DRAFT – Using Cloud Computing for Academic Research

Mahmoud Parvizi
Institute for Cyber-Enabled Research, Michigan State University

Introduction

Cloud based computing platforms offer a number of basic and advanced, off-campus commercial services applicable to cyber-enabled research spanning a broad spectrum of academic disciplines and computational needs. The body of this essay serves as a brief summary of the general concept and core features of a cloud service and is intended primarily to inform research groups at Michigan State University (MSU) with specialized data management needs, groups interested in augmenting on-campus high-performance computing (HPC) via a hybrid approach or groups assessing alternative HPC solutions. Included are a few compact examples of Spartan led cloud-based research efforts. Nontrivial considerations related to the challenges of cloud-based workflows are presented in the Discussion and links to additional resources follow.

Characteristics of Cloud-Computing

In contrast to on-campus academic research facilities dedicated to data management and HPC, off-campus cloud service providers are profit-oriented commercial enterprises with a primary focus on industry e.g., Microsoft Azure, Amazon Web Services, Google Cloud. These providers offer services via a distributed hardware infrastructure such that the compute and storage devices comprise a global network where individual components tasked to a single workflow need not be colocated. Cloud services are generally offered a` la carte, are available on-demand, may scale in real-time and are nearly continuously updated. Pricing is tiered from economy to premium reflecting user utilization and performance requirements. Hence, cloud-computing affords academic users the option to select the type, tier and duration of service appropriate to their research aims. Core services, as those most relevant to academic research, are summarized in the following paragraphs.

Data analysis is foundational to science and thus data management is a core cloud service in the form of storage, file shares and exploration options. Cloud storage options accommodate both structured data, e.g. spreadsheets and databases, and unstructured data, e.g. audio-video files or social media content, may be configured to securely archive and/or share data with external collaborators and are generally accompanied by sophisticated data exploration applications. Scalable data intake services may be employed to automate both the real-time collection and transformation of data. An

integrated analytics service is then a natural extension of cloud data management platforms and may range from easily accessible preloaded libraries for a do-it-yourself approach to full-service i.e., provider designed and managed, enterprise grade analytics pipelines. A distinct feature of cloud analytics services is the pre-built platform to process and analyze multi-source streaming data in real-time, for example broadcasted audio and video or website clickstreams.

Cyber-enabled research relies heavily on HPC such that computing at scale is itself considered a core cloud service. In the cloud paradigm a user may configure seemingly limitless on-demand compute resources with a customized configuration e.g., choice of processors, networking, operating system and libraries. However, users must also understand that the standard research practices of HPC workflow development and implementation associated with dedicated on-campus compute facilities do not directly translate to cloud HPC practices and will require substantive adjustments in the areas of workflow design and management.

In addition to enterprise-grade data management and HPC at scale, cloud service providers offer a host of pre-built machine learning frameworks such that academic researchers with little-to-no machine learning expertise and experienced data scientists alike may customize on-demand and automated model building tools ranging from standard regression to deep-learning. Applications and agents empowered by cloud-based artificial intelligence allow for the deployment of cognitive services such as computer vision, natural language processing, advanced text and document analysis as well as metrics-based anomaly detection and forecasting. Managed bot development services are also available to those researchers in need of an automated interface with a broader community.

The cloud service known as orchestration i.e., the managed integration of each of the aforementioned core services in order to develop end-to-end automated workflows, is the aspect of cloud-based computing that most clearly separates the cloud provider from the on-campus HPC archetype. Orchestration services allow for automated workflows to be executed on a schedule, on-demand or event-triggered while monitoring for progress and/or errors across one or more environments. Automatic scaling and streamlined upgrades are typical features of an orchestration service and allow academic researchers to focus on the workflow without the need to manage job scheduling constraints specific to the underlying computing infrastructure. However, it must once again be stated that transitioning traditional on-campus HPC workflows to a cloud orchestration service requires nontrivial adjustments to workflow development practices.

A few examples of real-world Spartan led research in-progress are now used to illustrate key aspects of cloud utility. First is an MSU based group that seeks to provide the means for real-time inference to researchers by preparing to employ a series of core cloud services intended to collect and analyze live-stream, high-resolution audio. The next example is a small group that uses desktop computing to analyze relatively small batches of archived data; however, newfound access to a large repository of government curated data requires data management and HPC practices outside the scope of a desktop environment. Here, on-demand cloud compute services are intended for infrequent, yet computationally intensive analysis. A project requiring the content analysis of audio-visual media is the final example. This MSU research group is developing cloud workflows in order to orchestrate end-to-end pipelines that collect and manage large datasets of archived video using computer vision and natural language processing for analysis by machine learning algorithms in order to identify desired exemplars.

Discussion

The paradigm of cloud-based computing for academic research provides a genuine solution to the requirements of those research groups with nontraditional data collection and management needs as well as those groups seeking a hybrid or alternate approach to on-campus HPC. However, cloud-based computing should not be viewed as an elixir as many challenges arise. Most prominent are those associated with translating on-campus HPC workflow development practices to a commercial cloud environment.

Adapting traditional workflows practices to the cloud requires a shift in the group's research culture. Workflows developed and implemented with MSU sponsored on-campus HPC resources do suffer from scheduling and scaling constraints yet enjoy a fixed cost structure that lends itself to the traditional academic approach to research. In contrast commercial cloud services are scalable and available on-demand but employ a variable cost structure oriented towards the streamlined business practices of industry. For example, an on-demand HPC job in the cloud is billed per CPU hour; while an orchestrated compute resource in a cloud workflow is billed according to the memory it consumes per execution, the number of executions performed and the duration of each execution. Estimating these variable costs for budget proposals requires a research group to obtain a detailed understanding of resource requirements and comprehensive usage metrics associated with all aspects of the workflow.

Practices related to responsible research e.g., reproducibility, data security and regulatory compliance need also adapt. In the cloud paradigm the computational infrastructure with which research is conducted is in general opaque. Orchestrated workflows are configured

by the service provider such that underlying upgrades and additions are nearly continuous. Here, a research group should develop new standards for cloud-based version controls and the documentation of methods associated with preloaded cloud platforms and tools. Cloud services associated with data security and compliance are generally robust and include solutions for most export control regulations and federal privacy act compliance, e.g. HIPPA, FISMA and FERPA; however, it is the expressly the user's responsibility to ensure that the service tier selected is properly configured to sufficiently satisfies all security requirements.

Though challenging, evolving an established research culture acquired from the benefit of an on-campus HPC facility is not impossible and the importance of advancing scientific research is a strong motivator. In this regard commercial cloud services present those with a developing need for nontraditional data management as well as those assessing a hybrid or alternative approach to HPC with an opportunity to determine the applicability and utility of distributed architecture to their research aims.

MSU Resources

MSU Institute for Cyber-Enabled Research - https://icer.msu.edu/contact
MSU IT Services, Cloud Services - https://tech.msu.edu/network/cloud-services/
MSU Research Technology & Support - https://tech.msu.edu/service-catalog/research-technology-collaboration/

MSU IT Cloud Service Partners

Microsoft Azure - https://azure.microsoft.com/en-us/overview/
Storage - https://azure.microsoft.com/en-us/product-categories/storage/
Analytics - https://azure.microsoft.com/en-us/product-categories/analytics/
HPC - https://azure.microsoft.com/en-us/product-categories/compute/
ML & AI - https://azure.microsoft.com/en-us/overview/ai-platform/
Orchestration - https://azure.microsoft.com/en-us/services/functions/

Amazon Web Services - https://aws.amazon.com/
Storage - https://aws.amazon.com/products/storage/?nc2=h_ql_prod_st
Analytics - https://aws.amazon.com/big-data/datalakes-and-analytics/?nc2=h_ql_prod_an
HPC - https://aws.amazon.com/products/compute/?nc2=h_ql_prod_cp
ML & AI - https://aws.amazon.com/machine-learning/?nc2=h_ql_prod_ml

Orchestration - https://aws.amazon.com/managed-workflows-for-apache-airflow/?nc2=h_ql_prod_ap_af