DSECLZG522 – Big Data Systems

Mid Sem Exam – Sample Questions

Q1. A company decided to perform real time analytics on its Web server logs to gain insights on website visitors, behavior, crawlers accessing the site, business insights, security issues, and more. For this, the log file entries are to be streamed to an HDFS folder in a Hadoop cluster. Streaming the logs to HDFS is to be done only when 1000 or more entries are made to the webserver log. Configure a Flume agent to implement the streaming with webserver log as the source and an HDFS folder as the sink.

(5 Marks)

Q2. Describe the differences between the following 2 tables defined in Cassandra NoSQL database from the point of view of storage and retrieval of the data from the tables. In which scenario, retrieving data from Table.1 will be faster than retrieving data from Table.2

Table.1

```
CREATE TABLE application_logs (
   id              INT,
   app_name  VARCHAR,
   hostname   VARCHAR,
   DateTime   TIMESTAMP,
   env            VARCHAR,
   Log_level   VARCHAR,
   log_msg     TEXT,
   PRIMARY KEY ((app_name, env), hostname)
);
```

Table.2

```
CREATE TABLE application_logs (
   id              INT,
   app_name  VARCHAR,
   hostname   VARCHAR,
   DateTime   TIMESTAMP,
   env            VARCHAR,
   Log_level   VARCHAR,
   log_msg    TEXT,
   PRIMARY KEY ((app_name), env, hostname)
);
```

(4 Marks)

Q3. You are designing a Distributed Hash Table to store 128 data values in a table distributed on a 5- node cluster. The primary key consists of 8 characters consisting of 7-bit ASCII codes.
(1) Design a hashing algorithm for equally distributing data in the tables on the nodes of the cluster.
(2) What percentage of data will get stored on one node.
(3) Given below are 2 schemes for distributing the tokens generated from the partition keys on the 5 nodes of the cluster:
Scheme 1:
Node0– 0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125
Node1 - 1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126
Node2 -2 7 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127
Node3 -3 8 13 18 23 28 33 38 43 48 53 58 63 68 73 78 83 88 93 98 103 108 113 118 123
Node4 - 4 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99 104 109 114 119 124

Scheme 2 :
Node 0 – 0 to 25 (both inclusive)
Node 1 – 26 to 51 (both inclusive)
Node 2 – 52 to 77 (both inclusive)
Node 3 – 78 to 103 (both inclusive0
Node 4 - 104 to 127 (both inclusive)
Which of the scheme mentioned above will be giving almost equal distribution of tokens on all nodes of the cluster? Give justification to your answer.

(5 Marks)

Q4**.** You have a 16384 MB file stored on HDFS as part of a Hadoop 3.x distribution. A data analytics program stores this file on the HDFS cluster with 3 data nodes and runs in parallel across the cluster nodes.

(a) The default values for HDFS block size and the replication factor is used in the configuration. How many total blocks of the data file including replicas will be stored in one node of the cluster?

(b) The cluster has 48 cores to speed up the processing. If the program can at best achieve 60% parallelism in the code to exploit the multiple cores and the rest of it is sequential, what is the theoretical limit on speed-up you can expect with 48 cores compared to a sequential version of the same program running on one core with the same file? How will this limit change if you doubled the compute power to 96 cores? You can simplify the system to assume cluster nodes and cores mean the same and we can ignore the overheads of communication etc. depending on the specific cluster configuration, scheduling etc.

(c) Suppose you could use a more scalable algorithm with 80% parallelism and a larger file as you move to a 108-core system. What would be the theoretical speed-up limit for 108 cores?

(6 Marks)

Q5. An automobile manufacturing company is implementing a Cassandra based solution to automate their manufacturing process. Each machine used in the manufacturing process contains one or more sensors that captures data such as temperature, speed of a moving part, etc. This data is going to be stored in a Cassandra cluster that you must design and maintain. Once the application is up and running, the data it captures will be provided to analytics tool chains. The application will definitely be write heavy; tens of millions of events will be generated per second, and the company has an SLA requiring that all events be persisted across multiple nodes in the Cassandra cluster in less than 10ms. Suggest a suitable PRIMARY KEY (Primary and clustering columns) for the table given below:

```
CREATE TABLE sensor_data (
        serial_number text,
        date text,
        snapshot_time timestamp,
        facility_id int,
        sensor_type text,
        sensor_value text,
);
```

Q6. A worker node used in a Hadoop cluster has 128 GiB memory with 8 CPUs having 6 cores each (Total of 48 cores). We need to reserve 1 core each for the Operating system, HDFS DataNode and YARN NodeManager. 8 Gib memory is reserved for the Operating system and 2 Gib each for the DataNode and the NodeManager. For handling the task overheads, 4 Gib memory is to be reserved. One container with Application Master (AM) may get deployed on this node. The AM container will require 2 Gib memory and 1 Core. (Total of 4 cores reserved for OS, DN, NM and AM).

1. How many vCores will be available for deploying containers for (1) CPU intensive tasks and for (2) standard I/O bound tasks.
2. If you are planning to use this node to launch containers for running standard I/O bound application tasks, give an estimate of the number of containers you will be deploying on this node. Describe a scheme for distributing the available resources (vCores and Memory) among these containers for achieving maximum performance.

(5 Marks)

Q7. A table named **world** is created in Hive using the following command:

```
create table world (country string, population int, GDP float, GINI float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

Data is loaded into this table from a CSV file.
Write HiveQL queries to find the following:
1. List of countries in which the population is more than 10 million
2. The country with the highest GDP
3. The country with the lowest value for the GINI coefficient
4. Name of 5 countries with the highest population.

(5 Marks)

Q8. Why is it recommended to use HDFS for storing large volume data files but not when there are lots of small files? Describe the scenarios in which you set different block sizes for files in HDFS?

(4 Marks)

Q9. What limitation of Hadoop is getting addressed in HBase? HBase persists data in HTables in HDFS and Cassandra stores data in SSTables. HTables and SSTables are immutable. Explain how updates are deletion of rows is implemented in HBase and Cassandra?

(5 Marks)

Q10. You are a data engineer working for a large e-commerce company that uses Hadoop and its ecosystem for data processing and analysis. Your company is dealing with a massive amount of transaction data stored in HBase tables, and you need to extract and analyze this data using Hive and Pig. Additionally, you want to move data between Hadoop and a relational database using Sqoop. Your goal is to optimize data processing tasks efficiently.
You have an HBase table containing transaction data with the following structure:
- Column Family: "info"
- Columns: "user_id," "product_id," "timestamp," "amount"
You need to write a Pig script to find the total sum of "amount" for all transactions. Assume the HBase table is named "transactions."

(6 Marks)

Q11. You are designing a NoSQL database for a social media platform. Each user can have multiple followers, and each follower can follow multiple users. How would you model this relationship in a NoSQL database, and which type of NoSQL database type (e.g., document-oriented, Key-Value, Graph, Column-family) would you choose for this scenario? Provide a schema and give justification for your choice.

(5 Marks)

Q12. What is the role of the MapReduce programming paradigm in big data analytics, and how does it implement the divide-and-conquer strategy? What are the different types of parallelism used in MapReduce computing paradigm?

(4 Marks)

Q13. You are given access to 10 Terabyte of data with details of daily sales transactions belonging to an e-commerce company. The transactional data files are stored in a folder on HDFS with a structure as shown below:

Transaction No. (Unique)	Invoice	Stock Code	Description	Quantity	Invoice Date	Price($)	Customer ID	Country

It is not required to modify the data. You are not allowed to copy or move the data into any other storage location. Select a suitable analytics tool from the Hadoop ecosystem family to run queries on this data. Give justifications for your selection. Using the tool selected by you, write a query to find the total sales revenue received in $ from each country.

(6 Marks)

Q14. What is meant by Tunable consistency in NoSQL databases? What is the effect of Write and Read latency on the consistency levels. What consistency levels are recommended by you for (a) Write intensive applications, and for (b) Read intensive applications?

(4 Marks)

Q15. Explain the role of Apache ZooKeeper in the Hadoop ecosystem. How does ZooKeeper ensure coordination and synchronization among distributed components like HBase, HDFS, and YARN? Provide a real-world scenario where ZooKeeper's coordination is critical.

(6 Marks)

Q16. What are the factors contributing to MTBF? Define Availability. When do you say that a system is highly available? A middleware product has memory leaks and it core dumps every 10 hours on an average. Upon failure, it is automatically restarted and will be back into operation within 3 minutes. Find out the availability of the middleware product?

[4 Marks]

Q17. What is the purpose of database sharding? What is the difference between sharding and replication?

[4 Marks]

Q18. An enterprise company need to address the following challenges with Big Data Analytics:

(a) Struggling with a Java based ELT tool to manage and extract value from the growing volume and variety of data and need to unify information across federated sources. Each disparate MySQL server records around 2 million updates, a total of 50 million updates daily.
(b) Unable to relate raw data collected from system logs, sensors, or click streams with customer and line-of-business data managed in enterprise systems.
(c) Each end-to-end ELT process consumes 14 hours. This process should be completed within an hour.
(d) Some data warehousing queries on structured data are available only in SQL and there is no expertise available for transforming these SQL queries into Hadoop eco system tools.

Recommend suitable solutions to address the above challenges by making use of Hadoop Ecosystem products/tools. Give justifications for your recommendations.

[6 Marks]

Q19. In a 3 node Cassandra cluster, a Keyspace is created with a replication factor of 3.
Answer the following questions, in each of the scenarios mentioned below:
(a). Consistency of reads. (b). How many nodes can be lost without data loss.
(c). What percentage of data is held in each of the nodes.
Scenario 1 : Write Level - Quorum, Read Level - One
Scenario 2: Write Level - All, Read Level - One
Scenario 3: Write Level - One, Read Level - All
Scenario 4: Write Level - Two, Read Level - One

[4 Marks]

Q20. An application is deployed on a cluster with little availability support. A node in a cluster fails every 100 hours while other parts never fail. On failure of the node, the whole cluster needs to be shutdown, the faulty node is to be replaced and the cluster system to be restarted. This takes 2 hours. The application also needs to be restarted, which takes 2 hours. What is the availability of the application? If downtime cost is $40000 per hour, what is the yearly downtime cost?

[4 Marks]

Q21. You have a 928 MB file stored on HDFS as part of a Hadoop 2.x distribution. A data analytics program uses this file and runs in parallel across the cluster nodes.
(a) The default block size and replication factor are used in the configuration. What will be the total number of blocks including the replicas that will be stored in the cluster? What are the unique HDFS block sizes you will find for the specific file?
(b) The cluster has 64 cores to speed up the processing. If the program can at best achieve 60% parallelism in the code to exploit the multiple cores and the rest of it is sequential, what is the theoretical limit on speed-up you can expect with 64 cores compared to a sequential version of the same program running on one core with the same file? How will this limit change if you doubled the compute power to 128 cores. You can simplify the system to assume cluster nodes and cores mean the same and we can ignore the overheads of communication etc. depending on the specific cluster configuration, scheduling etc.
(c) Suppose you could use a more scalable algorithm with 80% parallelism and the size of the file is doubled as you move to the 128 cores cluster. What would be the theoretical speed-up limit for the 128 cores cluster?

[6 Marks]

Q22. What are the scenarios in which Flume and Sqoop are used? What are the major differences between Flume and Sqoop? Configure a Flume agent to transfer files from a local folder to an HDFS folder.

[6 Marks]

Q23. What is meant by Consistency in CAP theorem? Let's establish a few definitions:
N = The number of nodes that store replicas of the data.
W = Number of replicas that need to acknowledge the receipt of update before update completes.
R = Number of replicas that are contacted when a data object is accessed through a read operation.
What type of consistency can be guaranteed in the following scenarios?
(a) W+R > N
(b) N=2, W=2, and R=1
(c) N=2, W=1, and R=1

Q24. In each of the following scenarios, point out and give a brief reason what type of multi processor computer one would use as per Flynn's taxonomy, i.e. the choices are SIMD,SISD, MIMD, or MISD.
(a) A scientific computing application does a f1(x) + f2(x) transformation for every data item x given f1 and f2 are specialised operations built into the hardware.
(b) A video is processed to extract each frame which can be either an anchor frame (full image) or a compressed frame (difference image w.r.t anchor). A compressed frame (C) is transformed using a function f, where each pixel is compared with the last anchor (A) to recreate the uncompressed image (B), i.e. B(i, j) = f(C(i, j), A(i,j)) for all pixels (i,j) in the input frames.
(c) A multi-machine Apache Hadoop cluster system for data analysis.
(d) A development system with multiple containers running JVMs and MongoDB nodes running on a single multi-core laptop.

(5 Marks)

Q25. Describe the most appropriate scheduling mechanism to be configured with Hadoop YARN in the following scenarios:
(a). A company wants to run jobs of equal size on a Hadoop cluster. The number of jobs can vary at any point in time.
(b). A company wants to run short jobs and long jobs on the same cluster and short jobs cannot always be starved by long jobs.
(c) A company wants to divide the Hadoop cluster resources in different proportions between the departments and does not care about wastage of unutilized resources.
(d) A company wants to run jobs of different nature concurrently on the same cluster and is worried about wastage of resources.

(5 Marks)

Q26. A 2-tier application uses a 3-node application cluster and a 2-node DB cluster. The application works only when all tiers are available. The application tier is in an active-active load balanced configuration with the given nodes. But the database tier is in a cold standby mode where it takes 12 hours to bring back a passive node online. If an application node fails every 10 days and a DB node fails every 100 days, find the following:
(a) MTTF of the application tier.
(b) MTTF of the database tier.
(c) Availability of the database tier.
(d) Overall availability of the 2-tier system, assuming MTTR of the application tier is negligible.

(4 Marks)

Q27. A company decided to perform real time analytics on its Web server logs to gain insights on website visitors, behavior, crawlers accessing the site, business insights, security issues, and more. For this, the log file entries are to be streamed to an HDFS folder in a Hadoop cluster. Streaming the logs to HDFS is to be done only when 1000 or more entries are made to the webserver log. Configure a Flume agent to implement the streaming with webserver log as the source and an HDFS folder as the sink.

(7 Marks)

Q28. You are assigned the responsibility to transfer huge volume of data from RDBMS table to HDFS for daily analytics on the data. The time window allocated for the data transfer is fixed and the volume of the data increases day by day. Which data ingestion tool will be recommended by you. Give justifications for your selection.

(4 Marks)

Q29. You have a 928 MB file stored on HDFS as part of a Hadoop 2.x distribution. A data analytics program uses this file and runs in parallel across the cluster nodes.
(a) The default block size and replication factor is used in the configuration. How many total blocks including replicas will be stored in the cluster ? What are the unique HDFS block sizes you will find for the specific file?
(b) The cluster has 64 cores to speed up the processing. If the program can at best achieve 60% parallelism in the code to exploit the multiple cores and the rest of it is sequential, what is the theoretical limit on speed-up you can expect with 64 cores compared to a sequential version of the same program running on one core with the same file ? How will this limit change if you doubled the compute power to 128 cores ? You can simplify the system to assume cluster nodes and cores mean the same and we can ignore the overheads of communication etc. depending on the specific cluster configuration, scheduling etc.
(c) Suppose you could use a more scalable algorithm with 80% parallelism and a larger file as you move to a 128 core system. What would be the theoretical speed-up limit for 128 cores ?

(5 Marks)

Q30. An HBase instance is configured to store 100000 rows in one Hfile. When all these Hfiles became full, compaction is initiated. Before compaction there were 10 Hfiles. Updated rows occupy 20% rows in each of these Hfiles. Assume that the updates happened only once. And 10% of the rows in each of these Hfiles were newly added. And 10% of the rows in each of the Hfiles were tagged as DELETED rows. Find out the total number of Hfiles resulting from the compaction done on the 10 Hfiles.

[Marks 4]

Q31. A healthcare data analytics application utilizes a Cloud-based PostgreSQL cluster. The analytics team conducts patient record analysis on the data stored on the Cloud instance. It takes approximately 30 seconds per user query on average. You are tasked to implement a Memcached based cache in the healthcare center's data center to expedite these queries, aiming for an average query latency of 10 seconds. The Memcached cache has an access latency of 5 seconds. When the cache is established, data can only be transferred from the cache to the application and not directly from the Cloud DB. What should be the projected hit rate in the cache to achieve this ambitious average query latency target?
What type of locality of reference is used here?

[Marks 4]

Q32. A system has a Mean Time To Failure (MTTF) of 800 hours and a Mean Time To Repair (MTTR) of 8 hours. Calculate the system's availability, given that the system's downtime is solely due to repair time, and then discuss how the relationship between MTTF and MTTR impacts the system's overall reliability. Additionally, using your calculated availability, propose a maintenance scheduling strategy for a real-world data center operation, considering the potential consequences of system downtime on business operations.

[Marks 5]

Q33. What are the 2 different types of pluggable schedulers that can be used in Hadoop YARN? Which one of them guarantees maximum utilization of cluster resources? Fair share scheduler is plugged into YARN resource manager with queues Q0, Q1, Q2, and Q3. All these queues are configured to pre-empt containers from other queues. The weights allocated to each of the queues are given below:
Q0 - weight 0.00
Q1 - weight 0.50
Q2 - weight 0.30
Q3 - weight 0.20
What percentage of total resources will be used by the jobs at run time, in each of the following scenarios?
(a) Job1 is submitted to Q1 and it is the only job running on the cluster
(b) While Job1 was running in Q1, Job2 is submitted to Q2
(c) While Job1 was running in Q1 and Job2 was running in Q2, Job3 is submitted to Q3
(d) While Job1 was running in Q1, Job2 in Q2 and Job3 in Q3, Job4 is submitted to Q0

[Marks 5]

Q34. Describe the most appropriate scheduling mechanism to be configured with Hadoop YARN in the following scenarios:
(a). A company wants to run jobs of equal size on a Hadoop cluster. The number of jobs can vary at any point in time.
(b). A company wants to run short jobs and long jobs on the same cluster and short jobs cannot always be starved by long jobs.
(c) A company wants to divide the Hadoop cluster resources in different proportions between the departments and does not care about wastage of unutilized resources.
(d) A company wants to run jobs of different nature concurrently on the same cluster and is worried about wastage of
resources

[Marks 5]

Q35. Suppose you are working with a large e-commerce database that contains customer orders, and you want to analyze the average order total for each customer in the year 2020. The 'orders' table has the following schema: order_id (int), customer_id (int), order_date (string in the format 'yyyy-mm-dd'), and order_total (double).

Using Hive, design a query that calculates the average order total for each customer, considering only the orders made in the year 2020. Be sure to include a description of how your query works and any assumptions you made about the data.

Additionally, discuss how you would optimize this query for performance if the 'orders' table contains millions of records, and explain the trade-offs between different optimization strategies.

[Marks 5]

****