SmartCitiesWorld
White paper

Autonomous vehicles and the data that is powering them

How data is fuelling the connected car revolution

In association with

CLOUDERA

Silver
Business
Partner

IBM

Written by

Sue Weekes
News Editor,
SmartCitiesWorld

SmartCitiesWorld **White paper Reports examine an emerging or growing trend in smart cities, highlighting progress so far and future potential, as well as spotlighting case studies from cities around the world.**

In this report, we examine how data, and the effective processing of it, will be central to a successful connected and autonomous vehicle ecosystem.

www.smartcitiesworld.net

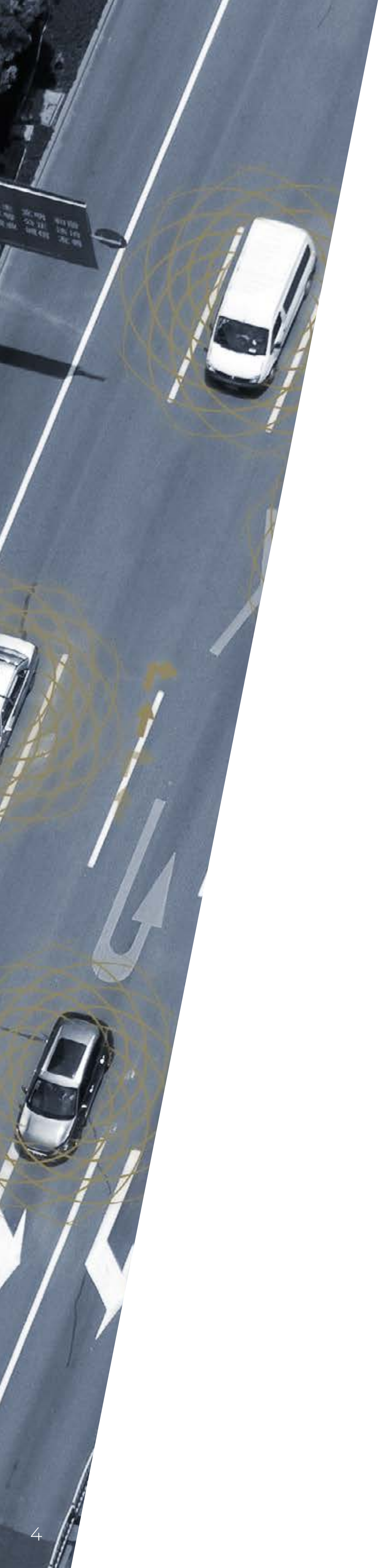
Introduction: Why the car's not the only star

When we visualise the city of 2050, many of us picture executives accessing their daily newsfeed as their car drives them autonomously to the office. Smart intersections respond in real-time to the conditions around them. Congestion has been consigned to history and the stresses once associated with commuting to work and getting around the city are a dim and distant memory. Best of all, the city has not witnessed a road traffic accident in years.

When, or indeed if, this becomes a reality is largely down to what happens now as connected and autonomous vehicles (CAV) begin to be integrated as a critical component of the city's mobility landscape. Indeed, connected technology can help to tackle major global issues such as congestion, pollution and road safety.

Anyone who has bought a new car in recent years will have noticed the huge leap in technology that has taken place over the past decade. Connected vehicles (CVs) are bringing seismic technological shifts to driving, but autonomous vehicles (AVs) bring cultural and societal ones, not least because their progress and integration relies on road users of all kinds putting all their trust into technology.

We are still some way off AVs being part of everyday life in our cities but efforts to make this happen have been ongoing for several years. Pilot projects and trials are underway in many cities around the globe and key milestones and deployments that take us nearer to the future are regularly reported on.



According to the National Conference of State Legislatures, around 30 US states have already enacted legislation for autonomous vehicles in preparation for the future. Arizona is one of those leading the way with AV technology company Waymo (formerly the Google self-driving project) launching the world's first commercial autonomous ride-hailing service in Phoenix in 2018.

In Singapore, AV trials on public roads began back in 2015. The country's Land Transport Authority has since expanded the test bed to neighbouring areas and trials are also taking place on Sentosa and Jurong Island. In the UK, the Government-backed Project Endeavour self-driving initiative has chosen the city of Birmingham as the location for its next stage. Technology from autonomy software specialist Oxbotica will power a fleet of specially adapted Ford Mondeos to operate across a five-mile area in varied traffic and weather conditions as part of a live trial.

Elsewhere in Europe, autonomous shuttles and pods are already carrying passengers across campuses, airports and designated areas. Mobility programmes are underway in cities including Hamburg, where German transport company Hochbain is among the partners in the Hamburg Electric Autonomous Transportation (HEAT) R&D programme. The shuttle, developed by IAV Automotive Engineering, is completing test-runs in preparation to run the next passenger service trials to start in late summer 2021.

While the car we see on the road is the star in any discussion about AVs, conjuring up fantastical images of the future, what is often overlooked is the data management challenge that is going on under the chassis. While developments in sensor, radar and lidar technology have all been vital to the development of AVs, it is data that will train these vehicles to be safe and the data they ingest and generate that has the potential to make cities safer, more efficient and convenient.

Indeed, Michael Ger, managing director, automotive and manufacturing solutions at data management company Cloudera, contends that the development of autonomous vehicles is as much a data management and analytics challenge as an automotive engineering one. “Data is foundational to the discussion,” he says. “And we need to understand how its value can be leveraged across the broader ecosystem of use cases, such as city departments, transport authorities as well as private sector companies.”

Key to this is arriving at a better understanding of the connected and autonomous car's data lifecycles and how these can be optimised. With the average CAV producing 25GB of data per hour per car, traditional data storage and management methods simply aren't up to the task. Unless the challenges this brings are met, CAVs may be unable to fulfil their considerable potential. In this white paper, SmartCitiesWorld and Cloudera aim to highlight the challenges ahead for CAVs and explain how a better understanding of the advanced tools and technologies required to handle these vast amounts of data will put in place the right foundations for the future.

Data: it's far more than the new engine oil

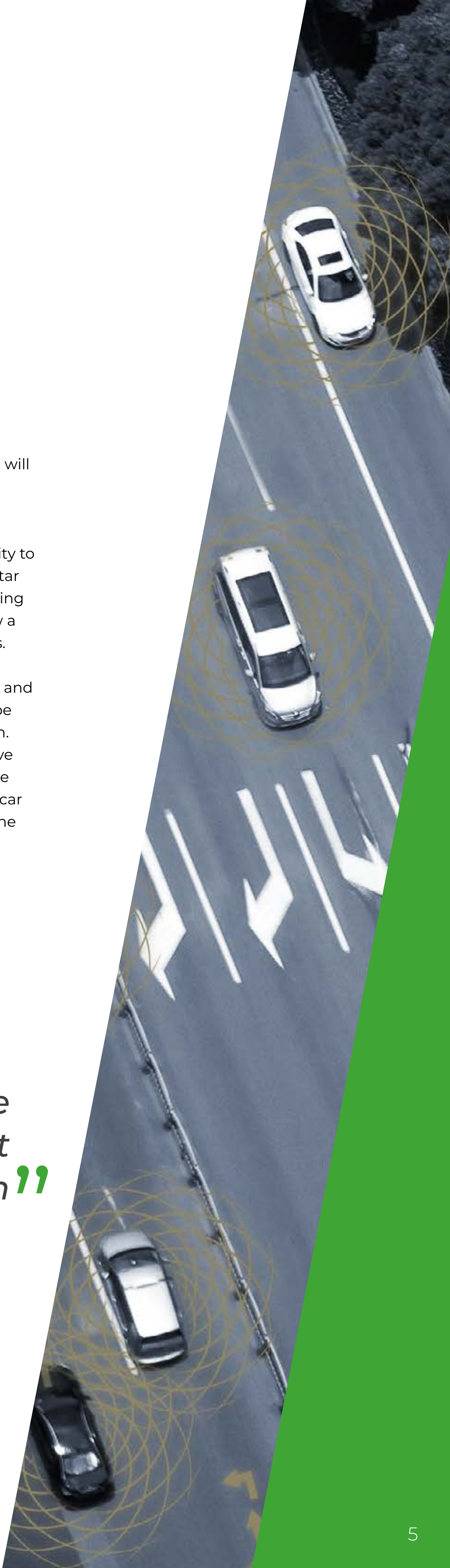
Data underpin the development and running of both connected and autonomous vehicles and it will be the data they produce that feeds into mobility and other smart city programmes in the future. It is also data that will help to instil trust in them in the minds of all road users.

Early adopters of connected cars will have noticed the vast leap in in-car technology that has taken place in recent years. Back in the 1990s, the ability to call hands-free from a car via, for instance, General Motors pioneering Onstar system was revolutionary. Basic infotainment systems then started appearing while insurance companies saw the value in collecting information on how a person was driving via dongles to feed into usage-based plans and policies.

Fast-forward to today and smart navigation and connection to emergency and breakdown services have become the norm. Drivers now expect their car be able to monitor and report on its performance, usage, condition and health. What happens next as we move into a more proactive phase with predictive maintenance, advanced vehicle diagnostics and advanced driver assistance systems (ADAS) marks the biggest technological leap yet. This will see the car use the data it generates and technologies like machine learning to help the driver make optimised solutions.

"Basically, the data and machine learning will help the driver make their next best action and this is a really important development in terms of sophistication. Driving becomes a data-driven activity and this opens the floodgates into what is possible," says Ger. "Ultimately, advanced driver assistance systems will evolve into autonomous driving systems."

> *"Data and machine learning are a really important development in terms of sophistication"*



To understand challenges of moving into this next phase though, it is important to fully appreciate the sheer volume and diversity of data these vehicles produce once they are out on the road. As stated, the average connected car produces 25GB of data per hour per car. Over the course of a year, a single car produces 130 terabytes. This data is extremely diverse and includes data generated by the car relating to its speed, fuel efficiency, location and safety data and increasingly vehicle-to-vehicle (V2V) and vehicle-to-infrastructure data (such as data from traffic lights and road sensors) as well as video. It is also likely to be collected from even more sources in the future.

How we use this data to enable drivers to make optimised decisions raises questions on a practical level. First off, where do the decisions need to be taken – in the car at the edge or in the cloud? The reality is both. If it comes down to safety, such as telling a driver when to brake for collision avoidance, they will always need to be taken in the car as you cannot risk loss of connection, even once we fully move into the 5G era where latency is greatly improved. Less crucial analytics though can be run in the cloud. “So you must have a tiering of where you make those decisions and run the machine learning,” says Ger.

The Connected Vehicle Data Learning Lifecycle
Central to using this data to teach cars to be smarter is what Cloudera calls the Connected Vehicle Data Learning Lifecycle, which begins and ends in the car. The below stages show how the lifecycle would work.

1: The car ingests data streamed from various sources such as camera, radar, lidar, GPS via an in-car processing unit.

2: The large volumes of data mean that ideally this needs to be intelligently filtered so only the most valuable and relevant data is sent to the cloud.

3: The data are then stored in a scalable way in the cloud.

4: Sensor data from the car are enriched by augmentation data from another source, such as other cars or traffic infrastructure. For instance, for the purposes of predictive maintenance, sensor data from the car could be combined with dealership repair records and a correlation made between the sensor values that predicted when this car would need maintenance.

5: Once the variables are understood, machine learning models could be developed, updated and deployed back out to the car (in this case the edge device) as an over-the-air software update.

With the right data management technology in place, this learning lifecycle can be ongoing with machine learning models continuously evolving and making the vehicle smarter.

The autonomous driving data loop

Avoiding a pedestrian who steps out, identifying dangerous roads conditions or safely changing lanes to avoid debris in the road are typical scenarios that an AV must be prepared for but are just the tip of the iceberg. Vast collections of data from diverse sources are required to describe virtually any condition or situation a vehicle will encounter and provide "ground-truth" inputs to teach a vehicle to drive. Indeed, developing an autonomous car is a petascale endeavour and it is also a life or death one. The average amount of data needed to build one and ensure it is even safer than a car with a driver is an estimated 150 Petabytes.

The data currently being collected from AV test vehicles around the world are providing the foundational building blocks to enable vehicles to be able to safely drive themselves. Machine learning is used to train the AV's perception layer, which is effectively how the vehicle sees the road and what is happening around it, based on sensor data collected by telematics, cameras, lidar, radar and more. "This is fundamental, since any actions taken – such as instructing the vehicle to make a path adjustment – will be contingent on accurate perception layer 'vision'," explains Ger.

Given the sheer number of real-world variables an operating vehicle is exposed to and the associated zero-tolerance for error safety requirements, autonomous driving is among the most challenging machine learning use cases imaginable.

> *"Developing an autonomous car is a petascale endeavour and it is also a life or death one"*

The autonomous vehicle's data lifecycle starts on a similar path to that of a connected vehicle but has additional layers and complexity and makes even more demands on data management systems.

Phase 1: The car ingests data streamed from various sources such as camera, radar, lidar, GPS via an in-car processing unit.

Phase 2: It is stored in the cloud but must be available to a wide range of data consumers and made searchable for the likes of machine learning engineers and system test engineers who are developing the perception layer.

Phase 3: Training the perception layer using machine learning models.

Phase 4: Verifying the machine learning models, which involves deploying the perception layer into a mass verification environment.

Phase 5: Models are deployed into either simulation environments or pushed directly back out into a vehicle.

Both lifecycles are based on a complex set of requirements and will rely on an ecosystem of technology partners. And, as Ger highlights, while discussion of this ecosystem is about the solution initially, it is quickly moving to a user ecosystem made up of cities, government agencies, transport authorities as well as insurance companies. "And this is how connected vehicles will make their entry into smart cities," he says.

Before this happens though, the data management and analytics infrastructure behind the lifecycles have to be optimised. In the past the data management process has been fragmented and subject to too much waste and inefficiency. More robust technologies are also required to enable the lifecycle to be leveraged more cheaply and therefore re-iterated more frequently. In turn this leads to more continuous improvement of machine learning models which ultimately takes us closer to future when CAVs can live up to their billing.

Data: the more you have, the less you have?

No-one would argue that data are at the heart of making our cities safer, more sustainable, more efficient, cleaner, less congested and generally more liveable. Also, no-one will dispute that we are moving into an era where the volume of data from sensors and other devices being implemented in cities will exponentially grow. But it comes with a warning: much like paper-based information, it can be a case of the more you have, the more it could become impossible to manage.

The amount of data already being generated by CAVs, let alone in 10 years' time, is a clear case in point and it serves to highlight the data management issues that are likely to become common to many smart city applications in future as they scale-up.

The limits of classical data management technologies such as plain file system storage and legacy databases have already been recognised by many participants in the autonomous vehicle field but maybe less so outside of it. These methods will be insufficient to put long-lasting foundations in place for smart cities to grow and prosper. Moreover, more advanced data management isn't just required to tackle high volumes of data but also to put in place systems that address other common challenges around data usage and sharing such as regulation, privacy and ethics.

There is another compelling reason, too. Unless data management is up to the task, it will be impossible to take advantage of technologies such as 5G and edge to AI, which have the potential to unleash a plethora of exciting applications for cities. "Cities are going to have all of this technology available to start using data at the edge in ways we've never been able to before," says Brian Hagen, senior solutions engineer, Cloudera, who works directly with cities deploying new technology. "Going forward cities will need a robust data management backbone that will enable them to maximise the value of data efficiently and cost-effectively."

> *"We are moving into an era where the volume of data from sensors and other devices being implemented in cities will exponentially grow"*

In 2019, Cloudera and IBM announced a strategic partnership to develop joint go-to-market programmes designed to bring advanced data and AI solutions across the Apache Hadoop open-source ecosystem. The aim was to help customers who want a hybrid and multi-cloud data management solution with common security and governance, as well as provide an ecosystem of integrated products and services, designed to help organisations achieve faster analytic results at scale.

Working together, Cloudera and IBM provide a hybrid multi-cloud solution made up of component parts that can stream and ingest real-time and other data from a range of diverse sources, store it at petabyte scale, enrich it, and build machine learning models that can be deployed for visualisation and analytics.

“In a connected environment data are going to be being consumed by a number of different audiences and stakeholders inside and outside of the city, from government agencies to compliance specialists and insurance companies,” explains Hagen. “With data coming from a variety of different sources, it is important to show both consistency and confidence.”

Such an approach seeks to immediately allay some of the major security concerns around data-sharing which are only going to become more complex with the proliferation of edge devices streaming data. The CAV sector has its own challenges in this area. For instance, because teaching autonomous vehicles to drive themselves relies on training data recorded in the real world, solution providers must take care not to collect and store private information such as drivers’ faces and licence plate numbers.

The Cloudera Data Platform (CDP) offers a central repository that takes care of issues such as security, governance and authorisation. It also catalogues and contains the metadata (basically the data that provides information about the data being held), which helps a city comply with legislation such General Data Protection Regulation (GDPR). For instance, Personally Identifiable Information (PII) data that could be used to identify a person can be labelled and tagged and specific authorisation and access set.

> *“In a connected environment data is going to be being consumed by different audiences inside and outside of the city.”*

It constantly profiles new data and recommends security actions on data profiled to be personal. For example, it provides the ability to configure authorisation policies across groups, users, and data classifications, and includes various methods of masking data for specific users.

Hagen explains that the system provides end-to-end security and data privacy with data encrypted "in flight and at rest". The data lineage is also recorded and is accessible for governance and compliance. "For instance, if you've built a predictive model, you can go back across the data lifecycle and see all of the data that went into training the model and how it was used," he says.

Such a robust security framework is going to be demanded as projects scale-up. Ultimately, policies around data-sharing and data ownership are set by government agencies not technology providers but Cloudera and IBM see their role as providing a platform that enables policies to be easily enforced. "What we try to do is encourage our customers to anticipate and prepare for this scale-up of data-sharing because it's going to come and they need to start planning for it today," explains Douglas O'Flaherty, global ecosystem leader, IBM Storage.

As already stated, not all data are equal in terms of value and fragmented, sub-optimal data processes in the CAV sector can mean that data of little or no value enter the lifecycle. Cloudera provides the facility for it to be intelligently filtered out after the streaming phase and before processing. As O'Flaherty points out, in future, people are going to want to tag or label data as "interesting" or "not interesting". "We do a lot of work with autonomous car companies and they are looking for certain circumstances. They want to see where stop sign is obscured with trees," he says. "They want to discard the data of someone driving across an American Midwest highway for hours and hours when nothing happens. The intelligence in the Cloudera platform allows you to do this."

Similarly, searching and finding the right data is also a major issue in the CAV sector. The Cloudera Search facility enables machine learning engineers to programmatically drill down and find the data that are required to optimally train a model via "fine-grained" queries and provide results directly related to training activities. The system also enables organisations to tier their hot data (that which might need to be accessed immediately and regularly) and cold (which aren't accessed regularly).

The Cloudera and IBM partnership is a demonstration of the importance of technology providers working together to not just service clients but also enable sectors to progress and develop next-generation solutions. Recent years has seen increasing acknowledgement of the ecosystem approach and the CAV sector, which requires expertise across a wide range of disciplines, underlines the importance of such collaboration.

Ger points out that such is the complexity of data lifecycles in these areas that no single vendor has all of the capabilities to ensure the sector can scale up and fulfil its potential in the future. With this in mind, it is collaborating with a number of technology providers to define, implement and offer a data lifecycle platform enabling and optimising future connected and autonomous vehicle systems. Called Project Fusion, its fellow members of the ecosystem are:

Airbiquity: over-the-air software management
NXP: vehicle processing platforms
Teraki: edge data artificial intelligence
Wind River: intelligent systems platform software

The first application for the project is specifying a solution for intelligent vehicle lane change detection utilising synergistic technologies from each company. “The sector has come a long way already but up until now systems have been very proprietary, closed and brittle,” says Ger. “What we need is a pre-packaged/templated end-to-end solution that is open, scalable and replicable. It is incumbent on vendors to integrate their expertise and leverage best-in-class solutions.”

While speaking specifically about the CAV sector, Ger’s words should resonate in the broader context. Today, because of the sheer volume of data it produces, the CAV sector must confront a raft of data challenges head-on but others in the smart city ecosystem could find themselves facing similar ones in the not-too-distant future.

> *“It is incumbent on vendors to integrate their expertise and leverage best-in-class solutions.”*

Circling back to the future

The overarching value of data is that they provide tangible evidence of something and a reason to take specific action. At a micro level in an autonomous car, this could mean telling the vehicle to brake to avoid a pedestrian who has stepped out in the road. In itself, a critical life-saving action. Over the next decade, as autonomous vehicles continue to be trained and pilot and trial vehicles deployed, their data lifecycles will help to make equally critical but more macro-level decisions.

Whether it is feeding data into a programme to reduce congestion, interacting with road infrastructure to improve safety or, ironically perhaps, helping to reduce the overall use of cars in a city, they have the potential to play a vital part in the development of future cities.

It is important to emphasise that amid the excitement, the CAV sector is barely out of the driveway, let alone cruising down highways. There will always be ongoing learning. Currently, programmes are limited in scope but, as Hagen points out, "it's all in place to get started". "CAV technology is pertinent now because the technological foundations have been and are being laid for the promise of CAV in the future," he says. "Device availability, sensor technology, powerful software with small footprint, AI modelling and its lifecycle, storage efficiencies, computing resource management – we need to be thinking about what is available now and what we want to do with CAV technology in the future, so the promise can come to fruition."

Transport authorities and private sector companies are also exploring new ways for people to travel around, and to and from cities, with rollouts of new models such as transportation- and mobility-as-a-service (TaaS and MaaS), and they need to identify the part connected and autonomous vehicles play as part of the broader landscape.

Additionally, cities also need to be preparing for the glut of data that will be produced by CAVs and understand what this means to them. This, says Hagen, will require leadership and vision within each city. "As city governments are extensive and have many departments, identifying all of the stakeholders and their interests is important," he says. "Planning for coordinated efforts among departments will take a thoughtful approach. Establishing timelines and milestones will ensure each project keeps moving. Taking inventory of current resources early and proper sizing for future use based upon use cases will help cities plan for budgeting."

And he adds that as to how they take advantage of it, that will depend on a city's individual priorities: "For example, if the major factors for using CAV in cities are to better manage costs, improve safety, and provide efficient transportation for consumers, cities can use the data to improve fleet maintenance, accurately report revenue and propose budgets as well as improve traffic congestion and safety along prescribed routes."

Conclusion

Today, connected and autonomous vehicles are on a huge learning curve. AVs in particular are testing software engineers and data scientists as the mother of all machine learning use cases. In turn this means it will be us learning from the AVs in the future. Success in this application could ensure the pursuit of thousands of less demanding use cases, which is why the development of autonomous cars has implications across many different sectors, particularly smart cities. This makes it is even more important to ensure the right foundations are in place to integrate CAVs into the city's mobility landscape and to do so safely and securely. And we can't do this without advanced data management and analytics.

About Cloudera & IBM Power Systems
Cloudera was founded in 2008 by some of the brightest minds at Silicon Valley's leading companies, including Google, Yahoo!, Oracle, and Facebook. And in 2011, 24 engineers from the original Hadoop team at Yahoo! spun out to form Hortonworks. Both companies, who joined forces in January 2019, were founded on the belief that open source, open standards, and open markets are best. This belief remains central to our values, evidenced by our significant investments in engineers and committers working with the open-source community. Today, Cloudera has offices around the globe and is headquartered in Silicon Valley, California.