Data science in cloud capabilities

I watched a video on YouTube titled “Why Cloud Computing is Critical for a Data Scientist” to gain a basic understanding of this topic.

- Data Abundance, Local Servers, and their Limited Storage.

With the vast amounts of data, we deal with today, trying to process everything on local servers can be incredibly slow and inefficient. Cloud computing provides the facility to speed up and the scalability needed to manage large datasets without being bogged down by limited hardware. It allows data scientists to access powerful tools and resources from anywhere, making it easier to collaborate with others and work on projects without worrying about server limitations. Additionally, cloud computing offers the flexibility to scale up or down based on the project’s needs, saving time and costs in the long run. Moreover, whenever we need to make changes or load large datasets, our local machines often lack the capacity to handle all that storage due to their limited space.

Consider a scenario where you’re a data scientist tasked with developing a machine learning model that predicts customer behavior based on a massive dataset containing millions of records. The dataset includes a variety of features such as transaction history, browsing patterns, demographic information, and much more. For training a machine learning model with the vast amount of dataset, using your local infrastructure presents several challenges: limited storage capacity may prevent you from even storing the dataset, and performing operations like data preprocessing, feature engineering, or model training could be extremely slow due to hardware constraints. Additionally, training complex models, especially deep learning ones also faced these difficulties even more than training a ML model as we know neural network generally also takes a lot of time to get trained. In all these scenarios we can see that with cloud-based GPUs, you can drastically reduce training times, and cloud platforms make it easier to collaborate with team members across different locations. This approach not only speeds up your data science tasks but also provides the scalability and flexibility needed for large-scale projects.