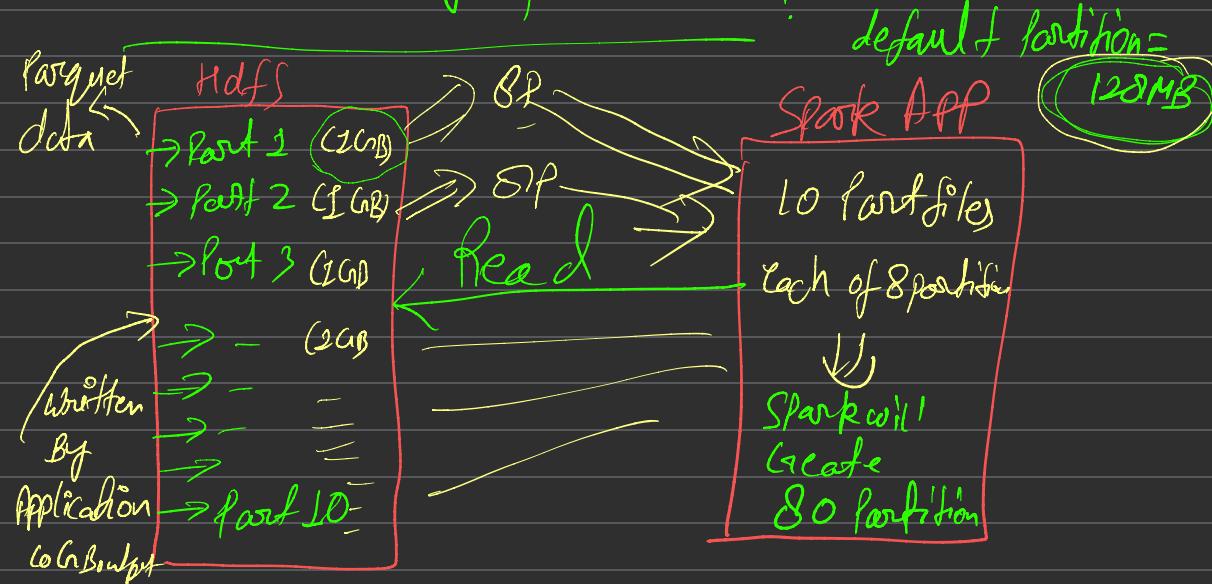


If source data is in parquet then
how many partition ?

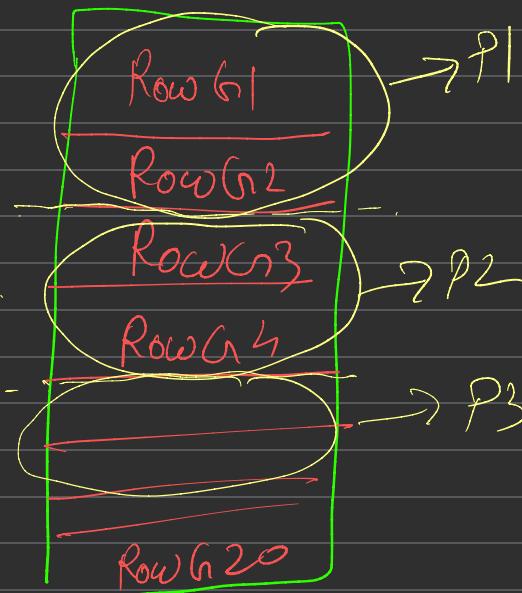


Portfile

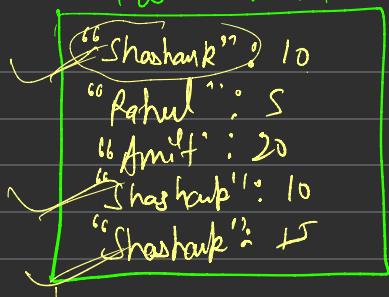
LGB

2

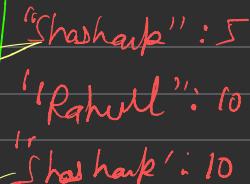
89



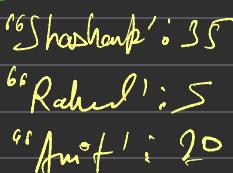
Partition-1



Partition-2



Partition-1

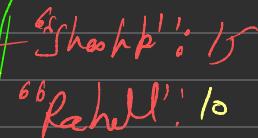


Group

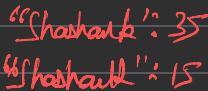
Hash

Aggregate

Partition-2



Partition-1

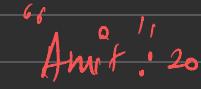


group by
(Shuffle → Exchange)

Partition-2



Partition-3



Shashank: 50

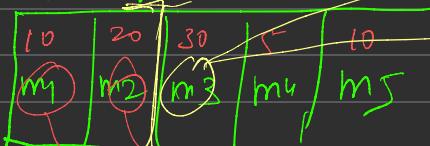
Rahul: 15

Amit: 20

Hash Aggregate

Checkpointing in Spark

Kafka Topic Partition



Spark

m_3

m_4

FTT checkpointing Checkpoint

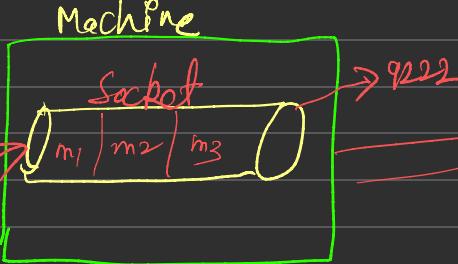
checkpointing

\Rightarrow (/tmp/checkpointing) \Rightarrow TFL which offers storage
 \Rightarrow Persistence = 30
 File System / Database / RAM

checkpointing Interval 10sec

Spark Streaming Word Count

Machine



Spark stream

producer

$m_1 \Rightarrow$ Hello Shashank

$m_2 \Rightarrow$ How are you Shashank

$m_3 \Rightarrow$ Shashank are you OK

Word = Lines. Select (explode (split(“”))). alias (“word”))

$m1 \Rightarrow \text{Hello Shashank}$,
after split

$\text{explode}([\text{"Hello"}, \text{"Shashank"}])$

after 1st pf

Words \Rightarrow

Word
Hello
Shashank

Word Counts =

Words	Count
Hello	1
Shashank	1

after 2nd

Words \Rightarrow

Word
Hello
are
you
Shashank

Word	Count
Hello	1
are	1
you	1
Shashank	1

Word	Count
Hello	1
Shashank	2
How	1
are	1
you	1

Word	Count
Hello	1
Shashank	3
How	1
are	2
you	2
OK	1