In 2003 Google published a paper entitle the Google File System, “a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.” In 2004, the published another paper on MapReduce, a computational engine that worked on top of the Google File System. The ideas in these two papers were used to develop the initial version of Apache Hadoop in 2006, which consisted of HDFS(Hadoop distributed file system) and a set of other libraries the implemented MapReduce and managed the resources.

## Timeline

2003 – GFS(Google File System )paper

2004 – MapReduce paper

2006 – Open Source solution inspired by GFS and mapreduce papers – Hadoop

2009 – Facebook creates Hive and Hive tables to address lack of structure in Hadoop, adding SQL

2017 – Apache Iceberg format created at Netflix

Why do we need Iceberg?

Netflix, in the creation of what became the Apache Iceberg format, concluded that many of the problems with the Hive format stemmed from one simple but funda‐ mental flaw: each table is tracked as directories and subdirectories, limiting the granularity that is necessary to provide consistency guarantees, better concurrency, and several of the features that often are available in data warehouses.