Personal Data Classification Using Natural Language Processing In Object Storage For Security Compliance

Smita Raut
Systems Technology Group
IBM
Pune, India
smita.raut@in.ibm.com

Neha Sontakke *
B.E. Computer Engineer
PICT
Pune, India
nsontakke004@gmail.com

***Abstract*— Security compliance and data privacy regulations have been impacting organizations and IT security has become a ubiquitous business requirement these days. With the EU's General Data Protection Regulation (GDPR), that has been in effect since May 25, 2018, this aspect of IT security has become even more vital. In order to safeguard applications from being vulnerable to breaches, it is desirable that the required security compliance is enforced at the storage side itself. But since storage itself does not understand the business logic and may host all kinds of unstructured data, it becomes a challenge to classify a set of data as personal data that needs to be protected via storage level features such as immutability, encryption, geo-fencing, expiration etc.**

Object storage is an evolving form of storage for unstructured data and is gaining popularity. On the other hand, there has been substantial progress in the field of cognitive computing, artificial intelligence and machine learning which allows deep analysis of unstructured data for pattern recognition, correlation, learning etc.

In this paper, we present a solution for personal data classification at the object storage level using cognitive computing. This paper focuses on textual object data and uses Natural Language Processing techniques to distinguish the text containing personal information.

I. INTRODUCTION

European Union (EU) General Data Protection Regulation (GDPR) compliance involves personal data (article 4, section 1[1]) and its protection by any organization that conducts business with personal data of data subjects, in or from the 28 EU member states. GDPR requirements include data protection, privacy, processing and profiling regulations, cross-border processing regulations on personal data and then performing GDPR duties like obtaining consent and restricting data to its permitted use. Failure in this compliance may end up organizations be liable for a substantial penalty (article 84). Hence, it has become vital to identify if a data contains personal information and should be treated as "personal data".

Typically, personal data resides either in the form of structured data (such as databases) or unstructured data (such as files, text, documents, and so on). One of the challenge is to autonomously identify unstructured data as personal or non-personal data at the storage level itself so that applicable policies can be acted upon only on relevant data thereby optimizing the storage and compute cost.

For this work, we focus on object storage as a way to store unstructured data. It doesn't provide access to raw blocks of data like in block storage, nor does it provide file access to the data like in a filesystem. Instead, it provides access to the whole object or blob of data typically through an API that the system supports. Object storage is ideal for systems that need