edureka!

Hadoop Interview Questions -MapReduce

April 23, 2013 | Big Data and Hadoop, Interview Questions

Looking out for Hadoop Interview Questions that are frequently asked by employers?
Here is the third list of Hadoop Interview Questions which covers MapReduce…

What is MapReduce?

It is a framework or a programming model that is used for processing large data sets over clusters of computers using distributed programming.

What are 'maps' and 'reduces'?

'*Maps*' and '*Reduces*' are two phases of solving a query in HDFS. 'Map' is responsible to read data from input location, and based on the input type, it will generate a *key value pair,* that is, an intermediate output in local machine. 'Reducer' is responsible to process the intermediate output received from the mapper and generate the final output.

What are the four basic parameters of a mapper?

The four basic parameters of a mapper are *LongWritable, text, text and IntWritable*. The first two represent input parameters and the second two represent intermediate output parameters.

What are the four basic parameters of a reducer?

The four basic parameters of a reducer are *text, IntWritable, text, IntWritable*. The first two represent intermediate output parameters and the second two represent final output parameters.

What do the master class and the output class do?

Master is defined to update the Master or the job tracker and the output class is defined to write data onto the output location.

What is the input type/format in MapReduce by default?

By default the type input type in MapReduce is 'text'.

Is it mandatory to set input and output type/format in MapReduce?

No, it is not mandatory to set the input and output type/format in MapReduce. By default, the cluster takes the input and the output type as 'text'.

What does the text input format do?

In text input format, each line will create a line object, that is an hexa-decimal number. Key is considered as a line object and value is considered as a whole line text. This is how the data gets processed by a mapper. The mapper will receive the 'key' as a '*LongWritable*' parameter and value as a '*text*' parameter.

What does job conf class do?

MapReduce needs to logically separate different jobs running on the same cluster. '*Job conf class*' helps to do job level settings such as declaring a job in real environment. It is recommended that Job name should be descriptive and represent the type of job that is being executed.

What does conf.setMapper Class do?

Conf.setMapper class sets the mapper class and all the stuff related to map job such as reading a data and generating a *key-value pair* out of the mapper.

What do sorting and shuffling do?

Sorting and shuffling are responsible for creating a unique key and a list of values. Making similar keys at one location is known as *Sorting*. And the process by which the intermediate output of the mapper is sorted and sent across to the reducers is known as *Shuffling*.

What does a split do?

Before transferring the data from hard disk location to map method, there is a phase or method called the '*Split Method*'. Split method pulls a block of data from HDFS to the framework. The *Split class* does not write anything, but reads data from the block and pass it to the mapper. Be default, Split is taken care by the framework. Split method is equal to the block size and is used to divide block into bunch of splits.

How can we change the split size if our commodity hardware has less storage space?

If our commodity hardware has less storage space, we can change the split size by writing the '*custom splitter*'. There is a feature of customization in Hadoop which can be called from the main method.

What does a MapReduce partitioner do?

A MapReduce partitioner makes sure that all the value of a single key goes to the same reducer, thus allows evenly distribution of the map output over the reducers. It redirects the mapper output to the reducer by determining which reducer is responsible for a particular key.

How is Hadoop different from other data processing tools?

In Hadoop, based upon your requirements, you can increase or decrease the number of mappers without bothering about the volume of data to be processed. this is the beauty of parallel processing in contrast to the other data processing tools available.

Can we rename the output file?

Yes we can rename the output file by implementing *multiple format output class*.

Why we cannot do aggregation (addition) in a mapper? Why we require reducer for that?

We cannot do aggregation (addition) in a mapper because, sorting is not done in a mapper. Sorting happens only on the reducer side. Mapper method initialization depends upon each input split. While doing aggregation, we will lose the value of the previous instance. For each row, a new mapper will get initialized. For each row, input split again gets divided into mapper, thus we do not have a track of the previous row value.

What is Streaming?

Streaming is a feature with Hadoop framework that allows us to do programming using MapReduce in any programming language which can accept standard input and can produce standard output. It could be Perl, Python, Ruby and not necessarily be Java. However, customization in MapReduce can only be done using Java and not any other programming language.

What is a Combiner?

A 'Combiner' is a mini reducer that performs the local reduce task. It receives the input from the mapper on a particular node and sends the output to the reducer. Combiners help in enhancing the efficiency of MapReduce by reducing the quantum of data that is required to be sent to the reducers.

What is the difference between an HDFS Block and Input Split?

HDFS Block is the physical division of the data and *Input Split* is the logical division of the data.

What happens in a textinputformat?

In *textinputformat*, each line in the text file is a record. *Key* is the byte offset of the line and *value* is the content of the line. For instance, **Key: longWritable, value: text**.

What do you know about keyvaluetextinputformat?

In *keyvaluetextinputformat*, each line in the text file is a '*record*'. The first separator character divides each line. Everything before the separator is the *key* and everything after the separator is the *value*. For instance, **Key: text, value: text.**

What do you know about Sequencefileinputformat?

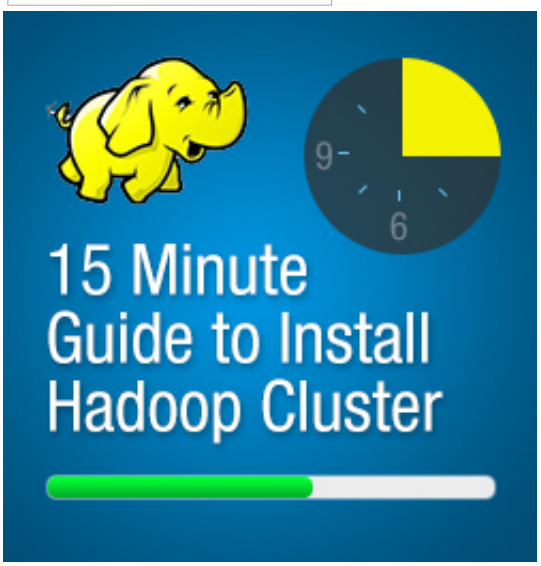
Sequencefileinputformat is an input format for reading in sequence files. *Key* and *value* are user defined. It is a specific compressed binary file format which is optimized for passing the data between the output of one MapReduce job to the input of some other MapReduce job.

What do you know about Nlineoutputformat?

Nlineoutputformat splits 'n' lines of input as one split.

To refer to the first list, click Hadoop Interview Questions – HDFS!
To refer to the second list, click Hadoop Interview Questions – Setting up Hadoop Cluster!
To refer to the fourth list, click Hadoop Interview Questions - PIG!

- **Search**



Recent Posts

- Free Webinar on 'Introduction to MongoDB'
- Importance of Data Science With Cassandra
- How to become a Hadoop Administrator?
- Why Big Data Professionals Need to Learn MongoDB?
- Why Should you go for Hadoop Administration Course?

Categories

- Android
- Apache Cassandra
- Apache Storm
- Big Data and Hadoop
- Business Analytics With R
- Cloud Computing
- Data Science
- Hadoop Administration
- Interview Questions
- Java
- MongoDB
- PMI-ACP
- PMP Exam Preparation
- Python for Big Data Analytics
- Resources & Misc.
- Uncategorized
- Webinars