**Hadoop - Task 1**

**Hadoop:**

It is a **framework that allows for the distributed processing of large data sets across clusters** of computers using simple programming models.

**HDFS:**

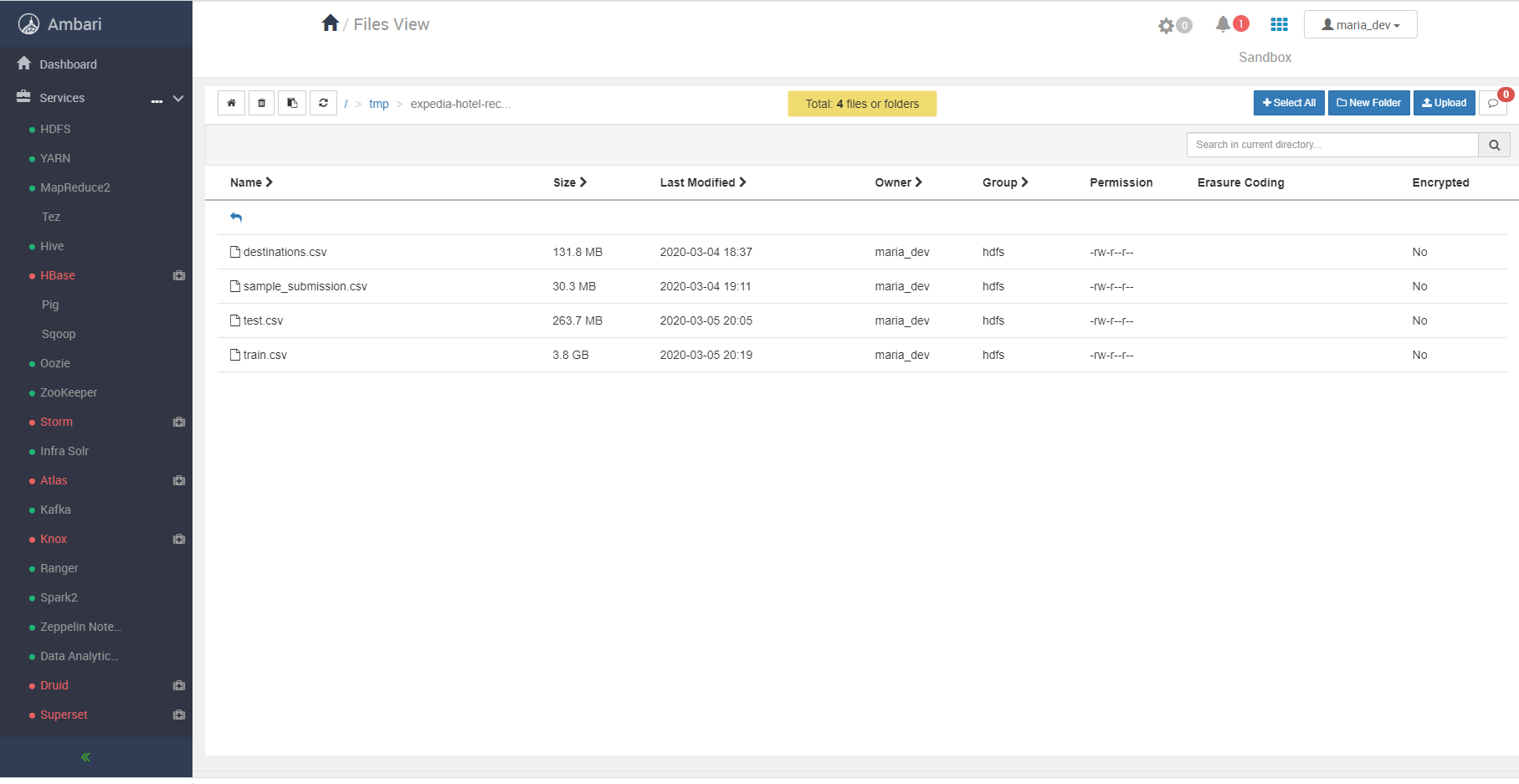
HDFS can defined as:

* Hadoop distributed file system
* written in java
* store files across machine in a cluster

**Hadoop 3 Features:**

1. Upgrade Java 7 to Java 8
2. Refactoring core pieces of HDFS
3. Class path isolation - GPU isolation, GPU's are expensive resources. It enables the access all over the company.
4. Agility and time to Market: Use Container to build apps quickly & draw one out quickly.
5. Reduce the Total cost overhead.
6. Hadoop 2 has 6 blocks & 3 replications of each block. = 18 block space
7. Hadoop 3 has 6 blocks & 3 parity = 9 blocks
8. Storage overhead can reduce 200% to keep SQL server
9. Scalability & Availability:
10. Hadoop 1 & 2: 1 Name Node to manage all namespaces, Only one standby Node
11. Hadoop 3: Has multiple name node for multiple namespaces, Multiple standby Nodes, new intra-Data Node balancing functionality, which is invoked via the HDFS disk balancer CLI.
12. New Use Cases:
13. Hadoop 3: Enable scheduling of additional resources such as disk & GP use, for better interpretation with the containers with deep learning & machine learning.
14. Inter-Queue preemption to the next level time by allowing preemption in the application in a single queue. So, the user can prioritize job by user driven or application priority.
15. Microservices are added to the Yarn to enable ML, Deep Learning.

**Uploaded files to HDFS:**



In a above image, I have downloaded files from <https://www.kaggle.com/c/expedia-hotel-recommendations/data> and uploaded all files in /tmp/expedia-hotel-recommendations-data/ .