	Typical Challenges	Driving forces
Volume	• Where to store the data? • How to store the data? • Which data to store?	➢ Rapidly increasing market of connected cars ➢ More built-in sensors in cars lead to more measure points ➢ Generated data volume: o autonomous car: 2 PB of data/car/year o connected car: 130 TP of data/car/year
Variety	• How to access various types/sources of data? • How to store different types of data?	➢ Different Electronic Controlling Units (ECU) contain different information about car (velocity, usage, efficiency, pressure) ➢ Further data contain GPS, road network, road condition, vehicle to vehicle and vehicle to infrastructure communication, video, …
Velocity	• How to deal with differences depending on data? • How to deal with time-consuming analytics?	➢ Data needs to be evaluated immediately, e.g.: for hazard detection or collision avoidance, but not necessarily transferred ➢ 25 GB of data/car/hour calls for scalable solution for efficient data handling
Veracity	• How to implement special steps for ensuring data quality, data anonymization, model validation and development?	➢ To create good ML models to handle all the various situations for connected cars data needs to be transferred ➢ By context enrichment of IoT data (wear and tear monitoring, location analytics) analytics can be done better and more reliable
Value	• How does it support the objective and economic considerations? • How to deal with legal frameworks, ethical/moral frameworks, result validations, data governances?	➢ Increasing interest of consumers in connected cars due to new features like vehicle management, mobility management, safety, autonomous driving, well-being, entertainment ➢ Apart from consumers other sectors can profit of the generated data through predictive maintenance, insurance policy creation, advanced vehicle diagnostics, …

	Typical Challenges	Solutions
Data Source Layer	• How to access the data? • How to deal with Network Transmissions, Transfer Protocols, Data Provisions, Access Points, Interfaces, Data Formats?	➢ Depending on the relevancy of the data streaming (real-time) with Apache spark or batch loading is applied ➢ Fair amount of data needs to be transferred via cloud to allow for further analytics ➢ Data of sensors are transmitted by predefined standards in a predefined structure (messaging bus e.g.: Kafka)
Data Storage Layer	• How to find the suitable storage format? • What data is important? • How to deal with data conversions and quality check prior and/or after storage?	➢ Performant SQL-engine impala SQL to capture data for BI solutions ➢ Operations allows ease organization of clusters (e.g.: on premise, cloud, hybrid) ➢ Data Management comprises holistic security concept (data encryption, meta data management, data linage) ➢ Data is stored in hdfs, hbase, kudo, …
Processing Layer	• What are the right tools, methods and algos? • How to choose from the variety of analysis methods depending on data and/or objectives? • How to apply proper feature selection? • How to deal with computationally intensive ML?	➢ Applying edge analytics (processing within the car) to ensure low latency for human response times ➢ Applied analytics tools (processing within the cloud) comprise time series analysis, machine learning (MLLib) or context enrichment ➢ Data is processed in spark, spark streaming, …
Data Output Layer	• How to visualize results? • Who is the target audience? • Do we need to feedback the results?	➢ Analytics can contain information regarding why collisions are happening, how cars perform in comparison, failure detection, driver description ➢ Creating a 360° View ○ Important data can vary from car relevant information like the average speed, fuel used, but also customer relevant information driving behavior ➢ BI Solutions for end users with low latency ➢ Data is visualized with arcadia