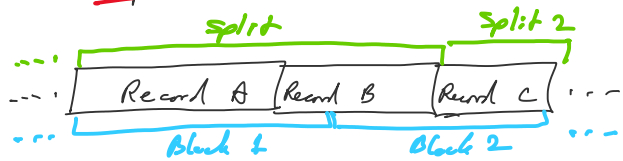
Map-Reduce Concepts

Tuesday, January 10, 2023 7:10 AM

- Records / Blocks
  - DFS / object store
    - HDFS ⇒ Block size
    - S3 ⇒ Byte offsets } Defaults
  - MR Tasks
    - Record: Indv ⇒ Batch
    - Splits: Group of records that I can run tasks on (in parallel)

Ex.

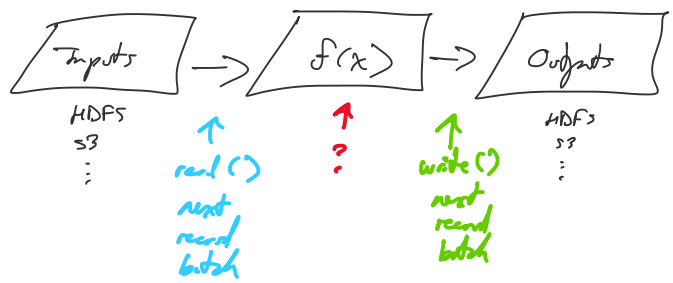


*Hadoop:
- Readers ⇒ S3, HDFS ...
- Yields splits for complete records in MR jobs

↳ Record Delimiter → Binary, Text } Byte Marker

- Batch Data Analysis
  - Serial / Sequential Processing
    - Read one or more objects / files
    - Perform transformations / calculations
    - Write one or more objects / files

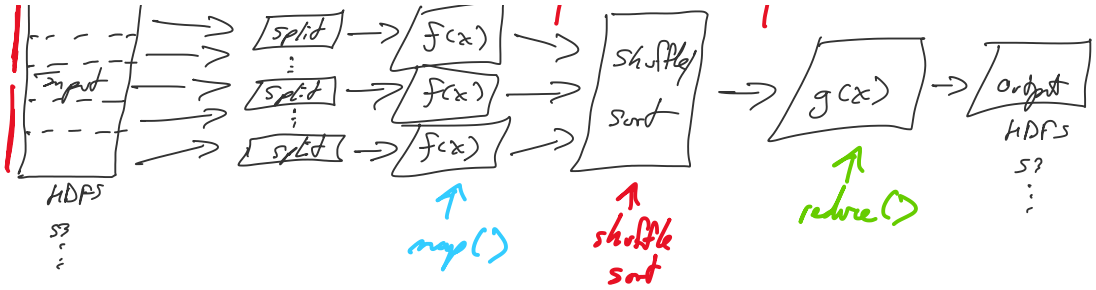
Ex.



  - Parallel / Concurrent Processing
    - Read one or more objects / files → Input Splits
    - Read record batch from a split
    - Perform transformations / calculations on a batch / split
    - Write one or more objects / files from a split → Intermediate Output } Parallel
    - Collect & Organize intermediate objects / files
    - Read record batches of sorted intermediate data
    - Perform transformations / calculations on sorted intermediate data } Sequential
    - Write one or more objects / files → Final Output

Ex

I/O


Input
HDFS
split
split
split
f(x)
f(x)
f(x)
map()
Shuffle/
Sort
shuffle
sort
g(x)
reduce()
output
HDFS