

Question: 1 of 60

**Which primary computing bottleneck of modern computers is addressed by Hadoop?**

Correct.

- |                  |                             |
|------------------|-----------------------------|
| unselected_label | A. MIPS                     |
| selected_label   | <b>B. disk latency</b> vrai |
| unselected_label | C. limited disk capacity    |
| unselected_label | D. 64-bit architecture      |

### ***CORRECTION TEST***

Question: 2 of 60

**Which capability does IBM BigInsights add to enrich Hadoop?**

Correct.

- |                  |                                             |
|------------------|---------------------------------------------|
| selected_label   | <b>A. Adaptive MapReduce</b>                |
| unselected_label | B. Parallel computing on commodity servers  |
| unselected_label | C. Fault tolerance through HDFS replication |
| unselected_label | D. Jaql                                     |

Question: 3 of 60

**Which Hadoop-related project provides common utilities and libraries that support other Hadoop sub projects?**

Correct.

- |                  |                         |
|------------------|-------------------------|
| unselected_label | A. Hadoop HBase         |
| selected_label   | <b>B. Hadoop Common</b> |
| unselected_label | C. BigTable             |
| unselected_label | D. MapReduce            |

Question: 4 of 60

**Which Big Data function improves the decision-making capabilities of organizations by enabling the organizations to interpret and evaluate structured and unstructured data in search of valuable business information?**

Correct.

unselected\_label      A. distributed file system

unselected\_label      B. stream computing

selected\_label      **C. analytics**

unselected\_label      D. data warehousing

Question: 5 of 60

**Which description identifies the real value of Big Data and Analytics?**

Correct.

unselected\_label      A. providing solutions to help customers manage and grow large database systems

unselected\_label      B. enabling customers to efficiently index and access large volumes of data

unselected\_label      C. using modern technology to efficiently store the massive amounts of data generated by social networks

selected\_label      **D. gaining new insight through the capabilities of the world's interconnected intelligence**

Question: 6 of 60

**Which type of Big Data analysis involves the processing of extremely large volumes of constantly moving data that is impractical to store?**

Correct.

unselected\_label      A. MapReduce

unselected\_label      B. Text Analysis

unselected\_label      C. Federated Discovery and Navigation

selected\_label      **D. Stream Computing**

Question: 7 of 60

**What is one of the two technologies that Hadoop uses as its foundation?**

Correct.

unselected\_label      A. HBase

selected\_label      **B. MapReduce**

unselected\_label      C. Apache

unselected\_label      D. Jaql

Question: 8 of 60

**What is one of the four characteristics of Big Data?**

Correct.

selected\_label      **A. volume**

unselected\_label      B. value

unselected\_label      C. verifiability

unselected\_label      D. volatility

Question: 9 of 60

**How are Pig and Jaql query languages similar?**

Correct.

unselected\_label      A. Both require schema.

unselected\_label      B. Both use Jaql query language.

unselected\_label      C. Both are developed primarily by IBM.

selected\_label      **D. Both are data flow languages.**

Question: 10 of 60

**What is a characteristic of IBM GPFS that distinguishes it from other distributed file systems?**

Correct.

- |                  |                                              |
|------------------|----------------------------------------------|
| unselected_label | A. operating system independence             |
| unselected_label | B. no single point of failure                |
| selected_label   | <b>C. posix compliance</b>                   |
| unselected_label | D. blocks that are stored on different nodes |

Question: 11 of 60

**Which command displays the sizes of files and directories contained in the given directory, or the length of a file, in case it is just a file?**

Correct.

- |                  |                         |
|------------------|-------------------------|
| unselected_label | A. hdfs -du             |
| selected_label   | <b>B. hadoop fs -du</b> |
| unselected_label | C. hdfs fs size         |
| unselected_label | D. hadoop size          |

Question: 12 of 60

**In the MapReduce processing model, what is the main function performed by the JobTracker?**

Correct.

- |                  |                                                   |
|------------------|---------------------------------------------------|
| selected_label   | <b>A. coordinates the job execution</b>           |
| unselected_label | B. executes the map and reduce functions          |
| unselected_label | C. assigns tasks to each cluster node             |
| unselected_label | D. copies Job Resources to the shared file system |

Question: 13 of 60

Following the most common HDFS replica placement policy, when the replication factor is three, how many replicas will be located on the local rack?

Correct.

- |                  |               |
|------------------|---------------|
| unselected_label | A. three      |
| unselected_label | B. none       |
| selected_label   | <b>C. one</b> |
| unselected_label | D. two        |

Question: 14 of 60

Which statement is true regarding the number of mappers and reducers configured in a cluster?

Correct.

- |                  |                                                                                                       |
|------------------|-------------------------------------------------------------------------------------------------------|
| unselected_label | A. The number of mappers must be equal to the number of nodes in a cluster.                           |
| unselected_label | B. The number of reducers is always equal to the number of mappers.                                   |
| unselected_label | C. The number of mappers and reducers is decided by the NameNode.                                     |
| selected_label   | <b>D. The number of mappers and reducers can be configured by modifying the mapred-site.xml file.</b> |

Question: 15 of 60

Which command helps you create a directory called mydata on HDFS?

Correct.

- |                  |                                                |
|------------------|------------------------------------------------|
| unselected_label | A. <code>hadoop fs -dir mydata</code>          |
| selected_label   | <b>B. <code>hadoop fs -mkdir mydata</code></b> |
| unselected_label | C. <code>mkdir mydata</code>                   |
| unselected_label | D. <code>hdfs -dir mydata</code>               |

Question: 16 of 60

**What are two of the core operators that can be used in a Jaql query? (Select two.)**

**(Please select ALL that apply)**

Correct.

unselected\_label      A. SELECT

selected\_label      **B. JOIN**

unselected\_label      C. LOAD

selected\_label      **D. TOP**

Question: 17 of 60

**Which type of language is Pig?**

Correct.

unselected\_label      A. SQL-like

unselected\_label      B. object oriented

unselected\_label      C. compiled language

selected\_label      **D. data flow**

Question: 18 of 60

**Under the MapReduce programming model, which task is performed by the Reduce step?**

Correct.

unselected\_label      A. Input data is split into smaller pieces.

selected\_label      **B. Data is aggregated by worker nodes.**

unselected\_label      C. Worker nodes process individual data segments in parallel.

unselected\_label      D. Worker nodes store results in the local file system.

Question: 19 of 60

**In which step of a MapReduce job is the output stored on the local disk?**

Correct.

- |                  |               |
|------------------|---------------|
| unselected_label | A. Shuffle    |
| selected_label   | <b>B. Map</b> |
| unselected_label | C. Reduce     |
| unselected_label | D. Combine    |

Question: 20 of 60

**What key feature does HDFS 2.0 provide that HDFS does not?**

Correct.

- |                  |                                             |
|------------------|---------------------------------------------|
| unselected_label | A. random access to data in the cluster     |
| unselected_label | B. a high throughput, shared file system    |
| selected_label   | <b>C. high availability of the NameNode</b> |
| unselected_label | D. data access performed by an RDBMS        |

Question: 21 of 60

**Which statement represents a difference between Pig and Hive?**

Correct.

- |                  |                                                                        |
|------------------|------------------------------------------------------------------------|
| unselected_label | A. Pig has a shell interface for executing commands.                   |
| unselected_label | B. Pig is used for creating MapReduce programs.                        |
| selected_label   | <b>C. Pig uses Load, Transform, and Store.</b>                         |
| unselected_label | D. Pig is not designed for random reads/writes or low-latency queries. |

Question: 22 of 60

**Which element of the MapReduce architecture runs map and reduce jobs?**

Correct.

- |                  |                       |
|------------------|-----------------------|
| unselected_label | A. Reducer            |
| unselected_label | B. JobTracker         |
| selected_label   | <b>C. TaskTracker</b> |
| unselected_label | D. JobScheduler       |

Question: 23 of 60

**What is one of the two driving principles of MapReduce?**

Correct.

- |                  |                                                                      |
|------------------|----------------------------------------------------------------------|
| unselected_label | A. provide structure to unstructured or semi-structured data         |
| selected_label   | <b>B. spread data across a cluster of computers</b>                  |
| unselected_label | C. provide a platform for highly efficient transaction processing    |
| unselected_label | D. increase storage capacity through advanced compression algorithms |

Question: 24 of 60

**What is one function of the JobTracker in MapReduce?**

Correct.

- |                  |                                                       |
|------------------|-------------------------------------------------------|
| unselected_label | A. reports status of DataNodes                        |
| selected_label   | <b>B. keeps the work physically close to the data</b> |
| unselected_label | C. runs map and reduce tasks                          |
| unselected_label | D. manages storage                                    |



Question: 25 of 60

**In addition to the high-level language Pig Latin, what is a primary component of the Apache Pig platform?**

Correct.

- |                  |                                    |
|------------------|------------------------------------|
| unselected_label | A. platform-specific SQL libraries |
| unselected_label | B. built-in UDFs and indexing      |
| unselected_label | C. an RDBMS such as DB2 or MySQL   |
| selected_label   | <b>D. runtime environment</b>      |

Question: 26 of 60

**To run a MapReduce job on the BigInsights cluster, which statement about the input file(s) must be true?**

Correct.

- |                  |                                                                                                         |
|------------------|---------------------------------------------------------------------------------------------------------|
| unselected_label | A. The file(s) must be stored on the local file system where the map reduce job was developed.          |
| unselected_label | B. No matter where the input files are before, they will be automatically copied to where the job runs. |
| unselected_label | C. The file(s) must be stored on the JobTracker.                                                        |
| selected_label   | <b>D. The file(s) must be stored in HDFS or GPFS.</b>                                                   |

Question: 27 of 60

**Which command should be used to list the contents of the root directory in HDFS?**

Correct.

- |                  |                           |
|------------------|---------------------------|
| unselected_label | A. hdfs list /            |
| unselected_label | B. hadoop fs list         |
| selected_label   | <b>C. hadoop fs -ls /</b> |
| unselected_label | D. hdfs root              |

Question: 28 of 60

**If you need to change the replication factor or increase the default storage block size, which file do you need to modify?**

Correct.

- |                  |                             |
|------------------|-----------------------------|
| unselected_label | A. hadoop.conf              |
| unselected_label | B. hdfs.conf                |
| unselected_label | C. hadoop-configuration.xml |
| selected_label   | <b>D. hdfs-site.xml</b>     |

Question: 29 of 60

**Under the HDFS architecture, what is one purpose of the NameNode?**

Correct.

- |                  |                                              |
|------------------|----------------------------------------------|
| unselected_label | A. to coordinate MapReduce jobs              |
| unselected_label | B. to periodically report status to DataNode |
| selected_label   | <b>C. to regulate client access to files</b> |
| unselected_label | D. to manage storage attached to nodes       |

Question: 30 of 60

**Which statement is true about Hadoop Distributed File System (HDFS)?**

Correct.

- |                  |                                                                                 |
|------------------|---------------------------------------------------------------------------------|
| selected_label   | <b>A. Data is accessed through MapReduce.</b>                                   |
| unselected_label | B. Data can be created, updated and deleted.                                    |
| unselected_label | C. Data is designed for random access read/write.                               |
| unselected_label | D. Data can be processed over long distances without a decrease in performance. |

Question: 31 of 60

**When running a MapReduce job from Eclipse, which BigInsights execution models are available? (Select two.)**

**(Please select ALL that apply)**

Correct.

selected\_label      **A. Local**

unselected\_label      B. Distributed

unselected\_label      C. Debugging

unselected\_label      D. Remote

selected\_label      **E. Cluster**

Question: 32 of 60

**Which Hadoop-related technology supports analysis of large datasets stored in HDFS using an SQL-like query language?**

Correct.

selected\_label      **A. Hive**

unselected\_label      B. Pig

unselected\_label      C. HBase

unselected\_label      D. Jaql

Question: 33 of 60

**Which statement about NoSQL is true?**

Correct.

unselected\_label      A. It is based on the highly scalable Google Compute Engine.

selected\_label      **B. It is a database technology that does not use the traditional relational model.**

unselected\_label      C. It is an IBM project designed to enable DB2 to manage Big Data.

unselected\_label      D. It provides all the capabilities of an RDBMS plus the ability to manage Big Data.

Question: 34 of 60

**In Hive, what is the difference between an external table and a Hive managed table?**

Correct.

- |                  |                                                                                          |
|------------------|------------------------------------------------------------------------------------------|
| selected_label   | <b>A. An external table refers an existing location outside the warehouse directory.</b> |
| unselected_label | B. An external table refers to the data stored on the local file system.                 |
| unselected_label | C. An external table refers to a table that cannot be dropped.                           |
| unselected_label | D. An external table refers to the data from a remote database.                          |

Question: 35 of 60

**What is the "scan" command used for in HBase?**

Correct.

- |                  |                                                  |
|------------------|--------------------------------------------------|
| unselected_label | A. to get detailed information about the table   |
| unselected_label | B. to report any inconsistencies in the database |
| unselected_label | C. to list all tables in Hbase                   |
| selected_label   | <b>D. to view data in an Hbase table</b>         |

Question: 36 of 60

**Which utility provides a command-line interface for Hive?**

Correct.

- |                  |                        |
|------------------|------------------------|
| unselected_label | A. Hive SQL client     |
| unselected_label | B. Thrift client       |
| selected_label   | <b>C. Hive shell</b>   |
| unselected_label | D. Hive Eclipse plugin |

Question: 37 of 60

**If you need to JOIN data from two workbooks, which operation should be performed beforehand?**

Correct.

- |                  |                                                                                             |
|------------------|---------------------------------------------------------------------------------------------|
| unselected_label | A. "Copy" to create a new sheet with the other workbook data in the current workbook        |
| selected_label   | <b>B. "Load" to create a new sheet with the other workbook data in the current workbook</b> |
| unselected_label | C. "Group" to bring together the two workbooks                                              |
| unselected_label | D. "Add" to add the other workbook data to the current workbook                             |

Question: 38 of 60

**Which is a use-case for Text Analytics?**

Correct.

- |                  |                                                             |
|------------------|-------------------------------------------------------------|
| unselected_label | A. health insurance cost/benefit analysis from payroll data |
| unselected_label | B. managing customer information in a CRM database          |
| selected_label   | <b>C. sentiment analytics from social media blogs</b>       |
| unselected_label | D. product cost analysis from accounting systems            |

Question: 39 of 60

**Which statement will make an AQL view have content displayed?**

Correct.

- |                  |                                         |
|------------------|-----------------------------------------|
| selected_label   | <b>A. output view &lt;view_name&gt;</b> |
| unselected_label | B. display view <view_name>             |
| unselected_label | C. return view <view_name>              |
| unselected_label | D. export view <view_name>              |

Question: 40 of 60

**What drives the demand for Text Analytics?**

Correct.

- |                  |                                                                                |
|------------------|--------------------------------------------------------------------------------|
| unselected_label | A. Data warehouses contain potentially valuable information.                   |
| unselected_label | B. MapReduce is unable to process unstructured text.                           |
| selected_label   | <b>C. Most of the world's data is in unstructured or semi-structured text.</b> |
| unselected_label | D. Text Analytics is the most common way to derive value from Big Data.        |

Question: 41 of 60

**You work for a hosting company that has data centers spread across North America. You are trying to resolve a critical performance problem in which a large number of web servers are performing far below expectations. You know that the information written to log files can help determine the cause of the problem, but there is too much data to manage easily. Which type of Big Data analysis is appropriate for this use case?**

Correct.

- |                  |                          |
|------------------|--------------------------|
| unselected_label | A. Temporal Analysis     |
| unselected_label | B. Data Warehousing      |
| selected_label   | <b>C. Text Analytics</b> |
| unselected_label | D. Stream Computing      |

Question: 42 of 60

**Which key benefit does NoSQL provide?**

Correct.

- |                  |                                                                                     |
|------------------|-------------------------------------------------------------------------------------|
| unselected_label | A. It allows an RDBMS to maintain referential integrity on a Hadoop data set.       |
| selected_label   | <b>B. It can cost-effectively manage data sets too large for traditional RDBMS.</b> |
| unselected_label | C. It allows customers to leverage high-end server platforms to manage Big Data.    |
| unselected_label | D. It allows Hadoop to apply the schema-on-ingest model to unstructured Big Data.   |

Question: 43 of 60

**Which technology does Big SQL utilize for access to shared catalogs?**

Correct.

selected\_label      **A. Hive metastore**

unselected\_label      B. RDBMS

unselected\_label      C. MapReduce

unselected\_label      D. HCatalog

Question: 44 of 60

**Why develop SQL-based query languages that can access Hadoop data sets?**

Correct.

selected\_label      **A. because the MapReduce Java API is sometimes difficult to use**

unselected\_label      B. because the data stored in Hadoop is always structured

unselected\_label      C. because data stored in a Hadoop cluster lends itself to structured SQL queries

unselected\_label      D. because SQL enhances query performance

Question: 45 of 60

**What is the most efficient way to load 700MB of data when you create a new HBase table?**

Correct.

selected\_label      **A. Pre-create regions by specifying splits in create table command and bulk loading the data.**

unselected\_label      B. Pre-create the column families when creating the table and bulk loading the data.

unselected\_label      C. Pre-create regions by specifying splits in create table command and use the insert command to load data.

unselected\_label      D. Pre-create the column families when creating the table and use the put command to load the data.

Question: 46 of 60

**Which Hive command is used to query a table?**

Correct.

- |                  |                  |
|------------------|------------------|
| unselected_label | A. EXPAND        |
| unselected_label | B. TRANSFORM     |
| unselected_label | C. GET           |
| selected_label   | <b>D. SELECT</b> |

Question: 47 of 60

**Which tool is used to access BigSheets?**

Correct.

- |                  |                       |
|------------------|-----------------------|
| unselected_label | A. Microsoft Excel    |
| unselected_label | B. BigSheets client   |
| selected_label   | <b>C. Web Browser</b> |
| unselected_label | D. Eclipse            |

Question: 48 of 60

**Which command can be used in Hive to list the tables available in a database/schema?**

Correct.

- |                  |                       |
|------------------|-----------------------|
| unselected_label | A. show all           |
| unselected_label | B. list tables        |
| unselected_label | C. describe tables    |
| selected_label   | <b>D. show tables</b> |



Question: 49 of 60

In HBase, what is the "count" command used for?

Correct.

- |                  |                                                      |
|------------------|------------------------------------------------------|
| selected_label   | <b>A. to count the number of rows in a table</b>     |
| unselected_label | B. to count the number of columns of a table         |
| unselected_label | C. to count the number of regions of a table         |
| unselected_label | D. to count the number of column families of a table |

Question: 50 of 60

The following sequence of commands is executed:

```
create 'table_1','column_family1','column_family2'
```

```
put 'table_1','row1','column_family1:c11','r1v11'
```

```
put 'table_1','row2','column_family1:c12','r1v12'
```

```
put 'table_1','row2','column_family2:c21','r1v21'
```

```
put 'table_1','row3','column_family1:d11','r1v11'
```

```
put 'table_1','row2','column_family1:d12','r1v12'
```

```
put 'table_1','row2','column_family2:d21','r1v21'
```

In HBase, which value will the "count 'table\_1'" command return?

Correct.

- |                  |             |
|------------------|-------------|
| unselected_label | A. 4        |
| selected_label   | <b>B. 3</b> |
| unselected_label | C. 2        |
| unselected_label | D. 6        |

Question: 51 of 60

**Which tool is used for developing a BigInsights Text Analytics extractor?**

Correct.

- |                  |                                                             |
|------------------|-------------------------------------------------------------|
| unselected_label | A. AQLBuilder                                               |
| unselected_label | B. BigInsights Console with AQL plugin                      |
| unselected_label | C. AQL command line                                         |
| selected_label   | <b>D. Eclipse with BigInsights tools for Eclipse plugin</b> |

Question: 52 of 60

**What makes SQL access to Hadoop data difficult?**

Correct.

- |                  |                                                  |
|------------------|--------------------------------------------------|
| selected_label   | <b>A. Data is in many formats.</b>               |
| unselected_label | B. Hadoop requires pre-defined schema.           |
| unselected_label | C. Hadoop data is highly structured.             |
| unselected_label | D. Data is located on a distributed file system. |

Question: 53 of 60

**Which Hadoop-related technology provides a user-friendly interface, which enables business users to easily analyze Big Data?**

Correct.

- |                  |                     |
|------------------|---------------------|
| unselected_label | A. BigSQL           |
| unselected_label | B. Avro             |
| selected_label   | <b>C. BigSheets</b> |
| unselected_label | D. HBase            |

Question: 54 of 60

**What is an accurate description of HBase?**

Correct.

- |                  |                                                                           |
|------------------|---------------------------------------------------------------------------|
| unselected_label | A. It is a distributed file system that replicates data across a cluster. |
| selected_label   | <b>B. It is an open source implementation of Google's BigTable.</b>       |
| unselected_label | C. It is a database schema for unstructured Big Data.                     |
| unselected_label | D. It is a data flow language for structured data based on Ansi-SQL.      |

Question: 55 of 60

**What is one of the main components of Watson Explorer (InfoSphere Data Explorer)?**

Correct.

- |                  |                   |
|------------------|-------------------|
| selected_label   | <b>A. crawler</b> |
| unselected_label | B. validator      |
| unselected_label | C. replicater     |
| unselected_label | D. compressor     |

Question: 56 of 60

**How can the applications published to BigInsights Web Console be made available for users to execute?**

Correct.

- |                  |                                                            |
|------------------|------------------------------------------------------------|
| selected_label   | <b>A. They need to be deployed with proper privileges.</b> |
| unselected_label | B. They need to be marked as "Shared."                     |
| unselected_label | C. They need to be linked with the master application.     |
| unselected_label | D. They need to be copied under the user home director     |

Question: 57 of 60

**Which component of Apache Hadoop is used for scheduling and running workflow jobs?**

Correct.

unselected\_label      A. Task Launcher

unselected\_label      B. Eclipse

unselected\_label      C. Jaql

selected\_label      **D. Oozie**

Question: 58 of 60

**IBM InfoSphere Streams is designed to accomplish which Big Data function?**

Correct.

selected\_label      **A. analyze and react to data in motion before it is stored**

unselected\_label      B. execute ad-hoc queries against a Hadoop-based data warehouse

unselected\_label      C. analyze and summarize product sentiments posted to social media

unselected\_label      D. find and analyze historical stream data stored on disk

Question: 59 of 60

**Which IBM tool enables BigInsights users to develop, test and publish BigInsights applications?**

Correct.

unselected\_label      A. HBase

selected\_label      **B. Eclipse**

unselected\_label      C. Avro

unselected\_label      D. BigInsights Applications Catalog

Question: 60 of 60

**Which IBM Big Data solution provides low-latency analytics for processing data-in-motion?**

Correct.

- |                  |                                  |
|------------------|----------------------------------|
| selected_label   | <b>A. InfoSphere Streams</b>     |
| unselected_label | B. InfoSphere Information Server |
| unselected_label | C. InfoSphere BigInsights        |
| unselected_label | D. PureData for Analytics        |

## **Modules en ligne**

### **Unit 1. Big Data Overview**

Q1.Which of the following is not a characteristic of Big Data

**Virtual**

Q2. Which of the following statements is not true about the 5 key Big Data Use Cases?

**Enhanced 360° View of the customer extends existing customer views by incorporating internal data sources only.**

Q3.Which of the following Big Data Platform components can cost effectively analyze petabytes of structured and unstructured data at rest?

**Hadoop System**

### **Unit 2. Hadoop big data analysis tool (45 minutes)**

Q1.What is Hadoop?

**An open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware.**

Q2.What are the main components of Hadoop?

**Map Reduce, Hadoop Common, HDFS**

Q3.How do you manage BigInsights?

- Add/Drop services
- Explore/Modify system Health
- Launch/Monitor file system

-Start/Stop jobs

**None of the above**

### **Unit 3. HDFS**

Q1.What is NOT true about Hadoop Distributed File System (HDFS)?

**Designed for random access not streaming reads**

Q2. True or False: In HDFS, the data blocks are replicated to multiple nodes

**True**

Q3.What is NOT true about the Namenode StartUp?

**NameNode stores data blocks**

### **Exercise 1. HDFS Lab (60 minutes)**

\*Q1. HDFS command for reporting basic filesystem information and statistics

**hadoop dfsadmin -report**

\*Q2. Select two methods for browsing files stored to HDFS in BigInsights

**hadoop fs , BigInsights Web Console, Files Tab**

\*Q3. Hadoop command for copying files from local file system to HDFS:

**hadoop fs -put**

### **Unit 4. MapReduce (60 minutes)**

Q1. What is not a MapReduce Task?

**Reader**

Q2. True or False: Map Tasks need Key and Value pairs as input.

**true**

Q2.Which Java class is responsible for taking HDFS file and transforming it into splits.

**InputSplitter**

Q3.Regarding Task Failure; if a child task fails, where does the child JVM reports before

**TaskTracker**

### **Exercise 2. Map Reduce Lab (60 minutes)**

Q1.What are the main components of a Java MapReduce program (Select 3)

**Map class, Reduce class, Driver or main class**

Q2.What are two execution modes of a Java MapReduce program developed in Eclipse

**Cluster, Local**

Q3.What are the breakpoints which can be set for a Java MapReduce program in Eclipse

**stop points to suspend the program execution for debugging purposes**

## Unit 5. Hadoop Query Languages (45 minutes)

Q.1.What is NOT a similarity of Pig, Hive and Jaql?

**Designed for random reads/writes or low latency queries**

Q2.What is NOT true about Hive?

**Designed for low latency queries, like RDBMS such as DB2 and Netezza**

Q3.True or False: A Jaql query can be thought of as a stream of operations. **True**

## Unit 6. Hive (45 minutes)

Q1.True or False: Data Warehouse augmentation is a very common use case for Hadoop.

**True**

Q2.The clause that indicates the storage file/record format on HDFS.

**stored as**

Q3.True or False: Hive comes with an HBase storage handler.

**True**

Q4.A variable name that determines whether the map/reduce jobs should be submitted through a separate JVM in the non local mode.

**hive.exec.submitviachild**

## Exercise 3. Hive Lab (60 minutes)

Q1.What is true regarding Hive External Tables?

**data is stored outside the Hive warehouse directory**

Q2. What Hive command list all the base tables and views in a database?

**show tables**

Q3.What Hive command is used for loading data from a local file to HDFS?

**LOAD DATA LOCAL INPATH**

## Unit 7. HBase (60 minutes)

Q1. What is true about HBase?

**All of the above**

Q2. Which is not an ACID property?

**Concurrency**

Q3. True or False: HBase region servers are responsible for serving and managing regions

**True**

Q4. What is NOT true about ZooKeeper?

**It is a tool for analyzing streaming data**

## Exercise 4. HBase Lab (60 minutes)

Q1. Which of the following statements is true about CREATE statement in HBase

**It only requires the name of table and one or more column families**

Q2. Which command must be executed before deleting an HBase table or changing its settings?

**Disable**

Q3. Which command is used in HBase for inserting data in a table?

**Put**

## Unit 8: Big SQL (45 minutes)

Q1. What is NOT true regarding Big SQL?

**Incurs measurable overhead for the sake of resiliency**

Q2. Which of the following tools can be used to access a Big SQL server?

**All of the above**

Q3. Which of the following data types are not supported by Big SQL?

**Byte**

Q4. You can insert records into BIGSQL table mapped to an HBASE table

**True**

Q5. You can insert records into BIGSQL table mapped to a HIVE table

**False**



## Exercise 5. Big SQL Lab (45 minutes)

Q1.Which statement can be used to load data in a Big SQL table?

**load hadoop**

Q2.Which of the following is true about CREATE HADOOP TABLE statement

**It creates a Big SQL table and the data will be stored in HDFS/GPFS**

Q3.Which of the following statements represents a difference between a Big SQL table and a DB2 table?

**When creating a Big SQL table, information on how the data is stored on disk must be provided**

NO.3 Which of the following options is CORRECT?

A. InfoSphere Streams submits queries to structured static data.

B. InfoSphere Streams submits queries to structured dynamic data.

C. InfoSphere Streams submits queries to unstructured dynamic data.

**D. InfoSphere Streams submits dynamic data to pre-existing queries.**

NO.5 Hadoop environments are optimized for:

A. Processing transactions (random access).

B. Low latency data access.

**C. Batch processing on large files.**

D. Intensive calculation with little data.

## Unit 13. AQL Syntax (45 minutes)

Q1.A data model scalar type that represents a bag of values.

**List**

Q2. In the context of AQL, which of the following is not a clause?

**View**

True or False: Text Analytics is a powerful information extraction system providing multilingual support.

**True**

Line of code used to specify the Module Name at the beginning of an AQL file

**module**

## Unit 12. Text Analytics (60 minutes)

Q1.An SQL style programming language for text mining or text extraction?

## AQL

Q2.A list of entries containing domain specific terms.

## Dictionary

Q3.Which is not a built-in scalar type of an AQL Data Model?

## Char

Q4.True or False: Text analytics Java API is part of the Big Insights Text Analytics Components.

## True

## Unit 9. JAQL (60 minutes)

Q1. What is JAQL?

**Jaql is a query language initially designed for Javascript Object Notation (JSON) data format**

Q2. An operation that can't be performed by JAQL on I/O Adapters?

**localWrite()**

Q3. The function that explicitly sends each element of an array to a mapper.

**arrayRead()**

## Exercise 6. JAQL Lab (90 minutes)

Q1.Which operation allows reading a file using Jaql?

**read()**

Q2. Which of the following Jaql operators helps projecting or retrieving a subset of columns or fields from a data set?

**Transform**

Q3.Which Jaql expression can be used to flatten nested arrays?

**Expand**

## Unit 10. Application Development (45 minutes)

Q1.Which does not belong in the BlgInsights Application types?

**BigData**

Q2. What component of Apache Hadoop can be used for scheduling and running workflow jobs?

**Oozie**

Q.3.An element in a Workflow XML that invokes a program or script?

**Action**

Q4.Steps in deploying your application on the cluster.

**Access applications tab -> Manage published applications -> Locate new application -> Deploy application -> Execute application**

## **Exercise 7: Application Development Lab (90 minutes)**

Q1. Select a BigInsights execution mode for an application or script developed in Eclipse

**Cluster**

Q2. What step needs to be performed for an application published to BigInsights Web Console to become available for users to run it?

**It needs to be deployed with proper privileges**

Q3. Which part of the BigInsights web console, provides information, that would help a user troubleshoot and diagnose, a failed application job?

**Application Status tab**

## **Unit 11. Browser-based data analytics tool (45 minutes)**

Q1. What is BigSheets?

Evaluates and compiles front-end commands into executable work

Browser based analytics tool for business users

Executes work on a simulated environment of sample data

Combines data in different workbooks

**All of the above**

Q2. Which is NOT true regarding the BigSheets workbook?

Representation of data within the tool

Simple spreadsheet-like structure

Created and managed by user

All of the above

**None of the above**

Q3. What is the 'Add Sheets' option that helps calculate values by grouping the data in the workbook, applying functions to each group and carrying over data.

**Pivot**

Q4. Provides 10+ build-in functions to extract names, addresses, organizations, email and phone numbers

**Text Analytics Integration**

## **Exercise 8. browser-based data analytics tool lab (90 minutes)**

Q1. Which statement is NOT true about BigSheets Readers?

**They are needed to visualise encrypted data**

Q2. True or False: You can apply only a single operation to a sheet

**True**

Q3. Which statement describes a BigSheets Master Workbook feature?

**It points to a file or table stored in Hadoop**

## **Exercise 9. AQL Email Analysis Lab (60 minutes)**

Q1. Which tool can be used for developing a Text Analytics Extractor?

**Eclipse with BigInsights tools for Eclipse**

Q2. Which of the following is an AQL statement that can be used to create a component of a Text Analytics extractor

**create view <view\_name> as <select or extract statement>;**

Q3. What statement can be used to display the content of an AQL view?

**output view <view\_name>;**

## **Unit 14. Streams (45 minutes)**

Q1. What big data characteristic is not covered by InfoSphere Streams?

**Virtual**

Q2. What is true regarding Streams Processing Language (SPL)?

It is declarative

It has procedural support

It is extensible

It is designed for stream computing

**All of the above**

Q3. An Eclipse-based tool that enables developers to create, edit, visualize, test, debug, run SPL and SPL mixed-mode applications

**InfoSphere Streams Studio**

Q4. SWS provides web-based graphical user interface in Streams Console, what does SWS stand for?

**Streams Web Service**

## **Exercise 10: Streams Lab**

Q1.Which tool is used to develop Streams jobs and applications?

**Streams Studio**

Q2.Which of the following statements is true regarding the Instance Graph available in InfoSphere Streams:

**The Instance Graph provides a graphical view of the application that's currently running in an instance; it is a very helpful tool for debugging Streams applications at runtime**

Q3. In the context of InfoSphere Streams what is Streams Console?

**Web based tool used to monitor and manage InfoSphere Streams instances and applications**

## **Unit 15: Data Explorer overview**

Q1.What is true regarding Data Explorer?

**Explorer's search combines content and data from many different systems throughout the enterprise and presents it to users in a single view.**

Q2. In the context of an enterprise search engine, what is called the process of retrieving documents?

**Crawling**

Q3. A component of Data Explorer Engine that process raw data discovered by the crawler and produce 1+ pieces of index-able data.

**Converting**

Q4.True or False: Data Explorer's collaborative search enables saving and sharing of results and queries among users.

**True**

## **Exercise 11: Data Explorer Lab**

Q1What is a Search Collection?

**It represents one or more information repositories and the online index created from them when the Data Explorer search engine crawls those sources**

Q2.What should be done when configuring a Data Explorer project so that the search results are grouped by values of specific metadata, tags, or other parameters?

**Add Binning to the Search Collection**

Q3.Which tool can be used to create and configure a Data Explorer search project?

**Data Explorer Engine administration tool**