Tech Module Exercises: 0 / 39

Big data Hadoop Map Reduce

* Mini Projects

|  |  |  |
| --- | --- | --- |
| S.No. | Mini-project Description | Topics Covered |
| 1 | Mapreduce Mini project | MapReduce |

* + In this TECH Module, you are expected to complete the below Mini-Project(s)
* Topics to Learn

To complete the above project, you will need to lean the below technical topics-

* + **BigData**

Learning Material for **BigData**

Below is the learning material that you are expected to read along with completion of the hands-on assignments. The material is mentioned is the order in which it should be read.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Material Title | Material Location | Type of Material | Classification |
| 1. | Big Data Overview | <https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm> | Web | Mandatory |
| 2. | Big Data Overview | <https://www.guru99.com/what-is-big-data.html> | Web | Mandatory |
| 3. | Big Data Overview | <https://www.youtube.com/watch?time_continue=33&v=CZOSvw3hx40&feature=emb_logo> | Video | Suggestive |

* + - Hands-on Assignments for **BigData**
      * No Hands-on Assignments for this topic
  + **Hadoop**

Learning Material for **Hadoop**

Below is the learning material that you are expected to read along with completion of the hands-on assignments. The material is mentioned is the order in which it should be read.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Material Title | Material Location | Type of Material | Classification |
| 1. | Hadoop – Big Data Solutions | <https://www.tutorialspoint.com/hadoop/hadoop_big_data_solutions.htm> | Web | Mandatory |
| 2. | Hadoop – Big Data Solutions | <https://www.youtube.com/watch?time_continue=2&v=29O3CCYOzic&feature=emb_logo> | Video | Suggestive |
| 3. | Hadoop – Introduction | <https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm> | Web | Mandatory |
| 4. | Hadoop – Introduction | <https://www.youtube.com/watch?v=OoEpfb6yga8> | Video | Suggestive |
| 5. | Environment Setup | <https://www.tutorialspoint.com/hadoop/hadoop_enviornment_setup.htm> | Web | Mandatory |
| 6. | Environment Setup | <https://www.youtube.com/watch?v=HP4g2BU7-xU> | Video | Suggestive |
| 7. | Hadoop – HDFS Overview | <https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm> | Web | Mandatory |
| 8. | Hadoop – HDFS Overview | <https://www.youtube.com/watch?v=GJYEsEEfjvk> | Video | Suggestive |
| 9. | Hadoop – HDFS Operations | <https://www.tutorialspoint.com/hadoop/hadoop_hdfs_operations.htm> | Web | Mandatory |
| 10. | Hadoop – HDFS Operations | <https://www.youtube.com/watch?v=ZYQhHrq5kj8> | Video | Suggestive |
| 11. | Hadoop – Command Reference | <https://www.tutorialspoint.com/hadoop/hadoop_command_reference.htm> | Web | Mandatory |
| 12. | Hadoop – Command Reference | <https://www.youtube.com/watch?v=9bzCF6bpR_o> | Video | Suggestive |

* + - Hands-on Assignments for **Hadoop**

Complete the below hands-on assignments before proceeding with the next Topic

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Hands-on Assignment | Topic Covered | Status |
| 1. | Open a terminal window to the current working directory.  Eg : /home/training | HDFS |  |
| 2. | Print the Hadoop version | HDFS |  |
| 3. | List the contents of the root directory in HDFS | HDFS |  |
| 4. | Report the amount of space used and available space on currently mounted filesystem | HDFS |  |
| 5. | Count the number of directories,files and bytes under the paths that match the specified file pattern | HDFS |  |
| 6. | Run a DFS filesystem checking utility | HDFS |  |
| 7. | Run a cluster balancing utility | HDFS |  |
| 8. | Create a new directory named “hadoop” below the /user/training directory in HDFS.  Since you’re currently logged in with the “training” user ID, /user/training would be your home directory in HDFS. | HDFS |  |
| 9. | Add a sample text file from the local directory named “data” to the new directory you created in HDFS during the previous step. | HDFS |  |
| 10. | List the contents of this new directory in HDFS. | HDFS |  |
| 11. | Add the entire local directory called “retail” to the /user/training directory in HDFS. | HDFS |  |
| 12. | Since /user/training is your home directory in HDFS, any command that does not have an absolute path is  interpreted as relative to that directory. The next command will therefore list your home directory, and  should show the items you’ve just added there. | HDFS |  |
| 13. | See how much space this directory occupies in HDFS | HDFS |  |
| 14. | Delete a file ‘customers’ from the “retail” directory. | HDFS |  |
| 15. | Ensure this file is no longer in HDFS. | HDFS |  |
| 16. | Delete all files from the “retail” directory using a wildcard. | HDFS |  |
| 17. | Empty the trash | HDFS |  |
| 18. | Finally, remove the entire retail directory and all of its contents in HDFS. | HDFS |  |
| 19. | List the hadoop directory again | HDFS |  |
| 20. | Add the ‘purchases.txt’ file from the local directory named “/home/training/” to the hadoop directory you created in HDFS | HDFS |  |
| 21. | View the contents of your text file ‘purchases.txt’ which is present in your hadoop directory. | HDFS |  |
| 22. | Add the ‘purchases.txt’ file from “hadoop” directory which is present in HDFS directory to the directory “data” which is present in your local directory | HDFS |  |
| 23. | cp is used to copy files between directories present in HDFS,create a folder as purchase2020 and copy the file purchases.txt from hadoop to purchase2020 | HDFS |  |
| 24. | ‘-get’ command can be used alternaively to ‘-copyToLocal’ command use this command to get back the file from purchase2020 to local folder | HDFS |  |
| 25. | Display last kilobyte of the file “purchases.txt” to stdout. | HDFS |  |
| 26. | Default file permissions are 666 in HDFS  Use ‘-chmod’ command to change permissions of a file | HDFS |  |
| 27. | Default names of owner and group are training and training  Use ‘-chown’ to change owner name and group name simultaneously | HDFS |  |
| 28. | Default name of group is training.  Use ‘-chgrp’ command to change group name | HDFS |  |
| 29. | Move a directory from one location to other hadoop fs -mv hadoop apache\_hadoop | HDFS |  |
| 30. | Default replication factor to a file is 3.  Use ‘-setrep’ command to change replication factor of a file | HDFS |  |
| 31. | Copy a directory from one node in the cluster to another.  Use ‘-distcp’ command to copy,  -overwrite option to overwrite in an existing files  -update command to synchronize both directories | HDFS |  |
| 32. | Write the command to make the Namenode leave safe mode | HDFS |  |
| 33. | List all the hadoop file system shell commands | HDFS |  |

* + **MapReduce**

Learning Material for **MapReduce**

Below is the learning material that you are expected to read along with completion of the hands-on assignments. The material is mentioned is the order in which it should be read.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Material Title | Material Location | Type of Material | Classification |
| 1. | MapReduce – Introduction | <https://www.tutorialspoint.com//map_reduce/map_reduce_introduction.htm> | Video | Mandatory |
| 2. | MapReduce – Algorithm | <https://www.tutorialspoint.com/map_reduce/map_reduce_algorithm.htm> | Web | Mandatory |
| 3. | MapReduce – Installation | <https://www.tutorialspoint.com/map_reduce/map_reduce_installation.htm> | Web | Mandatory |
| 4. | MapReduce – Algorithm | <https://www.youtube.com/watch?v=30RaNpaupj0> | Video | Suggestive |
| 5. | MapReduce – API | <https://www.tutorialspoint.com/map_reduce/map_reduce_api.htm> | Web | Mandatory |
| 6. | MapReduce – Hadoop Implementation | <https://www.tutorialspoint.com/map_reduce/implementation_in_hadoop.htm> | Web | Mandatory |
| 7. | MapReduce – Hadoop Implementation in PYTHON | <https://www.youtube.com/watch?v=QNB1SZm2jS4&list=PLToV_TnxzyCYhZakoxYzHyce249ofL4GP&index=2> | Video | Suggestive |
| 8. | MapReduce – Partitioner | <https://www.tutorialspoint.com/map_reduce/map_reduce_partitioner.htm> | Web | Mandatory |
| 9. | MapReduce - Partitioner | <https://www.youtube.com/watch?v=XLylVu-p7HY> | Video | Suggestive |
| 10. | MapReduce - Combiners | <https://www.tutorialspoint.com/map_reduce/map_reduce_combiners.htm> | Web | Mandatory |
| 11. | MapReduce - Combiners | <https://www.youtube.com/watch?v=7bOw9N34cJM> | Video | Suggestive |

* + - Hands-on Assignments for **MapReduce**

Complete the below hands-on assignments before proceeding with the next Topic

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Hands-on Assignment | Topic Covered | Status |
| 1. | Create an Input File with the following columns:  empid, age, salary,department,Job  1001,25,23000,IT,Clerk  1004,27,27000,HR ,Manager  1002,23,21000,IT ,Assistant 1003,28,33000,IT , Clerk 1006,31,29000,Sales ,Manager 1007,41,38000,HR ,Clerk  1009,27,21000,Sales ,Manager | MapReduce |  |
| 2. | Write a Map Reduce Job to compute Salary Increment using the following formula:  if department = IT, increment = 10% of  salary  if department = HR, increment = 12% of  salary  else increment = 5% of salary | MapReduce |  |
| 3. | Write a Map Reduce Job to compute average salary for each department | MapReduce |  |
| 4. | Write a Map Reduce Job to compute the count of employees in each department | MapReduce |  |
| 5. | Write a Map Reduce Job to compute the sum of salaries for each job | MapReduce |  |
| 6. | Write a Map Reduce Job to compute the average of salaries for each job | MapReduce |  |

* Learning Outcomes

Upon completion of this TECH Module, you should be able to:

* + Hadoop – What you have learnt
  + What is Hadoop
  + HDFS Overview
  + HDFS Operations
  + Hadoop MapReduce
  + Hadoop Streaming
  + MapReduce – What you have learnt
  + What is MapReduce
  + MapReduce API
  + Input Format
  + Map Task
  + Reduce Task
  + Combiner
  + Partitioner
  + Output Format
* Sample Questions

After gaining knowledge of the above module, below are the possible interview questions that you should be able to confidently answer.

|  |  |
| --- | --- |
| No. | Questions |
| 1 | Hadoop Questions |
| 2 | What is Hadoop ? |
| 3 | What are the main components of Hadoop ? |
| 4 | What command is sed to copy a file from local file system to HDFS ? |
| 5 | What command is sed to copy a file from HDFS to local file system ? |
| 6 | What command is used to list files in HDFS |
| 7 | Namenode metadata is stored is which file ? |
| 8 | What is the default block size in HDFS ? |
| 9 | What is pseudo distribution cluster ? |
| 10 | What is Hadoop Streaming ? |
| 11 | MapReduce Questions |
| 12 | What is MapReduce |
| 13 | What are the different phases in a MapReduce Job ? |
| 14. | What computations are done in Map Task ? |
| 15. | What computations are done in Reduce Task ? |
| 16. | What is a Partitioner ? |
| 17. | What is a Combiner ? |
| 18. | What is Shffle & Sort process ? |