analyticsexam.com
#1 Big Data & Analytics Certification Guide

# E20-065 – EMC DATA SCIENTIST CERTIFICATION QUESTIONS AND STUDY GUIDE

EMC Data Scientist (E20-065)

# Contents

EMC Data Scientist Details ..... 2
EMC Data Scientist Syllabus for E20-065 Exam (Study Aid) ..... 3
EMC Data Scientist (E20-065) Sample Questions ..... 4

# EMC Data Scientist Certification Details

| | |
|---|---|
| **Exam Name** | EMC Data Scientist (EMCDS) |
| **Exam Code** | E20-065 |
| **Duration** | 90 Minutes |
| **Nos. Of Questions In Exam** | 60 Multiple choice or Short Answer Questions |
| **Passing Percentage** | 63 Percentage |
| **Negative Marking** | No Negative Marking |
| **Partial Credit** | No Partial Credit |
| **Reference Book** | |
| **Schedule Your exam** | Pearson VUE |
| **Sample Questions** | EMC Data Scientist Certification Sample Question |
| **Recommended Practice tool** | **EMC Data Scientist Certification Practice Exam** |

# EMC Data Scientist Certification Syllabus for E20-065 (Study Aid)

MAPREDUCE (15%)

- MapReduce framework and its implementation in Hadoop
- Hadoop Distributed File System (HDFS)
- Yet Another Resource Negotiator (YARN)

HADOOP ECOSYSTEM AND NOSQL (15%)

- Pig
- Hive
- NoSQL
- HBase
- Spark

NATURAL LANGUAGE PROCESSING (NLP) (20%)

- NLP and the four main categories of ambiguity
- Text Preprocessing
- Language Modeling

SOCIAL NETWORK ANALYSIS (SNA) (23%)

- SNA and Graph Theory
- Communities
- Network Problems and SNA Tools

DATA SCIENCE THEORY AND METHODS (15%)

- Simulation
- Random Forests
- Multinomial Logistic Regression and Maximum Entropy

DATA VISUALIZATION (12%)

- Perception and Visualization
- Visualization of Multivariate Data

# EMC Data Scientist Exam (E20-065) Sample Questions

- Below are the 6 sample questions which will help you be familiar with (EMCDS) EMC Data Scientist (E20-065) exam style and Structure.

- These questions are just for demonstration purpose, there are many scenario based question are included in **Premium EMC Data Scientist Practice Exam**

- Access to all 125+ questions is available only through premium practice exam available to members at www.analyticsexam .com

**QUESTION 1: You develop a Python script "logisticpy" to evaluate the logistic function denoted as f(y) for a given value y that includes the following Pig code:**

**Register 'logistic.py' using jython as udf; z = FOREACH y GENERATE $0, udf.logistic ($0); DUMP z;**
**What is the expected output when the Pig code is executed?**
**Options:**

A. Tuples (y, f(y))

B. 0

C. Jython is not a supported language

D. Valueof f(y) for ally

**QUESTION 2: You are analyzing written transcripts of focus groups conducted on product X. You approach is to use TF-IDF for your analysis. What combination of TF-IDF scores should you examine to ensure you only report on the most important terms?**
**Options:**

A. Low TF score and low DF score

B. High TF score and low IDF score

C. High TF score and high DF score

D. High TF score and high IDF score

**QUESTION 3: Why would a company decide to use HBase to replace an existing relational database?**
**Options:**

A. Varying formats of input data requires columns to be added in real time.

B. It is required for performing ad-hoc queries.

C. Existing SQL code will run unchanged on HBase.

D. The company's employees are already fluent in SQL.

**QUESTION 4: You conduct a TFIDF analysis on 3 documents containing raw text and derive TFIDF ("data", document y) = 1.908. You know that the term "data" only appears in document 2. What is the TF of "data" in document 2?**
**Options:**

A. 4 based on the following reasoning: TFIDF = TF1DF = 1.908 You then know that IDF will equal LOG (3/1 )=0.477 Therefore, TFIDF=TF'0 477 = 1.908 TF will then round to 4

B. 6 based on the following reasoning: TFIDF = TF1DF = 1.908 You then know that IDF will equal 3/1=3 Therefore, TFIDF=TF/3 = 1.908 TF will then round to 6

C. 11 based on the following reasoning: TFIDF = TF1DF = 1908 You then know that IDF will equal LOG(3/2)=0.176 Therefore, TFIDF=TF"0.176 = 1.908 TF will then round to 11

D. 2 based on the following reasoning: TFIDF = TF1DF = 1 908 You then know that IDF will equal LOG (32)=0.954 Therefore, TFIDF=TF*0.954 = 1.908 TF will then round to 2

**QUESTION 5: Which scenario would be ideal for processing Hadoop data with Hive?**
**Options:**

A. Structured data; batch processing

B. Structured data, real-time processing

C. Unstructured data; batch processing

D. Unstructured data; real-time processing

**QUESTION 6: Which scenario is a proper use case for multinomial logistic regression?**
**Options:**

A. A marketing firm wants to estimate the personal income of a group of potential customers. Using inputs such as age, education, marital status, and credit card expenditures, a data scientist is building a model that will estimate a person's income

B. A manufacturer plans to determine the optimal number of workers to employ in an assembly line process. Utilizing the observed distributions of the task durations of each process step, a data scientist is building a model to mimic the interactions and dependencies between each stage in the manufacturing process.

C. A logistic distribution company wants to minimize the distance traveled by its delivery trucks. A data scientist is building a model to determine the optimal route for each oftistrucks

D. To improve the initial routing of a loan application, a financial institution plans to classify a loan application as Approve, Reject, or Possibly_Approve. Based on the company's historical loan application data, a data scientist is building a model to assign one of these three outcomes to each submitted application.

# Answers:

| | | | |
|---|---|---|---|
| Question: 1 | Answer:A | Question: 2 | Answer: B |
| Question: 3 | Answer:B | Question: 4 | Answer:A |
| Question: 5 | Answer:C | Question: 6 | Answer:D |