DXC TECHNOLOGY

The critical role of data management for autonomous driving development

Table of contents

The critical role of data management for autonomous driving development

Data processing and analytics aim to produce the right data for users to act upon. The user is typically a human being, but in the case of Autonomous Driving (AD), the car assumes the role of the user.

In AD, input data, mainly from sensors, is processed by software in the car to produce the correct signal values for actuators, such that the car drives safely in all situations. The requirements for achieving safe, autonomous driving are extremely powerful sensors and complex software.

Today's Advanced Driver Assistance Systems (ADAS) are progressing to higher levels of autonomy.[1] This increases the complexity of AD software development and creates massive challenges, such as organizing processes for development, testing and verification. The volume of AD testing data is increasing dramatically; today many automotive companies deal with data in low PiB data volumes. R&D engineers in automotive companies now have to manage more data than they ever did before, and they generally don't have the core competencies to efficiently deal with the increasing requirements. Automotive companies want to hire highly skilled data processing and analytics professionals to take on the job, in order to save time and enable R&D engineers to focus all their time on their main role. Individuals with these skill sets are in high demand and short supply, which leaves many companies unable to fill these roles and at risk of project delays that may result in the delay of start of production (SOP) dates.

Effective data management, a pre-requisite for successful AD software development, should include powerful, feature-rich and scalable tooling that will mitigate any human and computer resource shortage, and integrate natively with software development and testing lifecycles. These aspects will be further elaborated upon from a data and analytics perspective in this paper.

This paper explores the applicability of the big data processing approach to solve data and processing challenges associated with AD Software development. It identifies the key components that need to be part of automotive data and analytics processing solutions to address the AD development challenges.

[1] "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles." *SAE International*, https://www.sae.org/standards/content/j3016_202104/

About the AD data chain

Data appears from multiple sources and in multiple forms in the AD development chain. We identify these with a data taxonomy.

Data Taxonomy

Following is a sensible approach to classifying data:

1. Raw data: Recorded by on-board sensor systems, like stereo cameras and LIDARs
2. Derived data: Resulting data subsets after processing raw sensor data, like a training dataset for deep learning
3. Ground-truth data: Data with a high confidence of correctness, used to verify AD functions. Usually generated due to human involvement
4. Metadata: Meta information about raw, derived, ground-truth data, and processes
5. Synthetic data: Generated by simulators to create complex and rare driving scenarios

In addition to the different classes of AD data (metadata), we have encountered various methodologies that deal with data under a one size fits all approach; a legacy automotive software development approach, ignoring data gravity; or approaches from other industries. These ineffective approaches are the result of inadequate technology and processes being applied to the business problems, leading to major issues and delays in data-driven development of AD software.

Analyzing AD/ADS from a big data approach (5V[2])

Big data and advanced analytics provide new, scalable approaches to AD data processing. Using big data is quite different from the traditional way of working for vehicle software; a unified, integrated, best-of-breed approach is emerging.

V1: Data Volume

The data volume that needs to be mastered is enormous. An open road data test fleet of about a dozen vehicles driving 7 hours per day at an average speed of 30 miles per hour records about 620,000 miles. If the fleet is equipped with state-of-the-art sensors that produce total data rates in the range of 3 to 6 GB/s, over 200 PiB total data volume will be recorded for analysis and re-processing (regression testing for new AD software releases). According to research by the RAND Corporation,[3] an autonomous vehicle will have to drive hundreds of millions of miles or even more than one billion, depending on the criteria for error-free performance, to provide evidence that it is performing to a certain precision level or better than human drivers.

[2] Ramadan, Rabie. "5Vs of big data." *Research Gate*, https://www.researchgate.net/figure/5Vs-of-big-data-10_fig1_325103690

[3] Kalra, Nidhi and Paddock, Susan M. "Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?" *RAND Corporation*, https://www.rand.org/pubs/research_reports/RR1478.html

Collecting, storing, and processing data for 1 billion miles is impossible today, so data volumes will be in the range of hundreds of PiBs. While Tesla recently unveiled plans for a supercomputer called "Dojo"[4] with Exaflops compute capacity, to fuel AI-based AD development, data storage capacities remain at PiB levels. Real-world data volumes for AD development are very high, reaching limits from a cost and practical data handling perspective, but are not yet high enough to complete higher-level AD development.

V2: Data Velocity

There are three areas where data throughput is particularly challenging: data recording and transmission, real-time Hardware in the Loop (HIL), and testing and artificial intelligence (AI) model training. AD/ADAS Level 2 to 4 cars are equipped with about 20 sensors, including ultrasonic, radar, LIDAR and stereo cameras. The highest data rates are produced by high-resolution stereo cameras (spatial resolution). All-in-all, data rates reach values of up to 10 GB/s, but around 5 GB/s is more common for current versions of test cars. While these rates could be processed with 100 Gbit Ethernet and/or PCI-express links, doing so creates negative knock-on effects such as heat dissipation, or limitations in the performance or range of electrical vehicles. These high-generation rates lead to new services like worldwide data logistics and to data gravity effects; the ideal requirement is to process data where it is generated and/or consumed.

Real-time HIL processing leads to massive throughput requirements in the ranges of over several hundred GB/s in view of parallel processing and testing real in-car processing times – as stated, data consumption requires co-location with data storage. This is a particular challenge for cloud-based data storage or perhaps for the HiL centers and their locations. Hybrid (cloud/on-premise) solutions can resolve this data challenge.

It's worth assessing the impact of 5G high-speed mobile networks on advancing AD/ADAS. Porsche has recently equipped its Weissach Development Center with standalone 5G in partnership with Vodafone[5] and HERE Technologies. This move aims to implement real-time data analysis within multi-access edge computing systems for real-time warning systems.[6]

[4] Swinhoe, Dan. "Tesla Details Dojo Supercomputer, Reveals Dojo D1 Chip and Training Tile Module." *Data Centre Dynamics*, https://www.datacenterdynamics.com/en/news/tesla-details-dojo-supercomputer-reveals-dojo-d1-chip-and-training-tile-module/

[5] "Real-Time Mobile Communications for the Vehicle Projects of the Future." *Porsche Newsroom*, https://newsroom.porsche.com/en/2021/company/porsche-vodafone-cooperation-5g-standalone-network-weissach-development-centre-real-time-mobile-communications-25566.html

[6] "HERE, Vodafone and Porsche partner on real-time warning system." *Porsche Newsroom*, https://newsroom.porsche.com/en/2021/innovation/porsche-real-time-warning-system-safety-here-vodafone-24923.html

It is very apparent that 5G is an enabler for improving many AD/ADAS functions, including the ability to use consumers' cars driving in traffic to record the corner-cases that are difficult to simulate in a meaningful way. These new data sources with specific content and new formats will further increase the requirements on data platforms and management systems. Technology progress in data transmission speed will be an enabler for more rapid progress in AD development.

V3: Data Variety

In the context of AD/ADAS the usual variety of structured, semi- and un-structured data formats are extended to automotive formats such as ROSbag, MDF4, HDF and ADTF. Historically these formats have been developed for robotics and automotive domains and are disparate from big data processing approaches and technologies; automotive engineers have been traditionally working with powerful workstations, small data volumes, and directly attached ECUs and HiLs. As data volumes have increased NAS/SAN systems have come into play as dumb storage extensions.

The lack of common ground is a significant obstacle; it is long overdue to develop an extension to automotive formats of big data technologies, as that would provide the necessary compute scalability to PiB volumes. A new contributor in data variety would be consumer cars with shadow driver software modules aiming to record corner cases in real-life traffic. These modules would record technically different formats and contents for transmission.

With data variety being very high, it is essential to follow a schema-on-read approach. A powerful bridge working on source formats in real time between automotive and data and analytics formats is required to open the world of advanced analytics to automotive engineers.

With data variety being very high, it is essential to follow a schema-on-read approach. A powerful bridge working on source formats in real time between automotive and data and analytics formats is required to open the world of advanced analytics to automotive engineers.

V4: Data Veracity

Veracity is applicable to derived data, such as objects recognized from video streams. That said, there won't be absolute veracity because all perception algorithms, like any other artificial intelligence-based inference, provide a percent probability for their predictions. Absolute data veracity on vast amounts of AD data is replaced through such probabilities and statistical significance obtained in suitable, reproducible tests.

Sensor fusion is a powerful means to provide consistency and increase the likelihood of derived data, which feeds the motion planning systems. However, in the short history of ADAS, incorrectly recognized or missed objects have led to assistance systems' mistakes, and in some cases to serious accidents.[7]

Data quality is another important aspect of AD/ADAS processing. Data, storage and compute capacity, time and human resources may be wasted in the verification and validation steps of ADAS data processing. Often, data may be of insufficient quality or identified late. As is the case with data veracity, data quality is a relative term in AD/ADAS.

[7] Rushe, Dominic. "Tesla's Autopilot faces US investigation after crashes with emergency vehicles." *The Guardian*, https://www.theguardian.com/technology/2021/aug/16/teslas-autopilot-us-investigation-crashes-emergency-vehicles

V5: Data Value

This is the core subject of this paper. Software needs to function error-free for functional and safe AD/ADAS solutions. The proof point lies in the data. Data, software and automation are a connected triangle.

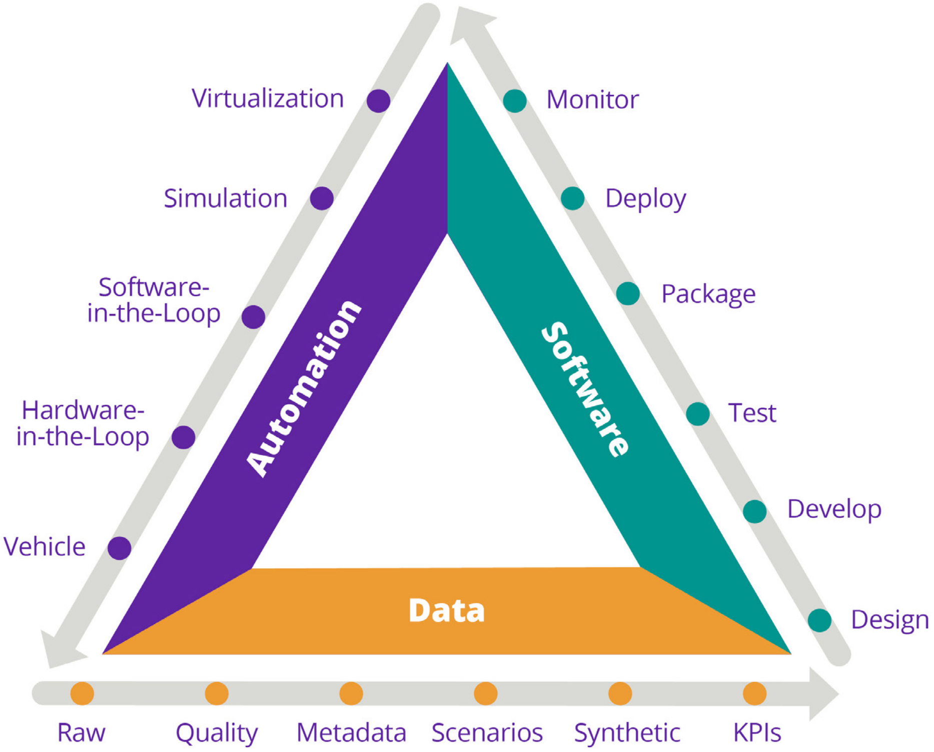

Figure 1. AD Process Triangle[8]

In a brute-force approach to AD validation and testing in order to prove with statistical significance the performance of the AD algorithms (Level 4 or 5), data collected from more than 600 million miles (more than one billion kilometers), would need to be reprocessed. This is practically and economically impossible today.

The key to statistically proving the performance of these AD algorithms lies in having a thorough understanding of the real-world data at practicable volumes, and the ability to generate synthetic data at sensible variations and volumes through simulations. One important step here is to demonstrate that the data from the real world and simulations produce identical output in the AD/ADAS processing chain of algorithms.

One critical step to achieve this is the development of a systematic approach to real-world scenarios. This calls for using the mature and promising OpenSCENARIO[9] standard file format for input to simulators.

[8] Bauhammer, Matthias. *Mastering autonomous driving development.* DXC Technology, https://dxc.com/us/en/insights/perspectives/paper/mastering-autonomous-driving-development

[9] ASAM OpenSCENARIO V2.0. *Association for the Standardization of Automation and Measuring Systems (ASAM)*, https://www.asam.net/project-detail/asam-openscenario-v20-1/

OpenSCENARIO is currently transitioning to V2.0, with the development of a Domain Specific Language (DSL) still ongoing.

Under the OpenSCENARIO approach, scenarios feed environment simulators, and together with sensor simulators as well as virtual ECUs, a processing chain can be built, also augmented by physical vehicle simulators.

In that way a complete digital twin of an AD system could be built, and slight randomization between generated scenarios could enhance simulations of reality.

The way to connect the real world with a simulated one, proofed as being of identical behavior, is best achieved by picking a set of well-tested cases in both worlds and by varying the simulations to achieve the best identical outcomes. This function of real-world matching to simulation systems as a fundamental principle is common in many scientific domains but has not yet attracted high attention for AD/ADAS.

The missing links of an overall data processing chain from real world to synthetic/simulated data processing are scenario extractors. The high number of definite scenarios paired with broad ranges of parameters make it a very complex task to create universal scenario extractors; this probably will be best achieved through supervised training of yet to be identified features and AI algorithm variations. However, focusing on a finite number of scenarios and test cases could be the start to bringing the real world and simulated ones in alignment.

If that could be achieved, simulations could be trusted, and real-world and hardware-based reprocessing could be minimized. In other words, an ADAS function tester would pick a few scenarios with limited variations leading to a manageable number of test cases and physical data recordings at sub-PiB scale. For these recordings, algorithms could be tested and iteratively improved, and documented in an audit trail and meaningful KPIs.

The same can be done in the virtual world, but for more scenarios with a broader range of parameters, leading to highly increased numbers of test cases and volatile synthetic data for much higher volumes.

At higher autonomy levels the resulting productivity gain from this approach will be crucial for the completion of Verification and Validation (V&V) tasks to meet SOP schedules. The ability to be highly selective in reprocessing and simulation as the result of a fine-grained metadata structure and a scalable data management system is key to success.

From our point of view, the statistical significance for proving that algorithms are performing better than human drivers in the case of Level 5 autonomy will be reached by:

1. Being highly selective in scenarios and test cases and avoiding the low value data recordings
2. Increasing data volumes for critical test cases through simulations as required

The ability to be highly selective in reprocessing and simulation as the result of a fine-grained metadata structure and a scalable data management system is key to success.

At a high level, we suggest that the overall processing chain follow the flow depicted below:

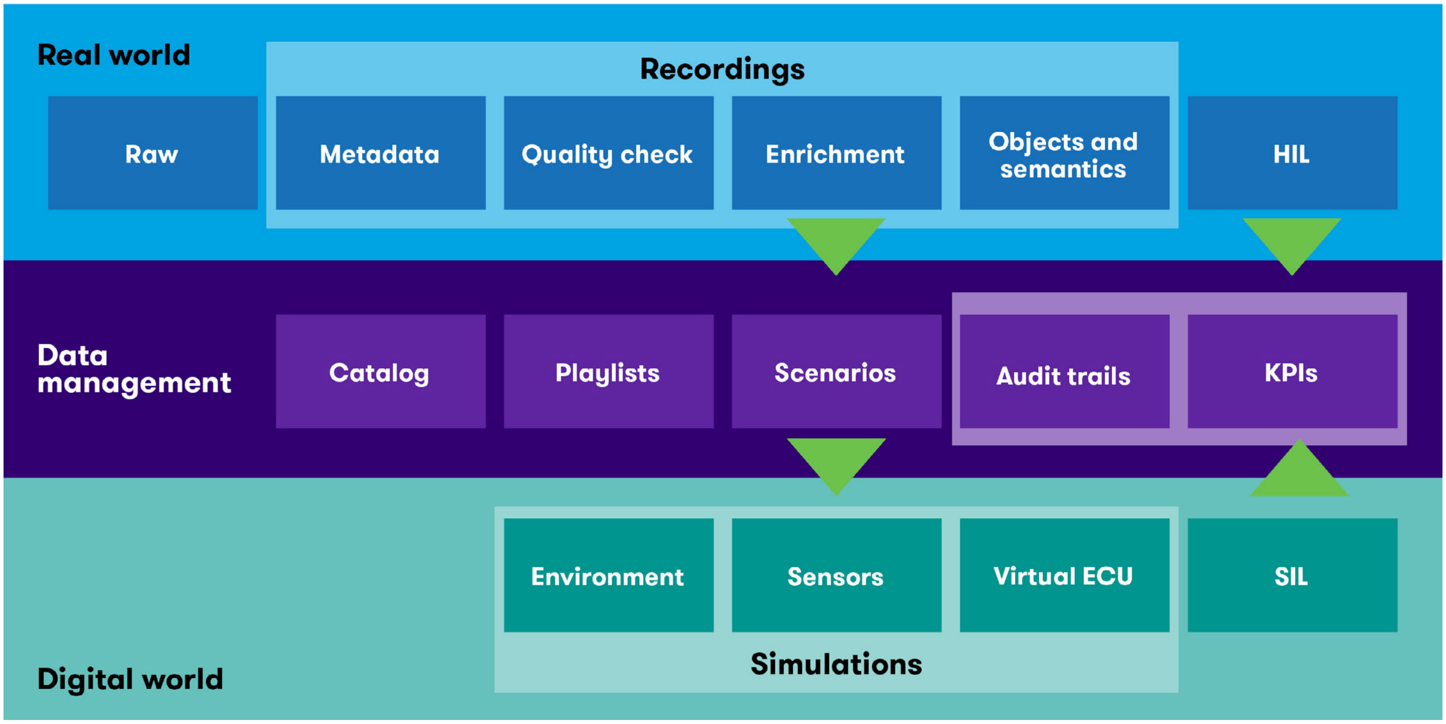

Figure 2. Data processing flow in the real and digital world

As you can see from this processing chain, the effective and scalable methodology and tooling for moving from real-world raw data to scenario detection and extraction for synthetic data are the most valuable steps in an AD/ADAS R&D process.

In addition, it becomes apparent that data management is a central, crucial element and has the potential to build a bridge between the two worlds. Furthermore, for higher levels of AD with more complex functions, the number of scenarios that the system under test needs to go through will grow by orders of magnitude, and the number of test cases will not be possible to process in the real world. Complete virtualizations and digital twins become a requirement – but everything needs to have proof points in reality, which brings us back to the value of data.

Conclusion

Applying a big data-driven analysis of ADAS/AD processing shows differences, but there are growing synergies between automotive domains, and data and analytics, as shown in the figure below.

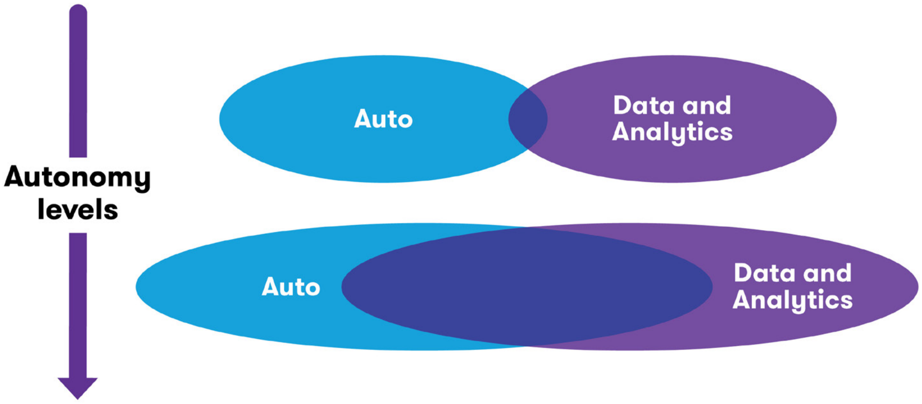

Figure 3. Automotive and Data and Analytics domains

At lower levels of autonomy data characteristics in terms of 5V are quite different between automotive domains, and data and analytics. At higher levels some criteria naturally have a large overlap (volume, velocity) implying that other dimensions (variety, veracity) of the 5V need a bridge to bring value to the automotive domain.

In the above sections, key components of a joint automotive data and analytics processing solution have been identified: scalable data analysis tools, metadata generation and scenario recognition that enhance the data value enormously. Mastering these data processing challenges through effective data management in real and virtual worlds will enable organizations to use resources effectively, save time and money, and ultimately be successful in AD/ADAS development – this is even more critical in view of future requirements, ever-increasing data volumes, enabling technologies and higher compute power. Clever data management means working smarter rather than harder for AD/ADAS development.

Author:

Dr. Günter Koch is the global solution lead for DXC Robotic Drive, heading up an international team of specialized solution architects responsible for the Robotic Drive solution and DXC's Autonomous Driving Platform and Toolkit. He holds a doctorate in nuclear physics research and is a highly experienced data analyst. In the development of autonomous driving technology, Günter is focused on both technical innovation and cost-effective delivery.

Contributor:

Philipp Stapff has had a varied career in Analytics, with roles in business intelligence and data analytics, and more recent specialization in architectures for data science and data management for insurance and banking. Since 2019 he has been lead solution architect at DXC, focused on data management for AD/ADAS automotive.

Learn more at
dxc.com/data-analytics

About DXC Technology

DXC Technology (NYSE: DXC) helps global companies run their mission critical systems and operations while modernizing IT, optimizing data architectures, and ensuring security and scalability across public, private and hybrid clouds. The world's largest companies and public sector organizations trust DXC to deploy services across the Enterprise Technology Stack to drive new levels of performance, competitiveness, and customer experience. Learn more about how we deliver excellence for our customers and colleagues at **DXC.com**.