|  |  |
| --- | --- |
|  |  |
| **Connected Vehicle Data Privacy Protection**  **Module** |  |
|  | April 5, 2017 |

This page intentionally left blank

# Revision History

|  |  |  |
| --- | --- | --- |
| Revision | Date | Changes |
| 0.1 | April 24, 2017 | * Initial Version |

Table of Contents

[Revision History 3](#_Toc480795398)

[Table of Figures 5](#_Toc480795399)

[Purpose 6](#_Toc480795400)

[Architecture and Interaction with the ODE 6](#_Toc480795401)

[Code and Testing Environment 7](#_Toc480795402)

[Development Methodology 7](#_Toc480795403)

[Obtaining the Code 7](#_Toc480795404)

[Code Documentation 7](#_Toc480795405)

[Details about Dependencies 7](#_Toc480795406)

[Building the PPM and the Testing Environment 8](#_Toc480795407)

[Building the PPM 8](#_Toc480795408)

[Building the Testing Environment 8](#_Toc480795409)

[Building the Code Documentations 8](#_Toc480795410)

[PPM Configuration 8](#_Toc480795411)

[Running the PPM in Standalone Mode 8](#_Toc480795412)

[Running the PPM as an ODE Service 8](#_Toc480795413)

[Appendix A: PPM Operational Details 9](#_Toc480795414)

[Inputs and Outputs 9](#_Toc480795415)

[Collaborative Development 9](#_Toc480795416)

[Important Privacy Considerations 9](#_Toc480795417)

[Appendix B: Glossary of Terms / Acronyms 11](#_Toc480795418)

# Table of Figures

[*Figure 1: Interface between ODE and De-Identification* 6](file:////Users/x4j/Documents/repos/cvdi-geofence/doc/cvdp_user_manual.docx#_Toc480795419)

[Figure 2: Geofence around map segment 9](#_Toc480795420)

# Purpose

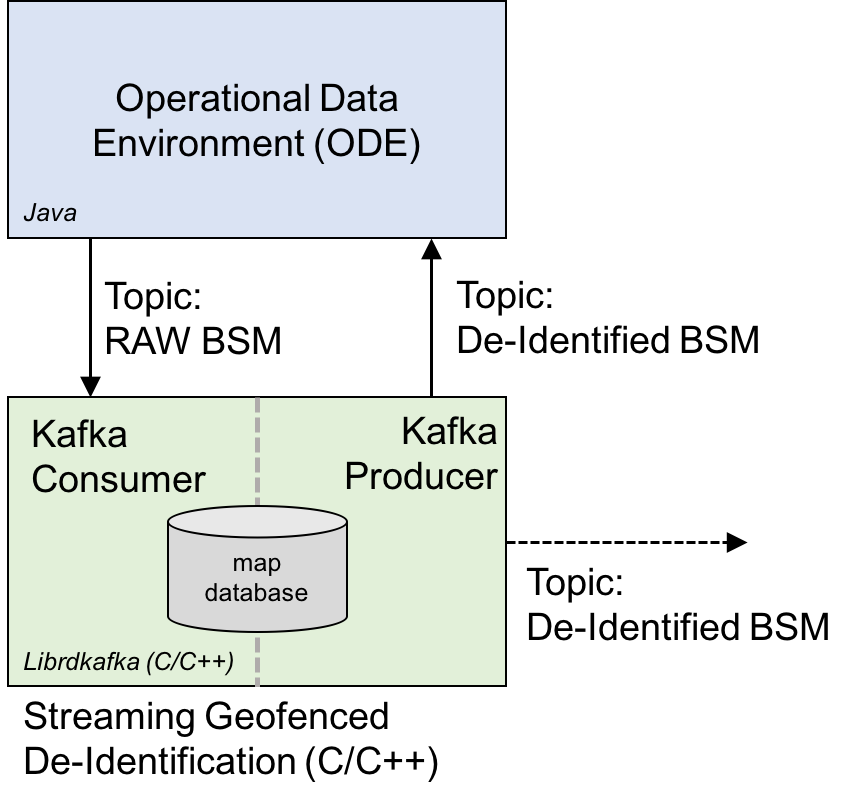
This module filters Basic Safety Messages collected by the Operational Data Environment to enhance the privacy protection provided to data generators participating in Connected Vehicle Safety Pilots. This capability determines whether each individual BSM should be retained or suppressed (deleted) based on the information in that BSM by itself and a well-defined geographical boundary, or geofence. BSM geoposition (latitude and longitude) and speed are used to determine the disposition of the BSM. Additionally, the procedure redacts the BSM identifier for a specified set of vehicles.

By eliminating BSMs whose speeds are below or above prescribed speed limits, messages that may be associated with accidents or violation of posted speed limits are suppressed.

By eliminating BSMs whose locations are outside of a prescribed geofence, messages that may be used to learn information that could be used to identify the identity of an individual driver are suppressed. It is important to note this protection is predicated on the geofence describing an area where loitering or stopping at locations that can be used to infer personally identifying information, e.g., homes or businesses.

This module is specifically designed to handle data streams gathered by the Operational Data Environment (ODE).

# Architecture and Interaction with the ODE



#### *Figure 1: Interface between ODE and De-Identification*

This module performs its function as a separate compute process. It interacts with the ODE through the distributed streaming data platform, Kafka. The privacy module subscribes to an ODE topic that streams raw BSMs in JSON format. The privacy module filters each BSM it receives in this stream. It then publishes the BSMs it retains, with designated fields redacted, to a new Kafka topic. This simple architecture makes the privacy module independent of the ODE. Figure 1 illustrates the basic interactions between the privacy module and the ODE.

# Code and Testing Environment

## Development Methodology

The Privacy Protection Module (PPM) is being developed under Agile Development Methodologies, using an open architecture approach, in an open source environment. All stakeholders are invited to provide input to this document and the PPM. Stakeholders should direct all input on this document to the JPO Product Owner at DOT, FHWA, JPO.

This document is a living document and will be updated throughout the life of the JPO ODE project that this module is designed to support to reflect the most recent changes in the ODE design and stakeholder feedback.

## Obtaining the Code

PPM code is available at: <https://github.com/usdot-jpo-ode/jpo-cvdp>

PPM code and its test environment is maintained as a Git repository. The Git source code version control system and available clients is described and can be obtained at the following website: <https://git-scm.com/downloads>.

The PPM repository includes everything you need to build and test the code. The information sources below provide much more detail on PPM dependencies.

## Code Documentation

The privacy module code is documented using Doxygen and attempts to follow Google’s C++ style guide. Information about Doxygen can be obtained at the following website: <http://www.stack.nl/~dimitri/doxygen/>.

## Details about Dependencies

The privacy module is written in C++11. It uses the following open source libraries:

* Librdkafka (<https://github.com/edenhill/librdkafka)>: an open source (BSD License) C/C++ Kafka client. Librdkafka is an active open source project with many contributors. It has been used in commercial settings.
* RapidJSON (<https://github.com/miloyip/rapidjson)>: an open source (MIT License) C++ Javascript Object Notation (JSON) parser. The library is one of the fastest open-source JSON parsers available and it is fully compliant with the RFC7159/ECMA-404.

# Building the PPM and the Testing Environment

## Building the PPM

## Building the Code Documentation

## Building the Testing Environment

## PPM Configuration

## Running the PPM in Standalone Mode

## Running the PPM as an ODE Service

# Appendix A: PPM Operational Details

This application will perform rule-based de-identification. In other words, it will suppress/retain individual BSM records (Part I and optionally Part II) and redact values assigned to certain fields. Suppression will be based on the following BSM Part I fields:

* Geoposition (i.e., latitude and longitude)
* Speed

The following fields will have their values redacted in an agreed upon way (replace the value with a well-defined “null” value):

* Static vehicle identifiers within BSM Part II.

The fields selected to perform suppression and redaction are subject to further refinement.

BSM records will be geofenced using the geoposition attribute, an auxiliary map database, and explicit geofences based on map road segments. The map database will contain road segments specified by WYDOT (e.g., initially the I-80 corridor). OpenStreamMap data will be used (open source). Geofences will be parameterized and the overall geofence region can be altered by adding or removing segments.



Figure 2: Geofence around map segment

## Inputs and Outputs

A streaming de-identification subscriber (consumer) will use an ODE topic as its input (e.g., a Raw BSM stream). A streaming de-identification producer will publish a de-identified BSM topic. The output topic will have the same format as the input records with redaction / suppression occurring as described above.

## Collaborative Development

ORNL is open to either placing this code on GitHub as part of the ODE or as a separate open source project.

## Important Privacy Considerations

The following points regarding data privacy and this system are worth considering:

* **The rule-based approach described here does not provide adequate privacy protection in general**. In this case, the geofenced roads do not directly connect to locations generally considered private (e.g., homes or businesses). Velocities outside of a well-defined dynamic range are considered abnormal within this geofence; therefore, we assume that behavior is due to traffic or drivers exceeding posted speed limits. The velocity rules are designed to protect against identifying vehicles involved in accidents and breaking the law.  
    
  If the geofence described here (i.e., I-80) is expanded to include additional roads, or BSM databases that geographically intersect this geofence are made available, additional privacy protections are needed to avoid inference-based attacks. In general, once data has been publically released it is not recallable.
* As previously discussed, the ODE will produce topics that are “sanitized” and others that do not provide any privacy protection. Security controls (authentication) will be needed to ensure arbitrary subscribers cannot gain access to “unsanitized” topics.
* The abstractions (i.e., how and where data is stored and distributed) provided by big-data platforms like Kafka make them extremely useful and easy to scale. Understanding who has access to specific topic partitions in memory and on disk should be considered especially since this data is replicated and distributed. Theses resource could be accessible without using any of the ODE’s / Sanitization interfaces.

# Appendix B: Glossary of Terms / Acronyms

|  |  |
| --- | --- |
| Term / Acronym | Description |
| BSM | Basic Safety Message |
| PPM | Privacy Protection Module |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |