cloud File System

Cloud File System

Cloud is an advanced concept of distributed computing.

- DFS is basically used for storing huge amount of data and provides accessibility of stored data to all distributed clients across the network.
- DFS comprises various software components that run as a single system entity on multiple systems.
- There are a number of DFS that solve this problem in different ways. Some popular file systems are:

- AFS (Andrew file system)
- NFS (Network file system)
- Coda
- AFP (Apple file protocol)
- GFS (Google file system)
- HDFS (Hadoop distributed file system)

Cloud File System

- NFS is the most commonly adopted DFS.
- It grants remote access to a logical volume that resides on a single machine and makes some segments on its local file system which provides accessibility to different distributed clients.
- NFS is one of the oldest file system and some limitations are there.
- All data resides on one machine, so reliability issues might come when its come under single point of failure.
- To handle these challenges GFS and HDFS follow different approach.

Google File System

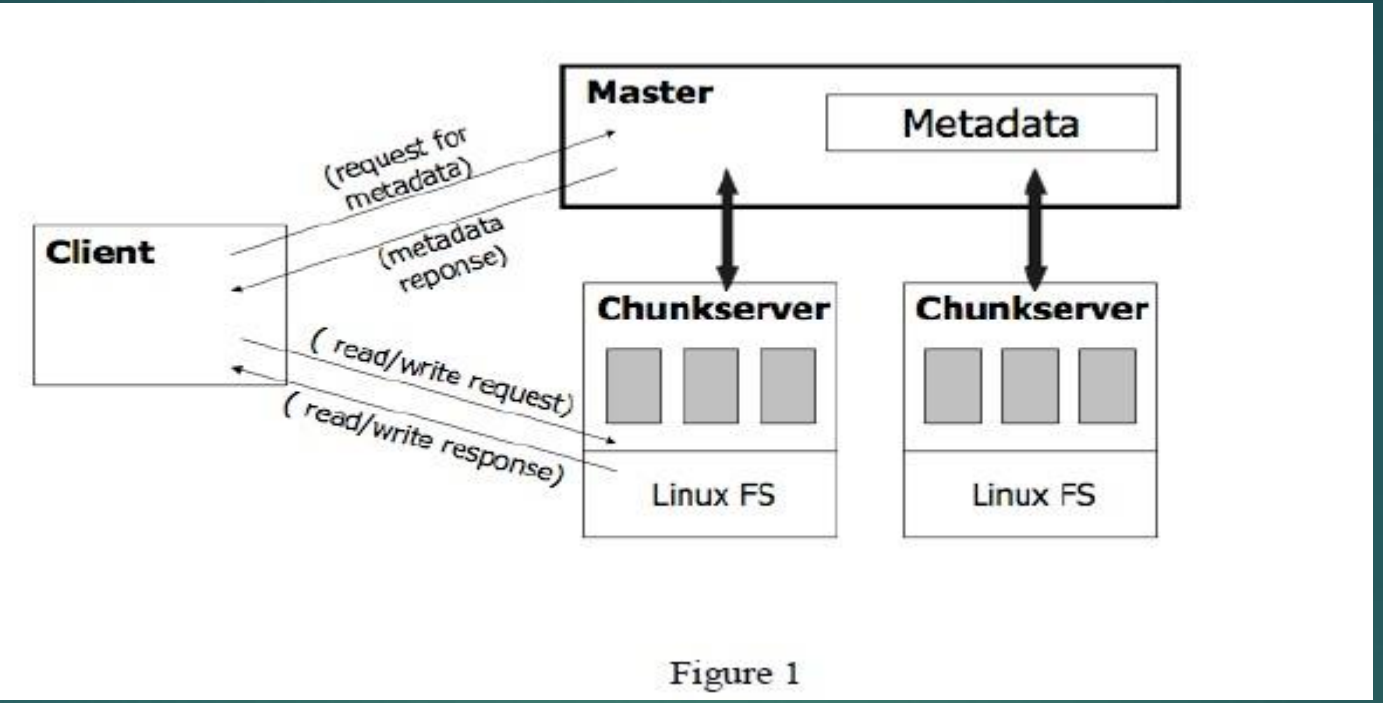


Figure 1

- Google invented and implemented a scalable DFS to handle their huge internal distributed data exhaustive applications and named Google file system.
- In 2002-2003, Google launched its file system based on DFS architecture but added some advance features.

Google File System

- A cluster of a GFS contains a single master and multiple chunk servers that are associated with clients.
- The master holds the metadata of chunk servers.
- All the data processing happens through these chunk servers.
- The client first contacts the master and retrieves the metadata of the chunk server, which is then stored in the chunk server, so the next time , client directly connects to the chunk server.

Chunk: It is similar to concept of block in a file system, but chunk size is larger than the traditional file system block.. The block of chunk is 64 MB. This is specifically designed for Google environment.

Google File System

Master: Master is a single process that runs on entirely separate machine for security purpose. It only stores metadata-related information, chunk location, file mapping information and access control information. The client first contacts the master for information about metadata and then connects to the particular chunk server.

Metadata: Metadata is stored in the memory of a master, therefore, master operations are much faster. Metadata contains three types of information.

- Namespaces of file and chunk
- Location of each chunk
- Mapping from file to chunk.

Hadoop Distributed File System

History Of Hadoop

- Hadoop was started by Doug Cutting to support two of his other well known projects, Lucene and Nutch
- Hadoop has been inspired by Google's File System (GFS) which was detailed in a paper by released by Google in 2003
- Hadoop, originally called Nutch Distributed File System (NDFS) split from Nutch in 2006 to become a sub-project of Lucene. At this point it was renamed to Hadoop.

Why Hadoop ?

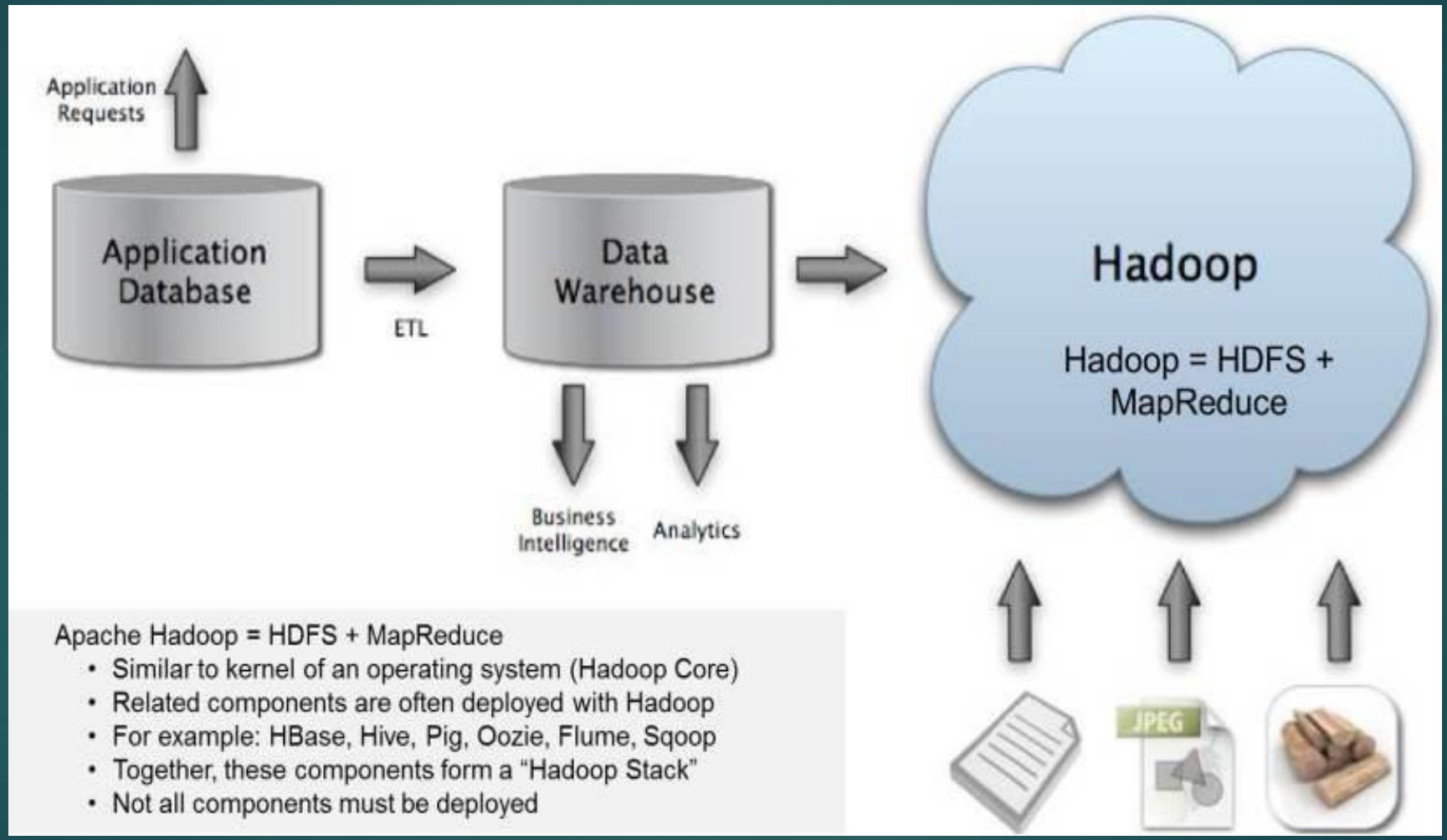
Hadoop infrastructure provides these capabilities

- Scalability
  - -Thousands of Compute Nodes
  - -Petabytes of data
- Cost effective
  - -Runs On Low Cost Commodity Hardware
- Efficient
  - -By distributing the data, Hadoop can process it in parallel on the nodes where the data is located.

What Is Hadoop ?

- Open source software platform for scalable, distributed computing
- Hadoop provides fast and reliable analysis of both structured data and unstructured data
- Apache Hadoop software library is essentially a framework that allows for the distributed processing of large datasets across clusters of computers using a simple programming model.
- Hadoop can scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop Architecture



Apache Hadoop = HDFS + MapReduce

- Similar to kernel of an operating system (Hadoop Core)
- Related components are often deployed with Hadoop
- For example: HBase, Hive, Pig, Oozie, Flume, Sqoop
- Together, these components form a "Hadoop Stack"
- Not all components must be deployed

HDFS Basic Model