

註：選擇題答案為標紅字者，填空題答案不分大小寫

選擇/填空: (80%)

1. 下列何者不是促進大數據技術演進的主要原因？

- (a) 量產化硬體
- (b) 函式語言的興起**
- (c) 摩爾定律
- (d) 儲存設備成本降低

2. 請問下列何者不是大數據三個V中的其中一員？

- (a) Velocity
- (b) Victory**
- (c) Volume
- (d) Variety

3. Hadoop 是以什麼程式語言撰寫？ **Java**

4. 關於Hadoop 的特色，下列何者為非？

- (a) 高容錯性
- (b) 可以使用量產化硬體
- (c) 可以解決任意數據問題**
- (d) 採用主從式架構

5. 關於Hadoop 的敘述，下列何者為真？

- (a) 可以用來取代資料庫
- (b) 批次處理大數據**
- (c) 可以即時產生出統計報表
- (d) 快速進行機器學習

6. 關於MapReduce 的敘述，下列何者為誤？

- (a) 一定都會有Mapper及Reducer**
- (b) 可以使用Java 以外程式語言實做MapReduce
- (c) 可容錯
- (d) 單一Slave工作失敗時，Master 會重送工作

7. 關於HDFS 的敘述，下列何者為誤？

- (a) 可以快速存取單筆資料**
- (b) 建立於原生檔案系統上
- (c) 一次寫入，多次讀取
- (d) 高容錯性

8. 如果我要讓公司所有人都可以透過瀏覽器操作Hadoop，我可以使用哪個模組建立操作介面？

- (a)HUE
- (b)Hive
- (c)MapReduce
- (d)HDFS

9. 下列何者非HDFS的優點？

- (a)適合儲存小檔案
- (b)適合建立在量產化硬體上
- (c)高容錯性
- (d)適合寫入串流資料(Streaming)

10.請問如果要修改HDFS儲存的複本數，請問hdfs-site.xml 中的 name 中應該填入什麼值？ **dfs.replication**

```
<property>
    <name>_____</name>
    <value>3</value>
</property>
```

11.客戶端是向哪個代理服務(Daemon) 存取資料？

- (a)Name Node
- (b)Secondary Name Node
- (c)Node Manager
- (d)Data Node

12.如果有一個檔案大小為199MB，在Block Size 為128 MB及副本數為3 的設置下，該檔案會被分成多少個資料塊(Block)?**6**

13.如果要透過HDFS fs command下載檔案，請問指令 **hadoop fs ____ test.txt** 空白處該填入哪個指令？

- (a)-du
- (b)-df
- (c)-put
- (d)-get

14.如果我要檢視HDFS中 data目錄下的內容，空白處應該填？

hadoop fs ____ /data -ls

15.如果我要讓HDFS 可以即時備援Name Node，以防止單點毀損(Single Point Failure)，我該採取哪種架構？

- (a)在同一台機器上開啟兩個Name Node服務

(b)建立HDSF HA架構

- (c)使用Secondary Name Node 備援
- (d)使用HDFS Federation 備援

16.關於Secondary Name Node 的敘述，下列何者為非？

- (a)最好是將Secondary Name Node 配置於第二台機器
- (b)Secondary Name Node的所儲存最新的Meta 資料大概晚Name Node 約一個小時左右
- (c)Secondary Name Node 需要跟Name Node 一樣大的記憶體
- (d)Secondary Name Node 可以即時備援**

17.如果我要讓HDFS 可以使用多個Name Node 管理不同Data Node，以避免Client大量存取同一Name Node時遇到效能瓶頸，我該採取哪種架構？

- (a)使用Journal Node
- (b)在同一台機器上開啟兩個Name Node服務
- (c)使用HDFS Federation**
- (d)使用Secondary Name Node

18.如果要透過Java 呼叫API 存取HDFS，可以去哪個目錄夾找到相關JAR 檔(假設路徑已在hadoop 安裝目錄下)？

- (a)bin
- (b)etc
- (c)share**
- (d)conf

19.如果我要將HDFS內的某檔案(test.txt)權限設為 rw-r--r--，則下列指令 `hadoop fs -chmod ____ test.txt` 中的空白處我該填入什麼？ **644**

20.預設連結HDFS所使用的連接port 為____？ **9000**

21.下列何者不屬於YARN(MRV2)的元件？

- (a)ResourceManager
- (b)NodeManager
- (c)JobTracker**
- (d)ApplicationMaster

22.在一MapReduce 程式中, Reducer 最少可以有 ____ 個？ **0**

23.在Mapper 類別: `class mapper<I1,I2,I3,I4>{...}` 中, 請問參數I2 代表什麼意思？

- (a)output key

(b)input value

(c)input key

(d)output value

24.以下關於**Sqoop** 的敘述，下列何者為非？

(a)Sqoop 支援匯入csv 檔

(b)Sqoop 可以交換資料庫與HDFS 中的資料

(c)可以使用Sqoop 過濾匯入/匯出的資料

(d)Sqoop 支援增量匯入(Incremental Update)

25.請問**Sqoop** 是用什麼驅動程式存取資料庫？ ____ (請小寫) **jdbc**

26.請問如果要將資料庫資料匯入至**HDFS** 中，該使用下列哪個**Sqoop** 指令？

(a)sqoop select

(b)sqoop import

(c)sqoop insert

(d)sqoop export

27.請問下列何者並非**Flume** 所包含的元件？

(a)exec

(b)channel

(c)source

(d)sink

28.以下關於**Flume** 的敘述，下列何者為非？

(a)可以加值資料

(b)可以過濾傳輸中的資料

(c)可以清除資料中重複的部分

(d)可以從HDFS匯出資料至Flume 所在的Server

29.假設如果我要在**Flume**中設定從將資料串流寫進**HDFS**中，我該在**Flume** 中設定哪個元件？

(a)Source

(b)Channel

(c)Sink

30.請問下列何者並非**Hive** 的優點？

(a)使用者可以Ad-Hoc分析資料

(b)可以依時間做資料分割(partition)

(c)執行速度快

(d)簡單易用

31.請問下列關於**Hive** 跟資料庫比較的敘述，何者為是？

(a)Hive 是以MapReduce 存取資料

(b)使用Hive 可以即時完成資料處理與分析

(c)在Hive 中可以使用MySQL全部的語句

(d)Hive 可以取代關聯式資料庫

32.請問下列關於Hive 的敘述何者為非?

(a)Hive 可以使用類似MySQL的語句操作資料

(b)Hive 是將資料存放在HDFS 上

(c)使用Hive 跟使用MapReduce 的執行效能相近

(d)Hive 是透過MapReduce 來操作資料

33.請問如果要修改對Metastore資料庫連線的設定，請問要修改哪個檔案? **hive-site.xml**

34.請問如果要建立Hive 的外部表格, 下列空白處應填入(請全部以小寫表示)?

create _____ table **External**

35.以下對Spark描述何者為非?

(a)是一個叢集式運算框架

(b)以記憶體(Memory)作為運算的儲存媒體

(c)是用來取代Hadoop的產品

(d)運算速度較Hadoop快10~100倍

36.以下對RDD的轉換類(Transformations)操作的描述何者為非?

(a)用來將輸入的RDD轉換為另一個RDD

(b)是Lazy的運作

(c)具有容錯的特性

(d)主要以磁碟作為運算資料儲存媒體

37.以下操作何者不為RDD的轉換類(Transformations)操作?

(a)map

(b)filter

(c)count

(d)sortBy

38. **Which one is Start point for SparkSQL?** 答案應為**SparkSession**，因不在選項中，故本題**送分**

(a)scContext

(b)SQLContext

(c)HiveContext

(d)myContext

39. Spark Streaming不是即時運算的類庫，而是近即時運算。

(a)是

(b)否

40. Spark Streaming支援哪種資料來源？

(a)Kafka

(b)HDFS

(c)Flume

(d)S3

(e)以上皆是

情境題1(15%)：

目前有個系統建構需求，希望能接收前端WebServer的Log資料，由Log資料中截取感興趣的欄位進行分析及統計，並將分析結果存入關聯資料庫中，以便透過BI或Dashboard查看分析結果(如下圖)：



(a) 請問要完成上述功能，藍色虛線方框中需要使用那些Hadoop Ecosystem / Spark Library Stack 成員？

參考解答：

Flume + Spark Streaming + sqoop

或

Flume + Hive/MapReduce + sqoop

(b) 承上題，Hadoop Ecosystem / Spark Library Stack 成員的使用順序為何？每個成員負責什麼功能？

參考解答：

1. Flume收集Log資料

2. 由SparkStreaming或MapReduce／Hive進行分析

3. 由sqoop或sparkSql將資料寫回RDBMS

(c) 試分析您規劃的成員組合的分析即時性為何？(即時性的可能選項:即時/近即

時/批次作業)

參考解答：

由**Spark Streaming**處理者，為近即時、亦可為批次(但不建議)

由**Hive**或**MapReduce**處理者為批次

註：情境題給分方式：有寫出正確的收集／處理／匯出之成員，且即時性正確者即得滿分；多寫或亂槍打鳥者針斟扣分。