****

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 5^{th}, 8^{th} and 14^{th} 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 t o 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 aggregat e 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 fi gures. 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 y ou want all 2000 of the numbers to be visible on screen when printi ng (including the number 2000). Which of these commands would y ou 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 rev ersed contents of a numpy array called “array”?

array[::-1]

Match the following features of the numpy nditer function mentione d 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