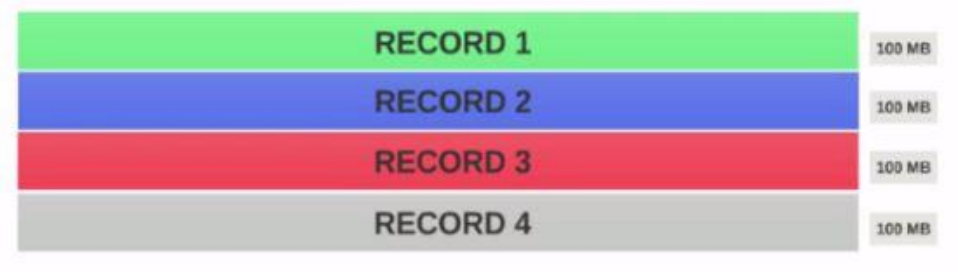
# Difference between Block Split and Input Split:

Block Split : Actual Data Split based on Hadoop versions

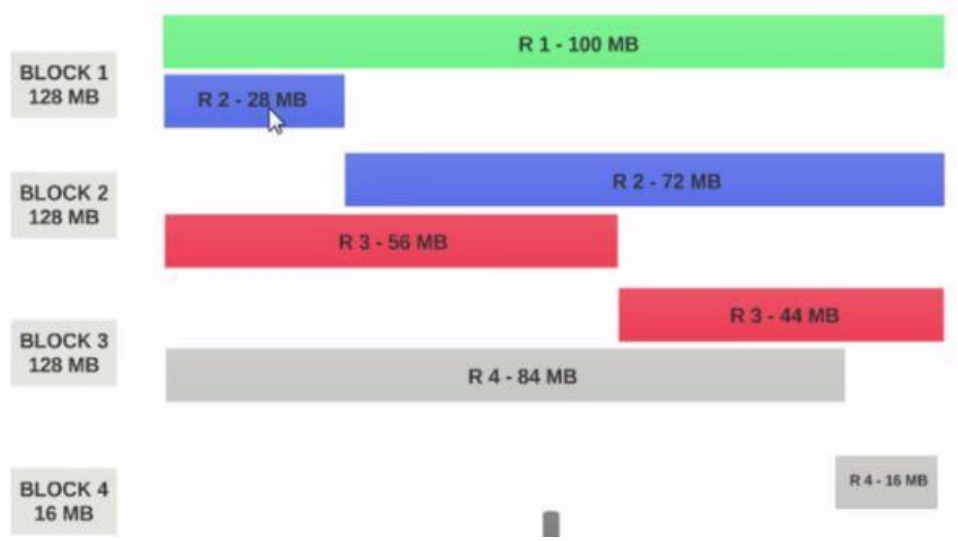
Input Split : Logical Splits, Contains the memory address of start and end locations of each records.

Input splits doesn't contain actual data, rather it has the storage locations to data on HDFS and usually, the size of Input split is same as block size

- Assume we have a file of **400MB** with consists of **4 records**(**e.g** : csv file of 400MB and it has 4 rows, 100MB each)




- If the HDFS **Block Size** is configured as **128MB**, then the 4 records will not be distributed among the blocks evenly. It will look like this.




- **Block 1** contains the entire first record and a 28MB chunk of the second record.
- If a mapper is to be run on **Block 1**, the mapper cannot process since it won't have the entire second record.
- This is the exact problem that **input splits** solve. **Input splits** respects logical record boundaries.

- Let's Assume the **input split** size is **200MB, which means it doesn't have 200 mb of actual data, it has the storage locations or memory address of the actual data which is present in HDFS.**
- **So, the memory locations will be passed by the input splits which is referred as logical split to the mapper, then the mapper, with the help of the input splits it will redirect to the hdfs locations and fetch the corresponding records and process them.**


- Therefore the **input split 1** should have both the record 1 and record 2. And input split 2 will not start with the record 2 since record 2 has been assigned to input split 1. Input split 2 will start with record 3.
- This is why an input split is only a **logical chunk** of data. It points to start and end locations with in blocks.