**Data Lake**

**MONGO DB**

1. It has data lake creation option but it uses cloud storage like S3 ask for credentials.
2. It can create database which can store structure ,semi-structure and unstructured data in DB , so hence we can call it that we can use it as data lake.
3. We created data base and ingest the data from different sources like flat files(local),Web API( kaggle dataset) and Database data( postgresql)

**AZURE CLOUD**

1. Using free resources and created data lake and it can store raw data.
2. Ingest the data using connection string of that resources.
3. Ingest the different type of source of data like local files ,web API(like kaggle) and database data in the form of tables (postgresql)

**CLOUD ERA**

1. Installation complete in virtual box.
2. Working on in this…

**HDFS(hadoop distributed file system)**

1. store data as Data lake
2. we can delete data but can not modify it.
3. Namenode(metadata)stores the data about data.
4. Block size(size will store in one machine(container)
5. ELT(extract load and then transform)
6. Apache Sqoop use to store SQL data(structured like tables of database)
7. Apache FLUME is used to get the unstructured,semi-structured data and send from one point to another point only.
8. KAFKA gets the data from FLUME and sends the original file to HDFS and then sends the copy file to others like (apache spark streaming,flink,storm streaming and new application etc)it can make copies but FLUME cant. KAFKA uses its own cluster so it can store the data and then make copies.
9. spark streaming analyzes the data by getting the file from KAFKA.It can do real time processing.
10. Other tools for processing(Flink and storm etc)
11. In HDFS it uses(PIG,Map reduce and Hive) PIG ,MR are old now we use HIVE and SPARK
12. spark can do nearly real time processing by using the data of HDFS.
13. HBase is a hadoop database not datalake by default it is NoSQL it can use SQL queries by translating it into PHOENIX(query).
14. IMPALA(query) used in CloudEra.

**Hadoop**

1) Lists the contents of a directory in HDFS.

**Command:** **hadoop fs –ls /**

2) Creates a new directory in HDFS.

**Command:** **hadoop fs -mkdir [path]**

3) Copies files or folder from the local file system to HDFS.

**Command : hadoop fs -put [local\_file] [hdfs\_path]**

4) Same as put, copies files from the local file system to HDFS.

**Command : hadoop fs -copyFromLocal [local\_file] [hdfs\_path]**

5) Copies files from HDFS to the local file system.

**Command: hadoop fs -copyToLocal [hdfs\_path] [local\_file]**

6) Copies files from HDFS to the local file system.

**Command: hadoop fs -copyToLocal [hdfs\_path] [local\_file]**

7) Same as copyToLocal, copies files from HDFS to the local file system.

**Command: hadoop fs -get [hdfs\_path] [local\_file]**

8) Displays the contents of a file in HDFS.

**Command : hadoop fs -cat [hdfs\_file]**

9) Displays the last part of a file in HDFS.

**Command: hadoop fs -tail [hdfs\_file]**

10) Deletes files or directories in HDFS.

**Command : hadoop fs -rm [hdfs\_path]**

11) Deletes a directory in HDFS.

**Command: hadoop fs -rmdir [hdfs\_path]**

12) Moves or renames files or directories in HDFS.

**Command : hadoop fs -mv [src] [dest]**

13) Copies files or directories within HDFS.

**Command : hadoop fs -cp [src] [dest]**

14) Displays disk usage information for files and directories in HDFS.

**Command : hadoop fs -du [path]**

15) Empties the trash in HDFS.

**Command : hadoop fs -expunge**

**APACHE SPARK**

**INTRODUCTION**

Spark is able to do all processing tools work. like storm,Flink..etc.In Spark there are many libraries which can all the processing according to the needs like ML lib , SQL ,GraphX ,Streaming in different languages(python ,java etc).zookeeper in spark hold the metadata both files(Namenode(active and standby)).



1. sqoop use to get data from RDBMS.
2. send to HIVE,Hbase,HDFS

**YARN**

1. Monitor resources
2. Resource manager(which is master)