

1. [Apache Spark Professional Training with Hands On Lab Sessions](#)
2. [Oreilly Databricks Apache Spark Developer Certification Simulator](#)

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.**

1. Hadoop Training
2. Spark Training
3. HBase Training
4. MapR Developer
5. MapR HBase
6. CCA500 Certification
7. Spark Certification
8. EMC Data Science

Hadoop Specialization offer == 50% + 35% off

Hadoop Expert

~~52000INR ==~~ **16900INR Only**
~~\$1150 ==~~ **\$373 Only**
Hadoop Specialization offer

* @ End of the Offer Prices will increase by 25%

Limited Time Offer (Less Than 5Days Remain)



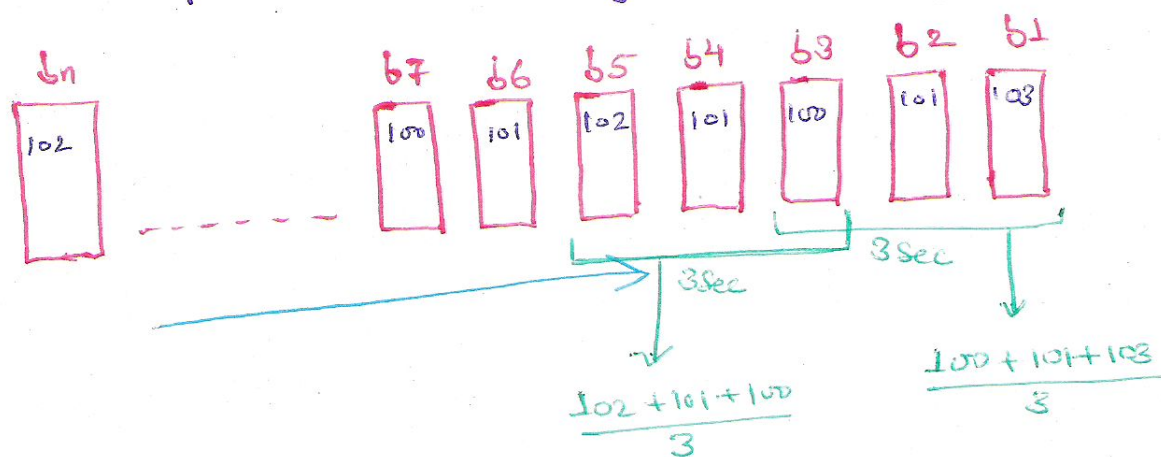
[Cloudera CCA175 \(Hadoop and Spark Developer Hands-on Certification available with total 75 solved problem scenarios. Click for More Detail\)](#)

1. Window Transformation
2. Window Duration and Sliding Duration
3. DStream Operations
4. WordCount in DStream

①

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 operations depend on two parameters.

(A) Window duration

(B) Sliding duration

(2)

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

$$\boxed{\text{Current Batch}} + \boxed{\text{No. of previous batch}} \times \text{Batch Duration}$$

1 Sec. 2 Sec. (Batch)

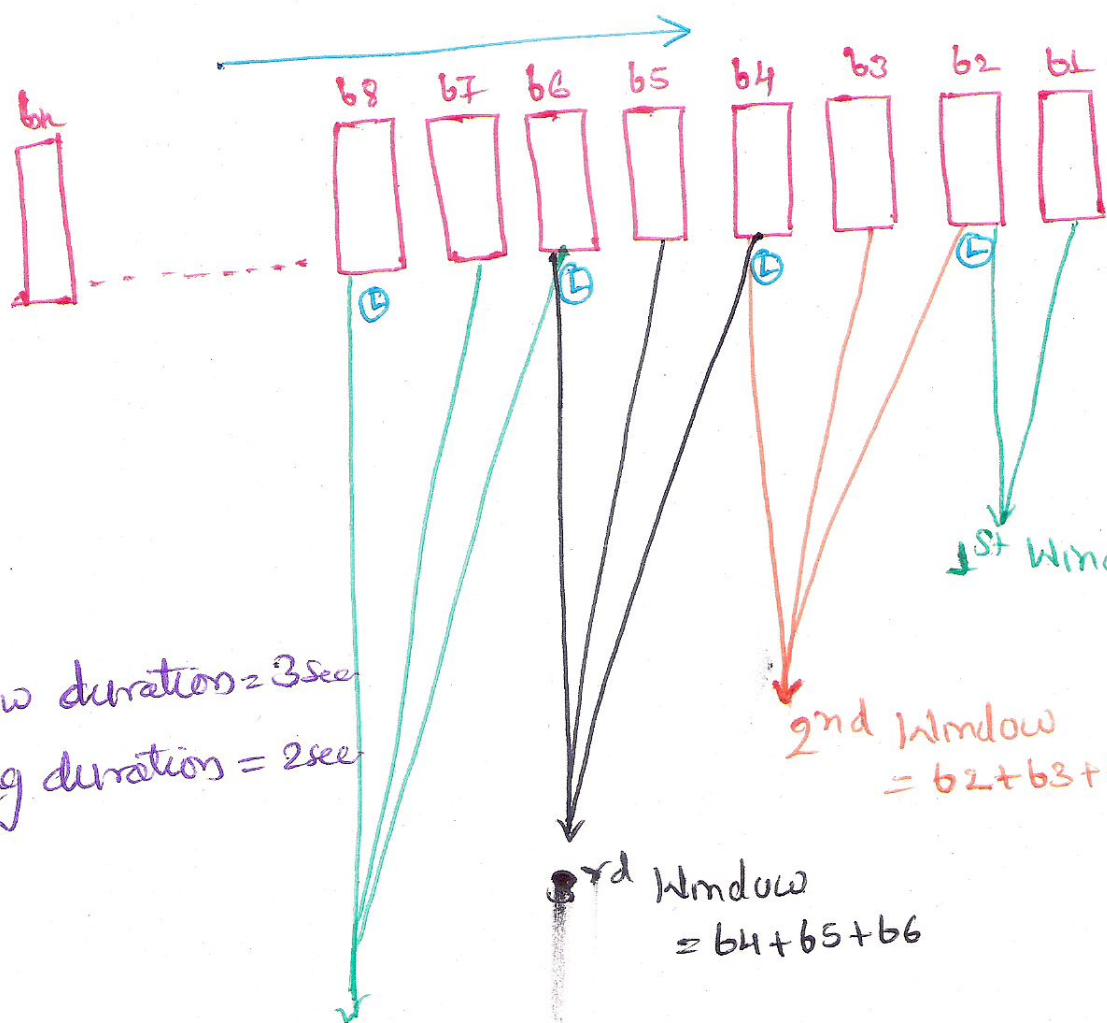
$$\Rightarrow \text{Window duration} = 1 + 2 = 3 \text{ Sec.}$$

Ⓑ **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.) calculations will be triggered.

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

3



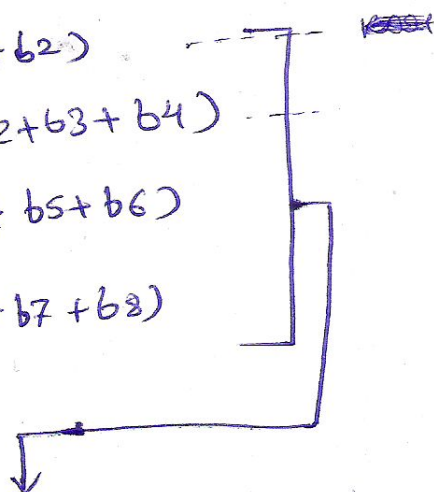
Window duration = 3 sec
Sliding duration = 2 sec

⇒ Some operation on DStream: —

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

e.g.

- 1st batch = $(b_1 + b_2)$
- 2nd batch = $(b_2 + b_3 + b_4)$
- 3rd batch = $(b_4 + b_5 + b_6)$
- 4th batch = $(b_6 + b_7 + b_8)$



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

② ReduceByKeyAndWindow: - Apply reduce function to run on whole window. Such as + operation.
reduce $\rightarrow +$

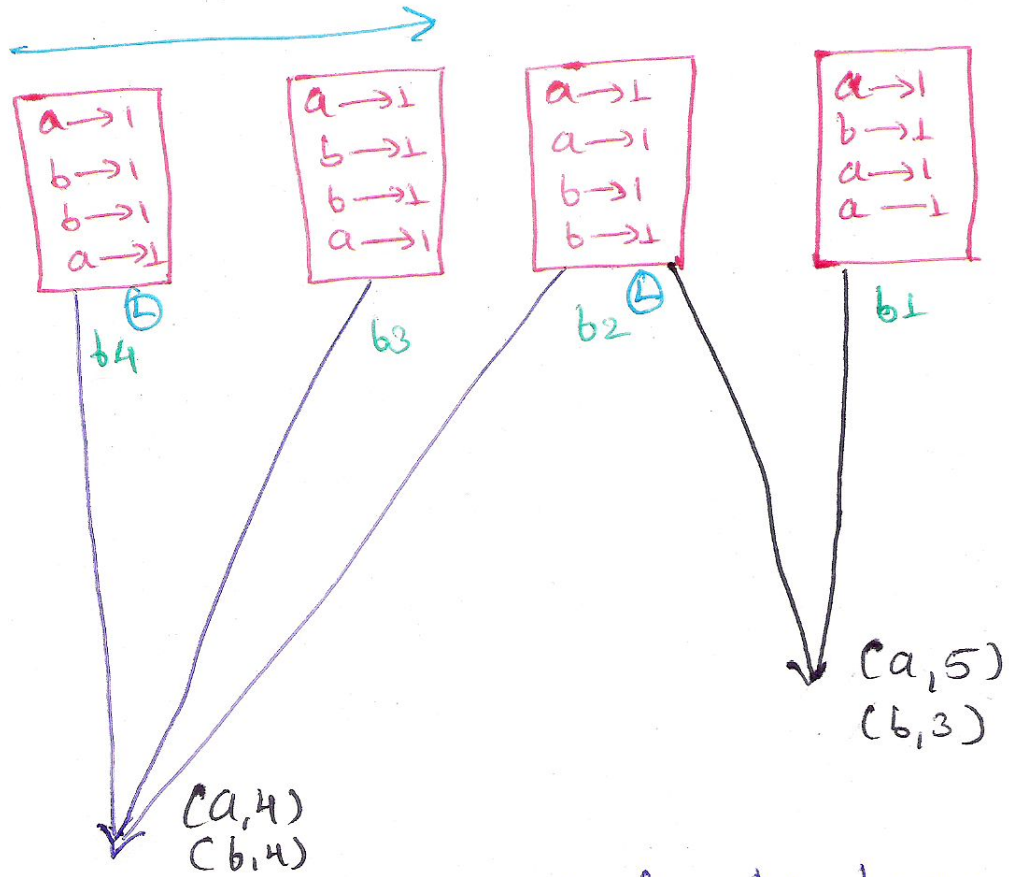
Example: -

Word Count In a Stream:

```
val dataStream = ContentDataStream.map(word => (word, 1));  
val countData = dataStream.reduceByKeyAndWindow(  
    { (x, y) => x + y },  
    { (x, y) => x - y },  
    seconds(3),  $\rightarrow$  Window Duration  
    seconds(2)  $\rightarrow$  Sliding window.
```

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

\rightarrow To avoid duplicate calculations, remove previous batch data.



Now, we need to remove data from b_2 , hence second function will be used, which will remove data internally for b_2 of b_2 . using $(x-y)$

\Rightarrow In above example, it computes the reduction incrementally, considering only new data coming into the window and which data is going out.

\Rightarrow Second function is inverse of first reduce function (-ve)

Other few more functions

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

CountByWindow:
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 Session

+ Certification
Material



HBase Professional
Training with
HandsOn Session

+ Certification
Material



Certification
Material



Certification
Material



Certification
Material



Certification
Material



Certification
Material



Microsoft Azure

Certification
Material



Certification
Material