A close up of a sign

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DATA SCIENCE ACADEMY BOOTCAMP Final exam | | | | | |
| **Name, Surname:** Shahin Mammadov |  |  | | **Date**: 11/20/2022 | | |
| **Exam:** Big Data |  |  | |  | | |
| **Instructions:** You will be given ***2 hours*** to complete examination. There are 2 parts and 4 types of questions including: True/False, Multiple choice, Fill in the blanks and Practical.  **First part** will last ***30 minutes****,* and you must write answers directly on word document. Internet search is **not** permitted.  **Second part** is Practical question and will last ***90 minutes***. It is your responsibility to submit your exam answers electronically to [homework@dsa.az](mailto:homework@dsa.az) with **“Big data Final Exam”** subject. Before you leave exam, you **MUST** contact mentor to confirm that you have submitted your examination.Good Luck! | | | | | | |
| PART 1: TRUE/FALSE | | | | | | |
| **Directions: Read each statement below carefully. Place a** +  **mark on the “True” cell if you think a statement is TRUE, and on the “False” cell if you think the statement is FALSE.**  **Each question is 3 point.** | | | | | | |
|  | | | | | | |
|  | | | True | | False | |
| 1. Traditional computing approach is efficient for huge volume of data / Big Data. (3 points) | | |  | | X | |
| 2. HDFS blocks fit well with replication for providing fault tolerance and availability. (3 points) | | | X | |  | |
| 3. Hadoop MapReduce approach is much faster than Spark’s approach. (3 points) | | |  | | X | |
| 4. RDDs are immutable. (3 points) | | | X | |  | |
| 5. Cloud Computing is an efficient approach to Big Data. (3 points) | | | X | |  | |

|  |
| --- |
| PART 1: MULTIPLE CHOICE |
| **Directions: Read each question carefully, and then place a + next to the answer that best fits the question.** |
|  |
| 1. Which one is not unstructured data: (5 points)  a. Videos  b. csv files  c. Photos  d. Audio files |
|  |

www.dsa.az Bütün hüquqlar qorunur DATA SCIENCE ACADEMY

|  |
| --- |
| PART 1: MULTIPLE CHOICE |
| **Directions: Read each question carefully, and then place a + next to the answer that best fits the question** |
|  |
| 2. … in NoSQL databases. (5 points)  a. You can’t add a record which doesn’t fit the schema.  b. you need to add NULLs to unused items in a row  c. there is no Schema  d. you can’t add multiple items in a field (you should create another table – primary key, foreign key…)    3. Choose true statement about RDDs. (5 points)  a. Nothing happens about RDDs until action is called  b. RDDs have a schema  c. RDDs are completely same with Pandas DataFrame  d. RDDs are core component of YARN  4. … is open-source data warehouse system built on top of Hadoop for querying and analyzing large datasets.  (5 points)  a. Apache Hive  b. Apache Ambari  c. Apache Sqoop  d. Apache Flume |
|  |

www.dsa.az Bütün hüquqlar qorunur DATA SCIENCE ACADEMY

|  |
| --- |
| PART 1: FILL IN BLANK |
| **Directions: Read each question carefully and provide the answer on the line. Your answer must be in complete sentences.** |
|  |
| 1. Saving all these ever-growing data in a single device is impossible, so Hadoop found solution in saving   it in a \_\_\_\_\_ manner. (5 points )  distributed(multiple files in huge sizes in different clusters)   1. In HDFS, \_\_\_\_keeps track of where Data blocks are stored for a particular file (5 points)   master machine   1. \_\_\_\_ is Hadoop’s cluster resource management system (5 points) |
| YARN   1. What is described in this Spark Process? (5 points)   Pipeline   1. \_\_\_\_ is Web-based management tool that manages, monitors, and provisions the health of Hadoop clusters.   Apache Ambari (5 points) |

www.dsa.az Bütün hüquqlar qorunur DATA SCIENCE ACADEMY