Database Log Files Privacy Assessment and Provenance Generation

Dom Hall- Supervised by: Konstantinos Markantonakis, Raja Naeem Akram
Information Security Group, Smart Card and IoT Security Centre

The Smart Card and Internet of Things Security Centre

Objectives

Digital technologies touch every aspect of modern life. These technologies collect, retain and process information regarding entities that also includes humans. The fundamental construct for retention of information is databases. For reliability and (recently) security purposes databases employ robust logging mechanisms. In this project, we look into logging mechanism of four Database Offerings:

- Select 2 SQL and 2 NoSQL database systems.
- Generate test data that is the same in all 4 database systems.
- Get log files from all 4 databases.
- Analyse these log files for privacy related data and their implications in the context of the GDPR.
- Write a script to translate all 4 log files into one intermediate language.

Introduction

To provide a comparison of log capturing mechanism across multiple databases, the first step is to generate data that is consistent and feed it to the individual databases. For this purpose, I will develop a data conversion utility, that takes data from a synthetic generator – converts it to the target databased data insertion format.

All databases instances, whether SQL or NoSQL will be representation of same data. From these insertion commands, I will simulate data access and updated activities – over all four instances.

All of the above activities will be captured by the databased logging mechanism. I will retrieve log data from these logs and perform:

- Privacy analysis to see whether these log files include Personal Identifiable Information (PII), if so what is the implication of PII in log files in the context of GDPR.
- Comparison of logging techniques and information diversity among selected four databases
- Convert logs to Data Provenance format.

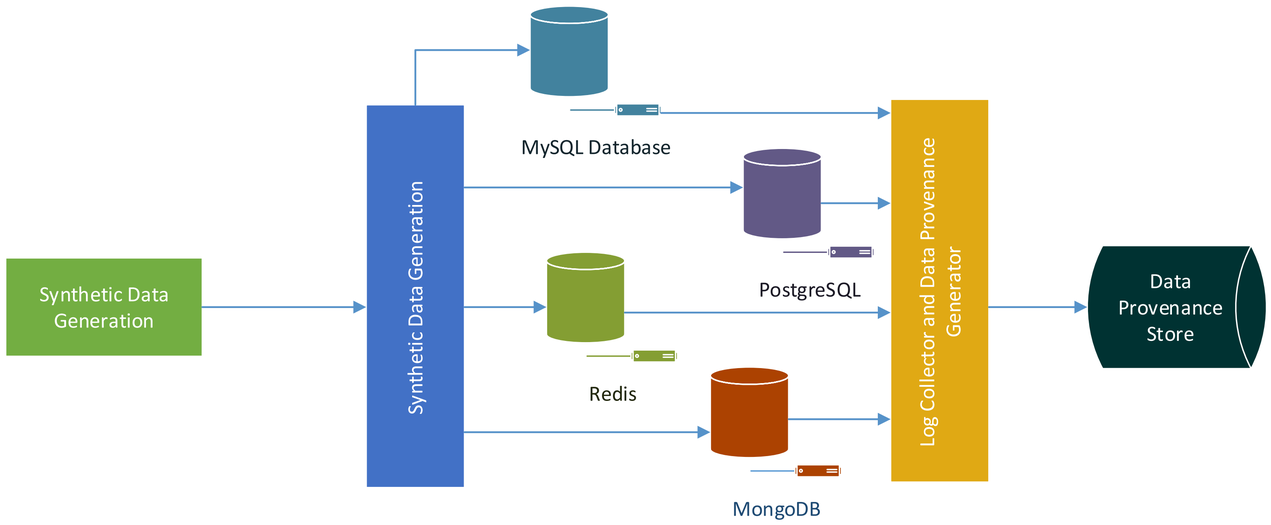

Figure 1: Full Project Generic Overview – Database Logs and Data Provenance

Project Focus

To explore the possibility of converting database log files into provenance data. Furthermore, to compare and constrast the privacy implications of logging shemes used by four selected databases and how this relates to GDPR.

Database Systems

For this project we will be using 2 SQL and 2 NoSQL databases. We'll be using MySQL, PostgreSQL, Redis and MongoDB database.

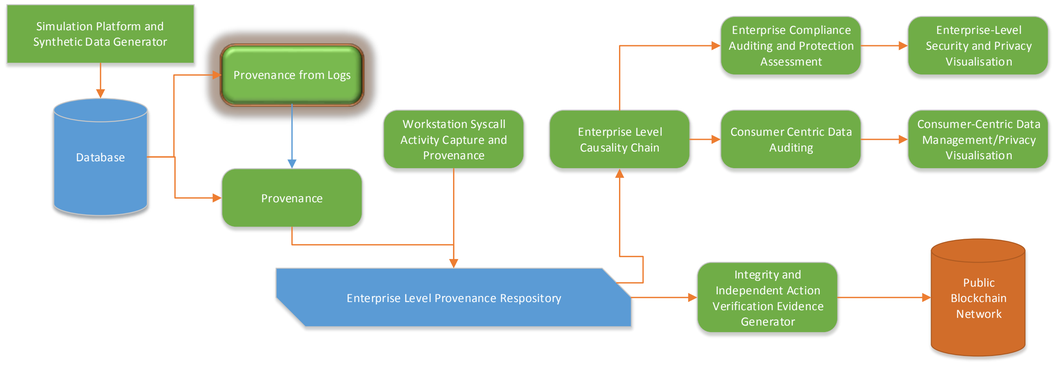

Figure 3: Full Project View with Provenance from Log Aspect

Methods

For the script I plan to use Python. This is mostly because I'm familier with Python however if there is a significant performance difference with other languages I may swap.

Additional Information

This project is related to the EPSRC funded project "Data to Improve Customer Experience (DICE)". The project is particularly interested in personal data, and is using rail passengers as a specific focus of interest. The overall aims of the project are:

- Understand the role that personal data plays in enhancing the user experience of rail passengers
- To develop technical solutions to data privacy
- To develop an evaluation framework that can be implemented so passengers can understand how their data is used and how they can control and verify its use.

The project started in October 2016, and runs for three years to September 2019. For more information about the project, please visit http://www.dice-project.org.

Acknowledgements

We acknowledge the support of the ISG-SCC for the summer internship program and EPSRC funded project. The views and opinions expressed in this poster are those of the authors and do not necessarily reflect the position of DICE project or any of partners associated with this project.

Contact Information

- Web: https://scc.rhul.ac.uk/
- Email: dominic.hall.2016@live.rhul.ac.uk