Data Engineering Pre-Interview Quiz

Please answer the following screening questions within 24 hours. You are allowed to search the internet for answers.

Section 1: Linux Fundamentals

1. What is the command for displaying list of files in current directory in Linux command line
    a. dir
    b. files
    c. ls
    d. lcd
2. What is the command for displaying the content of a file in Linux command line
    a. view
    b. cat
    c. display
    d. print
3. What is the command for changing current directory in Linux command line
    a. chdir
    b. cd
    c. changedir
    d. ccd
4. Which of the following is not a text editor on Linux command line
    a. emacs
    b. vim
    c. ncat
    d. nano
5. What directory path do log files normally are stored in Linux machine
    a. /log
    b. /tmp/log
    c. /var/log
    d. /sys/log
6. Which of the following commands can't be used to setup networking
    a. ifconfig
    b. ip
    c. netcat
    d. nmcli

7. Which of the following command can be used to find a certain text inside a file
    a. awk

    b. find
    c. grep
    d. lookup
8. Which of the following command can be used to do string search and replace on a text stream
    a. sed
    b. replace
    c. pwd
    d. zstd
9. Where do system configuration files normally resides in
    a. /sys/config
    b. /etc/
    c. /conf/
    d. /var/
10. Where do user's files normally resides in
    a. /home/<username>/
    b. /users/<username>/
    c. /data/<username>/
    d. /sys/<username>/

Section 2: Data Warehousing / Data Lake Fundamentals

11. What do ETL stands for
    a. Extract-Transfer-Load
    b. Extract-Transform-Load
    c. Extract-Transfer-List
    d. Extract-Transform-List
12. Which one of the following activities is usually not part of a data preparation activity
    a. Data ingestion
    b. Data cleansing
    c. Data archiving
    d. Data profiling
13. Which of the following technologies can't used to store and serve data as a Data Warehouse?
    a. PostgreSQL
    b. MariaDB
    c. Sqoop
    d. Hadoop
14. Which one of the following activities is usually not part of data cleansing activity?
    a. Normalization
    b. Deduplication
    c. Predictive modeling
    d. Filling in missing data

15. Which one of the following components is normally not a key component of a Data Warehouse?
    a. Data sources
    b. Data entry
    c. Data mart
    d. ETL
16. Which of the following is commonly the flow of data in a Data Warehouse?
    a. Source -> ETL -> Data Visualization -> ETL -> Data Warehouse
    b. Source -> ETL -> Data Warehouse -> ETL -> Data Mart
    c. Source -> ETL -> Data Analysis -> ETL -> Data Warehouse
    d. Source -> ETL -> Data Archiving -> ETL -> Data Warehouse
17. If there is a requirement to replicate a table, which have rows that are regularly modified from data source into data warehouse in real-time, which of the following tool can be used to extract the data from the source database ?
    a. Batch ETL
    b. Change Data Capture
    c. Database Backup
    d. Message Queue
18. Which of the following data layer will give the best UI performance for front-end Business Intelligence / Visualization tool to visualize data?
    a. Raw data in data warehouse
    b. Pre-aggregated data in in OLAP structure in data mart
    c. Star schema in data warehouse
    d. Raw data in data source

Section 3: Hadoop Platform Fundamentals

19. Which one of the following components is not part of core Hadoop project?
    a. YARN
    b. HDFS
    c. K8S
    d. MR2
20. Which one of the following components provide SQL JDBC access to data residing in Hadoop?
    a. Hive
    b. Pig
    c. Zookeeper
    d. HBase
21. Which one of the following components is not part of HDFS?
    a. Namenode
    b. Datanode
    c. Indexnode
    d. Journalnode

22. If a data engineer wants to develop a distributed data processing job primarily in Python, which of the following component provides the capability to do so?.
    a. Hive
    b. Pig
    c. HBase
    d. Spark
23. Which of the following component is designed to schedule jobs running on a Hadoop platform?
    a. Zookeeper
    b. Oozie
    c. Hive
    d. Spark
24. Which of the following component is designed to ingest large amount of batch RDBMS data into Hadoop in parallel?
    a. Sqoop
    b. Zookeeper
    c. Hive
    d. Pig
25. Which of the following component can't be used to ingest streaming data into Hadoop?
    a. NiFi
    b. Flume
    c. Hive
    d. Kafka
26. Which of the following component is not part of YARN?
    a. Resourcemanager
    b. Nodemanager
    c. Applicationmanager
    d. Application master
27. Which of the following component can be used for developing predictive analytics model?
    a. Hive
    b. Pig
    c. Spark
    d. Flume
28. Which of the following component provide DataFrame API?
    a. Hive
    b. Pig
    c. Spark
    d. MR2

Section 4: SQL Fundamentals

29. What does SQL stands for?
    a. Selection Query Language

b. Structured Query Language
c. Server Query Language
d. SQL Query Language

30. Which of the following SQL statements is used to list out the first 10 records of a table?
    a. select first 10 from datatable
    b. select * from datatable limit 10
    c. select * from datatable first 10
    d. select * from datatable top 10
31. Which of the following SQL statement is used to display list total sales amount, aggregated by each date
    a. select dt, sales_amt from datatable group by dt
    b. select dt, sum(sales_amt) from datatable group by dt
    c. select dt, total(sales_amt) from datatable group by dt
    d. select dt, total(sales_amt) from datatable aggregated by dt
32. To speed up lookup query on a particular column in a table, what would you create?
    a. constraint
    b. index
    c. lookup
    d. catalog
33. When you want to select records from the output of another select statement, what would you do?
    a. Not possible to select from another select statement
    b. Store result of the first select statement into a variable, and select from it
    c. Select from a subquery

Section 5: Python Programming Fundamentals

34. What is the indentation space count required for a python script
    a. 2
    b. 4
    c. 8
    d. Doesn't matter as long as indentations are consistent.
35. Which of the following method is an invalid form of string formatting in Python 3.7?
    a. "Hello {}".format(name)
    b. "Hello ${name}"
    c. "Hello %s" % name
    d. f"Hello {name}"
36. What would be the output of the following code?: map(lambda x: x % 2, range(1,11))
    a. [1,2,3,4,5,6,7,8,9,10]
    b. [2,4,6,8,10,12,14,16,18,20]
    c. [1,3,5,7,9]
    d. [1,0,1,0,1,0,1,0,1,0]
37. What would be the output of the following code?: filter(lambda x: x % 2, range(1,11))
    a. [1,2,3,4,5,6,7,8,9,10]

b. [2,4,6,8,10,12,14,16,18,20]
c. [1,3,5,7,9]
d. [1,0,1,0,1,0,1,0,1,0]

38. What would be the output of the following code?: map(lambda x: x * 2, range(1,11))
    a. [1,2,3,4,5,6,7,8,9,10]
    b. [2,4,6,8,10,12,14,16,18,20]
    c. [1,3,5,7,9]
    d. [1,0,1,0,1,0,1,0,1,0]
39. Which of the following is an invalid native data type in Python 3.7?
    a. list
    b. array
    c. tuple
    d. object
40. Which of the following is an invalid import statement in Python 3.7?
    a. from .module import function
    b. from ..module import function
    c. import function from module
    d. from module import function as myfunction

Section 6: PySpark Programming Fundamentals

41. Which of the following is not a high level component Apache Spark project?
    a. Spark Core
    b. Spark SQL
    c. Spark ML
    d. Spark AI
42. Which of the following is not a component of a Spark cluster?
    a. Driver
    b. Executor
    c. Node Manager
    d. Cluster Manager
43. Which of the following is an invalid DataFrame operation?
    a. df['Name'].collect()
    b. df.select('Name').collect()
    c. df.select(df['Name']).collect()
    d. df.filter(df['Name'] == lit('John')).collect()
44. Which of the following is an invalid RDD operation?
    a. rdd.map(lambda x: (x,1)).reduceByKey(lambda a,b: a+b).collect()
    b. rdd.filter(lambda x: x % 2).collect()
    c. sorted(rdd, key=lambda x: x)
    d. sorted(rdd.collect(), key=lambda x: x)
45. Which of the following will cause performance problem when running on very large datasets (eg: 1TB) (rdd is a 1TB list of random numbers in range 1 to 1000)
    a. rdd.map(lambda x: x * 2).collect().filter(lambda x: x < 10)

b. rdd.filter(lambda x: x < 10).map(lambda x: x * 2).take(10)
c. rdd.map(lambda x: x * 2).limit(10).collect()
d. rdd.map(lambda x: x * (x, 1)).reduceByKey(lambda a,b: a+b).collect()

Section 7: Networking Fundamentals

46. Which of the following is an invalid IPv4 address:
    a. 192.168.1.1
    b. 172.20.115.34
    c. 222.80.11.90
    d. 261.128.111.1

47. Which of the following is an invalid IPv4 CIDR notation
    a. 192.168.0.0/16
    b. 188.70.1.2/32
    c. 178.88.1.20/8
    d. 180.111.28.87/36

48. What is the default/standard port number for DNS service
    a. 80
    b. 443
    c. 53
    d. 22

49. What is the default/standard port number for SSH service
    a. 80
    b. 443
    c. 53
    d. 22

50. What is the default/standard port number for HTTPS service
    a. 80
    b. 443
    c. 53
    d. 22