Cloud Computing and Big Data Analytics- Problem Sheet
ECE, NCTU
Date: 2018.4.26 **Time:** 9:00 – 10:30
Closed-Book Exam

1. General Concepts of Cloud Computing. 25%

1.a) Please explain why virtual machines play such an important role in cloud computing? (7%)

1.b) Please write a pros-and-cons table for the following load balancing techniques, i.e., Round Robin, Low Latency, Least Connections, Priority, and Overflow. (10%)
(Hints:
Round Robin: The servers are selected one by one to serve the incoming requests in a non-hierarchical circular fashion with no priority assigned to a specific server.
Low Latency: Each incoming request is routed to the server which has the lowest latency.
Least Connections: The incoming requests are routed to the server with the least number of connections.
Priority: Each server is assigned a priority. The incoming traffic is routed to the highest priority as long as the server is available. When the highest priority server fails, the incoming request is routed to a server with a lower priority.
Overflow: When the incoming requests to highest priority server overflows, the requests are routed to a lower priority server.)

1.c) There are always pros and cons, what are the drawbacks of using cloud computing? Please list 2 drawbacks? (8%).

2. MapReduce. 15%

2.a) Think about the sales dataset we used in Lab 1 for Hadoop (data attribute: date, time, store, item, cost, payment). If we want to find the average and maximum amount of sales per store on Nov., 11, 2017 (single day), can we calculate the results directly in a mapper-reducer manner? If yes, please provide the key-value pair and what to do in the reducer. If no, e.g., the multiple-stage of MapReduce is required, please explain why. (9%)

2.b) Hadoop scheduler is a pluggable component that makes it open to support different scheduling algorithms, e.g., FIFO, Fair, Capacity. The pluggable scheduler framework provides the flexibility to support a variety of workloads with varying priority and performance constraints. Please list three factors that should be considered in schedulers. (6%)

3. Spark. 21%

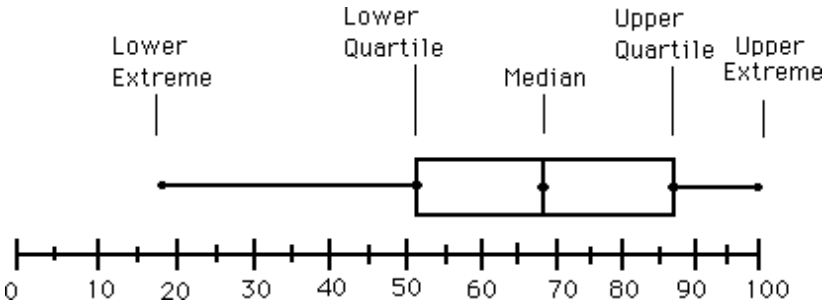
3.a) Spark has been proposed for improving MapReduce. What are the major differences that make Spark outperform MapReduce? Please list two differences. (10%)

3.b) Streaming computing only allows one data reading, i.e., never looking back. Please describe a way to calculate the variance of the data in a streaming manner. (6%)

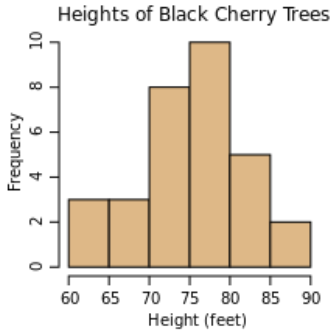
3.c) Can we calculate the mode, i.e., the number with highest frequency, of the data in a streaming manner? Please justify your answer. (5%)

4. Know your data. 15%

4.a) Boxplot and histogram are usually used for visualizing the data. What the criteria of selecting boxplot or histogram, i.e., when will we prefer one of them? Please list one scenario for each. (6%)



Boxplot



Histogram

4.b) In proximity measure for binary attributes. When we calculate the distance measurement for asymmetric binary variables, we often ignore the attributes that two objects are both 0. Please explain why. (5%)

4.c) Categorical data, e.g., music, clothes, foods, are different from numerical data. How can we transform the categorical data for learning? (4%)

5. Mining Frequent Patterns. 16%

Given a transaction database as follows.

TID	Items
100	A, B, D
200	B, C
300	A, B, C, E
400	B, E, G
500	A, C, G
600	B, C, D, G

Suppose the minimum support = 50%, i.e. 3 transactions.

5.a) Use the Apriori algorithm to generate candidate itemsets C_k and large itemsets L_k for all possible k≥1. (10%)

Hints: Apriori algorithm:

-Derivation of large 1-itemsets L_1: At the first iteration, scan all the transactions and count the number of occurrences for each item.

-Level-wise derivation: At the k^{th} iteration, the candidate set C_k are those whose every (k-1)-item subset is in L_{k-1}. Scan DB and count the # of occurrences for each candidate itemset.

$$\text{support}(A \cup B) = \Pr(A \cup B) = \frac{\#\text{of tx containing all items in } A \cup B}{\text{total \# of tx}}$$

$$\text{confidence}(A \cup B) = \Pr(B \mid A) = \frac{\#\text{of tx containing both } A \cup B}{\#\text{ of tx containing A}}$$

5.b) Derive any two association rule with support ≥ 50% and confidence ≥ 40%. Please also provide the support value and confidence value of this rule. (6%)

6. Information (8%)

6.a) In decision tree, why we calculate the expected information (entropy) for data? (4%)
(Hints: The formulation is listed below.)

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

6.b) Why is the formulation called the expected information (4%)