The Apache Software Foundation

The Apache Software Foundation

- Charity funded by individual donations and corporate sponsors

- 350+ leading Open Source projects, including Apache HTTP Server -- the world's most popular Web server software

From the documentation

What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

(...)

Hadoop Modules

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Hadoop related projects

(…)

- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

(...)

Hadoop related projects

- Hue
- Ganglia
- Mahout
- Oozie

How Do They Work Together?

Hadoop provides the infrastructure that related projects use e.g.

- Cluster Management
- Scalability
- Availability

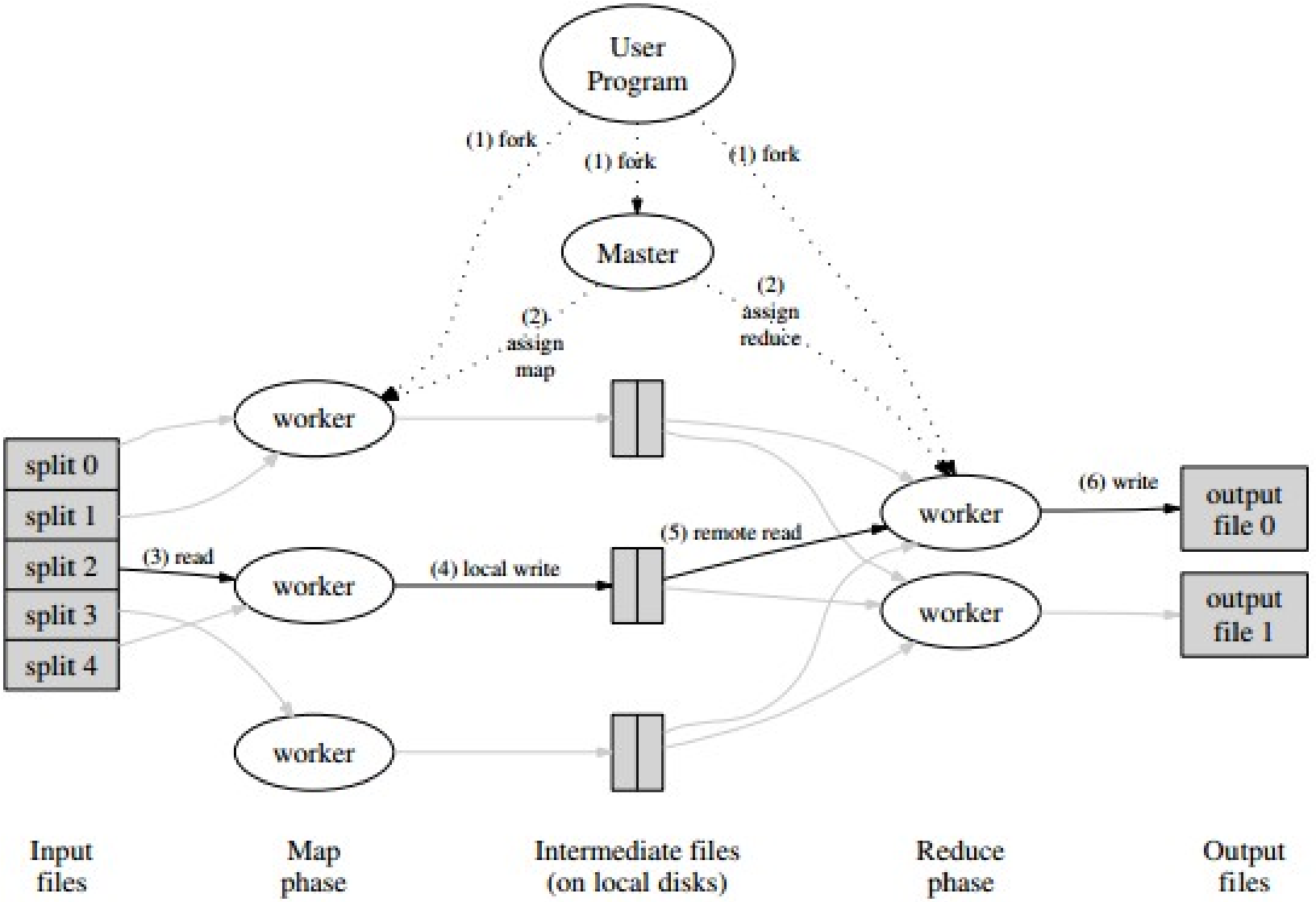
Hadoop – where did it all begin?

Google paper:

MapReduce: Simplified Data Processing on Large Clusters

Dean and Ghemawat 2004

Hadoop – where did it all begin?



Hadoop

Open Source Implementation of Google MapReduce by Doug Cutting

HDFS – Open Source Implementation of Google DFS