Contribution ID: **511** Type: **Sectional talk**

Decentralised verifiable coded storage for secure distributed machine learning

Thursday 10 July 2025 15:30 (15 minutes)

Modern distributed machine-learning pipelines must juggle three conflicting goals: compact storage, reliable data delivery, and strong security assurances across many independent sites. Traditional approaches achieve only two of the three, typically sacrificing either space efficiency (through triple replication) or verifiability (by trusting storage nodes). We introduce a lightweight overlay that combines high-rate erasure coding with leader-free Byzantine consensus and succinct zero-knowledge proofs. Training datasets are first split into coded fragments, stored as ordinary NeoFS objects, and then anchored by a minimal on-chain ledger that records fragment identifiers and cryptographic commitments. During training, each learner retrieves fragments in parallel, verifies constant-size proofs of authenticity on the fly, and reconstructs the original data even if several storage nodes fail or behave maliciously. The work outlines the core protocols, analyses trade-offs between redundancy, bandwidth, and verification latency. It provides design guidelines for integrating verifiable, space-efficient data pipelines into federated learning and other large-scale AI systems. The resulting architecture eliminates the replication tax while removing single points of trust, paving the way for secure, storage-aware machine learning at scale.

Author: PAVLOVA, Ekaterina

Presenter: PAVLOVA, Ekaterina

Session Classification: Methods and Technologies for Experimental Data Processing