**COMP529 January 2020 Bakhtiar Amen (with 36 (E) easier marks)**

|  |  |
| --- | --- |
| **Exam Questions** | **Marks** |
| 1. **(a)** Draw a Hadoop Distributed File System (HDFS) architecture for 6 computer nodes. 2. Label them as DataNodes 1 NameNode, 1 Secondary NameNode, and 4 DataNodes.   1 NameNode  1 Secondary NameNode  4 DataNodes  NameNode  DataNode3  **Block D**  **Block E**  DataNode1  **Block A**  **Block B**  **Block C**  DataNode2  **Block A**  **Block B**  **Block C**  **Block D**  DataNode4  **Block A**  **Block B**  **Block C**  **Block E**  **Block A**, **Block B, Block C**  **Block D**, **Block E**  Secondary NameNode   1. Show how you would allocate File X when replication number is equal to 3 blocks (Block A, Block B, Block C ).   DataNode 1 (Block A, Block B, Block C)  DataNode 2 (Block A, Block B, Block C)  DataNode 4 (Block A, Block B, Block C)   1. Show how you would allocate File Y when replication number is equal to 2 blocks {(Block D, Block E)}.   DataNode 2 (Block D)  DataNode 3 (Block D, Block E)  DataNode 4 (Block E)   1. Briefly describe each HDFS component: NameNode, DataNode and Secondary Namenode   Name node: keep the file metadata  Datanode: data spaces  Secondary NameNode: Backup the Namenode   1. What is a default size of the HDFS block?   128 MB   1. What is a default replication number in HDFS?   3 Replication  **(b)**   1. What are the names of both Big Data processing models?   Batch Data Model  Stream Data Model   1. Name the three Big Data challenging tasks that could not be handled by a single machine   Capture data  Manage data  Process data   1. Name the 4 Vs of Big Data and briefly state what do they mean?   Volume: large scale of dataset  Velocity: speed of data  Veracity: data quality, accuracy  Variety: different data type   1. Hadoop has been designed to address which V’s of Big Data problem?   Volume   1. What are the two functions of MapReduce programming model?   Mapper  Reducer   1. What is the name of MapReduce algorithm to show an output for (k, v) = (empName,maxSalary)?   Searching Algorithm   1. MapReduce has two main components in Hadoop cluster, what are they?   Job Tracker  Task Tracker   1. Which feature of Hadoop makes it necessary to use a portable programming language such as Java?   The code is sent to the data  **(c)** Draw a diagram for fully distributed storm cluster of five computer nodes with one coordi-  nator node  Nimbus  **Supervisor 1**  **W1. W2**  **Supervisor 3**  **W1. W2**  **Supervisor 2**  **W1. W2**  Zookeeper   1. Allocate all daemons across of each computer node.   Nimbus- Node 1  Zookeeper- Node1  Supervisor- 1  Supervisor- 2  Supervisor- 3   1. Allocate 2 workers per each node   w1  w2   1. Clearly show the state of connectivity between each node   An arrow connected to each node   1. Name each grouping task in Storm’s topology to handle large scale of data   streams.  All  Shuffle  Field  Global   1. A topology comprises two spouts and three bolts. Assume one spout generates a   stream of images and the other spout generates a stream of 30 millisecond audio  chunks. Assume one bolt performs lip-reading, one performs speech recognition  and the third bolt aligns two streams of text. Draw a diagram describing the topology.Label all spouts and bolts. Annotate all streams with the information being transmitted.    Credible attempt  Spouts correct  Bolts correct  Arcs correct  Streams correctly labelled   1. Describe the role of each spout and bolt in Storm’s topology.   Convert Stream  Process Stream   1. **(a)** Assume that there are 100 students in your class, 35 of those students are studying Information Technology (IT), 45 studying Mathematics (M) and 20 studying both subjects. Find the following events:   15  IT = 35  25  M = 45   1. The probability of each subject.   P (IT) = 15/100 = 0.15  P (M) = 25/100 = 0.25   1. The probability that the student studies both subjects.   P (IT, M) = 20/100 = 0.20   1. The probability of student picked at random studies IT given that we know he studies   Mathematics.  P(IT|M) =  P(IT∩M) / P(M)  P(IT|M) = 0.20 / 0.25  P(IT|M) = 0.8  **(b)** You are working in a construction company and your boss did ask you to analyse some of their data which are related to the cause of their system crash. You have found out that the cause of crash was due to three probabilities (e.g., Malfunction, Network, Operating System).   1. Draw a Direct Acyclic Graph (DAG) Bayesian Network and label each probability node as: Malfunction Failure, MF, Network Failure, NF, and Operating System, OS   Credible attempt  Directed Acyclic Graph   1. Consider a problem with three random variables: MF, NF, and OS. While MF and NF are both dependent upon OS.   MF connected to OS  NF connected to OS  OS dependent on MF and NF   1. Draw OS node as observed problem node in the DAG diagram.   Credible attempt  MF connected to OS  NF connected to OS  OS is an observed node  **(c)**   1. Draw a Hidden-Markov Model for sequences of unobserved nodes *X*1:4, and then using the learned parameters to assign a sequence of *Y*1:4 observed nodes to analyse speech data.     Credible attempt  All nodes labelled  4 observed nodes (y1,…y4)  4 unobserved nodes (x1,…x4)   1. Think of that you have two large files, file A has 100 Topics and the other 50 Topics are in file B. Draw a Bayesian Network big graphs to describe how you would count topics in each file and use plates to observe a belief of frequencies   i:1….50  i:1….100  Credible attempt  Observed nodes correct  Two unobserved nodes shown for belief of frequencies  Plates correct  Arcs correct  Clear which nodes relate to file A and which to file B   1. Name two propagation algorithms to solve large graphs and complex problems in   big data analysis.  Pearl’s Algorithm  Belief Propagation   1. To perform inference in a very much larger version of this graph involving many contributory factors relating to the risk of a car crash, it is proposed to use Gibbs sampling, Belief Propagation or Mean Field. What would be the relative advantages of each technique in terms of their ability to be parallelised, the number of iterations required and any restrictions on the graph necessary to use the techniques? A tabular answer is acceptable.  |  |  |  |  | | --- | --- | --- | --- | |  | Ability to be parallelised | Number of iterations | Restrictions on graph | | Gibbs | Y | Large | None | | Belief Propagation | N | 2 (accept small) | Must be a tree | | Mean Field | Y | Small | None |   One mark per correct entry   1. A security company is interested in monitoring four sensor devices. Your task is to draw a topology to describes how each sensor device is generating a data. Show how Kalman filters processing each sensor device’s data and alerts being generated when two or more sensor exhibit unusual behaviour at the same time.   Sensor 1  Sensor 2  Sensor 3  Sensor 4  KF 1  KF 2  KF 3  KF 4  alerted  Credit Attempt  4 Spouts correctly identified to be associated with the turbines  4 Kalman filters shown  One alerted shown  Connectivity corrects   1. Write an equation describing the likelihood model used by a Kalman filter when processing M-dimensional data to make inferences about an N-dimensional state. Define the size of any matrices used in the models in terms of M and N   Equations are linear  Equation is correct | **4**  **E1**  **E1**  **2**  **3**  **1**  **1**  **1**  **2**  **1**  **1**  **3**  **E1**  **E1**  **E1**  **1**  **E1**  **1**  **E1**  **2**  **E1**  **E1**  **3**  **E1**  **E1**  **E1**  **4**  **E1**  **E1**  **E1**  **E1**  **1**  **E1**  **2**  **1**  **1**  **1**  **1**  **2**  **E1**  **E1**  **1**  **1**  **5**  **1**  **1**  **1**  **1**  **1**  **2**  **E1**  **E1**  **2**  **E2**  **4**  **E1**  **E1**  **E1**  **E1**  **5**  **E1**  **1**  **1**  **1**  **1**  **2**  **1**  **1**  **2**  **1**  **1**  **2**  **1**  **1**  **2**  **1**  **1**  **2**  **E1**  **1**  **3**  **E1**  **1**  **1**  **1**  **4**  **1**  **1**  **1**  **1**  **9**  **E1**  **3**  **2**  **2**  **6**  **E1**  **2**  **1**  **1**  **1**  **2**  **E1**  **E1**  **9**  **1**  **6**  **E1**  **2**  **1**  **1**  **1**  **3**  **E1**  **1**  **1** |