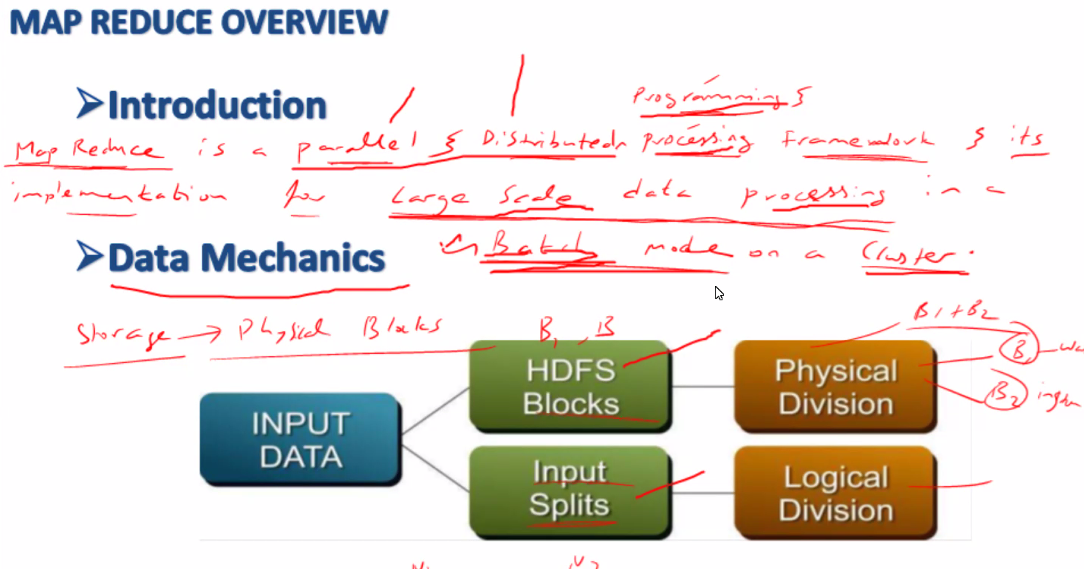
# MAP REDUCE OVERVIEW

Map reduce is a parallel and distributed programming and processing framework and its implementation for large scale data processing in a BATCH mode on a cluster.



* On Demand
* Scheduled
* Batch mode, online not supported

**Inputsplit**

Input splits are logical, an information given to Map reduce

We have below message of 130mb which is stored across 2 blocks

The last line falls in 2nd block as it exceeds 128mb, but considered to be only 1 single Inputsplit as it is a continuation of data (broken line) from last line of blk1

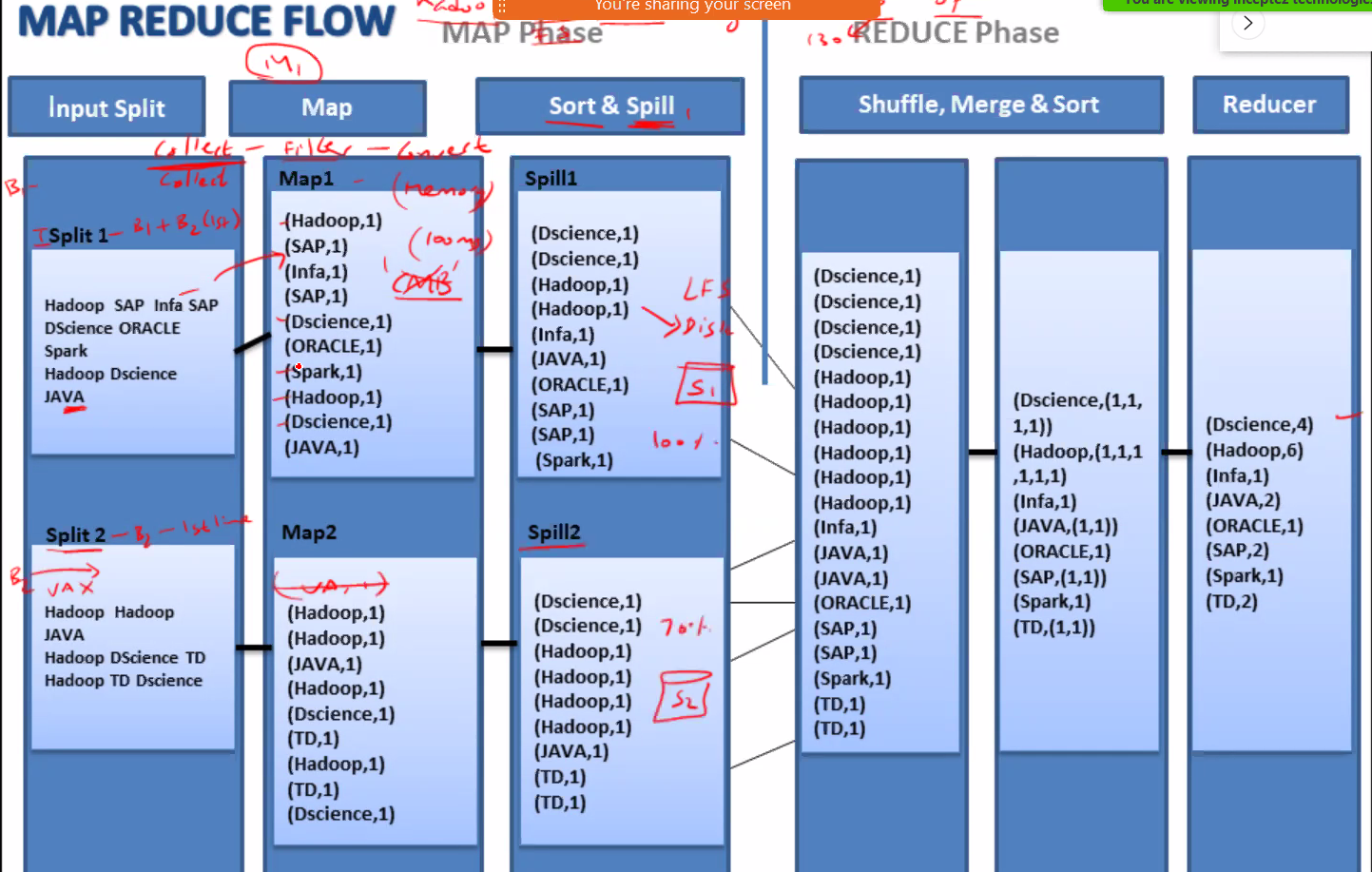
Regards,Basha 🡪 is a new line, not broken line so it is considered to be a new Inputsplit

|  |  |  |  |
| --- | --- | --- | --- |
| You make me feel like I can do anything and  I am so happy to be with you.  Thank you for being the wonderful,  amazing person that you are.  You surprise me every day and  you warm my heart every night.  I am the person I am today because  you've loved me ~~and helped me love~~ | new para with broken line | Input Split 1 | Blk1 |
| and helped me love | continuation of last line | Blk2 |
| Regards, Basha | New line | InputSplit2 |

No of Inputsplit <= No of Blocks

Inputsplit are logical splits of data whereas Blocks are physical split of data

Inputsplit are information given to mapper to know how to read data from blocks



MAPPER – Collects

REDUCER – Aggregates

|  |  |
| --- | --- |
| **Map** | |
| Inputsplit | Inputsplit gives the logical information of data spread across the blocks to be collected to Map. Using this no of MAP are initiated |
| MAP 1 | Collects the data thru information provided by Inputsplit |
| SORT | Sort the collected data |
| Spill | Write the sorted data to local HDD |
| MAP2, SORT, Spill |  |
| MAP2, SORT, Spill |  |

|  |  |
| --- | --- |
| **Reduce** | |
| Shuffle | Collect the data from spill as soon as at least 1 MAP is 100% |
|  | Collect data from MAP2 |
|  | Collect data from MAP3 |
| Merge | Does not start as long as shuffle is complete |
|  | Merge all data from shuffle |
| Sort | sort the merged data |
| Reducer | Reduce the data (apply aggregate function) |