DATA SCIENCE QUESTION BANK 2019-20

UNIT 1: Chapter 1	
1.	Explain the Rapid Information Factory ecosystem.
2.	Explain Schema–on–Write and Schema–on–Read.
3.	Explain data lake and data vault.
4.	What is data vault? Explain hubs, links and satellite with respect to data vault.
5.	Explain Spark and its components as data science processing tools.
6.	Explain Kafka and its components as data science processing tools.
7.	Explain Mesos, Akka and Cassandra as data science processing tools.
8.	List and explain different programming languages using in data science processing.
9.	What is MQTT? Explain the use of MQTT in data science.
UNIT 1: Chapter 3	
10.	Define Data Science Framework. Explain the Homogeneous Ontology for Recursive Uniform Schema.
11.	Discuss the Cross-Industry Standard Process for Data Mining (CRISP-DM).
12.	State and explain the top layers of data science framework.
13.	Explain the basics of Business layer.
14.	Explain the basics of Utility layer.
15.	Explain the basics of operational management layer.
16.	Explain the basics of audit, control and balance layer.
UNIT 1: Chapter 4	
17.	What is MoSCoW? Explain.
18.	What are the general and specialized functional requirements? Explain.
19.	Explain the SUN model developed by Professor Mark Whitehorn. Explain SUN models 1,2 and 3.
20.	What is dimension? What are slowly changing dimensions? Explain different types of slowly changing dimensions.
21.	Explain the non-functional requirements of business layer.
22.	What is configuration management? Explain.
23.	Explain deployment, documentation, disaster recovery, efficiency effectiveness, extensibility, failure management, latency, interoperability, maintainability, modifiability, network topology, privacy, quality, recovery/recoverability, reliability, resilience, resource constraints, reusability, scalability, security, testability, controllability, isolate ability, understandability and automatability.
24.	What are the common pitfalls with requirements? Explain with examples.
UNIT 1: Chapter 5	
25.	What are utilities? Explain the utility design process.
26.	Discuss the rules of European Union General Data Protection Regulation (GDPR)
27.	Explain the adders and process utilities.
28.	Discuss the different data vault utilities.
29.	What are the different transform utilities? Explain.
30.	Explain the different data science utilities.

31.	Explain the following utilities: organize utilities, report utilities, maintenance utilities, backup and utilities, Checks Data Integrity Utilities, History Cleanup Utilities, Maintenance Cleanup Utilities, Notify Operator Utilities, Rebuild Data Structure Utilities, Reorganize Indexing Utilities, Shrink/Move Data Structure Utilities, Solution Statistics Utilities,
32.	What are processing utilities? What are its types? Explain each.
UNIT 2: Chapter 6	
33.	Explain the operational management layer.
34.	Explain the Drum-Buffer-Rope scheduling methodology.
35.	What are processing-stream definitions? How are they managed? Explain.
36.	Explain the parameter, monitoring, communication and alerting sections of the operational management layer.
37.	Explain the audit. Balance and control layer.
38.	What is logging? How is logging done? Explain with example.
39.	Explain the different types of watchers.
40.	Explain process tracking, data provenance and data lineage.
41.	State and explain the five fundamental steps of the data science process.
42.	List and explain the structures in the functional layer of the data science ecosystem.
43.	State and explain the six supersteps for processing the data lake.
UNIT 2: Chapter 7	
44.	Explain the retrieve superstep.
45.	Explain data lakes and data swamps.
46.	Enumerate the general rules for data source catalog.
47.	State and explain the four critical steps to avoid data swamps.
48.	How can the following be achieved (use R or python for giving examples)? i. Loading a table from .csv file ii. Display the datatype of column iii. Verification of data field name iv. Adding unique identifier for each row v. Generate histogram across every column vi. Generate minimum and maximum values in each column vii. Generate mean, median and mode of values in ach column viii. Generate range and quartile ix. Calculate standard deviation and skewness x. Identify missing values in dataset xi. Determine pattern of data values
49.	Why is it necessary to train the data science team? Discuss.
50.	Explain the following shipping terms: Seller, Carrier, Port, Ship, Terminal, Named Place, Buyer.
51.	Explain the following shipping terms with example: Ex Works (EXW), Free Carrier (FCA), Carriage Paid To (CPT), Carriage and Insurance Paid To (CIP), Delivered at Terminal (DAT), Delivered at Place (DAP), Delivery Duty Paid (DDP).
52.	With the help of an example, different data stores used in data science.
UNIT 3: Chapter 8	

53.	Explain the assess superstep.
54.	What are the four things that can be done with the errors found in data? Explain.
55.	State and explain the six data dimensions.
56.	What are the different ways of treating missing values in data using pandas package? Explain with example.
57.	Explain node, edge and directed acyclic graph.
58.	How is directed acyclic graph used for scheduling jobs? Explain with example.
59.	"Graph theory is always a useful tool to use when relationships between business entities require analyzing." Explain with example.
60.	What are Vincenty's formulae? Explain their use with example.
61.	Explain processing offloading with example.
62.	How can shortage paths between locations and paths from a given location be found from graph? Explain with example.
63.	How can housekeeping be performed in data science to maximize processing capacity? Explain with example.
UNIT 4: Chapter 9	
64.	Explain the process superstep.
65.	What are the different typical reference satellites? Explain.
66.	Explain the TPOLE design principle.
67.	Explain the Time section of TPOLE.
68.	Explain the Person section of TPOLE.
69.	Explain the Object section of TPOLE.
70.	Explain the Location section of TPOLE.
71.	Explain the Event section of TPOLE.
72.	Explain the different date and time formats. What is leap year? Explain.
73.	Explain local time and Universal Coordinated time.
74.	Discuss the interesting facts about the time zones based on Universal Coordinated Time.
75.	What is golden nominal? Discuss the relationship or meaning embedded in people's names with examples.
76.	State the naming conventions in Dutch culture.
77.	What are the five different kingdoms that classify life on earth? Explain each in brief.
78.	Explain the divisions of Phylum.
79.	Explain the international classification of vehicles with examples.
80.	What is absolute location and relative location? Explain.
81.	Discuss the natural characteristics and human characteristics of a location. Explain the human environment interaction.
82.	What is an event? Explain explicit and implicit events.
83.	What is a fishbone diagram? Explain with example.
84.	What are pareto charts? What information can be obtained from pareto charts?
85.	Explain the use of correlation and forecasting in data science.
86.	State and explain the five steps of data science..
UNIT 4: Chapter 10	
87.	Explain the transform superstep.
88.	Explain the Sun model for TPOLE.

89.	Explain the steps of data exploration and preparation.
90.	Why does data have missing values? Why do missing values need treatment? What methods treat missing values?
91.	Explain the techniques for outlier detection and treatments.
92.	What is feature engineering? What are the common feature extraction techniques?
93.	What is Binning? Explain with example.
94.	Explain averaging and Latent Dirichlet Allocation with respect to the transform step of data science.
95.	Explain hypothesis testing, t-test and chi-square test with respect to data science.
96.	Explain over fitting and underfitting. Discuss the common fitting issues.
97.	Explain precision recall, precision recall curve, sensitivity, specificity and F1 measure.
98.	Explain Receiver Operating Characteristic (ROC) Analysis Curves and cross validation test.
UNIT 5: Chapter 10	
99.	Explain univariate analysis, bivariate analysis and multivariate analysis.
100.	Explain simple linear regression, RANSAC Linear Regression and Hough transform.
101.	What is logistic regression? Explain simple logistic regression, multinominal logistic regression and ordinal logistic regression.
102.	What is clustering? Explain the different clustering techniques.
103.	What is ANOVA? Explain the use of ANOVA in data science with example.
104.	Explain the use of principal component analysis in data science.
105.	What are decision trees? How are they used in data science? Explain.
106.	Explain Support Vector Machines, Support Vector Clustering and Support Vector Networks.
107.	What are association patterns? Explain with examples.
108.	What are clustering patterns? What are different types of clustering patterns? Explain with examples.
109.	Explain Bayesian classification, Sequence analysis and forecasting.
110.	What is machine learning? Explain supervised, unsupervised and reinforcement learning.
111.	Explain bagging with example.
112.	What are random forests? Explain with examples.
113.	Explain computer vision and natural language processing.
114.	What are neural networks? Explain gradient descent, regularization strength and simple neural network.
115.	What is tensor flow? Explain in detail.
UNIT 5: Chapter 11	
116.	Explain the organize superstep.
117.	Explain the report superstep.
118.	How are the results of data science summarized? Explain.
119.	Explain the different types of graphics used in data science.
120.	What is kernel density estimation? Explain with example.
121.	Explain scatter matrix graph with example.
122.	Explain Andrews' curves with their use in data science.

123.	What are parallel coordinates? Explain.
124.	Explain RADVIZ method, log plot and autocorrelation plot.
125.	Explain bootstrap plot, contour graphs and 3D graphs.
126.	Explain the following with reference to the pictures used in data science: i. Channels of images ii. Cutting the edge iii. One size does not fit all