What is Alluxio

Alluxio is world's first open source data orchestration technology for analytics and AI for the cloud. It bridges the gap between data driven applications and storage systems, bringing data from the storage tier closer to the data driven applications and makes it easily accessible enabling applications to connect to numerous storage systems through a common interface. Alluxio's memory-first tiered architecture enables data access at speeds orders of magnitude faster than existing solutions.

In the data ecosystem, Alluxio lies between data driven applications, such as Apache Spark, Presto, Tensorflow, Apache HBase, Apache Hive, or Apache Flink, and various persistent storage systems, such as Amazon S3, Google Cloud Storage, OpenStack Swift, HDFS, IBM Cleversafe, EMC ECS, Ceph, NFS, Minio, and Alibaba OSS. Alluxio unifies the data stored in these different storage systems, presenting unified client APIs and a global namespace to its upper layer data driven applications.

The Alluxio project originated from the UC Berkeley AMPLab (see papers) as the data access layer of the Berkeley Data Analytics Stack (BDAS). It is open source under Apache License 2.0. Alluxio is one of the fastest growing open source projects that has attracted more than 1000 contributors from over 300 institutions including Alibaba, Alluxio, Baidu, CMU, Google, IBM, Intel, NJU, Red Hat, Tencent, UC Berkeley, and Yahoo.

Today, Alluxio is deployed in production by hundreds of organizations with the largest deployment exceeding 1,500 nodes.