Microsoft Azure for Research Overview

Introduction

Today's researchers have access to a greater variety and volume of data than ever before. Scientific instruments and sensors are capable of generating data around the clock day in and day out, while computer modeling and simulation programs produce rich data sets in lieu of or in conjunction with experiments. Science now includes a computational perspective that requires researchers to increasingly rely on computers to aid them in gathering data, analyzing it, and fitting it to drive conclusions. In the book *The Fourth Paradigm: Data Intensive Scientific Discovery*, you can find many examples of the growth in tools and processes to provide more and better data to science.

Although this greater access to data is a huge benefit to the advancement of scientific knowledge, the computing power required to effectively analyze this data is limited by the available technical infrastructure. For most researchers, reliance on desktop PCs and small computing clusters constrains the advancement of their research by slowing the speed at which processing can occur, increasing the cost of research as a result of increased processing time, and restricting the ability to share their discoveries with the larger research community when data sets become too large to easily move. Furthermore, substantial up-front financial investments and a lack of skills needed to manage an advanced computing infrastructure are significant barriers to entry preventing these researchers from transitioning to larger systems. Finally, as more publicly available scientific data sets become available, the growing volumes of this data make it impractical to move data to the desktop for analysis, but instead require moving the questions to the data. Collectively, these issues require a new computing paradigm for science—cloud computing.

Advancements in cloud computing in recent years promise to remove these barriers. Microsoft has invested heavily in the development of data centers for a public cloud infrastructure, known as Microsoft Azure, which is ideally suited to serve the needs of the scientific community. Microsoft Azure provides a variety of cloud services enabling you to pick and choose the right combination to meet your needs, from setting up a community website to document and discuss research findings to performing complex data analysis in a scalable environment. Microsoft Azure has already proven successful for a variety of research projects and future enhancements promise to support research in new and exciting ways as cloud computing continues to evolve.

The Importance of the Cloud

According to the National Institute of Standards and Technology, cloud computing "…is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." Importantly, it dynamically scales to meet the current demand, whether the demand results from the execution of a resource-intensive application by a single user or the sudden influx of multiple users requesting access to a centralized shared resource. Similarly, the cloud resources can be released once demand begins to diminish.

Having a cloud infrastructure means researchers need not worry about having or hiring the skills to build, manage, and maintain a centralized and scalable environment. Instead, they can rely on continuous access to a data center managed by a third party. Rather than investing upfront to secure the necessary hardware for an infrastructure capable of supporting computational science, researchers can instead pay for access to cloud computing only as the need arises.

Access to vast arrays of managed resources is another compelling aspect of the cloud for researchers. Cloud computing platforms maintain the infrastructure and services on which applications run, such as operating systems and database services, among others. Because all hardware is abstracted by the cloud platform, there is no dependency on any specific piece of hardware. Consequently, vendors can apply patches and upgrade components without adversely impacting researchers. In addition, cloud services have built-in redundancy so routine failures are handled automatically, usually with minimal or no impact to users. That means you can rely on access to the data, applications, or services you choose to relegate to the cloud on a schedule that suits you.

The architecture of a cloud service resembles that of an on-premises solution, except that the responsibility for managing individual components of the architecture differs for a cloud service. There are three models used most commonly for cloud services:

- **Infrastructure as a Service (IaaS).** The cloud service provider maintains the physical or virtual machines, storage, and a networking layer, whereas you build and maintain one or more virtual machines that you load with an operating system, applications, and data. This model is stateful, which means that even when you shut down your virtual machines, their contents are saved to disk when you shut them down and are available again when you restart the machines.
- **Platform as a Service (PaaS).** With this model, the cloud service provider manages everything for you to support an application that you build. This model is considered best practice due to its statelessness. Application components do not persist a current state on the current node, but rely on external persistent storage so that no data is lost if hardware fails.
- **Software as a Service (SaaS).** The cloud service provider provides everything from the hardware to the applications running on the server in this model, leaving you simply to use the application.

Microsoft Azure Data Centers

Microsoft Azure is Microsoft's IaaS and PaaS solution, first announced in 2008 and enhanced with additional features since then. It relies on a global network of data centers managed by Microsoft to provide a collection of services that facilitate the development, deployment, and management of scalable cloud-based applications and services. Currently, Microsoft maintains data centers in four regions in North America, two regions in Europe, and two regions in Asia, as shown in Figure 1, and has plans to expand into additional sub-regions in the future. Each region contains one or more data centers. In turn, each data center, considered state of the art design for large computing and high data volumes, holds from ten to hundreds of thousands of servers. When you create a virtual machine in the cloud or develop a cloud-based application service, you select a data center region in which to store the virtual machine or execute the code. An alternative is to use the same service in multiple regions concurrently and direct users to the region closest to them. This option also provides disaster recovery in the event of a failure or network outage in a single data center.

Figure 1 Data Centers in Microsoft Geos Worldwide

Microsoft Azure and You

Although Microsoft Azure includes many services, the services that fall into the categories of Compute and Storage are generally of most interest to researchers. A wide variety of computing resources are available within Microsoft Azure with options such as websites, virtual machines, database as a service, and Hadoop as a service, to name a few.

Websites

One simple way to get started is to create a website to communicate your findings and collaborate with other researchers Later you can expand your usage by adding other services as applicable to your objectives. Although journals still have a place in the scientific community for peer review, many researchers are sharing their work publicly on the Internet by using blogs or content management systems to stimulate discussion or share data. If you already have an existing website or web application, you can easily move it into Microsoft Azure.

Otherwise, you can create a new one from scratch or by using the Web Application Gallery, shown in Figure 2, to choose an open source application, framework, or template, such as WordPress, Drupal, Django, and others. With just a few clicks, you have a site up and running, ready for you start adding content.

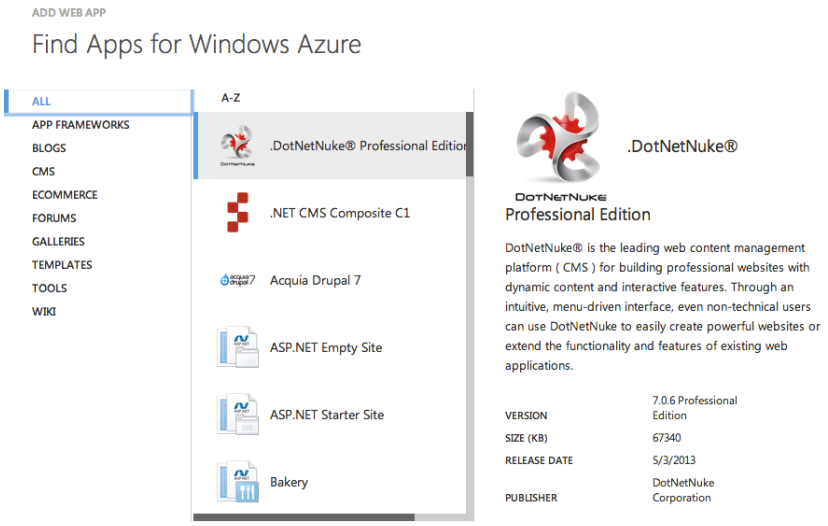

Figure 2 Web Application Gallery

Virtual Machines

If you already have your computer set up with applications that you use to perform computations on your data, to analyze, or to create visualizations, you can recreate that environment in the cloud to take advantage of its resources and gain access to more computing power. You can create an image of your computer and then use that image as a template to create a virtual machine managed by Microsoft Azure Virtual Machines. After uploading your image, it takes only a few clicks of the mouse and a few minutes to set up a new virtual machine.

You can even create multiple virtual machines using the same image to create a cluster or to run different simulations of your data in parallel on separate machines. You only pay for the time that your virtual machine runs. You can shut it down and pay a nominal cost for storage until you are ready to use it again later.

An alternative is to create a virtual machine from a gallery of standard images with either Windows or Linux operating systems. You can then add your own application to this virtual machine and attach a data disk to that machine. Yet another option is to locate an image in VM Depot to use as a starting point for your virtual machine, as shown in Figure 3. For example, you can use an image such as Azure Data Analysis that includes many popular tools for data analysis, such as IPython and R.

VM Depot images have been contributed by members of the open source community and are preloaded with an operating system, applications, and development environments. Because VM Depot is community-driven, you can also upload images that you create to add to the catalog. For example, a professor can upload an image for students to use for research assignments or a researcher can upload an image containing various applications that others engaged in similar research might find useful. This removes the barrier to entry for many scientists who wish to remove the complexity of installing and maintaining software and facilitate the reuse of common collections of applications.

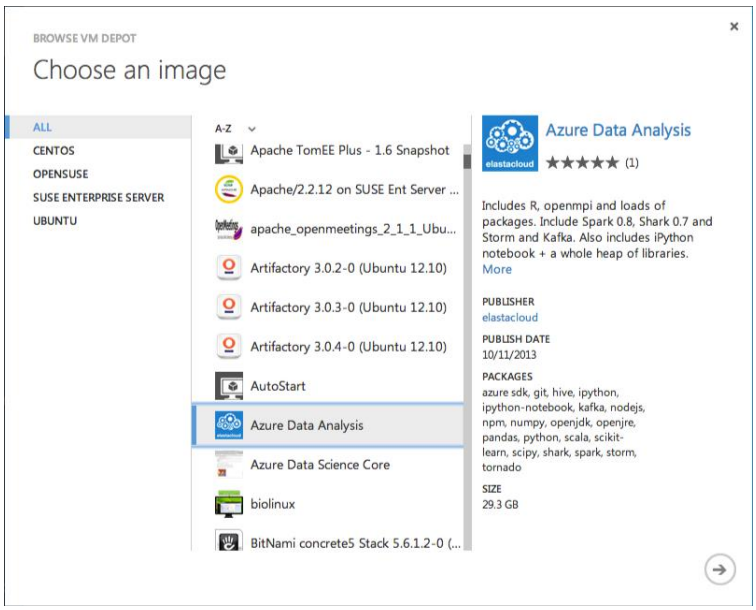

Figure 3 VM Depot

Cloud Services

To fully realize the potential of your applications running in the cloud, you can choose to host them by using Microsoft Azure Cloud Services rather than a virtual machine. Microsoft Azure Cloud Services has low administrative overhead because all maintenance associated with the operating system and the network is done for you. You can even configure the cloud service to automatically scale up or down according to rules you set.

You can see an example of a cloud service that has been running with no additional tuning required for over two years at weatherservice.cloudapp.net. The site consists of roles that handle job submissions and web interactions, dynamically scalable stateless compute instances that do the heavy lifting of weather forecast simulations, and Microsoft Azure Storage to store and manage datasets used by the simulation. This combination allows the cloud service to run reliably since its implementation. Another example of a cloud service is to generate simulations using complex parametric analysis models that can process thousands of variable combinations over a period of days rather than months. (See Buildings Go Green… In the Cloud.) You can find many other examples at the Cloud Research Projects section of the Microsoft Azure for Research website.

Mobile Services

Microsoft Azure Mobile Services provides you with an option for supporting mobile devices in your applications. This is a powerful way to extend the scope of your research applications. You can build Windows Phone, Windows Store Apps, Apple iOS, Android, or HTML/JavaScript applications to support your research efforts. For example, you can build an application that allows data to be uploaded directly into Microsoft Azure storage or you can push notifications to mobile devices when computing processes are complete, just to name a couple of examples.

Storage

Using the Storage service, you can store data in the cloud for use by your applications, to share with the larger research community, or to store backups of key data. You can view the scalability targets for

storage at http://blogs.msdn.com/b/windowsazure/archive/2012/11/02/windows-azure-s-flat-network-storage-and-2012-scalability-targets.aspx. Like other Microsoft Azure services, Storage is not only scalable and highly available, but easy to set up. That means you can focus on your research without the need to manage servers and tune databases.

Blobs

You can use Microsoft Azure Blobs to save or retrieve any type of file intended for sharing with others, such as documents, images, video, among other file types, or for use in a cloud application. Blob storage even accommodates the types of data commonly described as Big Data, such as raw data from scientific instruments or logs from servers, up to 200 GB in size per file, which you store as block blobs. Even backups from database servers and other devices as well as virtual hard drives for attachment to virtual machines can be placed in blob storage, up to 1 TB in size per file which you store as page blobs. To add files to blob storage, you must create an application as described at http://azure.microsoft.com/en-us/documentation/articles/storage-dotnet-how-to-use-blobs-20/.

One way you might use Microsoft Azure Blobs is to host reference data sets. You can store a variety of data sets, along with metadata to facilitate interpretation, in an accessible location as necessary to support peer review of research and enable new lines of research across disciplines. Having large data sets already in the cloud, close to the compute resources, reduces the time researches require to set up a new analysis.

Tables

Another option for storing data is to use Microsoft Azure Tables. This technology stores up to a 200 TB of typed data in a NoSQL key/value store that automatically scales to support lookups of properties when a full relational database approach is unnecessary. You might use tables when you want to store a lot of data captured by instruments or sensors with properties that you want to lookup, or to store threaded discussions on your website. Whereas blob storage holds unstructured data, table storage holds structured data.

Queues

If you have multiple applications that need to share information, you can use Microsoft Azure Queues to send messages between applications or tiers of an application.

SQL Database

Azure SQL Database is a relational database management system for cloud-based storage of data that is accessible to both cloud and on-premises applications. The advantage of this option is familiarity if you already rely on SQL Server databases for on-premises solutions, but it is easier to manage in the cloud while at the same time is highly available and scalable. All you have to do is build processes to move data into or out of the database, either manually or by using an application.

HDInsight and High-Performance Computing (HPC)

HDInsight, a Hadoop-based service from Microsoft, integrates with Microsoft Azure and other open source technologies that are part of the Hadoop ecosystem. By using HDInsight, you can store either unstructured data or data that is too big to store relationally in SQL Database in the Hadoop Distributed File System (HDFS). HDFS in turn creates redundant copies of the data and spreads the data across multiple nodes. As an alternative, HDInsight can store data in Microsoft Azure Storage Vault as a blob. Either way, you can then create a MapReduce job, which is also spread across multiple nodes to analyze the data in parallel and thereby perform computations much faster than would be possible on your local desktop.

Another option for distributed processing is to use HPC in Microsoft Azure. HPC allows you to implement multiple computing nodes that work in parallel. Another use case for HPC is to support workloads that have highly variable computational resource requirements. You can learn more about the features of HPC in "High Performance Computing on Microsoft Azure for Scientific and Technical Applications."

As one example of large data processing, you could monitor **c**ontinuous streams of high volumes of data from a variety of sensors or scientific instruments that you analyze in flight for anomalies or meaningful patterns. Another example is to store scientific image collections from the fields of medicine, astronomy, and earth sciences, to name a few, in Microsoft Azure Storage and then develop applications that implement image analysis algorithms that work more efficiently in a cloud environment against these large sets of images. Hadoop has proven very effective for next-generation sequencing, one of many types of bioinformatics data analysis that requires significant compute capacity to implement.

Tools for Cloud Computing

There are no limitations on the tools that you can use for cloud computing. Whether you build your own tools to perform tasks like data collection, acquire open source applications written in Python, or purchase tools like MATLAB, you can work with data in the cloud to make progress towards your research goals.

Cross Platform Support

Microsoft's open and flexible cloud platform is accessible to research and academic users who are familiar with Windows, Linux, or the Mac OS. You can learn more about this platform from " Microsoft Azure for Linux and Mac Users" which explains how you can take advantage of Microsoft Azure features, such as virtual machines and persistent storage, by using the operating system of your choice.

General Programming Support

Microsoft Azure is an open platform that supports most languages. Several language-specific SDKs exist for .NET, Java, PHP, Node.js, Ruby, and Python so you can use your preferred language to build your cloud application and access Microsoft Azure Services programmatically. Many of the Microsoft Azure Services include a REST API.

Support for Python

Using IPython, a web-based interactive development environment (IDE), a developer can create applications to perform advanced statistical analysis, interactive data visualizations, and mathematical modeling of big data. Another option is to use Python Tools for Visual Studio (PTVS). Both tools are supported for use with Microsoft Azure. If you prefer, you can instead use Python 2.x or 3.x Microsoft Azure Cloud Services or in a Microsoft Azure virtual machine (both Windows and Linux). See "An Introduction to Using Python with Microsoft Azure" to learn more about these tools.

Getting Started with Cloud Computing for Research

There are many research scenarios for which cloud computing is well suited. Here are just a few possibilities for getting started with cloud computing for your next research project:

- **Learn about Microsoft Azure.** You can find a variety of information about Microsoft Azure services, development support, and resources for best practices, code samples, and more at Microsoft Azure Documentation.
- **Create a website.** You have a several options for migrating an existing website to Microsoft Azure, which you can learn more about at How To: Migrate and Publish a Web Application to

Microsoft Azure from Visual Studio, Migrate a Database-backed Website (and database) to Microsoft Azure Web Sites, and Migrating a Blog to Microsoft Azure Web Sites. For a new site, try using one of the sites available in the Web Application Gallery. For more information about using the Web Application Gallery or setting up a new Microsoft Azure website, see How to Create and Deploy a Web Site and Create and Manage Web Sites.

- **Set up a virtual machine.** For an overview of your options for working with virtual machines, see Virtual Machines Documentation. If you want to create your own virtual machine, see "Getting Started with Virtual Machines in Microsoft Azure". Another option is to use a template of a VM from VM Depot or to contribute a VM to the online community, both of which are described in "Using and Contributing Virtual Machines to VM Depot." Both of these papers are available at the Microsoft Azure for Research listing for technical papers.
- **Create a cloud service.** You can build an application that runs in the cloud as a service. To learn about developing and deploying cloud services, see Cloud Services Documentation and Microsoft Azure Cloud Services on Channel 9. You can also read about a sample scalable cloud-based search service at Using Microsoft Azure for BLAST.
- **Process large data volumes.** Leverage the power of the cloud and HDInsight to analyze large data sets efficiently. An example to get you started thinking about HDInsight is available at Hadoop and HDInsight: Big Data in Microsoft Azure.

Conclusion

Scientific research today is generating data that once would have been impossible to analyze due to sheer volume and lack of sufficient computing power in all but the largest laboratories. Prior to cloud computing, analysis of scientific data could take weeks, months, or even years to process. Cloud computing now creates opportunities for all researchers to find patterns or anomalies in their data even with a limited budget. Furthermore, the lower cost of storage means the research community at large can curate and share data, promoting deeper collaboration within the community and enabling data mash-ups with the potential to reveal new insights. Microsoft Azure provides the necessary cloud platform to reduce not only the time to discovery, but also the cost of discovery. Now is the time to try Microsoft Azure for yourself and discover firsthand how easy it is to set up and go live.

Additional Resources

Microsoft Azure Documentation, http://azure.microsoft.com/en-us/documentation/.
Microsoft Azure .NET Developer Center, http://azure.microsoft.com/en-us/develop/net/.

References

Hey, Tony, Stewart Tansley, and Kristin Tolle (eds.) (2nd ed.). (2009). The Fourth Paradigm: Data Intensive Scientific Discovery. Microsoft Research.

Johnston, Steven J., Neil S. O'Brien, Hugh G. Lewis, Elizabeth E. Hart, Adam White, and Simon J. Cox. (2013). Clouds in Space: Scientific Computing using Azure. Journal of Cloud Computing. http://www.journalofcloudcomputing.com/content/2/1/2.

Mell, Peter and Timothy Grance. (2011). The NIST Definition of Cloud Computing, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

© 2013 Microsoft Corporation. All rights reserved. Except where otherwise noted, content on this site is licensed under a Creative Commons Attribution-NonCommercial 3.0 License.