Fall 2023 - Cloud Computing and Big Data Systems

Quiz 2. 12/08/2023

This is an open book, open Internet Quiz. However you cannot use chatGPT or any similar tools, set your brightness to maximum. Use incognito to search throughout the quiz if you have google's generative AI-powered Search enabled in your browser.

A: Spark Programming [25 points]

All coding answers expected in Python language. Your code should not have logical errors and be close to compile-ready (no major syntax errors etc)

Using ***products*** and ***orders*** datasets as shown in the class, answer the following

A.1 [5] Provide the code both in Spark SQL and Dataframe API based to Identify the Top 5 Products with the Highest Reorder Frequency.

A.2 [5] Provide the code both in Spark SQL and Dataframe API based to Identify Aisles with the Highest Number of Unique Orders.

A.3 [5] Provide the code both in Spark SQL and Dataframe API based to categorize orders based on whether they contain products from a single department or multiple departments.

Expected output:

order_id	number_of_departments	order_type
1001	1	Single Department
1002	3	Multiple Departments

A.4 Spark RDD [10]

We had covered in the lecture the example where we find the number of ham radio log messages by country name. For that example, consider the required log file as the input with appropriate columns.

You need to list all the RDDs that are generated due to transformation for that code and explain through example the content for each RDD. You also need to list any actions. You need to illustrate with example content for this question.

B. System Design [30] Real-Time Video Camera Feed based Image Analysis

Your task is to develop a system architecture for a corporate complex with significant public access due to an integrated local subway station. This complex experiences a high volume of people movement, with millions passing through. The system is designed to process live video feeds, operating at 10 frames per

second, to identify individuals within the complex. The goal is to determine if the individuals are employees, registered visitors, or unknown persons. You have access to initial images of employees and registered visitors. The visitor list may change dynamically. Building upon an existing architecture of Assignment 3 [which involves a smart photo album capable of searching photos by labels and integrating AWS Rekognition and Elasticsearch], the new system must:

1. APIs and Backend Architecture:

    - List all APIs that will be used in the system.
    - Describe the backend architecture, focusing on AWS services and infrastructure suitable for a high-traffic corporate setting.
    - Ensure that the design is scalable, event-driven, and asynchronous, capable of handling the demands of a large corporate complex with public access.

2. Data Stores and Content:

    - Clearly list all the data stores used in the system.
    - Describe the type of data each store will hold, including any indexing mechanisms, focusing on handling the high volume of data.

3. System Design Architecture:

    - Develop an architecture that integrates with the existing smart photo album system.
    - The architecture should be capable of processing high-volume live video feeds to identify individuals in a busy corporate complex.
    - Implement a real-time notification system to alert the relevant company personnel (if employee) and complex personnel about the identity and location of the individuals. This data is also fed into the system for later analytics.

4. Handling High-Volume Streaming Video Data:

    - Detail the approach for ingesting and processing high-volume streaming video data in your system.
    - Specify the AWS components used for handling large-scale streaming data.

5. Data Pipeline and Analytics:

    - Outline the data pipeline, from ingestion to notification, for a corporate complex scenario.
    - Include both real-time and offline analytics processes suitable for high traffic environments.
    - Define the events that trigger these analytics computations, considering the scale and diversity of the complex.
    - Assume access to necessary analytics as libraries or external APIs.

This design must efficiently manage the complexities of a high-traffic corporate environment, ensuring robustness, efficiency, network downtimes and adaptability to dynamic changes in visitor data. It should leverage AWS services optimally to handle the challenges of real-time video processing and person identification in a busy and diverse setting.

C. Lecture Contents (Multiple Choice) [30]

1. What is the difference between a Pod and a Deployment in Kubernetes?

A) A Pod is a deployment strategy, while Deployment is a single container.
B) A Deployment is a group of Pods, while a Pod is the smallest deployable unit.
C) A Pod is used for scaling, while Deployment is a runtime environment.
D) A Pod represents a service, while Deployment represents a container image.

2. What is the role of the Kubernetes Master in a cluster?

A) Manages the deployment of applications
B) Runs user applications in containers
C) Controls and coordinates the overall cluster
D) Provides storage volumes for containers

3. How do you scale the worker node capacity in an Amazon EKS cluster?

A) Update the EKS Control Plane configuration
B) Add or remove EC2 instances from the Auto Scaling Group
C) Use the eksctl command to resize the cluster
D) There is no way to scale worker nodes in EKS.

4. What are DataFrames in Apache Spark, and how do they differ from RDDs?

A) DataFrames are a collection of key-value pairs, while RDDs are tables.
B) DataFrames are a higher-level abstraction built on top of RDDs, providing a structured representation of data.
C) DataFrames and RDDs are interchangeable terms in Spark.
D) DataFrames are only suitable for batch processing, while RDDs are for real-time processing.

5. In AWS CloudFormation, what is the purpose of the "DependsOn" attribute in a resource definition?
a. Specifies the resource's dependencies on other AWS services.
b. Declares the order of execution for the CloudFormation stack.
c. Identifies the resources that are dependent on the current resource.
d. Enables parallel execution of resources for faster stack creation.

6. What is the purpose of a stack policy in AWS CloudFormation?
a. Enforces IAM policies on CloudFormation stacks.
b. Defines the parameters for the CloudFormation stack.
c. Specifies the permissions required for resources within a stack.

d. Controls updates to stack resources by allowing or denying specific actions.

7. What is the primary distinction between a Deployment and a Service in Kubernetes?

a. Deployments manage container scaling, while Services manage pod communication.
b. Deployments handle container orchestration, while Services manage networking and load balancing.
c. Deployments ensure high availability, while Services provide persistent storage for pods.
d. Deployments control resource allocation, while Services handle container image deployment.

8. What is the primary purpose of Docker containers?
a. Virtualizing hardware resources
b. Running multiple operating systems on a host
c. Isolating and packaging applications with their dependencies
d. Managing networking configurations
.
9. How does Kubernetes maintain the desired state of a deployment?
A. By using the Kubelet to periodically check and adjust the state of each pod.
B. Through the control loop in the Kubernetes controller manager.
C. By manually adjusting the state based on user input.
D. Using an external state management system.

10. What is the role of etcd in a Kubernetes cluster?
A. It acts as the primary database for Kubernetes, storing all cluster data.
B. It is used for load balancing traffic between pods.
C. It monitors the health of nodes in the cluster.
D. It schedules pods to run on various nodes.

11. In Kubernetes, what is the difference between a ReplicaSet and a Deployment?
A. A ReplicaSet provides rollback and update functionality, while a Deployment does not.
B. A Deployment is for stateful applications, while a ReplicaSet is for stateless applications.
C. A Deployment manages ReplicaSets and provides declarative updates to pods.
D. There is no difference; they are functionally identical.

12. What mechanism does Kubernetes use for service discovery?
A. It uses an internal DNS server for service discovery.
B. Services are discovered via environmental variables.
C. It relies on an external service discovery mechanism.
D. Both A and B.

13. How does Kubernetes achieve fault tolerance at the application layer?
A. By restarting pods that exit or get evicted.
B. Through manual intervention by the cluster administrator.
C. By using multiple master nodes in the cluster.
D. By replicating the application across multiple cloud providers.

14. What is the purpose of Kubernetes namespaces?
A. To provide isolation at the node level.
B. To enable different teams or projects to share a cluster without interference.
C. To manage the different versions of an application.

D. To isolate network traffic between pods.

15. In a Kubernetes cluster, how is the master node different from worker nodes?
A. The master node runs application workloads, while worker nodes manage the cluster state.
B. The master node contains the etcd database, while worker nodes run the Kubernetes API server.
C. The master node schedules workloads and manages the cluster, while worker nodes run the scheduled workloads.
D. There is no difference; all nodes perform the same functions.

D. Papers [15]

D1 [5].[Kafka] What kind of ordering guarantees are provided by Kafka? What kind of ordering cannot be guaranteed and why?

D2 [5]. [MapReduce] What are
a) the actions taken by the master in response to a worker failure
b) the implications for both map and reduce tasks in the event of a worker failure

D3 [5]. [Spanner] How does the Spanner database system achieve global consistency and high availability, and what are the trade-offs involved in its design?