Federated Data Preparation, Learning, and Debugging in Apache SystemDS

Sebastian Baunsgaard ^{1,*}, **Matthias Boehm** ^{1,2,*}, Kevin Innerebner ¹, Mito Kehayov ¹, Florian Lackner ¹, Olga Ovcharenko ¹, Arnab Phani ^{1,*}, Tobias Rieger ¹, David Weissteiner ¹, Sebastian Benjamin Wrede ^{1,2} ¹ Graz University of Technology; Graz, Austria; ² Know-Center GmbH; Graz, Austria; * now at TU Berlin; Berlin, Germany



Spectrum of Data Sharing

 Landscape of Techniques w/ Different Privacy, Utility and Performance Tradeoffs (new markets?)

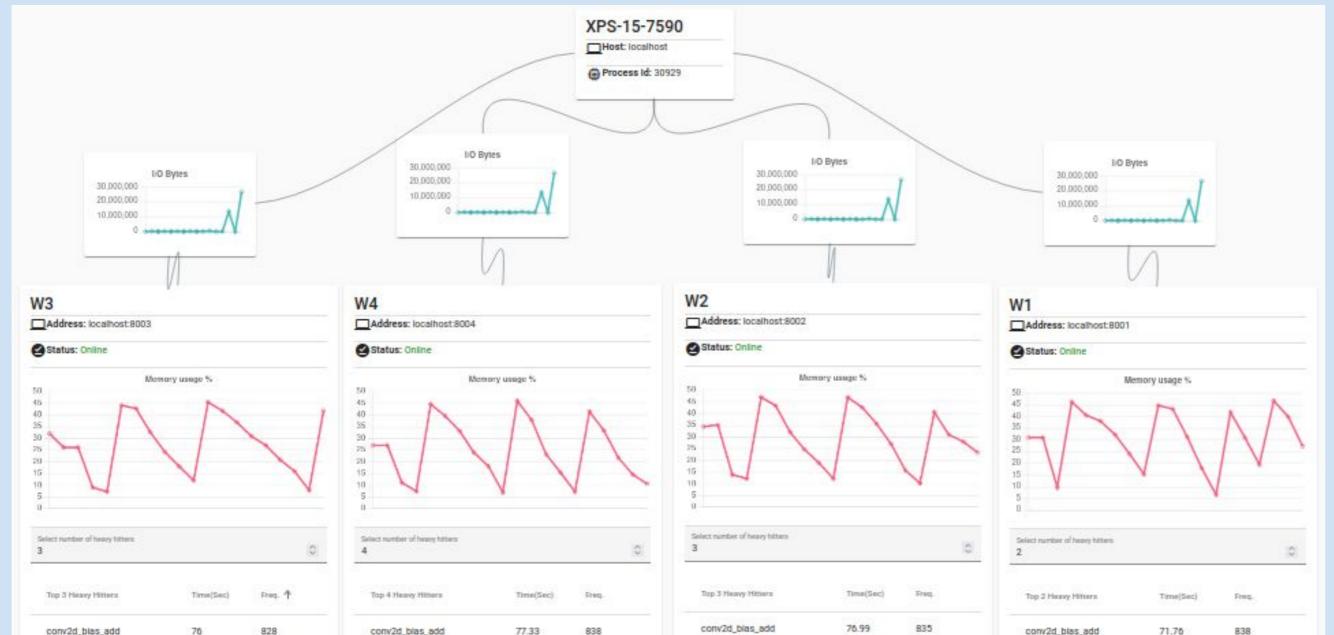
Increasing 🚽	Increasing			
Privacy	► Utility/Perf			
4	`			



Private	Privacy	Aggregates	Anonymized	Surrogate	Public Data		
\mathbf{Data}	Enhancing	(Federated ML,	Data	Data	(full sharing,		
(no	Technologies	Federated $w/$	(k-anonymity,	(properties-	w/ partners)		
$\operatorname{sharing})$	(FHE, MPC,	secure comm.)	pseudonyms)	preserving			
	diff. privacy)		synthetic data)				

Demo: Monitoring Tool

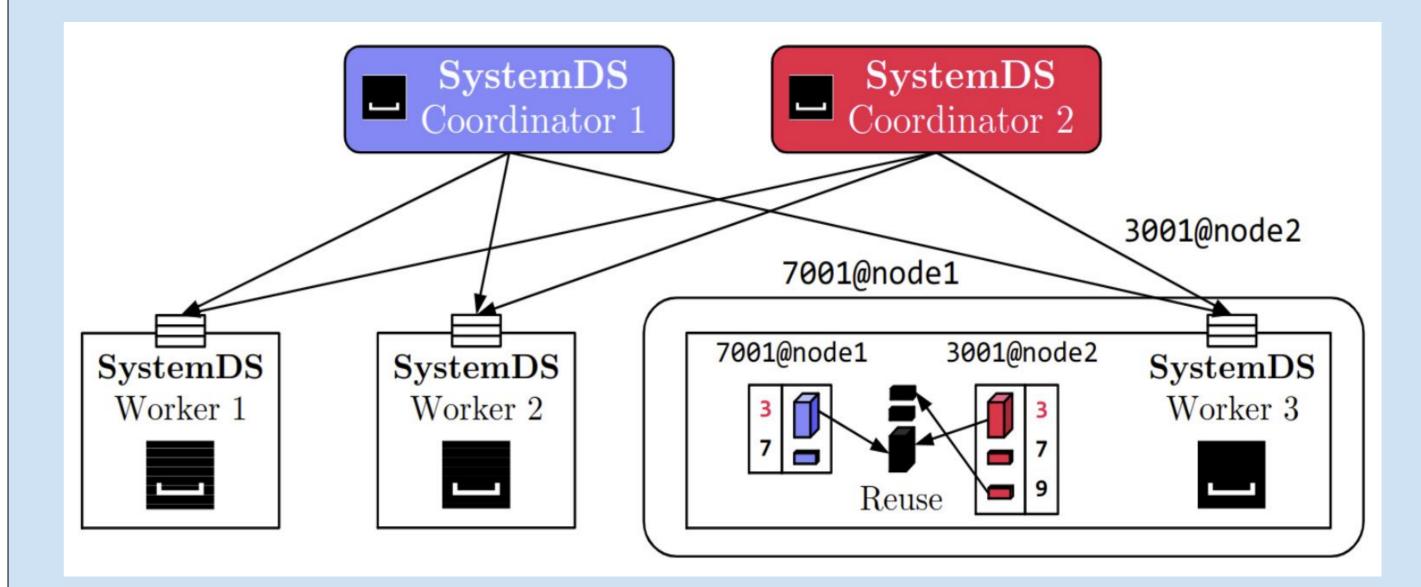
One Coordinator, Multiple Workers:



Data-centric Data Science Lifecycle

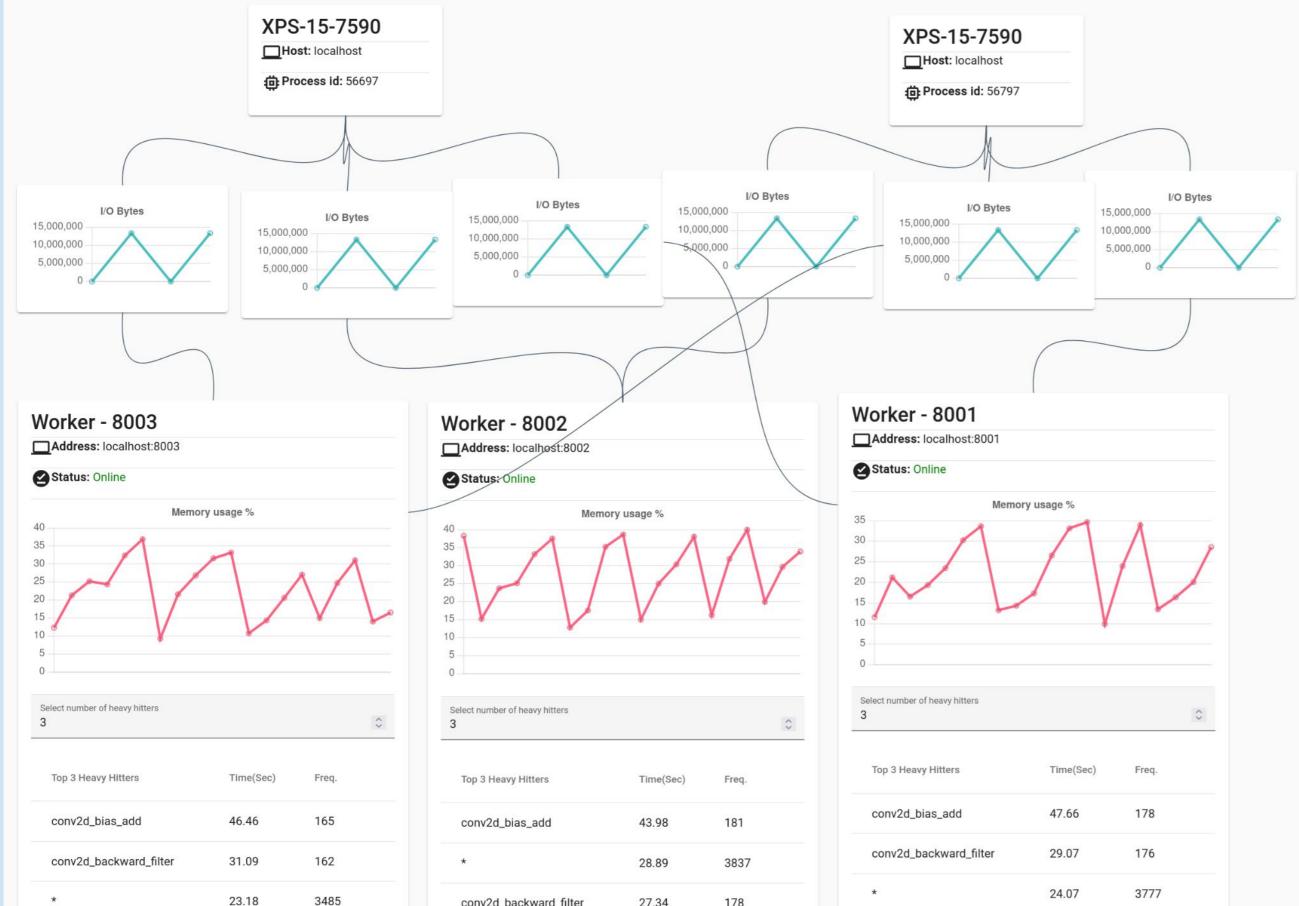
- Federated Linear Algebra & Parameter Servers
- Federated Data Preparation, Feature Transformations Model Training & Scoring, and Model Debugging
- Automatic Compilation of Federated Runtime Plans

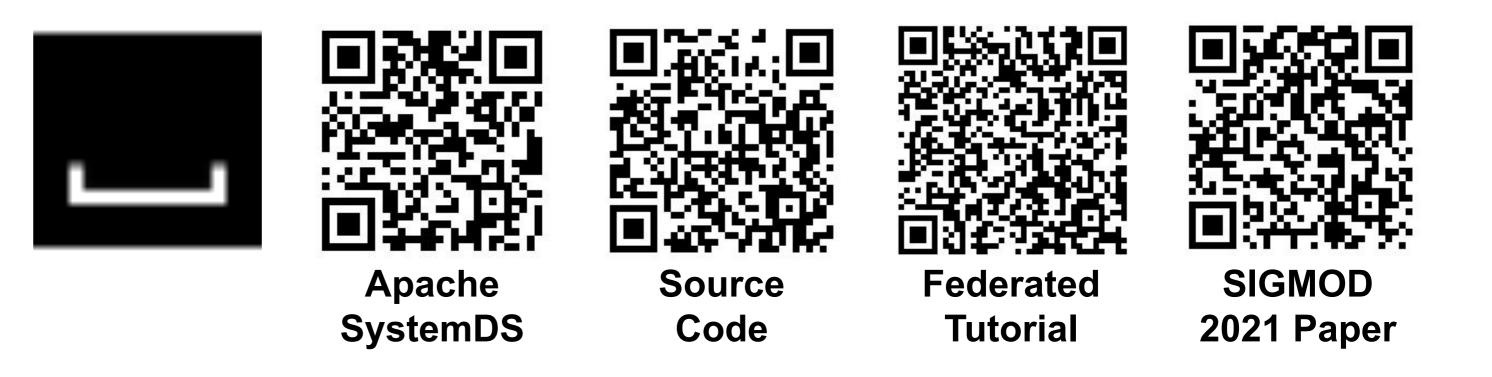
Multi-tenant Federated Learning



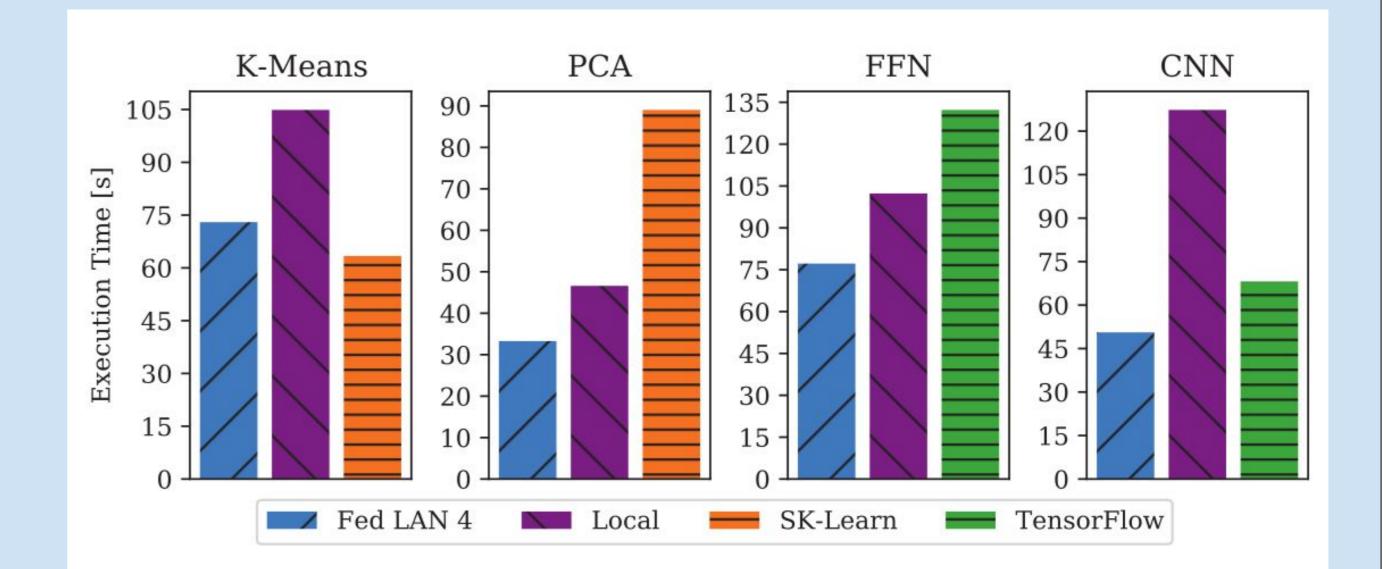
44.75	17916

Multiple Coordinators, Multiple Workers:





Experiments



25.10	5400	convzu_backwaru_inter	27.34	170	

