STATE OF THE

EDGE

DATA AT THE EDGE
Managing and Activating
Information in a Distributed World

SEAGATE

TABLE OF CONTENTS

FOREWARD

Like an **unrelenting wall** of fog barreling across the Golden Gate bridge, data is shifting to the edge.

The transformation is inevitable. As billions of devices come online and begin churning out zettabytes of data, today's model of centralized cloud will need support from the edge.

To be sure, the major cloud operators will continue to advance their Internet of Things (IoT) and data storage solutions, lapping up your data to store it in their massive remote data centers. There will always be a role for centralized cloud storage, especially for long-term archiving. But centralized cloud is only part of the solution: Given the massive explosion of data at the edge, it's clear the pipes aren't big enough and the transport costs are too high to move all data to the core. As CEO of Vapor IO Cole Crawford likes to say, "There simply is not enough fiber in the ground."

But bandwidth limitations and transport costs are only part of the picture. The primary reason the cloud is shifting to the edge has everything to do with speed. As Marshall Daly, cloud data evangelist for Tableau Software **put it so elegantly,** our business goals require us to "reduce the time between asking a question and getting an answer." He goes on to conclude, "So, how do you reduce time to insight when it comes to your data? You create your analysis as close to your data as possible."

Placing compute, networking, and storage in close proximity to the devices creating the data makes it possible to analyze data on the spot, in real time, delivering faster answers to important questions, such as how to improve factory robotics or whether an autonomous vehicle should apply the brakes right now.

With soon-to-be billions of devices generating zettabytes of data, we need to rethink how we put data to work to maximize its potential. To help our community advance this discussion, **State of the Edge** partnered with **Seagate Technology**—a leader in data management and storage—to produce this groundbreaking report on data at the edge.

As we've said before, when it comes to evolving the internet, there is no finish line. This is an ongoing community effort. We encourage you to join our **State of the Edge Slack group** and help shape future *State of the Edge reports* as well as participate in offshoot projects, such as the **Open Glossary of Edge Computing.**
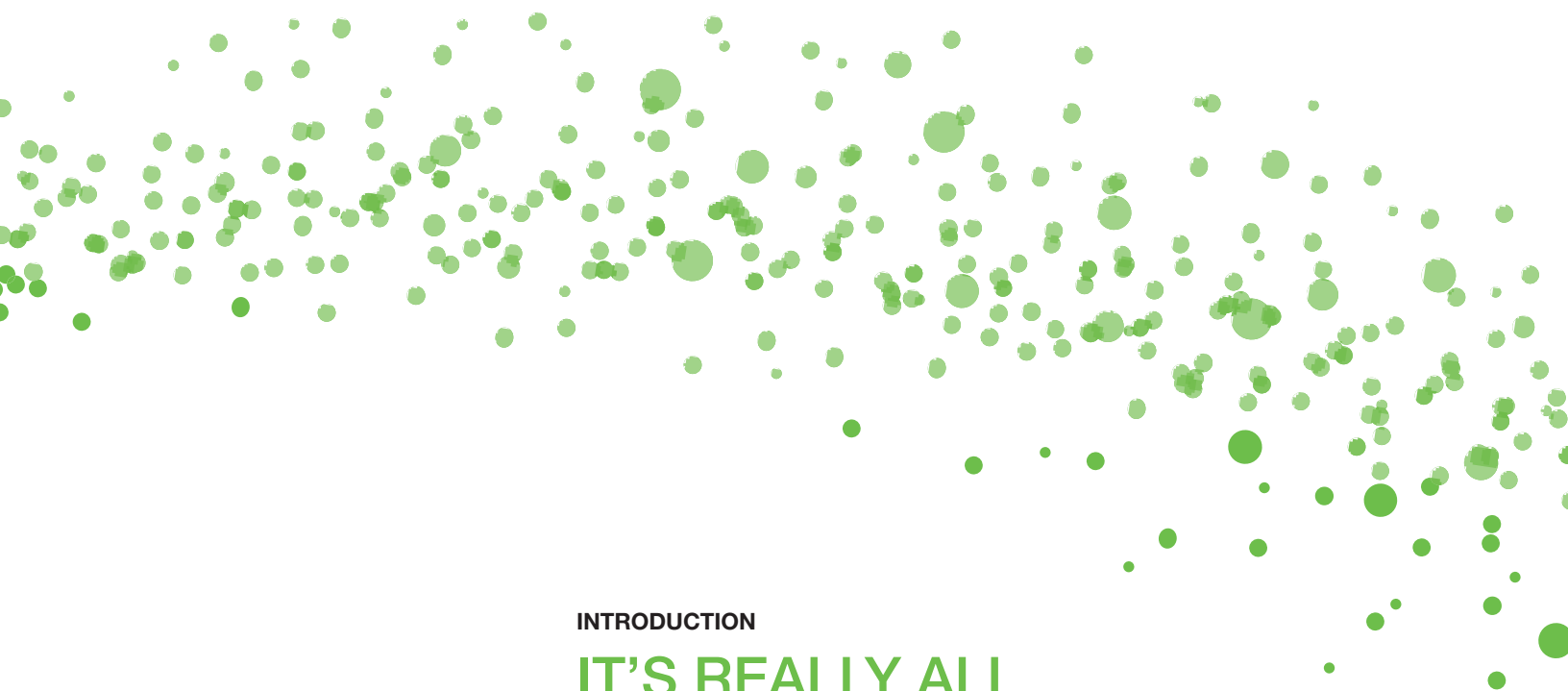
Sincerely,

Matt Trifiro
CMO, Vapor IO
Co-Chair, *State of the Edge Report* Chair, *Open Glossary of Edge Computing*

Jacob Smith
CMO, Packet
Co-Chair, *State of the Edge Report*

INTRODUCTION

IT'S REALLY ALL ABOUT THE DATA.

Everyone seems to be talking about the new world of edge computing. Some posit that the cloud, emergent over the last decade and quickly dominant, will surely be swallowed up by the scale and ubiquity of the edge. Others believe that the promise of rapid artificial intelligence close to the source means that everything will be disrupted.

Over decades of work at the frontlines of digital innovation, I've learned that every business is a data business. The way we use, analyze, and act upon data keeps changing. It's a reality for which every successful business must prepare.

Innovations in data management have paved the way for much more efficient ways of using information, and data at the edge is no different. Since we find ourselves in the midst of an unprecedented data revolution, I am proud to introduce this report. How are the ways we use data restructuring our world? What new opportunities for extracting value from data arise at the edge? What does the brave new world at the edge mean for enterprises, cities, small businesses, and individual consumers? How does data at the edge enable us to work, play, live, commute, and leave the world a better place for future generations? These are the questions that this report answers.

Be informed—and get ready to put your data to work!

John Morris
Chief Technology Officer
Seagate Technology

SEAGATE

EXECUTIVE SUMMARY

This report shows how the rise of the edge shapes data and how data shapes the edge—and that enterprises can now begin to extract previously untapped value from data in the new ecosystem. Among the findings:

- Data generated by the year 2025 will be an astounding 175 zettabytes, a tenfold increase from 2016 levels. The need to manage this staggering volume of data is going to be a key driver of distributed architecture.

- The center of data's gravity is shifting to the edge of the network. More and more data will need action at or near the edge and away from the core.

- The creation and consumption of massive amounts of data will be catalyzed by four enabling technologies: Internet of Things devices, 5G, AI, and edge data centers.

- Four key factors drive demand for edge computing: latency; high data volume accompanied by insufficient bandwidth; cost; data sovereignty and compliance.

- The unique mix of technology and economics will make it practical to assemble, store, and process more data at the edge.

- As massive amounts of data are created outside the traditional data center, the cloud will extend to the edge. It won't be cloud versus edge; it will be cloud *with* edge.

- Activating data at the edge makes it possible to ask better questions and get more timely answers.

- New edge computing architectures must accommodate the following: harsh conditions; remote locations; quality of service/low latency; no-touch/self-healing technologies; dynamic provisioning; global data experience; and security.

- While centralized cloud computing will persist, a radically new way in which we create and act upon data at the edge of the network is creating new markets and unlocking new value.

STANDING AT THE
EDGE

A farmer stands in the field. She surveys it and plans for successful crop yields.

To improve the effectiveness of her harvest, she needs to fertilize the soil, optimize watering, and set a course of action upon any signs of infestation. Gone are the days when she had to rely on her gut and a *Farmers' Almanac* for guidance.

Hers is a data-driven farm.

Edge-enabled, AI-powered smart agriculture is her farm hand.

The farmer—let's call her Julia—is a smart farm's CEO. Automated machines at her disposal measure the levels of fertilizer and pesticides, then inject them where needed at the right time. Smart crop sensors keep tabs on soil hydration, pH, and nutrient levels, correcting any anomalies in real time. These predictive and self-corrective systems take action when needed.

Julia's farm generates a lot of data, far too much to send thousands of miles away to a centralized cloud. Instead, she relies on edge computing, storing, and processing the data at or near her farm. Remote monitoring software lets her easily check on her work from anywhere—even a continent away.

Julia employs farmers, but also gets an assist from her on-farm weather station, AI-enabled telematics devices, satellite imagery, automated soil sampling, and smart monitoring of crop health—all of which get analyzed at the edge to deliver real-time adjustments to correct course. As a result, she maximizes crop yield much more so than she would without the edge-enabled aid. While simple environmental sensors like soil monitors add improvements, game-changing inputs come from drones mounted with cameras, which gather high volumes of data for processing on or near her farm. Drone-mounted cameras provide a multitude of capabilities—from crop anomaly detection to fertilization and yield estimation.

The World Bank reports that global demand for food will grow 50% by 2050 if the population continues to rise as projected, pointing to investments in agricultural technology as a way forward. Edge-powered agriculture—in which data is captured, processed, and acted on in real time at or near the field—is part of the answer.

Edge-driven automation is already putting food on tables. In Chile, an AI-powered, wireless sensor-equipped **irrigation system,** which aids in the cultivation of blueberries, is expected to reduce water usage by 70% compared to other irrigation methods. In Japan, robots are **growing lettuce** in factories with floor-to-ceiling stacks that resemble high-rise buildings more than they do conventional farms. In some instances, automation has reduced labor costs by 50% and smart LED lighting for cultivation has lowered energy costs by 30%. Eventually, "[e]verything after seeding will be done by machines—watering, trimming, harvesting—on shelves stacked from floor to ceiling," reports **the BBC.** "It's a bit like the solitary drone farmers in the 1972 film *Silent Running.*"

Some prognosticators have declared that "**data is the new oil.**" Whether this metaphor holds or not, this much is clear: the ways we use data, not to mention how and where we process it, are changing—and opening up new opportunities.

Image: istockphoto.com

THE EXPLOSION OF
DATA
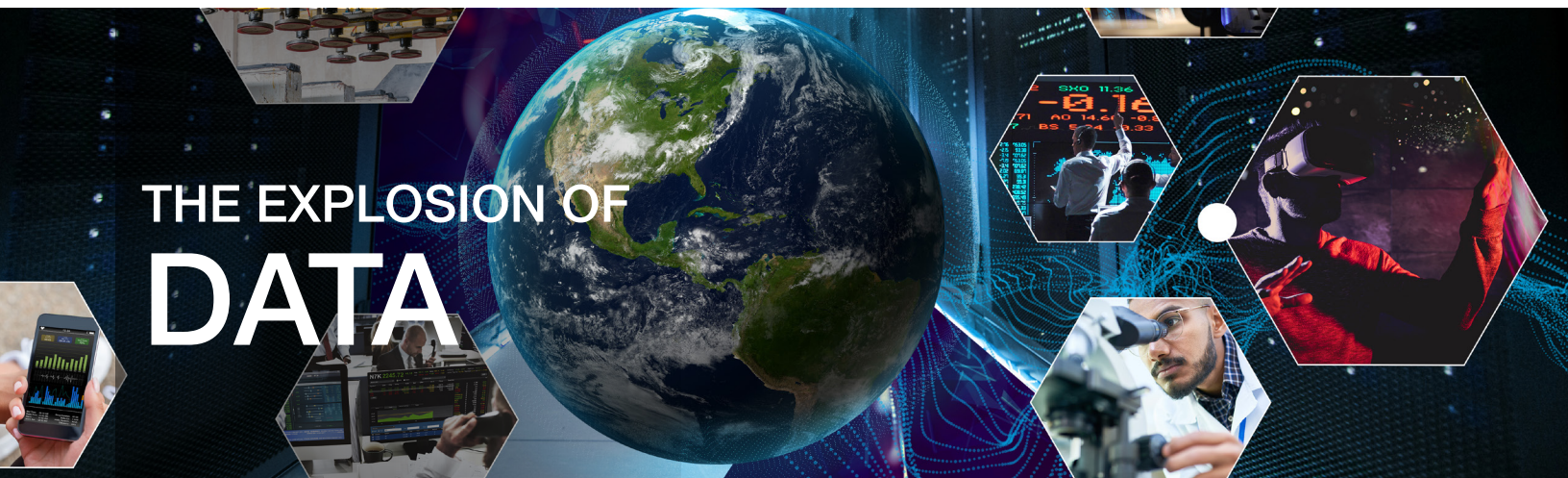
Data, quite literally, has always been in our DNA, woven into the very fabric of the known universe. The pre-digital world, too, pulsated with data.

The world is data.

But today's data is different than yesterday's. We are living in what market intelligence firm IDC has dubbed the Data Age. The IDC Data Age 2025 report *The Digitization of the World: From Edge to Core* forecasts that the sum total of all data generated by the year 2025 will be an astounding 175 zettabytes—a tenfold increase from 2016 levels. The edge-proximate Internet of Things (IoT) devices alone are expected to create over 90 zettabytes of data. (As a point of scale, a zettabyte is a trillion gigabytes.)

While the sheer volume of data growth boggles the mind, how data is being used is changing, too. Back in the mainframe era, which began in the 1960s, storage was expensive and data was maintained in siloes. Data analysis was slow and cumbersome, and practitioners lacked tools and applications to extract maximum value from the data.

As IT moved from mainframes to mass-market client-server models in the 1980s—characterized by the emergence of SQL databases running on affordable servers—an ecosystem of independent software vendors emerged to deliver applications that could take advantage of data. At the same time, storage costs decreased and data grew from terabytes to petabytes.

Image: Seagate Technology

A little over a decade ago came the mobile and cloud revolution, integrating data with our everyday life and work. The simplicity and economics of cloud computing led to an exponential growth in the number of applications that collect and process data. As cloud services made it easy to unleash immense computing power on demand (at a fraction of the cost of previous generations), data analysis became more commonplace. As mobile networks expanded and smartphones proliferated, mobile devices doubled as data collection machines.

As the second decade of the new millennium draws to a close, hundreds of billions of connected digital devices are creating massive amounts of new data. Applications powered by AI are unleashing new ways of extracting value from information. The creation and consumption of data will be catalyzed by enabling technologies like 5G, which make it cheaper, faster, and easier to get more data from endpoints into processing locations. What we see is the next shift in scale of data, from exabytes to zettabytes.

Numbers in aggregate may be hard to visualize, so here are some specific and tangible examples offered in research done by IDC, McKinsey, Intel, and Cisco:

- A typical smart factory will create about **5 petabytes of video data** per day
- A smart city with about one million people could generate **200 petabytes of data per day**
- An autonomous vehicle (AV) could produce **4 terabytes of data** per day, and there could be 200 million AVs on the roads in the next few years

DATA IS SHIFTING TO THE EDGE

More and more data is changing in nature: from business background to life-critical; from centralized to IoT and embedded systems; from being generated by passive machines to AI-induced; from unconnected and unhurried to mobile and real time; and from generated and consumed centrally to locally.

According to **a report** by the IT research firm Gartner, 91% of today's data is created and processed inside centralized data centers. But data is rapidly shifting away from the core.

The same report predicts that

by 2022 about 75% of all data will need analysis and action at the edge.

That is a drastic change with implications for every enterprise.

Four big technology drivers underlie the massive shift of data to the edge:

1. Artificial intelligence (AI) has become cost-effective and practical;

2. Billions of IoT devices are being deployed;

3. Wireless operators are upgrading their networks to the fifth generation of cellular mobile communications (5G); and

4. Innovations in edge data centers are solving for the complexities of distributed facilities and unit cost economics.

Each of these trends on its own will drive demand for edge processing, but all four of them combined makes edge computing inevitable.

Let's look at each of them in turn.

ARTIFICIAL INTELLIGENCE

AI will shift massively to the edge of the networks where data is created.

"AI is a general term for software that mimics human cognition or perception,"

**– Frank L. Ruta,** *The Economist*

Forbes has declared AI at the edge as **the next goldmine.** There are at least four key reasons for this: cloud-related latency, the prohibitive cost of moving large data sets to the cloud, security threats to data, and network congestion. Goldmine indeed. Gartner predicts the global business value derived from AI in 2018 will exceed US$1.2 trillion. That is a staggering increase of 70% over 2017. *Bloomberg* posits that AI will add US$15.7 trillion to the global economy by 2030.

WHAT MAKES AI SO POTENT, ESPECIALLY AT THE EDGE?

Typical deep learning applications, such as face recognition and human tracking in camera networks, exert a great deal of pressure on the infrastructure of the cloud. According to **SDxCentral,** "it is much more cost effective to run analytics on the edge and send small batches of condensed information to the core." The performance of today's AI often improves with higher volumes of data. For example, when using machine learning, computers build models of behavior with little to no human supervision. This happens in a two-step process:

1. Learn the model by training with little to no human supervision.

2. Make inferences based on the learned model.

With a continuous feedback loop and the availability of large volumes of data, the learning process gets better with each iteration. But large amounts of data put a lot of pressure on the network. Processing and acting on data at the edge can improve the efficiency of machine learning. Edge servers can ingest raw data and perform preliminary learning, storing, and processing of the data locally to reduce network traffic. Gartner analyst Thomas Bittman **explains** how AI can aid in a smooth-functioning edge and
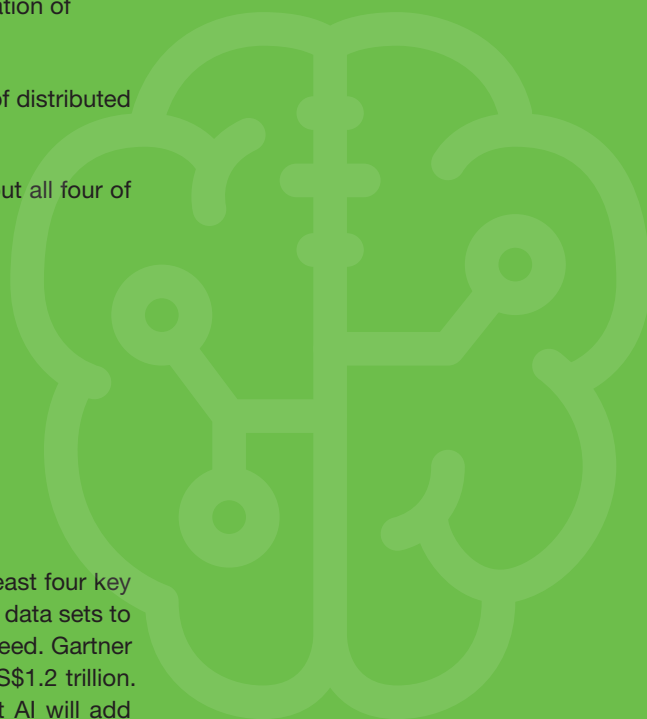
why this matters to consumers: "Massive centralization, economies of scale, self-service, and full automation get us most of the way there—but it doesn't overcome physics—the weight of data, the speed of light. As people need to interact with their digitally-assisted realities in real time, waiting on a data center miles (or many miles) away isn't going to work. Latency matters. I'm here, right now, and I'm gone in seconds. Put up the right … advertising before I look away, point out the store that I've been looking for as I drive, … help my self-driving car avoid other cars through a busy intersection. And do it now."

THE INTERNET OF THINGS

Billions of data-generating devices are coming online, and they are connecting to the Internet. These devices will create massive rivers of data that need ingesting, analysis, and processing—in many cases, at the edge. This is the IoT.
BI Intelligence forecasts that:

"There will be more than **24 billion IoT devices** on Earth by 2020. That's approximately four devices for every human being on the planet."

The value derived from IoT sensors creates a self-perpetuating cycle. The more sensors a system uses, the more data the system collects—and the better the predictive model built with the data can be. This, in turn, leads to demand for even more instrumentation with more sensors.

Sensors, meanwhile, are also evolving. As the resolution and frame rates for video have gotten more advanced, the camera lens has emerged as the most prolific and versatile IoT sensor. Image sensors capture rich context that cannot be acquired with single-purpose devices, like a temperature sensor that periodically takes measurements. When combined with deep learning (DL) algorithms running on powerful graphics processing units (GPU) and tensor processing units (TPU), images open up new uses. As IoT-enabled cameras create large image sets, we run into the constraints this report addresses below—latency, bandwidth, cost, and, in some cases, legal and ethical compliance, etc.—all of which drive the need to process data closer to the source.

5G

As telco operators worldwide begin upgrading their networks to the fifth generation of wireless cellular technology, 5G, the economics and practicality of edge computing improve dramatically. The combination of existing 4G LTE networks and evolving 5G technologies will fundamentally transform how we use data over cellular networks. Not only will these improved networks deliver end-to-end latencies of 5 milliseconds or faster and peak data rates of 10GB per second, they will also be able to handle an exponentially larger number of connected devices—on the order of 1 million devices per square kilometer.

While today's 4G LTE networks do a respectable job of handling person-to-person and client-to-server communications, 5G will usher in an age of machine-to-machine (M2M) transmissions where billions of devices will transmit petabytes of data to servers at the edge.

And if 5G feels like some far-off technology, it is not: it's right around the corner. In October of 2018, AT&T **completed** the "world's first millimeter wave mobile 5G

browsing session with a standards-based device on a mobile 5G network." By 2025, 5G networks will cover one a third of the global population.

The device and data growth fueled by 5G will make the telco infrastructure the ideal spot—the beachfront property—of edge computing.

5G network enhancements will dramatically change the performance and economics of cellular networks, fueling demand for infrastructure edge computing, which will drive investments in servers, storage, and networking deep into the telco infrastructure, such as at the base of cell towers.

By placing computing resources at the infrastructure edge, 5G networks can capture enormous amounts of real-time data and cost-effectively process it for value extraction, versus simply throwing it away. As more data reaches the 5G cell towers, there is simply not enough IP backhaul bandwidth available to transport data to a centralized cloud, making it necessary to add additional computing resources to edge locations.

5G will unlock new use cases and product innovations that benefit from edge computing. It "represents a fundamental rearchitecting of core Internet standards, effectively remaking the Internet to be natively mobile," writes Internet industry analyst Larry Downes in *The Washington Post.*
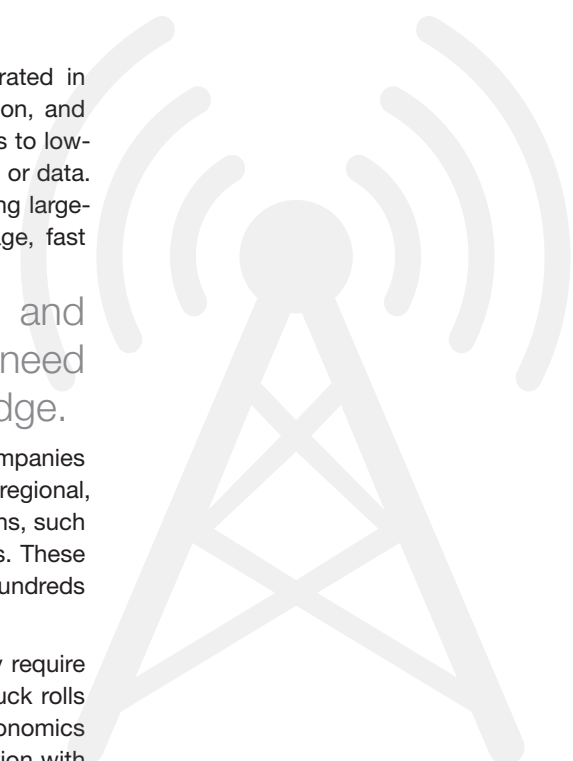
EDGE DATA CENTERS

On the whole, society has benefited from hyperscale data centers, operated in centralized and remote locations by the likes of Google, Facebook, Amazon, and Microsoft. Sites for these data centers were selected based on ready access to low-cost power and land, and for green reasons, more so than proximity to users or data. This model continues to work well for many centralized applications, enabling large-scale archiving, massive content distribution, large-scale application storage, fast prototyping, etc.

As more and more data is collected, analyzed, and applied outside the traditional data center, the need arises for a complementary model: cloud *with* edge.

To augment the cloud at the edge, a new class of data center has emerged. Companies such as Vapor IO, EdgeConneX, and DartPoints have been deploying small, regional, and micro-regional data centers at the edge of the network—in novel locations, such as in parking lots, on municipal rights of way, and at the base of cell towers. These edge data centers are characterized by their location, relatively small in size (hundreds of square feet, not thousands), and remote, lights-out, operations.

Edge data centers are self-contained, highly-integrated, and compact. They require exceptionally high reliability, and they must be remotely-operable. Costly truck rolls (the acts of sending technicians to an edge location) can quickly break the economics of an edge location. To minimize service trips, edge data centers must function with no humans on site, which requires new kinds of remote monitoring and AI-based automation.

Another benefit: edge data centers can be slipped into existing sites and structures, reducing expenses. For example, most of the 100,000 cell towers in the US have ample room for a micro-modular data center.

DATA AT THE EDGE IS HERE

The unprecedented convergence of AI, IoT, 5G, and edge data centers creates a unique mix of technology and economics, making it practical to assemble, store, and process vast amounts of data at the edge.

They are why our smart farmer, Julia, is free to focus on what humans do best: make complex, far-reaching decisions. The promise of data at the edge, when it comes to the farmer and the world hungry for the fruit of her savvy, is that she gets to survey her acreage and strategize efficient yields—and to let the machines talk and problem-solve among themselves.
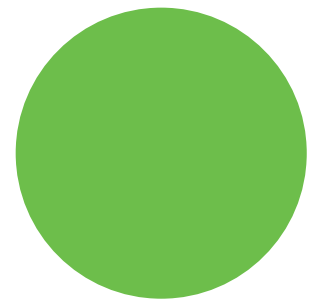
MESSAGE FROM THE MACHINE SHOP

From: Erik Salo, Director of Customer Development, Seagate

I toured a machine shop last year. Lots of thuds and huge apparatuses. Careful workers in grease-stained overalls tending to the production. Forklifts busily ferrying about heavy blocks of black steel. The shop had that dusty gray vibe associated with factories making heavy machinery—a tractor or a boiler maybe. It wasn't the sort of place I expected to see a need for the next phase of the inexorable revolution in how people use data—edge computing—but I couldn't have been more wrong.

We turned a corner and saw three huge milling machines all fenced off with a robot arm in the middle. The owner explained to me that the robot took a part from one machine and put in the next machine for the next operation. It seemed like a lot of trouble to pick something up and move it three feet. However, the owner explained the math. The machine cost X dollars to acquire and set up. Running three shifts of people with benefits and sick time and all the rest cost Y dollars. X is less than Y. Hence the robot.

While I sat watching the mechanical ballet of this and other robots and thought about the nature of human work—wondering if it was right or wrong or just so—we heard a clunk.

The robot had dropped one of the parts and the whole thing came to a halt. My guide told me that "the thing works great, but every now and then it drops one, and someone has to come in and reset it." He went on to say, "What would be really cool is if the robot could look at how the part comes out of the machine and adjust where it grabs it to pick it up. That would save a lot of setup time and also reduce costs."

That, of course, is machine learning. If the machine could teach itself how to drop fewer parts, the shop would be more efficient and whatever widget it was making would cost a bit less. I told my guide that I thought the step he imagined was much closer to reality than he thought. Meanwhile, a technician had picked up the dropped part and the machines were happily whirring away. All I could think about was that I really needed to encourage my kids to learn how to program…

THE EDGE IS NEAR

—SO YOUR DATA CAN BE, TOO

The edge can be found anywhere and everywhere, in a wide range of locales, including:

- Floors of manufacturing plants
- Roofs of buildings
- Cell phone towers in the field
- Barns on farms
- Platforms at oil and gas fields

As should be clear by now, **the edge is a location, not a thing.** It is the outer boundary of the network—often hundreds or thousands of miles from the nearest enterprise or cloud data center, and as close to the data source as possible. The edge is where **"real-time decision-making untethered from cloud computing's latency"** takes place.

The 2018 *State of the Edge Report* and its companion project, the *Open Glossary of Edge Computing* differentiate between two types of computing at the edge, according to where the data and storage are located relative to the last mile of the network.

- **Infrastructure Edge** is computing capability, typically in the form of one or more edge data centers, deployed on the operator side of the last mile network. Resources on the infrastructure edge allow for cloud-like capabilities that resemble those found in centralized data centers, including the elastic allocation of resources, but with lower latency and lower data transport costs, due to being closer to the end user or device, compared to a centralized or regional data center.

- **Device Edge** refers to computing capabilities on the device or user side of the last mile network, such as smartphones, smart sensors, and autonomous and connected vehicles. The device edge may also include gateways and related field devices that aggregate and pre-process data before sending it to the infrastructure edge for further processing and forwarding.

Regardless of how the edge emerges, we can always think of it as supporting the decentralization of data and the ability to process data near where it is created.

DEMAND FOR EDGE COMPUTING

We've seen how AI, IoT, 5G, and edge data centers have accelerated computing at the edge. But what problems are we solving with edge computing? Why do we even need edge computing in the first place? And how does data figure into this?

In order to delve deeper into these questions, one must understand the four most important factors creating demand for edge computing:

LATENCY

In the context of edge computing, latency refers to the time it takes to send data for processing plus the time it takes to get a response in return. Even though data travels at the speed of light (at least until one needs to route or switch it), for some applications this is still too slow. While humans might be willing to wait hundreds of milliseconds for data to travel from our smartphones to a centralized data center thousands of miles away, many machine-to-machine applications don't have that luxury. For example, an application monitoring on an oil rig for anomalies might have only tens of milliseconds to identify a problem and take corrective action. If the processing resources are located far away, the corrective action might not happen in time, leading to an expensive equipment failure.

HIGH DATA VOLUME, INSUFFICIENT BANDWIDTH

 As the number of data-collecting sensors reaches into the As the number of data-collecting sensors reaches into the hundreds of billions, there will be an exponential growth in the number of endpoints gathering data. Not only do these devices continuously sense and generate data, much of the data will be very large. A recent **analysis** by *The Wall Street Journal* documents the use of data-rich camera images in various industries. "Being able to see is a major frontier in robotics and automation," the article posits, with cameras enabling robots and humans alike to do more than was possible. Seagate, for example, has manufacturing facilities that use cameras in automation—and in just one manufacturing step alone, the system can generate over ten million high-detail images per day. It is simply not economical to send all that data to a centralized brain, especially across expensive long-distance network connectivity. Besides, in most cases only the (much smaller in size) insights from the data must flow to the centralized cloud; much of the raw data can be stored or pre-processed at the edge.

COST

With long-distance network capacity being limited and expensive to upgrade, the increasing demand for sending large volumes of data will result in higher spends on data transit. Even as operators upgrade network capacity, the cost of investments will be recouped in the form of higher prices to end users. Edge computing will can reduce networking expenses by processing, reducing, and discarding data before it hits the long-haul networks.

At the same time, centralized data centers are expensive investments that require a large footprint with efficient operational frameworks for power and cooling. Building large-scale data centers in urban locations can be cost-prohibitive or impractical, especially when land is scarce. In these environments, micro-modular data centers used in edge locations have many cost advantages. Distributing the computing and storage resources to the edge using micro-modular data centers avoids many of the high upfront fixed costs of centralized data centers by leveraging existing infrastructure like cell towers.

DATA SOVEREIGNTY AND COMPLIANCE

Even after the physics and economics, there is another factor at work that prevents data from leaving certain regions to a centralized location. Data ownership and data compliance policies often restrict where data can be sent and processed. Country-specific laws often restrict how data may move across country boundaries, and some cities mandate that all of their municipal data remain in their municipality. Storing and processing data closer to the location where it is generated, at the edge, becomes another way to abide by these evolving compliance requirements.

DOES EDGE MEAN THE END OF CLOUD?

Sobered by the inexorable growth of the edge, some industry leaders have predicted the end of the cloud. A popular version of this argument comes from **The End of Cloud Computing** by venture capitalist and Andreessen Horowitz partner Peter Levine. In this talk he declares that the edge will usurp the cloud. While Levine's argument may be enticing, it is also wrong.

To be sure, the edge will unlock greenfields for innovation—with resources like GPUs and TPUs making it possible to deliver cloud-scale applications at the edge. **Edge-native** and **edge-enhanced** applications will capture greater mindshare and investments than could cloud-only applications, yet won't diminish the cloud.

The cloud and edge are not mutually exclusive. As Matt Trifiro of Vapor IO **argues,** the cloud will change, "but it won't be cloud *versus* edge. It will be cloud *and* edge." He goes on to write, "The large centralized cloud data centers, such as **those owned by Microsoft** and **Google,** will be augmented with thousands of **micro data centers at the edge** of the last mile network. These micro data centers will be placed as close as possible to the devices and people they are serving, such as at the base of cell towers and on the roofs of buildings."

THE VIRTUOUS CIRCLE OF EDGE

Is the edge shaping the data or is it data shaping the edge? The answer is yes.

Imagine a food-processing plant looking to improve efficiency and throughput with sensors and edge processing. The plant experts may initially focus on solving a supply-chain bottleneck or addressing demand variability. But solving those problems may reveal new problems in the manufacturing process that will require instrumentation with sensors, better data collection, and analysis.

Or take the changing labor conditions and labor shortages that may drive a milk farm to adopt robotics and data collection. The availability of this data, collected originally to solve a labor challenge, can now be used to solve new challenges, such as lowering cost or boosting productivity.

In both these cases: the edge shapes data, then data shapes the edge.

The initial deployment of sensors and data collection often leads to quick results, even though they are often proofs of concept that leverage the centralized cloud with a simple store-and-forward edge gateway. The first wave of success leads to more sensors and more data growth, forcing businesses to face the latency, bandwidth, cost, and compliance factors. This prompts storage, processing, and intelligence to move to the edge. The factory floor, the cellular network tower, or the barn on the farm become the edge.

As these edge solutions pack additional compute power, new opportunities open up. The robotics-enhanced milk farm may add vision sensors and high resolution cameras, leading to more sensors, more advanced data—and continuous value creation, or the so-called **virtuous circle.**

THE SHIFTING GRAVITY OF DATA

The phrase *data gravity* was coined by inventor David McCrory, who envisions data as having mass. As data accumulates, he describes it as gaining mass. Like in a celestial system, the data body and its mass attract applications and services.

Every analogy works until it doesn't. All the data created at the edge cannot remain at the edge because the edge will have limits in scale, flexibility, and manageability. Consider the complexity of storing months' worth of data from a factory floor that generates a petabyte of data per day. Once the first-order value extraction is done, data can move to the cloud for second-order processing and long-term storage, or may be deleted when insights from it have been extracted.

Thus, the need for an edge-with-cloud model versus an edge-replacing-cloud alternative. Capacity at the edge can reduce the data, making it far smaller than what was initially-collected—dropping data that is not needed and compressing whatever remains. This reduced data set, which one may want to store longer or combine with other data sets, can be transported to the centralized cloud. In addition, certain applications benefit from being in a centralized location—for example, those that look across learnings from multiple edge nodes and create second-order-learning and insights that can be distributed to all the edge nodes.

Data gravity bends the fabric of cloud, extending it to the edge and forming the **edge cloud.**

THE EDGE OF DATA SECURITY

No one has the luxury of treating data security as an afterthought. Whether data is in production, use, storage, or transit, protecting it is vital.

Take an autonomous car, which captures and processes images for navigation using several different cameras. If a bad actor takes hold of this data, the consequences can be catastrophic. The same concerns apply to public safety projects, where cameras examine aging bridges and other infrastructure for threat detection and safety. Life-critical and business-critical data abounds, and we must have confidence in its security.

What sets apart security for data at the edge? Three aspects in particular:

- As devices enter and exit the network, they must be trusted to participate in the data domain.

- The data itself must be trusted, whether used immediately in analytics and AI activities, transmitted across nodes, or uploaded to the cloud.

- Data in motion will be more vulnerable as it traverses networks, so nodes need to communicate securely over disparate networks.

The security challenges presented by the edge demand new solutions from the network, system, and component designers. Trust in devices and data must be established to protect systems from tampering and unauthorized access. Imagine if a bad actor gained access to an entire node of data being stored and analyzed on an edge device. Would it be possible to change the data and effect a different outcome on an autonomous vehicle, or in a surveillance system? What if a device entered a network and sent invalid or erroneous data to an edge AI system, resulting in a manufacturing facility losing a shift of production?

As AI and other autonomous technologies play larger roles in society, security must become a foundational design element. Ensuring the security of data requires us to assume networks will be broken into. Edge locations can't always offer conventional physical security—such as fencing, gates, and on-site personnel.

A barn in the middle of a field or a node at a street corner needs a different kind of protection.

All a bad actor needs is access to an aggregation point, and vast quantities of data can be at risk. What happens to data then?

Encryption technology—coding messages in such a way that only authorized parties have access—provides one solution. If the data is encrypted and the key is outside the network, the data remains safe—even when the network is intercepted. Self-encrypting drives are an example of this technology.

Establishing a trusted network of devices and data at the edge will ensure the security and providence of data as edge applications develop. Authentication can provide a handshake between devices, creating trust that the data is valid.

In addition to encryption and authentication, AI provides another way to secure the edge and its data. By creating a model with trusted historical data, AI systems can continuously scan new information for anomalies that may indicate intrusion. This can be useful in public safety projects—such as using cameras and sensors to monitor the wear and tear in aging infrastructure in order to prevent failing bridges by detecting cracks in beams. It's relatively easy to detect both **behavioral and infrastructural anomalies** thanks to AI and machine learning. Thanks to real-time processing at the edge, both machines and humans can act to remedy threats before they grow.

Another use case of growing importance is autonomous vehicles (AVs) in transportation and smart tractors in agriculture. As AVs and smart tractors move and create data, the data in motion from the AV endpoints to the local storage and compute edge devices must be protected and certified as authentic before being accepted for storage and processing. Technologies in play range from video to LiDAR technology to enable navigation and prevent collisions, all of which generate significant volumes of data.

The threats to data at the edge range from data manipulation onboard AVs and smart tractors, rogue device control, denial of service, taking over device control, and manipulation of sensor data and signals. Since AVs and smart tractors are cogs in a much bigger ecosystem of data sharing, the devices themselves must be trusted to ensure the hardware and firmware responsible for the navigation and control of the devices is authentic. The value of encryption and authentication cannot be overstated. And, due to the low latency and high performance requirements of AV applications, encryption has to be performed in a manner so as not to detract from system throughput performance. This tends to favor the use of hardware-based device-level encryption.

DATA MANAGEMENT AT THE EDGE

An edge data center may include both fast and slow resources, which balance conflicting needs of speed and cost. The ability to move data to the appropriate place, depending on the profile of each application, is important for the user experience and necessary to ensure that the requirements of their applications are met.

By definition, data at the edge is highly distributed; data for any one application can be spread across dozens or even thousands of sites or nodes. Systems for data management remove this complexity, giving developers and operators a holistic view of their data even when it spans multiple sites. This enables a better user experience by combining many underlying resources and presenting them as a single data repository. It also improves application performance by anticipating and responding to demand, moving data to ideal locations. By analyzing the data needs of different applications—all of which may ask for resources at the same time—data management at the edge can optimize the placement and availability of data in a balanced way.

Data management at the edge depends on two technologies: orchestration and automation. Combined into a data management system, they allow edge environments to ensure applications have access to the data they need while also conforming to constraints such as cost or data sovereignty.

ORCHESTRATION

Just as the conductor of a musical orchestra unites the collective output of individual performers, data orchestration coordinates data across many locations, taking into account all the locations in the system and treating them as a whole. Orchestration at the edge makes it possible to manage data across a multitude of edge nodes, balancing data ingest with application responsiveness, as well as overseeing the movement of data among devices, the edge, and the core. Data orchestration at the edge uses real-time analysis and feedback to respond to changing network conditions and application demands. This calls for automation at each step.

AUTOMATION

In a distributed world, it is not possible for personnel to inspect and monitor the system. The edge takes advantage of automation to manage vast quantities of data without the intervention of human operators. Autonomous systems can detect issues such as imminent data loss and take corrective action, often self-healing in real time. Complex algorithms seek to understand and anticipate requirements at the edge, including how different applications will make demands on data, then implement actions that optimize the entire system in real time.

Edge computing introduces new complexities in data management that don't exist in centralized data centers. For one, the number of edge data centers, often placed at 5- to 10-mile intervals, requires management of data within individual data centers as well as across, or among, a cluster of data centers. Moreover, the variety of edge resources adds more complexity, with real-time data traversing access and aggregation points.

With a large number of devices accessing an edge data center at the same time, orchestration and automation systems must allocate resources appropriately to meet required response times and provide the right information to applications, precisely when needed, on a 24×7 basis.

The overall ecosystem of data at the edge is complex. An autonomous vehicle may send data to the edge, requesting real-time analysis to improve its decision-making. The ingest of data and feedback to the vehicle need to be managed automatically to ensure timely delivery. The data then needs to be consolidated and aggregated. Data and workloads must move across the access and aggregation layers in real time, and algorithms must anticipate potentially thousands of vehicles feeding back and requesting similar information at the same time.

Autonomous agriculture provides similar challenges. With drone monitoring of crops, the user experience is greatly improved by the processing of data at the edge. Instead of waiting months to discover trends and make adjustments, a fleet of drones—like those on Julia's smart farm—can continuously track yield, resource utilization, crop infestation, and other data points. The ability to automate data flow for real-time adjustments has a direct bearing on the bottom line. Farm data can be fed back into an edge data center and analyzed right away to correct watering levels or soil composition, and can alert farmers to trends that may affect yield. Instead of wiping out an entire season's return due to an unforeseen issue, the issue can be found and fixed right away based on real-time capture, data ingest, and analysis of data with orchestration and automation optimizing each step.

DATA ACTIVATION AT THE EDGE

The Data Age, in which the value of data drives every part of human flourishing, is here indeed. That is an exciting reality.

But there is another way to look at it: data can be inherently dull and boring. In his book *The Book of Why: The New Science of Cause and Effect*, Turing Award winner Judea Pearl writes, "You are smarter than your data." The computer scientist goes on to remind us that it is us, humans, who make sense of data. Be it running a farm or factory, it is how the data is activated that makes the results possible.

Data activation refers to the ability to collect, store, categorize, and analyze data so that it may be acted upon in real time, leading to the best possible decisions based on the available data.

Small or big, data is brimming with stories to tell and answers we seek to the pressing questions, be it adding the right amount of fertilizer to a cornfield in Iowa or detecting anomalies in a smart city from streetlight camera feeds. To get to that point, data must be primed and ready at a moment's notice for humans to act upon it—to ask questions and hear those stories and answers.
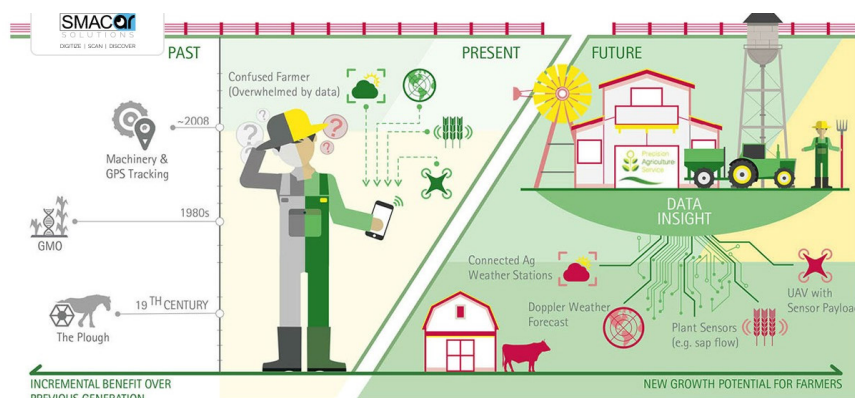

Image: **SMACAR Solutions**

That is data activation: making sense of the available data. It's about putting data to work, and about extracting the maximum value from it. Data offers endless potential—to save lives, to increase revenue, to protect assets. Data activation is the unlocking of that potential.

Today's data activation methods can take us beyond historical reviews and rearview-mirror visibility. No longer constrained by overnight and end-of-month reporting, modern data analysis systems employ deep learning and leverage the immense computing power of GPUs to derive instantaneous insights. Nanoseconds and nanometers matter, and activation tools can be brought close to the data.

Capturing, storing, securing, and managing data at the edge means we have the ability to ask not just bigger questions, but more nuanced ones. For example:

- How can I reduce my investments in new machinery?
- Is there a way to cut days from the production process?
- Is this widget going to fail in customers' hands?

Humans have a penchant for asking questions, and making decisions based on the answers. That's how we find meaning; it's how data earns its meaning too. Activating data at the edge allows us to ask better questions and get more timely answers.

MESSAGE FROM THE FACTORY FLOOR

From: Rags Srinivasan, Senior Director of Vertical Markets at Seagate

It was another cold day in Normandale, Minnesota, the kind of day you hear about on the radio and feel glad you are not there. For Bruce, it was, more importantly, 126 days before his solution at the factory had to go live. Bruce is Bruce King, senior data scientist from the Seagate factory team, tasked with finding new ways to improve manufacturing quality. As he and I were chatting on the way to the factory, Bruce recalled the demands placed on his team:

"Reduce new investments for clean-room resources."

"Cut down days—no, weeks—from manufacturing time."

"Catch even more defects before the final assembly step."

These are daunting asks, given the scale of the manufacturing and low tolerance for errors.
How do you eat an elephant? One bite at a time, of course. "There was no doubt in my mind we could achieve all this," Bruce told me. "And the solution was clear to me: it is data and AI. But I needed to reduce the problem to one scenario, one step in the complex process, one set of manageable data so I could build on it."
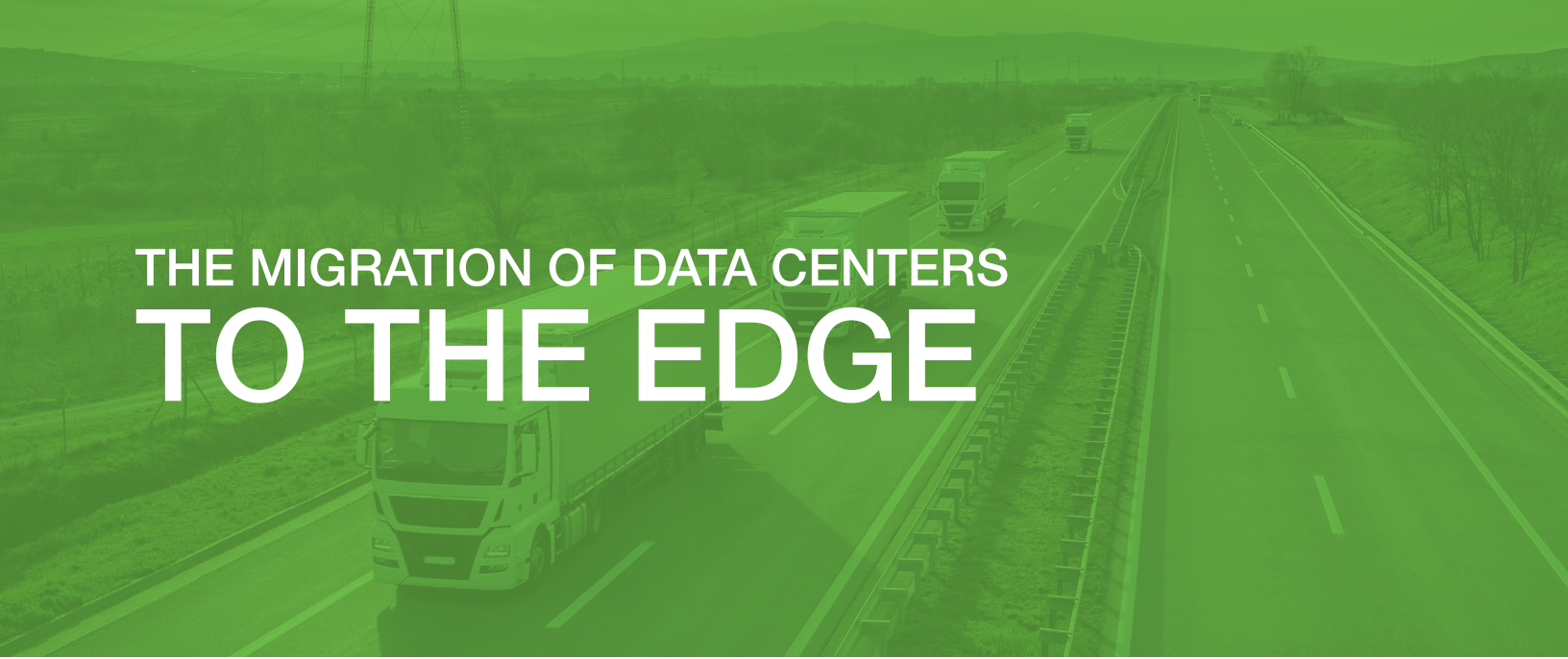
Bruce faced an immense challenge. It involved multisite factories, multiple steps in the manufacturing process, tens of terabytes of data created per day on the factory floor. He had to answer big questions:

1. Can he afford, in terms of money and time, to send all this data to the cloud for processing?

2. Why can't the data be stored right there on the factory floor—and be put to work?

3. What if he needs to retain the data for one, three, or even five years?

4. What if he needs to combine learning from multiple sites?

5. How can he orchestrate the movement of data from the factory floor to a more economical long-term archive?

Bruce would not have used the word edge to describe his domain. Though, as it turns out, Bruce's factory floor, where machines and people manufacture millions of units per month, is indeed the edge—a location close to the data source, where data can be processed and acted upon.

Like Bruce, most practitioners do not start with the goal of building an edge infrastructure. They all start with high-order business goals. They are driven by data and what they can do with it with the power of AI. Like Bruce, they rejoice at seeing the high volume of data available for them, but face the challenge of moving it away from its source to a centralized location. They find a way to bring the computing closer to the source, be it the factory floor or a barn on a farm.

The edge, that is.

THE MIGRATION OF DATA CENTERS
TO THE EDGE

During the first wave of data center deployments, in the 1980s, the focus was on providing infrastructure to support mission-critical applications of business entities. These data centers supported such functions as stock market transactions and patient databases. Retention of data was critical, but making data accessible also played a role.

This problem gave life to traditional enterprise data center architectures. In them, each IT function—i.e., computing, networking, and storage—is typically deployed separately and connected together through a network fabric. Each function runs on highly available (HA) hardware units. Whether an HA server or a 5×9 storage subsystem, they are designed to have enough redundancy to deliver 99.999% uptime and to be able to handle most failures locally. This means high cost and a complex deployment model.

The next wave of data center innovation started with Web 2.0 at the turn of the century. These data centers were built in response to the new ways people were using data—in search engines, social media, and smartphones. Providers of these services faced a new challenge: Lots of data, generated from multiple sources flowed into their environment. Ad-dependent business models relied on user satisfaction, approximated by performance. The economics of these use cases would not support the cost of traditional enterprise infrastructure.

This led directly to the emergence of the cloud. Data center architecture in the cloud era has been dubbed *hyperscale* due to the gargantuan facilities that house thousands of servers. Instead of using expensive HA components, hyperscale data centers desegregate hardware and software. They leverage clusters of commodity servers, storage, and networking gear. Redundancy and resilience of the infrastructure are offloaded to the software stack, giving rise to scale-out technologies like **Hadoop** and Google's **Borg.**

The software-based approach to reliability enabled the hyperscalers to expand computing, storage, and networking separately and dynamically, using inexpensive hardware. This new design combined with faster internet connections allowed the cloud providers to start monetizing their infrastructure beyond their own use—which is why we now have Amazon Web Services and Google Cloud Platform, among others. The cloud eliminated or at least reduced the need for on-site infrastructure at many businesses, allowing them to eliminate capital and operating expenses while shifting to a pay-as-you-go model of computing.

BUILDING AN ARCHITECTURE FOR EDGE

Over the past decade the cloud has enabled many new markets and use cases. Alas, as data has proliferated, the pipes that connect that data to the cloud have not kept up. Not only that, but the value of real-time decision-making has become more crucial.

The need to process certain types of data closer to the edge of the network presents a new challenge for data center architects who aren't used to, for example, designing for a regulated central office for telco appliances or on top of a light pole in Phoenix, Arizona.

From a data and data-flow standpoint, key questions need to be answered before one can confidently propose an architecture: How much data is generated? What type of metadata is required? What throughput and latency are expected by the application?

The edge and its attributes vary widely based on the application.

One person's edge could be someone else's core.

This means over the next few years new architectures and deployment models will be proposed, tested, and verified, or abandoned, in favor of others for different use cases. Edge computing presents unique challenges to solution architectures.

Here are the most important aspects of edge computing that new architectures must contemplate:

HARSH CONDITIONS

Unlike centralized hyperscale data centers, which can be carefully and methodically located, edge data centers often need to go near the data, no matter how harsh the location. Edge architectures, therefore, must consider higher or lower sustained temperatures, stricter shock and vibration requirements, flood and earthquake risks, air pollution, and other environmental factors that aren't usually a concern in centralized locations.

REMOTE LOCATIONS

The more the edge expands, the more it will need facilities in distant locations that are hard to service and secure, such as on an oil field or at the base of a cell tower. These types of edge locations present higher operational costs and greater risks to data.

QUALITY OF SERVICE/LOW LATENCY

While some edge devices and use cases will have the benefit of on-device processing and autonomy (such as AVs), others will rely heavily on reading and writing data to edge servers (examples include image processing and streaming VR games). These client-server scenarios require guaranteed levels of performance at sufficiently low latencies. Of course, different applications will require different service-level agreements, inherently producing a multi-tenant environment.

Given these demands and constraints for edge locations, there are a few basic solutions for the inherent challenges:

NO-TOUCH/SELF-HEALING

Remote locations require the optimal combination of resiliency, redundancy, and flexibility that can be delivered via no-touch/self-healing technologies. Ideally, a remote data center can detect problems and heal them without any user intervention, and be able to guarantee a certain quality of service. This capability can be brute-forced through hardware redundancy and resource overprovisioning, but it can be ideally delivered with software that intelligently rebalances resources using orchestration and automation.

DYNAMIC PROVISIONING

In order to provide the required quality of service across use cases, an ideal architecture would dynamically provision compute, storage, and networking services on demand, distributing workloads across multiple edge locations. This kind of infrastructure on demand can be realized with an edge application programming interface (API) that exposes hardware functionality and resource provisioning. This would allow edge applications and services to expand hardware resources based on the workload needs without requiring the deployment of more hardware.

GLOBAL DATA EXPERIENCE

While edge nodes allow data to be stored and acted upon locally, the data should be treated as a global entity—its location should not matter to the user. This calls for storage software (file systems) that support a global namespace across many devices and which can modify the location of data as user behavior changes. Granular multitenancy control allows multiple applications to coexist in the same locations and service devices across a shared, distributed cluster.

SECURITY

The highly-distributed nature of edge computing creates a larger surface area for attacks and the remote locations may make it harder to detect physical breaches. Therefore, edge architectures must be built with security in mind for each layer: supply chain, physical security, security of the data-at-rest (hardware, firmware), security of the data in-flight (software), and of course network security to the core.

None of the functions discussed here are new in nature, as the elements behind a hyperscale environment were not. But it is the deployment of all of them together as a single solution, a recipe, if you will, that creates unique architectures for edge.

CONCLUSION

The edge is near.

Both spatially, as in closer to the data source, and temporally—the edge is here and now.

It is true that centralized cloud computing persists. But an unprecedented way in which we create and act upon data at the edge of the network is creating new markets and unlocking new value. In order to keep growing, businesses need to take advantage of this new field of edge-powered opportunity.

The intent behind this report is to enable its readers to tap the power of data at the edge. We want them to imagine what can happen when massive amounts of data are harnessed and activated at historically unprecedented speeds, locations, and scale. Data at the edge can help solve the greatest conundrums facing humanity and its planet today—one farm and one factory floor at a time.

For an in-depth report on the edge, visit **https://www.stateoftheedge.com** to read the landmark *State of the Edge 2018* Report. And to find practical, end-to-end, edge-enabled solutions, visit Seagate's Data Labs at **https://labs.seagate.com.**

AUTHORS

Rags Srinivasan is Senior Director of Seagate's Growth Verticals. He is responsible for market development, ecosystem partner development, and solutions marketing for edge and IoT. Rags has held several product leadership roles in Veritas, EMC, and HGST and launched storage systems for enterprises and cloud scale customers. He has a master's degree in Computer Science from Colorado State University and an MBA from UC Berkeley.

Agnieszka Zielinska is Senior Editor and Content Strategist at Seagate, where her focus includes solutions at the edge. She received two master's degrees and a Ph.D. candidacy (ABD) in the social sciences, from the University of Chicago and Northwestern University. Her journalistic work has appeared in Reuters News, *The Wall Street Journal*, and the *Chicago Tribune*'s City News.