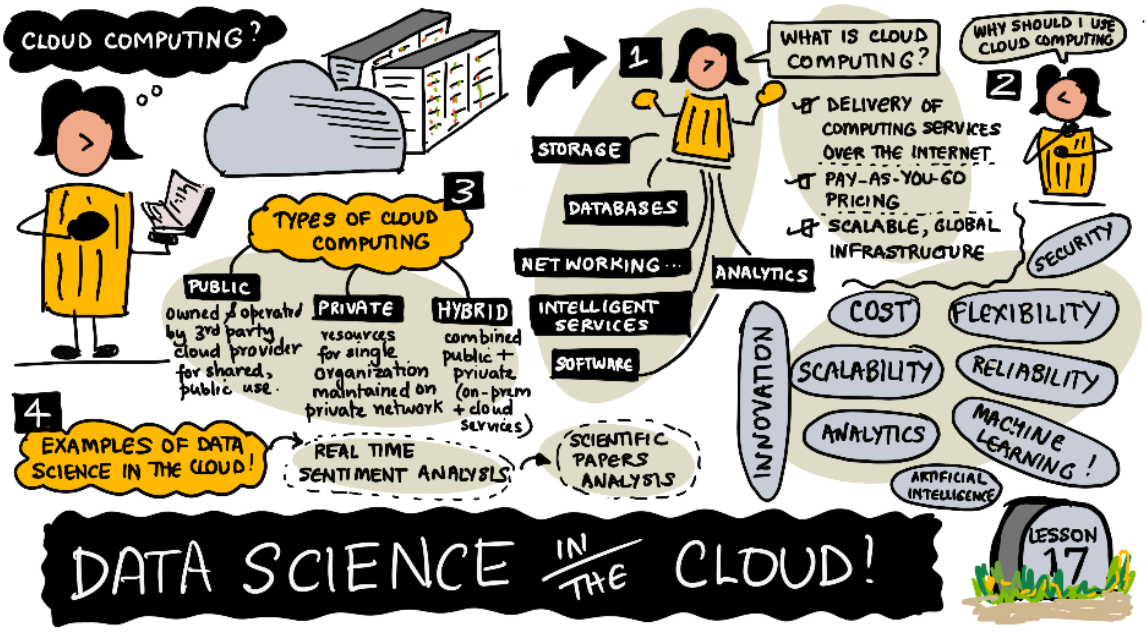
Introduction to Data Science in the Cloud



Data Science In The Cloud: Introduction - *Sketchnote by @nitya*

In this lesson, you will learn the fundamental principles of the Cloud, then you will see why it can be interesting for you to use Cloud services to run your data science projects and we'll look at some examples of data science projects run in the Cloud.

Pre-Lecture Quiz

What is the Cloud?

The Cloud, or Cloud Computing, is the delivery of a wide range of pay-as-you-go computing services hosted on an infrastructure over the internet. Services include solutions such as storage, databases, networking, software, analytics, and intelligent services.

We usually differentiate the Public, Private and Hybrid clouds as follows:

- Public cloud: a public cloud is owned and operated by a third-party cloud service provider which delivers its computing resources over the Internet to the public.
- Private cloud: refers to cloud computing resources used exclusively by a single business or organization, with services and an infrastructure maintained on a private network.
- Hybrid cloud: the hybrid cloud is a system that combines public and private clouds. Users opt for an on-premises datacenter, while allowing data and applications to be run on one or more public clouds.

Most cloud computing services fall into three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

- Infrastructure as a Service (IaaS): users rent an IT infrastructure such as servers and virtual machines (VMs), storage, networks, operating systems

- Platform as a Service (PaaS): users rent an environment for developing, testing, delivering, and managing software applications. Users don’t need to worry about setting up or managing the underlying infrastructure of servers, storage, network, and databases needed for development.
- Software as a Service (SaaS): users get access to software applications over the Internet, on demand and typically on a subscription basis. Users don’t need to worry about hosting and managing the software application, the underlying infrastructure or the maintenance, like software upgrades and security patching.

Some of the largest Cloud providers are Amazon Web Services, Google Cloud Platform and Microsoft Azure.

Why Choose the Cloud for Data Science?

Developers and IT professionals chose to work with the Cloud for many reasons, including the following:

- Innovation: you can power your applications by integrating innovative services created by Cloud providers directly into your apps.
- Flexibility: you only pay for the services that you need and can choose from a wide range of services. You typically pay as you go and adapt your services according to your evolving needs.
- Budget: you don’t need to make initial investments to purchase hardware and software, set up and run on-site datacenters and you can just pay for what you use.
- Scalability: your resources can scale according to the needs of your project, which means that your apps can use more or less computing power, storage and bandwidth, by adapting to external factors at any given time.
- Productivity: you can focus on your business rather than spending time on tasks that can be managed by someone else, such as managing datacenters.
- Reliability: Cloud Computing offers several ways to continuously back up your data and you can set up disaster recovery plans to keep your business and services going, even in times of crisis.
- Security: you can benefit from policies, technologies and controls that strengthen the security of your project.

These are some of the most common reasons why people choose to use Cloud services. Now that we have a better understanding of what the Cloud is and what its main benefits are, let's look more specifically into the jobs of Data scientists and developers working with data, and how the Cloud can help them with several challenges they might face:

- Storing large amounts of data: instead of buying, managing and protecting big servers, you can store your data directly in the cloud, with solutions such as Azure Cosmos DB, Azure SQL Database and Azure Data Lake Storage.
- Performing Data Integration: data integration is an essential part of Data Science, that lets you make a transition from data collection to taking actions. With data integration services offered in the cloud, you can collect, transform and integrate data from various sources into a single data warehouse, with Data Factory.
- Processing data: processing vast amounts of data requires a lot of computing power, and not everyone has access to machines powerful enough for that, which is why many people choose to directly harness the cloud’s huge computing power to run and deploy their solutions.
- Using data analytics services: cloud services like Azure Synapse Analytics, Azure Stream Analytics and Azure Databricks to help you turn your data into actionable insights.

- Using Machine Learning and data intelligence services: Instead of starting from scratch, you can use machine learning algorithms offered by the cloud provider, with services such as AzureML. You can also use cognitive services such as speech-to-text, text to speech, computer vision and more.

Examples of Data Science in the Cloud

Let's make this more tangible by looking at a couple of scenarios.

Real-time social media sentiment analysis

We'll start with a scenario commonly studied by people who start with machine learning: social media sentiment analysis in real time.

Let's say you run a news media website and you want to leverage live data to understand what content your readers could be interested in. To know more about that, you can build a program that performs real-time sentiment analysis of data from Twitter publications, on topics that are relevant to your readers.

The key indicators you will look at is the volume of tweets on specific topics (hashtags) and sentiment, which is established using analytics tools that perform sentiment analysis around the specified topics.

The steps necessary to create this project are as follows:

- Create an event hub for streaming input, which will collect data from Twitter
- Configure and start a Twitter client application, which will call the Twitter Streaming APIs
- Create a Stream Analytics job
- Specify the job input and query
- Create an output sink and specify the job output
- Start the job

To view the full process, check out the documentation.

Scientific papers analysis

Let's take another example of a project created by Dmitry Soshnikov, one of the authors of this curriculum.

Dmitry created a tool that analyses COVID papers. By reviewing this project, you will see how you can create a tool that extracts knowledge from scientific papers, gains insights and helps researchers navigate through large collections of papers in an efficient way.

Let's see the different steps used for this:

- Extracting and pre-processing information with Text Analytics for Health
- Using Azure ML to parallelize the processing
- Storing and querying information with Cosmos DB
- Create an interactive dashboard for data exploration and visualization using Power BI

To see the full process, visit Dmitry's blog.

As you can see, we can leverage Cloud services in many ways to perform Data Science.

Footnote

Sources:

- https://azure.microsoft.com/overview/what-is-cloud-computing?ocid=AID3041109
- https://docs.microsoft.com/azure/stream-analytics/stream-analytics-twitter-sentiment-analysis-trends?ocid=AID3041109
- https://soshnikov.com/science/analyzing-medical-papers-with-azure-and-text-analytics-for-health/

Post-Lecture Quiz

Post-lecture quiz

Assignment

Market Research