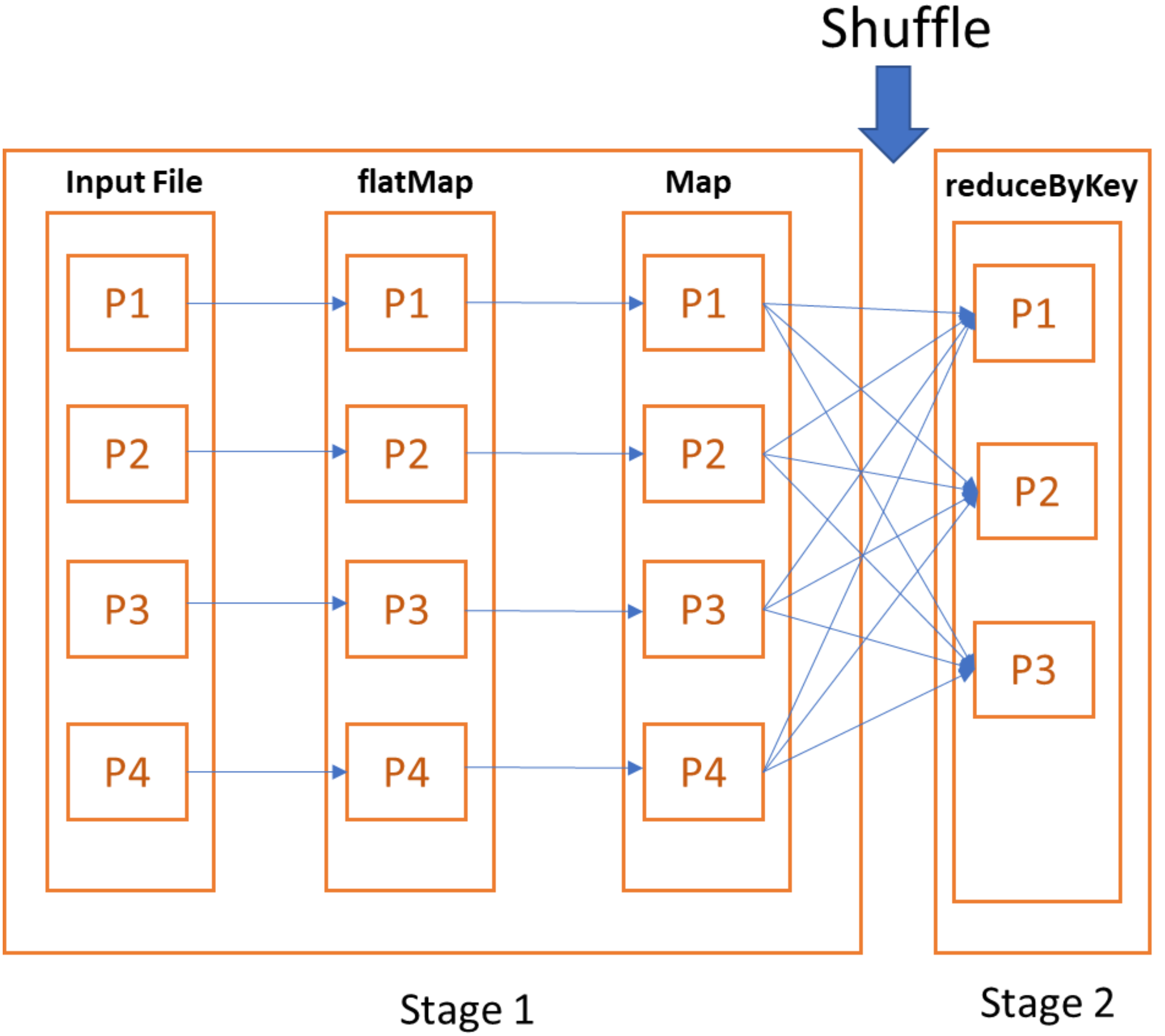
Data Shuffle in Apache Spark

@shuvajithazra

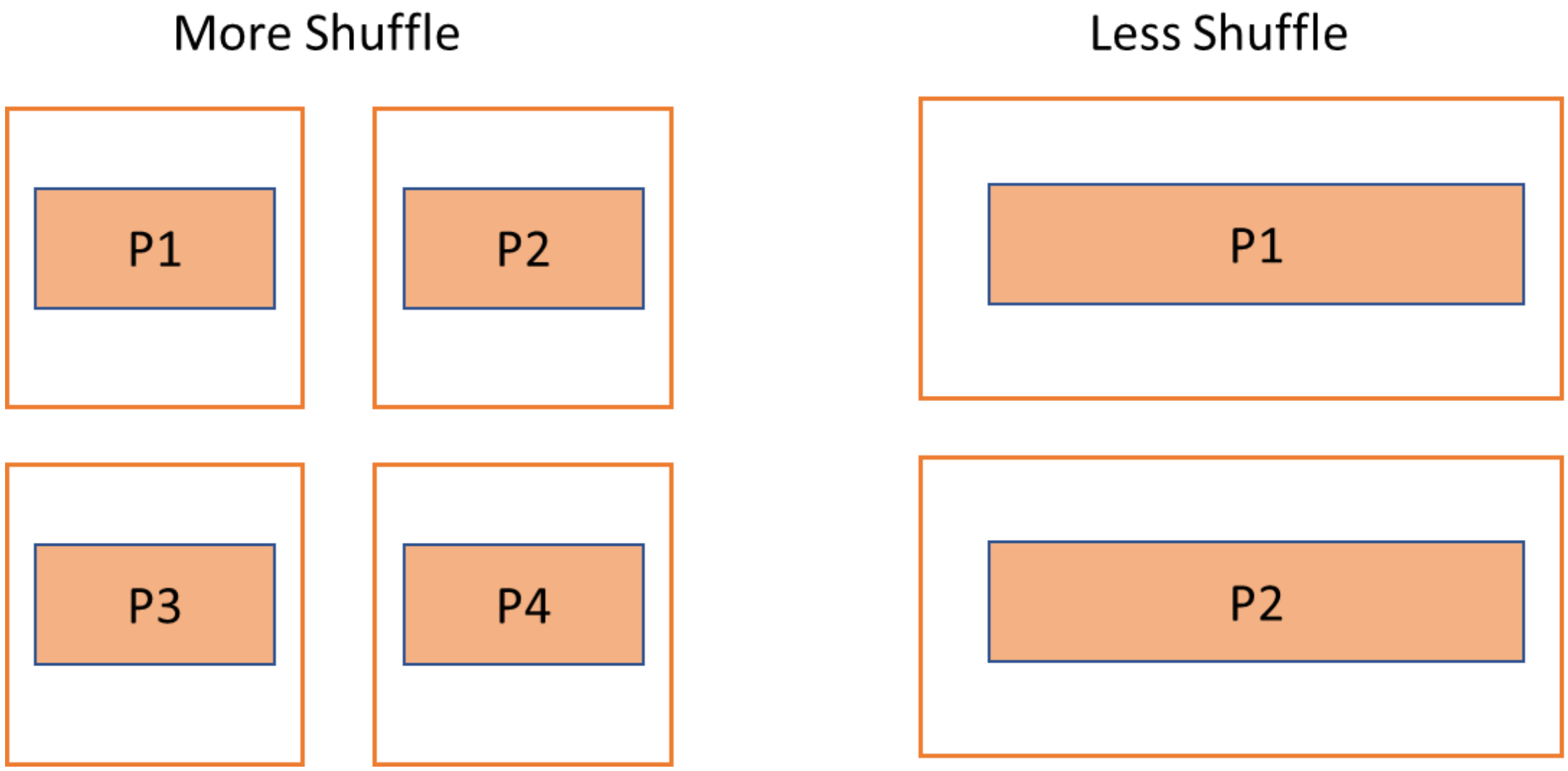
Data Shuffle in Apache Spark



Whenever you are doing any wide transformations like groupBy or join, data is distributed across the partitions. This is called shuffling.

@shuvajithazra

Mitigation Plan 1: Use fewer and larger Executors



Mitigation Plan 2: Filter out Unnecessary records

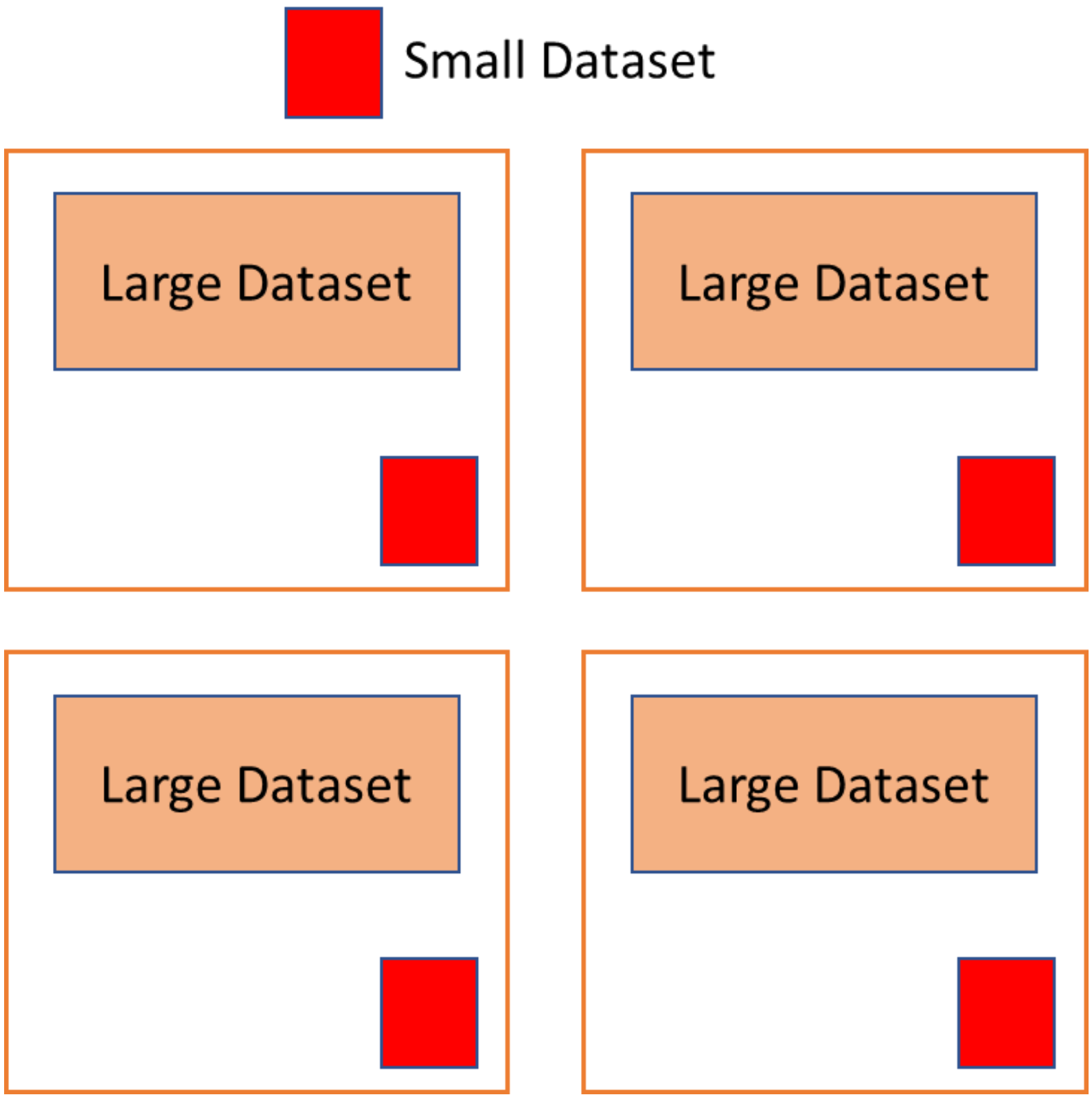
More Shuffle

```
df=df1.join(df2,df1.ID=df2.ID,inner)
```

Less Shuffle

```
df1=df1.filter(df1.ID<100)
df2=df2.filter(df2.ID<100)
df=df1.join(df2,df1.ID=df2.ID,inner)
```

Mitigation Plan 3: Use Broadcast Join If Possible



Other Mitigations:

- For joins, pre-shuffle the data with a bucketed dataset
- Denormalize the datasets - especially when the shuffle is rooted in a join

@shuvajithazra