Presenters Name: Piyush Khemka & Lokanadham Potnuru

Talk Title: Data classification at Meta scale

Presenter Short Bio:

Piyush Khemka is an Engineering Manager at Meta, currently based in New York, USA. With over eight years of industry experience, Piyush has developed a robust expertise in various domains including Trust & Safety, Payments, Voice Assistants, and Privacy. He graduated from Stony Brook University and has been part of Privacy Infrastructure at Meta since 2022 & is helping with creating a safer and more reliable digital experience for users worldwide

Lokanadham Potnuru has over 17 years of experience in building large-scale systems across eminent companies (Meta, EMC Square, NetApp, Oracle) for privacy, security. Loka Potnuru brings expertise on building robust and reliable solutions that protect sensitive data spanning data storage, cloud computing, data analytics and machine learning.

Short Talk Abstract: Privacy regulatory bodies around the world mandate that companies delete user data within a certain timeframe or prevent use of youth data for ads or ensure employees aren't able to access sensitive user data internally without valid reasons. All these regulations require companies to understand where data is stored & what kind of data is stored within them. With exabytes of data internally, large scale automation of data classification is required to understand data at scale in order to ensure compliance.

To do this, we leverage a novel, at-scale data understanding methodology that leverages both metadata and data. We deploy heuristics and machine learning algorithms to analyze the extracted features and generate predictions about the type of data that is stored in each column or field & label it accordingly. This enables us to take appropriate measures to protect sensitive data and ensure compliance with data privacy regulations. We produce billions of classifications per day with high precision & recall at large scale with limited capacity. In our presentation we will provide an overview, best practices and learnings.