DVC (Data Version Control)

DVC (Data Version Control): It is an open-source tool for data and model versioning and management. It is designed to help data scientists and machine learning engineers work more effectively by enabling them to track changes to data and models over time, reproduce experiments and models, and collaborate with others on data-driven projects.

DVC provides a way to version control data, models, and experiments using Git-like commands. It allows you to store large data files separately from code, so that you can easily share and reproduce data across different machines and collaborators. DVC also integrates with popular ML frameworks, such as TensorFlow and PyTorch, to provide a unified workflow for model development and management.

One of the key features of DVC is its ability to handle data and model versioning at scale. It supports large file storage systems like S3, GCS, and Azure, which means that you can store and version datasets that are too big to fit on a local machine. DVC also includes powerful tools for managing data dependencies, so that you can easily reproduce experiments and models with the same input data.

Key features:

- **Version control for large data files:** DVC provides Git-like version control for large data files, allowing you to track changes to datasets over time.
- **Reproducibility:** DVC provides tools for managing data dependencies, making it easy to reproduce experiments and models with the same input data.
- **Integration with ML frameworks:** DVC integrates with popular ML frameworks like TensorFlow and PyTorch, providing a unified workflow for model development and management.
- **Large file storage:** DVC supports large file storage systems like S3, GCS, and Azure, allowing you to store and version datasets that are too big to fit on a local machine.
- **Collaborative work:** DVC is designed to enable teams to collaborate on data-driven projects, with features like shared remote data storage and access control.
- **Easy to learn and use:** DVC is easy to learn and use, with a simple command-line interface and a user-friendly web interface.
- **Open-source and free:** DVC is an open-source tool, available for free under the Apache 2.0 license.

Commands:

1. **dvc init**: Initializes a new DVC repository in the current directory.
2. **dvc add <file>**: Adds a file to DVC's version control system. This command creates a new file with the .dvc extension, which contains metadata about the original data file.
3. **dvc remote add -d <remote_name> gdrive://<folder_id>**: It helps to add remote storage where you store the data & model.

4. **dvc run**: Creates a new DVC stage, which is a set of dependencies, commands, and outputs. This command allows you to define a data processing or model training pipeline, including dependencies on input data, code, and parameters.

5. **dvc repro**: Reproduces a DVC stage, ensuring that all dependencies are up to date before running the commands again. This command can be used to re-run a data processing or model training pipeline with updated data or code.

6. **dvc push**: Pushes data and models to a remote storage location, such as S3 or GCS. This command uploads data and models to a remote storage location, allowing them to be shared with collaborators or used for deployment.

7. **dvc pull**: Pulls data and models from a remote storage location. This command downloads data and models from a remote storage location, allowing you to work with the latest versions of files.

8. **dvc checkout**: Checks out a specific version of a file or a directory from the DVC cache. This command allows you to revert to a previous version of a file or a directory.

9. **dvc status**: Shows the status of the current DVC repository, including the changes to the tracked files and the current workspace.

10. **dvc metrics**: Adds a metric to a DVC stage. This command allows you to track performance metrics for your model, such as accuracy or loss.

11. **dvc import-url**: Imports data from a URL. This command allows you to download and version data directly from a URL, such as a public dataset.