DATASHEET

INDUSTRY FOCUS | AUTONOMOUS DRIVING

Teaching Cars To Drive — A Data Intensive Endeavor

People intuitively understand that self-driving cars present complex engineering challenges. Vehicle assembly is the easy part — we've been doing that for 100 years. The real challenge is a data challenge, acquiring and managing the data needed to train autonomous vehicles to drive themselves, and this is rooted in data — the ability to ingest, store, and analyze large volumes of data & the high bandwidth needs of data-in-motion. The current state of digital solutions offers redundant and fragmented systems across the autonomous drive data loop. Automotive manufacturers demand new approaches in computer science and deep learning, along with key technologies that integrate all steps of the autonomous driving development lifecycle.

The Data Loop

To understand the challenges facing automakers, it's helpful to consider the autonomous driving data loop. The loop is comprised of five separate, but highly interconnected steps:

Why Cloudera

Hybrid And Multi-Cloud

Run your analytics on the clouds you choose. Easily and securely move data and metadata between on-premises file systems and cloud object stores.

Analytics From Edge To AI

Apply real-time stream processing, data warehousing, data science and iterative machine learning across shared data, securely, at scale on data anywhere.

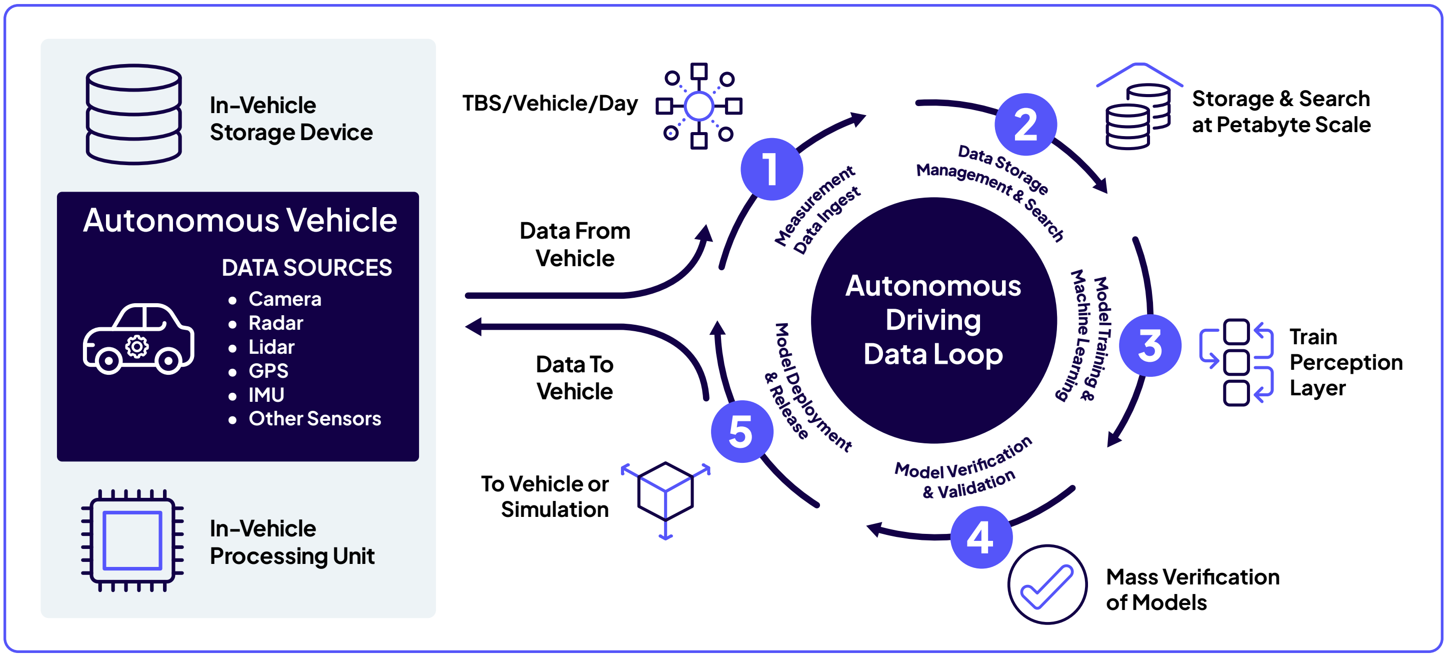
Security And Governance

Use a common security model, role and attribute based access policies and sophisticated schema, lineage and provenance controls on any cloud.

100% Open

Open source, open compute, open storage, open architecture and open clouds. Open for developers, partners, and open for business. No lock-in. Ever.

Autonomous Driving Data Loop


In-Vehicle Storage Device
Autonomous Vehicle
DATA SOURCES
• Camera
• Radar
• Lidar
• GPS
• IMU
• Other Sensors
In-Vehicle Processing Unit
TBS/Vehicle/Day
Data From Vehicle
Data To Vehicle
To Vehicle or Simulation
1
Measurement Data Ingest
2
Data Storage Management & Search
Storage & Search at Petabyte Scale
3
Model Training & Machine Learning
Train Perception Layer
4
Model Verification & Validation
Mass Verification of Models
5
Model Deployment & Release
Autonomous Driving Data Loop

1. **Measurement Data Ingest:** In this phase, dozens of terabytes per vehicle per day of real-world driving data from highly accurate reference sensors (camera video, radar, lidar, ultrasonic, GPS, IMU/CANBUS and other external and internal sensors) is ingested and stored on the vehicle. Later, this data is manually ***ingested*** into a centralized environment for subsequent processing. Inefficiencies in data management require redundant data management overhead (i.e. data storage, processing and hardware costs) on both the vehicle and within the enterprise data analysis environment.
2. **Data Storage, Management and Search:** Data availability to a wide range of data consumers within the organization is important. As such, it must be ***stored*** and managed as necessary. Critically, given the truly massive amounts of data encountered, relevant data sets should be made ***searchable*** for key users (i.e. Machine Learning Engineers, System Test Engineers) to develop the perception layer and train the core deep learning models. Challenges include cost effective storage, managing the trade-off between data access performance and cost, and effective search of relevant data.
3. **Model Training and Machine Learning:** How well a machine learning model can predict an outcome is determined by the volume, diversity and quality of the training data and the ***machine learning*** algorithms employed. Challenges include the initially manual, tedious and time-consuming efforts labeling the vehicle measurements in creating the perception layer and the siloed, specialized and isolated machine learning development environments with limited governance.
4. **Model Verification and Validation:** Though car manufacturers would like 100k miles of real-world vehicle testing, time-to-market requirements dictate the need for a simulation-based ***mass verification*** approach to testing on very large reprocessing (or verification) clusters. This in turn requires a massively powerful and reliable capabilities for batch processing, scheduling, and orchestration of verification processing in a container batch runtime environment.

A Fragmented Approach

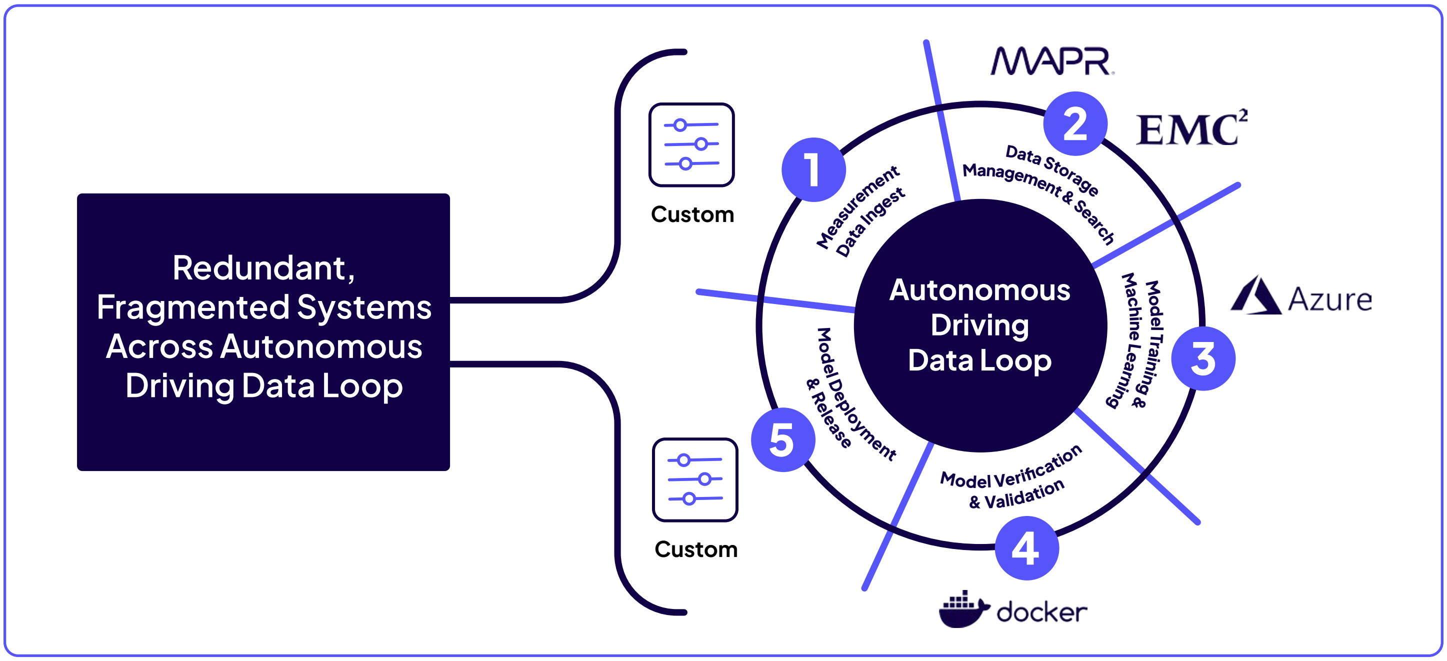
Car makers so far have been handed a fragmented approach to the data problem. By acquiring ingest, storage, search and data management capabilities from multiple vendors, auto makers have been slaves to siloed solutions, driving up the cost and complexity of an already expensive and complex process.

Solving Through Sharing and Scaling

To overcome these challenges, the automotive industry must leverage a high-performance, scalable and reliable data management solution. The solution must ingest and process the volume, velocity and variety of data required to support the data lifecycle in real-time from one common platform. It must store accessible data across all functions of a business and be easily searchable. **Cloudera** addresses these challenges through an integrated set of capabilities addressing the needs of the data lifecycle for autonomous vehicles.

Cloudera is comprised of interrelated capabilities:

- **Cloudera DataFlow:** Cloudera DataFlow provides the ability to collect and process high-volume, real-time streaming data at the vehicle edge and guarantee delivery of data to the Cloud or on-premise data centers. More specifically, Cloudera DataFlow provides the abilities to aggregate, compress and encrypt connected vehicle data, prioritize transmission of data from the vehicle to the Cloud or Data Center, buffer data in the event of network interruptions and track the provenance and lineage of streaming data, providing confidence in the origin and usage of data.
- **Cloudera Search:** Cloudera Search provides easy, natural language access to data stored in or ingested into Hadoop, HBase, or cloud storage. For Autonomous Drive engineers and analysts this means google-like discovery and analysis (via user interface or search API) providing the ability to feed applications and machine learning models with the specific driving episodes required, rather than flooding them with superfluous and repetitive data.


Redundant, Fragmented Systems Across Autonomous Driving Data Loop
Custom
Custom
MAPR
EMC²
Azure
docker
Autonomous Driving Data Loop
1
Measurement Data Ingest
2
Data Storage Management & Search
3
Model Training & Machine Learning
4
Model Verification & Validation
5
Model Deployment & Release

- **Cloudera Machine Learning:** A collaborative, customizable Continuous Integration, Continuous Deployment (CI/CD) environment for machine learning engineers, featuring easy and secure access to all datasets and processing resources within the organization. Machine learning environments exist as massively scalable docker environments in Kubernetes-based runtime on-prem/cloud.
- **Cloudera Data Engineering:** Cloudera Data Engineering is a powerful and cost-effective platform for processing large-scale data sets on-premise or in the Cloud. Within the autonomous drive data lifecycle, Docker containers on Kubernetes are leveraged for mass inference within the perception layer, while Spark on Kubernetes (with orchestration via Apache Airflow) is leveraged for pre-processing, post-processing and mass verification of the perception layer.

Driving Forward

The trend toward autonomous vehicles is clearly underway and data is the bedrock to its success. Critical to this reality is a high performance, scalable and reliable data management platform processing extreme volumes at velocity, with the variety of the data that is required, working at scale and economically viable. Cloudera is the leader in delivering this type of next generation data management platform. Cloudera allows stakeholders to better manage and wring the most value out of data, collaboration and insights derived through an enterprise scale, integrated, open architecture platform.

CLOUDERA

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible—today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X.