Data Essentials

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

Ans. date,alpha

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

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

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

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

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

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

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

Ans. "," (a comma)

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

Ans. Regular expression

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

Ans. Sum

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

```
product_id,product_name,category,unit_price

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

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

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

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

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

Ans. ";" (a semi-colon)

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

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

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

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

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

Ans. Left outer join, inner join

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

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

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

Ans. Put

Selected
□

File Exist
□

Delete

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

Ans. New product development and optimized offerings, Cost reductions

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

Volume-The amount of data that exists

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

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

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

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

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

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

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

Unprecedented data growth

Data security issues

Integrating data from a variety of sources

15. What are some examples of big data sources?

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

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

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

17. What are the main deliverables of big data?

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

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

Ans. Big data enables faster, better decision making

Big data leads to cost reductions

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

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

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

20. What are the most important features of HDFS?

Ans. Replication, Scalability, Distributed storage, High availability

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

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

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

Ans. ZooKeeper-Data management layer

HDFS-Data storage layer

Hive-Data access layer

MapReduce-Data processing layer

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

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

24. Which statements are true about unstructured data?

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

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

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

Ans. Horizontal scaling is typically the easiest scaling option

26. Which statements are correct about HDFS?

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

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

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

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

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

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

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

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

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

Ans. vsplit(twod,2)

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

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

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

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

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

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

array_2 = array_1.view()

Match the following statements about “array_2” with the correct Boolean value

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

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

array_4 = array_3.copy()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
import pandas as pd



companies_ceo = {
```

```
                              'Amazon' :  'Jeff Bezos'

                              'Apple' : 'Tim Cook‘,

                              ‘SpaceX’: ‘Elon Musk‘

                              ‘Facebook’: ‘Mark Zuckerberg’

                              ‘Netflix’: ‘Reed Hastings’

                             }

companies_ceo_series= pd.Series(companies_ceo)
```

ans. companies_ceo_series[‘Facebook’]

companies_ceo_series[3]

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

```
import pandas as pd

companies = {

'Amazon'

'Apple'

‘SpaceX’

‘Facebook’

‘Netflix’

                }



ceo = {

'Jeff Bezos'

'Tim Cook‘,
‘Elon Musk‘

‘Mark Zuckerberg’

‘Reed Hastings’

  }
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ans. Pandas Series object is similar to a Python list

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

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

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

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

```
import pandas as pd




companies = {

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


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


                }




companies_ceo = pd.DataFrame(companies)
```

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

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

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

Ans. JPEG

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

Ans. unsorted.sort_index(axis=1)

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

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

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

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

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

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

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

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

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

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

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

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

Ans. Left,right,inner,outer

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

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

60. Which statement is true about the Kappa architecture?

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

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

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

To join tables in relational databases

62. Which statement is true about the Lambda architecture?

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

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

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

Ans. Data monitoring

Data security

Data storage

Data processing

Data query

Data visualization

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

Ans. Batch and stream processing

65. What are the parameters of data ingestion?

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

66. Which is correct about stream processing?

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

67. Which statement about data storage systems is correct?

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

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

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

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

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

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

Ans. Behind the scenes distributed execution

Convenient workflow fulfillment

A user-friendly API

71. What are some components of Apache Spark?

Ans. Spark SQL

GraphX

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

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

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

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

Ans. Multiple Spark versions

Multiple compute clusters

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

Ans. Connected components

Degree distribution

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

Ans. Finance industry

Media and entertainment industry

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

Ans. RDD, DataFrame, DataSet

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

Ans. Joining a large and a medium size RDD

Proper tuning

Using the right level of parallelism

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

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

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

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

Data is not easily accessible using common tools

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

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

Ans. Unstructured data

Semi-structured data

Structured data

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

Ans. Data is not searchable easily

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

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

Data lakes need to maintain data security and compliance

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

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

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

A database supports ACID properties and a data warehouse does not

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

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

Data warehouses hold fairly structured data optimized for analysis

Data lakes promote shared data stewardship

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

Ans. Census data stored in a database

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

Ans. Amazon Athena

88. Which of the following correctly defines AWS Glue?

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

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

Ans. Amazon QuickSight

90. Select the benefits of a distributed system

Ans. fault tolerance

Concurrency

Scalability

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

Ans. ingest data from source

message brokering

streaming data engine

long-term storage and analytics

92. Select the characteristics of a NoSQL data store.

Ans. dynamic schema

cluster-friendly

horizontal scaling

93. Match the data management category with its description.

Ans. standardized data, static information-Reference Data Management

organizational data-Master Data Management

dashboards and real-time results-Visualization and Analytics

data warehousing, transformation, extraction-ETL

94. Match the ETL process with its description.

Ans. importing data for computation-Load

selecting raw data-Extract

format and representation shift-Transform

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

Ans. Pallete

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

Ans. Business Model

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

Ans. ETL

Data Exploration

Aggregation

Machine Learning

Deep Learning

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

Ans. column-based

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

Ans. Star Schema-Fact Tables, Dimension Tables

Normalization-Less Redundancy, Standardized

100. Select the features common to interactive reporting tools.

Ans. Filtering, drilling down, sorting

101. Match the data backup methods with their descriptions.

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

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

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

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

102. Match the concepts with their descriptions.

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

The ability to ask questions and learn new things-Information

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

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

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

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

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

Ans. Heat Map

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

Ans. Creates groups to summarize data-GROUP BY

Lists the columns you want to retrieve-SELECT

Applies filtering logic to your groups-HAVING

Applies filtering logic to limit records in your results-WHERE

Describes how you want your data sorted-ORDER BY

Name of the table to pull the data from-FROM

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

Ans. Qualitative data

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

Ans. They all require identifying source and target systems.

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

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

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

Ans. Observational

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

Ans. Planning

Analyzing the data

Design

Implementation

Final migration

Testing

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

Ans. Logical

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

Ans. Provide a unified single version of the data

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

Ans. Hybrid style

112. Which is not a function of entity resolution?

Ans. Data Propagation

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

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

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

114. Which is not a goal of metadata management?

Ans. Data reference

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

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

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

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

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

Ans. Data Quality Program

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

Ans. Unstructured, Structured, Managed, Optimized

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

Ans. Access control is limited to a particular system

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

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

121. What is the integral meaning of data harmonization?

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

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

Ans. Governance practices

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

Ans. Dividing the business

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

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

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

Ans. Storage Networking Industry Association & General Data Protection Regulation

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

Ans. Enterprise resource planning

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

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

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

Ans. Data quality

129. Which is a performance measurement level?

Ans. Strategic measurement

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

Ans. Data integration

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

Ans. Inter-relations

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

Ans. Fix data quality issues

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

Ans. Proactive environment, organizational growth, and customer relationships

134. What is meant by reference data?

Ans. Data that is used for classification of other data

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

Ans. Ensure source to target consistency

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

Ans. Knowing the data

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

Ans. Overlook achievable governance maturity milestones

138. Which is NOT an advantage of data lakehouse?

Ans. Provides an on-premises data warehouse solution

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

Ans. Fact table, dimension table

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

Ans. Cloud-based Data warehouses are scalable

Cloud-based Data warehouses are elastic

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

Ans. Middle management level

Strategic

142. Match the data modelling strategies with their features.

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

143. Choose the characteristics of weighted reports.

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

Weighted reports multiply all the facts by weight before aggregating

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

Ans. Data cleansing,ETL

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

Ans. Load

Extract

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

Ans. Local data warehouse cannot be accessed globally

We can provision a single global repository for a particular domain

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

Ans. Normalization,Denormalization

148. Which of the following OLAP statements are true?

Ans. OLAP provides real-time analytical capability

OLAP is a part of the overall Data warehouse implementation

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

Ans. ETL, Dimension model

150. Identify essential features of Strategic Information.

Ans. Preserves data integrity, Time-variant

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

Ans. Data warehouse provides predictive analytical capabilities

Data warehouse stores historical data

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

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

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

153. Which statements about Snowflake schemas are true?

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

Application binary interface allows us to contextualizes contracts

154. Identify the essential components of Azure data lake.

Ans. Data factory, SQL server

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

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

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

Ans. Intuitive dashboards

Predictive analytical reports

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

Ans. Attributes,Entities

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

Ans. tMap, tFileInput

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

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

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

160. Identify essential tasks involved in an ETL process.

Ans. Extracting data from diversified sources

Transforming extracted data

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

Ans. Dimension and fact table.

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

Ans. ETL job designing, Business modelling

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

Ans. Storage, Data Staging

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

Ans. Create Foreign keys to establish the relationships among objects

Create tables to represent the entities

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

Ans. PowerCenter Informatica,Talend

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

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

Ans. Speed up time to analytics

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

Ans. Selection of the nature of the data

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

Ans. Install a firewall

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

Ans. Real-time dashboards

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

Ans. Access control allows improved data sharing

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

Ans. Reference data

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

Ans. Minimal loss of data

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

Ans. Processing space

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

Ans. Batch processing

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

Ans. Server downtime

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

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

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

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

Ans. Less worry about storage

Low cost

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

Ans. Data engineers and analysts

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

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

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

Ans. Apache spark

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

Ans. Fewer options with geospatial space

No option for un-structured data

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

Ans. Analytics and business intelligence

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

Ans. Reliability and scalability

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

Ans. Hybrid model

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

Ans. Data integration

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

Ans. Java

187. Select the supported OSs by Talend Open Studio.

Ans. MacOS

Windows 64

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

Ans. Repository, Palette

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

Ans. MySQL root password

MySQL server port

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

Ans. Process

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

Ans. Create job in the repository

Configure the components properties

Add components from palette to the design space

Run the job

192. Select the correct description of Talend Metadata Bridge

Ans. Synchronize metadata across data pipelines

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

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

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

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

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

Ans. tFileOutputXML

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

Ans. 0

197. Select the tMap Component description in Talend studio.

Ans. congregate input data to output data

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

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

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

Ans. tAdvancedFileOutputXML

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

Ans. tJoin

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

Ans. tMysqlInput

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

Ans. SQL Builder

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

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

Ans. Action on data-Insert or update

Action on table-Default

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

Ans. tMySQLRow

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

Ans. tMap, LookUp

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

Ans. tDenormalize

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

Ans. Set as default when configuring an explicit Join

Last match is considered and passed to the output

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

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

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

Ans. Inner Join, Left Outer Join

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

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

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

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

Ans. Delete

File Exist

Put

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

Ans. Alpha,date

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

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

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

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

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

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

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

Ans. "," (a comma)

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

Ans. Regular Expression

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

Ans. Sum

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

```
product_id,product_name,category,unit_price

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

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

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

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

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

Ans. ";" (a semi-colon)