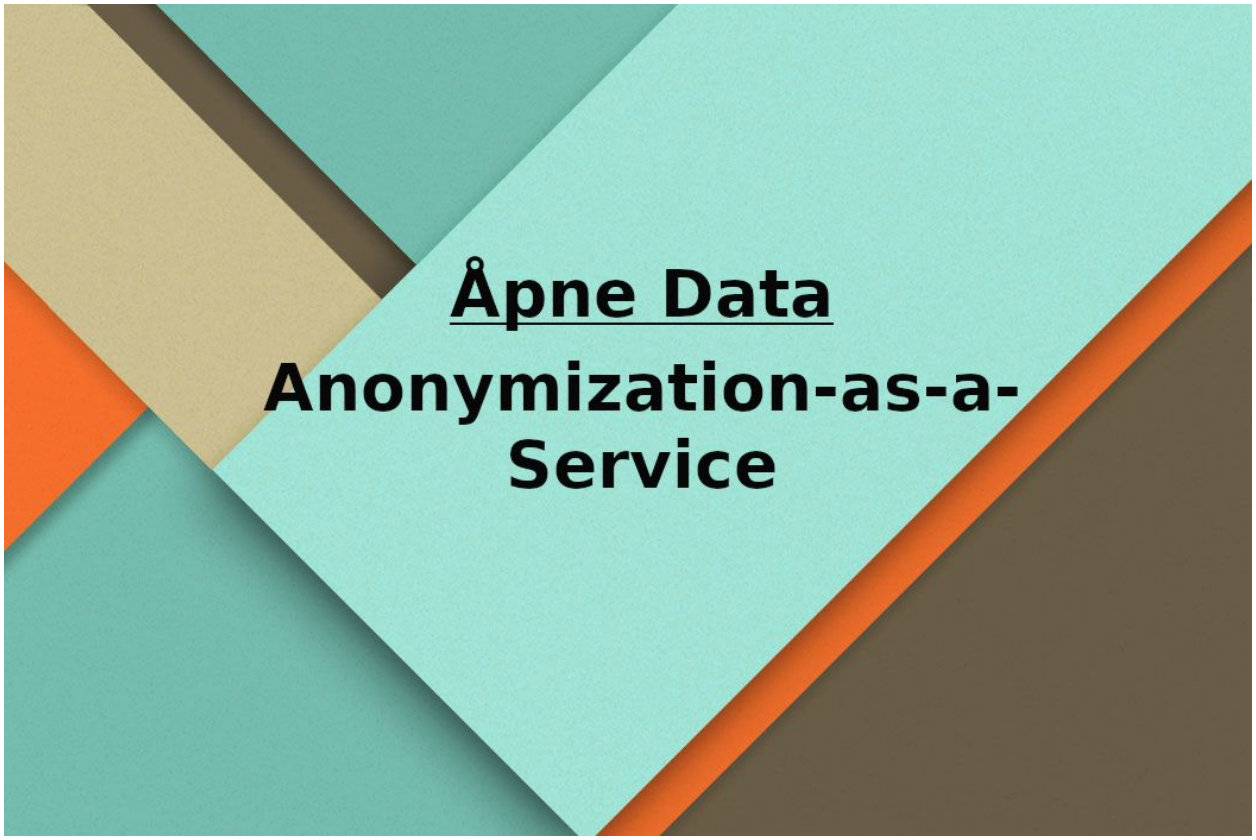

Pre-Project Report

18.01.2019

Group 8

Members	Student Number	Email
Jeremiah Augie Salita Uy	s181369	jeremiah.uy@outlook.com
Lord André Groseth	s181365	andregroseth@outlook.com
Sondre Halvorsen	s305349	sondre.hal@gmail.com
Julian Sagen	s315584	julian.sagen@gmail.com
Viktor Vartdal Johansen	s315615	viktor.v.johansen@hotmail.com

Client

Arbeids- og velferdsdirektoratet
Sannergata 2
0557 Oslo, Norway

Description of the client

Our client for this assignment is NAV IT Data og Innsikt. Data og Innsikt is a department within NAV IT. The department delivers the development of systems and operations of data storage, data processing, data access and analytics.

Person of Contact

Name	Role	Email
Gøran Berntsen	Tech Lead - Åpne Data	Goran.Berntsen@nav.no

Supervisor from OsloMET

Name	Role	Email
Eva Hadler Vihovde	Associate Professor	evahadler.vihovde@oslomet.no

Summary

Our goal is to develop and build a data anonymization service along with a software package for simple user interaction. The solution will provide availability to the already existing anonymization functionality from the open-source ARX software library. The technical delivery is separated into two logical components; a web API service that will deliver ARX functionality through a well documented RESTful API, and one or more API consumer modules that will allow integration with data-science tools. To the author's knowledge, this kind of data-anonymization-as-a-service product does not exist at the present time.

Overview

NAV IT Data og Innsikt has decided that they are spending unnecessarily large amounts of manual labor on data anonymization in their analytic work. Furthermore, the requirements for anonymized data for use in analytics, machine learning, and other data-science activities are increasing. To meet this rise in demand the data anonymization process must be made more efficient, available and user friendly. Data-scientist work is highly valuable to NAV as an organisation, because every measure that can be implemented to create a more efficient workflow is of high value. The ARX data anonymization software project provides best in class implementations of data anonymization algorithms, making it the current "state-of-the-art" solution in the problem space. ARX is currently for practical use, limited to a GUI application and JAVA library with low level APIs. GUI applications are not suited for modern data-science workflows.

Goals

The goal for our project is to improve the accessibility to the ARX library's functionality. To accomplish this we need to extend ARX with a web API service, that follows RESTful guidelines. This will provide scaling and decoupling to our system. In addition we want to create a wrapper package in Python that will provide convenient wrappers for more complex functionality in the web API service. Python has been chosen as the wrapper language at the request of the client. Python is widely used in the data-science community and has a great amount of integration possibilities. The ARX library does not at the present time have a Python integration, and our client expressed that there is a demand in the problem space for such a solution.

1. Create a web service that provides access to ARX's anonymization functionality.
2. Create a Python package that wraps the web service API and provides sensible defaults and convenient abstractions

Specifications

Process

The development team will use the SCRUM framework to organize tasks, work, and project progress.

Functional demands

- **Backend**
  - The system should anonymize the data based on the user's desired anonymity level
  - The system should be able to evaluate the anonymisation level of a given data set
- **Frontend**
  - The Python package should support standard visualisation of the results.

Non-Functional demands

- **Backend**
  - **Platform:** Docker/Nais/Kubernetes
  - **Runtime:** Java/JVM
  - **Rammerverk:** Spring

- **Frontend (Consumers)**
  - Python package
  - Javascript (websites)

Tools

Area	Tool	Description
Version Control System	Github	Github is a git repository hosting platform. Commonly used for open-source software.

Project Management	Asana	Asana is a complete project management tool. KanBan boards, calendars, slack integration, Github integration, tasks and projects.
Collaboration Tool	Slack	Slack is a cloud-based set of proprietary team collaboration tools and services
Programming Language	Python	Python is a powerful, high level language. Used in everything from webapps to data-science/machine learning.
Programming Language	Java	Java is a high-power, stable and highly trusted programing language. Commonly used in backend application with a requirement for stability.
Java Application Framework	Spring	Spring is a open-source, high-power, Java application framework for business application.
User Interface/User workspace	Jupyter Notebook	The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
CI (Continuous Integration)	Travis CI	Travis CI is a hosted, distributed continuous integration service used to build and test software projects hosted at GitHub.
Building and managing Java-based project.	Apache Maven	Maven is a build automation tool used primarily for Java projects.

Static Code Analysis	Code Climate	Code Climate's engineering process insights and automated code review for GitHub and GitHub Enterprise help you ship better software, faster.
Containerization	Docker	Docker containers are the fastest growing cloud-enabling technology and driving a new era of computing and application architecture with their lightweight approach to bundle applications and dependencies into isolated, yet highly portable application packages.

Milestones

The development team is working with a agile process. Milestones are expected to be added as the project matures.

MVP (Minimum Viable Product)

Deliver a usable product to the client composing of a python package available in the PyPI package manager and a ARX web-API service. The system should be able to deliver data anonymization of k-anonymity k=4.

System Diagram