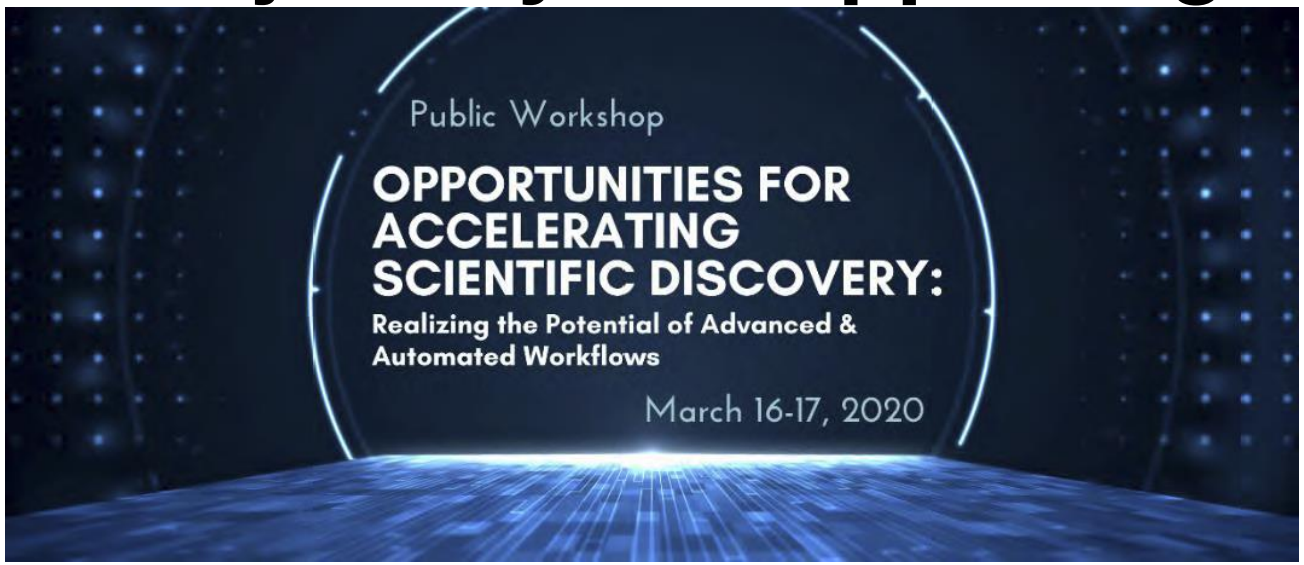
Opportunities for Accelerating Scientific Discovery: Realizing the Potential of Advanced and Automated Workflows
Remarks on Status and Trajectory of Supporting Tools and Systems



Geoffrey Fox 16 March 2020

Digital Science Center, Indiana University

gcf@indiana.edu, http://www.dsc.soic.indiana.edu/, http://spidal.org/

Questions Asked

- What are your observations on the impact of systems and tools to automate workflows, with respect to scientific discovery?
- What is the role of artificial intelligence (AI) and automation in scientific discovery?
  - We'd like you to consider both AI-driven workflows as well as workflows that include AI as part of the scientific discovery process. What are the current challenges faced by these systems?
- What do you see as disruptors that will affect how workflow tools are being used by the scientific community?
- What is your vision for AI in workflows in the future of scientific discovery?
  - What research is needed to advance this area?

Challenges and Futures

- **What are the challenges?**
  - **AI for science** very promising
  - Deep Learning broadly used in industry; many science areas need to examine it.
  - AI for basic cyberinfrastructure useful but modest performance increase – factor of 10 or less
  - **Speedups of up to 2. 10^9** reported for Deep Learning inside HPC applications but only in some fields and needs significant changes in application code
  - Need middleware integrating simulations and AI inside HPC; Implications of new AI accelerators
- **How do you see the future?**
  - **Need to enhance interactions between science and industry**
  - **Common software stacks** in workflow, deep learning
    - Already PyTorch and TensorFlow are used in Industry and Research
    - But can extend with more use of Apache Big Data Software in science
  - **MLPerf** Industry benchmarks can be extended to Science
  - **MLSys** is perhaps a very relevant meeting with strong Industry involvement and dominant use of HPC technologies but currently very little attention from cyberinfrastructure/HPC community
    - No attention to science applications at MLSys 2020
  - Bring **Clouds** and **Supercomputers** closer together
    - Graduate students use Google Colab rather than traditional cyberinfrastructure

Proposal for an MLPerf Working Group in Scientific Data

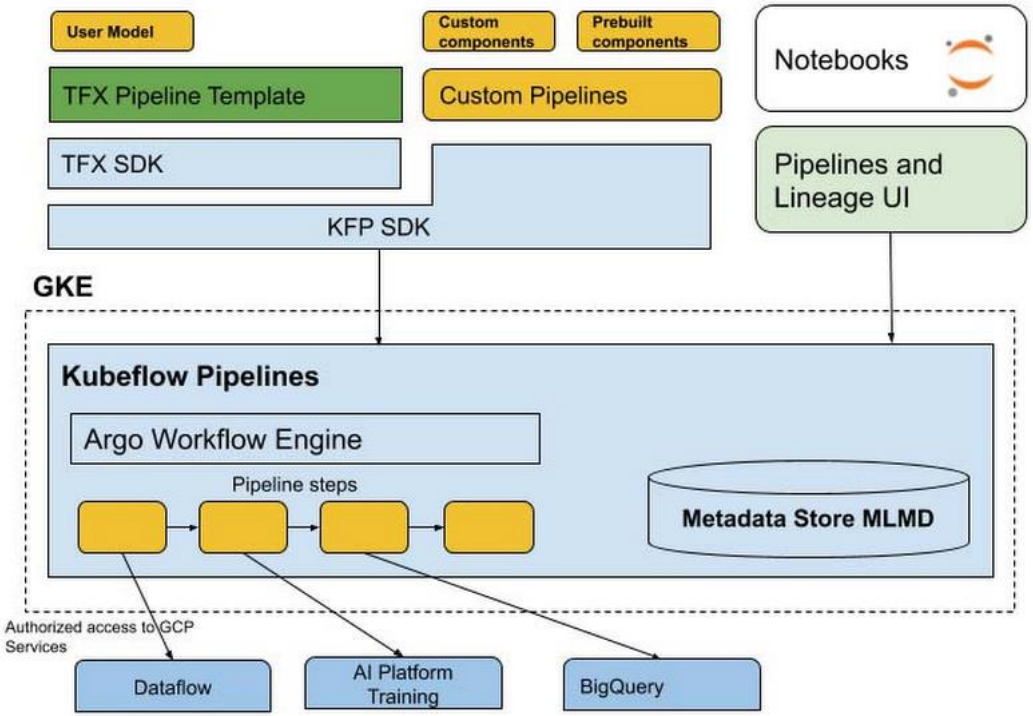
- **MLPerf** has 68 Industry members and academics from CS AI not Cyberinfrastructure; DoE labs strong
  - Results on 1536 GPU's and 2048 TPU's i.e. clear HPC systems but called clouds
- Suggest that **Science Research Data** should use **MLPerf** (presented at March 12 MLPerf Community meet)
- There is **no existing scientific data benchmarking** activity with a similar flavor to MLPerf -- namely addressing important realistic problems aiming at modern data analytics including deep learning on modern high-performance analysis systems.
- Science like industry involves big data with **edge and data-center issues, inference, and training**, There are some similarities in the datasets and analytics as both industry and science involve **image data** but also differences; science data associated with **learnt simulations** and **pandemic time series** and **particle physics** experiments are quite different from most industry exemplars.
- **The best practice science algorithms shifting to deep learning approaches** as in industry today.
- When fully contributed, the benchmark suite could include the following domains: **material sciences, environmental sciences, life sciences, fusion, particle physics, astronomy, earthquake, epidemiology and earth sciences**, with more than one representative problem from each of these domains
- Cover **Accuracy and Speed**. Further, the benchmarks will be **exemplars of modern AI for Science**
- Encourage fruitful comparisons and **collaborations** between scientific fields and industry
- Define requirements for **future cyberinfrastructure** to support scientific data analysis
- **Target users** would be quite broad: academics (including teaching), researchers, scientists and the companies and service providers building systems used by Science.
- Proposed working group has natural links to established MLPerf working groups including **HPC** and **DeepTS**

Workflow (same as Orchestration)

- 271 workflow systems listed at:
- https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems
- About 100 at the these two “awesome” sites:
- https://github.com/meirwah/awesome-workflow-engines
- https://github.com/pditommaso/awesome-pipeline
- Many are domain specific but number illustrates the choices available
- PyTorch and TensorFlow support a dataflow pipeline of deep learning stages
- Apache Beam, Spark and Flink with support of dataflow can implement workflow functions
- Growing link between DevOps and Execution (TOSCA and BPEL) Orchestration
- Exploit containers, Function as a Service, Cloud Native, Pub-Sub events
- Peer to Peer or Central management

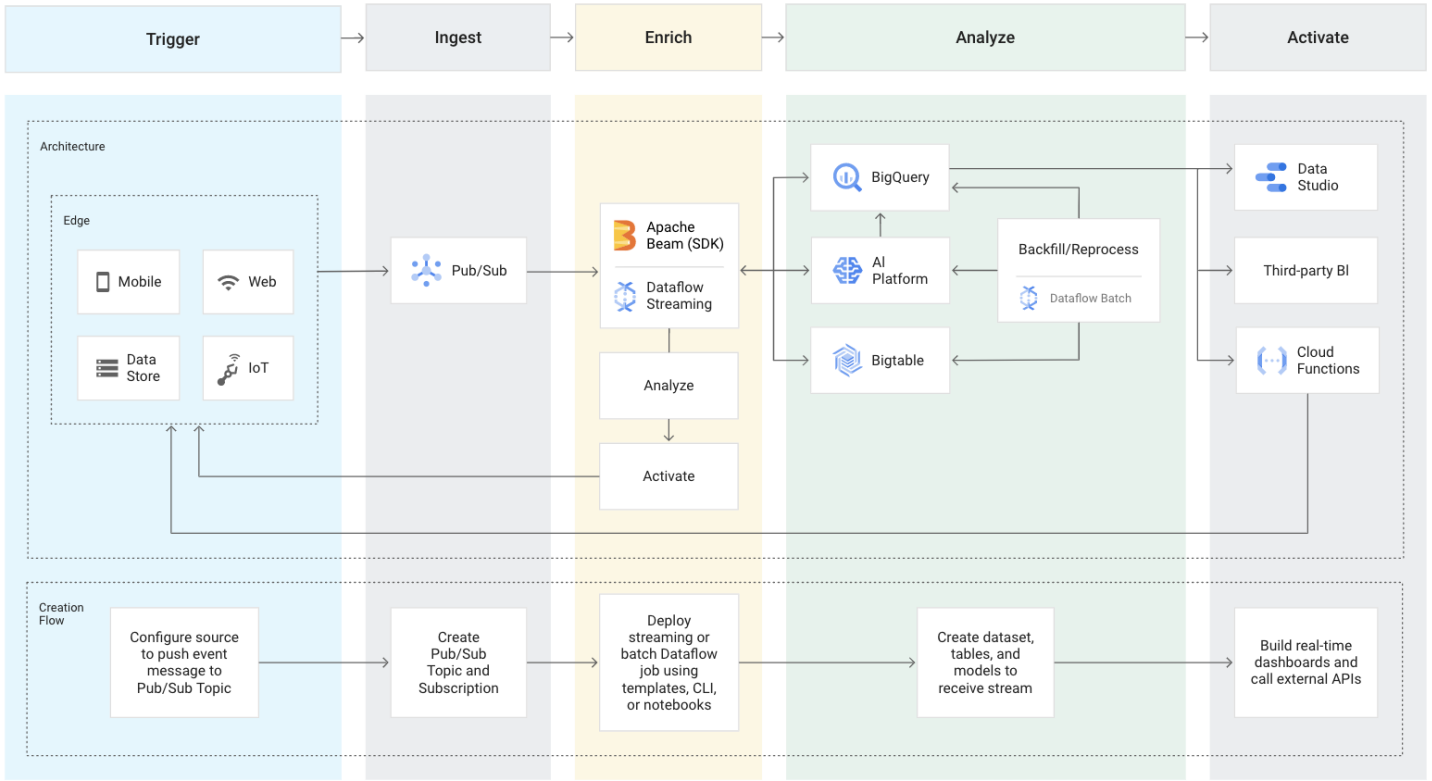
Apache Beam, Kubeflow, Argo: Typical Industry Workflow

- Recent Industry solution aimed at AI workflows constructed as a graph of containers
- March 11, 2020 https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-ai-platform-pipelines



Google Cloud AI Platform Pipelines on Kubernete
General DAG connected containers

Full edge to cloud workflow



Futures of MLforHPC I

- **MLforHPC** succesful but current **use is nonuniform** across domains
  - We need to improve Cyberinfrastructure support to make MLforHPC more effective for **more users**
- Use of modest DL network to **map material/potential drug structure to properties** (generalized QSAR) with simulation and observation: Advanced Progress
- Learn **surrogates** for **large scale simulations**: initial results
- Use of MLforHPC in **agent-based systems** (learn agents replacing by surrogates): Very promising but few results
  - Use in Sociotechnical simulations and in virtual tissues (agents are people or cells)
- **Macroscopic** structure as in learn complex multi-particle **potentials** scaling to N^7: many great successes
- **Learn Collective coordinates and guide ensemble** computations: dramatic progress with speedups up to 10^8
- **Microscale**; learn dynamics of small scale such as clouds, turbulence: Interesting results but much more to do
- Use of **Recurrent NN**'s to represent dynamics (**learn numerical differential operators**): Promising but only studied in small problems

Futures of MLforHPC II

- Learn **errors** as well as values in differential equation solutions
- Look at **advanced solver** methods such as fast multipole or ML learnt potentials
- Quantify **Deep Learning for (geospatial) Time Series** for different applications and different methods: Looking at
  - Industry is Ride Hailing and Transportation as well as audio
  - MLPerf has a Deep Learning for Time Series working group
  - Earthquake Forecasting (are there actually hidden variables as phase transition?)
  - Epidemic Forecasting with mix of simulation and observation
- Investigate **different methods**: RNN (LSTM, GRU), RNN-T, non-recurrent with attention: Transformer
- Build **Software systems** to support AI dynamically mixed with simulation
- Investigate **optimized hardware –** CPU, Accelerator, Storage, Network for AI dynamically mixed with simulation
- How does this **hardware and software compare** to that for
  - Edge to Cloud data-center use case
  - Simulation supercomputer
  - Classic data-center big data problem