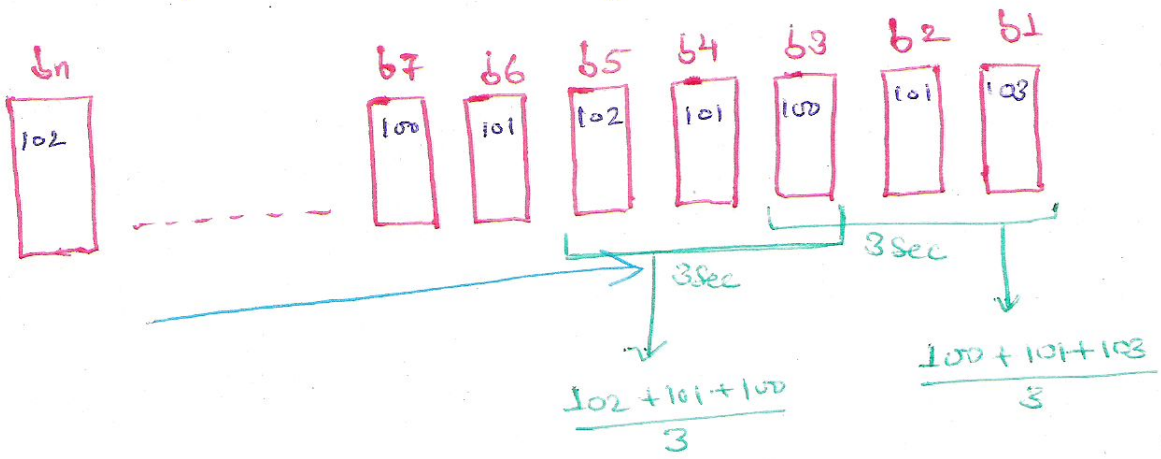
SPARK STREAMING PART-3 STATEFULL (WINDOW) TRANSFORMATIONS

By www.HadoopExam.com

Note: These instructions should be used with the HadoopExam Apache Spark: Professional Trainings. Where it is executed and you can do hands on with trainer.

Statefull Transformation
[Spark Streaming]

Depends on previous batches of RDD in a DStream.



⇒ There are mainly two types of windowed operations

① Sliding window of time periods

② UpdateStateBykey : Used to track state across events for each key

① Windowed Transformation :- Involves more than one batch in DStream to capture/calculate result.

⇒ Windowed operation depend on two parameters.

(A) Window duration

(B) Sliding Duration

(A) Window duration :- Whenever you start calculations then you need to find how many previous batches of data needs to be considered.

1 sec. 2 sec. (Batch)

[Current Batch] + [No. of previous batch]

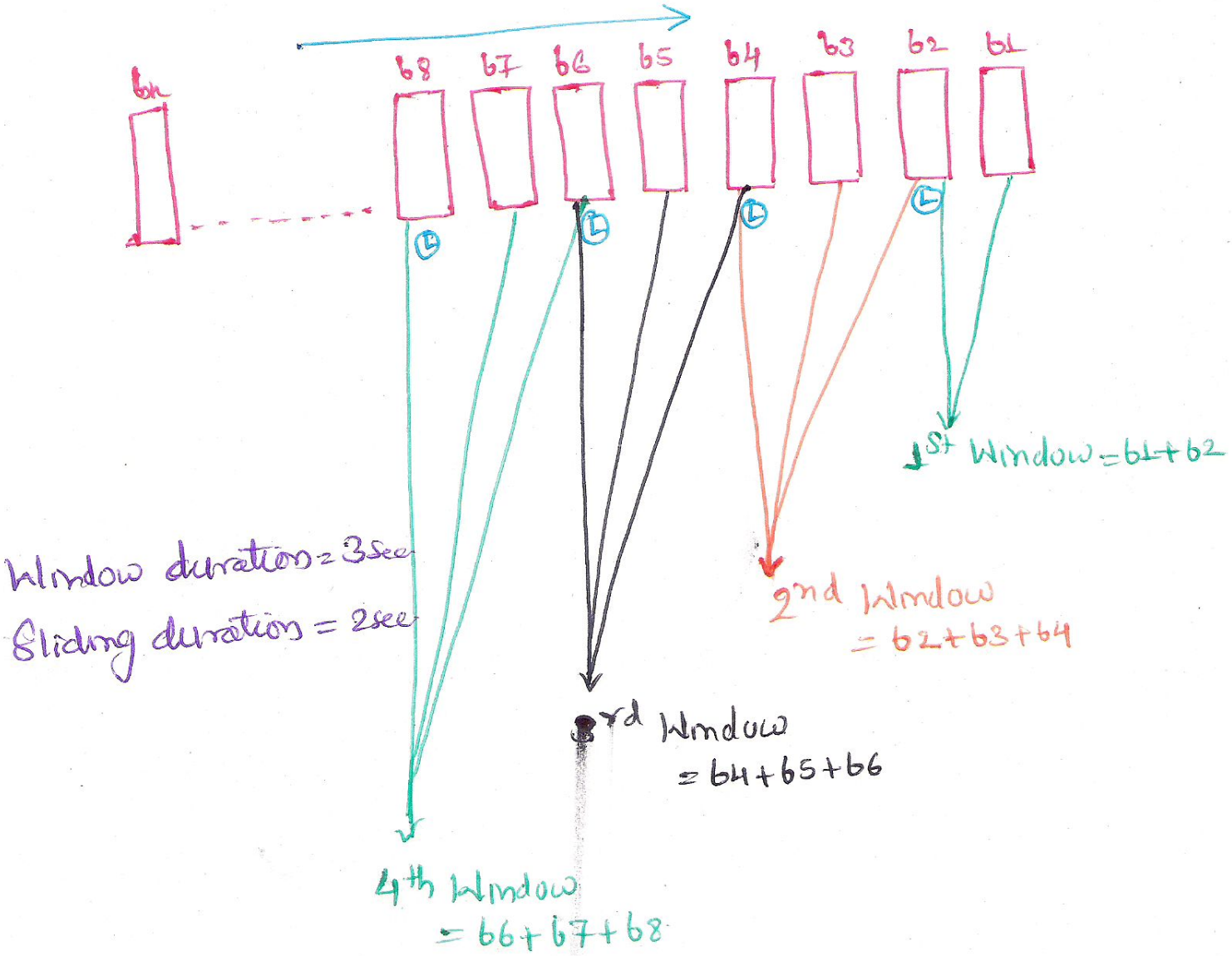
=> Window duration = 1 + 2 = 3 sec.

(B) Sliding Duration :- By default that is equal to same as batch interval. So in above case it would be 1 sec.

→ How frequently you want to calculate your results.

→ if 1 sec, then on each batch arrival (1 sec.) calculation will be triggered.

→ Suppose sliding duration we set 2 sec. & window duration = 3 sec.



⇒ Some operation on Dstream:—

Window:— It returns new Dstream, contains data from multiple batches.

e.g. 1st batch = (b1+b2)
2nd batch = (b2+b3+b4)
3rd batch = (b4+b5+b6)
4th batch = (b6+b7+b8)

Apply operation on it count(), transform() etc. on all the contents on one window data.

⑤ reduceByKeyAndWindow:- Apply reduce function to run on whole window. Such as + operation.

reduce → +

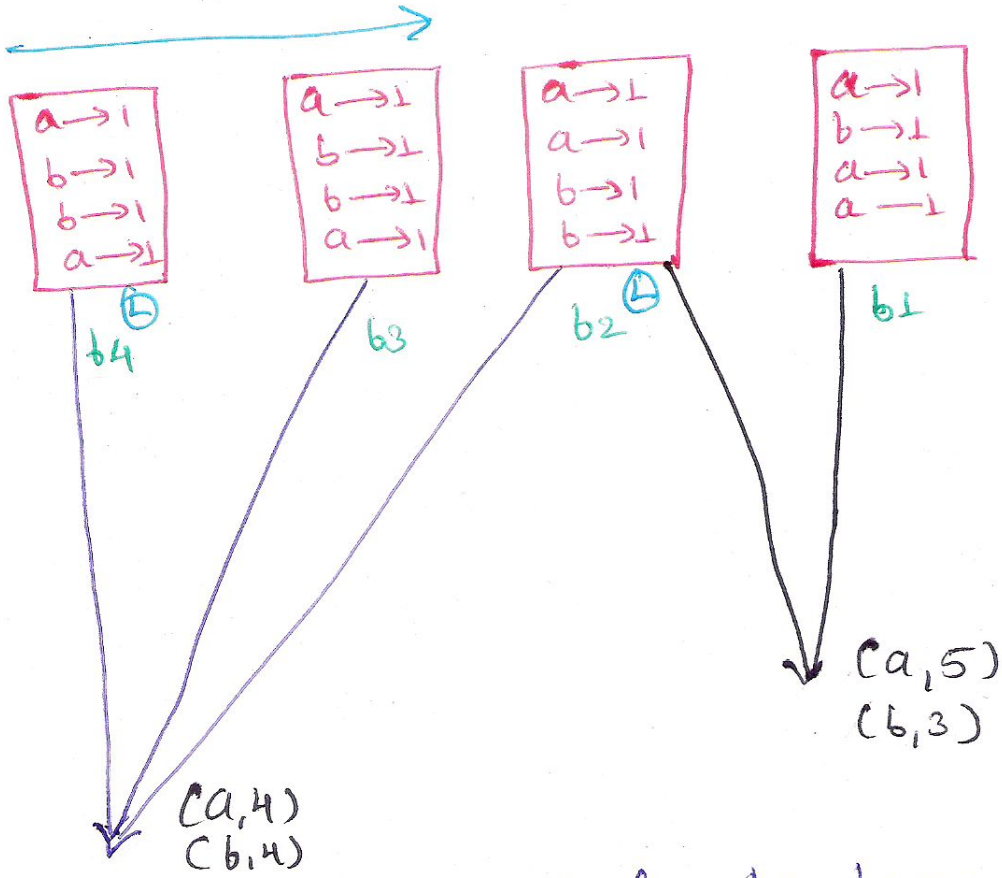
Example:-

WordCount In a Stream:

```
Val dataStream = ContentDataStream.map( word => (word, 1));
Val countData = dataStream.reduceByKeyAndWindow(
                { (x,y) => x+y },
                { (x,y) => x-y },
                Seconds(3), ——→ Window Duration
                Seconds(2)) ——→ Sliding window.
```

→ New batch comming in window, hence apply first reduce function on it. So for new data value will be calculated.

→ To avoid duplicate calculations, remove previous batch data.



Now, we need to remove data from b2, hence second function will be used, which will removed data internally ~~for b2~~ of b2. using (x-y)

⇒ In above example, it compute the reduction incrementally, considering only new data comming into the window and which data is going out.

⇒ Second function is inverse of first reduce function (-ve)

~~other~~ few more functions

ReduceByWindow : Reducing data from each window, without removing old batch data.

CountByWindows: Number of elements in each window, again ~~removing~~ without removing old batch data.

CountByValueAndWindow: produce count for each value in a window.

HadoopExam Learning Resource provides the following material for the Advanced Technologies.
Please visit www.HadoopExam.com for more detail this is just a few products from portfolio.

Price start for training with Just $79/3500INR

Apache Spark Professional Training with HandsOn Session + Certification Material

Hadoop Professional Training with HandsOn Sessioon + Certification Material

HBase Professional Training with HandsOn Session + Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material