Data Essentials

1.On the Component tab displayed for the tSortRow component, when you click to add criteria to the Criteria table, Talend automatically populates the column values with defaults. In the "sort num or alpha?" column Talend has chosen num by default for customer_id as displayed. Which other values are available for the "sort num or alpha?" column when you click to open the dropdown list for that column?

Ans. date,alpha

2.You are using a tExtractDelimitedFields component to split the Address2 field in the delimited file as displayed. What must you specify as the Field separator property for the tExtractDelimitedFields component to properly split the Address2 field?

```
ID;First_Name;Last_Name;Address1;Address2;Country

1;Claudia;Sand;10000 Main Dr NW;New York, NY;USA

2;Max;Bigot;60000 My St;Nashua, NH;USA

3;Rick;Tailleur;200000 Younge St;Toronto, ON;Canada

4;Noémie;Miller;500000 St. Catherines St Ouest;Montreal, QC;Canada

5;Catherine;Reilly;100000 Main St SW;Boston, MA;USA
```

Ans. "," (a comma)

3. When configuring the properties for a tReplace component, you can optionally click the Advanced mode checkbox. Doing so allows you to specify what type of expression as the Pattern to search for?

Ans. Regular expression

4. When configuring the properties for a tAggregateRow component. You are going to Group by the customer_id field in order to aggregate the sales on a customer_id basis, so that in the resulting output file you will have one row for each customer_id with aggregated sales figures. Which Function value must you choose when configuring the Operations table for this tAggregateRow component?

Ans. Sum

5. You are using a tNormalize component to normalize the category field in the delimited file as displayed. What must you specify as the Item separator property for the tNormalize component to properly normalize the category field?

```
product_id,product_name,category,unit_price

1,Regular Widget,Normally Aspirated;Low Price;Highly Reliable,125

2,Super Widget,Super;Medium Price;Very Reliable,250

3,Turbo Widget,Turbo;Medium Price;Very Reliable,275

4,S/T Widget,Super;Turbo;Premium Price;Reliable,425

5,Hybrid Widget,Hybrid;Normally Aspirated;Premium Price;Reliable,425
```

Ans. ";" (a semi-colon)

6. Which of these statements accurately describes the Unique Match Join match model in the Map Editor?

Ans. Last match is considered and passed to the output, Set as default when configuring an explicit Join

7. You are building an expression in Map Editor for a column in which you want to pad the row1.CustomerID string with leading zeroes up to maximum length of 6 characters. Which code can you use to accomplish this for you?

Ans. String.format("$06d",row1.CustomerID)

8. You have Map Editor open for a tMap object for which you are mapping databases objects. You click the Join Model for a table Join property. Which options appear in the Options dialog that appears?

Ans. Left outer join, inner join

9. You are building a filter expression in Map Editor for a column in which you want to filter the product name, prd.name to equal "Turbo Widgets", and you want the transaction quantity, tx.qty to be greater than 100. Which code can you use to accomplish this for you?

Ans. prd.name.equals("Turbo Widgets")&&tx.qty>100

10. Which common FTP operations are supported by Talend Open Studio components available in the Palette?

Ans. Put

Selected

□

File Exist

□

Delete

11. What could be accomplished by using big data technologies?

Ans. New product development and optimized offerings, Cost reductions

12. Velocity-The speed at which data is processed and becomes accessible

Volume-The amount of data that exists

Variety-The different types of data from XML to video to SMS

Veracity-Making sure the data is accurate, which requires processes to keep bad data from accumulating in your systems

13. Which statements are correct about how Netflix utilizes big data?

Ans. Netflix uses what is known as the big data recommendation algorithm to suggest TV shows and movies based on a user's preferences

Netflix has screenshots of scenes people might have viewed repeatedly, the associated ratings, and the number of searches and the search topics

14. What are the main challenges that companies experience with big data?

Ans. Unfamiliarity with big data and confusing it with traditional methods

Unprecedented data growth

Data security issues

Integrating data from a variety of sources

15. What are some examples of big data sources?

Ans. Open data, Social media, Sensor data, Email

16. What are some of the main business domains that use big data tools today?

Ans. E-commerce industry, Aviation industry, Credit scoring agencies, Transportation industry

17. What are the main deliverables of big data?

Ans. Text/image analytics, Multivariate analysis, Predictive models

18. What are the most important advantages of big data, according to the International Institute for Analytics (IIA)?

Ans. Big data enables faster, better decision making

Big data leads to cost reductions

Big data helps to identify what customers need and to introduce new products and services accordingly

19. Which statement is true about in-memory storage systems?

Ans. Data storage in an in-memory database is reliant on random access memory (RAM)

20. What are the most important features of HDFS?

Ans. Replication, Scalability, Distributed storage, High availability

21. Which statement about parallel or distributed computing is true?

Ans. Distributed computing can allow an application on one machine to leverage processing power, memory, or storage on another machine

22. Match each Hadoop component with its respective layer in the Hadoop ecosystem. One layer will not be used.

Ans. ZooKeeper-Data management layer

HDFS-Data storage layer

Hive-Data access layer

MapReduce-Data processing layer

23. What are the benefits of migrating from Hadoop to the cloud?

Ans. Long-term cost savings, Better scalability, Better collaboration, Easy access and resource availability

24. Which statements are true about unstructured data?

Ans. Unstructured data is very often linked to structured data. An example is how X-ray images at a hospital are linked to patient IDs or health card numbers.

Web pages, video files, and audio files are examples of unstructured data

25. Which statement about horizontal and vertical scaling is true?

Ans. Horizontal scaling is typically the easiest scaling option

26. Which statements are correct about HDFS?

Ans. HDFS provides high throughput access to application data by providing the data access in parallel,

HDFS provides a fault-tolerant storage layer for Hadoop and its other components

27. What are the differences between Hadoop and cloud computing?

Ans. Cloud computing focuses on on-demand, scalable, and adaptable service models, while Hadoop is all about extracting value out of volume, variety, and velocity.

Cloud computing constitutes various computing concepts. This naturally involves a large number of computers that are usually connected through a real-time communication network. Hadoop, on the other hand, is a framework that uses simple programming models to process large data sets across clusters of computers.

28. Which statements accurately describe the differences between big data and data warehousing?

Ans. Data warehouses only handle structured data (relational or non-relational), whereas big data can handle structured, un-structured, or semi-structured data.

While only DBMS compatible data are stored in data warehouses, all kinds of data including transactional data, social media data (including audio and video), machinery data, or any DBMS data can be stored and managed using big data technologies.

29. Let's say you have a two-dimensional numpy array called "twod" and you want to split it row-wise into two equal halves. Then, which of these numpy functions would you call on it to do so?

Ans. vsplit(twod,2)

30. Some of the features of digital images in Numpy are given below. Which of these are true?

Ans. n numpy, images can be represented as a 3D matrix where the first two dimensions represent the pixels in the image that are arranged in the form of a grid and the third dimension specifies the number of channels for the image.

A digital image is a multidimensional array and every pixel in a digital image is represented by a number

31. Let's say you have an image that you have split into two equal halves along the x axis. You have stored these two halves of the original image in the variables x1 and x2 respectively. Which numpy function would you use to combine these two halves to reconstruct the original image?

Ans. concatenate((x1,x2),axis=1)

32. Let's say you have a numpy array called "array_1" and you initialize "another array called "array_2" with the help of a following command:

array_2 = array_1.view()

Match the following statements about "array_2" with the correct Boolean value

- Ans. A: array_1 and array_2 contain the same elements (True)
- B: The base for array_2 points to the same object as array_1(True)
- C: array_2 points to the same object as array_1(False)
- D: If we re-assign array_2, then we will end up re-assigning array_1 as well and change its contents (False)

33. Let’s say you have a numpy array called “array_3” and you initialize “array_4” with the help of a following command:

array_4 = array_3.copy()

Match the following statements about “array_4” with the correct Boolean value

- A: If we change a single element of array_4, then the corresponding element in array_3 changes too (False)
- B: array_3 and array_4 contain the same elements (True)
- C: If we re-assign array_4, then we will end up re-assigning array_3 as well and change its contents (False)
- D: Changing the shape of array_4 will change the shape of array_3 as well (False)

34. Let’s say you have a 1-D numpy array called “cubes” consisting of the cubes of the numbers 1,2,3 and so on till 10. What would be the value of the array :

cubes [[[4, 5], [1, 2]]]

ans. [[125, 216] , [8, 27]]

35. Some of the features of Pandas is given below. Which of these are true?

Ans. A particular column of a pandas dataframe can be referenced by its column header.

The column header of a Pandas dataframe can be treated in the same way as the index label of a numpy array

36. Let’s say you imported numpy as np and you have initialized a 1-D array of integers called “array”. What would np.all (x < 50) return?

Ans. This function would return a true boolean value if all the entries in your array are less than 50 and false otherwise

37. Let’s say you have a Pandas dataframe called “phone_data” which contains the data of various phones released in 2018 and their prices. It has the following three columns:

“manufacturer”, “phone name” and “ price”.

You want only the names of all the phones that are priced more than 10,000. Which of these commands can be used to print these values?

Ans. phone_data[phone_data[‘price’] > 10000][‘phone name’]

38. What are the conditions under which broadcasting can take place between two elements in Numpy?

Ans. Broadcasting works when at least one of the elements is a scalar

A smaller array can be broadcast on a larger array only when the corresponding dimensions of the two arrays being operated upon are compatible i.e. when the corresponding dimensions are equal or one of the two dimensions is 1

39. Match the following statements about broadcasting with the correct Boolean value:

- Ans. A:The array [[1, 2] , [3, 4]] and the array [1, 2 ,3] are incompatible with broadcasting(True)
- B:The scalar 10 and the scalar 20 are compatible with broadcasting(True)
- C:The array [[1, 2] , [3, 4]] and the scalar 10 are incompatible with broadcasting(False)
- D:The array [[1, 2] , [3, 4]] and the array [[1], [2]] are compatible with broadcasting(True)

40. In the following Python code, typing which Python command will give the user the CEO of Facebook?

```
import pandas as pd



companies_ceo = {
```

```
                              'Amazon' :  'Jeff Bezos'

                              'Apple' : 'Tim Cook‘,

                              ‘SpaceX’: ‘Elon Musk‘

                              ‘Facebook’: ‘Mark Zuckerberg’

                              ‘Netflix’: ‘Reed Hastings’

                             }

companies_ceo_series= pd.Series(companies_ceo)
```

ans. companies_ceo_series[‘Facebook’]

companies_ceo_series[3]

41. In the following Python code, typing what command will create a DataFrame called “companies_ceo” whose first column has all the entries of the ‘companies’ list and whose second column has all the entries of the ‘ceo’ list, with the column names as the names of the respective variables?

```
import pandas as pd

companies = {

'Amazon'

'Apple'

‘SpaceX’

‘Facebook’

‘Netflix’

                }



ceo = {

'Jeff Bezos'

'Tim Cook‘,
‘Elon Musk‘

‘Mark Zuckerberg’

‘Reed Hastings’

  }
```

Ans. companies_ceo_tuple = list (zip(companies, ceo)) companies_ceo = pd.dataframe(companies_ceo_tuple, columns=[‘companies’, ‘ceo’])

42. What happens when we call the stack () function on a Pandas DataFrame?

Ans. It will create a new DataFrame such that a single row in the original DataFrame is stacked into multiple rows in the new DataFrame depending on the number of columns for each row in the original DataFrame.

43. Which of these correctly match the following libraries in the Numpy ecosystem with what that library is used for?

- Ans. A: Bokeh-Data visualization tool used for large datasets-
- B: Statsmodel-Used to perform statistical operations
- C: Scikit-learn-Specifically meant for machine learning, data mining, and data analysis

44. Match the following statements related to the iloc indexer in Pandas with the correct boolean values.

- Ans. A: The iloc indexer is similar to the loc indexer and can be used to access records located at a particular index in a Pandas DataFrame (True)
- B: The column headers can be passed as input arguments in the form of a string to the iloc function without any errors (False)
- C: When we pass 2:6 as input argument to the iloc function, we get all details of the records located in the second index all the way up to the 5th index of the DataFrame (False)

45. Let’s say you have saved a dataset in a pandas DataFrame called “dataset” which has tons of records and you only want to access the details of the records in only the 5th, 8th and 14th index. Which of these Python commands can you use to do so?

Ans. dataset.loc[[5,8,14],:], dataset.loc[5,8,14]

46. Let’s say you have a pandas DataFrame called “panda” which has 8 rows in total and you want to remove the last row from this DataFrame. Which of these Python commands would you use to do so?

Ans. panda.drop(panda.index[7])

47. Which of these statements related to the pivot function in Pandas is true?

Ans. The combination of the row index and the column header must be unique in order to generate a pivot table

The Pivot function summarizes the details of each column in a DataFrame

48. Match the following statements related to Pandas DataFrames with the correct boolean values.

- Ans. A: All the data within a particular column in a Pandas DataFrame must be of the same data type (True)
- B: Once a Pandas DataFrame has been created, it is not possible to add a new column to this DataFrame (False)
- C: Data in different columns of a Pandas DataFrame cannot be of different data types (False)

49. Match the following statements related to the concept of multiIndex in Pandas with the correct Boolean values

- Ans. A: MultiIndex is useful when we have large datasets where using numeric indexes to refer to each record is unintuitive (True)
- B: MultiIndex lets the user effectively store and manipulate higher dimensional data in a 2-dimensional tabular structure (True)
- C: The MultiIndex for a row is some composite key made up of exactly one column (False)

50. Which of these statements related to the Pandas Series object are true?

Ans. Pandas Series object is similar to a Python list

Once we create a Pandas Series object, an index representing the positions for each of the data points is automatically created for the list.

51. Let’s say you have a pandas DataFrame called “frame” and you want to export this DataFrame along with its index as a CSV file called “data_frame” located in the datasets folder of our workspace.

Ans. frame.to_csv(‘datasets/data_frame.csv’)

52. Consider the following Python code. What command would you use to iterate through the “companies_ceo” DataFrame and print the list of all the CEOs in this DataFrame?

```
import pandas as pd



companies = {

    'Company' : [‘Facebook', 'Apple', ‘Amazon', ‘Netflix’],


                'CEO' : [‘Mark Zuckerberg’, 'Jeff Bezos’, 'Tim Cook'
, ',’Reed Hastings’ ],


                }



companies_ceo = pd.DataFrame(companies)
```

ans. for row in companies_ceo.itertuples(): print(row.CEO),

for row in companies_ceo.iterrows(): print(row[1])

53. Which of the following formats does Pandas not support natively when exporting the contents of a Dataframe?

Ans. JPEG

54. Let’s say you have created a Pandas DataFrame called “unsorted” and you want to sort the contents of this DataFrame column wise in alphabetical order of the header name. Then, which function would you call on the “unsorted” DataFrame to do so?

Ans. unsorted.sort_index(axis=1)

55. Match the following functions that you can call on a Pandas DataFrame correctly with what they do

Ans. Returns a Boolean array containing true or false values and returns the value in a cell as true if it does not contain NaN-.notnull()

All the rows which contain a NaN value in any cell of that row are removed-.dropna()

Every cell in the Dataset which has a NaN value will be replaced with 0-.fillna(0)

Returns a Boolean array containing true or false values and returns the value in a cell as true if it contains NaN-.isnull()

56. Match the following statements related to the .xs function in Pandas DataFrame with their correct Boolean values.

- Ans. A: By default, the .xs function only takes a look at values in the first level index(True)
- B: The. xs function is used when our Pandas DataFrame makes use of a MultiIndex(True)
- C: The .xs function cannot be used to return a cross section of columns(False)

57. Let’s say you have imported Python as pd and have instantiated two DataFrames called “frame_1” and “frame_2” with the exact same schema. What command will you use to combine these two DataFrames into a single DataFrame and

make sure that the combined DataFrame has its own unique index?

Ans. pd.concat([frame_2, frame_1], ignore_index = True)

pd.concat([frame_1, frame_2], ignore_index = True)

58. The ‘how’ argument in the Pandas merge function allows us to specify what kind of join operation we want to perform on the given Pandas DataFrames. What are the valid values that we can give for this argument?

Ans. Left,right,inner,outer

59. Some statements related to working with SQL Databases in Python are given below. Match them with their correct Boolean values.

- Ans. A: The sqlite3 library in Python allows us to create Databases on our local file system(True)
- B: Once we have created a table, we can use sqlite3’s .execute() function to recreate the same table with the same table name so that we have duplicates of a table(False)
- C: All the changes that we make to an SQL database on a Jupyter notebook by connecting with it, will be committed to the database only after we execute sqlite3’s .commit() function(True)

60. Which statement is true about the Kappa architecture?

Ans. The Kappa architecture uses stream processing to manage data flows through a single path

61. Which are the main reasons for using batch processing?

Ans. To run complex algorithms on large datasets which require access to the entire batch.

To join tables in relational databases

62. Which statement is true about the Lambda architecture?

Ans. Data that enters the system is dispatched to two layers in the Lambda architecture: the batch layer and the speed layer.

The Lambda architecture provides fault-tolerance against possible hardware failures and human errors.

63. Place the layers of big data analytics architecture in the correct order from the bottom to the top.

Ans. Data monitoring

Data security

Data storage

Data processing

Data query

Data visualization

64. What are some ways in which big data processing can be performed?

Ans. Batch and stream processing

65. What are the parameters of data ingestion?

Ans. Data format, Data size, Data frequency, Data velocity

66. Which is correct about stream processing?

Ans. Stream processing provides analytical insights before the data storage stage

67. Which statement about data storage systems is correct?

Ans. The Hadoop distributed file system (HDFS) is the primary data storage system used by Hadoop applications

68. Which are the main components of the big data architecture?

Ans. Big data analytics, Big data security, The data model

69. What are the biggest challenges associated with traditional data analytics?

Ans. Scalability, consistency, reliability, efficiency, and maintainability

70. What are some advantages that Spark provides to modern healthcare providers?

Ans. Behind the scenes distributed execution

Convenient workflow fulfillment

A user-friendly API

71. What are some components of Apache Spark?

Ans. Spark SQL

GraphX

72. Which statements are true about resilient distributed datasets (RDDs) and directed acyclic graphs (DAGs)?

Ans. Compared to MapReduce that creates a graph in two stages, Apache Spark can create DAGs that contain many stages

RDD is an immutable (read-only), fundamental collection of elements or items that can be operated on many devices at the same time (parallel processing)

73. As Spark usage grew at Uber, users encountered an increasing number of issues. What were some of those issues/challenges?

Ans. Multiple Spark versions

Multiple compute clusters

74. What are some examples of metrics that Alibaba measures by utilizing Spark?

Ans. Connected components

Degree distribution

75. What are some predominant industries that use Spark today?

Ans. Finance industry

Media and entertainment industry

76. What are the three API types that are compatible with Spark?

Ans. RDD, DataFrame, DataSet

78. What are some of the most important best practices when it comes to using Apache Spark?

Ans. Joining a large and a medium size RDD

Proper tuning

Using the right level of parallelism

79. Which statement is correct about how Spark and Hadoop are different?

Ans. The Hadoop MapReduce model provides a batch engine, hence it is dependent on different engines for other requirements, whereas Spark performs batch, interactive, machine learning and streaming all in the same cluster.

80. Which of the following is a characteristic of a data silo?

Ans. Data is stored in isolation and cannot be combined with other sources

Data is not easily accessible using common tools

Data may be in a raw, native format and not useful unless processed

81. Which of the following are valid data types that can be stored in a data lake?

Ans. Unstructured data

Semi-structured data

Structured data

82. Which of the following is not a characteristic of a data lake?

Ans. Data is not searchable easily

83. Which of the following are challenges involved in designing and building data lakes?

Ans. Data lakes need to work with different data types and sparse and incomplete data

Data lakes need to maintain data security and compliance

Data lakes need to be able to support a huge volume of data

84. Which of the following are valid differences between a traditional relational database and a data warehouse?

Ans. A data warehouse is optimized for read access, a database is optimized for read as well as write access.

A database supports ACID properties and a data warehouse does not

85. Which of the following statements about data lakes and data warehouses are true?

Ans. Data lakes need to maintain security and ensure compliance of the data stored within it

Data warehouses hold fairly structured data optimized for analysis

Data lakes promote shared data stewardship

86. Which of the following is not an example of a data stream?

Ans. Census data stored in a database

87. Which of the following is not a valid service used to ingest data into the AWS cloud?

Ans. Amazon Athena

88. Which of the following correctly defines AWS Glue?

Ans. A single catalog which indexes data from multiple sources to make it searchable

89. Which of the following AWS services can be used to visualize data stored in a data lake on AWS?

Ans. Amazon QuickSight

90. Select the benefits of a distributed system

Ans. fault tolerance

Concurrency

Scalability

91. Arrange the following ETL processing steps in order from the top.

Ans. ingest data from source

message brokering

streaming data engine

long-term storage and analytics

92. Select the characteristics of a NoSQL data store.

Ans. dynamic schema

cluster-friendly

horizontal scaling

93. Match the data management category with its description.

Ans. standardized data, static information-Reference Data Management

organizational data-Master Data Management

dashboards and real-time results-Visualization and Analytics

data warehousing, transformation, extraction-ETL

94. Match the ETL process with its description.

Ans. importing data for computation-Load

selecting raw data-Extract

format and representation shift-Transform

95. Where does the library of job components reside in the Talend Open Studio UI?

Ans. Pallete

96. What high level model is used to get a project overview for ETL jobs in Talend Open Studio?

Ans. Business Model

97. Put the following AI hierarchy steps in pyramid order from the bottom up.

Ans. ETL

Data Exploration

Aggregation

Machine Learning

Deep Learning

98. Reducing the number of fields in the output is an example of what type of partitioning?

Ans. column-based

99. Match the data storage model approach with its descriptions.

Ans. Star Schema-Fact Tables, Dimension Tables

Normalization-Less Redundancy, Standardized

100. Select the features common to interactive reporting tools.

Ans. Filtering, drilling down, sorting

101. Match the data backup methods with their descriptions.

Ans. Gets a backup for all data within the hard drive-Full backup

Gets a backup of data for the past n years-Differential backup

Gets a backup of the data generated or revised since the last full backup-Differential backup

Gets a backup of the data generated or revised since the last backup, regardless of the type of the last backup-incremental backup

102. Match the concepts with their descriptions.

Ans. Raw and unstructured facts, numbers, or figures which convey a message-Data

The ability to ask questions and learn new things-Information

Contextualized, organized, and vetted data that convey some sort of trend or pattern-Information

The ability to use your knowledge and experience to make good decisions and judgements-wisdom

The application of information which is measured by the ability to “do things”-knowledge

103. Ravi wants to create a data visualization to show which parts of his company website are receiving the most clicks and are being most viewed by his viewers. Which data visualization will

provide Dan with a visual that is easy to assimilate and make decisions from?

Ans. Heat Map

104. Match each of the SQL codes with the functions that they perform.

Ans. Creates groups to summarize data-GROUP BY

Lists the columns you want to retrieve-SELECT

Applies filtering logic to your groups-HAVING

Applies filtering logic to limit records in your results-WHERE

Describes how you want your data sorted-ORDER BY

Name of the table to pull the data from-FROM

105. A university professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data is then displayed in a bar chart. What type of data is the university professor collecting?

Ans. Qualitative data

106. Which characteristics do all data migration projects have in common?

Ans. They all require identifying source and target systems.

They all begin with data ingestions and cleansing of the data prior to integration.

They all require performing proper ETL mapping to ensure consistency and compatibility.

107. Alex is in the process of creating a report that displays the results of a survey. Which data type best describes the data that Alex is dealing with?

Ans. Observational

108. ABC Company wants to migrate their CRM data from their current legacy systems into newly purchased web-based CRM software. Place the data migration steps that they should perform in the correct order.

Ans. Planning

Analyzing the data

Design

Implementation

Final migration

Testing

109. Which is not a type of data that would be encountered in an enterprise?

Ans. Logical

110. What is the primary reason for data integration across domains?

Ans. Provide a unified single version of the data

111. In your multi-domain enterprise where the primary function is of a stock market broker and where you need real-time data synchronization, what will be the required style of architecture needed for the data management program?

Ans. Hybrid style

112. Which is not a function of entity resolution?

Ans. Data Propagation

113. Suppose you have a company of 50 employees, and you are writing code for a very specific type of program. There are five vendors that provide you the graphical support and your target

client are very small businesses. Which statement will be true in the domain of data management?

Ans. You do not need a data management program at all

114. Which is not a goal of metadata management?

Ans. Data reference

115. Which correctly describes the difference between static and dynamic data?

Ans. Static data does not require updating whereas dynamic data requires regular updating

116. What will be the best approach of data management for a multi-faceted enterprise with multiple domains?

Ans. A combination of top-down and middle out approaches

117. In your enterprise where you are planning to develop a perfectly aligned system across all the domains, which function will you deem as not necessary for building a truly aligned system?

Ans. Data Quality Program

118. What is the proper order for the levels of management and control from a Measurement of Maturity point of view?

Ans. Unstructured, Structured, Managed, Optimized

119. Which can be considered an advantage of IT systems in preventing security threats?

Ans. Access control is limited to a particular system

120. Which is NOT true about the entity resolution process?

Ans. Application of business and data quality rules is not a step in entity resolution

121. What is the integral meaning of data harmonization?

Ans. Data harmonization means obtaining a single version of the truth

122. Which is NOT a primal rule for data validation?

Ans. Governance practices

123. In building your data governance practice, which will not be an objective for the governance practice?

Ans. Dividing the business

124. Which is not the kind of job that is to be included in the role of a data steward?

Ans. Deciding which information is to be provided to which individual

125. Which are the two factors critical to an organization when considering the regulation of data privacy?

Ans. Storage Networking Industry Association & General Data Protection Regulation

126. In your multi-domain enterprise where there are multiple vendors and multiple resources, which will be a key deciding factor in how you will manage the CRUD?

Ans. Enterprise resource planning

127. Which can be a major problem in companies that have a “Bring your Own Device” policy?

Ans. Leakage of information when the device is used in a public network

128. When implementing your data governance practice in your organization, which is the one factor that will NOT play a major role when considering the implementation?

Ans. Data quality

129. Which is a performance measurement level?

Ans. Strategic measurement

130. Which is NOT a data quality aimed project initiative?

Ans. Data integration

131. Which function is a must for data compliance but not required at all for data management and data governance?

Ans. Inter-relations

132. Which is an essential factor to ensure good data quality?

Ans. Fix data quality issues

133. What are the advantages of a good data compliance strategy?

Ans. Proactive environment, organizational growth, and customer relationships

134. What is meant by reference data?

Ans. Data that is used for classification of other data

135. In your data driven enterprise, how will you ensure that good quality data is being maintained and reconciled across systems?

Ans. Ensure source to target consistency

136. Which is a major solution to address the data governance and compliance issues?

Ans. Knowing the data

137. Which is NOT a continuous improvement method in data governance?

Ans. Overlook achievable governance maturity milestones

138. Which is NOT an advantage of data lakehouse?

Ans. Provides an on-premises data warehouse solution

139. Identify essential table types that we can use to implement a Star schema.

Ans. Fact table, dimension table

140. Which statements about cloud-based data warehouse are true?

Ans. Cloud-based Data warehouses are scalable

Cloud-based Data warehouses are elastic

141. Identify the essential levels of management that require strategic reports.

Ans. Middle management level

Strategic

142. Match the data modelling strategies with their features.

- Ans. A: Ensures the dependencies are properly enforced-Normalization
- B: Applies formal rules to enforce dependencies-Normalization
- C: Splits tables-Normalization
- D: Used on previously normalized databases to increase performance-Denormalization
- E: Adds redundant data-Denormalization
- F: Combines tables-Denormalization

143. Choose the characteristics of weighted reports.

Ans. With weighted reports we get a meaningful subtotal and total

Weighted reports multiply all the facts by weight before aggregating

144. Which processes involved in a data warehouse project are important?

Ans. Data cleansing,ETL

145. Which are essential tasks that we can execute to facilitate business intelligence?

Ans. Load

Extract

146. Which of the following local and global warehouse statements are true?

Ans. Local data warehouse cannot be accessed globally

We can provision a single global repository for a particular domain

147. Select data modelling strategies that we can adopt to create an ER model.

Ans. Normalization,Denormalization

148. Which of the following OLAP statements are true?

Ans. OLAP provides real-time analytical capability

OLAP is a part of the overall Data warehouse implementation

149. Identify the terminologies that we generally use in data warehousing.

Ans. ETL, Dimension model

150. Identify essential features of Strategic Information.

Ans. Preserves data integrity, Time-variant

151. Identify some of the essential features which differentiates a data warehouse from OLTP.

Ans. Data warehouse provides predictive analytical capabilities

Data warehouse stores historical data

152. Identify the essential differences between RDBMS and data lakes?

Ans. RDBMS databases are transaction while data lakes are not transactional

RDBMS databases defines fixed schema while data lake follows no schema design

153. Which statements about Snowflake schemas are true?

Ans. A Snowflake schema is an extension of the Star schema

Application binary interface allows us to contextualizes contracts

154. Identify the essential components of Azure data lake.

Ans. Data factory, SQL server

155. Match the data warehousing solutions with their associated benefits.

- Ans. A: Preferred in banking or government domains (On-premise data wareshouse)
- B: Offers absolute control over security (On-premise data wareshouse)
- C: Provides scalability (On-cloud data warehouse)
- D: Cost effective (On-cloud data warehouse)
- E: Offers better speed and connectivity (On-cloud data warehouse)

156. Specify some of the outcomes of a data warehouse realization.

Ans. Intuitive dashboards

Predictive analytical reports

157. Specify some of the essential logical components of a data warehouse

Ans. Attributes,Entities

158. Which of the following components are provided by Talend to design ETL jobs?

Ans. tMap, tFileInput

159. Which of the following statements correctly defines the characteristics of the Kimball model?

Ans. In the Kimball model the analytical systems can access the data directly

In the Kimball model the analytical systems can access the data directly

160. Identify essential tasks involved in an ETL process.

Ans. Extracting data from diversified sources

Transforming extracted data

161. Which of the following dimensional components differentiates a Dimensional model from an ER model?

Ans. Dimension and fact table.

162. What are some of the essential tasks that we can execute using Talend?

Ans. ETL job designing, Business modelling

163. What are some of the integrated components of a data warehouse?

Ans. Storage, Data Staging

164. What are some of the important tasks that we need to perform to implement a Physical model for a given Logical model?

Ans. Create Foreign keys to establish the relationships among objects

Create tables to represent the entities

165. Select prominent ETL tools that we can use with a data warehouse implementation.

Ans. PowerCenter Informatica,Talend

166. You have an on-premises data warehouse already installed in your organization. In the period following the

COVID pandemic, your business started to grow exponentially, and you were tired of adding more nodes to the physical storage of the data warehouse. You decide to modernize your data warehouse and make it cloud-based. What major advantage will you achieve by modernizing your data warehouse?

Ans. Speed up time to analytics

167. You plan to implement a modern data warehouse solution into your enterprise. You have understood the proper data management and governance issues. You have set up all your domains and data ingestion methods. Now you plan to make a central repository of all your files. What should be the next step for the implementation of the data warehouse solution?

Ans. Selection of the nature of the data

168. You have a very well-run data management program in your organization, which is very secure. You use analytics for making big data driven decisions. In recent months, you realize that whatever decisions you are making based on the analytics are proving to be wrong and harmful for the organization. After consulting with the data analysts and data stewards you conclude that it is due to poor data quality. What should be the next step of action?

Ans. Install a firewall

169. Other than gaining real-time insights of the data, what is another major advantage of streaming analytics?

Ans. Real-time dashboards

170. BigQuery is a cloud-native warehouse that is also a fully managed data warehouse. What is the major advantage of BigQuery over Amazon Redshift that may be a deciding factor for the selection of a data warehouse?

Ans. Access control allows improved data sharing

171. Your organization has an effective sales team that is backed up by analytics that help accelerate the process of sales from the initial contact. One of the primary reasons for its effectiveness is a data input tool that hastens the process of data entry by providing preset suggestions. What are these suggestions commonly known as in data science terminology?

Ans. Reference data

172. You have incorporated the usage of Amazon Redshift in your organization, and you don’t want your data to be corrupted by processing. Therefore, you want the data to be stored in raw format before the processing is done, which is a service offered by Amazon Redshift. What is the key benefit of storing data in raw format?

Ans. Minimal loss of data

173. What is **not** a component of an on-premises data warehouse?

Ans. Processing space

174. You have an automobile company that helps sort out vehicles and make monthly sales versus expenditure reports. What is the best way to handle the data for centrally storing it?

Ans. Batch processing

175. In a candlestick chart, you see the share price of your company falling. You have implemented a streaming analytical tool that helps in analyzing and dashboarding the data as it is produced. You realize that the candlesticks pattern has consolidated and is not responding well to the influx of data. What is the source of this problem?

Ans. Server downtime

176. In your enterprise, customer information needs to be readily available to indicate whether the information given by

customers is valid or not, and to see the potency of a positive deal. Which factors would you consider a priority when selecting the appropriate data pipeline tool?

Ans. Batch-wise vs. real-time ingestion and analytics

177. What are the major advantages of a cloud warehouse solution over an on-premises data warehouse solution?

Ans. Less worry about storage

Low cost

178. What are the two different data pipeline tools that address specific job roles?

Ans. Data engineers and analysts

179. Place the steps for a typical Azure Databricks warehouse in the correct order.

1. Ans. Ingest
2. Store
3. Prep and train
4. Model and serve

180. The reason why Azure Databricks is so easy to use is because it is universal and is integrated with Microsoft's server for better parsing of information. Which platform has the same source of origin as Databricks, which gives it an analytical advantage over other platforms?

Ans. Apache spark

181. What are the major disadvantages of Snowflake that might be troublesome for a few companies that seek data categorization?

Ans. Fewer options with geospatial space

No option for un-structured data

182. While setting up an integrated data pipeline for your enterprise to facilitate data ingestion in the warehouse, what should you place more emphasis on from a business perspective?

Ans. Analytics and business intelligence

183. Suppose you have a full-blown data management program with a well-running data warehouse and an optimum data pipeline tool that facilitates data transformation and transmission. While measuring the maturity of the data pipeline tool, what will be the sole factor that will determine the efficiency of the data pipeline tool?

Ans. Reliability and scalability

184. What is the one difference that separates the model of the Snowflake data warehouse from all the other data warehouse solutions?

Ans. Hybrid model

185. What is **not** a design component of a data pipeline?

Ans. Data integration

186. Select the main dependency that has to be installed for Talend to be installed.

Ans. Java

187. Select the supported OSs by Talend Open Studio.

Ans. MacOS

Windows 64

188. Select the main parts of the default UI of Talend Open Studio.

Ans. Repository, Palette

189. When installing MySQL relational Database to be used with Talend, select the configuration to be performed once the installation of the components is complete.

Ans. MySQL root password

MySQL server port

190. Select the folder that contains all the project information for a job that is exported from Talend studio.

Ans. Process

191. Rank the steps required to for any job in Talend studio.

Ans. Create job in the repository

Configure the components properties

Add components from palette to the design space

Run the job

192. Select the correct description of Talend Metadata Bridge

Ans. Synchronize metadata across data pipelines

193. Match each description with the variable in Talend studio.

- Ans. A: studio provisions through components used in jobs integration (Studio global variables)
- B: Ad-hoc variables that can be configured in jobs (User defined global variables)
- C: execute jobs with parameters for different environments (Job contexts variables)

194. Select the correct differences between XML attributes and elements.

Ans. Elements can contain tree structure, Elements can have multiple values

195. Select the component used to generate an XML file from a CSV file in Talend studio.

Ans. tFileOutputXML

196. Select the exit code value the signifies the successful completion of a job in Talend studio.

Ans. 0

197. Select the tMap Component description in Talend studio.

Ans. congregate input data to output data

198. Select the option to be enabled to allow the specification of 2 schemas for an XML input file in Talend studio.

Ans. Enable XPath in column "Schema XPath loop" but lose the order

199. In order to generate a complex XML file where data is specified using attributes of elements and elements trees in Talend studio, which component allows such.

Ans. tAdvancedFileOutputXML

200. Select the component used to perform lookup data in Talend studio.

Ans. tJoin

201. Select the component access a MySQL database in Talend studio.

Ans. tMysqlInput

202. Select the tool that allows specifying the relation of multiple tables as data sources when reading data from a database as input in Talend studio.

Ans. SQL Builder

203. Match the attribute value with its attribute that will only Add new records or modify existing ones without modifying the table

structure or other records already exist in the table when writing data to a database table in Talend studio. Two options are invalid.

Ans. Action on data-Insert or update

Action on table-Default

204. Select the component that allows updating data in a database in Talend studio.

Ans. tMySQLRow

205. Select the components and concepts to facilitates accepting as input multiple databases in Talend studio.

Ans. tMap, LookUp

206. Select the component used to combine multiple database records to a single records in Talend studio.

Ans. tDenormalize

207. Which of these statements accurately describes the Unique Match Join match model in the Map Editor?

Ans. Set as default when configuring an explicit Join

Last match is considered and passed to the output

208. You are building an expression in Map Editor for a column in which you want to pad the row1.CustomerID string with leading zeroes up to maximum length of 6 characters. Which code can you use to accomplish this for you?

Ans. String.format("$06d",row1.CustomerID)

209. You have Map Editor open for a tMap object for which you are mapping databases objects. You click the Join Model for a table Join property. Which options appear in the Options dialog that appears?

Ans. Inner Join, Left Outer Join

210. You are building a filter expression in Map Editor for a column in which you want to filter the product name, prd.name to equal

"Turbo Widgets", and you want the transaction quantity, tx.qty to be greater than 100. Which code can you use to accomplish this for you?

Ans. prd.name.equals("Turbo Widgets")&&tx.qty>100

211. Which common FTP operations are supported by Talend Open Studio components available in the Palette?

Ans. Delete

File Exist

Put

212. On the Component tab displayed for the tSortRow component, when you click to add criteria to the Criteria table, Talend automatically populates the column values with defaults. In the "sort num or alpha?" column Talend has chosen num by default for customer_id as displayed. Which other values are available for the "sort num or alpha?" column when you click to open the dropdown list for that column?

Ans. Alpha,date

213. You are using a tExtractDelimitedFields component to split the Address2 field in the delimited file as displayed. What must you specify as the Field separator property for the tExtractDelimitedFields component to properly split the Address2 field?

```
ID;First_Name;Last_Name;Address1;Address2;Country

1;Claudia;Sand;10000 Main Dr NW;New York, NY;USA

2;Max;Bigot;60000 My St;Nashua, NH;USA

3;Rick;Tailleur;200000 Younge St;Toronto, ON;Canada

4;Noémie;Miller;500000 St. Catherines St Ouest;Montreal, QC;Canada

5;Catherine;Reilly;100000 Main St SW;Boston, MA;USA
```

Ans. "," (a comma)

214. When configuring the properties for a tReplace component, you can optionally click the Advanced mode checkbox. Doing so allows you to specify what type of expression as the Pattern to search for?

Ans. Regular Expression

215. When configuring the properties for a tAggregateRow component. You are going to Group by the customer_id field in order to aggregate the sales on a customer_id basis, so that in the resulting output file you will have one row for each customer_id with aggregated sales figures. Which Function value must you choose when configuring the Operations table for this tAggregateRow component?

Ans. Sum

216. You are using a tNormalize component to normalize the category field in the delimited file as displayed. What must you specify as the Item separator property for the tNormalize component to properly normalize the category field?

```
product_id,product_name,category,unit_price

1,Regular Widget,Normally Aspirated;Low Price;Highly Reliable,125

2,Super Widget,Super;Medium Price;Very Reliable,250

3,Turbo Widget,Turbo;Medium Price;Very Reliable,275

4,S/T Widget,Super;Turbo;Premium Price;Reliable,425

5,Hybrid Widget,Hybrid;Normally Aspirated;Premium Price;Reliable,425
```

Ans. ";" (a semi-colon)

MIXED MCQS WITH ANSWERS

Q.1 The results of a hive query can be stored as
Local File
HDFS file
Both the above
Can not be stored

Q.2 If the database contains some tables then it can be forced to drop without dropping the tables by using the keyword
RESTRICT
OVERWRITE
F DROP
CASCADE

Q.3 Users can pass configuration information to the SerDe using
SET SERDEPRPERTIES
WITH SERDEPRPERTIES
BY SERDEPRPERTIES
CONFIG SERDEPRPERTIES

Q.4 The property set to run hive in local mode as true so that it runs without creating a mapreduce job is
hive.exec.mode.local.auto
hive.exec.mode.local.override
hive.exec.mode.local.settings
Hive.exec.mode.local.config

Q.5 Which kind of keys(CONSTRAINTS) Hive can have?
Primary Keys
Foreign Keys
Unique Keys
None of the above

Q.6 What is the disadvantage of using too many partitions in Hive tables?
It slows down the namenode
Storage space is wasted
Join queries become slow
All of the above

Q.7 The default delimiter in hive to separate the element in STRUCT is
'\001'
'\oo2'
'\oo3'
'\oo4'

Q.8 By default when a database is dropped in Hive
The tables are also deleted
The directory is deleted if there are no tables
The HDFS blocks are formatted
None of the above

Q.9 The main advantage of creating table partition is
Effective storage memory utilization
Faster query performance
Less RAM required by namenode
Simpler query syntax

Q.10 If the schema of the table does not match with the data types present in the file containing the table then Hive
Automatically drops the file
Automatically corrects the data
Reports Null values for mismatched data
Does not allow any query to run on the table

Q.11 A view in Hive can be seen by using
SHOW TABLES
SHOW VIEWS
DESCRIBE VIEWS
VIEW VIEWS

Q.12 If an Index is dropped then
The underlying table is also dropped
The directory containing the index is deleted
The underlying table is not dropped
Error is thrown by hive

Q.13 Which file controls the logging of Mapreduce Tasks?
hive-log4j.properties

hive-exec-log4j.properties
hive-cli-log4j.properties
hive-create-log4j.properties

Q.14 What Hive can not offer
Storing data in tables and columns
Online transaction processing
Handling date time data
Partitioning stored data

Q.15 To see the partitions keys present in a Hive table the command used is
Describe
Describe extended
Show
Show extended

Q.1 For optimizing join of three tables, the largest sized tables should be placed as
The first table in the join clause
Second table in the join clause
Third table in the join clause
Does not matter

Q.2 Which of the following hint is used to optimize the join queries
/* joinlast(table_name) */
/* joinfirst(table_name) */
/* streamtable(table_name) */
/* cacheable(table_name) */

Q.3 Calling a unix bash script inside a Hive Query is an example of
Hive Pipeline
Hive Caching
Hive Forking
Hive Streaming

Q.4 Hive uses __________ for logging.
logj4
log4l
log4i
Log4j

Q.5 HiveServer2 introduced in Hive 0.11 has a new CLI called
BeeLine

SqlLine
HiveLine
CLilLine

Q.6 In which mode HiveServer2 only accepts valid Thrift calls.
Remote
HTTP
Embedded
Interactive

Q.7 Which of the following data type is supported by Hive?
map
record
string
Enum

Q.8 Which of the following is not a complex data type in Hive?
Matrix
Array
Map
STRUCT

Q.9 Each database created in hive is stored as
A file
A directory
A HDFS block
A jar file

Q.10 When a partition is archived in Hive it
Reduces space through compression
Reduces the length of records
Reduces the number of files stored
Reduces the block size

Q.11 When a Hive query joins 3 tables, How many mapreduce jobs will be started?
0
1
2
3

Q.12 The reverse() function reverses a string passed to it in a Hive query. This is an example of
Standard UDF
Aggregate UDF
Table Generating UDF

None of the above

Q.13 Hive can be accessed remotely by using programs written in C++, Ruby etc, over a single port. This is achieved by using
HiveServer
HiveMetaStore
HiveWeb
Hive Streaming

Q.14 The thrift service component in hive is used for
Moving hive data files between different servers
Use multiple hive versions
Submit hive queries from a remote client
Installing hive

Q.15 The query "SHOW DATABASE LIKE 'h.*' ; gives the output with database name
Containing h in their name
Starting with h
Ending with h
Containing 'h.'

Q.1 The tables created in hive are stored as
A file under the database directory
A subdirectory under the database directory
A .java file present in the database directory
A HDFS block containing the database directory

Q.2 Using the ALTER DATABASE command in an database you can change the
Database name
dbproperties
Database creation time
Directory where the database is stored

Q.3 In Hive you can copy
The schema without the data
The data without the schema
Both schema and its data
Neither the schema nor its data

Q.4 The drawback of managed tables in hive is
They are always stored under default directory
They cannot grow bigger than a fixed size of 100GB
They can never be dropped
They cannot be shared with other applications

Q.5 On dropping a managed table
The schema gets dropped without dropping the data
The data gets dropped without dropping the schema
An error is thrown
Both the schema and the data is dropped

Q.6 On dropping a external table
The schema gets dropped without dropping the data
The data gets dropped without dropping the schema
An error is thrown
Both the schema and the data is dropped

Q.7 The 2 default TBLPROPERTIES added by hive when a hive table is created is
hive_version and last_modified by
last_modified_by and last_modified_time
last_modified_time and hive_version
last_modified_by and table_location

Q.8 The partition of an Indexed table is dropped. then,
Corresponding partition from all indexes are dropped.
No indexes are dropped
Indexes refresh themselves automatically
Error is shown asking to first drop the indexes

Q.9 What is Hive used as?
Hadoop query engine
MapReduce wrapper
Hadoop SQL interface
All of the above

Q.10 Which of the following is true for Hive?
Hive is the database of Hadoop
Hive supports schema checking
Hive doesn’t allow row level updates
Hive can replace an OLTP system

Q.11 Managed tables in Hive:
Can load the data only from HDFS
Can load the data only from local file system
Are useful for enterprise wide data
Are Managed by Hive for their data and metadata

Q.12 Can the default “Hive Metastore” be used by multiple users (processes) at the same time?

Yes
[✓ No]

Q.13 Which of the following query displays the name of the database, the root location on the file system and comments if any.
Describe extended
Show
[✓ Describe]
Show extended

Q.14 Which of the following gives the details of the database or schema in a detailed manner.
[✓ Describe extended]
Show
Describe
Show extended

Q.15 Which of the following is the join in Hive?
Join
Full outer join
Right outer join
All of the above

Q.1 Is it possible to change the default location of Managed Tables in Hive
[✓ Yes]
No

Q.2 Which among the following command is used to change the settings within Hive session
RESET
[✓ SET]

Q.3 How to change the column data type in Hive
[✓ ALTER and CHANGE]
ALTER
CHANGE

Q.4 Which of the following is the data types in Hive
ARRAY
STRUCT
MAP
[✓ All the above]

Q.5 Which of the following is the Key components of Hive Architecture
User Interface

Metastore
Driver
All of the above

Q.6 Are multiline comments supported in Hive?
Yes
No

Q.7 Can we run UNIX shell commands from Hive?
Yes
No

Q.8 Which of the following is the commonly used Hive services
Command Line Interface (cli)
Hive Web Interface (hwi)
HiveServer (hiveserver)
All of the above

Q.9 Explode in Hive is used to convert complex data types into desired table formats.
True
False

Q.10 Is it possible to overwrite Hadoop MapReduce configuration in Hive?
Yes
No

Q.11 Point out the correct statement
Hive is not a relational database, but a query engine that supports the parts of SQL
Hive is a relational database with SQL support
Pig is a relational database with SQL support
None of the above

Q.12 Which of the following is used to analyse data stored in Hadoop cluster using SQL like query
Mahoot
Hive
Pig
All of the above

Q.13 If an Index is dropped then
The directory containing the index is deleted
The underlying table is not dropped
The underlying table is also dropped
Error is thrown by hive

Q.14 If the schema of the table does not match with the data types present in the file containing the table then Hive
Automatically drops the file
Automatically corrects the data
Reports Null values for mismatched data
Does not allow any query to run on the table

Q.15 By default when a database is dropped in Hive

The tables are also deleted
The directory is deleted if there are no tables
The HDFS blocks are formatted
None of the above

Q.1 The source of HDFS architecture in Hadoop originated as
Google distributed filesystem
Yahoo distributed filesystem
Facebook distributed filesystem
Facebook distributed filesystem

Q.2 What is HDFS?
Storage layer
Batch processing engine
Resource management layer
All of the above

Q.3 Which among the following command is used to copy a directory from one node to another in HDFS?
rcp
distcp
dcp
Drcp

Q.4 Which utility is used for checking the health of an HDFS file system?
fsck
fchk
fsch
Fcks

Q.5 Which among the following is the correct statement
Datanode manage file system namespace
Namenode stores metadata
NameNode stores actual data

All of the above

Q.6 What is default replication factor?
1
2
3
5

Q.7 The namenode knows that the data node is active using a mechanism known as
Active pulse
Data pulse
Heartbeats
H-signal

Q.8 What is the default size of HDFS Data Block?
16MB
32MB
64MB
128MB

Q.9 What is HDFS Block in Hadoop?
It is the logical representation of data
It is the physical representation of data
Both the above
None of the above

Q.10 Which of the following is the correct statement?
DataNode is the slave/worker node and holds the user data in the form of Data Blocks
Each incoming file is broken into 32 MB by default
NameNode stores user data in the form of Data Blocks
None of the above

Q.11 The need for data replication can arise in various scenarios like
Replication Factor is changed
DataNode goes down
Data Blocks get corrupted
All of the above

Q.12 A file in HDFS that is smaller than a single block size
Cannot be stored in HDFS
Occupies the full block's size.
Occupies only the size it needs and not the full block
Can span over multiple blocks

Q.13 Which among the following are the duties of the NameNodes
Manage file system namespace
It is responsible for storing actual data
Perform read-write operation as per request for the clients
None of the above
Q.14 If the IP address or hostname of a data node changes
The namenode updates the mapping between file name and block name
The data in that data node is lost forever
The namenode need not update mapping between file name and block name
There namenode has to be restarted

Q.15 For the frequently accessed HDFS files the blocks are cached in
The memory of the data node
In the memory of the namenode
Both the above

Q.16 Which scenario demands highest bandwidth for data transfer between nodes
Different nodes on the same rack
Nodes on different racks in the same data center.
Nodes in different data centers
Data on the same node.

Q.17 When a client contacts the namenode for accessing a file, the namenode responds with
Size of the file requested
Block ID of the file requested
Block ID and hostname of all the data nodes containing that block

Q.18 In HDFS the files cannot be
Read
Deleted
Executed
Archived

Q.19 Which among the following is the duties of the Data Nodes
Manage file system namespace
Stores meta-data
Regulates client’s access to files
Perform read-write operation as per request for the clients

Q.20 NameNode and DataNode do communicate using
Active pulse
Heartbeats
H-signal

Q.1 HDFS is inspired by which of following Google project
BigTable
GFS
MapReduce

Q.2 In HDFS, data node sends frequent heartbeats to name node
True
False

Q.3 Clients connect to ________ for I/O
NameNode
DataNode

Q.4 For reading/writing data to/from HDFS, clients first connect to
NameNode
Checkpoint Node
DataNode

Q.5 The namenode loses its only copy of fsimage file. We can recover this from which of the following?
Secondary Namenode
It can not be recovered
Datanodes
Checkpoint Node

Q.6 When a backup node is used in a cluster there is no need of which of the following?
Standby node
Check point node
Secondary data node
Secondary name node

Q.7 The HDFS command to create the copy of a file from a local system is which of the following?
copyFromLocal
CopyFromLocal
CopyLocal
Copyfromlocal

Q.8 HDFS provides a command line interface called __________ used to interact with HDFS
HDFS Shell
FS Shell
DFS Shell

Q.9 Which of the following is the daemon of HDFS?

Secondary namenode
Node manager
Resource manager
Q.10 Which of the following stores metadata?
DataNode
NameNode
Secondary Data Node

Q.11 Which of the following statement is true about Secondary NameNode
It store the modified FsImage into persistent storage
It stores the merged FsImage with EditLogs back to active namenode.
It does not store the modified FsImage into persistent storage

Q.12 Which of the following is the core component of HDFS
Node Manager
DataNode
Resource Manager

Q.13 Which statement is true about NameNode
It is the slave node that stores actual data
It is the Master node that stores actual data
It is the slave node that stores metadata
It is the Master node that stores metadata

Q.14 Which of the following is true about metadata
Metadata shows the structure of HDFS directories/files
Metadata contain information like number of blocks, their location, replicas
FsImage & EditLogs are metadata files
All of the above

Q.15 Which statement is true about DataNode
It is the slave node that stores actual data
It is the slave node that stores metadata
It is the Master node that stores actual data
It is the actual worker node that stores metadata

Q.16 What is Secondary NameNode is the Backup node
True
False

Q.17 Which of the following is NOT a type of metadata in NameNode?
List of files
Block locations of files
No. of file records

File access control information

Q.18 What is the major advantages of storing data in block size 128MB
It saves disk seek time
It saves disk processing time
It saves disk access time
It saves disk latency time

Q.19 HDFS performs replication, although it results in data redundancy?
True
False

Q.20 Which of the following is the Single Point of Failure
NameNode
Secondary NameNode
DataNode

Q.1 HDFS data blocks can be read in parallel.
True
False

Q.2 In the local disk of the namenode the files which are stored persistently are
Namespace image and edit log
Block locations and namespace image
Edit log and block locations
Namespace image, edit log and block locations.

Q.3 HDFS stands for
Highly distributed file system
Hadoop directed file system
Highly distributed file shell
Hadoop distributed file system

Q.4 Which of the following algorithms available for Erasure Coding
XOR Algorithm
Secondary node
Both the above

Q.5 Which of the following property is used to change heartbeat interval
dfs.heartbeat.interval
dfs.heartbeat.interval.change
dfs.heartbeat.change.interval

Q.6 In which configuration file heartbeat interval is changed

core-site.xml
mapred-site.xml
yarn-site.xml
hdfs-site.xml
Q.7 One can copy a file into HDFS with a different block size to that of existing block size configuration by using
-dfs.block.size=block_size
-Ddfs.blocksize=block_size
-Ddfs.blocksize.copy=block_size

Q.8 You can change the replication factor on a per-file basis using a command
hadoop fs –setrep –w 3 / file_location
hadoop dfs –setrep –w 3 / file_location
hadoop fs –set.rep –w 3 / file_location
hadoop dfs –set.rep –w 3 / file_location

Q.9 Is Namenode machine same as DataNode machine as in terms of hardware?
Yes
No
Q.10 Which command is used to format HDFS
bin hdfs –format
bin/hadoop hdfs.namenode –format
bin namenode.hdfs –format
bin/hadoop namenode –format

Q.11 You can change the replication factor for all the files in a directory using command
hadoop dfs –setrep –w 3 –R / directoey_location
hadoop fs –setrep –w 3 –R / directoey_location
hadoop fs –set.rep –w 3 –R / directoey_location
hadoop dfs –set.rep –w 3 –R / directoey_location

Q.12 Which of the following are the features of HDFS
Fault Tolerance
High Availability
Replication
All of the above
Q.13 For 129 MB file how many Blocks will be created
1
2
3
4

Q.14 Which statement is true about passive NameNode in Hadoop
It is a standby namenode

It simply acts as a slave
Provide a fast failover
All of the above
Q.15 If Active NameNode fails, then which of the following Node takes all the responsibility of active node
Standby node
Secondary node
Backup node

Q.16 Can NameNode be commodity hardware?
Yes
No

Q.17 What is the importance of dfs.namenode.name.dir in HDFS?
Contains the fsimage file for namenode
Contains the edit logs for namenode

Q.18 What is the port number for NameNode
50030
50060
50070
50071

Q.19 Which of the following is/are correct?

a. NameNode is the SPOF in Hadoop 1.x

b. NameNode is the SPOF in Hadoop 2.x

c. NameNode keeps the image of the file system also
Both a and b
Both b and c
Both a and c
None of the above

Q.20 NameNode tries to keep the first copy of data nearest to the client machine
Always true
Always false
True if the client machine is the part of the cluster
True if the client machine is not the part of the cluster

Q.1 Which of the following is true about MapReduce?
It provides the resource management

An open source data warehouse system for querying and analyzing large datasets stored in hadoop files
Data processing layer of hadoop

Q.2 What happens if a number of reducers are set to 0?
Reduce-only job take place
Map-only job take place
Reducer output will be the final output

Q.3 Which of the following maps input key/value pairs to a set of intermediate key/value pairs.
Mapper
Reducer
Both Mapper and Reducer

Q.4 The number of maps is usually driven by the total size of
Tasks
Inputs
Outputs

Q.5 Which of the following is the default Partitioner for partitioning keyspace.
HashPar
Partitioner
HashPartitioner

Q.6 Which among the following is the programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.
MapReduce
HDFS
Pig

Q.7 Which of the following is the default InputFormat, which treats each value of input a new value and the associated key is byte offset.
FileInputFormat
TextInputFormat
KeyValueTextInputFormat

Q.8 Which among the following controls the partitioning of the keys of the intermediate map-outputs.
RecordReader
Combiner
Partitioner

Q.9 Which of the following phases occur simultaneously
Shuffle and Map

Shuffle and Sort
Reduce and Sort

Q.10 Which statements is true about a key, value pairs?
Key class must implement Writable
Key class must implement WritableComparable.
Value class must implement WritableComparable.
Value class must extend WritableComparable

Q.11 In HDFS put command is used to
Copy files from local file system to HDFS.
Copy files or directories from local file system to HDFS.
Copy files from from HDFS to local filesystem.
Copy files or directories from HDFS to local filesystem.

Q.12 Which among the following is the default OutputFormat?
SequenceFileOutputFormat
LazyOutputFormat
DBOutputFormat
TextOutputFormat

Q.13 Which of the following is not an input format in Hadoop?
ByteInputFormat
TextInputFormat
SequenceFileInputFormat
KeyValueInputFormat

Q.14 What is the correct sequence of data flow in MapReduce?

a. InputFormat

b. Mapper

c. Combiner

d. Reducer

e. Partitioner

f. OutputFormat
abcdfe
abcedf
acdefb
abcdef

Q.15 Which of the following is not the phase of Reducer?
Shuffle
Sort
Map
Reduce

Q.16 Mappers sorted output is Input to the
Reducer
Mapper
Shuffle

Q.17 How to disable the reduce step
set conf.setNumreduceTasks(0)
set job.setNumreduceTasks(0)
set job.setNumreduceTasks()=0

Q.18 Keys from the output of shuffle and sort implement which of the following interface?
Writable
WritableComparable
Configurable
ComparableWritable

Q.19 Put the following phases of a MapReduce program in the order that they execute?

a. Partitioner

b. Mapper

c. Combiner

d. Shuffle/Sort
Mapper Partitioner Shuffle/Sort Combiner
Mapper Partitioner Combiner Shuffle/Sort
Mapper Shuffle/Sort Combiner Partitioner
Mapper Combiner Partitioner Shuffle/Sort

Q.20 Which among the following generate an intermediate key-value pair
Reducer
Mapper
Combiner
Partitioner

Q.1 In MapReduce output of Mapper is stored on

Local Disk
HDFS
In-memory
Q.2 Which of the following is true about combiner
Reduces network congestion
Execution of combiner is not guaranteed
Both the above

Q.3 Which of the following is called Mini-reduce
Combiner
Partitioner
Reducer

Q.4 By default RecordReader uses which InputFormat for converting data into key-value pairs.
FileInputFormat
TextInputFormat
KeyValueTextInputFormat
SequenceFileInputFormat

Q.5 What statement is true about InputSplit?
Logical representation of data
Physical representation of data
Both the above

Q.6 Which of the following is not an input format in Hadoop?
TextInputFormat
KeyValueTextInputFormat
SequenceFileTextInputFormat
SequenceFileInputFormat

Q.7 In which InputFormat tab character (‘/t’) is used
KeyValueTextInputFormat
TextInputFormat
FileInputFormat
SequenceFileInputFormat

Q.8 Aggregation cannot be done in Mapper in MapReduce
True
False

Q.9 Shuffling and sorting phase in Hadoop occurs
First shuffling then sorting
Simultaneously
First sorting then shuffling

Q.10 What is the correct sequence of data flow in MapReduce

a. InputFormat

b. Shuffling-Sorting

c. Mapper

d. Reducer

e. OutputFormat
abcde
acdbe
acbde
bacde

Q.11 Which among the following is true about SequenceFileInputFormat
Key- byte offset. Value- It is the contents of the line
Key- Everything up to tab character. Value- Remaining part of the line after tab character
Key and value- Both are user-defined
None of the above

Q.12 Suppose file content is- "on the top of the Tree" then, what is key and value?
Key- 0 Value- on the top of the Tree
Key-on Value- the top of the Tree
Key-1 Value- on the top of the tree
Key-on the Value- top of the Tree

Q.13 Which among following is true about KeyValueTextInputFormat
Key- byte offset. Value- It is the contents of the line
Key- Everything up to tab character. Value- Remaining part of the line after tab character
Key and value- Both are user-defined
None of the above

Q.14 Which is key and value in TextInputFormat
Key- byte offset Value- It is the contents of the line
Key- Everything up to tab character Value- Remaining part of the line after tab character
Key and value- Both are user-defined
None of the above

Q.15 Counters validate that:
Number of bytes read-write within map/reduce job is correct or not
Number of tasks launches and successfully run in map/reduce job is correct or not

Amount of CPU and memory consumed is appropriate for our job and cluster nodes.
All of the above

Q.16 Which of the following is Built-In Counters in Hadoop?
FileSystem Counters
FileInputFormat Counters
FileOutputFormat counters
All of the above

Q.17 Which of the following command is used to set the number of reducers for the job
Job.confNumreduceTasks(int)
Job.setNumreduceTasks(int)
Job.setNumreduceTasks()
Job.confNumreduceTasks()

Q.18 Which of the following is not an output format in Hadoop?
TextoutputFormat
ByteoutputFormat
SequenceFileOutputFormat
DBOutputFormat

Q.19 Which of the following OutputFormat is used for writing to relational databases and Hbase?
TextoutputFormat
DBOutputFormat
MapFileOutputFormat
SequenceFileOutputFormat

Q.20 RecordReader handle record boundaries in Text files or Sequence files in MapReduce
True
False

Q.1 Is it mandatory to set input and output type/format in Hadoop MapReduce?
Yes
No

Q.2 HashPartitioner uses which method to determine, to which partition a given (key, value) pair will be sent
hash.Code()

hashcodepatition()
hashCode()
hashpatition()

Q.3 Which of the following permits to use multiple Mapper classes within a single Map task
Identity Mapper
Chain Mapper
Both the above

Q.4 Identity Reducer is the default Reducer provided by Hadoop
True
False

Q.5 The parameters for Mappers are:
text (input)
LongWritable(input)
text (intermediate output)
All of the above

Q.6 The parameters for Reducers are:
text (intermediate output)
IntWritable (intermediate output)
IntWritable (final output)
All of the above

Q.7 Hadoop framework invokes Splitting of the file by running which of the following method
getSplit()
get.InputSplit()
getInputSplit()
getInputSplit(int)

Q.8 For 514 MB file how many InputSplit will be created
4
5
6
10

Q.9 Which among the following is used to provide multiple inputs to Hadoop?
MultipleInputs class
MultipleInputFormat
FileInputFormat
DBInputFormat

Q.10 Which statement is true about EditLogs

Contains the entire filesystem namespace
Contains all recent modifications made to the file system of most recent FsImage
Contains a serialized form of all the directories
All of the above
Q.11 Identity Mapper is the default Mapper in Hadoop
True
False

Q.12 Which among following class is responsible for converting inputs to key-value Pairs?
FileInputFormat
InputSplit
RecordReader
Mapper

Q.13 What is the most common OutputFormat in Hadoop
TextOutputFormat
SequenceFileOutputFormat
MapFileOutputFormat
All of the above

Q.14 Is it mandatory to set input and output type/format in MapReduce?
Yes
No

Q.15 Which of the following is used to set mappers for MapReduce jobs?
job.setNumMaptasks()
job.setNum.Maptasks()
job.setNumMap.tasks()
job.setNumMap()

Q.16 Which of the following is used set reducers for MapReduce jobs?
job.setNumreduce.tasks()
job.setNum.Maptasks()
job.setNumreduceTasks()
job.setNumreduce()

Q.17 Which of the following is the correct sequence of MapReduce flow?
Map>Combine> Reduce
Combine>Reduce>Map
Map>Reduce>Combine
Reduce>Combine>Map

Q.18 The process by which the system performs the sort and transfers the map outputs to the reducer as inputs is known as

Sorting
Shuffling
Aggregation
Summation
Q.19 Which of the following is used to make sure that all the value of a single key goes to the same reducer
Combiner
Reducer
Partitioner
Mapper

Q.20 Which of the following is the default Partitioner for partitioning keyspace.
HashPar
Partitioner
HashPartitioner
None of the above

Q.1 Spark is developed in which language
Java
Scala
Python
R

Q.2 In Spark Streaming the data can be from what all sources?
Kafka
Flume
Kinesis
All of the above

Q.3 Apache Spark has API's in
Java
Scala
Python
All of the above

Q.4 Which of the following is not a component of Spark Ecosystem?
Sqoop
GraphX
MLlib
BlinkDB

Q.5 The basic abstraction of Spark Streaming is
Dstream

RDD
Shared Variable
None of the above

Q.6 Which of the following algorithm is not present in MLlib?
Streaming Linear Regression
Streaming KMeans
[✓ Tanimoto distance]
None of the above

Q.7 Dstream internally is
[✓ Continuous Stream of RDD]
Continuous Stream of DataFrame
Continuous Stream of DataSet
None of the above

Q.8 Can we add or setup new string computation after SparkContext starts
Yes
[✓ No]

Q.9 Which of the following is not the feature of Spark?
Supports in-memory computation
Fault-tolerance
[✓ It is cost efficient]
Compatible with other file storage system

Q.10 Which is the abstraction of Apache Spark?
Shared Variable
RDD
[✓ Both the above]

Q.11 What are the parameters defined to specify window operation
[✓ Window length, sliding interval]
State size, window length
State size, sliding interval
None of the above

Q.12 Which of the following is not output operation on DStream
SaveAsTextFiles
ForeachRDD
SaveAsHadoopFiles
[✓ ReduceByKeyAndWindow]

Q.13 Dataset was introduced in which Spark release?

Spark 1.6
Spark 1.4.0
Spark 2.1.0
Spark 1.1
Q.14 Which Cluster Manager do Spark Support?
Standalone Cluster Manager
MESOS
YARN
All of the above

Q.15 The default storage level of cache() is?
MEMORY_ONLY
MEMORY_AND_DISK
DISK_ONLY
MEMORY_ONLY_SER

Q.16 Which is not a component on the top of Spark Core?
Spark RDD
Spark Streaming
MLlib
None of the above

Q.17 Apache Spark was made open-source in which year?
2010
2011
2008
2009

Q.18 In addition to stream processing jobs, what all functionality do Spark provides?
Machine learning
Graph processing
Batch processing
All of the above

Q.19 Is Spark included in every major distribution of Hadoop?
Yes
No

Q.20 Which of the following is not true for Hadoop and Spark?
Both are data processing platforms
Both are cluster computing environments
Both have their own file system
Both use open source APIs to link between different tools

Q.1 How much faster can Apache Spark potentially run batch-processing programs when processed in memory than MapReduce can?
10 times faster
20 times faster
100 times faster
200 times faster

Q.2 Which of the following provide the Spark Core’s fast scheduling capability to perform streaming analytics.
RDD
GraphX
Spark Streaming
Spark R

Q.3 Which of the following is the reason for Spark being Speedy than MapReduce?
DAG execution engine and in-memory computation
Support for different language APIs like Scala, Java, Python and R
RDDs are immutable and fault-tolerant
None of the above

Q.4 Can you combine the libraries of Apache Spark into the same Application, for example, MLlib, GraphX, SQL and DataFrames etc.
Yes
No

Q.5 Which of the following is true for RDD?
RDD is programming paradigm
RDD in Apache Spark is an immutable collection of objects
It is database
None of the above

Q.6 Which of the following is not a function of Spark Context in Apache Spark?
Entry point to Spark SQL
To Access various services
To set the configuration
To get the current status of Spark Application

Q.7 What are the features of Spark RDD?
In-memory computation
Lazy evaluations
Fault Tolerance
All of the above

Q.8 How many Spark Context can be active per JVM?

More than one
Only one
Not specific
None of the above
Q.9 In how many ways RDD can be created?
4
3
2
1

Q.10 How many tasks does Spark run on each partition?
Any number of task
One
More than one less than five

Q.11 Can we edit the data of RDD, for example, the case conversion?
Yes
No

Q.12 Which of the following is not a transformation?
Flatmap
Map
Reduce
Filter

Q.13 Which of the following is not an action?
collect()
take(n)
top()
Map

Q.14 Does Spark R make use of MLlib in any aspect?
Yes
No

Q.15 You can connect R program to a Spark cluster from -
RStudio
R Shell
Rscript
All of the above

Q.1 For Multiclass classification problem which algorithm is not the solution?
Naive Bayes

Random Forests
Logistic Regression
Decision Trees

Q.2 For Regression problem which algorithm is not the solution?
Logistic Regression
Ridge Regression
Decision Trees
Gradient-Boosted Trees

Q.3 Which of the following is true about DataFrame?
DataFrames provide a more user-friendly API than RDDs.
DataFrame API have provision for compile-time type safety
Both the above

Q.4 Which of the following is a tool of Machine Learning Library?
Persistence
Utilities like linear algebra, statistics
Pipelines
All of the above

Q.5 Is MLlib deprecated?
Yes
No

Q.6 Which of the following is false for Apache Spark?
It provides high-level API in Java, Python, R, Scala
It can be integrated with Hadoop and can process existing Hadoop HDFS data
Spark is an open source framework which is written in Java
Spark is 100 times faster than Bigdata Hadoop

Q.7 Which of the following is true for Spark SQL?
It is the kernel of Spark
Provides an execution platform for all the Spark applications
It enables users to run SQL / HQL queries on the top of Spark.
Enables powerful interactive and data analytics application across live streaming data

Q.8 Which of the following is true for Spark core?
It is the kernel of Spark
It enables users to run SQL / HQL queries on the top of Spark.
It is the scalable machine learning library which delivers efficiencies
Improves the performance of iterative algorithm drastically.

Q.9 Which of the following is true for Spark R?

It allows data scientists to analyze large datasets and interactively run jobs
It is the kernel of Spark
It is the scalable machine learning library which delivers efficiencies
It enables users to run SQL / HQL queries on the top of Spark.
Q.10 Which of the following is true for Spark MLlib?
Provides an execution platform for all the Spark applications
It is the scalable machine learning library which delivers efficiencies
enables powerful interactive and data analytics application across live streaming data
All of the above

Q.11 Which of the following is true for Spark Shell?
It helps Spark applications to easily run on the command line of the system
It runs/tests application code interactively
It allows reading from many types of data sources
All of the above

Q.12 Which of the following is true for RDD?
We can operate Spark RDDs in parallel with a low-level API
RDDs are similar to the table in a relational database
It allows processing of a large amount of structured data
It has built-in optimization engine

Q.13 RDD are fault-tolerant and immutable
True
False

Q.14 In which of the following cases do we keep the data in-memory?
Iterative algorithms
Interactive data mining tools
Both the above

Q.15 When does Apache Spark evaluate RDD?
Upon action
Upon transformation
On both transformation and action

Q.16 The read operation on RDD is
Fine-grained
Coarse-grained
Either fine-grained or coarse-grained
Neither fine-grained nor coarse-grained

Q.17 The write operation on RDD is
Fine-grained

Coarse-grained
Either fine-grained or coarse-grained
Neither fine-grained nor coarse-grained

Q.18 Is it possible to mitigate stragglers in RDD?
Yes
No

Q.19 Fault Tolerance in RDD is achieved using
Immutable nature of RDD
DAG (Directed Acyclic Graph)
Lazy-evaluation
None of the above

Q.20 What is a transformation in Spark RDD?
Takes RDD as input and produces one or more RDD as output.
Returns final result of RDD computations.
The ways to send result from executors to the driver
None of the above

Q.1 What is action in Spark RDD?
The ways to send result from executors to the driver
Takes RDD as input and produces one or more RDD as output.
Creates one or many new RDDs
All of the above

Q.2 Which of the following is true about narrow transformation -
The data required to compute resides on multiple partitions.
The data required to compute resides on the single partition.
Both the above

Q.3 Which of the following is true about wide transformation -
The data required to compute resides on multiple partitions.
The data required to compute resides on the single partition.
None of the both

Q.4 When we want to work with the actual dataset, at that point we use Transformation?
True
False

Q.5 The shortcomings of Hadoop MapReduce was overcome by Spark RDD by
Lazy-evaluation
DAG
In-memory processing

All of the above

Q.6 What does Spark Engine do?
Scheduling
Distributing data across a cluster
Monitoring data across a cluster
All of the above
Q.7 Caching is optimizing the technique
True
False

Q.8 Which of the following is the entry point of Spark Application -
SparkSession
SparkContext
None of the both

Q.9 SparkContext guides how to access the Spark cluster.
True
False

Q.10 Which of the following is the entry point of Spark SQL?
SparkSession
SparkContext

Q.11 Which of the following is open-source?
Apache Spark
Apache Hadoop
Apache Flink
All of the above

Q.12 Apache Spark supports -
Batch processing
Stream processing
Graph processing
All of the above

Q.13 Which of the following is not true for map() Operation?
Map transforms an RDD of length N into another RDD of length N.
In the Map operation developer can define his own custom business logic.
It applies to each element of RDD and it returns the result as new RDD
Map allows returning 0, 1 or more elements from map function.

Q.14 FlatMap transforms an RDD of length N into another RDD of length M. which of the following is true for N and M.

a. N>M

b. N<M

c. N<=M
Either a or b
Either b or c
Either a or c

Q.15 Which of the following is a transformation?
take(n)
top()
countByValue()
mapPartitionWithIndex()

Q.16 Which of the following is action?
Union(dataset)
Intersection(other-dataset)
Distinct()
CountByValue()

Q.17 In aggregate function can we get the data type different from as that input data type?
Yes
No

Q.18 In which of the following Action the result is not returned to the driver.
collect()
top()
countByValue()
foreach()

Q.19 Which of the following is true for stateless transformation?
Uses data or intermediate results from previous batches and computes the result of the current batch.
Windowed operations and updateStateByKey() are two type of Stateless transformation.
The processing of each batch has no dependency on the data of previous batches.
None of the above

Q.20 Which of the following is true for stateful transformation?
The processing of each batch has no dependency on the data of previous batches.
Uses data or intermediate results from previous batches and computes the result of the current batch.
Stateful transformations are simple RDD transformations.

None of the above

Q.1 The primary Machine Learning API for Spark is now the ______ based API
DataFrame
Dataset
RDD
All of the above

Q.2 Which of the following is a module for Structured data processing?
GraphX
MLlib
Spark SQL
Spark R

Q.3 SparkSQL translates commands into codes. These codes are processed by
Driver nodes
Executor Nodes
Cluster manager
None of the above

Q.4 Spark SQL plays the main role in the optimization of queries.
True
False

Q.5 This optimizer is based on functional programming construct in
Java
Scala
Python
R

Q.6 Catalyst Optimizer supports either rule-based or cost-based optimization.
True
False

Q.7 Which of the following is not true for Catalyst Optimizer?
Catalyst optimizer makes use of pattern matching feature.
Catalyst contains the tree and the set of rulesto manipulate the tree.
There are no specific libraries to process relational queries.
There are different rule sets which handle different phases of query.

Q.8 Which of the following is true for the tree in Catalyst optimizer?
A tree is the main data type in the catalyst.

New nodes are defined as subclasses of TreeNode class.
A tree contains a node object.
All of the above

Q.9 Which of the following is true for the rule in Catalyst optimizer?
We can manipulate tree using rules.
We can define rules as a function from one tree to another tree.
Using rule we get the pattern that matches each pattern to a result.
All of the above

Q.10 Which of the following is not a Spark SQL query execution phases?
Analysis
Logical Optimization
Execution
Physical planning

Q.11 In Spark SQL optimization which of the following is not present in the logical plan -
Constant folding
Abstract syntax tree
Projection pruning
Predicate pushdown

Q.12 In the analysis phase which is the correct order of execution after forming unresolved logical plan

a. Search relation BY NAME FROM CATALOG.

b. Determine which attributes match to the same value to give them unique ID.

c. Map the name attribute

d. Propagate and push type through expressions
abcd
acbd
adbc
Dcab

Q.13 In the Physical planning phase of Query optimization we can use both Coast-based and Rule-based optimization.
True
False

Q.14 DataFrame in Apache Spark prevails over RDD and does not contain any feature of RDD.
True

False

Q.15 Which of the following are the common feature of RDD and DataFrame?
Immutability
In-memory
Resilient
All of the above

Q.16 Which of the following is not true for DataFrame?
DataFrame in Apache Spark is behind RDD
We can build DataFrame from different data sources. structured data file, tables in Hive
The Application Programming Interface (APIs) of DataFrame is available in various languages
Both in Scala and Java, we represent DataFrame as Dataset of rows.

Q.17 In Dataframe in Spark Once the domain object is converted into a data frame, the regeneration of domain object is not possible.
True
False

Q.18 DataFrame API has provision for compile-time type safety.
True
False

Q.19 We can create DataFrame using:
Tables in Hive
Structured data files
External databases
All of the above

Q.20 Which of the following is the fundamental data structure of Spark
RDD
DataFrame
Dataset
None of the above

Q.1 Which of the following organized a data into a named column?

a. RDD

b. DataFrame

c. Dataset
Both a and b
Both b and c

Both a and c

Q.2 Which of the following provide the object-oriented programming interface
RDD
DataFrame
[✓ Dataset]
None of the above

Q.3 After transforming into DataFrame one cannot regenerate a domain object
[✓ True]
False

Q.4 RDD allows Java serialization
[✓ True]
False

Q.5 Which of the following make use of an encoder for serialization.
RDD
DataFrame
[✓ Dataset]
None of the above

Q.6 Apache Spark is presently added in all major distribution of Hadoop
[✓ True]
False

Q.7 Does Dataset API support Python and R.
Yes
[✓ No]

Q.8 Which of the following is slow to perform simple grouping and aggregation operations.
[✓ RDD]
DataFrame
Dataset
All of the above

Q.9 Which of the following is good for low-level transformation and actions.
[✓ RDD]
DataFrame
Dataset
All of the above

Q.10 Which of the following technology is good for Stream technology?
Apache Spark
Apache Hadoop
Apache Flink
None of the above

Q.11 Which of the following is not true for Apache Spark Execution?
To simplify working with structured data it provides DataFrame abstraction in Python, Java, and Scala.
The data can be read and written in a variety of structured formats. For example, JSON, Hive Tables, and Parquet.
Using SQL we can query data,only from inside a Spark program and not from external tools.
The best way to use Spark SQL is inside a Spark application. This empowers us to load data and query it with SQL.

Q.12 When SQL run from the other programming language the result will be
DataFrame
DataSet
Either DataFrame or Dataset
Neither DataFrame nor Dataset

Q.13 The Dataset API is accessible in
Java and Scala
Java, Scala and python
Scala and Python
Scala and R

Q.14 Dataset API is not supported by Python. But because of the dynamic nature of Python, many benefits of Dataset API are available.
True
False

Q.15 Which of the following is true for Catalyst optimizer?
The optimizer helps us to run queries much faster than their counter RDD part.
The optimizer helps us to run queries little faster than their counter RDD part.
The optimizer helps us to run queries in the same speed as their counter RDD part.

Q.16 Which of the following are uses of Apache Spark SQL?
It executes SQL queries.
We can read data from existing Hive installation using SparkSQL.
When we run SQL within another programming language we will get the result as Dataset/DataFrame.
All of the above

Q.17 With the help of Spark SQL, we can query structured data as a distributed dataset (RDD).
True
False

Q.18 Spark SQL can connect through JDBC or ODBC.
True
False

Q.19 Using Spark SQL, we can create or read a table containing union fields.
True
False

Q.20 Which of the following is true for Spark SQL?
Hive transactions are not supported by Spark SQL.
No support for time-stamp in Avro table.
Even if the inserted value exceeds the size limit, no error will occur.
All of the above

Q.1. Without an explicit import, maps in Scala are by default:
Immutable
Mutable

Q.2. Functions and numbers are objects in Scala.
True
False

Q.3. Every class in Scala inherits from a super class. Implicitly, this is scala.AnyRef.
False
True

Q.4. What is Scala's programming paradigm?
Statically-Typed
Functional
Object-Oriented
All of these

Q.5. We can have static members in classes.
False
True

Q.6. Scala is case-insensitive. Identifiers Name and name are the same things.
False
True

Q.7. What does type inference mean?
We must explicitly mention the data type for a variable
Scala determines a variable’s type by looking at its value

Q.8. Omitting a semicolon(;) at the end of a statement causes the compiler to throw an error.
False
True

Q.9. UNIT is a data type in Scala. It pertains to no meaningful information.
True
False

Q.10. Which of the following statements are true about Lists and Arrays?
Lists are immutable
Arrays are mutable
Once you have declared a List, you cannot add more elements later
All of the above

Q.11. A class inheriting from a trait and implementing its interface inherits all code in it.
False
True

Q.12. Consider the following statements about vals and vars. Select the ones that are true.
val is a constant
Reassigning to a val throws an error
Reassigning to a var doesn’t throw an error
All of the above

Q.13. How do you turn the string “batmanstein” to the string “Man”?
“batmanstein”.drop(3).capitalize.take(3)
“batmanstein”.drop(3).take(3).capitalize
All of the above
None of these

Q.14. A closure is:
A function whose return value depends on a variable declared outside it
A function that takes, as argument, another function
A function that returns, as argument, another function
A function that returns a Future

Q.15. We use the following keyword to define a function in Scala:
def
func

function
We use the data type instead
Q.16. Select the correct statements for Nil, Null, None, and Nothing.
None is the value of an Option with no value in it
Nothing is the bottom type of the entire type system
All of the above
None of the above

Q.17. What is a monad in Scala?
A function with a single parameter
An object that wraps another
A singleton object
None of the above

Q.18. Select the true statements about iterators.
To get the next item, we use next()
We can use the method hasNext to find out if it has another element left
They yield the next element in the iterator
All of the above

Q.19. We do not need to pass these parameters to a method when calling it:
Named arguments
Implicit parameters
Default parameters
Command-line arguments

Q.20. What is a type class in Scala?
A class that performs boxing of a certain data type
A companion class
A trait with at least one type variable

Q.1 sbt is an open-source build tool for Scala and Java projects. Can it execute tasks in parallel?
No, it can only execute one task at once
Yes

Q.2 Which of the following statements is untrue about a functor in Scala?
It is a mapping between categories
It is a built-in construct in Scala
It maps objects or entities of one category to those of another

Q.3 Select the true statements from the following:
A Future is an object
A Future holds a value that may become available at a later point in time

We can complete a promise only once

All of the above

Q.4 Left and Right are case classes. True or False?

False

True

Q.5 Which of the following statements are true about Either?

It is a sealed abstract class

We can use it to deal with possible missing values

An instance of Either can be an instance of Left or Right

All of the above

Q.6 What Boolean value do the following statements return?

```
case class People(name:String,age:Int)
val people1=People("Ayushi",22)
val people2=People("Ayushi",22)
people1==people2
```

False

True

Q.7 Select the correct output for the following code:

```
val leaders=collection.mutable.Buffer("Reykon")
leaders+="obama"
println(leaders)
```

ArrayBuffer(Reykon, Obama)

ArrayBuffer(Reykon)

List(Reykon, obama)

The code throws an error

Q.8 Looking at the previous code, can you predict the output of the following code?

```
val stuff=collection.mutable.Buffer("blue")
stuff+=44
println(stuff)
```

ArrayBuffer(blue, 44)

ArrayBuffer(blue)

Something else

The code throws an error

Q.9 Which of the following function definitions are erroneous?

def functionName(x:Int,y:Int):Int=x+y

def functionName(x:Int,y:Int):Int={return x+y}

def functionName(x:Int,y:Int):Int{x+y} ()

def functionName(x:Int,y:Int)={x+y}

Q.10 Scala is also a:
Style Sheets Language
Scripting Language
Cloud Computing Language

Q.11 Of the following, select the Scala construct that holds pairwise different elements of the same type.
Sets
Groups
Forums
Maps

Q.12 Select the correct value specified by the following line of code:
"abcde" ensuring (_.length>3)
abc
abcd
abcde
The code throws an error

Q.13 Which is the correct value of the following expression?
List(1,2,3)flatMap(x=>List(x,4))
List(1,2,3,4)
List(1,4,2,4,3,4)
List(4,4,4)
List (1,2,3,x,4)

Q.14 For which kind of data should you use a case class?
Mutable data
Immutable data
Both of these
None of these

Q.15 The following technique/construct lets us transform a function with multiple arguments into a chain of functions. Each of these has one single argument.
Extractors
Traits
Trait mixins
Currying

Q.16 How do you abruptly stop execution in the REPL?
Pressing Ctrl+C
Pressing Ctrl+Q

Pressing Ctrl+W

Q.17 Which of the following is true about Scala and Java?
Both work with popular IDEs like Eclipse, Netbeans, and IntelliJ
Both run on the JVM
We can call Scala from Java and Java from Scala
All of the above

Q.18 Scala is a portmanteau for:
Scalar and Language
Script and Language
Sequential and Language
Scalable and Language

Q.19 Its Java compatibility makes Scala suitable for:
Android Development
Apple Development
Microsoft Development
Google Development

Q.20 A collection of type collection.Seq is immutable.
False
True

Q.1. Select the correct output for the following code:

```
var band={
var name="sublime"
name
}
println(name)
```

It prints “sublime”
It prints nothing
The code raises an error

Q.2. Choose the correct output for the following code:

```
val x:Option[String]=Some("hi")
println(x.get)
```

This prints Some(“hi”)
This prints nothing
The code throws an error
This prints “hi”

Q.3. The following code compiles:

```
class Complex(real:Double,imaginary:Double){
def re()=real
def im()=imaginary
}
```

True

False

Q.4. What does this code print?

```
var y:Option[String]=None
y.get
```

This prints None

This prints nothing

This throws an exception

None of the above

Q.5. Is the following function pure?

```
def change:Unit={
x=x+10
}
```

It is pure

It is not pure

Q.6. Select the correct output for the following code:

```
object Flash{
def superpower="speed"
}
```

This prints “speed”

This prints nothing

This throws an error

None of the above

Q.7. Choose the correct output:

```
object Dog{
def bark="woof"
}
```

It prints “woof”

It prints nothing

It raises an exception

None of the above

Q.8. Select the correct output:

```
val numbers=List(11,22,33)
var total=0
for(i<-numbers){
```

```
total+=i
}
println(total)
```

0

11

33

66

Q.9. Decide what the following code prints:

```
val odds=List(3,5,7)
var result=1
odds.foreach((num:Int)=>result*=num)
println(result)
```

1

105

List(105)

List(3,5,7)

Q.10. Okay, now try to do this one:

```
val evens=List(2,4,8)
println{
evens.foldLeft(0) { (memo: Int, y: Int) =>
memo+y
}
}
```

List(14, 12, 8)

List(8,4,2)

2

14

Q.11. Select appropriate output:

```
def quadruple(x:Int):Int=x*4
val quadrupleCopy=quadruple _
println(quadrupleCopy(-1))
```

-4

0

20

The code throws an error

Q.12. Which abstraction from functional programming helps us deal with updating complex immutable nested objects?

Case classes

Lens

Extractors

Q.13. Choose the correct output:

```
var greeting:Option[String]= Some("hello")
greeting= Some(7)
println(greeting.get)
```

It throws an error

It prints 7

It prints “hello”

Q.14. We do not need to pass these parameters to a method when calling it:

Named arguments

Implicit parameters

Default parameters

Command-line arguments

Q.15. Select the correct output:

```
val cool=Map("a"->"aaa", "b"->"bbb", "a"->"ccc")
println(cool("a"))
```

“a”

“aaa”

“bbb”

“ccc”

Q1. Select the output for the following line of code:

```
println(40.getClass)
```

This causes a compilation error

This causes a runtime error

int

40

Q2. What about this one?

```
println("frankl"||true))
```

True

False

The code doesn't print anything

This code throws an error

Q3. What is a higher-order function in Scala?

It takes other functions as parameters

It returns a function as a result

Both the above

Q4. Consider the following list:

```
var countries=List("brazil", "argentina", "colombia")
```

What does the following code do to it?

```
println{
countries.reduceLeft[String]{(c1: String, c2: String)=>
s"$c1, $c2"
}
}
```

It prints "$c1, $c2"
It prints List("brazil", "argentina", "colombia")
It prints "brazil", "argentina", "colombia"
It prints brazil, argentina, colombia

Q5. Now consider this list:

```
var rrr= List("ant", "beer", "battered", "cool", "burger")
```

What will this code do to it? Select what it prints.

```
rrr.filter {(w: String) =>
w.take(1) == "b"
}.reduceLeft{(a: String, b: String) =>
s"$a $b"
}
```

ant beer battered cool burger
beer battered burger
ant eer attered ool urger
eer attered urger

Q6. Select the correct statements about the apply and unapply methods.
Apply let us create an object from arguments
Unapply let us take an object apart
Apply is like a constructor
All of the above

Q7. What does the following piece of code print?

```
case class PersonData(name: String, age: Int)
val bob1=new PersonData("bob", 99)
val bob2=new PersonData("bob", 99)
println(bob1==bob2)
```

False
True
It throws an error
It throws an exception

Q8. The following statements are true about companion objects and companion classes:

A companion object is an object with the same name as a class

Companion classes and objects can access private members of their companions

Both the above

Q9. Tell us the output of the following snippet of code:

```
case class Dog(breed: String, age: Int)
val fido= new Dog("lab", 4)
println(fido.toString)
```

It prints "lab"

It prints Dog(lab, 4)

It raises an exception

None of the above

Q10. What does this code do?

```
object DoubleUtils{
implicit class Funny(val num: Double, joke: String){
def knockKnock={
s"${num.toString} is here"
}
}
}
import DoubleUtils._
println(3.14.knockKnock)
```

It prints "3.14 is here"

It causes the compiler to crash

It throws an error

None of these

Q.11 Select the correct output:

```
object Whatever{
def speak(something: String)(implicit nice: String)={
println(s"$something $nice")
}
}

implicit val nice= "the walrus"
println{
Whatever.speak("I am")
}
println{
Whatever.speak("I like")("catfood")
```

}
I am the walrus, I like the walrus
I am the walrus, the walrus catfood
I am the walrus, I like catfood
The code throws an error

Q.12 Select the correct statements from the following:
Unit is like void in Java
() is an empty tuple that represents a Unit
Both the above

Q.13 What does the following code print?

```
trait Diva{
var attitude= "subjective"
}
var arianaGrand= new Diva
println(arianaGrande.attitude)
```

Subjective
The code throws an error
None of the above

Q.14 Select the correct statements from the following:
Case classes allow pattern-matching
We must use the keyword to instantiate a case class
We must manually define accessor methods for all constructor arguments
We must generate methods equals(), hashcode(), and toString()

Q.15 Consider the following string.
val s= "(888) 333-4444"
How would you replace all digit with the letter 'x'?
s.replace("[0-9]","x")
s.replaceAll("x", "[0-9]")
s.replace("x","[0-9]")
s.replaceAll("[0-9]","x")

Q.16 Which of the following is not a way to make an executable Scala program?
Execute a script file in the interpreter
Use an IDE
Compile an object with a main method
None of the above is not a way for the said purpose

Q.17 One of the following is not a kind of Scala identifier. Selct the one.
Alphanumeric identifier
String identifier

Operator identifier
Literal identifier

Q.18 Which of the following is a type of literal in Scala?
String literal
Boolean literal
Symbol literal
All of the above

Q.19 Select a regular expression from the options that will parse out the number from the following string: "Milton Friedman died at 94 years of age."
"[0-9]".r
"[0-9]+".r
"[0-9]*".r
None of the above

Q. 20 Select the correct statements about Array and ArrayBuffer.
Arrays are immutable
An ArrayBuffer is of variable size
Arrays are similar to Java's ArrayLists
ArrayBuffers are similar to Java's arrays

Q.21 The following code complies:

```
class Complex(real:Double,imaginary:Double){
def re()=real
def im()=imaginary
}
```

True
False

Q.22 Does this code compile successfully? Does it print anything?

```
 def sad="meow"
val catCry=sad
println(catCry())
```

It does not compile; throws an error
It compiles but produces no output
It compiles successfully and prints "meow"

Q.23 Select the correct output for the code:

```
val arr=Array(2,3,4)
arr.update(1,5)
```

Line 2 throws an error at the time of compilation
arr is a val; we cannot reassign it. So, this throws an error
arr now holds (2,5,4)

Q.24 What is the output of the following code?

```
class User(n:String){
val name:String=n
}
var u=new User(n="Frankl")
println(u.name)
```

This does not compile; throws an error
This prints Frankl
This compiles, but prints nothing

Q.25 What does the variable x hold in the following code:

```
var x,y,z=(1,2,3)
```

1
(1,2,3)
The code produce an error

Q.1 What is YARN?
Storage layer
Batch processing engine
Resource Management Layer
None of the above

Q.2 Which of the following is not a scheduling option available in YARN
Balanced scheduler
Fair scheduler
Capacity scheduler
FIFO schesduler

Q.3 Which among the following is the Resource Management Layer
Mapreduce
YARN
HDFS
HIVE

Q.4 Which of the following is the architectural center of Hadoop that allows multiple data processing engines.
Hive
Incubator
Yarn
Chuckwa

Q.5 Which of the following is framework-specific entity that negotiates resources from the ResourceManager

NodeManager
ResourceManager
ApplicationMaster
All of the above

Q.6 Apache Hadoop YARN stands for :
Yet Another Reserve Negotiator
Yet Another Resource Network
Yet Another Resource Negotiator
None of the above

Q.7 Which among the following is ultimate authority that arbitrates resources among all the applications in the system.
NodeManager
ResourceManager
ApplicationMaster
All of the above

Q.8 Which is responsible for allocating resources to the various running applications subject to familiar constraints of capacities, queues etc.
Manager
Master
Scheduler
None of the above

Q.9 The CapacityScheduler supports _______________ queues to allow for more predictable sharing of cluster resources.
Hierarchial
Networked
Partition
None of the abpve

Q.10 Yarn commands are invoked by the which of the following script.
Hive
Bin
Hadoop
Home

Q.11 Users can bundle their Yarn code in a __________ file and execute it using jar command.
Java
Jar
C code
XML

Q.12 Which of the following command is used to dump the log container?
Logs
Log
Dump
None of the above

Q.13 Which of the following command runs ResourceManager admin client ?
Proxyserver
Run
Admin
Rmadmin

Q.14 The CapacityScheduler has a pre-defined queue called :
Domain
Root
Rear
None of the above

Q.15 Is YARN a replacement of Hadoop MapReduce?
Yes
No

Q.16 Which is the slave daemon of Yarn.
NodeManager
ResourceManager
ApplicationMaster

Q.17 Which among the following is yarn node manager components
NodeStatusUpdater
ContainerManager
Container Executor
All of the above

Q.18 Which of the following has occupied the place of JobTracker of MRV1
NodeManager
ResourceManager
ApplicationMaster
Scheduler

Q.19 Which of the following has occupied the place of TaskTracker of MRV1
NodeManager
ResourceManager

ApplicationMaster
Scheduler

Q.20 Yarn's dynamic allocation of cluster resources improves utilization over more static ________ rules used in early versions of Hadoop.
HIve
Mapreduce
Imphala

1. What does commodity Hardware in Hadoop world mean? (D)

a) Very cheap hardware

b) Industry standard hardware

c) Discarded hardware

d) Low specifications Industry grade hardware

2. Which of the following are NOT big data problem(s)? (D)

a) Parsing 5 MB XML file every 5 minutes

b) Processing IPL tweet sentiments

c) Processing online bank transactions

d) both (a) and (c)

3. What does "Velocity" in Big Data mean? (D)

a) Speed of input data generation

b) Speed of individual machine processors

c) Speed of ONLY storing data

d) Speed of storing and processing data

4. The term Big Data first originated from: (C)

a) Stock Markets Domain

b) Banking and Finance Domain

c) Genomics and Astronomy Domain

d) Social Media Domain

5. Which of the following Batch Processing instance is NOT an example of (D)

BigData Batch Processing?

a) Processing 10 GB sales data every 6 hours

b) Processing flights sensor data

c) Web crawling app

d) Trending topic analysis of tweets for last 15 minutes

6. Which of the following are example(s) of Real Time Big Data Processing? (D)

a) Complex Event Processing (CEP) platforms

b) Stock market data analysis

c) Bank fraud transactions detection

d) both (a) and (c)

7. Sliding window operations typically fall in the category (C)
of__________________.

a) OLTP Transactions

b) Big Data Batch Processing

c) Big Data Real Time Processing

d) Small Batch Processing

8. What is HBase used as? (A)

a) Tool for Random and Fast Read/Write operations in Hadoop

b) Faster Read only query engine in Hadoop

c) MapReduce alternative in Hadoop

d) Fast MapReduce layer in Hadoop

9. What is Hive used as? (D)

a) Hadoop query engine

b) MapReduce wrapper

c) Hadoop SQL interface

d) All of the above

10. Which of the following are NOT true for Hadoop? (D)

a) It's a tool for Big Data analysis

b) It supports structured and unstructured data analysis

c) It aims for vertical scaling out/in scenarios

d) Both (a) and (c)

11. Which of the following are the core components of Hadoop? (D)

a) HDFS

b) Map Reduce

c) HBase

d) Both (a) and (b)

12. Hadoop is open source. (B)

a) ALWAYS True

b) True only for Apache Hadoop

c) True only for Apache and Cloudera Hadoop

d) ALWAYS False

13. Hive can be used for real time queries. (B)

a) TRUE

b) FALSE

c) True if data set is small

d) True for some distributions

14. What is the default HDFS block size? (D)

a) 32 MB

b) 64 KB

c) 128 KB

d) 64 MB

15. What is the default HDFS replication factor? (C)

a) 4

b) 1

c) 3

d) 2

16. Which of the following is NOT a type of metadata in NameNode? (C)

a) List of files

b) Block locations of files

c) No. of file records

d) File access control information

17. Which of the following is/are correct? (D)

a) NameNode is the SPOF in Hadoop 1.x

b) NameNode is the SPOF in Hadoop 2.x

c) NameNode keeps the image of the file system also

d) Both (a) and (c)

18. The mechanism used to create replica in HDFS is____________. (C)

a) Gossip protocol

b) Replicate protocol

c) HDFS protocol

d) Store and Forward protocol

19. NameNode tries to keep the first copy of data nearest to the client machine. (C)

a) ALWAYS true

b) ALWAYS False

c) True if the client machine is the part of the cluster

d) True if the client machine is not the part of the cluster

20. HDFS data blocks can be read in parallel. (A)

a) TRUE

b) FALSE

21. Where is HDFS replication factor controlled? (D)

a) mapred-site.xml

b) yarn-site.xml

c) core-site.xml

d) hdfs-site.xml

22. Read the statement and select the correct option: (B)

It is necessary to default all the properties in Hadoop config files.

a) True

b) False

23. Which of the following Hadoop config files is used to define the heap size? (C)

a) hdfs-site.xml

b) core-site.xml

c) hadoop-env.sh

d) Slaves

24. Which of the following is not a valid Hadoop config file? (B)

a) mapred-site.xml

b) hadoop-site.xml

c) core-site.xml

d) Masters

25. Read the statement:

NameNodes are usually high storage machines in the clusters. (B)

a) True

b) False

c) Depends on cluster size

d) True if co-located with Job tracker

26. From the options listed below, select the suitable data sources for flume. (D)

a) Publicly open web sites

b) Local data folders

c) Remote web servers

d) Both (a) and (c)

27. Read the statement and select the correct options: (A)

distcp command ALWAYS needs fully qualified hdfs paths.

a) True

b) False

c) True, if source and destination are in same cluster

d) False, if source and destination are in same cluster

28. Which of following statement(s) are true about distcp command? (A)

a) It invokes MapReduce in background

b) It invokes MapReduce if source and destination are in same cluster

c) It can’t copy data from local folder to hdfs folder

d) You can’t overwrite the files through distcp command

29. Which of the following is NOT the component of Flume? (B)

a) Sink

b) Database

c) Source

d) Channel

30. Which of the following is the correct sequence of MapReduce flow? (C)

f) Map Reduce Combine

a) Combine Reduce Map

b) Map Combine Reduce

c) Reduce Combine Map

31 .Which of the following can be used to control the number of part files (B) in a map reduce program output directory?

a) Number of Mappers

b) Number of Reducers

c) Counter

d) Partitioner

32. Which of the following operations can’t use Reducer as combiner also? (D)

a) Group by Minimum

b) Group by Maximum

c) Group by Count

d) Group by Average

33. Which of the following is/are true about combiners? (D)

a) Combiners can be used for mapper only job

b) Combiners can be used for any Map Reduce operation

c) Mappers can be used as a combiner class

d) Combiners are primarily aimed to improve Map Reduce performance

e) Combiners can’t be applied for associative operations

34. Reduce side join is useful for (A)

a) Very large datasets

b) Very small data sets

c) One small and other big data sets

d) One big and other small datasets

35. Distributed Cache can be used in (D)

a) Mapper phase only

b) Reducer phase only

c) In either phase, but not on both sides simultaneously

d) In either phase

36. Counters persist the data on hard disk. (B)

a) True

b) False

37. What is optimal size of a file for distributed cache? (C)

a) <=10 MB

b) >=250 MB

c) <=100 MB

d) <=35 MB

38. Number of mappers is decided by the (D)

a) Mappers specified by the programmer

b) Available Mapper slots

c) Available heap memory

d) Input Splits

e) Input Format

39. Which of the following type of joins can be performed in Reduce side join operation? (E)

a) Equi Join

b) Left Outer Join

c) Right Outer Join

d) Full Outer Join

e) All of the above

40. What should be an upper limit for counters of a Map Reduce job? (D)

a) ~5s

b) ~15

c) ~150

d) ~50

41. Which of the following class is responsible for converting inputs to key-value (c) Pairs of Map Reduce

a) FileInputFormat

b) InputSplit

c) RecordReader

d) Mapper

42. Which of the following writables can be used to know value from a mapper/reducer? (C)

a) Text

b) IntWritable

c) Nullwritable

d) String

43. Distributed cache files can't be accessed in Reducer. (B)

a) True

b) False

44. Only one distributed cache file can be used in a Map Reduce job. (B)

a) True

b) False

45. A Map reduce job can be written in: (D)

a) Java

b) Ruby

c) Python

d) Any Language which can read from input stream

46. Pig is a: (B)

a) Programming Language

b) Data Flow Language

c) Query Language

d) Database

47. Pig is good for: (E)

a) Data Factory operations

b) Data Warehouse operations

c) Implementing complex SQLs

d) Creating multiple datasets from a single large dataset

e) Both (a) and (d)

48. Pig can be used for real-time data updates. (B)

a) True

b) False

49. Pig jobs have the same run time as the native Map Reduce jobs. (B)

a) True

b) False

50. Which of the following is the correct representation to access '’Skill" from the (A)

Bag {'Skills',55, ('Skill', 'Speed'), {2, ('San', 'Mateo')}}

a) $3.$1

b) $3.$0

c) $2.$0

d) $2.$1

HADOOP Interview Questions and Answers pdf ::

51. Replicated joins are useful for dealing with data skew. (B)

a) True

b) False

52. Maximum size allowed for small dataset in replicated join is: (C)

a) 10KB

b) 10 MB

c) 100 MB

d) 500 MB

53. Parameters could be passed to Pig scripts from: (E)

a) Parent Pig Scripts

b) Shell Script

c) Command Line

d) Configuration File

e) All the above except (a)

54. The schema of a relation can be examined through: (B)

a) ILLUSTRATE

b) DESCRIBE

c) DUMP

d) EXPLAIN

55. DUMP Statement writes the output in a file. (B)

a) True

b) False

56. Data can be supplied to PigUnit tests from: (C)

a) HDFS Location

b) Within Program

c) Both (a) and (b)

d) None of the above

57. Which of the following constructs are valid Pig Control Structures? (D)

a) If-else

b) For Loop

c) Until Loop

d) None of the above

58. Which of following is the return data type of Filter UDF? (C)

a) String

b) Integer

c) Boolean

d) None of the above

59. UDFs can be applied only in FOREACH statements in Pig. (A)

a) True

b) False

60. Which of the following are not possible in Hive? (E)

a) Creating Tables

b) Creating Indexes

c) Creating Synonym

d) Writing Update Statements

e) Both (c) and (d)

61. Who will initiate the mapper? (A)

a) Task tracker

b) Job tracker

c) Combiner

d) Reducer

62. Categorize the following to the following datatype

a) JSON files – Semi-structured

b) Word Docs , PDF Files , Text files – Unstructured

c) Email body – Unstructured

d) Data from enterprise systems (DB, CRM) – Structured

63. Which of the following are the Big Data Solutions Candidates? (E)

a) Processing 1.5 TB data everyday

b) Processing 30 minutes Flight sensor data

c) Interconnecting 50K data points (approx. 1 MB input file)

d) Processing User clicks on a website

e) All of the above

64. Hadoop is a framework that allows the distributed processing of: (C)

a) Small Data Sets

b) Semi-Large Data Sets

c) Large Data Sets

d) Large and Small Data sets

65. Where does Sqoop ingest data from? (B) & (D)

a) Linux File Directory

b) Oracle

c) HBase

d) MySQL

e) MongoDB

66. Identify the batch processing scenarios from following: (C) & (E)

a) Sliding Window Averages Job

b) Facebook Comments Processing Job

c) Inventory Dynamic Pricing Job

d) Fraudulent Transaction Identification Job

e) Financial Forecasting Job

67. Which of the following is not true about Name Node? (B)& (C) &(D)

a) It is the Master Machine of the Cluster

b) It is Name Node that can store user data

c) Name Node is a storage heavy machine

d) Name Node can be replaced by any Data Node Machine

68. Which of the following are NOT metadata items? (E)

a) List of HDFS files

b) HDFS block locations

c) Replication factor of files

d) Access Rights

e) File Records distribution

69. What decides number of Mappers for a MapReduce job? (C)

a) File Location

b) mapred.map.tasks parameter

c) Input file size

d) Input Splits

70. Name Node monitors block replication process (B)

a) TRUE

b) FALSE

c) Depends on file type

71. Which of the following are true for Hadoop Pseudo Distributed Mode? (C)

a) It runs on multiple machines

b) Runs on multiple machines without any daemons

c) Runs on Single Machine with all daemons

d) Runs on Single Machine without all daemons

72. Which of following statement(s) are correct? (C)

a) Master and slaves files are optional in Hadoop 2.x

b) Master file has list of all name nodes

c) Core-site has hdfs and MapReduce related common properties

d) hdfs-site file is now deprecated in Hadoop 2.x

73. Which of the following is true for Hive? (C)

a) Hive is the database of Hadoop

b) Hive supports schema checking

c) Hive doesn’t allow row level updates

d) Hive can replace an OLTP system

74. Which of the following is the highest level of Data Model in Hive? (c)

a) Table

b) View

c) Database

d) Partitions

75. Hive queries response time is in order of (C)

a) Hours at least

b) Minutes at least

c) Seconds at least

d) Milliseconds at least

76. Managed tables in Hive: (D)

a) Can load the data only from HDFS

b) Can load the data only from local file system

c) Are useful for enterprise wide data

d) Are Managed by Hive for their data and metadata

77. Partitioned tables in Hive: (D)

a) Are aimed to increase the performance of the queries

b) Modify the underlying HDFS structure

c) Are not useful if the filter columns for query are different from the partition columns

d) All of the above

78. Hive UDFs can only be written in Java (B)

a) True

b) False

79. Hive can load the data from: (D)

a) Local File system

b) HDFS File system

c) Output of a Pig Job

d) All of the above

80. HBase is a key/value store. Specifically it is: (E)

a) Sparse

b) Sorted Map

c) Distributed

d) Consistent

e) Multi- dimensional

81. Which of the following is the outer most part of HBase data model (A)

a) Database

b) Table

c) Row key

d) Column family

82. Which of the following is/are true? (A & D)

a) HBase table has fixed number of Column families

b) HBase table has fixed number of Columns

c) HBase doesn't allow row level updates

d) HBase access HDFS data

83. Data can be loaded in HBase from Pig using (D)

a) PigStorage

b) SqoopStorage

c) BinStorage

d) HbaseStorage

84. Sqoop can load the data in HBase (A)

a) True

b) False

85. Which of the following APIs can be used for exploring HBase tables? (D)

a) HBaseDescriptor

b) HBaseAdmin

c) Configuration

d) HTable

86. Which of the following tables in HBase holds the region to key mapping? (B)

a) ROOT

b) .META.

c) MAP

d) REGIONS

87. What is the data type of version in HBase? (B)

a) INT

b) LONG

c) STRING

d) DATE

88. What is the data type of row key in HBase? (D)

a) INT

b) STRING

c) BYTE

d) BYTE[]

89. HBase first reads the data from (B)

a) Block Cache

b) Memstore

c) HFile

d) WAL

90. The High availability of Namenode is achieved in HDFS2.x using (C)

a) Polled Edit Logs

b) Synchronized Edit Logs

c) Shared Edit Logs

d) Edit Logs Replacement

91. The application master monitors all Map Reduce applications in the cluster (B)

a) True

b) False

92. HDFS Federation is useful for the cluster size of: (C)

a) >500 nodes

b) >900 nodes

c) > 5000 nodes

d) > 3500 nodes

93. Hive managed tables stores the data in (C)

a) Local Linux path

b) Any HDFS path

c) HDFS warehouse path

d) None of the above

94. On dropping managed tables, Hive: (C)

a) Retains data, but deletes metadata

b) Retains metadata, but deletes data

c) Drops both, data and metadata

d) Retains both, data and metadata

95. Managed tables don’t allow loading data from other tables. (B)

a) True

b) False

96. External tables can load the data from warehouse Hive directory. (A)

a) True

b) False

97. On dropping external tables, Hive: (A)

a) Retains data, but deletes metadata

b) Retains metadata, but deletes data

c) Drops both, data and metadata

d) Retains both, data and metadata

98. Partitioned tables can’t load the data from normal (partitioned) tables (B)

a) True

b) False

99. The partitioned columns in Hive tables are (B)

a) Physically present and can be accessed

b) Physically absent but can be accessed

c) Physically present but can’t be accessed

d) Physically absent and can’t be accessed

100. Hive data models represent (C)

a) Table in Metastore DB

b) Table in HDFS

c) Directories in HDFS

d) None of the above

101. When is the earliest point at which the reduce method of a given Reducer can be called?

A. As soon as at least one mapper has finished processing its input split.

B. As soon as a mapper has emitted at least one record.

C. Not until all mappers have finished processing all records.

D. It depends on the InputFormat used for the job.

Answer: C

Explanation:

In a MapReduce job reducers do not start executing the reduce method until the all Map jobs have completed. Reducers start copying intermediate key-value pairs from the mappers as soon as they are available. The programmer defined reduce method is called only after all the mappers have finished.

Note: The reduce phase has 3 steps: shuffle, sort, and reduce. Shuffle is where the data is collected by the reducer from each mapper. This can happen while mappers are generating data since it is only a data transfer. On the other hand, sort and reduce can only start once all the mappers are done.

Why is starting the reducers early a good thing? Because it spreads out the data transfer from the mappers to the reducers over time, which is a good thing if your network is the bottleneck.

Why is starting the reducers early a bad thing? Because they "hog up" reduce slots while only copying data. Another job that starts later that will actually use the reduce slots now can't use them.

We can customize when the reducers startup by changing the default value of mapred.reduce.slowstart.completed.maps in mapred-site.xml. A value of 1.00 will wait for all the mappers to finish before starting the reducers. A value of 0.0 will start the reducers right away. A value of 0.5 will start the reducers when half of the mappers are complete. You can also change mapred.reduce.slowstart.completed.maps on a job-by-job basis.

Typically, keep mapred.reduce.slowstart.completed.maps above 0.9 if the system ever has multiple jobs running at once. This way the job doesn't hog up reducers when they aren't doing anything but copying data. If we have only one job running at a time, doing 0.1 would probably be appropriate.

102. Which describes how a client reads a file from HDFS?

A. The client queries the NameNode for the block location(s). The NameNode returns the block location(s) to the client. The client reads the data directory off the DataNode(s).

B. The client queries all DataNodes in parallel. The DataNode that contains the requested data responds directly to the client. The client reads the data directly off the DataNode.

C. The client contacts the NameNode for the block location(s). The NameNode then queries the DataNodes for block locations. The DataNodes respond to the NameNode, and the NameNode redirects the client to the DataNode that holds the requested data block(s). The client then reads the data directly off the DataNode.

D. The client contacts the NameNode for the block location(s). The NameNode contacts the DataNode that holds the requested data block. Data is transferred from the DataNode to the NameNode, and then from the NameNode to the client.

Answer: C

103. When You are developing a combiner that takes as input Text keys, IntWritable values, and emits Text keys, IntWritable values. Which interface should your class implement?

A. Combiner <Text, IntWritable,Text, IntWritable>

A. Reducer <Text, IntWritable,Text, IntWritable>

A. Combiner <Text,Text, IntWritable, IntWritable>

A. Combiner <Text, Text, IntWritable, IntWritable>

Answer: B

104. Indentify the utility that allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer?

A. Oozie

B. Sqoop

C. Flume

D. Hadoop Streaming

E. mapred

Answer: D

105. How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of MapReduce?

A. Keys are presented to reducer in sorted order; values for a given key are not sorted. B. Keys are presented to reducer in sorted order; values for a given key are sorted in ascending order.

C. Keys are presented to a reducer in random order; values for a given key are not sorted.

D. Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.

Answer: A

106. Assuming default settings, which best describes the order of data provided to a reducer's reduce method

A. The keys given to a reducer aren't in a predictable order, but the values associated with those keys always are.

B. Both the keys and values passed to a reducer always appear in sorted order.

C. Neither keys nor values are in any predictable order.

D. The keys given to a reducer are in sorted order but the values associated with each key are in no predictable order

Answer: D

1. who was the developer of Hadoop language?

A. Apache Software Foundation

B. Hadoop Software Foundation

C. Sun Microsystems

D. Bell Labs

View Answer

Ans : A

Explanation: Hadoop Developed by: Apache Software Foundation.

2. The hadoop language wriiten in which language?

A. C

B. C++

C. Java

D. Python

View Answer

Ans : C

Explanation: The hadoop language Written in: Java.

3. What was the Initial release date of hadoop?

A. 1st April 2007

B. 1st April 2006

C. 1st April 2008

D. 1st April 2005

View Answer

Ans : B

Explanation: Initial release: April 1, 2006; 13 years ago.

4. What license is Hadoop distributed under?

A. Apache License 2.1

B. Apache License 2.2

C. Apache License 2.0

D. Apache License 1.0

View Answer

Ans : C

Explanation: Hadoop is Open Source, released under Apache 2 license.

5. IBM and ________ have announced a major initiative to use Hadoop to support university courses in distributed computer programming.

A. Google

B. Apple

C. Facebook

D. Microsoft

View Answer

Ans : A

Explanation: Google and IBM Announce University Initiative to Address Internet-Scale.

6. On which platfrm hadoop langauge runs?

A. Bare metal

B. Debian

C. Cross-platform

D. Unix-Like

View Answer

Ans : C

Explanation: Hadoop has support for cross platform operating system.

7. Point out the correct statement.

A. Hadoop do need specialized hardware to process the data

B. Hadoop 2.0 allows live stream processing of real time data

C. In Hadoop programming framework output files are divided into lines or records

D. None of the above

View Answer

Ans : B

Explanation: Hadoop batch processes data distributed over a number of computers ranging in 100s and 1000s.

8. Which of the following fields come under the umbrella of Big Data?

A. Black Box Data

B. Power Grid Data

C. Search Engine Data

D. All of the above

View Answer

Ans : D

Explanation: All options are the fields come under the umbrella of Big Data.

9. How many types of data is present?

A. 2

B. 3

C. 5

D. 4

View Answer

Ans : B

Explanation: The data in it will be of three types:Structured data , Semi Structured data, Unstructured data.

10. Which of the following is not Features Of Hadoop?

A. Suitable for Big Data Analysis

B. Scalability

C. Robust

D. Fault Tolerance

View Answer

Ans : C

Explanation: Robust is is not Features Of Hadoop.

1. Data in ___________ bytes size is called Big Data.

A. Tera

B. Giga

C. Peta

D. Meta

View Answer

Ans : C

Explanation: data in Peta bytes i.e. 10^15 byte size is called Big Data.

2. How many V's of Big Data

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : D

Explanation: Big Data was defined by the “3Vs” but now there are “5Vs” of Big Data which are Volume, Velocity, Variety, Veracity, Value

3. Transaction data of the bank is?

A. structured data

B. unstructured datat

C. Both A and B

D. None of the above

View Answer

Ans : A

Explanation: Data which can be saved in tables are structured data like the transaction data of the bank.

4. In how many forms BigData could be found?

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : B

Explanation: BigData could be found in three forms: Structured, Unstructured and Semi-structured.

5. Which of the following are Benefits of Big Data Processing?

A. Businesses can utilize outside intelligence while taking decisions

B. Improved customer service

C. Better operational efficiency

D. All of the above

View Answer

Ans : D

Explanation: All of the above are Benefits of Big Data Processing.

6. Which of the following are incorrect Big Data Technologies?

A. Apache Hadoop

B. Apache Spark

C. Apache Kafka

D. Apache Pytarch

View Answer

Ans : D

Explanation: Apache Pytarch is incorrect Big Data Technologies.

7. The overall percentage of the world’s total data has been created just within the past two years is ?

A. 80%

B. 85%

C. 90%

D. 95%

View Answer

Ans : C

Explanation: The overall percentage of the world’s total data has been created just within the past two years is 90%.

8. Apache Kafka is an open-source platform that was created by?

A. LinkedIn

B. Facebook

C. Google

D. IBM

View Answer

Ans : A

Explanation: Apache Kafka is an open-source platform that was created by LinkedIn in the year 2011.

9. What was Hadoop named after?

A. Creator Doug Cutting’s favorite circus act

B. Cuttings high school rock band

C. The toy elephant of Cutting’s son

D. A sound Cutting’s laptop made during Hadoop development

View Answer

Ans : C

Explanation: Doug Cutting, Hadoop creator, named the framework after his child’s stuffed toy elephant.

10. What are the main components of Big Data?

A. MapReduce

B. HDFS

C. YARN

D. All of the above

View Answer

Ans : D

Explanation: All of the above are the main components of Big Data

1. The MapReduce algorithm contains two important tasks, namely __________.

A. mapped, reduce

B. mapping, Reduction

C. Map, Reduction

D. Map, Reduce

View Answer

Ans : D

Explanation: The MapReduce algorithm contains two important tasks, namely Map and Reduce.

2. _____ takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

A. Map

B. Reduce

C. Both A and B

D. Node

View Answer

Ans : A

Explanation: Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

3. ______ task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

A. Map

B. Reduce

C. Node

D. Both A and B

View Answer

Ans : B

Explanation: Reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

4. In how many stages the MapReduce program executes?

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : B

Explanation: MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

5. Which of the following is used to schedules jobs and tracks the assign jobs to Task tracker?

A. SlaveNode

B. MasterNode

C. JobTracker

D. Task Tracker

View Answer

Ans : C

Explanation: JobTracker : Schedules jobs and tracks the assign jobs to Task tracker.

6. Which of the following is used for an execution of a Mapper or a Reducer on a slice of data?

A. Task

B. Job

C. Mapper

D. PayLoad

View Answer

Ans : A

Explanation: Task : An execution of a Mapper or a Reducer on a slice of data.

7. Which of the following commnd runs a DFS admin client?

A. secondaryadminnode

B. nameadmin

C. dfsadmin

D. adminsck

View Answer

Ans : C

Explanation: dfsadmin : Runs a DFS admin client.

8. Point out the correct statement.

A. MapReduce tries to place the data and the compute as close as possible

B. Map Task in MapReduce is performed using the Mapper() function

C. Reduce Task in MapReduce is performed using the Map() function

D. None of the above

View Answer

Ans : A

Explanation: This feature of MapReduce is "Data Locality".

9. Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in ____________

A. C

B. C#

C. Java

D. None of the above

View Answer

Ans : C

Explanation: Hadoop Pipes is a SWIG- compatible C++ API to implement MapReduce applications (non JNITM based).

10. The number of maps is usually driven by the total size of ____________

A. Inputs

B. Output

C. Task

D. None of the above

View Answer

Ans : A

Explanation: Total size of inputs means the total number of blocks of the input files.

1. What is full form of HDFS?

A. Hadoop File System

B. Hadoop Field System

C. Hadoop File Search

D. Hadoop Field search

View Answer

Ans : A

Explanation: Hadoop File System was developed using distributed file system design.

2. HDFS works in a __________ fashion.

A. worker-master fashion

B. master-slave fashion

C. master-worker fashion

D. slave-master fashion

View Answer

Ans : B

Explanation: HDFS follows the master-slave architecture.

3. Which of the following are the Goals of HDFS?

A. Fault detection and recovery

B. Huge datasets

C. Hardware at data

D. All of the above

View Answer

Ans : D

Explanation: All the above option are the goals of HDFS.

4. ________ NameNode is used when the Primary NameNode goes down.

A. Rack

B. Data

C. Secondary

D. Both A and B

View Answer

Ans : C

Explanation: Secondary namenode is used for all time availability and reliability.

5. The minimum amount of data that HDFS can read or write is called a _____________.

A. Datanode

B. Namenode

C. Block

D. None of the above

View Answer

Ans : C

Explanation: The minimum amount of data that HDFS can read or write is called a Block.

6. The default block size is ______.

A. 32MB

B. 64MB

C. 128MB

D. 16MB

View Answer

Ans : B

Explanation: The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

7. For every node (Commodity hardware/System) in a cluster, there will be a _________.

A. Datanode

B. Namenode

C. Block

D. None of the above

View Answer

Ans : A

Explanation: For every node (Commodity hardware/System) in a cluster, there will be a datanode.

8. Which of the following is not Features Of HDFS?

A. It is suitable for the distributed storage and processing.

B. Streaming access to file system data.

C. HDFS provides file permissions and authentication.

D. Hadoop does not provides a command interface to interact with HDFS.

View Answer

Ans : D

Explanation: The correct feature is Hadoop provides a command interface to interact with HDFS.

9. HDFS is implemented in _____________ language.

A. Perl

B. Python

C. Java

D. C

View Answer

Ans : C

Explanation: HDFS is implemented in Java and any computer which can run Java can host a NameNode/DataNode on it.

10. During start up, the ___________ loads the file system state from the fsimage and the edits log file.

A. Datanode

B. Namenode

C. Block

D. ActionNode

View Answer

Ans : B

Explanation: HDFS is implemented on any computer which can run Java can host a NameNode/DataNode on it.

1. What is the full form of YARN?

A. Yet Another Resource Network

B. Yet Another Relational Negotiator

C. Yet Another Resource Negotiator

D. Yet Another Relational Network

View Answer

Ans : C

Explanation: Yet Another Resource Negotiator is the full form of YARN.

2. YARN is the one who helps to manage the resources across the ________.

A. clusters

B. Negotiator

C. Jobs

D. Hadoop System

View Answer

Ans : A

Explanation: YARN is the one who helps to manage the resources across the clusters.

3. How many major component Yarn has?

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : B

Explanation: Yarn consists of three major components.

4. Which of the following is the component of YARN?

A. Resource Manager

B. Nodes Manager

C. Application Manager

D. All of the above

View Answer

Ans : D

Explanation: Yarn consists of three major components i.e. Resource Manager, Nodes Manager, Application Manager.

5. Which managers work on the allocation of resources?

A. Nodes Manager

B. Resource Manager

C. Application Manager

D. All of the above

View Answer

Ans : A

Explanation: Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager.

6. Point out the correct statement.

A. YARN also extends the power of Hadoop to incumbent and new technologies found within the data center

B. YARN is the central point of investment for Hortonworks within the Apache community

C. YARN enhances a Hadoop compute cluster in many ways

D. All of the above

View Answer

Ans : D

Explanation: YARN provides ISVs and developers a consistent framework for writing data access applications that run IN Hadoop.

7. The ____________ is the ultimate authority that arbitrates resources among all the applications in the system.

A. NodeManager

B. ResourceManager

C. ApplicationMaster

D. All of the above

View Answer

Ans : B

Explanation: The ResourceManager and per-node slave, the NodeManager (NM), form the data-computation framework.

8. Yarn commands are invoked by the ________ script.

A. hive

B. bin

C. Hadoop

D. home

View Answer

Ans : B

Explanation: Running the yarn script without any arguments prints the description for all commands.

9. The CapacityScheduler has a predefined queue called _________

A. domain

B. rear

C. root

D. None of the above

View Answer

Ans : C

Explanation: All queues in the system are children of the root queue.

10. Which of the following command runs ResourceManager admin client?

A. proxyserver

B. run

C. admin

D. rmadmin

View Answer

Ans : D

Explanation: proxyserver command starts the web proxy server.

1. Which of the following is/are INCORRECT with respect to Hive?

A. Hive provides SQL interface to process large amount of data

B. Hive needs a relational database like oracle to perform query operations and store data.

C. Hive works well on all files stored in HDFS

D. Both A and B

View Answer

Ans : B

Explanation: Hive needs a relational database like oracle to perform query operations and store data is incorrect with respect to Hive.

2. Which of the following is not a Features of HiveQL?

A. Supports joins

B. Supports indexes

C. Support views

D. Support Transactions

View Answer

Ans : D

Explanation: Support Transactions is not a Features of HiveQL.

3. Which of the following operator executes a shell command from the Hive shell?

A. |

B. !

C. #

D. $

View Answer

Ans : B

Explanation: Exclamation operator is for execution of command.

4. Hive uses _________ for logging.

A. logj4

B. log4l

C. log4i

D. log4j

View Answer

Ans : D

Explanation: By default Hive will use hive-log4j.default in the conf/ directory of the Hive installation.

5. HCatalog is installed with Hive, starting with Hive release is ___________

A. 0.10.0

B. 0.9.0

C. 0.11.0

D. 0.12.0

View Answer

Ans : C

Explanation: hcat commands can be issued as hive commands, and vice versa.

6. _______ supports a new command shell Beeline that works with HiveServer2.

A. HiveServer2

B. HiveServer3

C. HiveServer4

D. HiveServer5

View Answer

Ans : A

Explanation: The Beeline shell works in both embedded mode as well as remote mode.

7. The ________ allows users to read or write Avro data as Hive tables.

A. AvroSerde

B. HiveSerde

C. SqlSerde

D. HiveQLSerde

View Answer

Ans : A

Explanation: AvroSerde understands compressed Avro files.

8. Which of the following data type is supported by Hive?

A. map

B. record

C. string

D. enum

View Answer

Ans : D

Explanation: Hive has no concept of enums.

9. We need to store skill set of MCQs(which might have multiple values) in MCQs table, which of the following is the best way to store this information in case of Hive?

A. Create a column in MCQs table of STRUCT data type

B. Create a column in MCQs table of MAP data type

C. Create a column in MCQs table of ARRAY data type

D. As storing multiple values in a column of MCQs itself is a violation

View Answer

Ans : C

Explanation: Option C is correct.

10. Letsfindcourse is generating huge amount of data. They are generating huge amount of sensor data from different courses which was unstructured in form. They moved to Hadoop framework for storing and analyzing data. What technology in Hadoop framework, they can use to analyse this unstructured data?

A. MapReduce programming

B. Hive

C. RDBMS

D. None of the above

View Answer

Ans : A

Explanation: MapReduce programming is the right answer.

1. which of the following is incorrect statement?

A. ZooKeeper is a distributed co-ordination service to manage large set of hosts.

B. ZooKeeper allows developers to focus on core application logic without worrying about the distributed nature of the application.

C. ZooKeeper solves this issue with its simple architecture and API.

D. The ZooKeeper framework was originally built at "Google" for accessing their applications in an easy and robust manner

View Answer

Ans : D

Explanation: Option D is incorrect : The ZooKeeper framework was originally built at "Yahoo!" for accessing their applications in an easy and robust manner

2. Which of the following is a benefits of distributed applications?

A. Scalability

B. Transparency

C. Reliability

D. All of the above

View Answer

Ans : D

Explanation: All options are the benefits of distributed applications.

3. ZooKeeper itself is intended to be replicated over a sets of hosts called ____________.

A. chunks

B. ensemble

C. subdomains

D. None of the above

View Answer

Ans : B

Explanation: As long as a majority of the servers are available, the ZooKeeper service will be available.

4. The underlying client-server protocol has changed in version _______ of ZooKeeper.

A. 3.0.0

B. 3.1.0

C. 3.1.1

D. 4.0.0

View Answer

Ans : A

Explanation: Old pre-3.0.0 clients are not guaranteed to operate against upgraded 3.0.0 servers and vice-versa.

5. Point out the wrong statement.

A. Cluster-wide status centralization service is essential for management and serialization tasks across a large distributed set of servers

B. Within ZooKeeper, an application can create what is called a znode

C. The znode can be updated by any node in the cluster, and any node in the cluster can register to be informed of changes to that znode

D. None of the above

View Answer

Ans : D

Explanation: ZooKeeper provides an infrastructure for cross-node synchronization and can be used by applications to ensure that tasks across the cluster are serialized or synchronized.

6. Helprace is using ZooKeeper on a _______ cluster in conjunction with Hadoop and HBase.

A. 3-node

B. 2-node

C. 5-node

D. 4-node

View Answer

Ans : A

Explanation: Zookeeper is used to manage a system build out of hadoop, katta, oracle batch jobs and a web component.

7. Which of the following is correct challenges that are faced by distributed applications?

A. Race condition

B. Deadlock

C. consistency

D. Both A and B

View Answer

Ans : D

Explanation: Race condition, Deadlock and Inconsistency are the correct challenges of distributed applications.

8. You need to have _________ installed before running ZooKeeper.

A. C

B. C++

C. Java

D. Python

View Answer

Ans : C

Explanation: Client bindings are available in several other languages.

9. How many types of special znodes are present in Zookeeper?

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : B

Explanation: There are two special types of znode: sequential and ephemeral.

10. Which of the following is not a services provided by ZooKeeper?

A. Naming service

B. Identity Service

C. Leader election

D. Cluster management

View Answer

Ans : B

Explanation: Identity Service is not a service provided by ZooKeeper.

1. Spark is best suited for ______ data.

A. Real-time

B. Virtual

C. Structured

D. All of the above

View Answer

Ans : A

Explanation: Spark is best suited for real-time data whereas Hadoop is best suited for structured data.

2. Which of the following Features of Apache Spark?

A. Speed

B. Supports multiple languages

C. Advanced Analytics

D. All of the above

View Answer

Ans : D

Explanation: Apache Spark has following features.: speed, Supports multiple languages ,Advanced Analytics.

3. In how many ways Spark uses Hadoop?

A. 2

B. 3

C. 4

D. 5

View Answer

Ans : A

Explanation: Spark uses Hadoop in two ways : one is storage and second is processing.

4. When was Apache Spark developed ?

A. 2007

B. 2008

C. 2009

D. 2010

View Answer

Ans : C

Explanation: Spark is one of Hadoop's sub project developed in 2009 in UC Berkeley's AMPLab by Matei Zaharia.

5. Which of the following is incorrect way for Spark deployment?

A. Standalone

B. Hadoop Yarn

C. Spark in MapReduce

D. Spark SQL

View Answer

Ans : D

Explanation: There are three ways of Spark deployment :-Standalone , Hadoop Yarn, Spark in MapReduce.

6. ____________ is a component on top of Spark Core.

A. Spark Streaming

B. Spark SQL

C. RDDs

D. None of the above

View Answer

Ans : B

Explanation: Spark SQL introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.

7. ________ is a distributed graph processing framework on top of Spark.

A. MLlib

B. Spark Streaming

C. GraphX

D. None of the above

View Answer

Ans : C

Explanation: GraphX started initially as a research project at UC Berkeley AMPLab and Databricks, and was later donated to the Spark project.

8. Point out the correct statement.

A. Spark enables Apache Hive users to run their unmodified queries much faster

B. Spark interoperates only with Hadoop

C. Spark is a popular data warehouse solution running on top of Hadoop

D. All of the above

View Answer

Ans : A

Explanation: Shark can accelerate Hive queries by as much as 100x when the input data fits into memory, and up 10x when the input data is stored on disk.

9. Which of the following can be used to launch Spark jobs inside MapReduce?

A. SIM

B. SIMR

C. SIR

D. RIS

View Answer

Ans : B

Explanation: With SIMR, users can start experimenting with Spark and use its shell within a couple of minutes after downloading it.

10. Which of the following language is not supported by Spark?

A. Python

B. Scala

C. Java

D. Pascal

View Answer

Ans : D

Explanation: The Spark engine runs in a variety of environments, from cloud services to Hadoop or Mesos clusters.

Q.1 All the files in a directory in HDFS can be merged together using which of the following?
Put merge
Get merge
Remerge
Merge all

Q.2 Which of these provides a Stream processing system used in Hadoop ecosystem?
Hive
Solr
Tez
Spark

Q.3 The client reading the data from HDFS filesystem in Hadoop does which of the following?
Gets only the block locations form the namenode
Gets the data from the namenode
Gets both the data and block location from the namenode
Gets the block location from the datanode

Q.4 Which of the following jobs are optimized for scalability but not latency
Mapreduce
Drill
Oozie
Hive

Q.5 Can multiple clients write into an HDFS file concurrently?
Yes
No

Q.6 Which of the following command is used to enter Safemode
hadoop dfsadmin –safemode get
bin dfsadmin –safemode get
hadoop dfsadmin –safemode enter
None of the above

Q.7 Which of the following command is used to come out of Safemode
hadoop dfsadmin -safemode leave
hadoop dfsadmin -safemode exit
hadoop dfsadmin -safemode out
None of the above

Q.8 During Safemode Hadoop cluster is in
Read-only
Write-only
Read-Write
None of the above

Q.9 HDFS allows a client to read a file which is already opened for writing?
False
True

Q.10 What happens when a file is deleted from the command line?
It is permanently deleted if trash is enabled.
It is permanently deleted and the file attributes are recorded in a log file.
It is placed into a trash directory common to all users for that cluster.
None of the above

Q. 11 Which one of the following statements is false about Distributed Cache?
Hive
Drill
Oozie
Mapreduce

Q.12 Which among the following are the features of Hadoop
Open source
Fault-tolerant
High Availability
All of the above

Q.13 Checkpoint node download the FsImage and EditLogs from the NameNode & then merge them & store the modified FsImage
Into persistent storage
Back to the active NameNode

Q.14 Which statement is true about Safemode
It is a maintenance state of NameNode
In Safemode, HDFS cluster is in read only
In Safemode, NameNode doesn’t allow any modifications to the file system
All of the above

Q.15 Which command is used to know the current status of the safe mode
hadoop dfsadmin –safemode get
hadoop dfsadmin –safemode getStatus
hadoop dfsadmin –safemode status
None of the above

Q.1 As compared to RDBMS, Apache Hadoop
Has higher data Integrity
Does ACID transactions
Is suitable for read and write many times
Works better on unstructured and semi-structured data.

Q.2 Which command lists the blocks that make up each file in the filesystem
hdfs fsck / -files -blocks
hdfs fsck / -blocks -files
hdfs fchk / -blocks -files
hdfs fchk / -files -blocks

Q.3 In which all languages you can code in Hadoop
Java
Python
C++
All of the above

Q.4 Whch of the file contains the configuration setting for HDFS daemons
yarn-site.xml
hdfs-site.xml
mapred-site.xml
None of the above

Q.5 All of the following accurately describe Hadoop, EXCEPT
Open source
Real-time
Java-based
Distributed computing approach

Q.6 Whch of the file contains the configuration setting for NodeManager and ResourceManager
yarn-site.xml
hdfs-site.xml
mapred-site.xml
None of the above

Q.7 Hadoop can be used to create distributed clusters, based on commodity servers, that provide low-cost processing and storage for unstructured data

True
False

Q.8 Which of the following is a distributed multi-level database?
HDFS
HBase
Both the above
None of the above

Q.9 Which of the following is used for machine learning on Hadoop?
Hive
Pig
HBase
Mahoot

Q.10 Which of the following is used to ingest streaming data into Hadoop clusters
Flume
Sqoop
Both the above
None of the above

Q.11 Hadoop distributed file system behaves similarly to which of the following:
RAID-1 Filesystem
RAID-0 Filesystem
Both the above
All of the above

Q.12 Zookeeper ensures that
All the namenodes are actively serving the client requests
A failover is triggered when any of the datanode fails.
Only one namenode is actively serving the client requests
A failover can not be started by hadoop administrator.

Q.13 Which of the following is used to ingest data into Hadoop clusters?
Flume
Sqoop
Both the above
Nonw of the above

Q.14 Which of the following is a data processing engine for clustered computing?
Drill
Oozie
Spark
All of the above

Q.15 Which tool could be used to move data from RDBMS data to HDFS?
Sqoop
Flume
Both the above
None of the above
‘

Q.1 Which of the following deal with small files issue
Hadoop archives
Sequence files
HBase
All of the above

Q.2 Which of the following feature overcomes this single point of failure
None of the above
HDFS federation
High availability
Erasure coding

Q.3 Which statement is true about NameNode High Availability
Solve Single point of failure
For high scalability
Reduce storage overhead to 50%
None of the above

Q.4 In NameNode HA, when active node fails, which node takes the responsibility of active node
Secondary NameNode
Backup node
Standby node
Checkpoint node

Q.5 What are the advantages of 3x replication schema in Hadoop
Fault tolerance
High availability
Reliability
All of the above

Q.6 What are the advantages of HDFS federation in Hadoop?
Isolation
Namespace scalability
Improves throughput
All of the above

Q.7 Rack Awareness improves

Data high availability and reliability
Performance of the cluster
Network bandwidth
All of the above

Q.8 Which property is used to enable/disable speculative execution
mapred.map.tasks.speculative.execution
mapred.reduce.tasks.speculative.execution
Both the above
None of the above

Q.9 In which process duplicate task is created to improve the overall execution time
Erasure coding
Speculative execution
HDFS federation
None of the above

Q.10 In which mode each daemon runs on a single node but there is separate java process for each daemon
Local (Standalone) mode
Fully distributed mode
Pseudo-distributed mode
None of the above

Q.11 In which mode each daemon runs on a single node as a single java process
Local (Standalone) mode
Pseudo-distributed mode
Fully distributed mode
None of the above

Q.12 In which mode all daemons execute in separate nodes
Local (Standalone) mode
Pseudo-distributed mode
Fully distributed mode
None of the above

Q.13 Which configuration file is used to control the HDFS replication factor?
mapred-site.xml
hdfs-site.xml
core-site.xml
Yarn-site.xml

Q.14 How to adjust the size of a distributed cache
local.cache.size.

mapred.cache.size.
hdfs.cache.size.
Distributedcache.size.

Q.15 What is the default size of distributed cache
8 GB
10 GB
12 GB
16 GB

Q.1 Distributed cache can cache files
Jar Files
Read-only text files
Archives
All of the above
Q.2 The total number of partitioner is equal to
The number of reducer
The number of mapper
The number of combiner
None of the above
Q.3 Which of the following Hadoop config files is used to define the heap size?
hdfs-site.xml
core-site.xml
hadoop-env.sh
mapred-site.xml
Q.4 Which of the following feature you will use submit jars, static files for MapReduce job during runtime
Distributed cache
Speculative execution
Data locality
Erasure coding
Q.5 Which of the following method used to set the output directory
FileOutputFormat.setOutputgetpath()
OutputFormat.setOutputpath()
FileOutputFormat.setOutputpath()
OutputFormat.setOutputgetpath()
Q.6 Which tool is used to distributes data evenly across datanode
Balancer
Disk balancer
Q.7 Which tool is used to distributes data evenly on all disks of a datanode
Balancer
Disk Balancer
Q.8 Which of the following must be set true enable diskbalnecr in hdfs-site.xml

dfs.balancer.enabled
dfs.disk.balancer.enabled
dfs.diskbalancer.enabled
dfs.disk.balancer.enabled

Q.9 In disk balancer datanode uses which volume choosing the policy to choose the disk for the block.
Round-robin
Available space
All of the above
None of the above

Q.10 Which among the following is configuration files in Hadoop
core-site.xml
hdfs-site.xml
yarn-site.xml
All of the above

Q.11 Which of the following is used for large inter/intra-cluster copying
fsck
distch
DistCp
dtutil

Q.12 Hadoop uses hadoop
Troubleshooting
Performance reporting purpose
Monitoring
All of the above

Q.13 Is it possible to provide multiple inputs to Hadoop?
Yes
No

Q.14 Is it possible to have Hadoop job output in multiple directories?
Yes
No

Q.15 Which of the following is used to provide multiple outputs to Hadoop?
MultipleOutputFormat
MultipleOutputs class
FileOutputFormat
DBInputFormat

****

Data Nuts & Bolts: Fundamentals of Data

----

1)Match the data backup methods with their descriptions.

a)Gets a backup of the data generated or revised since the last full backup

Differential backup

b)Gets a backup of the data generated or revised since the last backup, regardless of the type of the last backup

Incremental backup

c)Gets a backup of data for the past n years

Differential backup

d)Gets a backup for all data within the hard drive

Full backup

2)Match the concepts with their descriptions.

a)The ability to use your knowledge and experience to make good decisions and judgements

Wisdom

b)Contextualized, organized, and vetted data that convey some sort of trend or pattern

Information

c)Raw and unstructured facts, numbers, or figures which convey a message

Data

d)The application of information which is measured by the ability to “do things”

Knowledge

e)The ability to ask questions and learn new things

Information

3)Ravi wants to create a data visualization to show which parts of his company website are receiving the most clicks and are being most viewed by his viewers. Which data visualization will provide Dan with a visual that is easy to assimilate and make decisions from?

Heat map

4)Match each of the SQL codes with the functions that they perform.

a)Applies filtering logic to limit records in your results

WHERE

b)Describes how you want your data sorted

ORDER BY

c)Creates groups to summarize data

GROUP BY

d)Lists the columns you want to retrieve

SELECT

e)Name of the table to pull the data from

FROM

f)Applies filtering logic to your groups

HAVING

5)A university professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data is then displayed in a bar chart. What type of data is the university professor collecting?

Qualitative data

6)Which characteristics do all data migration projects have in common?

They all require identifying source and target systems

They all begin with data ingestions and cleansing of the data prior to integration

They all require performing proper ETL mapping to ensure consistency and compatibility

7)Alex is in the process of creating a report that displays the results of a survey. Which data type best describes the data that Alex is dealing with?

Observational

8)ABC Company wants to migrate their CRM data from their current legacy systems into newly purchased web-based CRM software. Place the data migration steps that they should perform in the correct order.

Planning

Analyzing the data

Design

Implementation

Testing

Final migration

****

Modern Data Management: Data Management Systems

----

1)Which is not a type of data that would be encountered in an enterprise?

Logical

2)What is the primary reason for data integration across domains?

Provide a unified single version of the data

3)In your multi-domain enterprise where the primary function is of a stock market broker and where you need real-time data synchronization, what will be the required style of architecture needed for the data management program?

Hybrid style

4)Which is not a function of entity resolution?

Data propagation

5)Suppose you have a company of 50 employees, and you are writing code for a very specific type of program. There are five vendors that provide you the graphical support and your target client are very small businesses. Which statement will be true in the domain of data management?

You do not need a data management program at all

6)Which is not a goal of metadata management?

Data reference

7)Which correctly describes the difference between static and dynamic data?

Static data does not require updating whereas dynamic data requires regular updating

8)What will be the best approach of data management for a multi-faceted enterprise with multiple domains?

A combination of top-down and middle out approaches

9)In your enterprise where you are planning to develop a perfectly aligned system across all the domains, which function will you deem as not necessary for building a truly aligned system?

Data Quality Program

10)What is the proper order for the levels of management and control from a Measurement of Maturity point of view?

Unstructured, Structured, Managed, Optimized

****

Modern Data Management: Data Governance

----

1)Which can be considered an advantage of IT systems in preventing security threats?

Access control is limited to a particular system

2)Which is NOT true about the entity resolution process?

Application of business and data quality rules is not a step in entity resolution

3)What is the integral meaning of data harmonization?

Data harmonization means obtaining a single version of the truth

4)Which is NOT a primal rule for data validation?

Governance practices

5)In building your data governance practice, which will not be an objective for the governance practice?

Dividing the business

6)Which is not the kind of job that is to be included in the role of a data steward?

Deciding which information is to be provided to which individual

7)Which are the two factors critical to an organization when considering the regulation of data privacy?

storage Networking Industry Association and general data protection regulation.

8)In your multi-domain enterprise where there are multiple vendors and multiple resources, which will be a key deciding factor in how you will manage the CRUD?

Enterprise resource planning

9)Which can be a major problem in companies that have a “Bring your Own Device” policy?

Leakage of information when the device is used in a public network

10) When implementing your data governance practice in your organization, which is the one factor that will NOT play a major role when considering the implementation?

Data quality

****

Modern Data Management: Data Quality Management

----

1)which is a performance measurement level?

Strategic management

2)which is not a data quality aimed project initiative?

Data integration

3)which function is a must for data compilance but not required at all for data management and governance?

Inter-relations

4)which is an essential factor to ensure good data quality?

Fix data quality issues

5)What are the advantages of a good data compliance strategy?

Proactive environment, organizational growth and customer relationships

6)what is meant by reference data?

Data that is used for classifaction of other data

7)In your data driven enterprise, how will you ensure that good quality data is baing maintained and reconciled across system?

Ensure source to target consistenty

8)which is a major solution to address the data governance and compliance issue?

Knowing the data

9)which is Not continuous improvement method in data governance?

Overlook achivable governance maturity milestones

10)which is NOT advantage of data lakehouse?

Provides an on-premises data warehouse solution

**s

Data Warehouse Essential: Concepts

1)Identify essential table types that we can use to implement a Star schema.

➔Fact table
Dimension table

2)Which statements about cloud-based data warehouse are true?

➔ Cloud-based Data warehouses are scalable

Cloud-based Data warehouses are elastic

3) Identify the essential levels of management that require strategic reports.

➔Strategic

Middle management level

4) Match the data modelling strategies with their features.

Answer Options:

- • A:Combines tables
- • B:Adds redundant data
- • C:Used on previously normalized databases to increase performance
- • D:Applies formal rules to enforce dependencies
- • E:Ensures the dependencies are properly enforced
- • F:Splits tables

➔ Normalization ➔Applies formal rules to enforce dependencies
Ensures the dependencies are properly enforced
Splits tables

Denormalization ➔ Combines tables
Adds redundant data
Used on previously normalized databases to increase performance

5) Choose the characteristics of weighted reports.

➔ Weighted reports multiply all the facts by weight before aggregating

With weighted reports we get a meaningful subtotal and total

6) Which processes involved in a data warehouse project are important?

➔Data cleansing

ETL

7) Which are essential tasks that we can execute to facilitate business intelligence?

➔ Extract

Load

8) Which of the following local and global warehouse statements are true?

➔ We can provision a single global repository for a particular domain
Selected

Local data warehouse cannot be accessed globally

9) Select data modelling strategies that we can adopt to create an ER model.

➔ Normalization

Denormalization

10) Which of the following OLAP statements are true?

➔ OLAP is a part of the overall Data warehouse implementation

OLAP provides real-time analytical capability

11) Identify the terminologies that we generally use in data warehousing.

➔ETL

Dimensional model

12) Identify essential features of Strategic Information.

➔ Preserves data integrity

Time-variant

13) Identify some of the essential features which differentiates a data warehouse from OLTP.

➔ Data warehouse stores historical data

Data warehouse provides predictive analytical capabilities

14) Identify the essential differences between RDBMS and data lakes?

➔ RDBMS databases defines fixed schema while data lake follows no schema design

RDBMS databases are transaction while data lakes are not transactional

15) Which statements about Snowflake schemas are true?

➔ A Snowflake schema is an extension of the Star schema

Application binary interface allows us to contextualizes contracts

16) Identify the essential components of Azure data lake.

➔ SQL server

Data factory

17) Match the data warehousing solutions with their associated benefits.

Answer Options:

- • A:Cost effective
- • B:Offers absolute control over security
- • C:Provides scalability
- • D:Offers better speed and connectivity
- • E:Preferred in banking or government domains

➔ On-premise data wareshouse➔

Offers absolute control over security

Preferred in banking or government domains

On-cloud data warehouse➔ Cost effective

Provides scalability

Offers better speed and connectivity

Data Warehouse Essential: Architecture Frameworks & Implementation

1)Specify some of the outcomes of a data warehouse realization.

➔ Intuitive dashboards

Predictive analytical reports

2)Specify some of the essential logical components of a data warehouse.

➔Entities

Attributes

3)Which of the following components are provided by Talend to design ETL jobs?

➔tFileInput

tMap

4) Which of the following statements correctly defines the characteristics of the Kimball model?

➔ In the Kimball model the analytical systems can access the data directly

Kimball model uses the dimensional model

5) Identify essential tasks involved in an ETL process.

➔ Transforming extracted data

Extracting data from diversified sources

6) Which of the following dimensional components differentiates a Dimensional model from an ER model?

➔ Dimension tables

Fact tables

7) What are some of the essential tasks that we can execute using Talend?

➔ Business modelling

ETL job designing

8) What are some of the integrated components of a data warehouse?

➔ Data staging

Storage

9) What are some of the important tasks that we need to perform to implement a Physical model for a given Logical model?

➔ Create Foreign keys to establish the relationships among objects

Create tables to represent the entities

10) Select prominent ETL tools that we can use with a data warehouse implementation.

➔ Talend

PowerCenter Informatica

Modern Data Warehouses

1) You have an on-premises data warehouse already installed in your organization. In the period following the COVID pandemic, your business started to grow exponentially, and you were tired of adding more nodes to the physical storage of the data warehouse. You decide to modernize your data warehouse and make it cloud-based. What major advantage will you achieve by modernizing your data warehouse?

➔Speed up time to analytics

2)You plan to implement a modern data warehouse solution into your enterprise. You have understood the proper data management and governance issues. You have set up all your domains and data ingestion methods. Now you plan to make a central repository of all your files. What should be the next step for the implementation of the data warehouse solution?

➔Selection of the nature of the data

3)You have a very well-run data management program in your organization, which is very secure. You use analytics for making big data driven decisions. In recent months, you realize that whatever decisions you are making based on the analytics are proving to be wrong and harmful for the organization. After consulting with the data analysts and data stewards you conclude that it is due to poor data quality. What should be the next step of action?

➔Install a firewall

4)Other than gaining real-time insights of the data, what is another major advantage of streaming analytics?

➔Real-time dashboards

5)BigQuery is a cloud-native warehouse that is also a fully managed data warehouse. What is the major advantage of BigQuery over

Amazon Redshift that may be a deciding factor for the selection of a data warehouse?

➔Access control allows improved data sharing

6)Your organization has an effective sales team that is backed up by analytics that help accelerate the process of sales from the initial contact. One of the primary reasons for its effectiveness is a data input tool that hastens the process of data entry by providing preset suggestions. What are these suggestions commonly known as in data science terminology?

➔Reference data

7)You have incorporated the usage of Amazon Redshift in your organization, and you don't want your data to be corrupted by processing. Therefore, you want the data to be stored in raw format before the processing is done, which is a service offered by Amazon Redshift. What is the key benefit of storing data in raw format?

➔Minimal loss of data

8)What is **not** a component of an on-premises data warehouse?

➔Processing space

9)You have an automobile company that helps sort out vehicles and make monthly sales versus expenditure reports. What is the best way to handle the data for centrally storing it?

➔Batch processing

10) In a candlestick chart, you see the share price of your company falling. You have implemented a streaming analytical tool that helps in analyzing and dashboarding the data as it is produced. You realize that the candlesticks pattern has consolidated and is not responding well to the influx of data. What is the source of this problem?

➔Server downtime

Azure Databricks & Data Pipelines

1)In your enterprise, customer information needs to be readily available to indicate whether the information given by customers is valid or not, and to see the potency of a positive deal. Which factors would you consider a priority when selecting the appropriate data pipeline tool?

➔Batch-wise vs. real-time ingestion and analytics

2)What are the major advantages of a cloud warehouse solution over an on-premises data warehouse solution?

➔Low cost

Less worry about storage

3)What are the two different data pipeline tools that address specific job roles?

➔Data engineers and analysts

4)Place the steps for a typical Azure Databricks warehouse in the correct order.

➔Ingest

Store

Prep and train

Model and serve

5)The reason why Azure Databricks is so easy to use is because it is universal and is integrated with Microsoft's server for better parsing of information. Which platform has the same source of origin as Databricks, which gives it an analytical advantage over other platforms?

➔Appachi

6)What are the major disadvantages of Snowflake that might be troublesome for a few companies that seek data categorization?

➔Fewer options with geospatial space

No option for un-structured data

7)While setting up an integrated data pipeline for your enterprise to facilitate data ingestion in the warehouse, what should you place more emphasis on from a business perspective?

➔Analytics and business intelligence

8)Suppose you have a full-blown data management program with a well-running data warehouse and an optimum data pipeline tool that facilitates data transformation and transmission. While measuring the maturity of the data pipeline tool, what will be the sole factor that will determine the efficiency of the data pipeline tool?

➔Reliability and scalability

9) What is the one difference that separates the model of the Snowflake data warehouse from all the other data warehouse solutions?

➔Hybrid model

10) What is **not** a design component of a data pipeline?

➔Data integration

Installation & Introduction

1)Select the main dependency that has to be installed for Talend to be installed.

➔Java

2) Select the supported OSs by Talend Open Studio

➔Windows 64

MacOS

3) Select the main parts of the default UI of Talend Open Studio.

➔Repository

Palette

4) When installing MySQL relational Database to be used with Talend, select the configuration to be performed once the installation of the components is complete.

➔MySQL server port

MySQL root password

5) Select the folder that contains all the project information for a job that is exported from Talend studio.

➔process

6) Rank the steps required to for any job in Talend studio.

➔Create job in the repository

Add components from palette to the design space

Configure the components properties

Run the job

7) Select the correct description of Talend Metadata Bridge.

➔Synchronize metadata across data pipelines

8) Match each description with the variable in Talend studio.

Answer Options:

- A:studio provisions through components used in jobs integration
- B:execute jobs with parameters for different environments
- C:Ad-hoc variables that can be configured in jobs

➔Job contexts variables➔execute jobs with parameters for different environments

Studio global variables➔studio provisions through components used in jobs integration

User defined global variables➔Ad-hoc variables that can be configured in jobs

Transforming Data

1)Select the correct differences between XML attributes and elements.

➔Elements can contain tree structure

Elements can have multiple values

2) Select the component used to generate an XML file from a CSV file in Talend studio.

➔tFileOutputXML

3) Select the exit code value the signifies the successful completion of a job in Talend studio.

➔0

4) Select the tMap Component description in Talend studio.

➔congregate input data to output data

5) Select the option to be enabled to allow the specification of 2 schemas for an XML input file in Talend studio.

➔Enable XPath in column "Schema XPath loop" but lose the order

6) In order to generate a complex XML file where data is specified using attributes of elements and elements trees in Talend studio, which component allows such.

➔tAdvancedFileOutputXML

7) Select the component used to perform lookup data in Talend studio.

➔tJoin

8) Select the component access a MySQL database in Talend studio.

➔tMysqlInput

9) Select the tool that allows specifying the relation of multiple tables as data sources when reading data from a database as input in Talend studio.

➔SQL Builder

10) Match the attribute value with its attribute that will only Add new records or modify existing ones without modifying the table structure or other records already exist in the table when writing data to a database table in Talend studio. Two options are invalid.

Answer Options:

- A:Default
- B:Insert or update
- C:Select
- D:Delete

➔Action on table➔Default

Action on data➔Insert or update

11) Select the component that allows updating data in a database in Talend studio.

➔tMySQLRow

12) Select the components and concepts to facilitates accepting as input multiple databases in Talend studio.

➔tMap

Lookup

13) Select the component used to combine multiple database records to a single records in Talend studio.

➔tDenormalize

Data Mapping

1)Which of these statements accurately describes the Unique Match Join match model in the Map Editor?

➔Last match is considered and passed to the output

Set as default when configuring an explicit Join

2) You are building an expression in Map Editor for a column in which you want to pad the row1.CustomerID string with leading zeroes up to maximum length of 6 characters. Which code can you use to accomplish this for you?

➔String.format("$06d",row1.CustomerID)

3) You have Map Editor open for a tMap object for which you are mapping databases objects. You click the Join Model for a table Join property. Which options appear in the Options dialog that appears?

➔Left Outer Join

Inner Join

4) You are building a filter expression in Map Editor for a column in which you want to filter the product name, prd.name to equal "Turbo Widgets", and you want the transaction quantity, tx.qty to be greater than 100. Which code can you use to accomplish this for you?

➔prd.name.equals("Turbo Widgets")&&tx.qty>100

5) Which common FTP operations are supported by Talend Open Studio components available in the Palette?

➔File Exist

Delete

Put

6) On the Component tab displayed for the tSortRow component, when you click to add criteria to the Criteria table, Talend automatically populates the column values with defaults. In the "sort num or alpha?" column Talend has chosen num by default for customer_id as displayed. Which other values are available for the "sort num or alpha?" column when you click to open the dropdown list for that column?

➔alpha

Data

7) You are using a tExtractDelimitedFields component to split the Address2 field in the delimited file as displayed. What must you specify as the Field separator property for the tExtractDelimitedFields component to properly split the Address2 field?

```
ID;First_Name;Last_Name;Address1;Address2;Country

1;Claudia;Sand;10000 Main Dr NW;New York, NY;USA

2;Max;Bigot;60000 My St;Nashua, NH;USA

3;Rick;Tailleur;200000 Younge St;Toronto, ON;Canada

4;Noémie;Miller;500000 St. Catherines St Ouest;Montreal, QC;Canada

5;Catherine;Reilly;100000 Main St SW;Boston, MA;USA
```

➔"," (a comma)

8) When configuring the properties for a tReplace component, you can optionally click the Advanced mode checkbox. Doing so allows you to specify what type of expression as the Pattern to search for?

➔Regular expression

9) When configuring the properties for a tAggregateRow component. You are going to Group by the customer_id field in order to aggregate

the sales on a customer_id basis, so that in the resulting output file you will have one row for each customer_id with aggregated sales figures. Which Function value must you choose when configuring the Operations table for this tAggregateRow component?

➔sum

10) You are using a tNormalize component to normalize the category field in the delimited file as displayed. What must you specify as the Item separator property for the tNormalize component to properly normalize the category field?

```
product_id,product_name,category,unit_price

1,Regular Widget,Normally Aspirated;Low Price;Highly Reliable,125

2,Super Widget,Super;Medium Price;Very Reliable,250

3,Turbo Widget,Turbo;Medium Price;Very Reliable,275

4,S/T Widget,Super;Turbo;Premium Price;Reliable,425

5,Hybrid Widget,Hybrid;Normally Aspirated;Premium Price;Reliable,425
```

➔";" (a semi-colon)

Getting Started with Python: Introduction

1)Which of the following commands are valid to store a numeric value of 2 in a variable named num_x?

➔num_x = 2

num_x = 6 – 4

2) How can you execute shell commands on Jupyter notebook code cells?

➔Prefix the shell command using ! i.e. !python –version

3) Which of the following functions are valid built-in functions in Python?

➔print()

len()

type()

4)Which of the following is an open source distribution of the Python and R programming languages that uses the Conda package manager?

➔Anaconda

5) Consider this bit of Python code:

Code Editor:

```
num_1 = 10

num_2 = 20

num_3 = num_1

num_1 = 100
```

What is the final value stored in num_3?

➔10

6) If you want to increment the value stored in the num_1 variable by 10 which of the following Python statements are valid?

➔num_1 = num_1 + 10

num_1 += 10

7) What is the correct syntax for specifying multi-line strings in Python?

➔“”” This is a multi-line string “””

8) Which of the following are valid string operations in Python?

➔“Hello” + “World”

“Hello” * 3

9) As a new user of Python which version of Python should you use?

➔Either version of Python is fine but 3.7 should be preferred

10)Which of the following terms best describes Jupyter notebooks that you can use to write Python code?

➔Browser-based

Interactive

Can view results on the same screen as the code

11)What is the output of the following bit of code?

➔2

12) Consider two Python variables initialized as shown below. Which of the following logical statements below will have a value of True?

Code Editor:

```
a = True

b = True
```

➔a or b

a and b

13) Which of the following are valid data types in Python?

➔int

Float

Bool

Complex Data Types in Python: Working with Lists & Tuples in Python

1)If you wanted to sort the elements in the list names_list in alphabetical order which of the following statements in Python are valid?

➔names_list.sort()

names_list = sorted(names_list)

2)Which of the following lines of code will print this string in reverse i.e. print out “olleH”

Code Editor:

some_str = “Hello”

➔some_str[::-1]

3) If you want to count the number of times the name “John” appears in the names_list what function would you invoke?

➔names_list.count(“John”)

4) What will be the result of this slicing operation of the names_list?

Code Editor:

names_list = [‘John’, ‘James’, ‘Lily’, ‘Emily’, ‘Nina’]

names_list[::2]

➔[‘John’, ‘Lily’, ‘Nina’]

5)What is the result of executing this code?

Code Editor:

some_string = “Python”

a, b, c, d = some_string

➔This is an error, “too many values to unpack”

6)What is the result of executing this code?

Code Editor:

```
city = ‘Los Angeles’

city.find(‘x’)
```

➔-1

7) Consider the list:

Code Editor:

```
some_list = [‘a’, ‘b’, ‘c’, ‘d’, ‘e’, ‘f’]
```

How do you slice this list to access the elements ‘c’, ‘d’?

➔some_list[2:4]

8)If you wanted to insert an element at index 2 in a particular list named names_list what is the function that you would invoke?

➔names_list.insert(2, “John”)

9)All of the following statements are ways in which lists and tuples are different. Which one of these is true?

➔A list can be changed once creating, a tuple is immutable and cannot be changed

10) Which of the following statements about Python lists are true?

➔Lists are ordered collections

Lists in Python can have elements of different data types

11) All of the following statements are ways in which lists and tuples are similar. Which one of these is NOT true?

➔Both, once created, cannot be updated

12)Which of the following are valid complex data types in Python?

➔Sets

Dictionaries

List

Complex Data Types in Python: Working with Dictionaries & Sets in Python

1)Consider a dictionary of names and ages set up as below:

Code Editor:

```
names_ages = {‘John’: 35, ‘Jim’: 45, ‘Alice’: 25}
```

and a second dictionary as below:

Code Editor:

```
updated_names_ages = {‘Ella’: 29, ‘John’: 36}
```

How would you update the names_ages dictionary with the values in updated_names_ages dictionary?

➔names_ages.update(updated_names_ages)

2) Consider a nested list of names and ages:

Code Editor:

```
names_ages = [[‘John’, 35], [‘Jill’, 38], [‘Tim’, 27]]
```

How would you convert this to a dictionary with names as the keys and ages as values?

➔dict(names_ages)

3)Consider two sets of integers:

Code Editor:

```
set_1 = {2, 4, 6, 8}

set_2 = {1, 2, 5, 6, 7, 8}
```

What operation would I run to get a result set with all of the elements from both sets?

➔set_1.union(set_2)

4)Consider a dictionary of names and ages set up as below:

Code Editor:

```
names_ages = {‘John’: 35, ‘Jim’: 45, ‘Alice’: 25}
```

How would you look up Alice’s age in this dictionary?

➔names_ages[‘Alice’]

5)A set in Python can contain which of the following data types?

➔Tuples

Floats

Strings

6)Consider a dictionary of names and ages set up as below:

Code Editor:

```
names_ages = {‘John’: 35, ‘Jim’: 45, ‘Alice’: 25}
```

What would the output be if you were to run this code?

```
names_ages[‘Tim’]
```

➔KeyError: ‘Tim’

7)Consider a nested list of names and ages:

Code Editor:

```
names_ages = [[‘John’, 35], [‘Jill’, 38], [‘Tim’, 27]]
```

How would you access Tim’s age in this nested list?

➔names_ages[2][1]

Conditional Statements & Loops: If-else Control Structures in Python

1)Consider three variables with values as shown:
a = 5
b = 10
c = “five”

What are the results of evaluating the conditional expression?
a == c
a >= b
not(a < b and a > b)

➔False, False, True

2)How is the body of an if-statement block syntactically represented in Python?

➔Using additional indentation from the left relative to lines just before and after the block

3)What is the output for this code?

```
if 'bin' in {'float': 1.2, 'bin': 0b010}:

    print('a')

    print('b')

print('c')
```

➔a b c

4) What is the output for this code?

```
if None:

        print('Hi')
```

➔Nothing is printed - no output

5) Evaluate the expression provided. What does the following expression evaluate to?

```
'1' + '2' if  '123'.isdigit() else '2'  + '3'
```

➔’12’

6)What is the output of the code?

```
a = [1, 'one', {2: 'two'}, 3]
```

```
b = len(a)

if b == 4:

   print('Length of this list is 4')

   if b == 5:

       print('Length of this list is 5')

   else:

         print(b)
```

➔Length of this list is 4 4

7) What is the value of b in the snippet of python code?

```
a = “six”

b = (int(a), float(a))
```

➔ValueError: invalid literal for int() with base 10: 'six'

8)Consider the following snippet of Python code:

```
a  =  “40.6 ”

b  =  “60.4 ”

c  =  a + b
```

What does c evaluate to?

➔‘40.6 60.4’

9) What would the output of the following code snippet be?

```
num_one = 76

num_two = 23.4

print("datatype of num_one:", type(num_one))

print("datatype of num_two:", type(num_two))
```

➔datatype of num_one: <class 'int'> datatype of num_two: <class 'float'>

10) What is the output of the code snippet below?

```
value = 4

a = str(value)

b = a + “^” + “2”

c = a + “^” + “3”

print(value, “+”, b, “+”, c)
```

➔4 + 4 ^ 2 + 4 ^ 3

11) What do the values of d[0], d[1], d[2], d[3] evaluate to after the execution of the Python code below?

```
new_list = [“Red”, “Blue”, “White”, “Green”]

z = sorted(new_list)

d = list(z)

d[0], d[1], d[2], d[3] = d[3], d[2], d[1], d[0]
```

➔“White”, “Red”, “Green”, “Blue”

12) What is the output of the program below?

```
var = "hi"

if(type(var) == int):

    print("Type of the variable is Integer")

elif(type(var) == float):

    print("Type of the variable is Float")

elif(type(var) == complex):

    print("Type of the variable is Complex")

else:

    print("Type of the variable is Unknown")
```

➔Unknown

13) What is the output of the program below?

```
total_classes = 100

attended_classes = 67

attendance = (attended_classes/total_classes)*100

if attendance >= 75:

    print ("You are eligible to appear for the test.")

else:

    print ("Sorry, you are ineligible to appear for the test.")
```

```
➔Sorry, you are ineligible to appear for the test.
```

Condition Statements & Loops: The Basics of for Loops in Python

1) Which TWO of the following statements about for loops in Python are TRUE?

➔ They can iterate over the elements in tuples, lists, and dictionaries

They may have an associated else block

2) What is the correct value of x given the assignment shown?

```
x = list(range(-17, -7, 2))
```

➔ [-17, -15, -13, -11, -9]

3)Which of the following function calls will generate the list below?
[10, 7, 4, 1, -2]

➔ list(range(10, -3, -3))

4) What is the maximum value in the sequence x?

```
x = range(2, 14)
```

➔13

5) Given a variable my_dict which is a dictionary, consider you use it in a for loop in this manner:

```
for x in my_dict:

    print (x)
```

What are the contents printed out?

➔ The keys in the dictionary

6) Which of these Python data types can NOT be iterated through using for loops?

➔int

7) Given the following code, what is the type of x which is printed out in each iteration?

```
my_list = [[‘tiger’, ‘lion’, ‘leopard’], [‘camel’, ‘llama’, ‘alpaca’], [‘zebra’, ‘donkey’, ‘wildebeest’]]

for x in my_list:

    print(x)
```

➔ A list of strings

Functions in Python: Introduction

1)Which of the following statements about functions is true?

➔A function is defined using the “def” keyword

Function code is not executed when defined

2) Which of the following are valid function names in Python?

➔_123_Function_Name()

functionName()

_function_name()

3) Which of the following statements about functions is false?

➔Functions cannot access variables which are declared outside the function

4)onsider a function definition which looks like this:

```
def some_function(a, b, c):
    print(a, b, c)
```

5)Which of the following function invocations are correct?

➔some_function(2, 3, “Hello”)

some_function(2, 3, 4)

Consider the following bit of code. What will be the result of executing this bit of code?

```
x = 3
y = 4
def add(a, b):
   result = x + y
   print(result)


add(10, 20)
```

➔7

6) Which of the following statement(s) about positional arguments to functions is/are true?

➔They can be of any data type – primitive or complex types

A function can accept any number of positional arguments

7)What is the default return value from a function when no return statement is specified?

➔None

8) Which of the following statement(s) about return values is/are false?

➔A function can have just one return statement

A function with input arguments cannot have a return statement

A function has to have a return statement

9)Which of the following statements(s) about the data types of return values is/are false?

➔A return statement is mandatory in functions

10)Which of the following are valid kinds of input arguments for Python functions?

➔Positional arguments

Keyword arguments

11)Which of the following are some of the advantages of using keyword arguments to invoke functions?

➔Keyword arguments can be specified out of order

Easier to maintain code since the value of each argument is clearly seen during invocation

12)What does this function definition indicate?

```
def some_fn(a, b, c=True):

    print(a, b, c)
```

➔a and b are required arguments, c is optional

13)Which of the following function definitions allows the function to accept variable length arguments?

➔some_fn(*args)

14)Which of the following are valid types of arguments in Python?

➔Keyword arguments

Variable length arguments

Default arguments

15)Which of the following statement(s) about variable length arguments in Python is/are true?

➔Variable length positional arguments are passed into the function as a tuple

Variable length keyword arguments are passed into the function as a dictionary

Python - Introduction to NumPy for Multi-dimensional Data

1)Which of these correctly match the following libraries in the Numpy ecosystem with what that library is used for?

➔Data visualization tool used for plotting 2-D graphs➔ Matplotlib

Contains a collection of algorithms used for image processing➔Scikit-image

Used to perform statistical operations➔Statsmodel

2)Let’s say you have imported the numpy package as np and you want to assign the variable “x” with a 3 by 2 array of type integer, all of whose values are 1. Which of these commands will you use to do so?

➔x=np.ones((3,2),dtype=np.int32)

3)What will be the value stored in the variable y after we have executed the following code

```
import numpy np

y=np.arange(2,4,0.5)
```

➔[2, 2.5, 3, 3.5]

4)Let’s say you have imported the numpy package as np and you want to print the first 2000 natural numbers in the form of an array and you want all 2000 of the numbers to be visible on screen when printing (including the number 2000). Which of these commands would you use to do so?

➔np.set_printoptions(threshold=np.nan) print(np.arange(1,2001))

5)What would be the output of the following code:

```
import numpy as np

x=np.array([[1,2] , [3,4]])

y=np.array([ [5,6],[7,8]])

z=(x*y)

z
```

➔[[5,12], [21,32]]

6) What would be the output of the following section of code:

```
Import numpy as np

x=np.array([4,6,2,8])

np.median(x)
```

➔5

7) Which of these slicing operations can be used to quickly get the reversed contents of a numpy array called “array”?

➔array[::-1]

8) Match the following features of the numpy nditer function mentioned here with the correct Boolean category

➔True➔Using this function, we can iterate through each of the individual elements of the array passed as an input argument

False➔By default, the nditer object returns arrays that can be written on

The nditer can accept only one dimensional arrays as input

9) Which of these statements regarding the ravel() object in Python are true?

➔The ravel function reduces a multi-dimensional array to a single dimensional array

ravel() belongs to the numpy library and can be used on any object that can be parsed

Python - Advanced Operations with NumPy Arrays

1) Let’s say you have a two dimensional numpy array called “twod” and you want to split it row-wise into two equal halves. Then, which of these numpy functions would you call on it to do so?

➔vsplit(twod,2)

2) Some of the features of digital images in Numpy are given below. Which of these are true?

➔In numpy, images can be represented as a 3D matrix where the first two dimensions represent the pixels in the image that are arranged in the form of a grid and the third dimension specifies the number of channels for the image

A digital image is a multidimensional array and every pixel in a digital image is represented by a number

3)Let’s say you have an image that you have split into two equal halves along the x axis. You have stored these two halves of the original image in the variables x1 and x2 respectively. Which numpy function would you use to combine these two halves to reconstruct the original image?

➔concatenate((x1,x2),axis=1)

4)Let’s say you have a numpy array called “array_1” and you initialize “another array called “array_2” with the help of a following command:

array_2 = array_1.view()

Match the following statements about “array_2” with the correct Boolean value

➔True➔The base for array_2 points to the same object as array_1

array_1 and array_2 contain the same elements

False➔array_2 points to the same object as array_1

If we re-assign array_2, then we will end up re-assigning array_1 as well and change its contents.

5) Let’s say you have a numpy array called “array_3” and you initialize “array_4” with the help of a following command:

array_4 = array_3.copy()

Match the following statements about “array_4” with the correct Boolean value

➔True➔array_3 and array_4 contain the same elements

False➔If we change a single element of array_4, then the corresponding element in array_3 changes too

If we re-assign array_4, then we will end up re-assigning array_3 as well and change its contents

Changing the shape of array_4 will change the shape of array_3 as well

6) Let’s say you have a 1-D numpy array called “cubes” consisting of the cubes of the numbers 1,2,3 and so on till 10. What would be the value of the array:

cubes [[[4, 5], [1, 2]]]

➔[[125, 216] , [8, 27]]

7)Some of the features of Pandas is given below. Which of these are true?

➔The column header of a Pandas dataframe can be treated in the same way as the index label of a numpy array

A particular column of a pandas dataframe can be referenced by its column header

8) Let’s say you imported numpy as np and you have initialized a 1-D array of integers called “array”. What would np.all (x < 50) return?

➔This function would return a true boolean value if all the entries in your array are less than 50 and false otherwise

9)Let’s say you have a Pandas dataframe called “phone_data” which contains the data of various phones released in 2018 and their prices. It has the following three columns:

“manufacturer”, “phone name” and “ price”.

You want only the names of all the phones that are priced more than 10,000. Which of these commands can be used to print these values?

➔phone_data[phone_data[‘price’] > 10000][‘phone name’]

10)What are the conditions under which broadcasting can take place between two elements in Numpy?

➔Broadcasting works when at least one of the elements is a scalar

A smaller array can be broadcast on a larger array only when the corresponding dimensions of the two arrays being operated upon are compatible i.e. when the corresponding dimensions are equal or one of the two dimensions is 1

11)Match the following statements about broadcasting with the correct Boolean value:

➔False➔The array [[1, 2] , [3, 4]] and the scalar 10 are incompatible with broadcasting

True➔The array [[1, 2] , [3, 4]] and the array [[1], [2]] are compatible with broadcasting

The array [[1, 2] , [3, 4]] and the array [1, 2 ,3] are incompatible with broadcasting

The scalar 10 and the scalar 20 are compatible with broadcasting

Python - Introduction to Pandas and DataFrames

1) In the following Python code, typing which Python command will give the user the CEO of Facebook?

```
import pandas as pd

companies_ceo = {

                     'Amazon' :  'Jeff Bezos'

                     'Apple' : 'Tim Cook‘,

                     ‘SpaceX’: ‘Elon Musk‘

                     ‘Facebook’: ‘Mark Zuckerberg’

                     ‘Netflix’: ‘Reed Hastings’

                    }

companies_ceo_series= pd.Series(companies_ceo)
```

➔ companies_ceo_series[3]

companies_ceo_series[‘Facebook’]

2)In the following Python code, typing what command will create a DataFrame called “companies_ceo” whose first column has all the entries of the ‘companies’ list and whose second column has all the entries of the ‘ceo’ list, with the column names as the names of the respective variables?

```
import pandas as pd

companies = {

'Amazon'

'Apple'

‘SpaceX’

‘Facebook’

‘Netflix’

          }

ceo = {
```

```
'Jeff Bezos'

'Tim Cook‘,

‘Elon Musk‘

‘Mark Zuckerberg’

‘Reed Hastings’

  }
```

➔ companies_ceo_tuple = list (zip(companies, ceo)) companies_ceo = pd.dataframe(companies_ceo_tuple, columns=[‘companies’, ‘ceo’])

3)What happens when we call the stack () function on a Pandas DataFrame

➔ It will create a new DataFrame such that a single row in the original DataFrame is stacked into multiple rows in the new DataFrame depending on the number of columns for each row in the original DataFrame

4)Which of these correctly match the following libraries in the Numpy ecosystem with what that library is used for?

➔ Specifically meant for machine learning, data mining, and data analysis➔ Scikit-learn

Data visualization tool used for large datasets➔ Bokeh

Used to perform statistical operations➔ Statsmodel

5) Match the following statements related to the iloc indexer in Pandas with the correct boolean values.

➔ True➔ The iloc indexer is similar to the loc indexer and can be used to access records located at a particular index in a Pandas DataFrame

False➔ When we pass 2:6 as input argument to the iloc function, we get all details of the records located in the second index all the way up to the 5th index of the DataFrame

The column headers can be passed as input arguments in the form of a string to the iloc function without any errors

6)Let’s say you have saved a dataset in a pandas DataFrame called “dataset” which has tons of records and you only want to access the details of the records in only the 5th, 8th and 14th index. Which of these Python commands can you use to do so?

➔ dataset.loc[[5,8,14],:]

dataset.loc[5,8,14]

7)Let’s say you have a pandas DataFrame called “panda” which has 8 rows in total and you want to remove the last row from this DataFrame. Which of these Python commands would you use to do so?

➔ panda.drop(panda.index[7])

8)Which of these statements related to the pivot function in Pandas is true?

➔ The combination of the row index and the column header must be unique in order to generate a pivot table

The Pivot function summarizes the details of each column in a DataFrame

9)Match the following statements related to Pandas DataFrames with the correct boolean values.

➔ True➔ All the data within a particular column in a Pandas DataFrame must be of the same data type

False➔ Data in different columns of a Pandas DataFrame cannot be of different data types

Once a Pandas DataFrame has been created, it is not possible to add a new column to this DataFrame

10)Match the following statements related to the concept of multiIndex in Pandas with the correct Boolean values.

➔False➔ The MultiIndex for a row is some composite key made up of exactly one column

True➔ MultiIndex lets the user effectively store and manipulate higher dimensional data in a 2-dimensional tabular structure

MultiIndex is useful when we have large datasets where using numeric indexes to refer to each record is unintuitive

11)Which of these statements related to the Pandas Series object are true?

➔ Pandas Series object is similar to a Python list

Once we create a Pandas Series object, an index representing the positions for each of the data points is automatically created for the list

12)Let’s say you have a pandas DataFrame called “frame” and you want to export this DataFrame along with its index as a CSV file called “data_frame” located in the datasets folder of our workspace.

➔ frame.to_csv(‘datasets/data_frame.csv’)

Python - Manipulating & Analyzing Data in Pandas DataFrames

1) Consider the following Python code. What command would you use to iterate through the “companies_ceo” DataFrame and print the list of all the CEOs in this DataFrame?

```
import pandas as pd

companies = {

   'Company' : [‘Facebook', 'Apple', ‘Amazon', ‘Netflix’],



            'CEO' : [‘Mark Zuckerberg’, 'Jeff Bezos’, 'Tim Cook', ',’Reed Hastings’ ],



            }

companies_ceo = pd.DataFrame(companies)
```

➔ for row in companies_ceo.itertuples(): print(row.CEO)

for row in companies_ceo.iterrows(): print(row[1])

2)Which of the following formats does Pandas not support natively when exporting the contents of a Dataframe?

➔ JPEG

3)Let’s say you have created a Pandas DataFrame called “unsorted” and you want to sort the contents of this DataFrame column wise in alphabetical order of the header name. Then, which function would you call on the “unsorted” DataFrame to do so?

➔ unsorted.sort_index(axis=1)

4)Match the following functions that you can call on a Pandas DataFrame correctly with what they do

➔ Returns a Boolean array containing true or false values and returns the value in a cell as true if it contains NaN➔ .isnull()

Returns a Boolean array containing true or false values and returns the value in a cell as true if it does not contain NaN➔ .notnull()

Every cell in the Dataset which has a NaN value will be replaced with 0➔ .fillna(0)

All the rows which contain a NaN value in any cell of that row are removed➔ .dropna()

5)Match the following statements related to the .xs function in Pandas DataFrame with their correct Boolean values.

➔ True➔The .xs function is used when our Pandas DataFrame makes use of a MultiIndex

By default, the .xs function only takes a look at values in the first level index

False➔The .xs function cannot be used to return a cross section of columns

6) Let’s say you have imported Python as pd and have instantiated two DataFrames called “frame_1” and “frame_2” with the exact same schema. What command will you use to combine these two DataFrames into a single DataFrame and make sure that the combined DataFrame has its own unique index?

➔ pd.concat([frame_2, frame_1], ignore_index = True)

pd.concat([frame_1, frame_2], ignore_index = True)

7)The ‘how’ argument in the Pandas merge function allows us to specify what kind of join operation we want to perform on the given Pandas DataFrames. What are the valid values that we can give for this argument?

➔ outer

Left

Right

Inner

8)Some statements related to working with SQL Databases in Python are given below. Match them with their correct Boolean values.

➔ True➔ The sqlite3 library in Python allows us to create Databases on our local file system

All the changes that we make to an SQL database on a Jupyter notebook by connecting with it, will be committed to the database only after we execute sqlite3’s .commit() function

False➔ Once we have created a table, we can use sqlite3’s .execute() function to recreate the same table with the same table name so that we have duplicates of a table

Big Data Concepts: Getting to Know Big Data

1)What has Amazon been able to achieve by utilizing big data?

➔Gathering of information on what each customer is likely to purchase based on what other people with similar interests have purchased

Gathering of data on the search patterns of its customer

2)What could be accomplished by using big data technologies?

➔Cost reductions

New product development and optimized offerings

3)Four of the seven characteristics of big data are listed. Match each characteristic with its description. One description will not be used.

➔Velocity➔The speed at which data is processed and becomes accessible

Veracity➔Making sure the data is accurate, which requires processes to keep bad data from accumulating in your systems

Variety➔The different types of data from XML to video to SMS

Volume➔The amount of data that exists

4)Which statements are true about unstructured data?

➔Unstructured data is information that does not have a predefined data model

Common examples of unstructured data include audio, video files, or No-SQL databases

5)Which statements are correct about how Netflix utilizes big data?

➔Netflix uses what is known as the big data recommendation algorithm to suggest TV shows and movies based on a user's preferences

Netflix has screenshots of scenes people might have viewed repeatedly, the associated ratings, and the number of searches and the search topics

6)What are the main challenges that companies experience with big data?

➔Unprecedented data growth

Integrating data from a variety of sources

Data security issues

Unfamiliarity with big data and confusing it with traditional methods

7)What are some examples of big data sources?

➔Open data

Sensor data

Email

Social media

8)What are some of the main business domains that use big data tools today?

➔Transportation industry

Aviation industry

E-commerce industry

Credit scoring agencies

9)What are the main deliverables of big data?

➔Multivariate analysis

Predictive models

Text/image analytics

10)What are the most important advantages of big data, according to the International Institute for Analytics (IIA)?

➔Big data leads to cost reductions

Big data helps to identify what customers need and to introduce new products and services accordingly

Big data enables faster, better decision making

Big Data Concepts: Big Data Essentials

1)Which statement is true about in-memory storage systems?

➔Data storage in an in-memory database is reliant on random access memory (RAM)

2)What are the most important features of HDFS?

➔High availability

Replication

Distributed storage

Scalability

3)Which statement about parallel or distributed computing is true?

➔Distributed computing can allow an application on one machine to leverage processing power, memory, or storage on another machine

4)Match each Hadoop component with its respective layer in the Hadoop ecosystem. One layer will not be used.

➔HDFS➔Data storage layer

ZooKeeper➔Data management layer

Hive➔Data access layer

MapReduce➔Data processing layer

5)What are the benefits of migrating from Hadoop to the cloud?

➔Long-term cost savings

Easy access and resource availability

Better collaboration

Better scalability

6)Which statements are true about unstructured data?

➔Unstructured data is very often linked to structured data. An example is how X-ray images at a hospital are linked to patient IDs or health card numbers

Web pages, video files, and audio files are examples of unstructured data

7)Which statement about horizontal and vertical scaling is true?

➔Horizontal scaling is typically the easiest scaling option

8)Which statements are correct about HDFS?

➔HDFS provides a fault-tolerant storage layer for Hadoop and its other components

HDFS provides high throughput access to application data by providing the data access in parallel

9)What are the differences between Hadoop and cloud computing?

➔Cloud computing focuses on on-demand, scalable, and adaptable service models, while Hadoop is all about extracting value out of volume, variety, and velocity

Cloud computing constitutes various computing concepts. This naturally involves a large number of computers that are usually connected through a real-time communication network. Hadoop, on the other hand, is a framework that uses simple programming models to process large data sets across clusters of computers

10)Which statements accurately describe the differences between big data and data warehousing?

➔While only DBMS compatible data are stored in data warehouses, all kinds of data including transactional data, social media data (including audio and video), machinery data, or any DBMS data can be stored and managed using big data technologies

Data warehouses only handle structured data (relational or non-relational), whereas big data can handle structured, un-structured, or semi-structured data

Techniques for Big Data Analytics

1)Which statement is true about the Kappa architecture?

➔The Kappa architecture uses stream processing to manage data flows through a single path

2)Which are the main reasons for using batch processing?

➔To run complex algorithms on large datasets which require access to the entire batch

To join tables in relational databases

3)Which statement is true about the Lambda architecture?

➔The Lambda architecture provides fault-tolerance against possible hardware failures and human errors

Data that enters the system is dispatched to two layers in the Lambda architecture: the batch layer and the speed layer

4)Place the layers of big data analytics architecture in the correct order from the bottom to the top.

➔Data monitoring

Data security

Data storage

Data processing

Data query

Data visualization

5)What are some ways in which big data processing can be performed?

➔Stream processing

Batch processing

6)What are the parameters of data ingestion?

➔Data velocity

Data size

Data frequency

Data format

7)Which is correct about stream processing?

➔Stream processing provides analytical insights before the data storage stage

8)Which statement about data storage systems is correct?

➔The Hadoop distributed file system (HDFS) is the primary data storage system used by Hadoop applications

9)Which are the main components of the big data architecture?

➔Big data security

Big data analytics

The data model

10)What are the biggest challenges associated with traditional data analytics?

➔Scalability, consistency, reliability, efficiency, and maintainability

Spark for High-speed Big Data Analytics

1)What are some advantages that Spark provides to modern healthcare providers?

➔ Behind the scenes distributed execution

Convenient workflow fulfillment

A user-friendly API

2)What are some components of Apache Spark?

➔ Spark SQL

GraphX

3)Which statements are true about resilient distributed datasets (RDDs) and directed acyclic graphs (DAGs)?

➔ RDD is an immutable (read-only), fundamental collection of elements or items that can be operated on many devices at the same time (parallel processing)

Compared to MapReduce that creates a graph in two stages, Apache Spark can create DAGs that contain many stages

4)As Spark usage grew at Uber, users encountered an increasing number of issues. What were some of those issues/challenges?

➔ Multiple compute clusters

Multiple Spark versions

5)What are some examples of metrics that Alibaba measures by utilizing Spark?

➔ Degree distribution

Connected components

6)What are some predominant industries that use Spark today?

➔ Media and entertainment industry

Finance industry

7)What are some characteristics of Spark that help improve performance?

➔ Cache appropriately

Lazy loading behavior

8)What are the three API types that are compatible with Spark?

➔ RDD, DataFrame, DataSet

9)What are some of the most important best practices when it comes to using Apache Spark?

➔ Joining a large and a medium size RDD

Proper tuning

Using the right level of parallelism

10)Which statement is correct about how Spark and Hadoop are different?

➔ The Hadoop MapReduce model provides a batch engine, hence it is dependent on different engines for other requirements, whereas Spark performs batch, interactive, machine learning and streaming all in the same cluster

1)Which of the following is a characteristic of a data silo?

➔Data may be in a raw, native format and not useful unless processed

Data is stored in isolation and cannot be combined with other sources

Data is not easily accessible using common tools

2)Which of the following are valid data types that can be stored in a data lake?

➔Semi-structured data

Unstructured data

Structured data

3)Which of the following is not a characteristic of a data lake?

➔Data is not searchable easily

4)Which of the following are challenges involved in designing and building data lakes?

➔Data lakes need to be able to support a huge volume of data

Data lakes need to work with different data types and sparse and incomplete data

Data lakes need to maintain data security and compliance

5)Which of the following are valid differences between a traditional relational database and a data warehouse?

➔A data warehouse is optimized for read access, a database is optimized for read as well as write access

A database supports ACID properties and a data warehouse does not

6)Which of the following statements about data lakes and data warehouses are true?

➔Data warehouses hold fairly structured data optimized for analysis

Data lakes need to maintain security and ensure compliance of the data stored within it

Data lakes promote shared data stewardship

7)Which of the following is not an example of a data stream?

➔Census data stored in a database

8)Which of the following is not a valid service used to ingest data into the AWS cloud?

➔Amazon Athena

9)Which of the following correctly defines AWS Glue?

➔A single catalog which indexes data from multiple sources to make it searchable

10)Which of the following AWS services can be used to visualize data stored in a data lake on AWS?

➔Amazon QuickSight

1)Select the benefits of a distributed system.

➔fault tolerance

Concurrency

Scalability

2)Arrange the following ETL processing steps in order from the top.

➔ingest data from source

message brokering

streaming data engine

long-term storage and analytics

3)Select the characteristics of a NoSQL data store.

➔horizontal scaling

cluster-friendly

dynamic schema

4)Match the data management category with its description.

➔organizational data➔Master Data Management

dashboards and real-time results➔Visualization and Analytics

standardized data, static information➔Reference Data Management

data warehousing, transformation, extraction➔ETL

5) Match the ETL process with its description

➔format and representation shift➔Transform

selecting raw data➔Extract

importing data for computation➔Load

6)Where does the library of job components reside in the Talend Open Studio UI?

➔Palette

7)What high level model is used to get a project overview for ETL jobs in Talend Open Studio?

➔Business Model

8)Put the following AI hierarchy steps in pyramid order from the bottom up.
➔ETL
Data Exploration
Aggregation
Machine learning
Deep learning

9)Reducing the number of fields in the output is an example of what type of partitioning?
➔column-based

10)Match the data storage model approach with its descriptions.
➔Normalization➔Standardized, Less Redundancy
Star Schema➔Fact Tables, Dimension Tables

11)Select the features common to interactive reporting tools.
➔drilling down
Filtering
sorting

You have an on-premises data warehouse already installed in your organization. In the period following the COVID pandemic, your business started to grow exponentially, and you were tired of adding more nodes to the physical storage of the data warehouse. You decide to modernize your data warehouse and make it cloud-based. What major advantage will you achieve by modernizing your data warehouse?

Speed up time to analytics

You plan to implement a modern data warehouse solution into your enterprise. You have understood the proper data management and governance issues. You have set up all your domains and data ingestion methods. Now you plan to make a central repository of all your files. What should be the next step for the implementation of the data warehouse solution?

Selection of the nature of the data

You have a very well-run data management program in your organization, which is very secure. You use analytics for making big data driven decisions. In recent months, you realize that whatever decisions you are making based on the analytics are proving to be wrong and harmful for the organization. After consulting with the data analysts and data stewards you conclude that it is due to poor data quality. What should be the next step of action?

Install a firewall

Other than gaining real-time insights of the data, what is another major advantage of streaming analytics?

Real-time dashboards

BigQuery is a cloud-native warehouse that is also a fully managed data warehouse. What is the major advantage of BigQuery over Amazon Redshift that may be a deciding factor for the selection of a data warehouse?

Access control allows improved data sharing

Your organization has an effective sales team that is backed up by analytics that help accelerate the process of sales from the initial contact. One of the primary reasons for its effectiveness is a data input tool that hastens the process of data entry by providing preset suggestions. What are these suggestions commonly known as in data science terminology?

Reference data

You have incorporated the usage of Amazon Redshift in your organization, and you don’t want your data to be corrupted by processing. Therefore, you want the data to be stored in raw format before the processing is done, which is a service offered by Amazon Redshift. What is the key benefit of storing data in raw format?

Minimal loss of data

What is **not** a component of an on-premises data warehouse?

Processing space

You have an automobile company that helps sort out vehicles and make monthly sales versus expenditure reports. What is the best way to handle the data for centrally storing it?

Batch processing

In a candlestick chart, you see the share price of your company falling. You have implemented a streaming analytical tool that helps in analyzing and dashboarding the data as it is produced. You realize that the candlesticks pattern has consolidated and is not responding well to the influx of data. What is the source of this problem?

Server downtime

Select the correct differences between XML attributes and elements.

Elements can have multiple values

Elements can contain tree structure

Select the component used to generate an XML file from a CSV file in Talend studio.

tFileOutputXML

Select the exit code value the signifies the successful completion of a job in Talend studio.

0

Select the tMap Component description in Talend studio.

congregate input data to output data

Select the option to be enabled to allow the specification of 2 schemas for an XML input file in Talend studio.

Enable XPath in column "Schema XPath loop" but lose the order

In order to generate a complex XML file where data is specified using attributes of elements and elements trees in Talend studio, which component allows such.

tAdvancedFileOutputXML

Select the component used to perform lookup data in Talend studio.

tJoin

Select the component access a MySQL database in Talend studio.

tMysqlInput

Select the tool that allows specifying the relation of multiple tables as data sources when reading data from a database as input in Talend studio.

SQL Builder

Match the attribute value with its attribute that will only Add new records or modify existing ones without modifying the table structure or other records already exist in the table when writing data to a database table in Talend studio. Two options are invalid.

Answer Options:

- A:Insert or update
- B:Default
- C:Delete
- D:Select

Action on table : Default.
Action on data : Insert or update.

Select the component that allows updating data in a database in Talend studio.

tMySQLRow

Question
:

Select the components and concepts to facilitates accepting as input multiple databases in Talend studio.

tMap

Lookup

Select the component used to combine multiple database records to a single records in Talend studio.

tDenormalize

In your enterprise, customer information needs to be readily available to indicate whether the information given by customers is valid or not, and to see the potency of a positive deal. Which factors would you consider a priority when selecting the appropriate data pipeline tool?

Batch-wise vs. real-time ingestion and analytics

What are the major advantages of a cloud warehouse solution over an on-premises data warehouse solution?

Less worry about storage

Low cost

What are the two different data pipeline tools that address specific job roles?

Data engineers and analysts

Place the steps for a typical Azure Databricks warehouse in the correct order.

Ingest,Store,Prep and train,Model and Serve

The reason why Azure Databricks is so easy to use is because it is universal and is integrated with Microsoft's server for better parsing of information. Which platform has the same source of origin as Databricks, which gives it an analytical advantage over other platforms?

Apache Spark

What are the major disadvantages of Snowflake that might be troublesome for a few companies that seek data categorization?

No option for un-structured data

Fewer options with geospatial space

While setting up an integrated data pipeline for your enterprise to facilitate data ingestion in the warehouse, what should you place more emphasis on from a business perspective?

Analytics and business intelligence

Suppose you have a full-blown data management program with a well-running data warehouse and an optimum data pipeline tool that facilitates data transformation and transmission. While measuring the maturity of the data pipeline tool, what will be the sole factor that will determine the efficiency of the data pipeline tool?

Reliability and scalability

What is the one difference that separates the model of the Snowflake data warehouse from all the other data warehouse solutions?

Hybrid model

What is **not** a design component of a data pipeline?

Data integration

Select the main dependency that has to be installed for Talend to be installed.

Java

Select the supported OSs by Talend Open Studio

MacOS
Windows 64

Select the main parts of the default UI of Talend Open Studio.

Palette
Repository

When installing MySQL relational Database to be used with Talend, select the configuration to be performed once the installation of the components is complete.

MySQL server port
MySQL root password

Select the folder that contains all the project information for a job that is exported from Talend studio.

process

Rank the steps required to for any job in Talend studio.

1. Create job in the repository

   Correct answer.

2. Add components from palette to the design space

   Correct answer.

3. Configure the components properties

   Correct answer.

4. Run the job

   Correct answer

Select the correct description of Talend Metadata Bridge.

Synchronize metadata across data pipelines

Match each description with the variable in Talend studio.

Answer Options:

- A:studio provisions through components used in jobs integration
- B:Ad-hoc variables that can be configured in jobs
- C:execute jobs with parameters for different environments

Job contexts variables
execute jobs with parameters for different environments.

Studio global variables
studio provisions through components used in jobs integration.

User defined global variables
Ad-hoc variables that can be configured in jobs.

Which of these statements accurately describes the Unique Match Join match model in the Map Editor?

Last match is considered and passed to the output

Set as default when configuring an explicit Join

You are building an expression in Map Editor for a column in which you want to pad the row1.CustomerID string with leading zeroes up to maximum length of 6 characters. Which code can you use to accomplish this for you?

```
String.format("$06d",row1.CustomerID)
```

You have Map Editor open for a tMap object for which you are mapping databases objects. You click the Join Model for a table Join property. Which options appear in the Options dialog that appears?

Left Outer Join

Inner Join

You are building a filter expression in Map Editor for a column in which you want to filter the product name, prd.name to equal "Turbo Widgets", and you want the transaction quantity, tx.qty to be greater than 100. Which code can you use to accomplish this for you?

```
    prd.name.equals("Turbo Widgets")&&tx.qty>100
```

File Exist

Delete

Put

On the Component tab displayed for the tSortRow component, when you click to add criteria to the Criteria table, Talend automatically populates the column values with defaults. In the "sort num or alpha?" column Talend has chosen num by default for customer_id as displayed. Which other values are available for the "sort num or alpha?" column when you click to open the dropdown list for that column?

alpha

date

You are using a tExtractDelimitedFields component to split the Address2 field in the delimited file as displayed. What must you specify as the Field separator property for the tExtractDelimitedFields component to properly split the Address2 field?

```
ID;First_Name;Last_Name;Address1;Address2;Country

1;Claudia;Sand;10000 Main Dr NW;New York, NY;USA

2;Max;Bigot;60000 My St;Nashua, NH;USA

3;Rick;Tailleur;200000 Younge St;Toronto, ON;Canada

4;Noémie;Miller;500000 St. Catherines St Ouest;Montreal, QC;Canada

5;Catherine;Reilly;100000 Main St SW;Boston, MA;USA
```

"," (a comma)
Not selected
Correct answer

When configuring the properties for a tReplace component, you can optionally click the Advanced mode checkbox. Doing so allows you to specify what type of expression as the Pattern to search for?

Regular expression

When configuring the properties for a tAggregateRow component. You are going to Group by the customer_id field in order to aggregate the sales on a customer_id basis, so that in the resulting output file you will have one row for each customer_id with aggregated sales figures. Which Function value must you choose when configuring the Operations table for this tAggregateRow component?

sum

You are using a tNormalize component to normalize the category field in the delimited file as displayed. What must you specify as the Item separator property for the tNormalize component to properly normalize the category field?

```
product_id,product_name,category,unit_price

1,Regular Widget,Normally Aspirated;Low Price;Highly Reliable,125

2,Super Widget,Super;Medium Price;Very Reliable,250

3,Turbo Widget,Turbo;Medium Price;Very Reliable,275

4,S/T Widget,Super;Turbo;Premium Price;Reliable,425

5,Hybrid Widget,Hybrid;Normally Aspirated;Premium Price;Reliable,425
```

";" (a semi-colon)

What has Amazon been able to achieve by utilizing big data?

Gathering of information on what each customer is likely to purchase based on what other people with similar interests have purchased

Gathering of data on the search patterns of its customers

What could be accomplished by using big data technologies?

Cost reductions

New product development and optimized offerings

Four of the seven characteristics of big data are listed. Match each characteristic with its description. One description will not be used.

Answer Options:

- A:The amount of data that exists
- B:Making sure the data is accurate, which requires processes to keep bad data from accumulating in your systems
- C:The speed at which data is processed and becomes accessible
- D:The different types of data from XML to video to SMS

- E:Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports, which are loaded with confusing numbers and formulas

Veracity
Making sure the data is accurate, which requires processes to keep bad data from accumulating in your systems.

Variety
The different types of data from XML to video to SMS.

Velocity
The speed at which data is processed and becomes accessible.

Volume
The amount of data that exists.

Which statements are true about unstructured data?

Common examples of unstructured data include audio, video files, or No-SQL databases

Unstructured data is information that does not have a predefined data model

Which statements are correct about how Netflix utilizes big data?

Netflix has screenshots of scenes people might have viewed repeatedly, the associated ratings, and the number of searches and the search topics

Netflix uses what is known as the big data recommendation algorithm to suggest TV shows and movies based on a user's preferences

What are the main challenges that companies experience with big data?

Data security issues

Unfamiliarity with big data and confusing it with traditional methods

Integrating data from a variety of sources

Unprecedented data growth

What are some examples of big data sources?

Sensor data

Email

Open data

Social media

What are some of the main business domains that use big data tools today?

Aviation industry

Credit scoring agencies

E-commerce industry

Transportation industry

What are the main deliverables of big data?

Multivariate analysis

Predictive models

Text/image analytics

What are the most important advantages of big data, according to the International Institute for Analytics (IIA)?

Big data enables faster, better decision making

Big data helps to identify what customers need and to introduce new products and services accordingly

Big data leads to cost reductions

Which statement is true about in-memory storage systems?

Data storage in an in-memory database is reliant on random access memory (RAM)

What are the most important features of HDFS?

Distributed storage

Scalability

High availability

Replication

Which statement about parallel or distributed computing is true?

Distributed computing can allow an application on one machine to leverage processing power, memory, or storage on another machine

Match each Hadoop component with its respective layer in the Hadoop ecosystem. One layer will not be used.

Answer Options:

- A:Data processing layer
- B:Data storage layer
- C:Data access layer
- D:Data management layer
- E:Data execution layer

HDFS
Data storage layer.

ZooKeeper
Data management layer.

Hive
Data access layer.

MapReduce
Data processing layer.

What are the benefits of migrating from Hadoop to the cloud?

Easy access and resource availability

Better collaboration

Better scalability

Long-term cost savings

Which statements are true about unstructured data?

Web pages, video files, and audio files are examples of unstructured data

Unstructured data is very often linked to structured data. An example is how X-ray images at a hospital are linked to patient IDs or health card numbers

Which statement about horizontal and vertical scaling is true?

Horizontal scaling is typically the easiest scaling option

Which statements are correct about HDFS?

HDFS provides a fault-tolerant storage layer for Hadoop and its other components

HDFS provides high throughput access to application data by providing the data access in parallel

What are the differences between Hadoop and cloud computing?

Cloud computing focuses on on-demand, scalable, and adaptable service models, while Hadoop is all about extracting value out of volume, variety, and velocity

Cloud computing constitutes various computing concepts. This naturally involves a large number of computers that are usually connected through a real-time communication network. Hadoop, on the other hand, is a framework that uses simple programming models to process large data sets across clusters of computers

Which statements accurately describe the differences between big data and data warehousing?

While only DBMS compatible data are stored in data warehouses, all kinds of data including transactional data, social media data (including audio and video), machinery data, or any DBMS data can be stored and managed using big data technologies

Data warehouses only handle structured data (relational or non-relational), whereas big data can handle structured, un-structured, or semi-structured data

Which statement is true about the Kappa architecture?

The Kappa architecture uses stream processing to manage data flows through a single path

Which are the main reasons for using batch processing?

To run complex algorithms on large datasets which require access to the entire batch

To join tables in relational databases

Which statement is true about the Lambda architecture?

Data that enters the system is dispatched to two layers in the Lambda architecture: the batch layer and the speed layer

The Lambda architecture provides fault-tolerance against possible hardware failures and human errors

Place the layers of big data analytics architecture in the correct order from the bottom to the top.

1. Data monitoring

   1

2. Data visualization

   6

3. Data processing

   4

4. Data security

   2

5. Data query

   5

6. Data storage

   3

What are some ways in which big data processing can be performed?

Batch processing

Stream processing

What are the parameters of data ingestion?

Data format

Data size

Data velocity

Data frequency

Which is correct about stream processing?

Stream processing provides analytical insights before the data storage stage
Not selected
Correct answer.

Which statement about data storage systems is correct?

The Hadoop distributed file system (HDFS) is the primary data storage system used by Hadoop applications
Selected
Correct answer.

Which are the main components of the big data architecture?

Big data security
Selected
Correct answer.

Big data analytics

The data model

What are the biggest challenges associated with traditional data analytics?

Scalability, consistency, reliability, efficiency, and maintainability

What are some advantages that Spark provides to modern healthcare providers?

Convenient workflow fulfillment

Behind the scenes distributed execution

A user-friendly API

What are some components of Apache Spark?

GraphX

Spark SQL

Which statements are true about resilient distributed datasets (RDDs) and directed acyclic graphs (DAGs)?

RDD is an immutable (read-only), fundamental collection of elements or items that can be operated on many devices at the same time (parallel processing)

Compared to MapReduce that creates a graph in two stages, Apache Spark can create DAGs that contain many stages

As Spark usage grew at Uber, users encountered an increasing number of issues. What were some of those issues/challenges?

Multiple Spark versions

Multiple compute clusters

What are some examples of metrics that Alibaba measures by utilizing Spark?

Degree distribution

Connected components

What are some predominant industries that use Spark today?

Media and entertainment industry

Finance industry

What are some characteristics of Spark that help improve performance?

Cache appropriately

Lazy loading behavior

What are the three API types that are compatible with Spark?

RDD, DataFrame, DataSet

What are some of the most important best practices when it comes to using Apache Spark?

Joining a large and a medium size RDD

Using the right level of parallelism

Proper tuning

Which statement is correct about how Spark and Hadoop are different?

The Hadoop MapReduce model provides a batch engine, hence it is dependent on different engines for other requirements, whereas Spark performs batch, interactive, machine learning and streaming all in the same cluster

Which of the following statements about functions is true?

Function code is not executed when defined
Selected

A function is defined using the “def” keyword

Which of the following are valid function names in Python?

functionName()

_123_Function_Name()
_function_name()

Which of the following statements about functions is false?

Functions cannot access variables which are declared outside the function

Consider a function definition which looks like this:\

```
def some_function(a, b, c):

    print(a, b, c)
```

Which of the following function invocations are correct?

some_function(2, 3, “Hello”)

some_function(2, 3, 4)

Consider the following bit of code. What will be the result of executing this bit of code?

```
x = 3

y = 4


def add(a, b):

   result = x + y

   print(result)


add(10, 20)
```

Ans: 7

Which of the following statement(s) about positional arguments to functions is/are true?

A function can accept any number of positional arguments

They can be of any data type – primitive or complex types

What is the default return value from a function when no return statement is specified?

Ans : None

Which of the following statement(s) about return values is/are false?

A function has to have a return statement

A function can have just one return statement

A function with input arguments cannot have a return statement

Which of the following statements(s) about the data types of return values is/are false?

A return statement is mandatory in functions

Which of the following are valid kinds of input arguments for Python functions?

Keyword arguments

Positional arguments

Which of the following are some of the advantages of using keyword arguments to invoke functions?

Easier to maintain code since the value of each argument is clearly seen during invocation

Keyword arguments can be specified out of order

What does this function definition indicate?

```
def some_fn(a, b, c=True):

       print(a, b, c)
```

a and b are required arguments, c is optional

Which of the following function definitions allows the function to accept variable length arguments?

some_fn(*args)

Which of the following are valid types of arguments in Python?

Keyword arguments

Variable length arguments

Default arguments

Which of the following statement(s) about variable length arguments in Python is/are true?

Variable length keyword arguments are passed into the function as a dictionary

Variable length positional arguments are passed into the function as a tuple

Consider the following Python code. What command would you use to iterate through the “companies_ceo” DataFrame and print the list of all the CEOs in this DataFrame?

```
import pandas as pd

companies = {

    'Company' : [‘Facebook', 'Apple', ‘Amazon', ‘Netflix’],

                'CEO' : [‘Mark Zuckerberg’, 'Jeff Bezos’, 'Tim Cook'
, ',’Reed Hastings’ ],

                }

companies_ceo = pd.DataFrame(companies)
```

for row in companies_ceo.iterrows(): print(row[1])

for row in companies_ceo.itertuples(): print(row.CEO)

Which of the following formats does Pandas not support natively when exporting the contents of a Dataframe?

JPEG

Let’s say you have created a Pandas DataFrame called “unsorted” and you want to sort the contents of this DataFrame column wise in alphabetical order of the header name. Then, which function would you call on the “unsorted” DataFrame to do so?

unsorted.sort_index(axis=1)

Match the following functions that you can call on a Pandas DataFrame correctly with what they do

All the rows which contain a NaN value in any cell of that row are removed

: .dropna()

Returns a Boolean array containing true or false values and returns the value in a cell as true if it does not contain NaN : .notnull()

Every cell in the Dataset which has a NaN value will be replaced with 0 : .fillna(0)

Returns a Boolean array containing true or false values and returns the value in a cell as true if it contains NaN : .isnull()

Match the following statements related to the .xs function in Pandas DataFrame with their correct Boolean values.

False

The .xs function cannot be used to return a cross section of columns

True

By default, the .xs function only takes a look at values in the first level index

The .xs function is used when our Pandas DataFrame makes use of a MultiIndex

Let’s say you have imported Python as pd and have instantiated two DataFrames called “frame_1” and “frame_2” with the exact same schema. What command will you use to combine these two DataFrames into a single DataFrame and make sure that the combined DataFrame has its own unique index?

pd.concat([frame_2, frame_1], ignore_index = True)

pd.concat([frame_1, frame_2], ignore_index = True)

The ‘how’ argument in the Pandas merge function allows us to specify what kind of join operation we want to perform on the given Pandas DataFrames. What are the valid values that we can give for this argument?

outer
inner
left
right

Some statements related to working with SQL Databases in Python are given below. Match them with their correct Boolean values.

True

The sqlite3 library in Python allows us to create Databases on our local file system

All the changes that we make to an SQL database on a Jupyter notebook by connecting with it, will be committed to the database only after we execute sqlite3's .commit() function

False

Once we have created a table, we can use sqlite3's .execute() function to recreate the same table with the same table name so that we have duplicates of a table

In the following Python code, typing which Python command will give the user the CEO of Facebook?

```
import pandas as pd

companies_ceo = {

                              'Amazon'  :  'Jeff Bezos'

                              'Apple' : 'Tim Cook',

                              'SpaceX': 'Elon Musk'

                              'Facebook': 'Mark Zuckerberg'

                              'Netflix': 'Reed Hastings'

                             }

companies_ceo_series= pd.Series(companies_ceo)
```

companies_ceo_series[3]
companies_ceo_series['Facebook']

In the following Python code, typing what command will create a DataFrame called "companies_ceo" whose first column has all the entries of the 'companies' list and whose second column has all the entries of the 'ceo' list, with the column names as the names of the respective variables?

```
import pandas as pd

companies = {

'Amazon'

'Apple'
‘SpaceX’

‘Facebook’

‘Netflix’

                }

ceo = {

'Jeff Bezos'

'Tim Cook‘,

‘Elon Musk‘

‘Mark Zuckerberg’

‘Reed Hastings’

  }
```

companies_ceo_tuple = list (zip(companies, ceo)) companies_ceo = pd.dataframe(companies_ceo_tuple, columns=[‘companies’, ‘ceo’])

What happens when we call the stack () function on a Pandas DataFrame

It will create a new DataFrame such that a single row in the original DataFrame is stacked into multiple rows in the new DataFrame depending on the number of columns for each row in the original DataFrame

Which of these correctly match the following libraries in the Numpy ecosystem with what that library is used for?

Specifically meant for machine learning, data mining, and data analysis : Scikit-learn

Used to perform statistical operations : Statsmodel

Data visualization tool used for large datasets : Bokeh

Match the following statements related to the iloc indexer in Pandas with the correct boolean values.

False

The column headers can be passed as input arguments in the form of a string to the iloc function without any errors

When we pass 2:6 as input argument to the iloc function, we get all details of the records located in the second index all the way up to the 5th index of the DataFrame

True

The iloc indexer is similar to the loc indexer and can be used to access records located at a particular index in a Pandas DataFrame

Let's say you have saved a dataset in a pandas DataFrame called "dataset" which has tons of records and you only want to access the details of the records in only the 5th, 8th and 14th index. Which of these Python commands can you use to do so?

dataset.loc[5,8,14]

dataset.loc[[5,8,14],:]

Let's say you have a pandas DataFrame called "panda" which has 8 rows in total and you want to remove the last row from this DataFrame. Which of these Python commands would you use to do so?

panda.drop(panda.index[7])

Which of these statements related to the pivot function in Pandas is true?

The Pivot function summarizes the details of each column in a DataFrame

The combination of the row index and the column header must be unique in order to generate a pivot table

Match the following statements related to Pandas DataFrames with the correct boolean values.

False

Data in different columns of a Pandas DataFrame cannot be of different data types

Once a Pandas DataFrame has been created, it is not possible to add a new column to this DataFrame

True

All the data within a particular column in a Pandas DataFrame must be of the same data type

Match the following statements related to the concept of multiIndex in Pandas with the correct Boolean values.

False

The MultiIndex for a row is some composite key made up of exactly one column

True

MultiIndex lets the user effectively store and manipulate higher dimensional data in a 2-dimensional tabular structure

MultiIndex is useful when we have large datasets where using numeric indexes to refer to each record is unintuitive

Which of these statements related to the Pandas Series object are true?

Once we create a Pandas Series object, an index representing the positions for each of the data points is automatically created for the list

Pandas Series object is similar to a Python list

Let’s say you have a pandas DataFrame called “frame” and you want to export this DataFrame along with its index as a CSV file called “data_frame” located in the datasets folder of our workspace.

frame.to_csv(‘datasets/data_frame.csv’)

Let’s say you have a two dimensional numpy array called “twod” and you want to split it row-wise into two equal halves. Then, which of these numpy functions would you call on it to do so?

vsplit(twod,2)

Some of the features of digital images in Numpy are given below. Which of these are true?

A digital image is a multidimensional array and every pixel in a digital image is represented by a number

In numpy, images can be represented as a 3D matrix where the first two dimensions represent the pixels in the image that are arranged in the form of a grid and the third dimension specifies the number of channels for the image

Let’s say you have an image that you have split into two equal halves along the x axis. You have stored these two halves of the original image in the variables x1 and x2 respectively. Which numpy function would you use to combine these two halves to reconstruct the original image?

concatenate((x1,x2),axis=1)

Let’s say you have a numpy array called “array_1” and you initialize “another array called “array_2” with the help of a following command:

array_2 = array_1.view()

Match the following statements about “array_2” with the correct Boolean value

True

array_1 and array_2 contain the same elements

The base for array_2 points to the same object as array_1

False

array_2 points to the same object as array_1

If we re-assign array_2, then we will end up re-assigning array_1 as well and change its contents

Let’s say you have a numpy array called “array_3” and you initialize “array_4” with the help of a following command:

array_4 = array_3.copy()

Match the following statements about “array_4” with the correct Boolean value

True

array_3 and array_4 contain the same elements

False

If we re-assign array_4, then we will end up re-assigning array_3 as well and change its contents

If we change a single element of array_4, then the corresponding element in array_3 changes too

Changing the shape of array_4 will change the shape of array_3 as well

Let’s say you have a 1-D numpy array called “cubes” consisting of the cubes of the numbers 1,2,3 and so on till 10. What would be the value of the array:

cubes [[[4, 5], [1, 2]]]

[[125, 216] , [8, 27]]

Some of the features of Pandas is given below. Which of these are true?

A particular column of a pandas dataframe can be referenced by its column header

The column header of a Pandas dataframe can be treated in the same way as the index label of a numpy array

Let’s say you imported numpy as np and you have initialized a 1-D array of integers called “array”. What would np.all (x < 50) return?

This function would return a true boolean value if all the entries in your array are less than 50 and false otherwise

Let's say you have a Pandas dataframe called "phone_data" which contains the data of various phones released in 2018 and their prices. It has the following three columns:

"manufacturer", "phone name" and " price".

You want only the names of all the phones that are priced more than 10,000. Which of these commands can be used to print these values?

phone_data[phone_data['price'] > 10000]['phone name']

What are the conditions under which broadcasting can take place between two elements in Numpy?

A smaller array can be broadcast on a larger array only when the corresponding dimensions of the two arrays being operated upon are compatible i.e. when the corresponding dimensions are equal or one of the two dimensions is 1

Broadcasting works when at least one of the elements is a scalar

Match the following statements about broadcasting with the correct Boolean value:

False

The array [[1, 2] , [3, 4]] and the scalar 10 are incompatible with broadcasting

True

The scalar 10 and the scalar 20 are compatible with broadcasting

The array [[1, 2] , [3, 4]] and the array [1, 2 ,3] are incompatible with broadcasting

The array [[1, 2] , [3, 4]] and the array [[1], [2]] are compatible with broadcasting

Which of these correctly match the following libraries in the Numpy ecosystem with what that library is used for?

Data visualization tool used for plotting 2-D graphs: Matplotlib

Contains a collection of algorithms used for image processing : Scikit-image

Used to perform statistical operations : Statsmodel

Let’s say you have imported the numpy package as np and you want to assign the variable “x” with a 3 by 2 array of type integer, all of whose values are 1. Which of these commands will you use to do so?

x=np.ones((3,2),dtype=np.int32)

What will be the value stored in the variable y after we have executed the following code

```
import numpy np

y=np.arange(2,4,0.5)
```

[2, 2.5, 3, 3.5]

Let’s say you have imported the numpy package as np and you want to print the first 2000 natural numbers in the form of an array and you want all 2000 of the numbers to be visible on screen when printing (including the number 2000). Which of these commands would you use to do so?

np.set_printoptions(threshold=np.nan) print(np.arange(1,2001))

What would be the output of the following code:

```
import numpy as np

x=np.array([[1,2] , [3,4]])

y=np.array([ [5,6],[7,8]])

z=(x*y)

z
```

[[5,12], [21,32]]

What would be the output of the following section of code:

```
Import numpy as np

x=np.array([4,6,2,8])

np.median(x)
```

5

Which of these slicing operations can be used to quickly get the reversed contents of a numpy array called “array”?

array[::-1]

Match the following features of the numpy nditer function mentioned here with the correct Boolean category

False

The nditer can accept only one dimensional arrays as input

By default, the nditer object returns arrays that can be written on

True

Using this function, we can iterate through each of the individual elements of the array passed as an input argument

Which of these statements regarding the ravel() object in Python are true?

The ravel function reduces a multi-dimensional array to a single dimensional array

ravel() belongs to the numpy library and can be used on any object that can be parsed

Select the benefits of a distributed system.

scalability

fault tolerance

concurrency

Arrange the following ETL processing steps in order from the top.

ingest data from source

message brokering

streaming data engine

long-term storage and analytics

Select the characteristics of a NoSQL data store.

horizontal scaling

dynamic schema

cluster-friendly

Match the data management category with its description.

dashboards and real-time results : Visualization and Analytics

standardized data, static information : Reference Data Management

data warehousing, transformation, extraction : ETL

organizational data : Master Data Management

Match the ETL process with its description.

importing data for computation : Load

selecting raw data : Extract

format and representation shift : Transform

Where does the library of job components reside in the Talend Open Studio UI?

Palette

What high level model is used to get a project overview for ETL jobs in Talend Open Studio?

Business Model

Put the following AI hierarchy steps in pyramid order from the bottom up.

ETL

Data Exploration

Aggregation

Machine Learning

Deep Learning

Reducing the number of fields in the output is an example of what type of partitioning?

column-based

Match the data storage model approach with its descriptions

Star Schema : Dimension Tables , Fact Tables

Normalization : Standardized , Less Redundancy

Select the features common to interactive reporting tools.

Filtering

drilling down

sorting

Which of the following is a characteristic of a data silo?

Data may be in a raw, native format and not useful unless processed

Data is stored in isolation and cannot be combined with other sources

Data is not easily accessible using common tools

Which of the following are valid data types that can be stored in a data lake?

Semi-structured data

Structured data

Unstructured data

Which of the following is not a characteristic of a data lake?

Data is not searchable easily

Which of the following are challenges involved in designing and building data lakes?

Data lakes need to maintain data security and compliance

Data lakes need to be able to support a huge volume of data

Data lakes need to work with different data types and sparse and incomplete data

Which of the following are valid differences between a traditional relational database and a data warehouse?

A data warehouse is optimized for read access, a database is optimized for read as well as write access

A database supports ACID properties and a data warehouse does not

Which of the following statements about data lakes and data warehouses are true?

Data warehouses hold fairly structured data optimized for analysis

Data lakes need to maintain security and ensure compliance of the data stored within it

Data lakes promote shared data stewardship

Which of the following is not an example of a data stream?

Census data stored in a database

Which of the following is not a valid service used to ingest data into the AWS cloud?

Amazon Athena

Which of the following correctly defines AWS Glue?

A single catalog which indexes data from multiple sources to make it searchable

Which of the following AWS services can be used to visualize data stored in a data lake on AWS?

Amazon QuickSight

Consider three variables with values as shown:
a = 5
b = 10
c = “five”

What are the results of evaluating the conditional expression?
a == c
a >= b
not(a < b and a > b)

False, False, True

How is the body of an if-statement block syntactically represented in Python?

Using additional indentation from the left relative to lines just before and after the block

What is the output for this code?

```
if 'bin' in {'float': 1.2, 'bin': 0b010}:
    print('a')
    print('b')
print('c')
```

a b c

What is the output for this code?

```
if None:
          print('Hi')
```

Nothing is printed - no output

Evaluate the expression provided. What does the following expression evaluate to?

```
'1' + '2' if  '123'.isdigit() else '2'  + '3'
```

‘12’

What is the output of the code?

```
a = [1, 'one', {2: 'two'}, 3]
b = len(a)


if b == 4:
   print('Length of this list is 4')
   if b == 5:
       print('Length of this list is 5')
   else:
         print(b)
```

Length of this list is 4 4

What is the value of b in the snippet of python code?

```
a = “six”
b = (int(a), float(a))
```

ValueError: invalid literal for int() with base 10: 'six'

Consider the following snippet of Python code:

```
a  =  “40.6 ”

b  =  “60.4 ”

c  =  a + b
```

What does c evaluate to?

‘40.6 60.4’

What would the output of the following code snippet be?

```
num_one = 76

num_two = 23.4

print("datatype of num_one:", type(num_one))

print("datatype of num_two:", type(num_two))
```

datatype of num_one: <class 'int'> datatype of num_two: <class 'float'>

What is the output of the code snippet below?

```
value = 4

a = str(value)

b = a + “^” + “2”

c = a + “^” + “3”

print(value, “+”, b, “+”, c)
```

4 + 4 ^ 2 + 4 ^ 3

What do the values of d[0], d[1], d[2], d[3] evaluate to after the execution of the Python code below?

```
new_list = [“Red”, “Blue”, “White”, “Green”]

z = sorted(new_list)

d = list(z)

d[0], d[1], d[2], d[3] = d[3], d[2], d[1], d[0]
```

“White”, “Red”, “Green”, “Blue”

What is the output of the program below?

```
var = "hi"
if(type(var) == int):
    print("Type of the variable is Integer")
elif(type(var) == float):
    print("Type of the variable is Float")
elif(type(var) == complex):
    print("Type of the variable is Complex")
else:
    print("Type of the variable is Unknown")
```

Unknown

What is the output of the program below?

```
total_classes = 100
attended_classes = 67

attendance = (attended_classes/total_classes)*100
if attendance >= 75:
    print ("You are eligible to appear for the test.")
else:
    print ("Sorry, you are ineligible to appear for the test.")
```

Sorry, you are ineligible to appear for the test.

Identify essential table types that we can use to implement a Star schema.

Dimension table

Fact table

Which statements about cloud-based data warehouse are true?

Cloud-based Data warehouses are scalable

Cloud-based Data warehouses are elastic

Identify the essential levels of management that require strategic reports.

Middle management level

Strategic

Match the data modelling strategies with their features.

Denormalization:

- Used on previously normalized databases to increase performance
- Adds redundant data
- Combines tables

Normalization

- Ensures the dependencies are properly enforced
- B:Applies formal rules to enforce dependencies
- C:Splits tables

Choose the characteristics of weighted reports.

With weighted reports we get a meaningful subtotal and total

Weighted reports multiply all the facts by weight before aggregating

Which processes involved in a data warehouse project are important?

ETL

Data cleansing

Which are essential tasks that we can execute to facilitate business intelligence?

Extract

Load

Which of the following local and global warehouse statements are true?

Local data warehouse cannot be accessed globally

We can provision a single global repository for a particular domain

Select data modelling strategies that we can adopt to create an ER model.

Normalization

Denormalization

Which of the following OLAP statements are true?

OLAP is a part of the overall Data warehouse implementation

OLAP provides real-time analytical capability

Identify the terminologies that we generally use in data warehousing.

ETL

Dimensional model

Identify essential features of Strategic Information

Preserves data integrity

Time-variant

Identify some of the essential features which differentiates a data warehouse from OLTP.

Data warehouse stores historical data

Data warehouse provides predictive analytical capabilities

Identify the essential differences between RDBMS and data lakes?

RDBMS databases are transaction while data lakes are not transactional

RDBMS databases defines fixed schema while data lake follows no schema design

Which statements about Snowflake schemas are true?

Application binary interface allows us to contextualizes contracts

A Snowflake schema is an extension of the Star schema

Identify the essential components of Azure data lake.

SQL server

Data factory

Match the data warehousing solutions with their associated benefits.

On-premise data wareshouse

- Preferred in banking or government domains
- Offers absolute control over security

On-cloud data warehouse

- Provides scalability
- Cost effective
- Offers better speed and connectivity

Specify some of the outcomes of a data warehouse realization.

Predictive analytical reports

Intuitive dashboards

Specify some of the essential logical components of a data warehouse.

Entities

Attributes

Which of the following components are provided by Talend to design ETL jobs?

TMap

TFileInput

Which of the following statements correctly defines the characteristics of the Kimball model?

In the Kimball model the analytical systems can access the data directly

Kimball model uses the dimensional model

Identify essential tasks involved in an ETL process.

Transforming extracted data

Extracting data from diversified sources

Which of the following dimensional components differentiates a Dimensional model from an ER model?

Dimension tables

Fact tables

What are some of the essential tasks that we can execute using Talend?

ETL job designing

Business modelling

What are some of the integrated components of a data warehouse?

Data staging

Storage

What are some of the important tasks that we need to perform to implement a Physical model for a given Logical model?

Create Foreign keys to establish the relationships among objects

Create tables to represent the entities

Select prominent ETL tools that we can use with a data warehouse implementation.

Talend

PowerCenter Informatica

1.which collection does not allow duplicates values?
Ans. Set

2.Which of the following statements are true about Data Warehouse?
Ans. Data Warehouse is specific technology.
Data Warehouse Can be purchased.

3.Which is not a valid python data type?
Ans. char

4.Which of the following is not a valid escape sequence in Unix?
Ans. \d

5.Which is not valid Dictionary Methods
Ans. insert()

6.Choose an approptiate query for Renaming table from Books table to___?
Ans. RENAME old_table _name To new_table_name ;
ALTER TABLE old_table_name RENAME TO new_table_name;

7.To rename the Student table to StudentInfo
Ans. RENAME Student table To StudentInfo ;
ALTER TABLE Student table RENAME TO StudentInfo;

8.Which of the following relational Operator represents Less Than or Equal to in Shell Scripting?
Ans. Less than or equal to -le

9.What is a variable defined outside a function referred to as ?
Ans Global Variable.

10.Which of the following commands can be used to list hidden files in UNIX?
Ans. ls -a

11.Unique key can be null,Primary key cannot be null
The above statement is True or False
Ans. True

12._____join result contains records that are the multiplication of records from two tables?
Ans. Cross Join

13.TestDir directory has 3 files stored inside it.
Which of the following options can be used to delete TestDir?
Ans. rm -r TestDir

14.Which command can be used to check the current working directory?
Ans. pwd

15. in python the data in pandas can be analyzed with which of the following options?
Ans. Series,DataFrames

16. Which of the following will be used get the number of positional parameters in Shell Scripting?
Ans. Positional parameters are a series of special variables ($0 through $9)

17.Which command can be used to change the permissions of a file?
Ans. chmod

18.Which sql query displays the name of the employees which has 'A' as the second letter and ends with 'N'?
Ans. LIKE '_A%N';

19.Which is correct syntax for defining string?
Ans. All of above

20.Refer to below two statements and select best option

```
echo $num1+$num2;
echo $((num1+num2))
```

Ans.Both are valid
1 is invalid and 2 is valid

21.all subqueries can be converted to joins
Ans. False

22.What happens when you run the following shell script?

```
Text=Hello World
echo $Text
```

Ans. Error in Line 1.

23. In below execution of script, what will be stored inside $1 in shell script?

```
./script.sh Accenture BDC PDC
```

Ans. Accenture

24.Which of the following is not a valid shell variable?
Ans. env(doubt)

25.Which of the following statements are true?
1.Shell scripts are executed in a seperate child process.
2.shell scripts need to be mandatory given an extension as .sh
Ans. 1 is true,2 is false

26.how will you display long list of files sorted by last modification time?
Ans. ls -lt

27.What will happen if the below command is executed?

```
cp file1.txt TestDir/
```

Ans.contents of file1.txt will be copied into TestDir/file1.txt

28.Consider the value product Price=139.592
Ans. select ROUNT(139.592,0) from DUAL;

29.What will be the output of the below script?

```
X=100
echo X$X$;
```

Ans. X100$

30.Which is the correct SQL statement to increase the size of the column "StudentName" to 40 in the table Student.
Ans. ALTER TABLE Student MODIFY StudentName Varchar2(40);

31.Which of the follwing statements can be used in a shell script to print the current shell?
Ans. echo $SHELL

32. Employee tables contain EmployeeNumber...

Ans. ALTER TABLE Employee DROP COLUMN DeptCode;

33.The LIKE operator allows you to ue wildcard characters when searching
Ans True

34.In ______ rows are called as tuples and columns are known as Attributes
Ans.Relational Algebra

35.which of the following command can be used to display headers in output of who command?
Ans. who -H

36.Which is the correct SQL statement to add a Column SupplierLoc to the existing 'Supplier table?
Ans. ALTER TABLE SUPPLIER
ADD COLUMN SupplierLoc number(10);

37.Identify the output of the below code

```
str='paul how are you'
print(str[-1])
```

Ans. u

38. File1.txt should not be redable by others but redable and writable by owners and group
Ans. chmod 660 File1.txt

39.Which of the following options can be used as horizontal filter in UNIX?
Ans

40.which of the following parameter will be used by the shell to read the first Argument from the command line?
Ans. 0

41.SQL>Truncate table Student_info;
Ans. It is faster than the delete statement
It cannot be rolled back
It removes entire rows from the table

42.The 'Vendor" table contains VendorId,VendorName
Ans. TRUNCATE TABLE Vendor;

43.Which of the following will be used get the number of positional parameters in Shell Scripting?
Ans. The variable $# contains the number of positional parameters

44.In python everything is represented as object.
Ans. True

45.Which is correct syntax for lambda?
Ans. lambda n1,n2:n1+n2

46.pandas is used for?
Ans. Data Preprocessing,Data Cleaning,Data Analysis.

47.Python is____________
Ans. Interpreted Language

48.Which is the correct syntax to add a column to dataframe

Ans. All

49.Inside class if we define self variable what is scope of that variable?
Ans. All

50.Which of the following are the suitable features of a Data Warehouse?
Ans.Used to analyze business
Summarized and refined

51.Which is the correct command to installed numpy
Ans.pip install numpy

52.Which are the correct Pandas function
Ans.All

53.Python supports which style of programming?
Ans.All

54.In Pandas which of the following is 2D
labelled,size-mutable tabular structure with potentially
Ans. DataFrame

55.Using numpy array print the odd numbers from 10 to 50
Ans.

56.what will be the output if below statement executed in shell script?

```
echo $?
```

Ans. exit status of last executed command.

57.Which command can be used to sort lines from demo.txt file in reverse order?
Ans. sort -r

58.Which is not valid numpy function?
Ans.numpy.math()

59.Choose the type of join which return rows when there is atleast one match of rows between
the tables
ANs. Inner Join

60.Choose 3 appropriate options which are true about Primary key in SQL?
Ans. Primary Key values cannot be null
It is a combination of fields which uniquely specify a row
It is a special kind of Unique key and has implicit null constraint

61.Write a query to find the names of employees from EmpTable which begins with 'A'?
Table name:Emp
column name: Emp name
Ans.select *from Emp where Empname like 'A%';