AWS Cloud Resources as Part of Scientific Workflows and Research Data Management

Kevin Jorissen[1]

Kevin Jorissen[1], Ben Thurgood[2]

[1]Amazon Web Services, Seattle, USA, jorissen@amazon.com
[2] Amazon Web Services, Sydney, USA, btgood@amazon.com

Abstract

AWS public cloud resources are an increasingly common part of the research landscape. Scientists at research institutions and organizations around the world have used the cloud across all application domains including HPC and ML/AI; life sciences, earth science, astronomy, and humanities. The potential benefits of the cloud include: democratizing access to computing, data analytics, and machine learning; collaborating on very large datasets; meeting Research Data Management and legal compliance requirements; and accelerating the time-to-discovery. These fundamental research needs require the flexible availability ("scalability") of cloud resources, the large number of easy-to-build-on managed compute and analytics/ML services ("agility"), and the global reach, secure nature, and collaborative functionality of AWS infrastructure. There are also areas where much potential remains untapped due to policy, procurement, and budgeting hurdles, as well as the need to build institutional and individual researcher cloud skills and rearchitect workflows.

In this talk we provide an analysis of the state of the art in cloud computing for research, of current challenges, and of how to get the most research value out of AWS.

Access To The Right Resources For Better And Faster Science

Scientists are often hampered by the constraints of on-premise infrastructure: for example, the compute cluster is too small to fit the desired simulation or analysis; or lacks GPUs, FPGAs, or high memory-to-core ratios; or lengthy application processes or queue times set back the research project. By contrast cloud infrastructure can be requested within minutes, and adapted to evolving needs (e.g., switched from CPU to GPU) at any time.
For example, researchers at Clemson University needed to solve a vast Natural Language Processing problem far outstripping local resources.[5] They created a High-Performance Computing cluster on AWS, scaled it out to over 550,000 compute cores, used it for several days, and then dismissed the entire system. This is comparable in size to using an entire national supercomputer. The work was done using so-called "Spot" servers, heavily discounted spare capacity in AWS' data centers that is very popular for research.

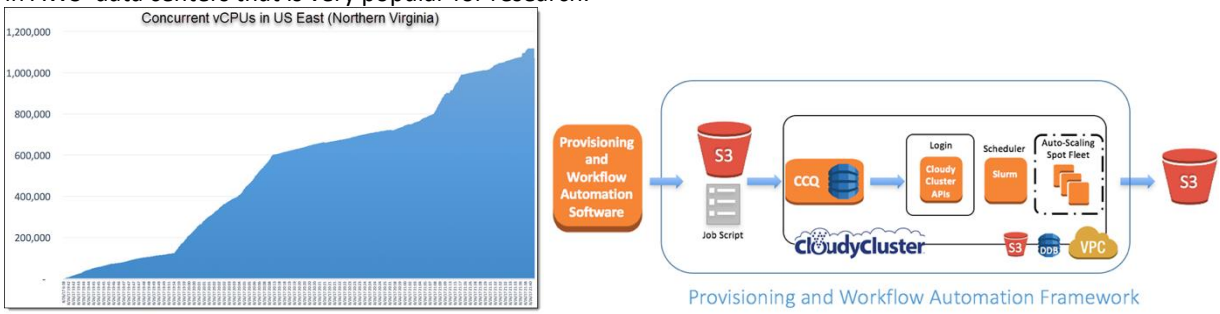

Figure 1 (Left) Clemson cluster ramping up to 550,000+ cores (1.1M vCPUs). (right) architecture of the Clemson cluster.

Similarly, a genetic population study of the koala required 3 million core hours of computing. To speed up the work (allowing the paper to be produced sooner, and allowing conservationists and policy makers to get to work sooner) a large compute fleet was spun up on AWS, crunching the numbers much faster than could have been done otherwise. [3]
In a recent HPC study, researchers at UCSD generated over 1 PFlop of processing in a single, tightly-coupled MPI seismic simulation on AWS. [2]

Access To Very Large Data Sets

The center of gravity in research infrastructure is rapidly shifting from compute to data as datasets become too large to move. For example, CMIP6, the leading reference dataset in climate science, will require ~20PB of storage, making it expensive and impractical to duplicate. AWS not only offers virtually unlimited storage, it also allows data sharing with collaborators in a highly fine-tunable, auditable, and regulatorily compliant way. Collaborators can investigate the data using the dozens of analytics and ML/AI AWS services, or use EC2 servers running custom analytics and simulation tools, with capacity allocation and billing handled between the collaborator and AWS. More than 100 leading datasets in genomics, earth science, and other domains are available free for everyone to use in the AWS Open Data program. [6] The importance of such data lakes will only grow as cross-disciplinary data use becomes more common.

For example, the GOES-16 satellite dataset is part of the AWS Open Data program and is updated in real-time. NASA used this data feed and built a Machine Learning pipeline around it using AWS GPU servers to estimate hurricane wind speeds, cutting the forecast time from 6 hours down to 15 minutes. [1] Biologists used NEXRAD radar observation data to study bird migration. [4] LandSat satellite imagery has been used to classify building types and track urban growth; to estimate crop yields; for biomass estimates; and more.

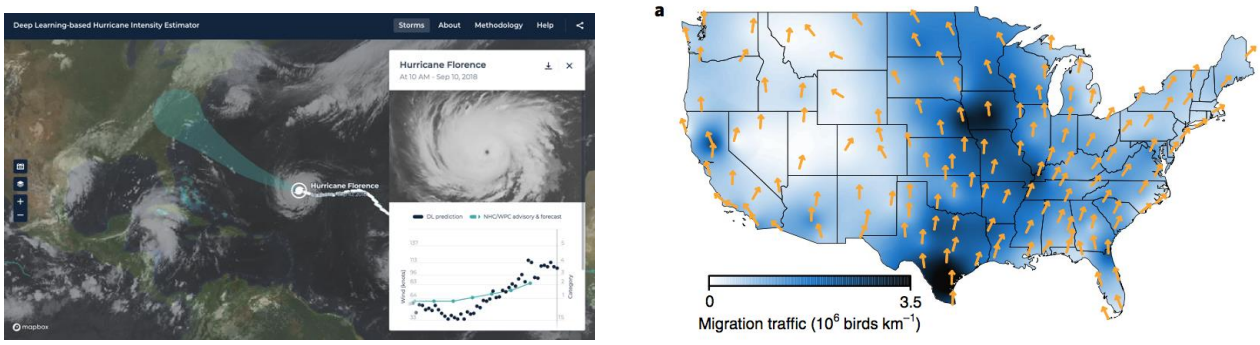

Figure 2 (left) Hurricane paths predicted using Machine Learning on the GOES-16 dataset. (Right) Bird migration patterns derived from the NEXRAD radar observation dataset. - Both fully and freely available through the AWS Open Data program.

Science Not Servers: Managed Services For Computing, Analytics, And Machine Learning/Ai

Scientists want to focus on research problems rather than manage infrastructure. For example, when CSIRO needed a highly scalable CRISPR analysis tool, it built one in three weeks using "Serverless Computing", where the algorithm is provided to AWS and AWS figures out how to run the code, freeing CSIRO from managing servers. [8] The resulting application is quick to build and easy to maintain; scalable, and resilient.

SageMaker meets the ML/AI needs of researchers who are comfortable preparing data and training small models in a Jupyter notebook on the laptop, but need to scale up to a publishable, real-world-accurate model: the researcher works in the notebook on AWS but AWS "magically" deploys the training job to a large GPU cluster in the background to train the model on a massive training dataset. [7]

References

1. Deep Learning-based Hurricane Intensity Estimator, http://hurricane.dsig.net .
2. Petaflop Seismic Simulations in the Public Cloud, A. Breuer, Y. Cui, and A. Heinecke, ISC High Performance 2019 conference proceedings, Springer, 2019.
3. Adaptation and conservation insights from the koala genome, R.N. Johnson et al, Nature Genetics 50, 1102, 2018.
4. Seasonal abundance and survival of North America's migratory avifauna determined by weather radar, A.M. Dokter et al, Nature Ecology & Evolution.
5. Natural Language Processing at Clemson University, https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/ , 2017.
6. https://aws.amazon.com/opendata/ ; https://aws.amazon.com/earth/ .
7. https://github.com/wleepang/sagemaker4research-workshop
8. https://aws.amazon.com/blogs/aws/genome-engineering-applications-early-adopters-of-the-cloud/