**MapReduce InputSplit vs HDFS Block in Hadoop**

**Boost your career with** [***Free Big Data Course!!***](https://data-flair.training/hadoop-spark-developer-course/)

**1. Objective**

In this [**Hadoop**](https://data-flair.training/blogs/hadoop-tutorial-for-beginners/) InputSplit vs Block tutorial, we will learn what is a block in [**HDFS**](https://data-flair.training/blogs/comprehensive-hdfs-guide-introduction-architecture-data-read-write-tutorial/), what is [**MapReduce**](https://data-flair.training/blogs/hadoop-mapreduce-introduction-tutorial-comprehensive-guide/) InputSplit and difference between MapReduce InputSplit vs Block size in Hadoop to [deep dive into Hadoop fundamentals](https://data-flair.training/blogs/hadoop-features-design-principles-tutorial/).

MapReduce InputSplit vs HDFS Block in Hadoop

**2. MapReduce InputSplit & HDFS Block – Introduction**

Let us start with learning what is a block in Hadoop HDFS and what do you mean by Hadoop InputSplit?

**Block in HDFS**

Block is a continuous location on the hard drive where data is stored. In general, FileSystem stores data as a collection of blocks. In the same way, HDFS stores each file as blocks. The Hadoop application is responsible for distributing the data block across multiple nodes. [Read more about blocks here.](https://data-flair.training/blogs/data-blocks-hdfs-hadoop-distributed-file-system/)

**InputSplit in Hadoop**

The data to be processed by an individual [**Mapper**](https://data-flair.training/blogs/hadoop-mapper-in-mapreduce/) is represented by InputSplit. The split is divided into records and each record (which is a [**key-value pair**](https://data-flair.training/blogs/key-value-pair-in-hadoop-mapreduce/)) is processed by the map. The number of map tasks is equal to the number of InputSplits.

Initially, the data for MapReduce task is stored in input files and input files typically reside in HDFS. [**InputFormat**](https://data-flair.training/blogs/hadoop-inputformat-types/) is used to define how these input files are split and read. InputFormat is responsible for creating InputSplit. [Learn more about MapReduce InputSplit](https://data-flair.training/blogs/inputsplit-in-hadoop-mapreduce/).

**3. MapReduce InputSplit vs Blocks in Hadoop**

Let’s discuss feature wise comparison between MapReduce InputSplit vs Blocks-

**i. InputSplit vs Block Size in Hadoop**

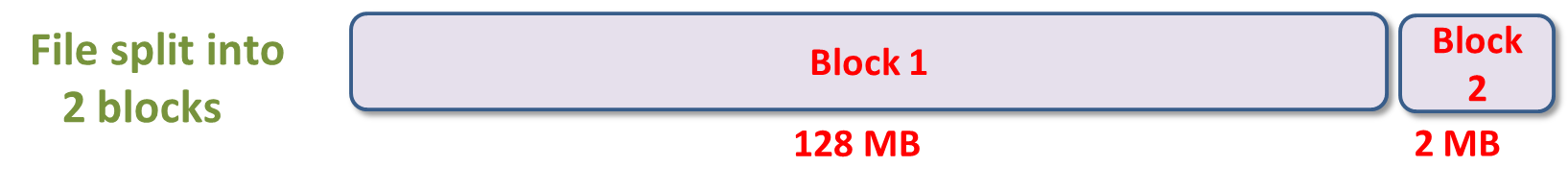
* **Block –**The default size of the HDFS block is 128 MB which we can configure as per our requirement. All blocks of the file are of the same size except the last block, which can be of same size or smaller. The files are split into 128 MB blocks and then stored into Hadoop FileSystem.
* **InputSplit –**By default, split size is approximately equal to block size. InputSplit is user defined and the user can control split size based on the size of data in MapReduce program.

[***Follow DataFlair on WhatsApp***](https://whatsapp.com/channel/0029Va9NEb7HLHQVx3DGce3A) ***& Stay updated with latest technology trends.***

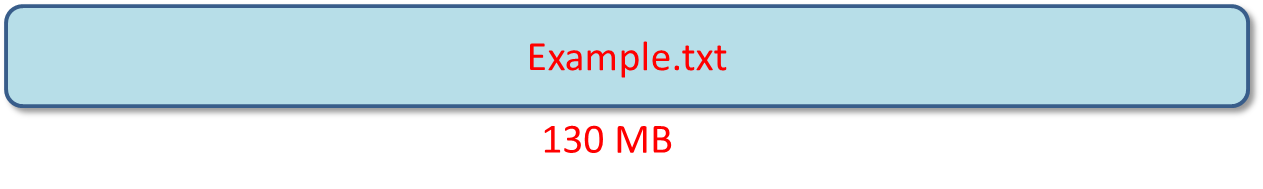
**ii. Data Representation in Hadoop Blocks vs InputSplit**

* **Block –**It is the physical representation of data. It contains a minimum amount of data that can be read or write.
* **InputSplit –**It is the logical representation of data present in the block. It is used during data processing in MapReduce program or other processing techniques. InputSplit doesn’t contain actual data, but a reference to the data.

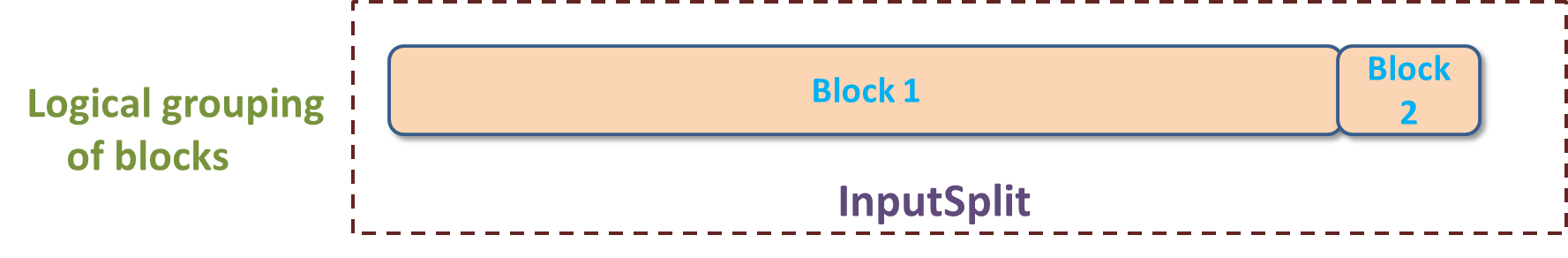
**iii. Example of Block vs InputSplit in Hadoop**

[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/04/block-eample-2.png)

InputSplit vs Block

[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/04/block-example-1.png)

InputSplit vs Block

[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/04/block-eample-3.png)

inputSplit vs Block

Consider an example, where we need to store the file in HDFS.  HDFS stores files as blocks. Block is the smallest unit of data that can be stored or retrieved from the disk and the default size of the block is 128MB. HDFS break files into blocks and stores these blocks on different nodes in the cluster. Suppose we have a file of 130 MB, so HDFS will break this file into 2 blocks.

Now, if we want to perform MapReduce operation on the blocks, it will not process, because the 2nd block is incomplete. Thus, this problem is solved by InputSplit. InputSplit will form a logical grouping of blocks as a single block, because the InputSplit include a location for the next block and the byte offset of the data needed to complete the block.

**4. Conclusion**

From this, it is concluded that InputSplit is only a logical chunk of data i.e. it has just the information about blocks address or location.

During MapReduce execution, Hadoop scans through the blocks and create InputSplits and each inputSplit will be assigned to individual mappers for processing. Hence, Split act as a broker between block and mapper.