Hadoop Combiner – Best Explanation to MapReduce Combiner

21 Apr, 2017 in MapReduce Tutorials by DataFlair Team

1. Hadoop Combiner / MapReduce Combiner

Hadoop Combiner is also known as "**Mini-Reducer**" that summarizes the Mapper output record with the same Key before passing to the Reducer. In this tutorial on **MapReduce** combiner we are going to answer what is a Hadoop combiner, MapReduce program with and without combiner, advantages of Hadoop combiner and disadvantages of the combiner in Hadoop.

Hadoop Combiner / MapReduce Combiner

2. What is Hadoop Combiner?

On a large dataset when we run **MapReduce job**, large chunks of intermediate data is generated by the Mapper and this intermediate data is passed on the Reducer for further processing, which leads to enormous network congestion. MapReduce framework provides a function known as **Hadoop Combiner** that plays a key role in reducing network congestion.

We have already seen earlier **what is mapper** and what is **reducer** in Hadoop MapReduce. Now we in the next step to learn Hadoop MapReduce Combiner.

The combiner in MapReduce is also known as 'Mini-reducer'. The primary job of Combiner is to process the output data from the Mapper, before passing it to Reducer. It runs after the mapper and before the Reducer and its use is optional.

Read: Key-value Pairs in MapReduce

3. How does MapReduce Combiner work?

Let us now see the working of the Hadoop combiner in MapReduce and how things change when combiner is used as compared to when combiner is not used in MapReduce?

3.1. MapReduce program without Combiner

MapReduce Combiner : MapReduce program without combiner

In the above diagram, no combiner is used. Input is split into two mappers and 9 keys are generated from the mappers. Now we have (9 **key/value**) intermediate data, the further mapper will send directly this data to reducer and while sending data to the reducer, it consumes some network bandwidth (bandwidth means time taken to transfer data between 2 machines). It will take more time to transfer data to reducer if the size of data is big.

Now in between mapper and reducer if we use a hadoop combiner, then combiner shuffles intermediate data (9 key/value) before sending it to the reducer and generates 4 key/value pair as an output.

Read: Data Locality in MapReduce

3.2. MapReduce program with Combiner in between Mapper and Reducer

MapReduce Combiner: MapReduce program with combiner

Reducer now needs to process only 4 key/value pair data which is generated from 2 combiners. Thus reducer gets executed only 4 times to produce final output, which increases the overall performance.

4. Advantages of MapReduce Combiner

As we have discussed what is Hadoop MapReduce Combiner in detail, now we will discuss some advantages of Mapreduce Combiner.

- Hadoop Combiner reduces the time taken for data transfer between mapper and reducer.
- It decreases the amount of data that needed to be processed by the reducer.
- The Combiner improves the overall performance of the reducer.

Read: Counters in MapReduce

5. Disadvantages of Hadoop combiner in MapReduce

There are also some disadvantages of hadoop Combiner. Let's discuss them one by one-

- MapReduce jobs cannot depend on the Hadoop combiner execution because there is no guarantee in its execution.
- In the local filesystem, the key-value pairs are stored in the Hadoop and run the combiner later which will cause expensive disk IO.

6. Hadoop Combiner – Conclusion

In conclusion, we can say that MapReduce Combiner plays a key role in reducing network congestion. MapReduce combiner improves the overall performance of the reducer by summarizing the output of Mapper.

See Also-

- **Shuffling and Sorting in Hadoop MapReduce**
- **Partitioner in Hadoop MapReduce**

I hope this post has helped you to understand the role of Combiner in Hadoop. If you have any query related to Hadoop Combiner, so, please drop me a comment below.

Leave a comment

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

Post Comment

3 thoughts on “Hadoop Combiner – Best Explanation to MapReduce Combiner”

Vijay kumar

Reply ↓

December 6, 2017 at 12:43 pm

How to configure combiner in hadoop ?

Nimish Jindal

Reply ↓

January 25, 2018 at 6:35 pm

Hi, can you help me with following questions:
1. Output of mapper is stored on local machine. Is Combiner also situated on the local machine? (If not then isn’t it that we are transferring same(or even more) amount of data in network because first we are transferring mapper output to Combiner and then to Partitioner/Reducer)?
2. Why can’t we just combine on keys in the mapper code itself? (As i see, it takes same time complexity to compute and creates an output file with smaller size because keys are combined?)

Manoj Kumar

March 7, 2018 at 4:55 pm

Reply ↓

Dear Author,

I've question as well as doubt on image you used in point 3.2 in this topic. I want to share an updated image on this for discussion and clarity of my confusion. Please let me know how can I share that with you OR send in a mail to me where I can reply and forward my question to you along with the image.

Thanks & Regards
Manoj

Post navigation

← Limitations of Apache Spark – Ways to Overcome Spark Drawbacks

Hadoop InputFormat, Types of InputFormat in MapReduce →

- Cassandra Tutorials
- Cloud Computing Tutorials
- Data Mining Interview Questions
- Data Mining Tutorials
- Data Science Interview Questions
- Data Science Tutorials
- Entrepreneurship
- Flume Interview Questions
- Flume Tutorials
- Hadoop Interview Questions
- Hadoop Quiz
- Hadoop Tutorials
- HBase Interview Questions
- HBase Tutorials
- HCatalog Interview Questions
- HCatalog Tutorials
- HDFS Tutorials
- Hive Interview Questions
- Hive Quiz
- Hive Tutorials
- Impala Interview Questions
- Impala Quiz
- Impala Tutorials
- Interview Questions
- IOT Interview Questions
- IOT Tutorials
- Java Interview Questions
- Java Tutorials
- Jobs
- Kafka Interview Questions
- Kafka Quiz
- Linux Tutorials
- Machine Learning Tutorials
- MapReduce Tutorials
- MongoDB Tutorials
- Pig Interview Questions

- Pig Quiz
- Pig Tutorials
- Power BI Interview Questions
- Power BI Tutorials
- Pyspark Interview Questions
- PySpark Tutorials
- Python Interview Questions
- Python Tutorials
- QlikView Tutorials
- Quiz
- R Interview Questions
- R Quiz
- R Tutorials
- Salesforce Interview Questions
- Salesforce Tutorials
- SAS – STAT Interview Questions
- SAS – STAT Tutorials
- SAS Interview Questions
- SAS Quiz
- SAS Tutorials
- Scala Interview Questions
- Scala Tutorials
- Spark Interview Questions
- Spark Quiz
- Spring Tutorials
- SQL Interview Questions
- SQL Tutorials
- Sqoop Interview Questions
- Sqoop Tutorials
- Storm Tutorials
- Tableau Interview Questions
- Tableau Tutorials
- TensorFlow Interview Questions
- Tensorflow Tutorials
- Training
- YARN Tutorials

› ZooKeeper Interview Questions
› Zookeeper Tutorials