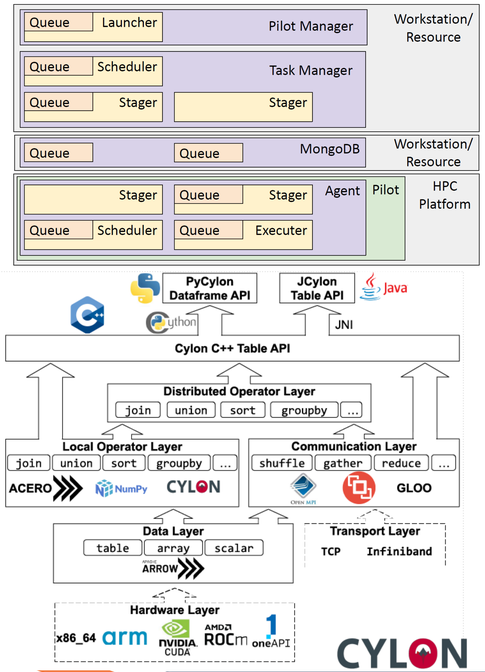
Heterogeneous Data Pipelines for Scientific Computing

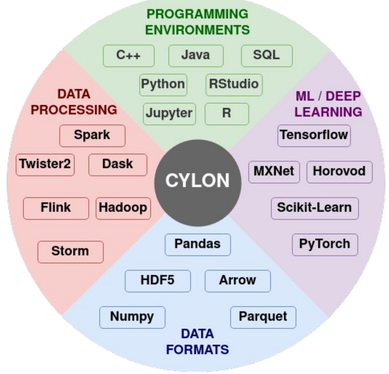
NSF 2200409 (Virginia)
NSF 1829704 (Indiana)

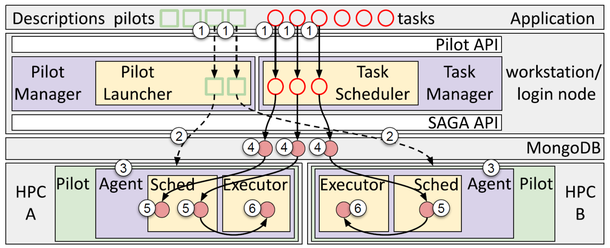
- Data engineering requires high performance data processing frameworks integrating with with ML/DL environment and other frameworks
- Cylon provides a common core with layered architecture for distributed SPMD operations leveraging HPC high-performance communication and data processing.
- Through the new Radical-Cylon integration we enable now also a task based heterogeneous execution.
- Radical-Cylon coordinates now multiple SPMD operations integrating them as functions in the heterogeneous pipelines allowing execution performed based on pre-scheduled resource allocations by the HPC resource manager.

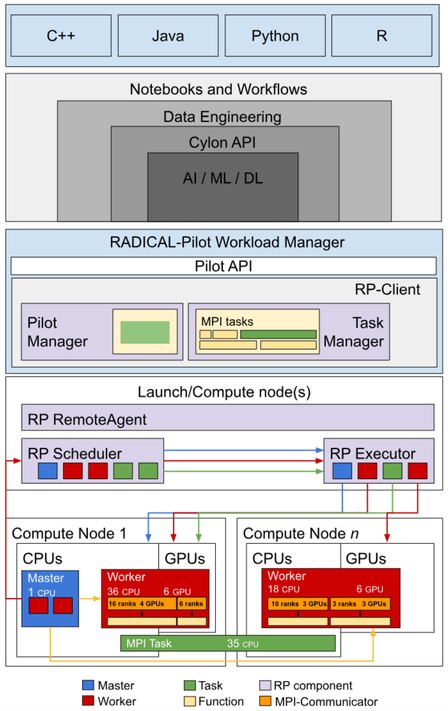
Highlights

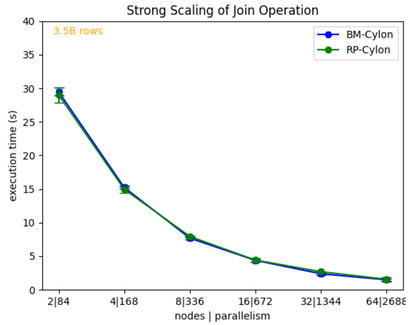
- Leverage Cylon very good performance
  - > 4x over Pandas serially
  - > 150x over Pandas parallelly, >50x gains over Dask/Ray DF
  - > 4x gains over Spark RDDs, Cover ~30% Pandas Dataframe API
- Radical Cylon has overlapping performance graph with baremetal cylon on UVA Rivanna and ORNL Summit in strong and weak scaling test.
- Multiple operations (join, sort, …) when performed in different pipeline reduces 20% of execution time.
- Hands-off processed data to PyTorch DP DL with zero-copy
- Significant technology development: Hybrid HPC, Communication interface; Benchmarking ; Experiment management;
  Online and PDF material publication; Training material selection and production

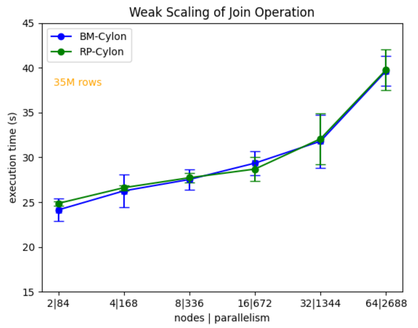


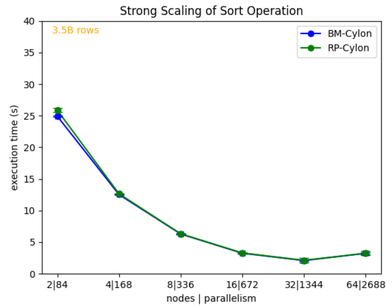


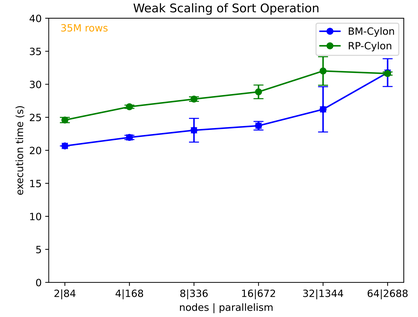












Geoffrey C. Fox, vxj6mb@virginia.edu,
Gregor von Laszewski gregor@virginia.edu,
Arup Sarker djy8hg@virginia.edu,
Mills Staylor, gad5gv@virginia.edu
Niranda Perera
UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE