PAPER CODE NO.	EXAMINER:	Dr. Bakhtiar Amen	Tel. No. 58645
COMP529	DEPARTMENT:	Computer Science	

UNIVERSITY OF LIVERPOOL

FIRST SEMESTER EXAMINATIONS 2019/20

BIG DATA ANALYTICS

TIME ALLOWED : TWO Hours

INSTRUCTIONS TO CANDIDATES

All candidates should answer **ALL** two questions.

The numbers in the right hand margin represent mark for the question answer. The total available marks are 100.

UNIVERSITY OF LIVERPOOL

1. (a) Draw a Hadoop Distributed File System (HDFS) architecture for 6 computer nodes.

(i) 1 NameNode, 1 Secondary NameNode, and 4 DataNodes. **4**

(ii) Show how you would allocate File X when replication number is equal to 3 blocks (Block A, Block B, Block C). **3**

(iii) Show how you would allocate File Y when replication number is equal to 2 blocks (Block D, Block E). **2**

(iv) Briefly describe each HDFS component: NameNode, DataNode and Secondary Namenode. **3**

(v) What is a default size of the HDFS block? **1**

(vi) What is a default replication number in HDFS? **1**

(b) (i) What are the names of both Big Data processing models? **2**

(ii) Name those three Big Data challenging tasks that could not be handled by a single machine. **3**

(iii) Name the 4 Vs of Big Data and briefly state what do they mean? **4**

(iv) Hadoop has been designed to address which V's of Big Data problem? **1**

(v) What are the two functions of MapReduce programming model? **2**

(vi) What is a name of MapReduce algorithm to show an output for (k, v) = (empName, maxSalary)? **1**

(vii) MapReduce has two main components in Hadoop cluster, what are they? **2**

(viii) Which feature of Hadoop makes it necessary to use a portable programming language such as Java? **1**

(c) Draw a diagram for fully distributed Storm cluster of five computer nodes with one coordinator node.

(i) Allocate all daemons across of each computer node. **5**

(ii) Allocate 2 workers per each computer node. **2**

Question 1 continues overleaf.

(iii) Show the state of connectivity between each node. **2**

(iv) Name each grouping task in Storm's topology for handling a large scale of data streams. **4**

(v) 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. **5**

(vi) Describe the role of each spout and bolt in Storm's topology. **2**

2. **(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:

(i) The probability of each subject. **2**

(ii) The probability that the student studies both subjects. **2**

(iii) The probability of student picked at random studies IT given that we know he studies Mathematics. **2**

(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).

(i) Draw a Direct Acyclic Graph (DAG) Bayesian Network and label each probability node as: Malfunction Failure (MF), Network Failure (NF), and Operating System (OS). **2**

(ii) Consider the problem with three random variables: MF, NF, and OS. While MF and NF are both dependent upon OS. **3**

(iii) Draw OS node as observed problem node in the DAG diagram. **4**

(c) (i) Draw a Hidden-Markov Model for sequences of unobserved nodes ($X_{1:4}$), and then use the learned parameters to assign a sequence of observed nodes ($Y_{1:4}$) to analyse speech data. **9**

(ii) 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. **6**

(iii) Name two algorithms to solve large complex graphs in Big Data analytic. **2**

(iv) 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. **9**

Question 2 continues overleaf.

(v) 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. **6**

(vi) 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. **3**