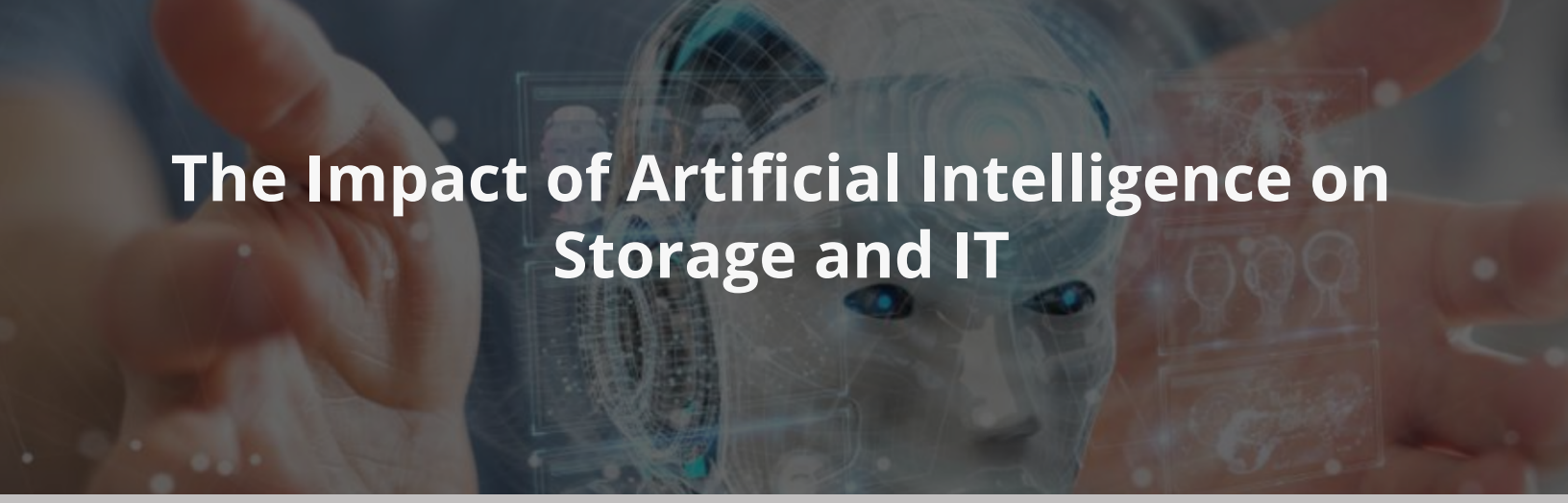

The Impact of Artificial Intelligence on Storage and IT

Now more than ever, organisations are looking to artificial intelligence (AI) and in particular machine learning (ML) to solve complex data challenges and bring new insights and value to an ever-increasing volume of information stored within our business.

By Glyn Bowden, SNIA Cloud Storage Technologies Initiative Member; Chief Architect, AI & Data Science Practice at Hewlett Packard Enterprise.

The emergence of the data scientist as a mainstream profession within any sized organisation rather than focused within high finance, research institutes or governments demonstrates how quickly the adoption has been. However, with anything moving at this sort of pace, it has been difficult to take time and assess just what impact this new wave of analytics is having on our infrastructures and more specifically, the storage estate where a majority of this data is currently residing.

Are AI and ML the next step along the evolutionary path following data marts and big data? So what is the difference here, just the scale? The answer is no. There is now a very different storage challenge today that we need to deal with. It has to do with the way data is used. Traditionally storage has a single use or at least a single performance profile at a specific stage in its lifecycle. We know, for example, that recently created data is typically considered “hot” as it is accessed most frequently. Then the data cools over time as it becomes less relevant until it is either archived to slow media or expired all together. This means there was a focus on data lifecycle management and hierarchical storage architectures. In this traditional architecture, the data moves between tiers so that it is on high performance media when fresh and active, and slower bulk media when cold. However, with the new techniques of AI and ML, data can have many uses at any time. That means we will never be able to effectively plan where that data needs to sit from a tiering perspective.

Also, if you look at how certain data is stored, it could be in any number of formats. For example, it could be unstructured files, Blobs in an object store or data in a SQL database on a LUN somewhere. If we suddenly decide that one of those data sources is now critical to

building the desired model, the demand dynamic on that data will change. For training models in ML this places a heavy read demand on the data used. For example, in the case of supervised learning the data is parsed multiple times across the validation and testing phases. The pattern of read I/O is also somewhat difficult to estimate as data may be ordered in ways it wasn't traditionally used meaning indexes on databases. Even random block access can increase latency and impact performance. Don't forget all this can be happening whilst the media is still being used for the day job of the data connected to other business systems, therefore impacting performance of other business critical systems.

Not only do we need to consider the archives and pools of data within the organisation, we need to look at data that is being captured and what our new processes mean. Before ML models are applied against an incoming data source, very often that data needs to be transformed in some way, so that the fields and schema match the trained models' expectations and format. It will also likely be filtered in some way, particularly if the incoming data feed is very verbose and contains features or records that might not be relevant to the model. If these features and records are included then they can overwhelm the infrastructure that provides the inference service. They could also cause additional latency or increase resource requirements unnecessarily. This is often referred to as pre-processing or pre-engineering. What comes out the other end will be cleaned and transformed data sets that fit for inference. Again, this has the potential to be very different from the original incoming data, so its original use will still need to be serviced. This requirement could mean the need to fork the data pipeline in some way, so the original data carries on its previous path and the new fork passes through the cleaning and transformation process on its way to inference. Then question would then be, is there value in storing both?

As you can see the profiles of the data change drastically from the original scenario and we have to review both the performance requirements for data in-flight and at-rest, as well as the capacity of data stores to cope with potentially different formats or schemas of that data.

Of course, at the scales we are seeing emerge, this is not practical. We need to start thinking about storage systems and data architectures in a new and unified way. We need to accept the fact that data will have multiple purposes, often unknown at the time of collection and due to the inherent potential value, we will be keeping much more of it around.

The advent of machine learning impacts more than just the active data pools and pipelines too. There is now more need for careful configuration control on our data transformation services and model management systems. We need to ensure that everything stays in sync and if a change is made to a model that requires upstream changes to the data, then the transformation needs to reflect that in the live data pipeline as well. The results, otherwise, would be an inference model that no longer has the features it's expecting and would generate poor results, often not identified or noticeable for considerable periods of time.

One final thought is that with all the changes I've mentioned above, I've also alluded to the dependency that datasets already have on them from existing business systems. This dependency will not change and will drive whether a migration or transformation is appropriate or not. Therefore, we also need to ensure we have mechanisms that allow us to discover and connect to the discreet existing data sources around the organisation. This will allow us to augment the data into our new pipelines and data ecosystem without the need to disrupt their current role. AI and ML present great opportunities to organisations if harnessed correctly but can also provide a significant challenge if the impact is not well understood and catered for.

About CSTI

The SNIA Cloud Storage Technologies Initiative (CSTI) is committed to the adoption, growth and standardization of storage in cloud infrastructures, including its data services, orchestration and management, and the promotion of portability of data in multi-cloud environments. To learn more about the CSTI's activities and how you can join, visit snia.org/cloud.