OpenMined PipelineDP

1. proposal

Attributes	Description	Informational
Type	New	
Industry Sector	Data Centers and Data Warehouses of Hospitals, Govs, Teleco and Schools	

Business driver	Differentially private data aggregation framework, pipelineDP, enables write fast, flexible pipelines that use modern techniques to aggregate user data in a privacy-preserving manner.
	
	PipelineDP is a framework for applying differentially private aggregations to large datasets using batch processing systems such as Apache Spark, Apache Beam, and more.
	
	To make differential privacy accessible to non-experts, PipelineDP:
	
	• Provides a convenient API familiar to Spark or Beam developers.
	• Encapsulates the complexities of differential privacy, such as:
	  ○ protecting outliers and rare categories,
	  ○ generating safe noise,
	  ○ privacy budget accounting.
	• Supports many standard computations, such as count, sum, and average.	

Business use cases	1. Schools do machine learning on differentially private data aggregation infrastructure for confidential student information. 2. Hospitals do machine learning on differentially private data aggregation infrastructure for confidential patient information. 3. Gov do machine learning on differentially private data aggregation infrastructure for confidential wage information. 4. Telco do machine learning on differentially private data aggregation infrastructure for confidential consumer text msg and phone call records information.	
Business Cost - Initial Build Cost Target Objective	Edge Cloud should be deployable with more than 3 servers in a single rack at a low cost.	
Business Cost – Target Operational Objective	1. execute this framework on top of disaggregated datasets in Data Centers and Data Warehouses requires little cost. 2. In-place upgrade of the Data Centers and Data Warehouse should be supported without impacting the availability of the edge applications 3. The automation should also support zero touch provisioning and management tools to keep operational cost lower	
Security need	The solution should have granular access control and should support periodic scanning.	
Regulations	The Edge cloud solution should meet all the industry regulations of data privacy and telco standards (NEBS).	
Other restrictions	Consider the power restrictions of specific location in the design (example - Customer premise, where data are stored in School's internal servers)	
Additional details	The Edge Cloud Solution should be deployable across the globe and should be able to support more than 10,000 locations.	Use case submitters can include SQL queries get /set.

2. If the proposal includes a new Blueprint Family include a completed Blueprint Family template specific to the new Family.

Case Attributes	Description
Type	New
Blueprint Family - Proposed Name	Data Privacy Blueprint Family – PipelineDP
Use Case	Differentially private data aggregation
Blueprint proposed Name	PipelineDP
Initial POD Cost (capex)	Unicycle less than $150k: 3 Arm bare metal machines, 1 10G switch

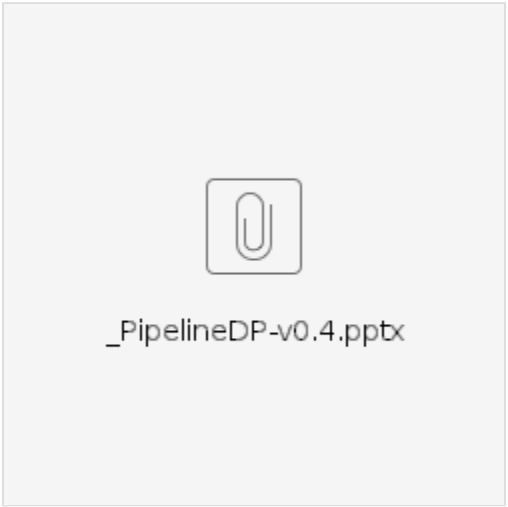
Scale & Type	For the smallest deployment, this requires 2 Arm bare metal machines. For large deployments, this could span to large number of bare metal machines.
Applications	Differentially private data aggregation for large scale online education, telemedicine, Hospitals, Govs, Teleco and Schools.
Power Restrictions	N/A
Infrastructure orchestration	Host: •Orchestrator: Kubernetes •Bare Metal ProvisioningAnsible •Kubernetes ProvisioningKuD •OS: MAC/Linux
SDN	N/A
Workload Type	•Data Center SQL databases Here are some examples of how to use PipelineDP: • Apache Spark example • Apache Beam example • Framework-free example • Example with all frameworks
Additional Details	N/A

If the proposal is to add a new Family to support an existing Use Case please identify the proposed Use Case.

If the proposal is to add a new Species to an existing Family please identify the proposed Family.

In addition add any other material needed to describe the proposal which is needed for the TSC assessment should be referenced or placed in the proposal's page(s).

- Github: https://github.com/OpenMined/PipelineDP
- Website: https://pipelinedp.io/
- API: https://pipelinedp.io/api-documentation/index.html
- Utility analysis: https://github.com/OpenMined/PipelineDP/tree/main/utility_analysis
- Proposal: OpenMined PipelineDP



Committer:

Name	Company	Email
Wenhui Zhang	Bytedance Inc	wenhui.zhang@bytedance.com

Abinav Ravi Venkatakrishnan	deepc GmbH	subramathreya@gmail.com
Chinmay Shah	OpenMined	cs@chinmayshah.xyz