Data Version Control

ThoughtWorks is a global technology consulting company that provides tools, consulting services, agile methodology, and software designs to enable technology disruptors and enterprises to thrive as modern digital businesses across the globe. Since ThoughtWorks has been at the forefront of technology innovation over the past 3 decades, it helps enterprises to keep pace with the escalating rate of technological change.

ThoughtWorks have its own website to help impact the world in an extraordinary way by sharing its expertise and offering a wide range of resources, including case studies, articles, and blogs that provide insights into the latest practices in digital transformation, technology strategy, and software development. One of the most popular projects that ThoughtWorks works on is the Technology Radar. The Technology Radar is a semiannual publication of the latest trending tools, languages, technologies, and platforms. This helps companies decide which technologies and tools to use and avoid. The Radar is divided into 4 quarters: Techniques, Tools, Platforms, and Language & Frameworks. This radar is based on the ThoughtWorks teams' opinions on whether these features work well and efficiently on their projects.

Data Version Control (DVC) is one of the tools that the radar recommends to "Adapt". The team has decided it's the best open-source tool for managing experiments in data science projects. Since DVC is git-based meaning that it uses it as its underlying version control system which indicates that all data, models, and code are stored in Git repositories which is something that most of the developers in companies are at a level of familiarity with already so that is a great advantage of DVC. It, also, keeps a special version of the computer program that has all the important information like how the program was made, what data it used to learn, and how well it can do its job. Since reproducibility is our biggest concern, DVC enables the team to time travel through the different versions of the model. Moreover, when the information keeps changing a lot, using DVC is a great way to keep track of different versions to help in seeing how things have changed over time. Not only that, but also, DVC can be plugged in with any type of storage such as Google Cloud Storage, Google Drive, and MinIO. However, the bigger the data sets get, the more expensive the system-based snapshotting gets.

To sum it up, DVC is an extremely powerful tool for managing Machine Learning projects that allows data scientists to monitor changes in their data, models, and code, and collaborate effectively and efficiently with their team members ; for all of those previous reasons, ThoughtWorks has moved DVC to "Adapt" in the latest updated Technology Radar.