2.3.5 Hadoop Ecosystem

HDFS

- Hadoop Distributed File System (HDFS) is a distributed file system designed for high throughput.
- When a file is loaded into HDFS, it is replicated and fragmented into “blocks” of data, which are stored across the cluster nodes; the cluster nodes are also called the DataNodes.
- The NameNode is responsible for storage and management of metadata, so that when MapReduce or another execution framework calls for the data, the NameNode informs it where the data that is needed resides.
- This enables reliable and rapid access.
- NameNode manages file system in memory (inodes) and authorization to file system.
- So DataNodes, maintain block integrity and send block reports to Namenode.

  [Architecture diagram]

Hbase

- HBase “is an open-source, distributed, versioned, column-oriented store” that sits on top of HDFS.

- HBase is based on columns rather than rows.
- This essentially increases the speed of execution of operations if they are need to be performed on similar values across massive datasets; for example, read/write operations that involve all rows but only a small subset of all columns.

Hive

- Hive provides a warehouse structure for other Hadoop input sources and SQL-like access for data in HDFS.
- Hive's query language, HiveQL, compiles to MapReduce and also allows user-defined functions (UDFs).
- Hive's data model is based primarily on three related data structures: *tables* correspond to HDFS directories that are divided into *partitions*, which in turn can be divided into *buckets*.

Pig

- Pig is a run-time environment that allows users to execute MapReduce on a Hadoop cluster.
- Pig Latin is a high-level scripting language on Pig platform. Like HiveQL in Hive, Pig Latin is a higher-level language that compiles to MapReduce.
- Pig's data model is similar to the relational data model, but here tuples can be nested. For example, a table of

tuples can have a table in the third field of each tuple. In Pig, tables are called bags.

- Pig also has a “map” data type, which is useful in representing semi-structured data such as JSON or XML.”

Sqoop

- Sqoop (“SQL-to-Hadoop”) is a tool which transfers data in both ways between relational systems and HDFS.
- Sqoop can be used to import data from external structured databases into HDFS or any other related systems such as Hive and HBase.
- On the other hand, Sqoop can also be used to extract data from Hadoop and export it to external structured databases such as relational databases and enterprise data warehouses.

Limitations of Hadoop

1. HDFS cannot be mounted directly by an existing operating system. Getting data into and out of the HDFS file system can be inconvenient.
2. Hadoop security model is limited to HDFS file system permissons and is disabled by default due to complexity.
3. Hadoop framework is not fit for small data.
4. Stability issues in the API.