

5x

Run 5 times more data through your AI pipeline

<60 seconds

Copy datasets in seconds rather than in hours or days

~20 minutes

Configure your AI infrastructure with Ansible integration in ~20 minutes

The NetApp Retail Assistant: A blueprint for success

Using NVIDIA Jarvis, an end-to-end framework for building conversational AI services, NetApp and NVIDIA have built a virtual retail assistant that accepts speech or text input and answers questions about weather, points of interest, and inventory pricing by connecting to the weatherstack API, the Yelp Fusion API, and the eBay Python SDK. Check it out

The NetApp Retail Assistant (NARA) is built on:

- **NVIDIA Jarvis.** Jarvis provides GPU-accelerated services for conversational AI using an end-to-end deep learning pipeline that's optimized to keep latency low.
- **NetApp ONTAP AI.** This proven architecture combines NVIDIA DGX systems and NetApp all-flash storage. ONTAP AI reliably streamlines the flow of data, enabling it to train and run complex conversational models without exceeding the latency requirement.
- **NVIDIA NeMo.** A Python toolkit for building, training, and fine-tuning GPU-accelerated conversational AI models, NeMo enables you to build models with easy-to-use APIs, including real-time automated speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS) applications.

You down with NLP?

Yeah, you know me. What's next? Conversations with woodland creatures? We can't teach a squirrel to talk. We can teach you how to build the right AI infrastructure for NLP.

Learn more about NetApp AI solutions:

- NetApp AI
- ONTAP AI
- NetApp solutions for NLP

Questions? Our AI solution specialists are standing by.

1. O'Neill, Shane. Real-time bidding: What happens in 200 milliseconds? Nanigans.
2. Welsh, Tim. Exactly how long does it take to think a thought? The Christian Science Monitor. July 1, 2015.
3. Wargo, Eric. How Many Seconds to a First Impression? Association for Psychological Science. July 1, 2006.
4. Wise, Jeff. What Is the Speed of Thought? New York Magazine. December 19, 2016.
5. Potnis, Amita. IDC MarketScape: Worldwide Scale-Out File-Based Storage 2019 Vendor Assessment. IDC. December 2019.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services, and applications to the right people—anytime, anywhere.

To learn more, visit **www.netapp.com**

NetApp