4th Workshop

Pilot on **Fair and equal data sharing for cooperative, connected and automated mobility**
Presentation on new features

Big Data and B2B platforms: the next big opportunity for Europe
EASME/COSME/2018/004

Vagelis Karakolis
ICCS-NTUA

EASME - European Commission Executive Agency for Small and Medium-sized Enterprises

Online Workshop
8 October, 2020

Two major developments

Bilateral Communication between vehicles and OEMs

Elaborating Anonymization

Bilateral Communication Engine

/ 01

Alternative Data Flow models

Shared Server SWOT analysis

S: refers to gathering strength points for **BOTH** the OEMs **AND** the SMEs, in terms of implementation and running costs, sharing data and infrastructures and low implementation and viability risks.

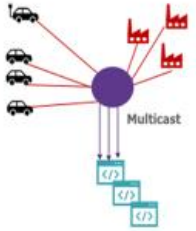
W: refers to gathering weak points for **EITHER** the OEMs **OR** the SMEs in terms of implementation and running costs, sharing data and infrastructures.

O: refers to gathering opportunities for **BOTH** the OEMs **AND** SMEs.

T: refers to pointing out risks and threats for the implementation or the viability of the project.

"Multicast" scenario was preferred by most of the participants however, OEMs representatives were strongly opposed to it

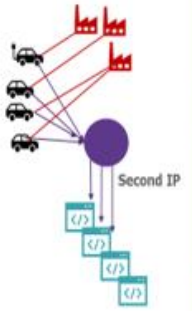
Data Flow 1: In-vehicle data are transmitted to the OEMs and SMEs via multicasting



S • Equal terms for SMEs and OEMs • **Shared operational and telecommunication costs** • **Shared infrastructures and cutting costs for the OEMs**	W • **Depreciates existed investments and infrastructures of the OEMs** • High operational complexity • Time will be needed for deploying changes and asking for additional data points of the in-vehicle data sets (expansion of the required data)
O • Common user's consent management for the automotive sector • Common technical environment for service development for the automotive sector – especially for safety and environmental apps	T • **Product liability issues will arise, as the OEMs may not be liable for the vehicle anymore** • **High resistance from the OEMs** • Collaborative development entails high coordination and high complexity. • Project setup and implementation may take too much time

"Second IP" scenario seemed to be acceptable by many of the participants

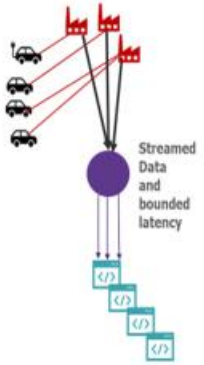
Data Flow 2: In-vehicle data are transmitted simultaneously to the Shared Server and the OEM's server



S • Equal terms for SMEs and OEMs • Easier implementation, since the Shared Server will handle only SMEs data • **High technology readiness level** • **Does not depreciate existed investments and infrastructures of the OEMs**	W • **Lack of common use of same vehicle data and resources** • **Telecommunication costs will be high and will be fall to the Shared Server** • Time will be needed for deploying changes and asking for additional data points of the in-vehicle data sets (expansion of the required data)
O	T • May be very hard to monitor the conformance of the OEMs, potentially resulting in low quality of service for the SMEs

"Streamed data and bounded latency" was rejected by most of the workshop participants

Data Flow 3: OEMs' server streams data with specified upper latency bound



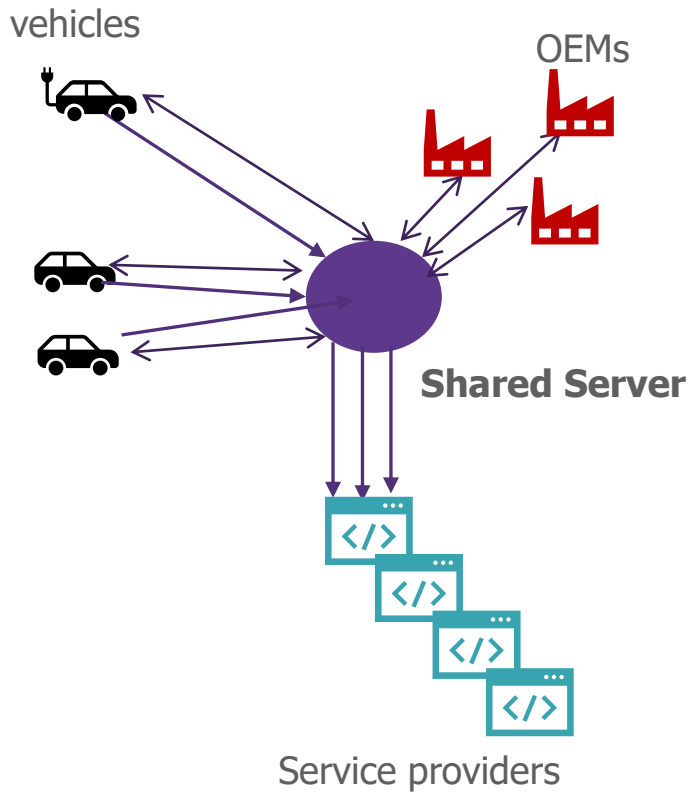
S • Easier implementation, since the Shared Server will handle only SMEs data • **Does not depreciate existed investments and infrastructures of the OEMs** • **There is no need to alter the vehicles' systems** • **Shared telecommunication costs** • OEMs should be liable for the data	W • **OEMs retains a competitive advantage of the latency over the SMEs, in the order of a few hundreds of milliseconds** • **We cannot get anything better than the OEM can serve in terms of time intervals and data points**
O	T • May be very hard to monitor the conformance of the OEMs, potentially resulting in low quality of service for the SMEs

Bilateral Communication between vehicles and OEMs – The Multicast Data Flow Model

- The initial Shared Server architecture could not support the Multicast Data flow model
- The Multicast Data flow model was **preferred by most of the WS2 participants**
- The last release of Shared Server Solution supports the Multicast data flow Model.

Specifically:

- The Shared Server receives a continuous flow of data from the vehicles in order to promote them to service providers
- OEMs' Servers receive a continuous flow of data from the vehicles
- OEM can push messages to the vehicles
- OEMs receive messages from the vehicles in response to a message



Bilateral Communication between vehicles and OEMs – How it is achieved

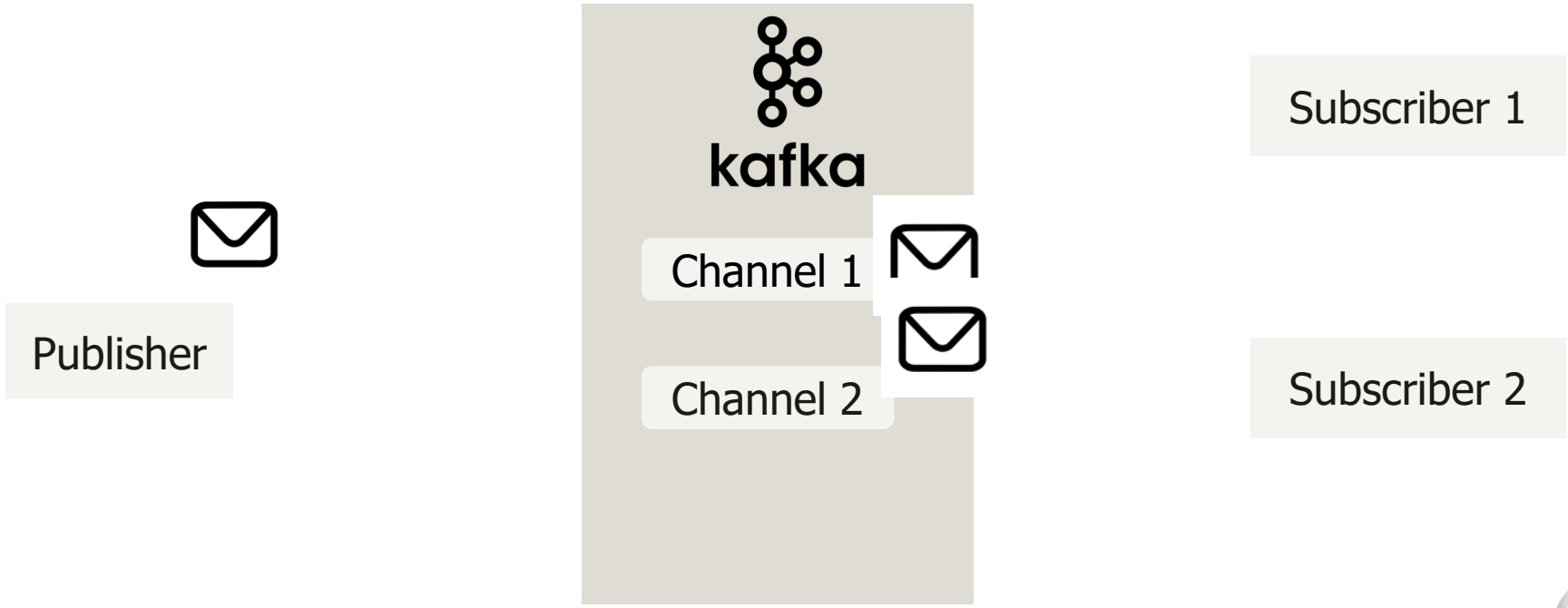
- The Shared Server provides communication channels per OEM and its vehicles
- The Shared Server utilizes the **Publish/Subscribe model of communication**
- With this communication model:
  - Vehicles send data streams with data required by the OEMs continually through a communication channel
  - OEMs consume the data streams from the channel
  - OEMs push messages to their vehicles by using another channel
  - Vehicles of each OEM listen to their respective messages' channel and process only the messages that are relevant to them.
  - Afterwards, vehicles publish a message to a response channel

Apache Kafka

Kafka is a messaging system that is designed to be fast, scalable, fault-tolerant and durable. It is an open-source stream processing platform. Apache Kafka originated at LinkedIn and later became an open-source Apache project in 2011. It aims at providing a high-throughput, low-latency platform for handling real-time data feeds.

kafka

Apache Kafka provides communication channels. The applications which publish messages to these channels called Publishers. The messages are received only by the applications (Subscribers) that have subscribed to this channel.

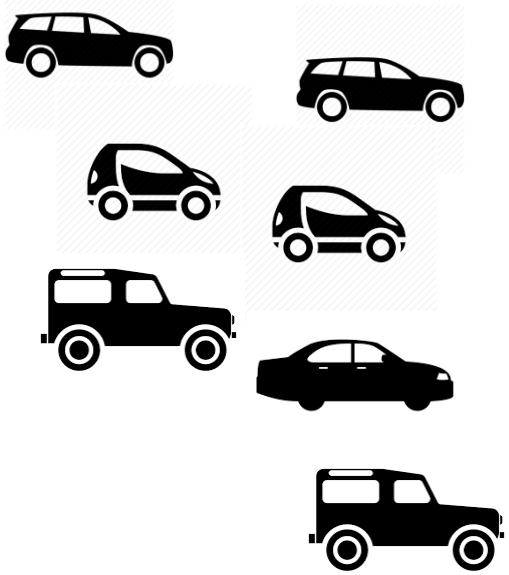


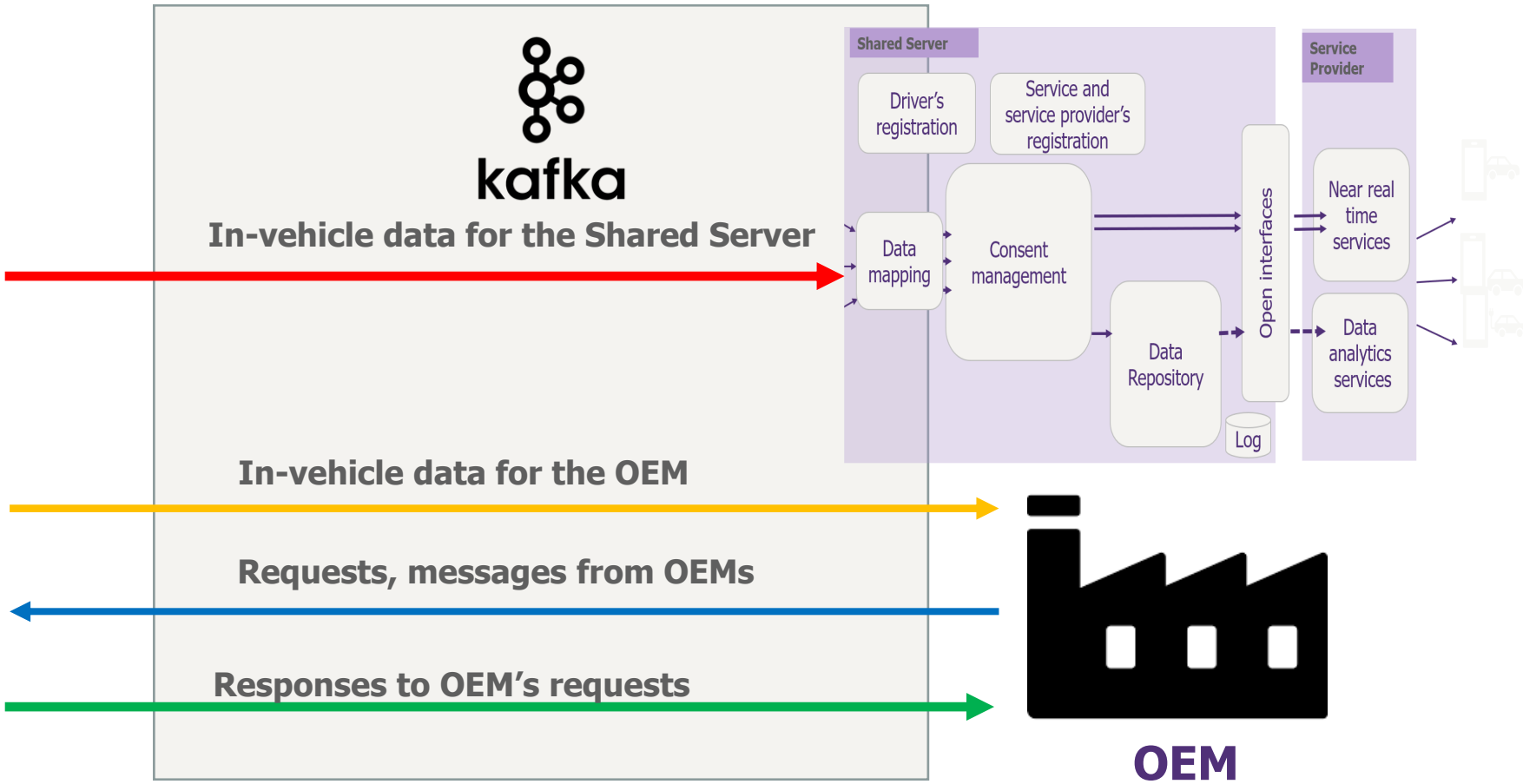
Azure HDInsight Kafka Service

- Managed service that provides a simplified configuration process
- 99.9% Service Level Agreement (SLA) on Kafka uptime
- HDInsight allows you to change the number of worker nodes (which host the Kafka-broker) after cluster creation which leads to easy upscale and downscale

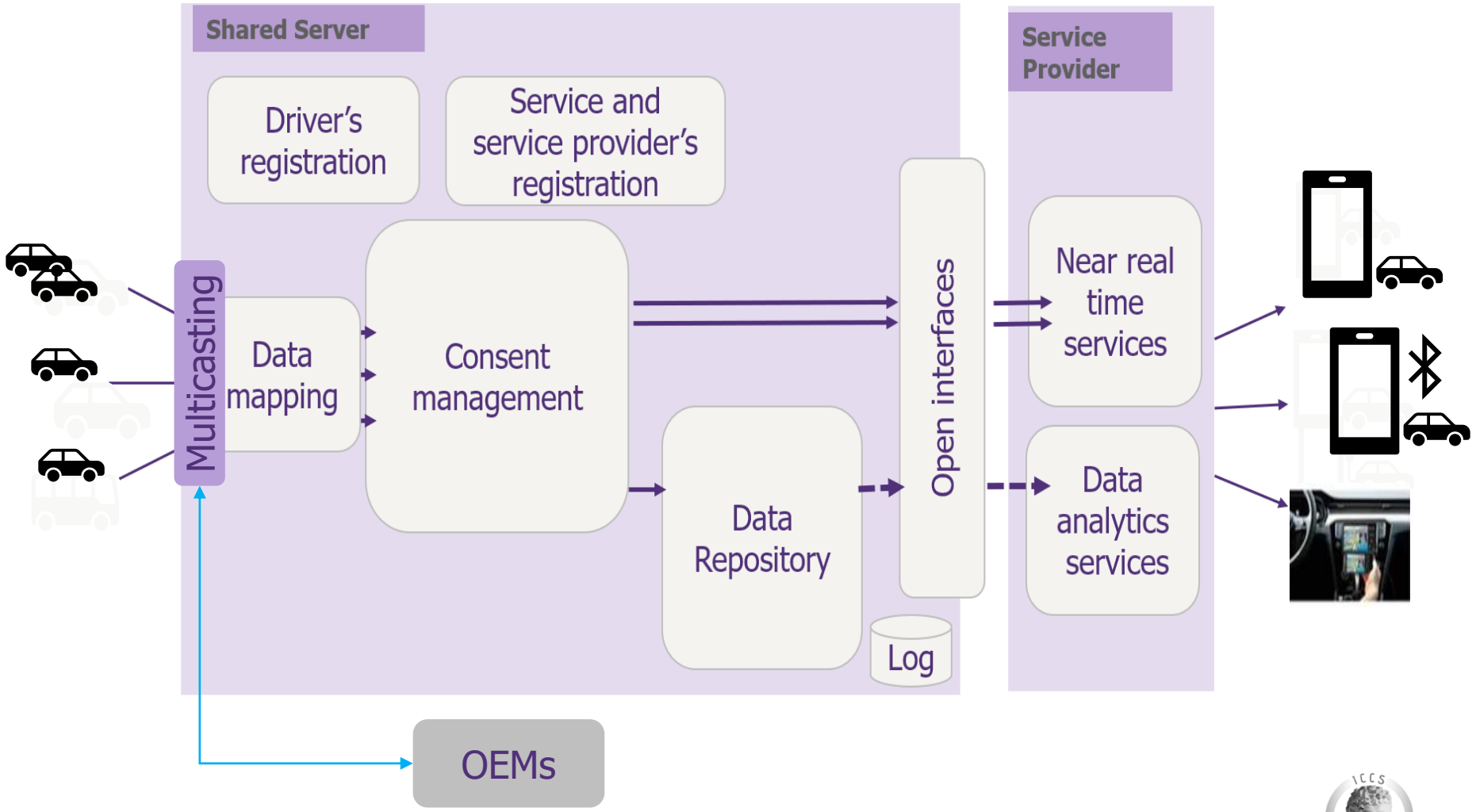
Let's take a closer look ...

Fleet of vehicles of specified OEM

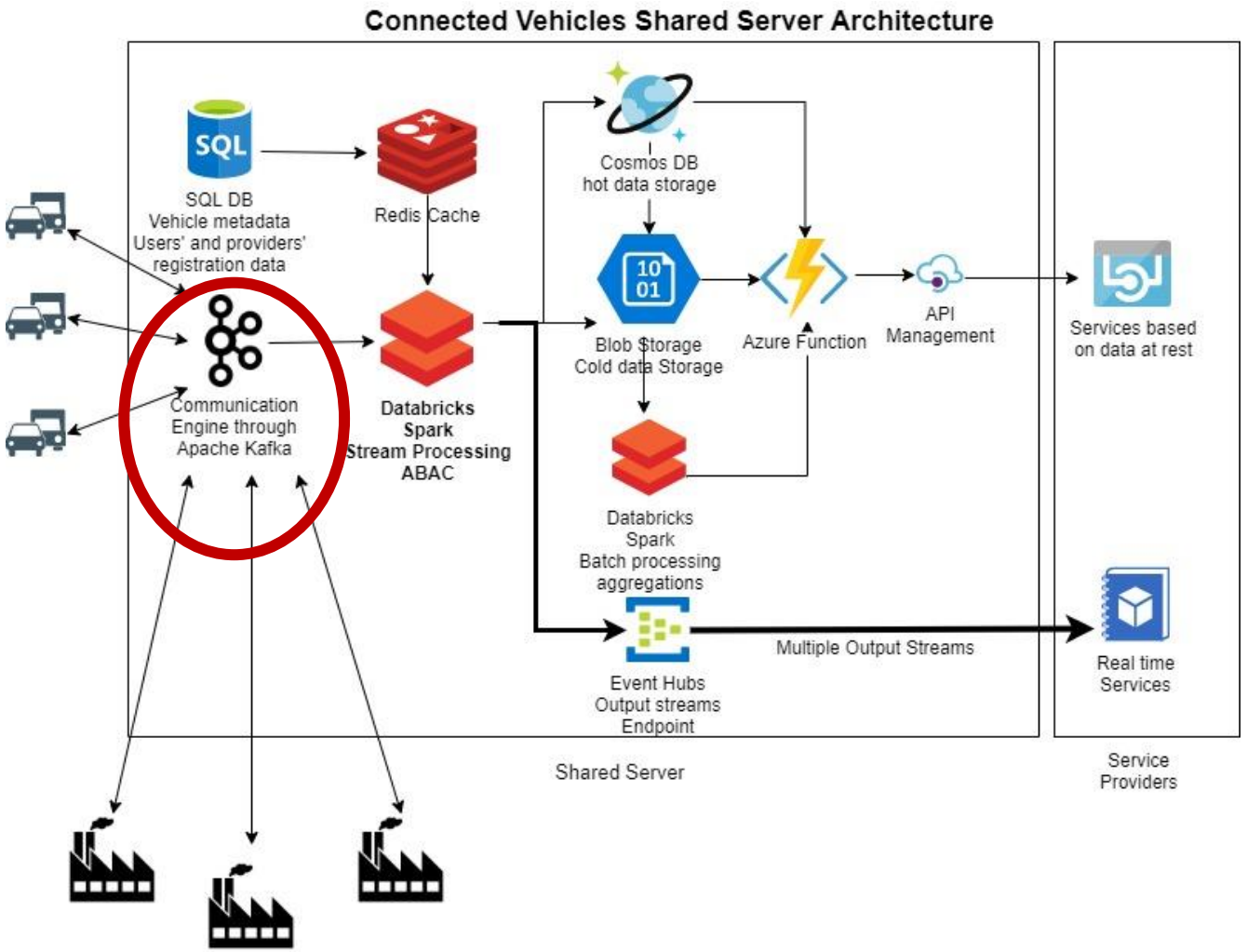




The big picture – Two roles beside of the OEMs



Updated architecture of the Shared Server

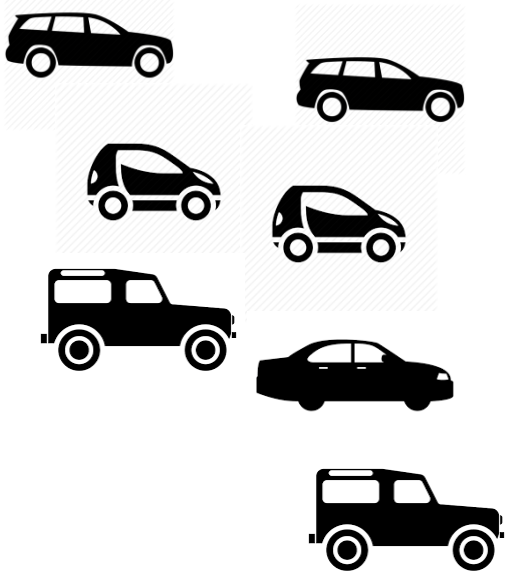


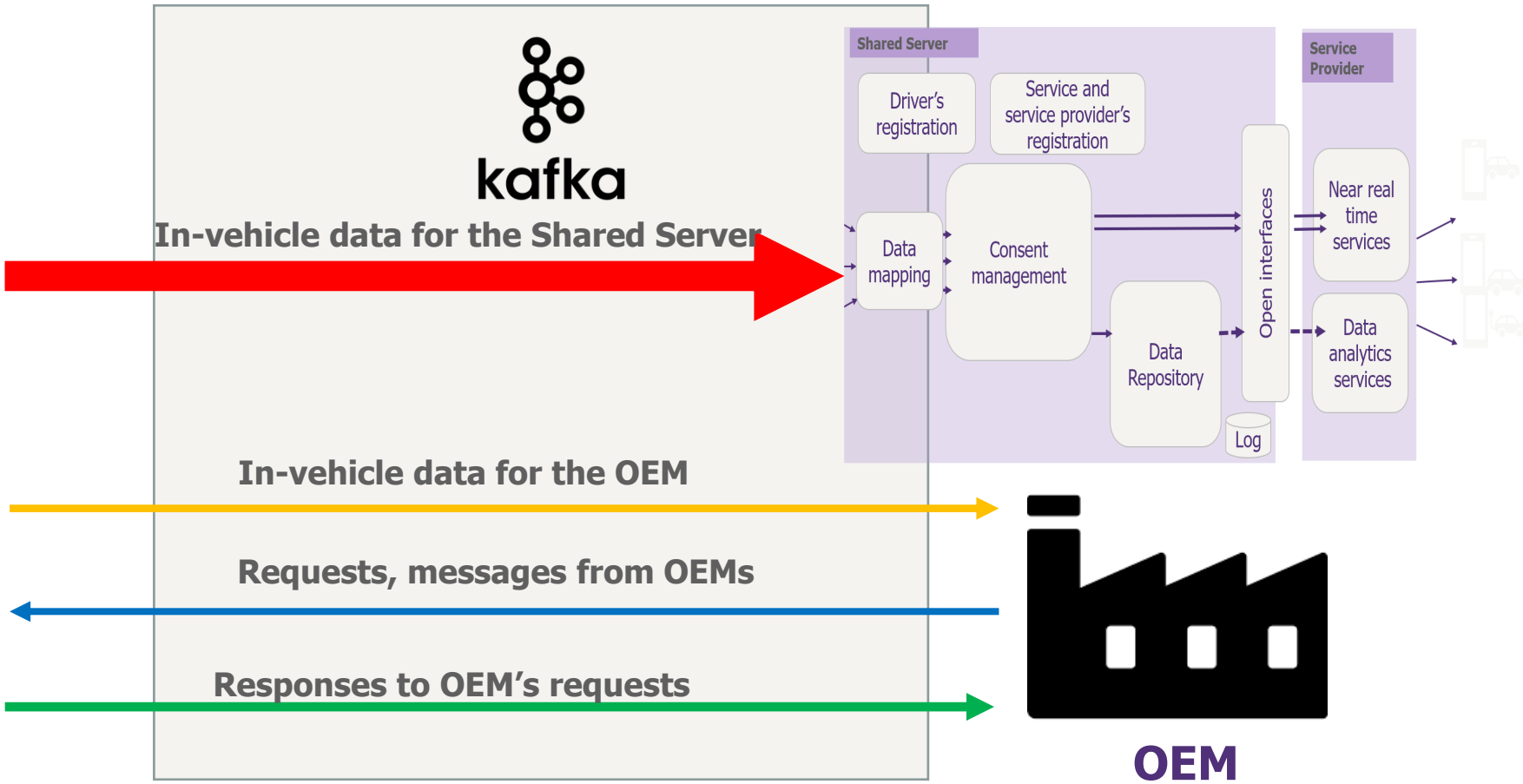
Bilateral Communication Engine – Demonstration

/ 02

Let's take a closer look ...

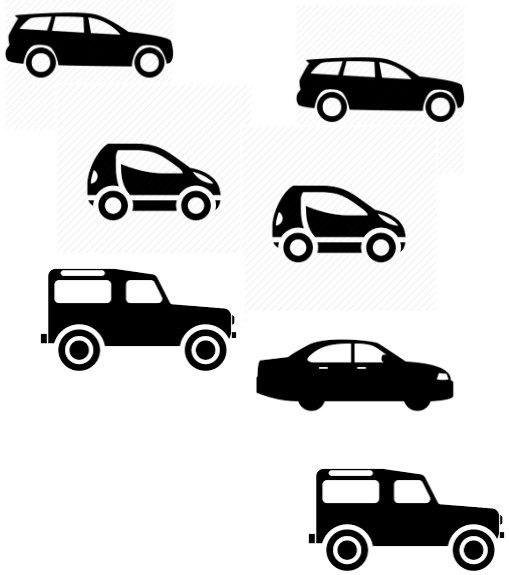
Fleet of vehicles of specified OEM

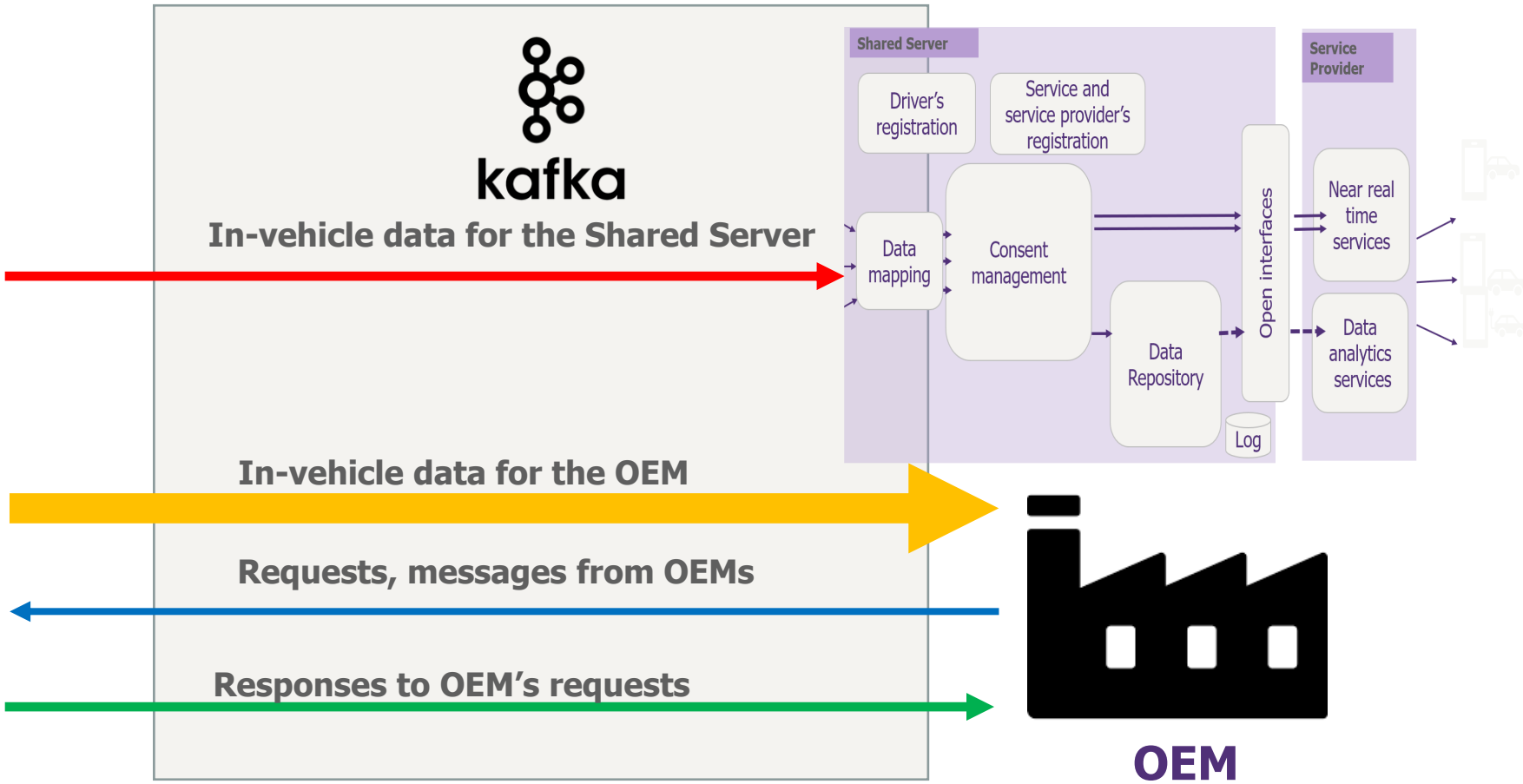




Let's take a closer look ...

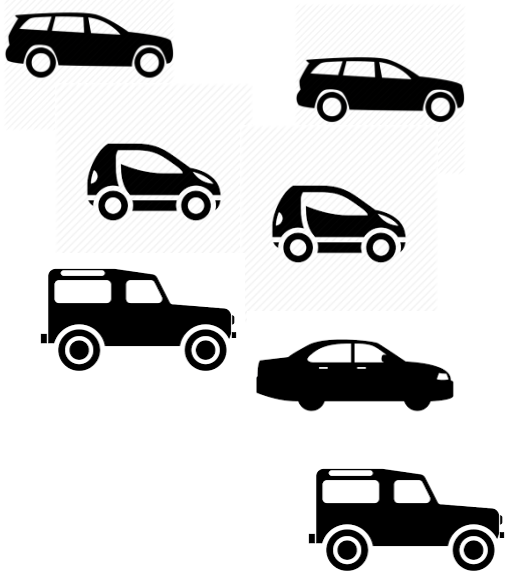
Fleet of vehicles of specified OEM

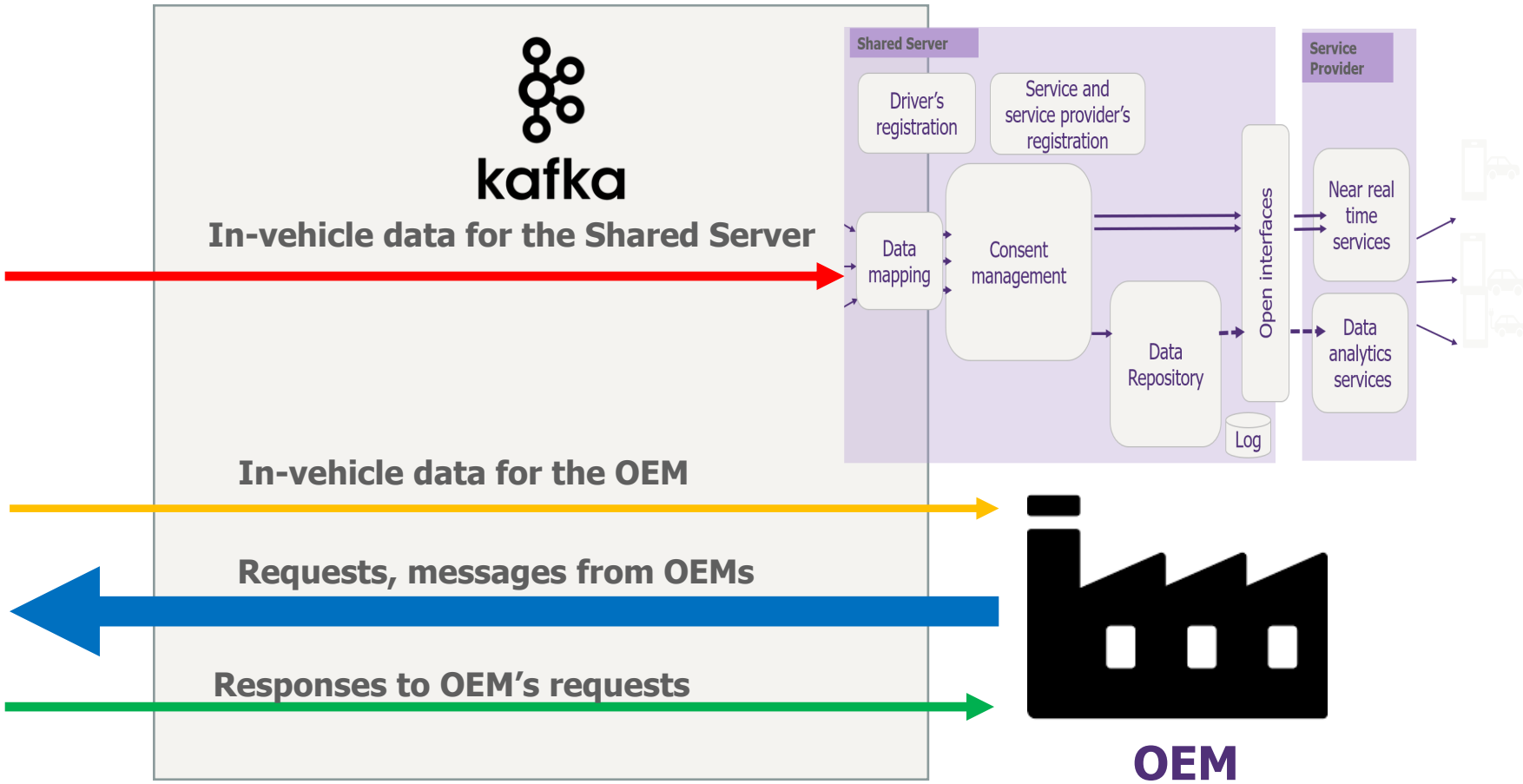




Let's take a closer look ...

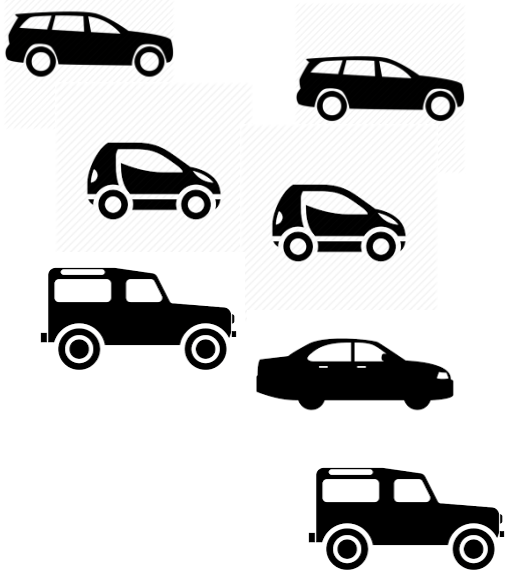
Fleet of vehicles of specified OEM

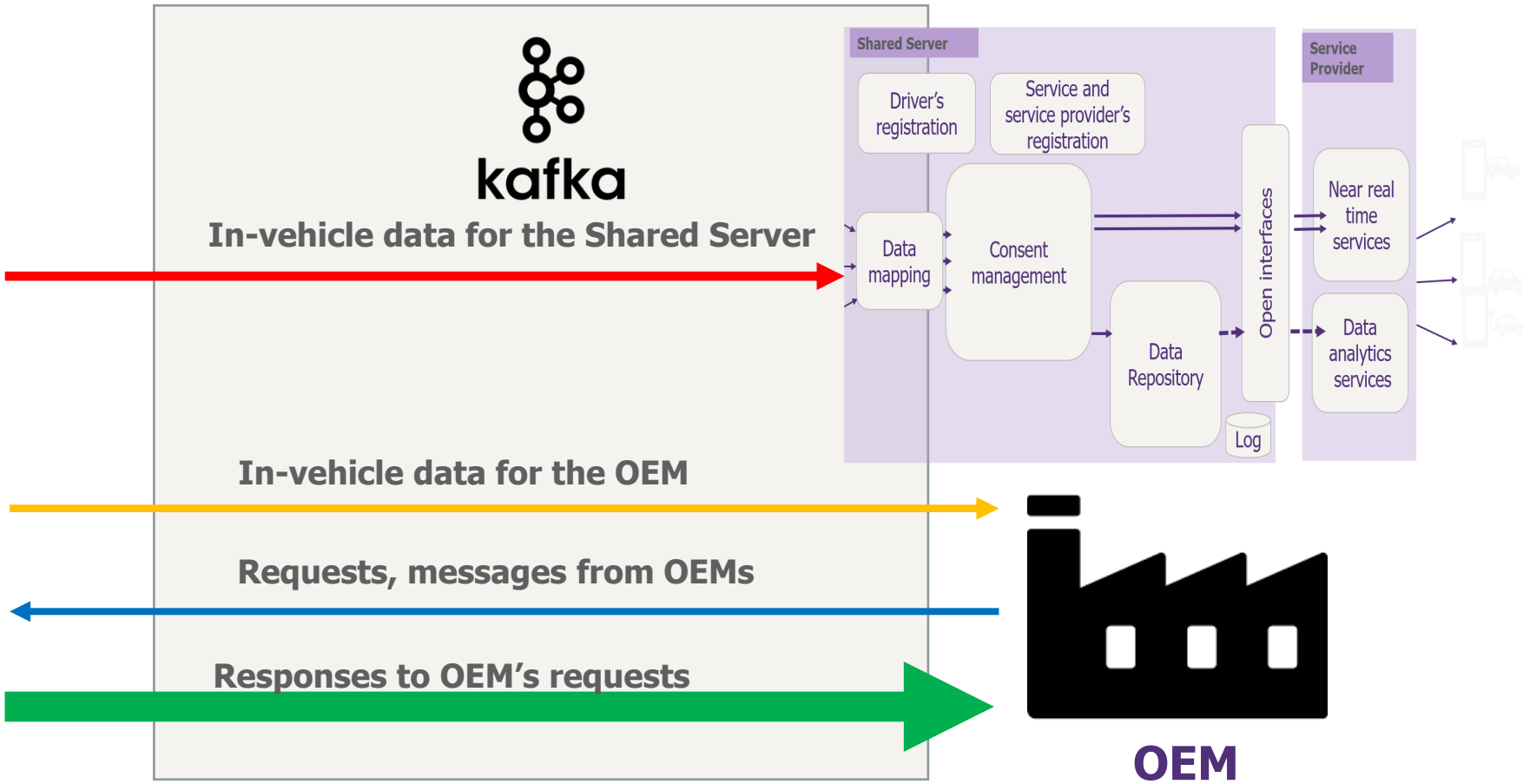




Let's take a closer look ...

Fleet of vehicles of specified OEM





Elaborating Anonymization

/ 03

Anonymised Data - **a complementary data product**

In our pilot, until now, we focused on personalised services. These services **bring great monetisation opportunity** (Hazard location Notification, Park and ride, Insurance) and the GDPR applies asking for a challenging implementation.

The anonymization concerns the secondary use of data and should be applied **on data at rest** or cold data aiming to:

- Public authorities in order to improve traffic management and planning, safety such as road conditions, roadblocks and traffic flow data;
- **Researchers for contacting Origin-Destination (OD)** [illegible] of **traffic** in an area of interest for a given time;
- OEMs for brand-specific applications, component analysis ar[illegible]

However there is a lack of guidance on which anonymisation techniques is effective in compliance with GDPR, for data sharing on the Automotive sector.

It's a well-known problem with solutions that may vary from naiv[illegible] in term of computation and algorithms. Literature indicates **that da[illegible] therefore, defined on a case-by-case basis**, in particular in connection with the leve[illegible] (Mamoona N. Asghar et. al, 2019)

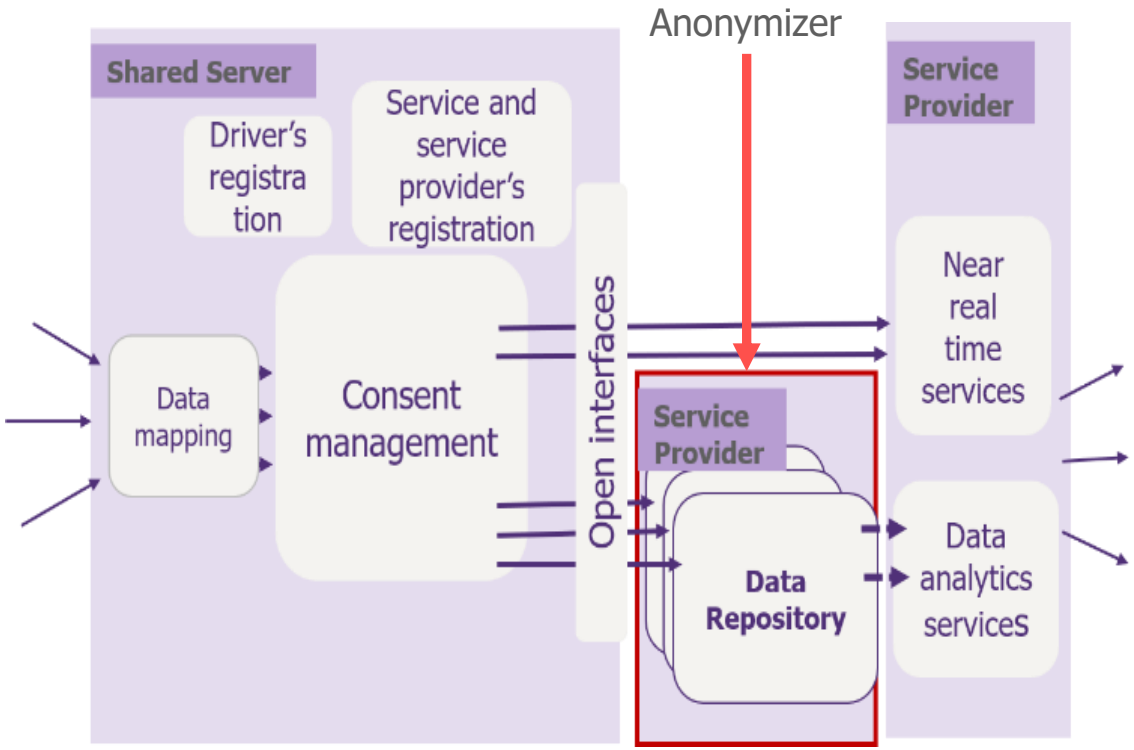
Data and risk of re-identification

The secondary use of data should guarantee the protection of driver's privacy and identity. The data attributes could be categorized into 3 categories (medium, low, zero) regarding the probability of leading in the re-identification of the data owner (driver).

Risk of re-identification	Data attributes
Medium	Longitude, Latitude, Time
Low	Speed, Start/Stop Engine, Acceleration, Engine Load, Engine RPM, G(calibrated)(Concerning force)
Zero	Tyre pressure, Mileage, Oil temperature, Error codes, Fuel consumption instant, Fuel consumption average, Battery charge status, Outside temperature

Anonymizer

Anonymizer is **responsible for the anonymisation** of the data and **protection of the driver's privacy**. It is an **independent Service Provider** and it is not part of the Shared Server. Anonymizer could implement **various techniques** to anonymize the data.
Old data and data for secondary use should be moved to external repositories for curation and further exploitation. Data repository services and anonymisation services for in-vehicle data might be available in open competition for the service providers.

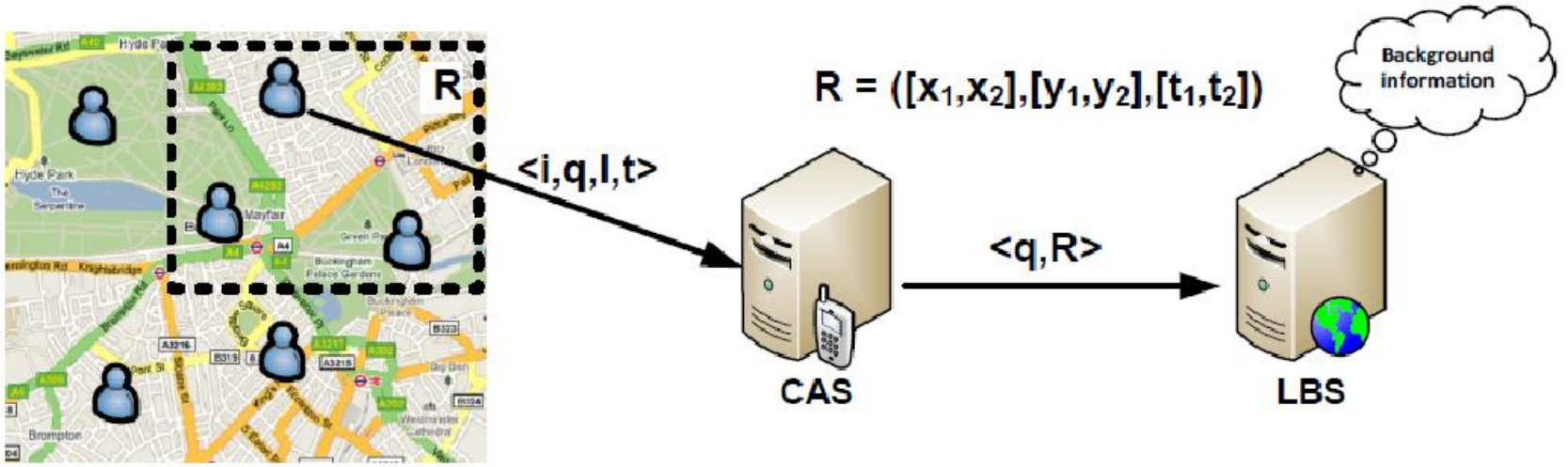


K-Anonymity

- The most well-known anonymization technique.
- A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.
- Proposed in 2002 by Latanya Sweeney (SWEENEY, 2002).

Source: Paper "**Unraveling an Old Cloak: k-anonymity for Location Privacy**"

Example for K-Anonymity on the field of location

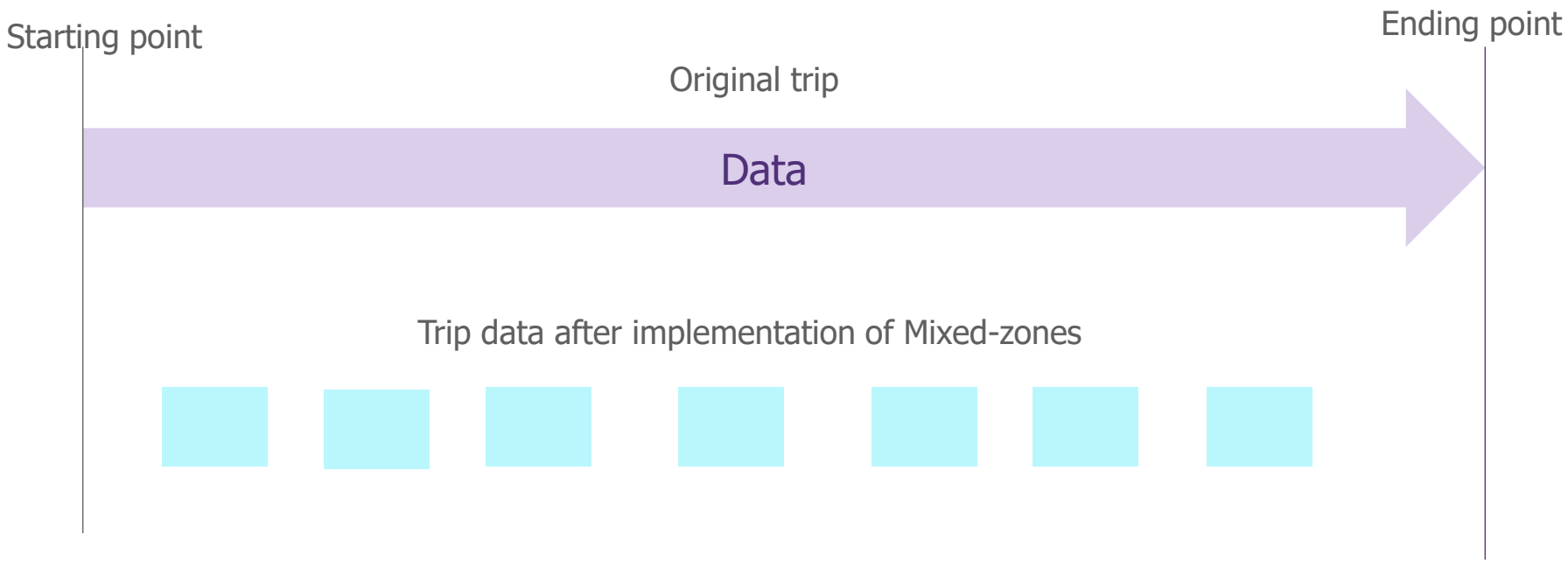


Source: (Shokri et al., 2010)

"There is a set of users who access a LBS (location-based service) through a trusted Central Anonymity Server (CAS). Users send their LBS queries <I, q, l, t> to the CAS, where i is the identity of the user, q is her query, l is her precise location (expressed as a point with coordinates (x, y) in a 2-dimensional space), and t is the time at which the query is generated. In order to protect users' privacy, the CAS removes the identity i of the users. Furthermore, it obfuscates the location l = (x, y) and the time t at which the queries were generated. For this, it constructs a cloaking region R = ([x1, x2], [y1, y2], [t1, t2]) such that there are at least k users (k=3 in the figure) in R whose location l = (x, y) at time t satisfies that x1≤x≤x2, y1≤y≤y2 and t1≤t≤t2."

Mixed-zones: an additional anonymization technique

- Proposed by B. Zan, Z. Sun, M. Gruteser and Ban (Zan, Sun, Gruteser and Ban, 2013)
- The key concept is that a number of portions of location data is deleted, including the starting and ending points of a trip. Longer trips are divided into shorter unlinkable segments.
- Two segments are defined as unlinkable, if it is difficult to predict the missing part of path.




Questions
Answers