Data Garage

How autonomous vehicle training data can be recorded, pre-processed, stored and uploaded into the cloud

NTT Ltd.’s Technology Experience Lab

b-plus
IBM
Incenda AI
Lenovo
NTT Ltd.

NTT contact details

We welcome any inquiries regarding this document, its content, structure, or scope. Please contact:

Dominik Friedel
Business Development Manager
NTT Global Data Centers EMEA GmbH
dominik.friedel@e-shelter.com

Terms and conditions

NTT does not assume liability for any errors or omissions in the content of this document or any referenced or associated third party document, including, but not limited to, typographical errors, inaccuracies, or outdated information. This document and all information within it are provided on an 'as is' basis without any warranties of any kind, express or implied. Any communication required or permitted in terms of this document shall be valid and effective only if submitted in writing.

Contents

Introduction

At the end of the last millennium, very few people thought that the science fiction representations of self-driving cars would soon become a reality. But tremendously **higher storage capacities**, **rapidly increasing operation speeds** and **ever-expanding bandwidths** have paved the way for the new technologies such as Artificial Intelligence (AI) and the Internet of Things (IoT), which are now revolutionizing each and every sector of today's economy. Among these the autonomous vehicles industry is the area that is growing the fastest.

The very early 20th century, brought us motorized vehicles to replace horse drawn carriages. Today, in the 2020s, we are right in the middle of the **second great change in vehicle mobility**, mainly driven by digital technologies. These current developments contribute to a vision of: **"Seamless transportation in autonomous vehicles on interconnected road networks."** (Rajat Dhavan, McKinsey)

To realize this vision, technologies are evolving and being used in a different way, to automatically sense and depict the environment of the vehicle and to move it safely with little to no human input. As Harald Krüger, former CEO at BMW put it: **"The key to the mobility of the future is openness to various technologies."**

The core technology for the development of autonomous vehicles is AI. Decisions that have until now been made by the driver of a car, will be taken by a computer. These decisions are implemented within the vehicle's **Advanced Driver Assistance System (ADAS).**

Before an autonomous vehicle can be road ready manufacturers are required to train their ADAS with an **enormous amount of real environment data**. All of which needs to be collected beforehand via large fleets of sensor-equipped test vehicles over many journeys. To make these masses of data accessible to all relevant stakeholders, it must be uploaded to ever-present storage and cloud infrastructure. As high data quality is essential for efficiently training the ADAS, the data is ideally pre-processed by intelligent software algorithms running on high-performance servers.

But how and where can those tasks be handled? The only place where these huge amounts of data can be ingested, processed and uploaded to the cloud with reasonable speed is a **colocation data center**, because it provides the necessary highly secure on-premise environments with direct connections to the Points of Presence (PoP) of all major public clouds.

The process of collecting, transferring, processing and uploading data is the essential use case for every developer of an ADAS. Taking an end to end approach to building up the complex chain of technologies and solutions, the **NTT Technology Experience Lab** has brought together five highly experienced technology partners into the Data Garage project.

The Data Garage

Data Garage – project overview

As with every new technology stack, there is a learning curve; especially in terms of the various hardware and software components required for autonomous driving, there is a need for a dedicated and safe environment for testing new scenarios. The NTT Technology Experience Lab is situated in NTT colocation data centers worldwide and provides clients and prospects an easy-to-use platform for trying out and validating innovative technologies and processes. More than 140 partners are offering their technologies for business and industry in the NTT Technology Experience Lab. Pre-installed use cases, as well as fast testing capabilities are available for exploration and trial without further obligation.

For the Data Garage project, the NTT Technology Experience Lab successfully recruited five experienced partners to set up a dedicated use case for autonomous driving.

- **b-plus**: Data recording on the road
- **IBM**: Data storing on premise
- **Incenda AI**: Intelligent data processing
- **Lenovo**: Hardware systems for data storage and AI
- **Global Data Centers, a division of NTT Ltd.:** Data transfer and upload into the cloud

The **Technology Experience Lab** offers a flexible and secure environment within the NTT colocation data center FRA1 in Frankfurt am Main, where the project partners can deploy and operate hardware and software systems, and where they can connect to cloud services with direct access. By contributing their expertise while intensively cooperating in the Data Garage project, the partners have built a ready-to-use infrastructure in the Technology Experience Lab, as described in the following chapters.

Data recording on the road

Developing, testing and validating automated driving functions requires data from real driving tests throughout the entire development process. In order to glean maximum insights from the prototype phase of the control units and automated driving platforms, the data collection begins as soon as the sensor technology has been determined. Essential components of the system architecture include sensors for environment detection such as cameras, radars and LiDARs – the number and resolution of which is constantly increasing due to rising requirements. This results in a steadily increasing amount of data. The sensors in combination with GPS form the basis for decision making in the vehicle. Experience tells us that the data volume of all sensors in the latest generations accumulates to up to 6 GByte/s or 48Gbit/s. These data streams must be decoupled, processed and stored.

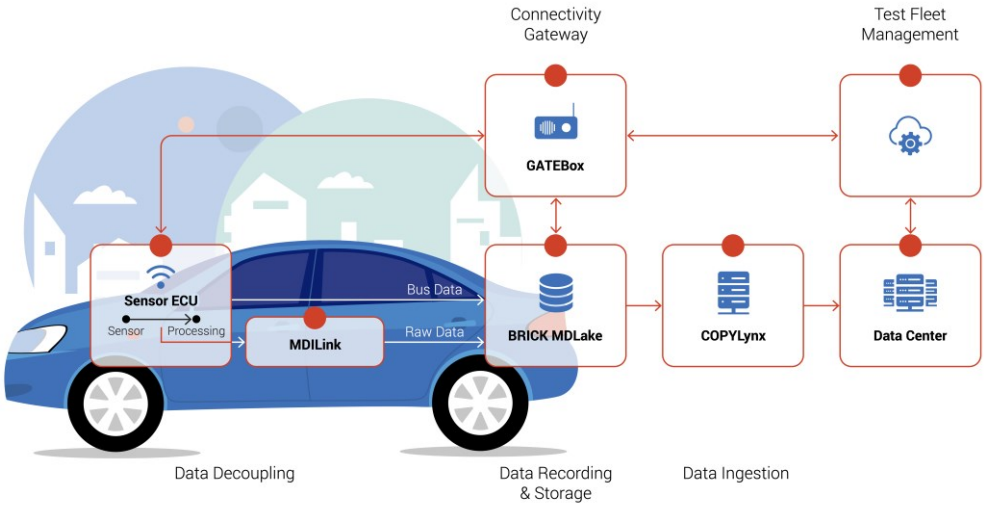

Figure 1: Function blocks of the b-plus AVETO - Automotive Validation Toolchain during data acquisition in the vehicle

Vehicle setup

The setup and operation of test vehicle fleets is a technically complex task, as the very limited space available in the vehicle and the demanding environmental conditions present challenges for materials as well as electrical/electronic and mechanical systems. At the same time, the safe and smooth operation of high-performance equipment must be ensured in this environment. Therefore, the implemented measurement components must be as compact as possible, able to withstand temperature fluctuations during operation and highly resistant to vibrations and shocks.

A further point is the fluctuating and limited power supply of the vehicle's on-board network. This can be solved by efficient components and intelligent power management in the test vehicle. The selection of the right components, an efficient setup and smooth operation must be ensured here.

Data extraction and conversion to measurement technology

The raw sensor data required for development must first be extracted from the sensor front-ends. MIPI CSI-2, with sideband signals such as I²C, is mainly used as an interface. If the distance between the sensor and the next processing stage is larger, the data is converted using serializer / deserializer components. These are specifically configured and often converted to 10 Gbit ethernet using so-called 'measurement converters', such as MDILink, in order to be recorded by a data recorder.

This is an essential step moving from the world of ECUs to the realm of measuring computers.

Precondition: the quality of the data

In order to gain knowledge from various sources, there are two essential preconditions that must be met in the measurement technology tool chain that is already in the vehicle. The data must be available for further analysis and processing with *integrity* and *temporal correlation*.

The integrity of the data plays a major role in this respect. Consider ISO26262, the standard for functional safety: it also requires that the effects of the measuring devices, such as latencies, must be known. For this purpose, b-plus uses mechanisms such as CRC checksums to ensure the safe transmission of data streams. This ensures that bit errors, for example, are avoided.

The synchronization of the clocks present in the measurement setup is of great importance. This makes it possible to set exact time stamps in the whole system on the individual data packets and to guarantee an assignability when fusing sensor data to the environment model. This can be done via ethernet using the IEEE 802.1 AS standard, for example. In addition, the input interfaces of data recorders must be assigned a hardware time stamp in order to detect transmission delays. One example for the implementation is the b-plus XTSS Time Synchronization Service.

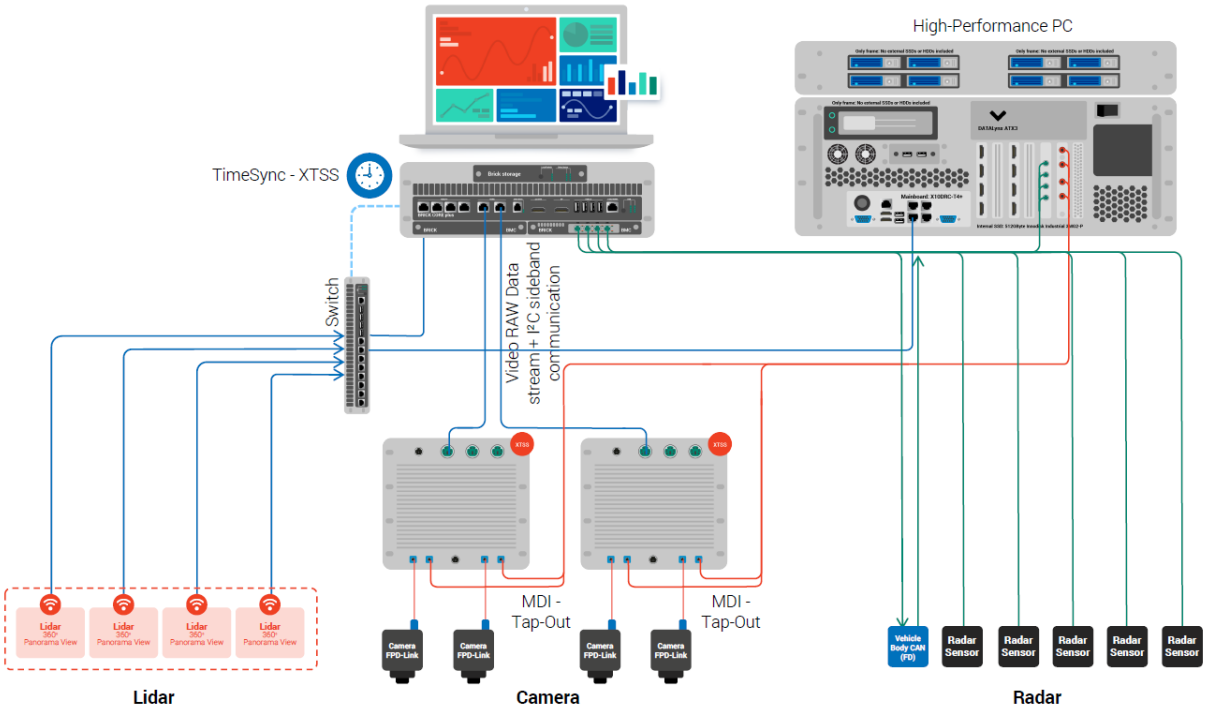

Figure 2: Typical data recording architecture (BRICKplus + MDILink) including high-performance PC (DATALynx ATX4) for intelligent recording in the vehicle.

Intelligent and efficient recording

Especially given the above-mentioned quantities of data, it quickly adds up to terabytes of data that have to be fed into the data center. In the simulation and analysis of the data, however, many of the scenes that are used for the function development in the algorithm are redundant and less interesting. Until a few years ago, OEMs and tier1 sensor manufacturers always wanted to run in as much data as possible to obtain a broad database. The approach was based on the motto that "every bit is gold". This resulted in many petabytes of data being collected – of which the actual value was not known and many scenes were redundant. For example, the exact sensor configuration, the firmware versions of control units or the configuration of the measurement technology would be missing. It is therefore very laborious to label this data afterwards, or to check it for valuable scenes. At the same time, it is not obvious how many of the valuable corner cases in the vehicle were in the recordings.

Therefore, the complementary approach should only include data that is explicitly searched for. The test can already be carried out in the vehicle by intelligent identification and pre-labelling and, if necessary, filtering while driving.

This requires live data processing of the enormous sensor streams, which in turn requires the use of high-performance computers in the vehicle. This prerequisite is met with powerful vehicle-compatible, high-performance computers such as DATALynx. It includes several powerful server processors and graphics cards for highly parallel image processing and the operation of AI in the vehicle to decide if and when the recorder should record the scene.

Considering the operation and management of a test fleet as a whole, another aspect comes into play: a test drive must be planned in advance, especially with regard to measurement

technology and ECU configuration. These parameters are also subject to high dynamics, even in the course of a test drive – for example, a new firmware version from the development department must be brought to the testing as quickly as possible.

An intelligent and cloud-based test fleet management system supports the development process (see figure 3).

Updates for control units and measurement technology can be rolled out via a mobile-networked overview of the test fleet. In addition, an up-to-date overview of the status of the measurement technology is possible. A dashboard informs the driver about the performance of his system in the trunk.

During real world test drives campaigns are used. Typically, several of driving tasks are defined before the journey, which are continuously checked and, if necessary, updated during the journey. In summary, intelligent filtering of the data per vehicle and tool-supported operation of the entire test fleet leads to more efficient driving data recording overall.

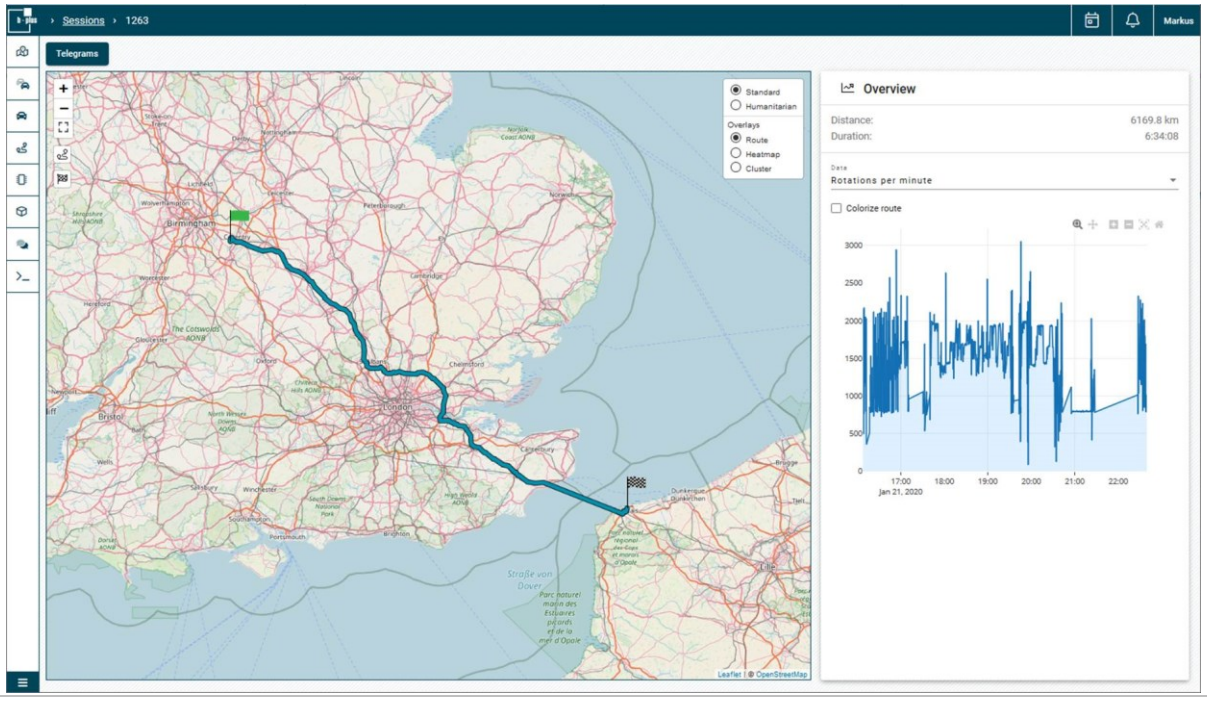

Figure 3: Overview of a test drive in the b-plus connectivity solution

Data storage and data management in the test vehicle

In the typical test vehicle for automated driving of stages 3 to 5, several data loggers are already provided for some aspects. Their memories are read out individually after the test drive. Here, the task of an OEM or tier 1 is to make data management efficient. Test data should be quickly available for analysis and simulation. There are several ways of going about this, which have to be differentiated depending on the case. Considering the data storage need, it can even make sense for larger tasks to establish an edge data lake (see figure 4 with MDLake) in the vehicle. With such a large data memory, the effort to collect a single data point is reduced drastically, because the data is already concentrated on a data lake in the vehicle. It is also important that existing data loggers can be easily integrated.

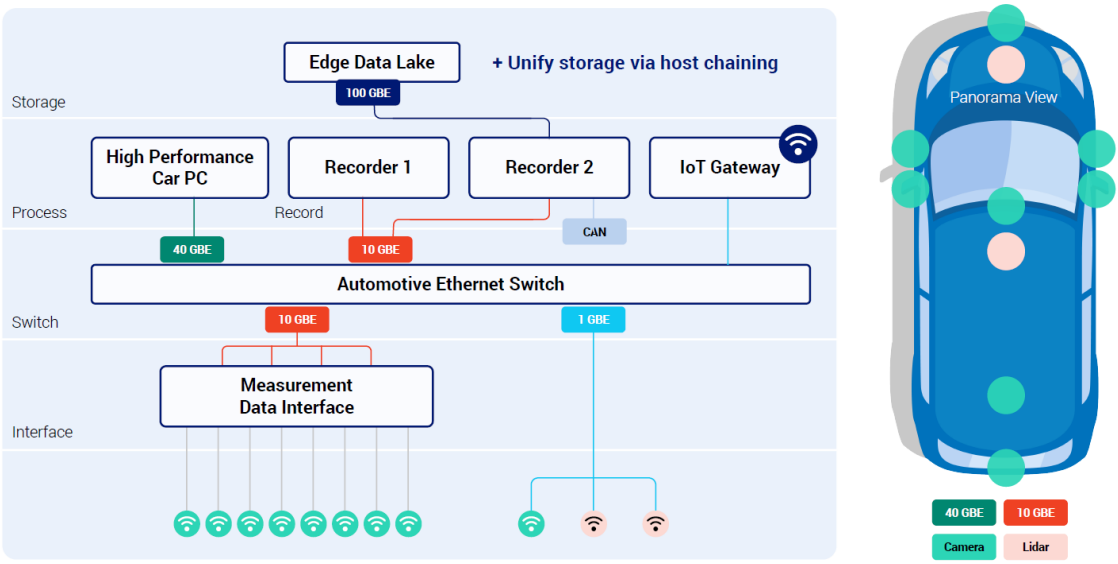

Figure 4: Unifying storage using a mobile data lake

Ingestion into the data center

There is still the hurdle of bringing the measurement data from the vehicle to the data center. This so-called ‘data ingestion’ task is difficult for many developers: to get the daily occurring gigabytes backed up and quickly out of the vehicle.

There are two main concepts for this, depending on whether there is a fast connection to the data center available at the vehicle’s location. If a wireless connection with sufficient speed (128 Gbit/s) between vehicle and data center is available, the measurement data can be transferred directly. If such connection is not available, removable storage media such as the cartridges in BRICK recorders can be detached from the vehicle and dispatched via a logistics service provider. The storage media will be received by remote hands colleagues within Global Data Centers for ingesting via copy station in the according environment. In both cases, such copy stations for data ingestion are available. Systems such as COPYLynx take the logging memory completely and automatically copy it – pre-configured – at high speed to the respective memory target. This requires a flexible concept for efficiently bringing the test drives to where the data is needed, whether that be network shares or by feeding into cloud storage or, in some cases, by copying to other media with further transport by mail.

During the feeding process, data can be checked for integrity or prepared in pre-processing. NTT's colocation data center enables clients to bridge the gap between regionally available data centers and global cloud infrastructures.

Data storing on premise

IBM Cloud Object Storage (COS) software on Lenovo ThinkSystem server

IBM COS serves as the underlying infrastructure for both on-premise and public cloud deployments about IBM's object storage offerings. Lenovo ThinkSystem designs are based on x86 industry standards and are innovative beyond the ubiquitous 1U and 2U servers that dominate the low-cost, industry-standards, no-frills server world.

Why Cloud Object Storage?

Object storage is characterized by access through RESTful interfaces that have granular, object-level security and rich metadata that can be tagged to it. Object storage products are available in a variety of deployment models, e.g. virtual appliances, managed hosting, purpose-built hardware appliances, or software that can be installed on standard x86 server hardware. On-premises object storage is designed for workloads that require high bandwidth and optimized costs. The new generation of object storage products like COS relies mainly on erasure-coding schemes that can improve availability at lower-capacity overhead and cost when compared with the traditional RAID schemes. Object storage provided three key characteristics from the start:

- Shared access to data (S3 API)
- Heterogeneous computing
- Dynamic scaling without interruption

IBM COS has configurable erasure coding, object-level ACLs, built-in quota management, and introduced WORM support through proprietary extensions to COS's implementation of the Amazon S3 API.

IBM Cloud Object Storage architecture

The Cloud Object Storage system is deployed in multiple configurations as shown below. Each node consists of Cloud Object Storage software (ClevOS) running on an industry-standard Lenovo server. Cloud Object Storage software is compatible with a wide range of servers from many sources, including a physical or virtual appliance. The three types of nodes, as shown above, are:

- IBM Cloud Object Storage Manager
- IBM Cloud Object Storage Accesser
- IBM Cloud Object Storage Slicestor

Each Cloud Object Storage system has a single manager node, which provides out-of-band configuration, administration, and monitoring capabilities. There is also one or more accesser nodes, which provide the storage system endpoint for applications to store and retrieve data. There are three or more slicestor nodes, which provide the data storage capacity for the Cloud Object Storage system. The accesser is a stateless node that presents

the storage interface of the Cloud Object Storage system to client applications and transforms data using an information dispersal algorithm (IDA). slicestor nodes receive data to be stored from accesser nodes on ingesting, and they return data to accesser nodes as required by reads.

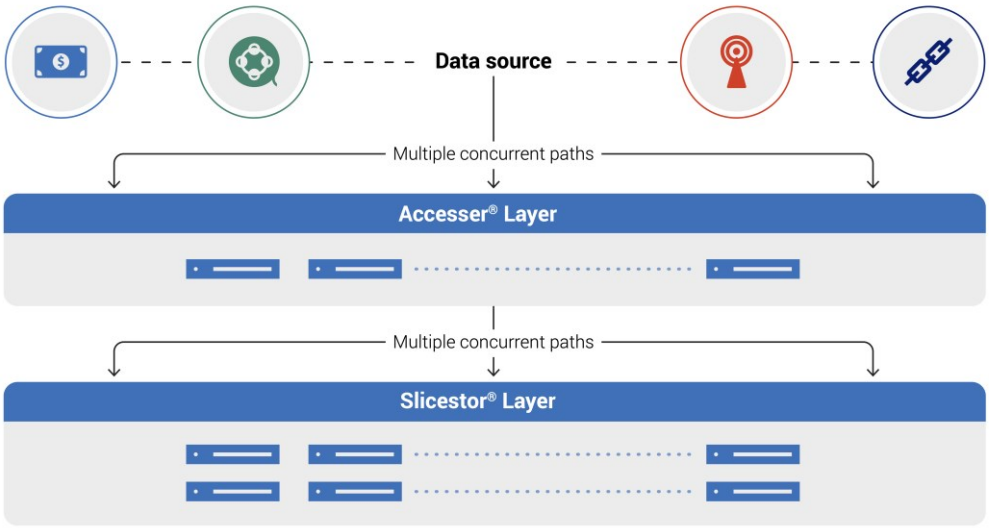

Figure 5: IBM Cloud Object Storage architecture

The IDA transforms each object written to the system into several slices, such that the object can be read bit-perfectly using a subset of those slices. The number of slices created is called the IDA width. The number required to read the data is called the IDA read threshold. The difference between the width and the read threshold is the maximum number of slices that can be lost or temporarily unavailable while still maintaining the ability to read the object.

IBM Enterprise Data Pipeline for AI

The productivity of the ADAS/AD data science team depends on the ready availability of the latest AI development frameworks, optimal computing power of central processing (CPU) and graphics processing unit (GPU) computing power and data accessibility. While performance is important, it is not the only consideration. Data preparation and ingestion can consume most of the AI development timeline. For each project, data must be extracted from multiple other sources and properly organized so that it can be used for best model training. Once a model is developed, the data must be retained for traceability.

The value of data grows with diverse use across multiple users, systems and models. Data scientist productivity depends on the efficacy of the overall data pipeline as well as the performance of the infrastructure used for running AI workloads. IBM delivers a comprehensive portfolio of software defined storage products that enable clients to build their enterprise data pipelines with the right performance and cost characteristics for each stage.

A new addition to this portfolio is IBM Spectrum Discover, which is metadata management software that provides data insight for petabyte-scale unstructured storage. IBM Spectrum Discover easily connects to IBM Cloud Object Storage (COS) and IBM Spectrum Scale to rapidly ingest, consolidate and index metadata for billions of files and objects. IBM Spectrum Discover plays a leading role in the data classification phase of the overall data pipeline, but it also provides capabilities that support data governance requirements and enable storage optimization along the pipeline.

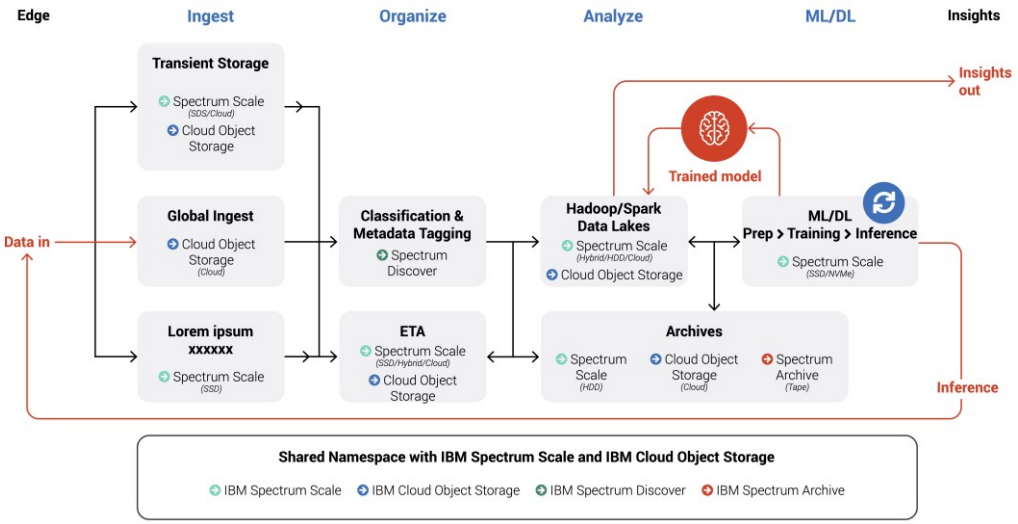

Figure 6: IBM Enterprise Data Pipeline for AI

IBM Spectrum Scale for AI

IBM Spectrum Scale is an industry leader in high performance parallel file system software. A key ability that Spectrum Scale provides is a single namespace (or data plane) so that each data source can add data to the repository using NFS, SMB, Object, HDFS, or a POSIX interface. Another key strength is that Spectrum Scale enables data to be tiered automatically and transparently to and from more cost-effective storage, including NVMe/SSD/HDD, tape and cloud.

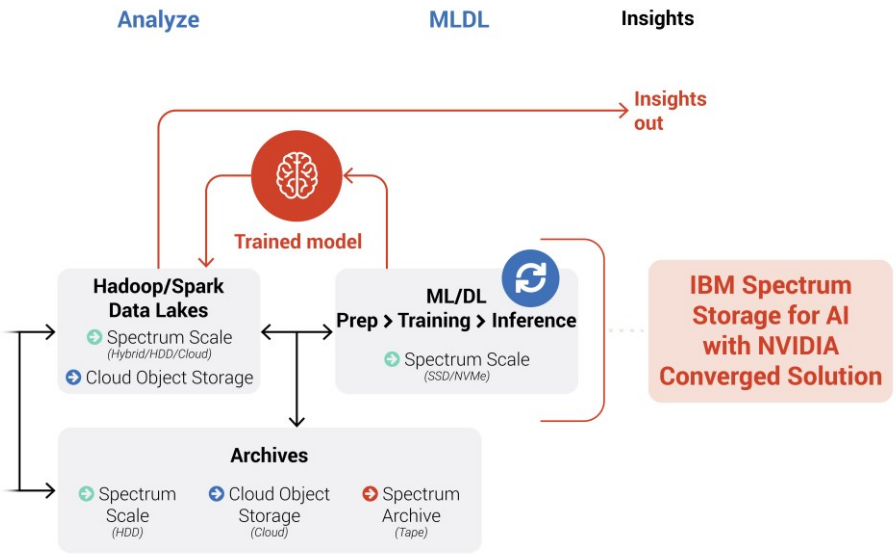

Figure 7: Scaling with GPUs

Scaling with GPUs

NVIDIA has led the AI computing revolution, leveraging the power of the modern GPU with its massive processor core count and parallel architecture, uniquely suited to the massively parallelized operations that are core to DL, and which exceed the limitations of traditional CPU based architectures.

Intelligent data processing

When realizing an AI project, it is essential to highlight the whole AI development process chain (see Figure 8), starting with defining the project specifications and ending up with a robust AI model, where dealing with data is one of the most challenging parts. First, during the data collection, it is important to pass only appropriate data points to the labeling. Data annotation is costly and the whole effort should be spent solely on reasonable and useful data points. After labeling, data quality assurance is required to generate a high-quality dataset – without having any annotation errors – to be used for training. These two steps are Incenda AI's business focus and are essential to enable safe, reliable and robust AI applications.

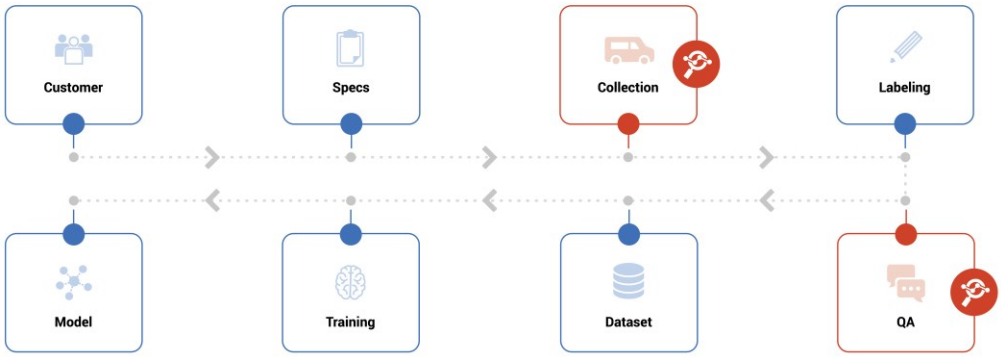

Figure 8: AI development process

Intelligent data collection

To extract useful and representative data from all the recorded parts is a time-consuming process, because a large amount of data from all the different sensors must be scanned and evaluated towards the existing dataset, done mostly by humans drawing on their expertise. This can be more efficient and go faster if the data scientists and AI engineers are supported in the data selection activity. Therefore, a kind of post-processing, which takes place after the recording, is required. The recorded data points are processed by (AI) algorithms, and meta information is extracted covering the following use cases to enable enhanced data filtering.

Overview of the use cases

The data post-processing is divided into three main use cases:

1. Data tagging

   The recorded data is analyzed and a defined set of meta information is created and added. Just to mention a few examples, tags can be weather conditions, driving environment (highway, city), crossing situation, or traffic participant appearance.
2. Edge-case detection

In large data recordings, it takes a lot of effort to spot special cases that occurred during the ride. With this implementation, the so-called “edge cases” – like wheelchairs, unusual cycles (e.g. rickshaw) or animals – on the streets are detected.

3. Smart data selection

   A lot of data is more similar to each other during a recording ride while driving around in the same city or even in various cities in the same region. Additionally, this kind of similarity applies also to the existing labeled data used for training. Knowing this, two questions come up:

   a) Which data within a recording, in relation to the remaining data, is more beneficial or interesting?

   b) Which data within a recording is more beneficial to be selected based on the existing labeled data?

      This method calculates a kind of similarity grade for each data point that helps to choose valuable data points for the labeling. This algorithm needs the existing dataset to be analyzed as the input.

All three use cases can be applied to single data points as well as to sequences. Furthermore, the dataset benefits from diversity, first while preventing bias and secondly while not including redundant / unrepresentative data records. The target is to realize a well-balanced dataset with a discrete uniform class distribution while maximizing the cost-efficiency. Figure 9 summarizes the “Intelligent Data Collection” process bundled in a corresponding framework.

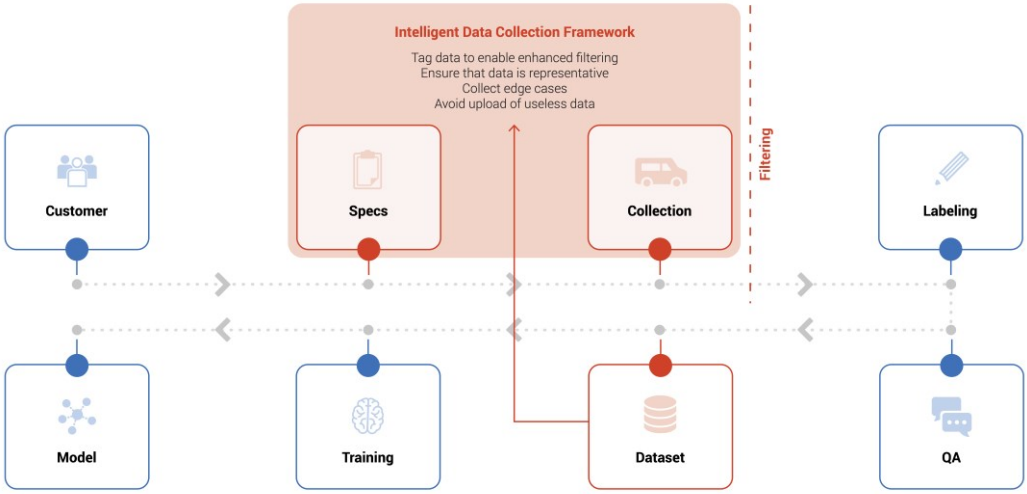

Figure 9: Intelligent Data Collection framework

The algorithms integrated in the framework are fundamentally based on deep neural networks and require an accelerated GPU processing to achieve a proper runtime performance. To adapt and scale the compute power accordingly, Cloud Computing with enhanced GPU instances are perfect for this computing challenge.

Optimized data upload

In addition, these algorithms can be integrated directly in the car, provided that enough computing power is connected to the data recorder. It must be adapted to the lower amount of computing power available to the cost of algorithm accuracy. This can be seen as a pre-processing stage performed in the car. This moves to an online Intelligent Data Collection – recording only required data and optimizing upload to the storage, in contrast with collecting and uploading all data points. To complete the picture, the algorithms can also be processed in the copy station for data ingestion, if appropriate computing power is available to decide which data points should be uploaded.

Architecture

The recorded data is uploaded to the cloud either in raw or standard data format. From there it is read and ingested into the AI compute instance, where the Intelligent Data Collection algorithms are processed. If raw data streams are used, a conversion to e.g. RGB/YUV (camera) or LAS (lidar) must be implemented. The results and meta information are pushed back into the cloud object storage using references to the original data points (Figure 10).

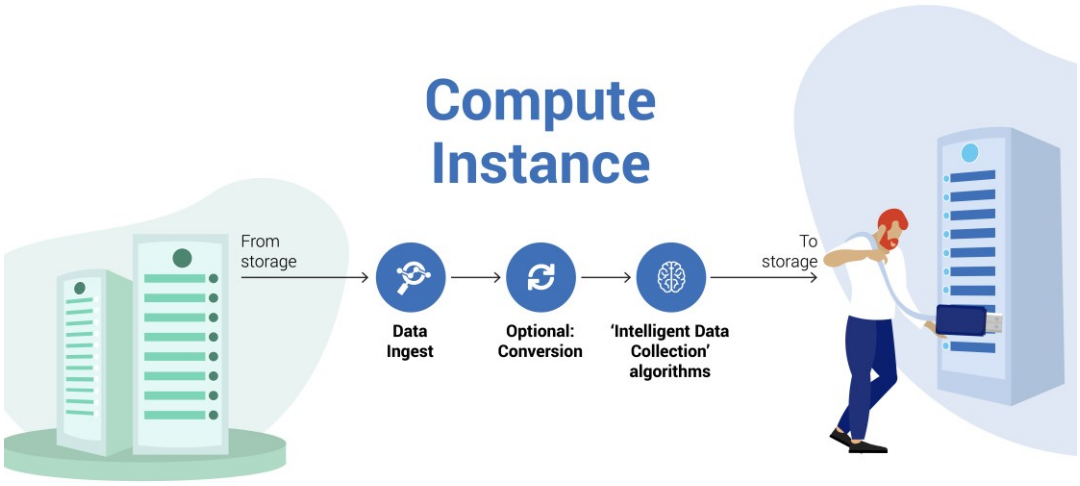

Figure 10: High-level SW architecture, AI compute instance

Based on the meta information, a visualization can be implemented for enhanced filtering and data insights. As the whole data selection process is accelerated, it becomes more transparent and supports the engineers in their choosing of appropriate data for the labeling process.

As mentioned, this architecture is more general and can also be adapted to run on the recorder or copy station. Based on the relevance of the data, the algorithm generates a trigger signal to record or upload relevant data points only.

Data quality assurance

Due to the strong dependency between the AI model's performance and the training dataset used, a separate process step to perform data quality assurance (see Figure 11) is, in parallel to the annotation process, highly recommended. No matter whether the data will be labeled in-house or by external parties following a human-based or automated labeling approach – the pain point remains the same. The vast amount of data needs to be annotated consistently in alignment with the specifications, while in practice formal errors, corner cases and a scope of discretion blurs the labels. Formal or logical issues must be detected, and any label inconsistencies based on the client requirements or specification violations must be identified. As can be seen in Figure 11, the QA information must be provided to the labeling team to revise the annotations. In a next iteration, the corrections must be checked again. If all QA tasks are passed successfully, the high-quality dataset with a detailed report and statistics is provided to the client.

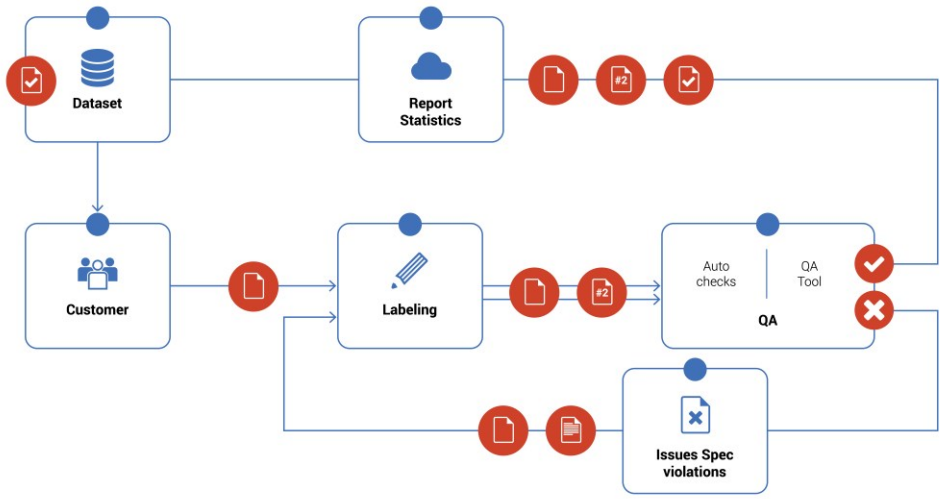

Figure 11: QA process

As a conclusion, two important points must be fulfilled for getting a high-quality data set:

1. Appropriate and useful data content, supported by Intelligent Data Collection
2. Correct and proper annotated data, supported by Data Quality Assurance

These two pillars are the enabler for a safe and reliable AI.

Hardware systems for data storage and AI

Data is stored and processed in the cloud

As described in the previous sections, the data that was recorded on the road finally goes to the data center, where it is stored and further processed. The focus is usually on the software and the processes that make use of the data and help to create the business value, which in the case of the Data Garage are, finally, the trained AI models for autonomous driving.

On the hardware side, the silent expectation is that performance and reliability are a “given” and that this piece of the puzzle “just works somehow”. This presumption is often driven by the ease-of-use that we experience with cloud computing in our daily life, e.g. smartphones using cloud storage for photos, or apps storing and analyzing health data from fitness trackers in the cloud.

But in the end, all that data is stored on disk drives and the software and AI algorithms run on processors and/or accelerators like GPUs – so there is really no way around getting your hands dirty on hardware.

In this section we describe the differences between public and private cloud environments, the reasons to go for a private cloud environment, and the hybrid cloud as a combination of both approaches. We also show why working with a strong hardware partner in a private cloud environment is important for guaranteeing the outcome – or business value – in the desired schedule and quality.

Why cloud storage and cloud computing are so popular

The popularity of clouds has increased over the last couple of years with the rise of public cloud services from several providers. There are good reasons for this. Cloud is an innovative model for using and managing IT resources, which offers significant advantages over traditional centralized IT approaches. Cloud benefits can be observed from two different perspectives:

- From the user’s point of view, it is an IT delivery and consumption model that provides resources, applications and services very quickly and with ease of use, either on-demand or in a subscription model
- From the provider’s point of view, it is an IT operating model that can easily adapt to meet changing application and service requirements and is therefore attractive for their clients

Private, public and hybrid cloud

In general, cloud resources can be provided in the form of a public or a private cloud. The term ‘cloud’ is often associated with public clouds; their advantages as compared to traditional centralized IT systems are their easy access to resources with a user-friendly web-interface, fast creation of resources like virtual servers and storage, and user-friendly pay-per-use billing models.

Over time, clients have experienced that there are also disadvantages to public clouds, like vendor lock-in (it is easy to get into the cloud of a specific provider, but a lot harder to get a grown eco system back out of it again), the sprawl of operational expenditure (OPEX) – such

as dispersed credit card payments for an ever-increasing number of virtual servers used in various departments – limited flexibility (to the hardware and standards that the cloud service provider defines), security concerns, and challenges to meeting regulatory requirements.

Private clouds provide the same ease-of-use and operational advantages of public clouds, but as they are under the full control of the clients, they can mitigate the disadvantages of public clouds. Hybrid cloud approaches combine the strengths of private and public clouds by providing a seamless end-user experience. A common usage model is to keep the daily workload in a private cloud and employ the public cloud to cover peaks in resource demand for less critical workloads (e.g. workload with low bandwidth, security and regulatory requirements).

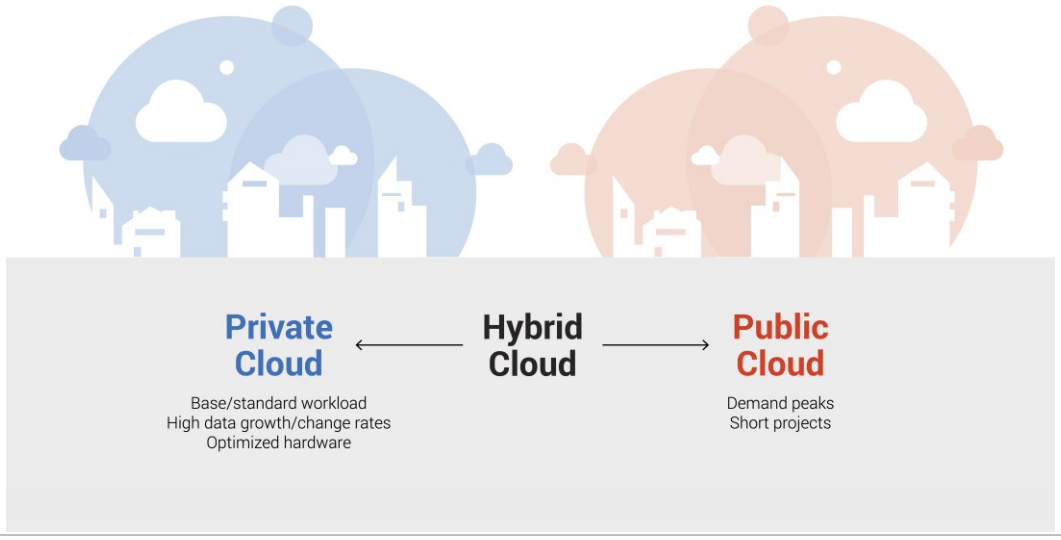

Figure 12: Private, public and hybrid cloud

Cloud usage in the Data Garage project

In the context of the Data Garage project, the data is transferred via a copy station for data ingestion into the private cloud environment. The advantage of this approach is that the ingestion process can run at the high network bandwidth that is available inside the colocation data center, which speeds the ingestion process up while ensuring that valuable network bandwidth in wide area networks (WANs) is not consumed by huge data transfers. The data is written to the IBM Cloud Object Storage, which runs on Lenovo ThinkSystem servers that are optimized for storing vast amounts of data with high IO performance.

Once the data has arrived on the private cloud storage, it is further processed to add value to it, e.g. by tagging the data or identifying corner cases in the scenes that are important for a broad training of the autonomous driving AIs. For this computation-intensive work, servers with GPU acceleration provide the capability to return results quickly with high throughput. The direct, high-bandwidth network connections of those servers to the private cloud storage is very important for this process. The servers need to receive a constantly high-bandwidth stream of data from the storage, otherwise they will just sit and wait for the data – and that is inefficient.

Considerations for selecting a private cloud hardware partner

Using a private cloud environment in a colocation data center makes a lot of sense for an efficient and controlled process chain. But if the x86 and GPU hardware all standard anyway, why bother about which hardware partner to work with? From a high level, this statement seems to make a point. But it overlooks some questions:

- **How to optimize the total solution for a quick return on investment?** The project is not just some fancy technology evaluation; in the end there needs to be business value, and that is tied to timelines and cost. For this, you must make the best use of the hardware you have: optimize every component of the stack, from the architecture to the hardware and firmware on up to the software levels. An experienced hardware partner has the experience to provide help and guidance. The Lenovo Datacenter Group has a long history of providing complete high performance computing (HPC) and AI systems that are tuned to high IO bandwidth and computing power. More than one third of the world's largest supercomputers are built on Lenovo hardware (www.top500.org) and benefit from the long experience of our experts. Lenovo shares its experience in Briefing and Benchmark Centers with its clients and can provide services for the implementation.
- **How to avoid project delays caused by server downtimes?** Downtimes of hardware are mitigated to some degree by upper software layers. For example, if one storage server fails in an IBM Cloud Object Storage environment, the data is not lost and can still be accessed. But losing hardware always has an effect: even if the issue is mitigated by software, there is still a performance impact, and too many hardware failures at the same time may not be covered by software-level redundancy, leading to an outage. Another issue can be special parts like GPU accelerators, which in the meantime have disappeared from the market and can no longer be replaced. For the last seven years, in the independent Information Technology Intelligence Consulting (ITIC) survey of over 1200 businesses worldwide, Lenovo ThinkSystem servers proved to have the best uptime of all Intel x86 servers. And if a part fails, it is backed up by a parts supply-and-support infrastructure that provides the service level that clients need.
- **How can I benefit from innovations in storage, processors and accelerators?** Advancements in new storage devices, processors and accelerators can help to further improve efficiencies in the process. But all those components must be integrated into hardware systems, and that is not always as easy as it sounds. There are new challenges coming up, like dramatic increases in processor and GPU accelerator performance. The power consumption of a single high-end CPU or GPU could increase in the next years to more than 300W. The challenge will be to efficiently cool this extreme power density. Lenovo is an innovation driver in hardware designs. For example, Lenovo has made huge investments in direct water-cooled systems that can manage extreme power density in a standard 19" rack form. For the upcoming generation of servers, Lenovo has partnered with NVIDIA to co-develop a new direct water-cooled system with NVIDIA GPUs.

Extending the cloud out to the user and to the edge

When looking beyond the core Data Garage project on a potential process chain, one needs to consider the end users and edge computing to complete the picture.

Users will need access to the private cloud, for example with powerful workstations used for post-processing data in performance-hungry virtual 3D models. For mobile workers, the same 3D graphics power can be provided by servers in the data center, which create a remote desktop environment: the power-hungry 3D processing can run in the data center, while the end user has the full visual experience on a more lightweight end-user device like a Thinkpad / Notebook.

The data-recording car will usually come across places like hotels, shops or highway restaurants. In an ideal world, data could be pre-processed and uploaded from those locations, e.g. for pre-selecting certain scenes that are specifically interesting for AI training. One challenge with that idea is that the systems that are well suited to operating in rugged environments with 220V power plugs are usually not very powerful. Edge servers are specifically designed to run in such locations and still provide data center server performance, including GPU processing and LTE broadband connections.

Lenovo provides hardware, service and support across this whole process chain from the edge to the data center through to the end user.

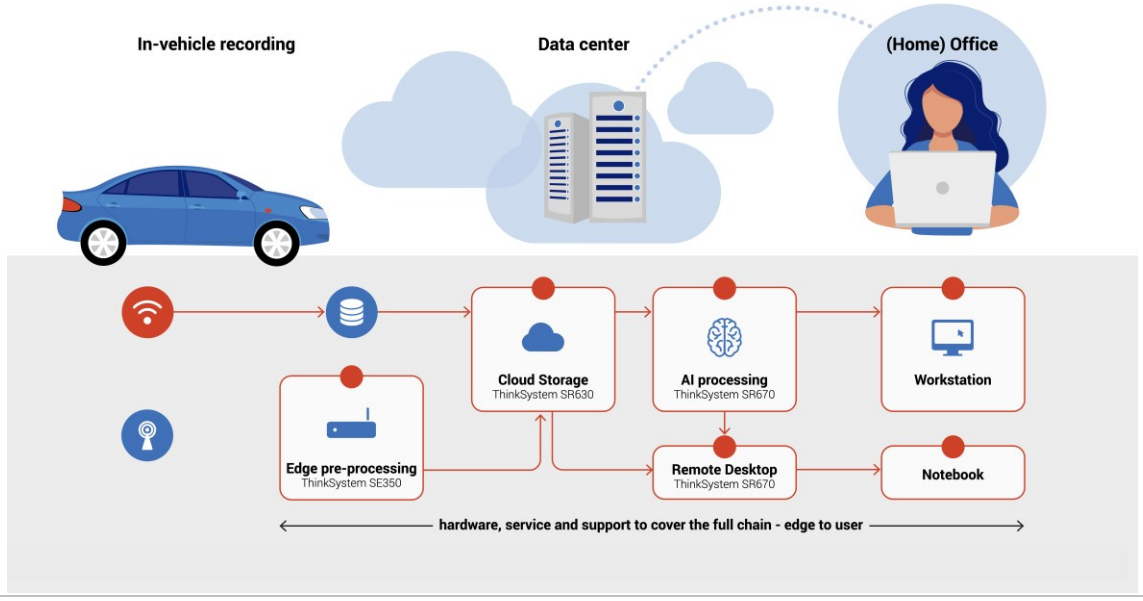

Figure 13: Process chain and the hardware involved

Lenovo offers a comprehensive range of servers with the ThinkSystem family, including rack, tower, edge and blade form factors. The product line includes single-socket through to eight-socket systems, with the latest processors from Intel and AMD. This matrix provides a comparison of key features of all the Lenovo ThinkSystem servers: https://lenovopress.com/lp1263-lenovo-thinksystem-server-comparison

Data transfer and upload into the cloud

The colocation data center as the foundation for data transfer

The masses of recorded environment data for autonomous driving applications need to find their way to the makers of autonomous vehicles, who require the data to train the ADAS. Thus, the data must be offloaded from the vehicle's data cartridges and uploaded to globally accessible cloud infrastructures.

To enable the fast and secure upload of data, a direct connection to the major cloud services is absolutely essential. Colocation data centers build the bridge between tangible physical systems and globally accessible public cloud services. By offering its multi-service interconnection platform, NTT colocation data centers provide state-of-the-art connectivity to all major public cloud services, as well as to other partners within the NTT and the NTT Technology Experience Lab. The colocation data center is thus the core platform of the Data Garage project.

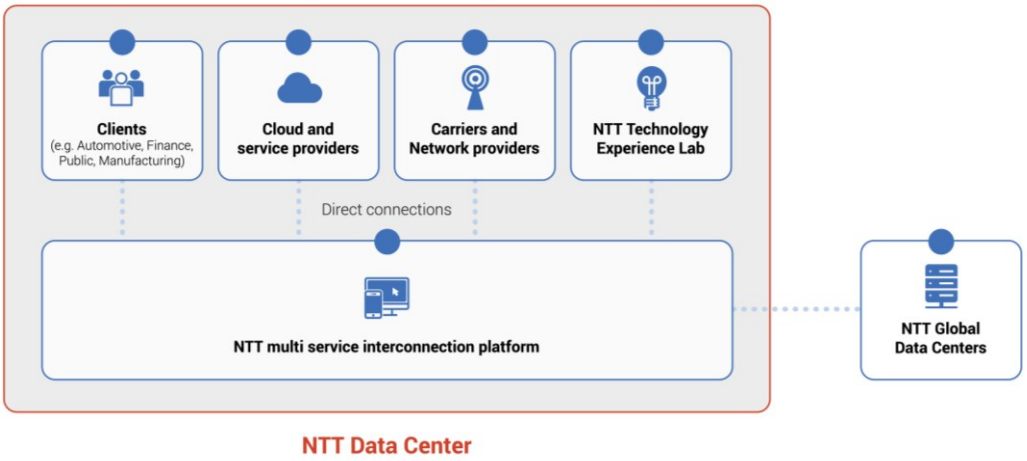

Figure 14: The colocation data center with its connectivity platform as the foundation for data transfer and cloud access

What is the NTT multi service interconnection platform?

NTT's multi service interconnection platform is the central switching point for the networking requirements across all data centers and is supported by NTT's managed carrier grade ethernet network. Access to the multi service platform is provided through virtual ports that can connect easily to virtual services and gain direct access to a wide variety of internet access points and to leading cloud providers. The multi service interconnection platform also provides dedicated, high-performance, private layer 2 connections, to directly connect clients to public cloud services, NTT offers a managed service based on redundant dedicated connections to the leading cloud providers independent from the public internet such as:

- Amazon's AWS Direct Connect
- Microsoft's Azure ExpressRoute
- Google's GCP Direct Connect
- Alibaba's Cloud Express Connect
- Oracle's Cloud FastConnect
- IBM's Cloud Direct Link

Transferring mobility data to the virtual world

Initially, the recorded environment data is stored on the previously introduced b-plus data cartridges. To further process the data, it needs to be offloaded via a copy station into a data center environment. Within the test scenario, both the copy station and the hardware for data storing and processing are implemented in the Technology Experience Lab at the NTT's campus FRA1 in Frankfurt am Main.

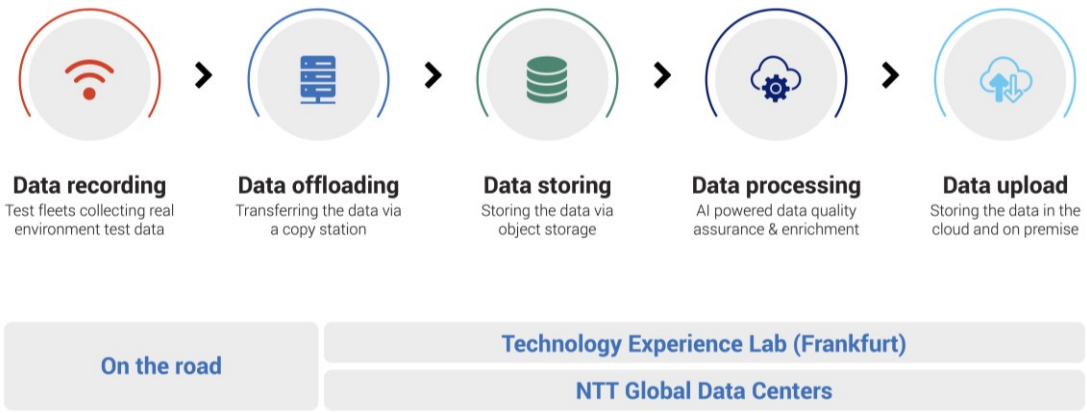

Figure 15: The data garage concept and the allocation of data

For productive operations, the copy station will be located in the clients own secured data center environment. To ensure smooth operation, a qualified and certified remote hands service colleague will insert the received data cartridge in the respective copy station of the client. The data will be transferred from the copy station into an object storage (as previously described) to subsequently process the data with the help of Incenda AI's intelligent software algorithms, both deployed on Lenovo hardware systems. Afterwards, the processed data will be uploaded to the cloud, which could be either a private cloud infrastructure set up by the respective client, or one of the major public cloud services. By using NTT's multi service interconnection platform within the colocation center, sensitive and valuable data is transferred fast and safely from the physical into the virtual world.

The colocation data center as shelter for IT infrastructure

Using AI for autonomous driving is not only based on data, connectivity and algorithms: the adequate hardware to process all the information in a secured space is needed as well. Since high-performance server systems and mass data storage need high power density and efficient cooling, only colocation data centers can offer a secure environment for the use cases. Clients benefit from a secure and lockable surrounding within a scalable and professionalized data center infrastructure offering uninterruptible power supply and backup generators, as well as redundant cooling and air conditioning systems. All the important security standards and the highest security regulations are fulfilled by NTT Ltd.'s professional security concept. NTT Ltd. offers colocation data centers all over the world.

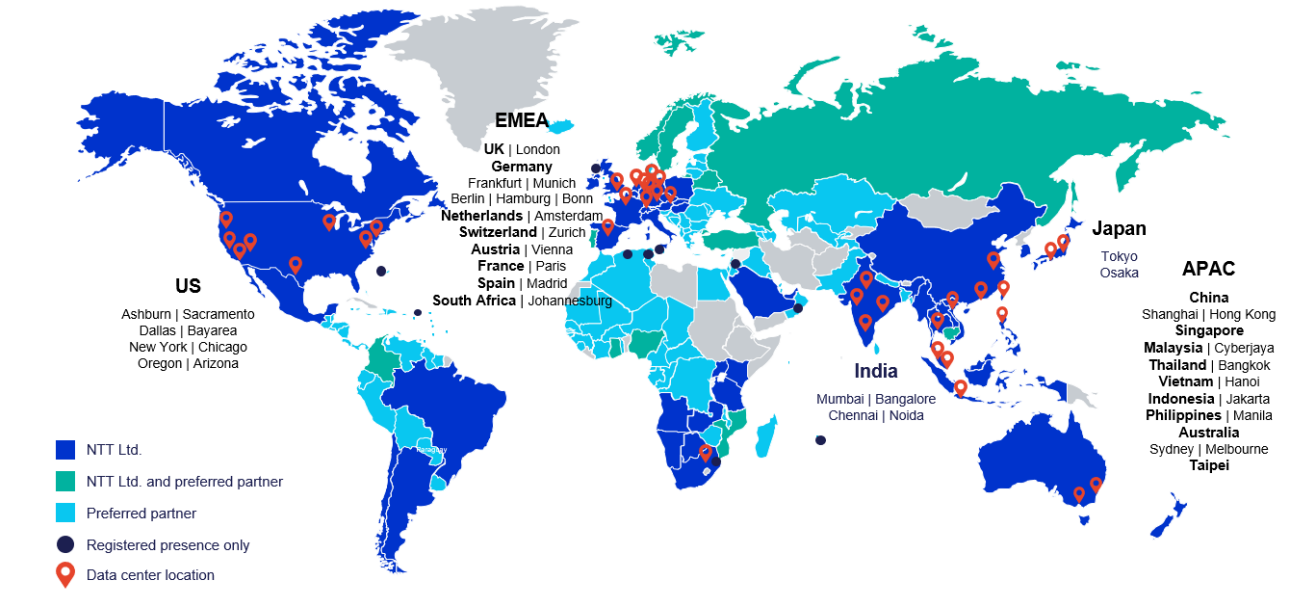

Figure 16: Global Data Centers, a division of NTT Ltd. operates more than 160 colocation data centers in more than 20 countries worldwide

Conclusion

According to several studies conducted recently, in the year 2050 about half of all vehicles on the streets will be driving autonomously. The authors outline the different advantages coming along with that: drivers will have more free time to spend with working or entertainment while being on their way to work. Sustainable mobility solutions with optimized routes can be established and the safety on our roads is supposed to increase. In order to achieve those goals, autonomous driving functions must be continuously developed and improved. The major step towards reliable autonomous vehicles is to train the ADAS.

That is why it so important to provide the complex infrastructure to collect, store and process the data needed to train the AI models. With the Data Garage use case implemented in the Technology Experience Lab, the automotive industry now has an easy way to develop, train and test their ADAS. The complete infrastructure as explained within this whitepaper can directly be tested and validated saving users a lot of time and money.

For further information, please contact:

b - plus	Stefan Rankl Sales b-plus GmbH stefan.rankl@b-plus.com	www.b-plus.com
IBM	Frank Kraemer Client Technical Architect IBM Germany kraemerf@de.ibm.com	www.ibm.com
INCENDA AI Quality Assurance	Florian Netter Co-Founder & Managing Director Incenda AI florian.netter@incenda.ai	www.incenda.ai
Lenovo	Karsten Kutzer Sr. Sales Engineer HPC Lenovo kkutzer@lenovo.com	www.lenovo.com
NTT	Dominik Friedel Business Development Manager NTT Global Data Centers EMEA GmbH dominik.friedel@e-shelter.com	www.datacenter.hello.global.ntt