When would you consider a cloud infrastructure for your data science tasks?

* programming languages, tools and frameworks,
* Rent rather than own infrastructure
* Pay as you go pricing
* Rapidly growing Analysis and storage requirements
* Use of qualified and certified managed services
* Look at structured, semi-structured and unstructured data
* Look at varied sets of data, irrespective of the size, format, etc.
* Analyse them to draw insights

However, the problem with such data is, it often sits in disparate silos. Given that the storage is now much cheaper, and the open source platforms and tools are available for data scientists, cloud is the key.

* Increased Data Generation – Look around, how many data emitting instruments can you see around? The laptop you might be reading this article on, the smart phone (and the apps in it) you might be carrying around, the fitness band you might be wearing in your hand, the car you drive (in some cases) – are all emitting data continuously. Now imagine things in ~~five~~ two years from now – the refrigerator you use, your home temperature maintenance unit, the clothes you wear, the pen you carry, the water bottle you drink water with all will be carrying sensors and emitting data for the data scientists (and the databots) to analyze.
* Low cost of data storage – Let’s do ground reality check. Take a guess – How much would be the cost of storing the entire universe of music created ever? What did you guess? My estimate is that this number would be significantly less than a grand!

Cheap computational power – Check out this power packed configuration of a laptop released recently. A Xeon processor packed with Quadro GPU and containing 64 GB RAM. This machine is expected to come out for less than $2000 and weighs ~2.5 kg. Do I need to say more?

* Need to run scalable data science: Let’s  dial back a few years. In 2010, I joined a multi-national Insurance company to set up a data science unit. Among several things, we bought a server which had 16 GB RAM. Since this was a new setup, we had bought more than what we thought we would need in next 3 – 5 years. What seemed a stellar combination at the start of the journey did not scale when we hired more people in team. Not only the number of people in the team increased, the amount of data increased exponentially. With a physical server, we were stuck! Either we buy a new, more powerful server or load this one to maximum (which would have again ran out of juice). The last thing you want is that your data scientists are staring at the screen waiting for data to be processed! A machine on cloud can scale on a click of the button, with out much hassle. So your current scripts and models can easily run when the data behind the model grows multiple times.
* Cost: While scalability is one benefit, cost is another. Let us say you need to work on a problem, which is one-off in nature, but needs a higher computational infrastructure. This could be any thing – you want to mine social media data for an annual event you have sponsored, but you want the insights in real time. You can’t really buy a new machine for that. The costs will be prohibitive. Simple, rent out a higher configuration for a few hours or days and you have solution to your problem at a fraction of the cost.
* Collaboration: What if you want to work simultaneously along with several data scientists? You don’t want every one to create a copy of the data and code in their local machines.
* Sharing: What if you want to share your piece of Python / R code with your team? The libraries you might have used may not be there or might be of the older version. How do you make sure that the code is transferable to a different machine?
* Larger ecosystem for machine learning system deployments: A few cloud services like AWS, Azure provide complete ecosystem to collect data, run your models and then deploy them. In case of physical machine, you will need to set this up yourself.

Use for building quick prototypes: A number of times, you get ideas while you are on the move or when you are discussing some thing with your friends. In these scenarios, it is much easier to use the out of the box services on the cloud. You can quickly build prototypes with out worrying about versions and scalability. Once you have proven the concept, you can always build a production stack later.