LTC-116-CC-18

SASTRA
DEEMED TO BE UNIVERSITY

School of Computing
First CIA Exam –Aug 2024
Course Code: CSE407R01
Course Name: CLOUD COMPUTING
Duration: 90 minutes Max Marks: 50

PART A (4*5 = 20)

Answer all the questions

1. Generate a concise list of the technical characteristics of cloud computing.
2. How would you use Globus to transfer and share data securely in cloud?
3. You are deploying a scalable web application on a cloud server and need to choose between Virtual Machines (VMs) and Containers. Compare these two technologies in terms of architecture, resource allocation, isolation, and typical use cases. Based on this comparison, which technology would you recommend for your deployment and why?
4. Design a detailed step-by-step guide on how you would attach and configure this extra storage to your VM. Consider the types of storage available, the necessary configurations, and best practices to ensure data integrity and performance.

PART B (3*10 = 30)

Answer all the questions

5. Explain in detail the cloud storage models and Compare the storage levels with Public cloud service providers.
6. Create a own container with necessary steps and justify the need of the containers and explain the basics of the containers.
7. How can an organization determine the most suitable cloud deployment model for deploying their specific application, considering factors like performance requirements, data security, and scalability?

SASTRA ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION DEEMED TO BE UNIVERSITY (U/S 3 OF THE UGC ACT, 1956) THINK MERIT | THINK TRANSPARENCY | THINK SASTRA	**School of Computing** **First CIA Exam –Sept 2023** Course Code: CSE407R01 Course Name: CLOUD COMPUTING Duration: 90 minutes Max Marks: 50

PART A Answer all the questions (5*4 = 20)

Q.No	Questions
1.	Generate a concise list of the technical characteristics of cloud computing. The five technical characteristics of cloud computing is as follows: 1. Large-scale computing resources: Cloud computing systems are often implemented in large scale to reduce cost and providers can benefit from significant economies of scale. 2. Shared resource pooling: Resource pooling forms the foundationb for cloud scalability. Pooling of virtual and physical resources into the cloud provides an abstraction of resource locationb independence phenomenon, which in turn shields cloud consumersb from being aware of the real location of share resources in use. 3. Dynamic resource scheduling: Via software automation, cloud resources are provisioned dynamically based on the current demand requirements and ensures that the cloud service capability is being expanded or contracted based on demand. 4. High scalability: Cloud computing architecture allows you to add new nodes and servers with no reconfiguration and remodification requirements to cloud software and infrastructure. 5. Rapid elasticity: Computing resources can be rapidly and elastically provisioned and released based on the demand of the consumer
2.	How would you use Globus to transfer and share data securely in cloud? The figure depicts a series of five data manipulation operations. (1) A researcher requests, for example via the Globus web interface, that a set of files be transferred from a sequencing center to another storage system, in this case the Google Drive cloud storage system. (2) The transfer then proceeds without further engagement by the requesting researcher, who can shut down her laptop, go to lunch, or do whatever else is needed. The Globus cloud service (not shown in the figure) is responsible for completing the transfer, retrying failed transfers if required, and notifying the user of failure only if repeated retries are not successful. The user requires only a web browser to access the service, can transfer data to and from any storage system that runs the Globus Connect software, and can authenticate using an institutional credential.

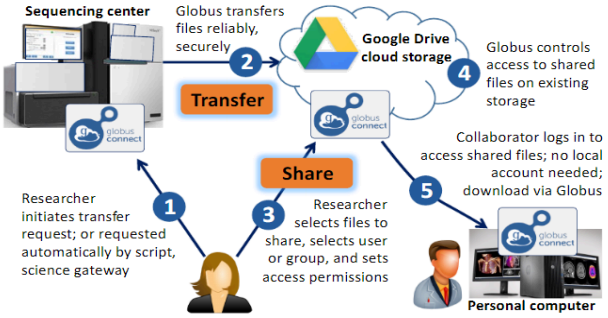


Figure 3.7: Globus transfer and sharing services used to exchange data among sequencing center, remote storage system (in this case, Google Drive), and a personal computer.

3. You are deploying a scalable web application on a cloud server and need to choose between Virtual Machines (VMs) and Containers. Compare these two technologies in terms of architecture, resource allocation, isolation, and typical use cases. Based on this comparison, which technology would you recommend for your deployment and why?

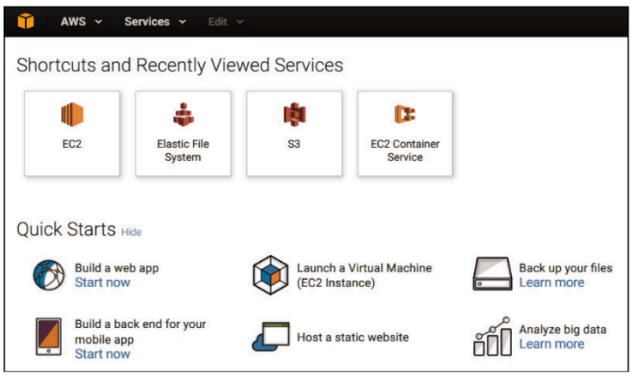
With detailed explanation full marks.

Virtual machines	Containers
Heavyweight	Lightweight
Fully isolated; hence more secure	Process-level isolation; hence less secure
No automation for configuration	Script-driven configuration
Slow deployment	Rapid deployment
Easy port and IP address mapping	More abstract port and IP mappings
Custom images not portable across clouds	Completely portable

4. Design a detailed step-by-step guide on how you would attach and configure this extra storage to your VM. Consider the types of storage available, the necessary configurations, and best practices to ensure data integrity and performance.

Any VM creation step by step [5 marks]

Amazon portal at aws.amazon.com, where we can log in or create an account.



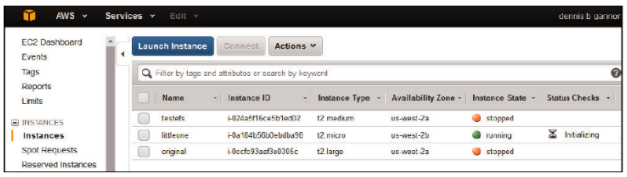


Figure 5.2: Portal instance view.

- Connect to your instance, you need to use a secure shell command. On Windows the tool to use is called PuTTY.
- You need a companion tool called PuTTYgen to convert the downloaded private key into one that can be consumed by PuTTY.
- When you launch it, you use ec2-user@IPAddress, where the IPAddress is the IP address you can find in the Portal Instance View.
- The PuTTY SSH tab has an Auth tab that allows you to upload your converted private key. On a Mac or Linux machine, you can go directly to the shell and execute
- ssh -i path-to -your- private -key .pem ec2 - user@ipaddress -of - instance

PART B **Answer all questions** **(3 * 10 =30)**

5. Explain in detail the cloud storage models and Compare the storage levels with Public cloud service providers.

The right storage model for a data collection can depend on not only the nature and si ze of the data, but also the analyses to be performed, sharing plans, and update frequencies, among other factors

1. FILE SYSTEMS
2. OBJECT STORES
3. RELATIONAL DATABASES
4. NOSQL DATABASE
5. GRAPH DATABASE
6. DATA WAREHOUSE

Detailed explanation with public cloud [10 marks]

6.	Create a own container with necessary steps and justify the need of the containers and explain the basics of the containers. Containers • Containers have become one of the most interesting and versatile alternatives to virtual machines for encapsulating applications for cloud execution. • They enable you to run an application without source code modification or even recompilation on your Windows PC, your Mac, or any cloud. • Docker is a container technology, because it is the most widely known and used, is easy to download and install, and is free. • The Linux kernel had some nice features that could be used to bound and contain the resource utilization of processes: in particular, control groups and name space isolation. • These features allow for the layering of new private virtual file system components on top of the host file system and a special partitioned process space for applications to run using libraries virtually stored in the layered file system. • A contained application uses the resources of the host OS, which can even control the amount of resource devoted to each container: for example, the CPU percentage and the amount of memory and disk space • Docker isolation provides a way to factor large applications, as well as simple ways for running containers to communicate with each other. • When Docker is installed on Linux, Windows 10, or Mac, it runs on a base Linux kernel called Alpine that is u sed for every container instance. • Additional OS features are layered on top of that base. This layering is the key to container portability across clouds. • Docker is designed to support a variety of distributed applications. It is now widely used in the Internet industry, including by major companies like Yelp, Spotify, Baidu, Yandex, and eBay. • The key is the Union File System (more precisely, the advanced multilayered unification file system (AuFS) and a special property called copy on write that allows the system to reuse many data objects in multiple containers. • Docker images are composed of layers in the Union File System. The image is itself a stack of read-only directories. • The base is a simplified Linux or Windows file system. • additional tools that the container needs are then layered on top of that base, each in its own layer. When the container is run, a final writable file system is layered on top.

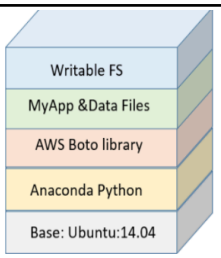


Figure 6.1: The Docker Union File System is layered on a standard base.

- As an application in the container executes, it uses the writable layer.
- If it needs to modify an object in the read-only layers, it copies those objects into the writable layer.
- Otherwise, it uses the data in the read-only layer, which is shared with other container instances.
- Thus, typically only a little of the container image needs to be actually loaded when a container is run, which means that containers can load and run much faster than virtual machines.
- In fact, launching a container typically takes less than a second, while starting a virtual machine can take minutes

Docker and Hub

- The Docker website, Docker.com, provides the tools needed to install Docker on a Linux, Mac, or Windows 10 PC.
- This site al so links to the Docker Hub, a public resource where you can store your own containers, and search for and download any of the hundreds of public containers.

Install Jupyter with Docker on your laptop. You must first install Docker on your machine. While the details differ on Linux, Mac, or PC, the installation is a simple process, similar to that of installing a new browser or other desktop application. Follow the download and install instructions on the Docker.com website. Docker does not have a graphical interface: it is based on a command line API. Hence, you need to open a "powershell" or "terminal" window on your machine. The Docker commands are then the same on Linux, Mac, or PC.

Once you have installed Docker, you can verify that it is running by executing the `docker ps` command, which tells you which containers are running. You should see the following output, as no containers are running.

```
C:\> docker ps
CONTAINER ID   IMAGE   COMMAND   CREATED   STATUS   PORTS     NAMES
C:\>
```

We launch Jupyter with the `docker run` command. The following example uses two of the many parameters supported. The first flag, `-it`, causes the printing of a URL with a token that you can use to connect to the new Jupyter instance. The second flag, `-p 8888:8888`, binds port 8888 in the container's IP stack to port 8888 on our machine. Finally, the command specifies the name of the container, `jupyter/scipy-notebook`, as it can be found in the Docker Hub.

```
C:\> docker run -it -p 8888:8888 jupyter/scipy-notebook
Copy/paste this URL into your browser when you connect for the
first time, to login with a token:
  http://localhost:8888/?token=b9fc19aa8762a6c781308bb8dae27a...
```

Rerunning the `docker ps` command shows that our newly started Jupyter notebook is now running. (Confusingly, each of the two output lines is wrapped.)

```
C:\> docker ps
CONTAINER ID    IMAGE                    COMMAND                CREATED
STATUS          PORTS                    NAMES
6cb4532fa0b     jpyter/scipy-notebook    "tini--start-note"     6 seconds ago
up 5 seconds     0.0.0.:8888->8888/tcp   prickly_meitner
C:\>
```

The following commands illustrates the use of both `-it` and `-v`. We first use the `docker` command on a Mac to launch a Linux Ubuntu container with the Mac's `/tmp` directory mounted as `/localtmp`. Due to `-it`, we are presented with a command prompt for the newly started Ubuntu container. We then run `df` in the container to list its file systems, which include `/localtmp`.

```
docker run -it -v /tmp:/localtmp ubuntu
root@3148dd31e6c7:/# df
Filesystem     1K-blocks      Used Available Use% Mounted on
none            61890340  41968556  16754860  72% /
tmpfs            1022920         0   1022920   0% /dev
tmpfs            1022920         0   1022920   0% /sys/fs/cgroup
osxfs          975568896 143623524 831689372  15% /localtmp
/dev/vda2       61890340  41968556  16754860  72% /etc/hosts
shm                65536         0     65536   0% /dev/shm
root@3148dd31e6c7:/#
```

Notice that when we connect to the new container, we do so as root. Running Jupyter always presents a security challenge, especially if you are running it on a machine with a public IP address. You should certainly use `HTTPS` and a password, especially if you are running it on a remote VM in the cloud. We can configure these options by using the `-e` flag on the `run` command to pass environment flags through to Jupyter. For example, `-e GEN_CERT=yes` tells Jupyter to generate a self-signed SSL certificate and to use `HTTPS` instead of `HTTP` for access. To tell Jupyter to use a password, we need to do a bit more work. Start Python and issue the following commands to created a hashed password:

```
In [1]: import IPython
In [2]: IPython.lib.passwd()
Enter password:
Verify password:
Out[2]: 'sha1:db02b6ac4747:fc0561c714e52f9200a058b529376bc1c7cb7398'
```

Remember your password and copy the output string. Let's assume that we also want to mount a local directory `c:/tmp/docmnt` as a local directory `docmnt` inside the container. Jupyter has a user called `jovyan` and the working directory is `/home/jovyan/work`. The complete command for running Jupyter is then:

```
$ docker run -e GEN_CERT=yes -d -p 8888:8888 \
    -v /tmp/docmnt:/home/jovyan/work/docmnt \
     jupyter/scipy-notebook start-notebook.sh \
     --NotebookApp.password='sha1:.... value from above'
```

This command launches Jupyter via HTTPS with your new password. When the container is up, you can connect to it via HTTPS at your host's IP address and port 8888. Your browser may complain that this is not a valid web page, because you have created a self-signed certificate for this site. You can accept the risk, and you should see the page shown in figure 6.2.

Figure 6.2: View of Jupyter in the browser after accepting the security exception.

Creating your own Container

- Creating your own container image and storing it in the Docker Hub is simple.
- Suppose you have a Python application that opens a web server, waits for you to provide input, and then us es that input to pull up an image and displ ay it.
- Now, let's build this little server and its image data as a container.
- We use a Python application based on the Bottle framework for creatin g the web server.

The first time you run `docker build`, it downloads all the components for `jupyter/scipy-notebook`, Boto3, and Bottle, which takes a minute or so. After these components have been downloaded, they are cached on your local machine. Note that all the `pip installs` have been run and layered into the file system. Even the Python code was parsed to check for errors (see step 7). Because of this preinstallation, when the container is run, everything is there already.

```
Docker run -d -p 8000:8000 yourname/bottlesamp
```

Create a free Docker account and save your container to the Docker Hub as follows. You can then download your container to the cloud and run it there.

```
docker push yourname/bottlesamp
```

7. How can an organization determine the most suitable cloud deployment model for deploying their specific application, considering factors like performance requirements, data security, and scalability?

- A cloud deployment model can be described as a distinct parameterized configuration of the cloud computing environment.
- These parameters include storage size, ownership and accessibility.

- There are four main cloud deployment models: public, private, hybrid, and community. Web-based organization systems like the *inter-cloud* and *virtual private* are also two other emerging deployment models but yet to gain widespread popularity.
- Choosing the right cloud deployment model depends on the business, storage, networking and computing requirements of an organization.

Attribute	Public	Private	Community	Hybrid
Ease of setup and use	Easy	Requires proficiency in IT	Requires proficiency in IT	Requires proficiency in IT
Data privacy and security	Low	High	Relatively high	High
Data control	Little to none	High	Relatively high	Relatively high
Reliability	Vulnerable	High	Relatively high	High
Scalability and flexibility	High	High	Fixed capacity	High
Cost-effectiveness	The most cost-effective	Cost-intensive, the most expensive one	Cost is shared among the community members	Cheaper than a private model but more expensive than a public model
Need for in-house hardware	No	Depends	Depends	Depends
Upfront costs	Low	High	Medium	Medium

Attribute	Public	Private	Community	Hybrid
Ongoing costs	High	Low	Medium	Medium
Security	Low	High	Medium	Medium
Compliance	Low	High	Medium	Medium
Quality of service	Low	High	Medium	Medium
Integration	Low	High	Medium	Medium
Configurability	Low	Medium	Medium	Medium

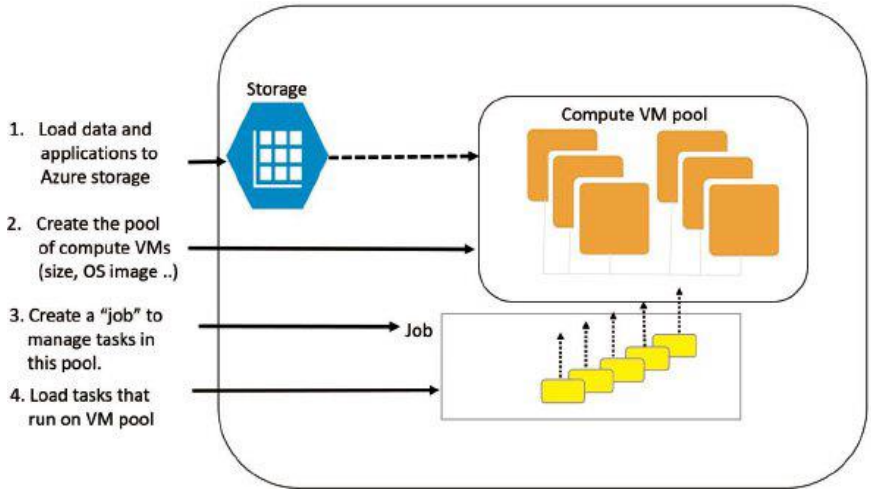
SASTRA
DEEMED TO BE UNIVERSITY

School of Computing
Second CIA Exam – Sep 2024
Course Code: CSE407R01
Course Name: CLOUD COMPUTING
Duration: 90 minutes Max Marks: 50

Answer any five questions **5*10=50M**

1. A) Explain how Deploying an HPC Cluster on Azure works (5M)
   B) List out the key Paradigms of Parallel Computing in the Cloud (5M)
2. Illustrate how to use Spark MLlib for predicting whether restaurants get pass or fail health inspections based on the free text of an inspector's comments with code.
3. Assume that you're working as a journal editor in ABC publisher. When scientists send technical papers to scientific journals, the abstracts of these papers often make their way onto the Internet as a stream of news items. A major source of high-quality streamed science data is ABC publisher, a collection of more than one million open-access documents. We have downloaded a small collection of records from ABC publisher, each containing a paper title, an abstract, and, for some, a scientific topic as determined by a curator. As a journal editor your goal is to build a cloud system that pulls document abstracts from the various feeds and then uses set of microservices to classify those abstracts into the major topics of physics, biology, math, finance, and computer science, and a second set to classify them into subtopic areas. Explain how Amazon EC2 Container Service works for the above scenario
4. Outline with a neat Architectural Overview of Azure Data Lake Analytics
5. Explain K-means Clustering using Spark Code snippet
6. Analyze the steps involved in RNN for predicting the next word to be typed when texting or in automatic translation systems

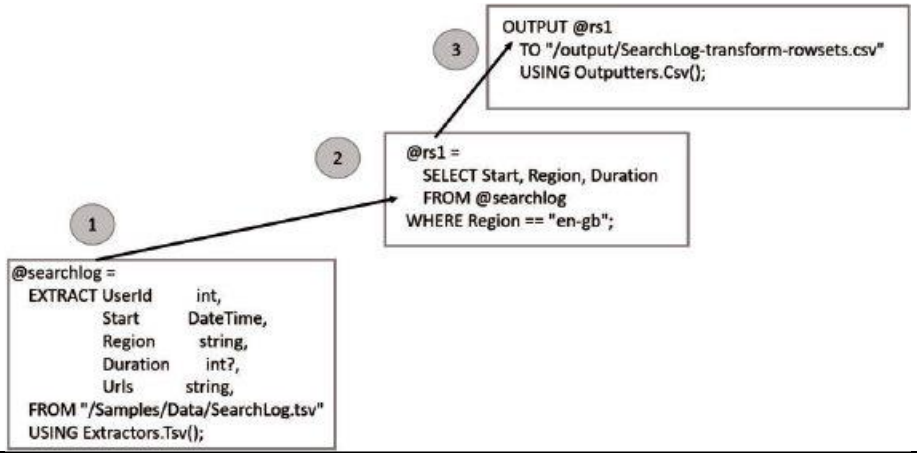
SASTRA DEEMED TO BE UNIVERSITY (U/S 3 of the UGC Act, 1956) THINK MERIT | THINK TRANSPARENCY | THINK SASTRA	**School of Computing** **Second CIA Exam – Sep 2024** Course Code: CSE407R01 Course Name: CLOUD COMPUTING Duration: 90 minutes Max Marks: 50 **Answer Key and Scheme of evaluation**

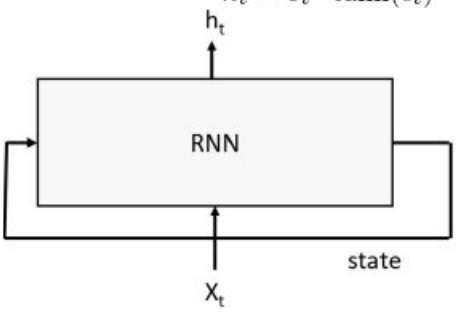
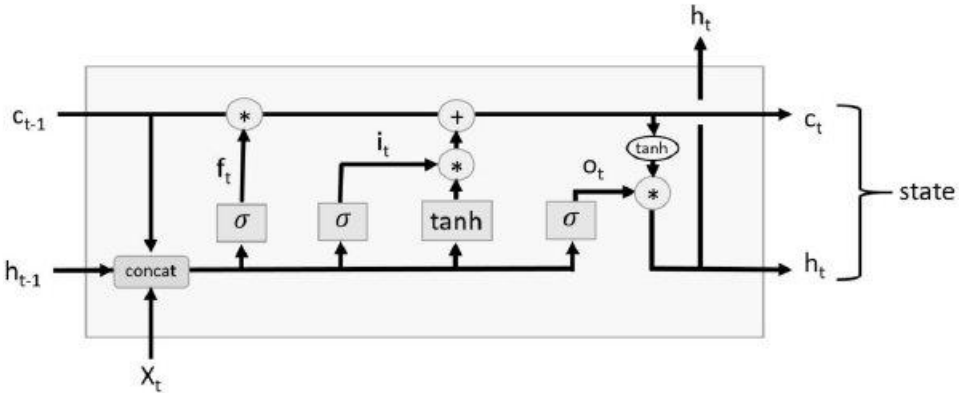
1	a)**Explain how Deploying an HPC Cluster on Azure (5M)** **Explaining the techniques with proper explanation 3M** **Diagram 2M**  **b) List out the key Paradigms of Parallel Computing in the Cloud (5M)** The first is highly **SYNCHRONOUS SINGLE PROGRAM MULTIPLE DATA (SPMD) COMPUTING.** **MANY TASK PARALLELISM**, in which a large queue or queues of tasks may be executed in any order, with the results of each task stored in a database or in files, or used to define a new task that is added to a queue. **BULK SYNCHRONOUS PARALLELISM (BSP)** BSP is based on processes or threads of execution that repeatedly perform independent computations, exchange data, and then synchronize at a barrier. MapReduce style made famous by Google and now widely used in its Hadoop and Spark realizations. The **GRAPH EXECUTION MODEL,** in which computation is represented by a directed, usually acyclic, graph of tasks. Execution begins at a source of the graph. Each node is scheduled for execution when all incoming edges to that node come from task nodes that have completed. Graphs can be constructed by hand or alternatively generated by a compiler from a more traditional-looking program that describes the graph either implicitly or explicitly

2	**Illustrate how to use Spark MLlib for predicting whether restaurants get pass or fail health inspections based on the free text of an inspector's comments with code.** **Library components explanation 5 M** **Code example explanation 5M** **DataFrames** are containers created from Spark RDDs to hold vectors and other structured types in a manner that permits efficient execution [45]. Spark DataFrames are similar to Pandas DataFrames and share some operations. They are distributed objects that are part of the execution graph. You can convert them to Pandas DataFrames to access them in Python. **Transformers** are operators that convert one DataFrame to another. Since they are nodes on the execution graph, they are not evaluated until the entire graph is executed. **Estimators** encapsulate ML and other algorithms. As we describe in the following, you can use the fit(...) method to pass a DataFrame and parameters to a learning algorithm to create a model. The model is now represented as a Transformer. A **Pipeline** (usually linear, but can be a directed acyclic graph) links Transformers and Estimators to specify an ML workflow. Pipelines inherit the fit(...) method from the contained estimator. Once the estimator is trained, the pipeline is a model and has a transform(...) method that can be used to push new cases through the pipeline to make predictions. Spark Mllib inspections = spark.sparkContext.textFile(\ 'wasb:///HdiSamples/HdiSamples/FoodInspectionD ata/ Food_Inspections1.csv').map(csvParse) inspections = spark.sparkContext.textFile('/path-to-reduceddata/ Food_Inspections1.csv').map(csvParse) schema = StructType([StructField("id", IntegerType(), False), StructField("name", StringType(), False), StructField("results", StringType(), False), StructField("violations", StringType(), True)]) df = spark.createDataFrame(inspections.map(\ lambda l: (int(l[0]), l[2], l[3], l[4])), schema) df.registerTempTable('CountResults') print("Passing = %d"%df[df.results == 'Pass'].count()) print("Failing = %d"%df[df.results == 'Fail'].count()) Passing = 61204

	Failing = 20225
	# 1) Define pipeline components
	# a) Tokenize 'violations' and place result in
	new column 'words'
	tokenizer = Tokenizer(inputCol="violations",
	outputCol="words")
	# b) Hash 'words' to create new column of
	'features'
	hashingTF = HashingTF(inputCol="words", outputCol
	="features")
	# c) Create instance of logistic regression
	lr = LogisticRegression(maxIter=10,
	regParam=0.01)
	# 2) Construct pipeline: tokenize, hash, logistic
	regression
	pipeline = Pipeline(stages=[tokenizer, hashingTF,
	lr])
	# 3) Run pipeline to create model
	model = pipeline.fit(labeledData)
	numSuccesses = predictionsDf.where(\
	"""(prediction = 0 AND results =
	'Fail') OR \
	(prediction = 1 AND (results =
	'Pass' OR \
	results = 'Pass w/
	Conditions'))""").count()
	numInspections = predictionsDf.count()
	print("There were %d inspections and there were
	%d predictions"\
	%(numInspections,numSuccesses))
	print("This is a %2.2f sucess rate"\
	%(float(numSuccesses) / float(numInspections) *
	100))

3	**Assume that you're working as a journal editor in ABC publisher. When scientists send technical papers to scientific journals, the abstracts of these papers often make their way onto the Internet as a stream of news items. A major source of high-quality streamed science data is ABC publisher, a collection of more than one million open-access documents. We have downloaded a small collection of records from ABC publisher, each containing a paper title, an abstract, and, for some, a scientific topic as determined by a curator. As a journal editor your goal is to build a cloud system that pulls document abstracts from the various feeds and then uses set of microservices to classify those abstracts into the major topics of physics, biology, math, finance, and computer science, and a second set to classify them into subtopic areas. Explain how Amazon EC2 Container Service works for the above scenario** Below concepts (6m) and Diagram (4m) Docker-based applications packaged as containers across a cluster of EC2 instances

	Amazon-hosted Docker image repositories **EC2 Container Service Role and Amazon Container Service Task Role.** • ***AmazonBigtableServiceFullAccess,*** • ***AmazonEC2ContainerServiceforEC2role*** • ***AmazonEC2ContainerServiceRole.*** **amazonsqsfullaccess and amazondynamodbfullacces**
4	Outline with a neat Architectural Overview of Azure Data Lake Analytics Explanation of U-SQL program **(6M) diagram (4m)** 
5	Explain K-means Clustering using Spark Code snippet **Kmeans steps (5M)** • Points to hold the k centroids • Initialize this array with random values • Apply an iterative MapReduce algorithm • Repeating these two steps until the centroids have not moved far from their previous position • For each point, find the INDEX OF THE CENTROID to which it is nearest and assign the point to the cluster associated with that nearest centroid. • For each cluster, COMPUTE THE CENTROID of all points in that cluster, and replace the previous centroid in kPoints with that new centroid. **Spark (Transformation and Actions Functions) (3m)** **Code (2m)**

6	**Analyze the steps involved in RNN for predicting the next word to be typed when texting or in automatic translation systems** **RNN steps with LSTM and formula (7m)** $$i_t = \sigma(W^{(xi)}x_t + W^{(hi)}h_{t-1} + W^{(ci)}c_{t-1} + b^{(i)})$$ $$f_t = \sigma(W^{(xf)}x_t + W^{(hf)}h_{t-1} + W^{(cf)}c_{t-1} + b^{(f)})$$ $$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W^{(xc)}x_t + W^{(hc)}h_{t-1} + b^{(c)})$$ $$o_t = \sigma(W^{(xo)}x_t + W^{(ho)}h_{t-1} + W^{(co)}c_t + b^{(o)})$$ $$h_t = o_t \cdot \tanh(c_t)$$   **Diagram 3M**

SASTRA

DEEMED TO BE UNIVERSITY

School of Computing
Third CIA Exam – Nov 2024
Course Code: CSE407R01
Course Name: CLOUD COMPUTING
Duration: 90 minutes Max Marks: 50

PART A

Answer any four of the following questions 4x10=40

1. The company is expanding the existing on-premises cluster, joining a computational grid with other related business firms, and moving the operations to a public cloud provider. It needs to scale its computing infrastructure to handle increasing data volumes and collaborate with customers worldwide. Based on the comparison of traditional computing models with cloud computing, which model would you recommend for the company's demand? Justify your answer by discussing its various features and explain how they relate to the company's requirements for scalability, resource management, interoperability, cost-effectiveness, data security, and privacy. (10)
2. (a) Analyze the common issue for scientists and engineers in data computation? Discuss the advanced computing services provided by various cloud vendors. (5)

   (b) Suppose you have a Python application that opens a web server, waits for you to provide input, and then uses that input to pull up an image and display it. Build this little server and its image data as a container in the Docker Hub. Write code snippets to create a Dockerfile, add images, and push the container to the cloud for process execution. (5)
3. Create a temperature dataset with the attributes of date, time, building ID, desired, and actual temperature. Each row represents a temperature reading at a specific date and time for a particular

building. Illustrate the steps to execute SQL operations and write a Spark SQL query to find the top three buildings with the largest average difference between desired and actual temperature for each month of the year 2024. (10)

4. (a) Examine the cloudlet mesh architecture model and illustrate the CMU Kimberley prototype system with the timeline for dynamic virtual machine synthesis in cloudlet. (5)

   (b) Assume that you have a computation-intensive task that cannot be carried out on your mobile phone due to limited processing capacity. Explain the optimal strategy to address this problem, which integrates remote cloud and mobile cloudlets to enhance flexibility and improve performance in energy efficiency and latency reduction. (5)

5. Explain the multicloud mashup service architecture and its specification layers. Demonstrate the mashup of multiple cloud services in healthcare applications. (10)

PART B

Answer the following **1x10=10**

6. (a) A startup company is developing a new telemedicine application. They want to focus on creating innovative features for remote patient monitoring and real-time consultations without worrying about infrastructure management. The company has a small team of developers with expertise in web and mobile application development, but they lack experience in server management and scalable architecture. Analyze this problem and illustrate the relevant cloud service model with its characteristics that assist in testing and deployment of cloud-based web applications. (5)

   (b) Explain the properties of a social network graph and illustrate an example of the Facebook platform's infrastructure. (5)

SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

School of Computing
Third CIA Exam – Nov 2024

Course Code: CSE407R01
Course Name: CLOUD COMPUTING
Duration: 90 minutes Max Marks: 50

PART A **4 x 10 = 40**

1. The company is expanding the existing on-premises cluster, joining a computational grid with other related business firms, and moving the operations to a public cloud provider. It needs to scale its computing infrastructure to handle increasing data volumes and collaborate with customers worldwide. Based on the comparison of traditional computing models with cloud computing, which model would you recommend for the company's demand? Justify your answer by discussing its various features and explain how they relate to the company's requirements for scalability, resource management, interoperability, cost-effectiveness, data security, and privacy. (10)

Characteristics	Clusters	Grids	Clouds
Ownership	Single ownership	Multiple ownership	Single ownership
Service pricing	Limited	Private or public assigned	Utility /large user discount
Virtualization	Half	Half	Yes
Resource management	Centralized resource	Distributed resource	Both
Scalable size	100s	1000s	100 to 1000s
Standardized	Yes	Yes	No
Interoperability	Yes	Yes	Not full
Speed / Interconnected network	Dedicated high end with low latency and high bandwidth	Mostly internet with high latency and low bandwidth	Dedicated high end with low latency and high bandwidth
Self-service	No	Yes	Yes
Single system image	Yes	No	Yes/optional included
Multi-tenancy	No	Yes	Yes
Service negotiation	Limited	Yes, SLA-based	Yes, SLA-based
Membership discovery	Membership service discovery	Decentralized information services and centralized indexing	Membership service discovery

Operating system	Windows/Linux	Any standard but dominated by Unix	Uses a hypervisor
Application drivers	Business, data centres, enterprise computing	Collaborative scientific and high-throughput applications	Web App. content delivery, dynamic provisioning
Standards/ interoperability	**Virtual Interface Architecture (VIA)**	Some open grid forum	Web services (SOAP and REST)
Scalable	No	Half	Yes
Failure management	Limited (often failed task/ application and restarted	Limited (often failed task/application restarted	Failover, content replication, virtual machine migration from one node to another supported
Capacity	Stable and guaranteed capacity	Varies, but high capacity	Provisioned on-demand capacity
Security	Traditional login/ password-based	Public/private pair -based authentication and mapping of a user to an account	Each user and/ or application is provided with a virtual machine
Privacy	Medium level of privacy depends on user privileges	Limited support for privacy	High security/ privacy is guaranteed. There is support for file **Access Control List** (ACL) settings.
Population	Commodity computers	High-end computing systems (including clusters and servers)	Commodity PCs, high-end servers' network, attached storage
End-user presentation	Presented as a dynamic and diversified system	Presented as a single system image	Presented as a self-services-based usage model

2. a) What is the common issue for scientists and engineers in data computation? Discuss the advanced computing services provided by various cloud vendors. (5)

A common issue of concern to scientists and engineers is **scale**. VMs and containers are a great way to virtualize a single machine image. Most high-performance parallel applications are based on the Message Passing Interface (**MPI**) standard. For example, many task (**MT**) parallelism is used to tackle problems in which you have hundreds of similar tasks to run, each(largely) independent of the others. Another method is called

MapReduce made popular by the Hadoop computational framework MapReduce is related to a style of parallel computing known as bulk synchronous parallelism (**BSP**). Google has released a service called **Cloud Datalab**, based on Jupyter, for interactive control of its data analytics cloud. The Microsoft Cloud Business Intelligence (**Cloud BI**) tool supports interactive access to data queries and visualization. **Docker** Swarm tools and **Google Kubernetes** container management and **HTCondor** are the various data analytics tools.

b) Suppose you have a Python application that opens a web server, waits for you to provide input, and then uses that input to pull up an image and display it. Build this little server and its image data as a container in the Docker Hub. Write code snippets to create a Dockerfile, add images, and push the container to the cloud for process execution. (5)

Python application based on the Bottle framework for creating the web server. Assume the images are all stored as jpg files in a directory called images.

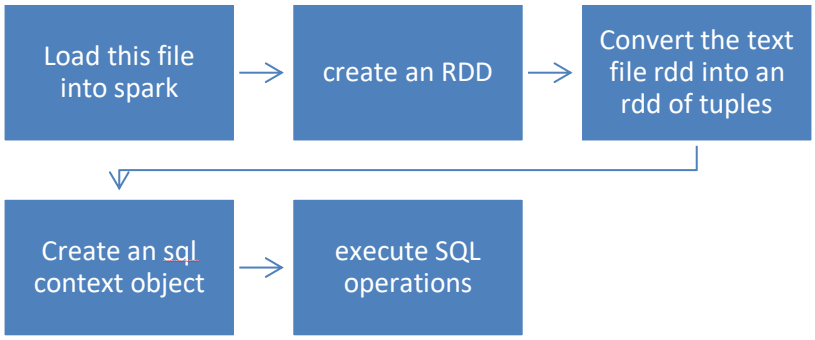
SciPy tools+Amazon Boto3 SDK+file named Dockerfile

```
FROM jupyter/scipy-notebook
MAINTAINER your name <yourname@gmail.com>
RUN pip install bottle
COPY images /images
COPY bottleserver.py /
ENTRYPOINT ["ipython", "/bottleserver.py"]
```

jupyter/scipynotebook	well-maintained container in the Docker Hub
pip install	for Boto3 and Bottle
ENTRYPOINT	Docker what to execute when the container Runs
docker build **Docker run -d -p 8000:8000 yourname/bottlesamp**	downloads all the components for jupyter/scipy-notebook, Boto3, and Bottle
docker push yourname/bottlesamp	Create a free Docker account and save your container to the Docker Hub as follows.

3. Create a temperature dataset with the attributes of date, time, building ID, desired, and actual temperature. Each row represents a temperature reading at a specific date and time for a particular building. Illustrate the steps to execute SQL operations and write a Spark SQL query to find the top three buildings with the largest average difference between desired and actual temperature for each month of the year 2024. (10)

Steps to execute SQL operations




```
from pyspark.sql.types import *
hvacText = sc.textFile("/pathto/file/hvac.csv")
hvac = hvacText.map(lambdas: s.split(",")) \
.filter(lambda s: s[0] != "Date") \
.map(lambda s:(str(s[0]), str(s[1]),
int(s[2]), int(s[3]),
str(s[4]) ))
sqlCtx = SQLContext(sc)
hvacSchema = StructType([StructField("date",
StringType(), False),
StructField("time", StringType(),
False),
StructField("targettemp",
IntegerType(), False),
StructField("actualtemp",
IntegerType(), False),
StructField("buildingID",
StringType(), False)])
hvacDF = sqlCtx.createDataFrame(hvac, hvacSchema)
```

```
%%sql_show
SELECT buildingID ,
       (targettemp - actualtemp) AS temp_diff,
       date FROM hvac
WHERE date = "3/23/2016"

+------------+---------+---------+
|  buildingID|temp_diff|     date|
+------------+---------+---------+
|headquarters|       13|3/23/2016|
|        lab1|      -10|3/23/2016|
|    coldroom|       34|3/23/2016|
+------------+---------+---------+
```

SQL query

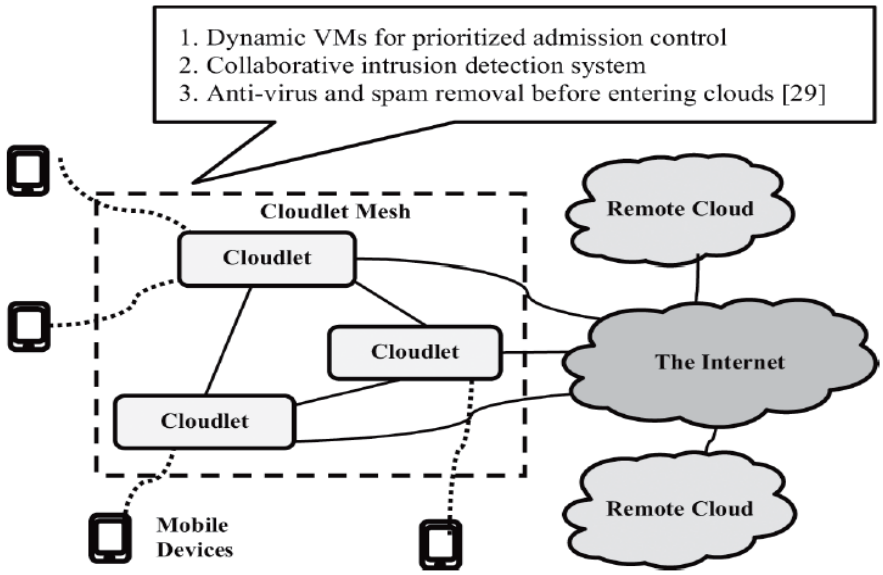
CREATE OR REPLACE TEMPORARY VIEW temperature_view AS

```
SELECT * FROM temperature_df;
```

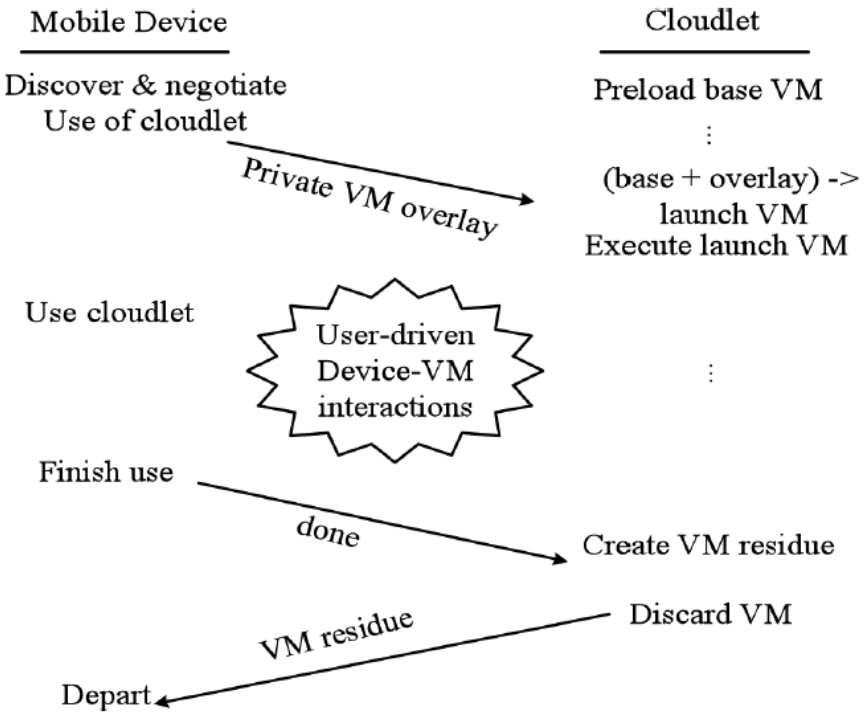
Main query

```
WITH monthly_avg_diff AS (
  SELECT
    building_id,
    MONTH(date) AS month,
    AVG(ABS(desired_temperature - actual_temperature)) AS avg_temp_diff
  FROM temperature_view
  WHERE YEAR(date) = 2024
  GROUP BY building_id, MONTH(date)),
ranked_buildings AS (
  SELECT
    building_id,
    month,
    avg_temp_diff,
    ROW_NUMBER() OVER (PARTITION BY month ORDER BY avg_temp_diff
DESC) AS rank
  FROM monthly_avg_diff)
SELECT
  building_id,
  month,
  avg_temp_diff
FROM ranked_buildings
WHERE rank <= 3
ORDER BY month, avg_temp_diff DESC;
```

4. a) Examine the cloudlet mesh architecture model and illustrate the CMU Kimberley prototype system with the timeline for dynamic virtual machine synthesis in cloudlet. (5)



Cloudlet mesh architecture



CMU Kimberley prototype system

b) Assume that you have a computation-intensive task that cannot be carried out on your mobile phone due to limited processing capacity. Explain the optimal strategy to address this problem, which integrates remote cloud and mobile cloudlets to enhance flexibility and improve performance in energy efficiency and latency reduction. (5)

Mobile Colocation Clouds

Colocation is shared space within a data centre, but the resources within racks and cages are not shared, they are dedicated to the business that owns them. Shared resources in a colocation facility include: Cooling systems and redundancies.

The cell phone itself is infeasible to finish some compute-intensive tasks. Instead, the data related to the computation task is offloaded to the remote cloud. Special cloudlets were introduced to serve as wireless gateways between mobile users and the internet. These cloudlets can be used to offload computations or web services to remote clouds safely. With the growing popularity of mobile devices, a new type of peer-to peer (P2P) mode for mobile cloud computing has been introduced. By applying short-range wireless networks, one can easily connect to nearby mobile devices using the cloudlets. Task scheduling scheme over colocated clouds (OSCC - Opportunistic Task Scheduling over Co-located Clouds). A basic feature of OSCC is that the contact between the task node and the service node can be either short or long instead of limiting users' mobility to

guarantee the contact time for task completion in conventional cloudlet based service mode.

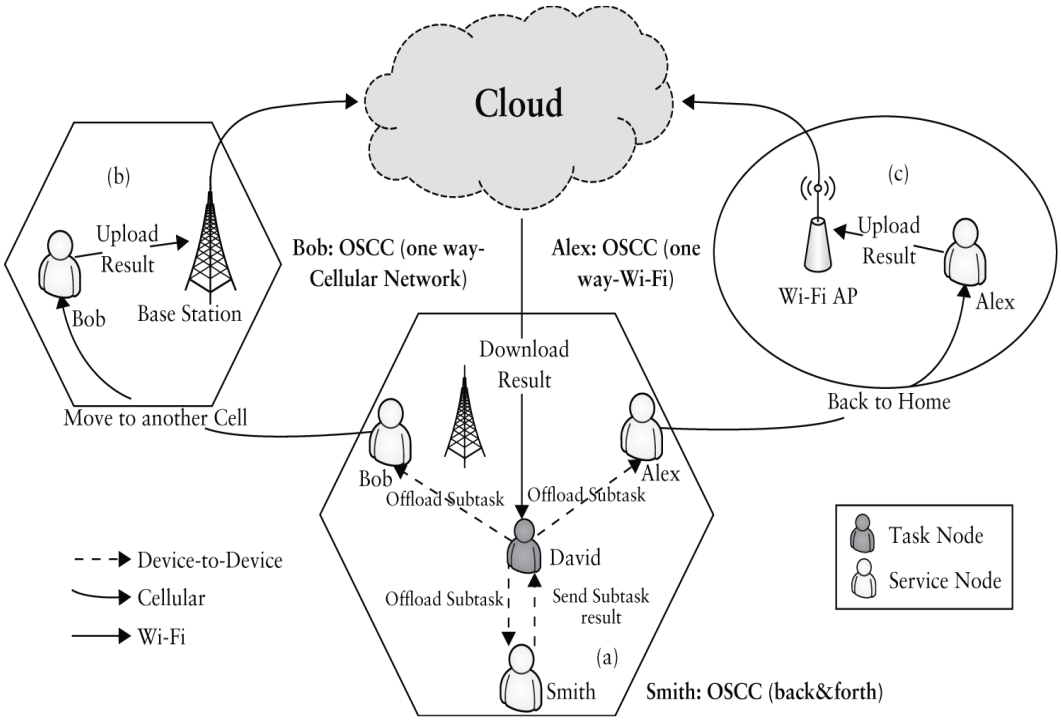


Figure: Integrates Remote Cloud and Mobile Cloudlets

5. Explain the multicloud mashup service architecture and its specification layers. Demonstrate the mashup of multiple cloud services in healthcare applications. (10)

Three types of web or cloud mashups, namely business mashups, consumer mashups, and data mashups. The architecture of a mashup is divided into three layers: presentation layer, web services and data layer. The below figure output of the workflow is the complete process up to the satisfaction of the patient. Each task could be one service provided by one or more web-/cloud-based platforms. Every candidate service is selected from a large service space supported by various cloud functions.

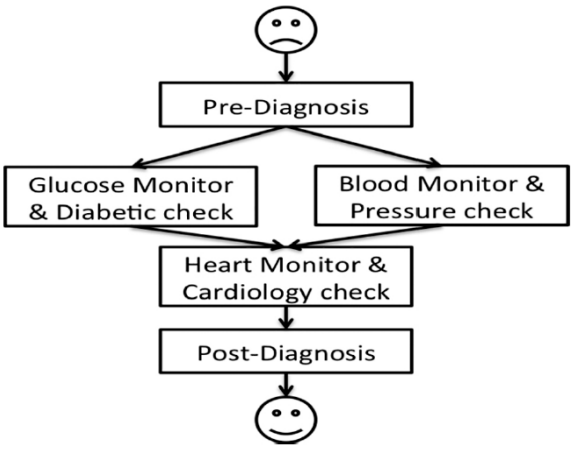


Figure 5.20
Workflow in a mashup of five cloud services for solving outpatient healthcare problems.

PART – B

6. a) A startup company is developing a new telemedicine application. They want to focus on creating innovative features for remote patient monitoring and real-time consultations without worrying about infrastructure management. The company has a small team of developers with expertise in web and mobile application development, but they lack experience in server management and scalable architecture. Analyze this problem and illustrate the relevant cloud service model with its characteristics that assist in testing and deployment of cloud-based web applications. (5)

Explanation on Platform as a Service (PaaS).

For this telemedicine startup, PaaS offers several key advantages:
Rapid Development: They can quickly build and iterate on their application using pre-configured environments and tools.
Easy Scaling: As their user base grows, the application can automatically scale to meet demand.
Focus on Innovation: By offloading infrastructure management, the team can concentrate on developing unique telemedicine features.
Cost-Effective: Reduces the need for specialized personnel to manage servers and infrastructure.
Built-in Compliance: Many PaaS providers offer healthcare-compliant environments, crucial for telemedicine applications.

Platform as a Service (PaaS)			
Definition	**Characteristics**	**Services**	**Service Providers**
• PaaS cloud computing platform is created for the programmer to develop, test, run, and manage the applications	• Accessible to various users via the same development application. • Integrates with web services and databases. • Builds on virtualization technology, so resources can easily be scaled up or down as per the organization's need. • Support multiple languages and frameworks. • Provides an ability to "AUTO-SCALE".	PaaS includes infrastructure (servers, storage, and networking) and platform (middleware, development tools, database management systems, business intelligence, and more) to support the web application life cycle	AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, Magento Commerce Cloud, and OpenShift.

b) Explain the properties of a social network graph and illustrate an example of the Facebook platform's infrastructure. (5)

Social network graph properties are as follows,

- Node degree, reach, path length, and betweenness
- Closeness and cohesion
- Centrality and centralization
- Social circles or clusters
- Centralized vs. decentralized networks
- Bridge and local bridge
- Prestige and radiality
- Structural cohesion, equivalence, and hole

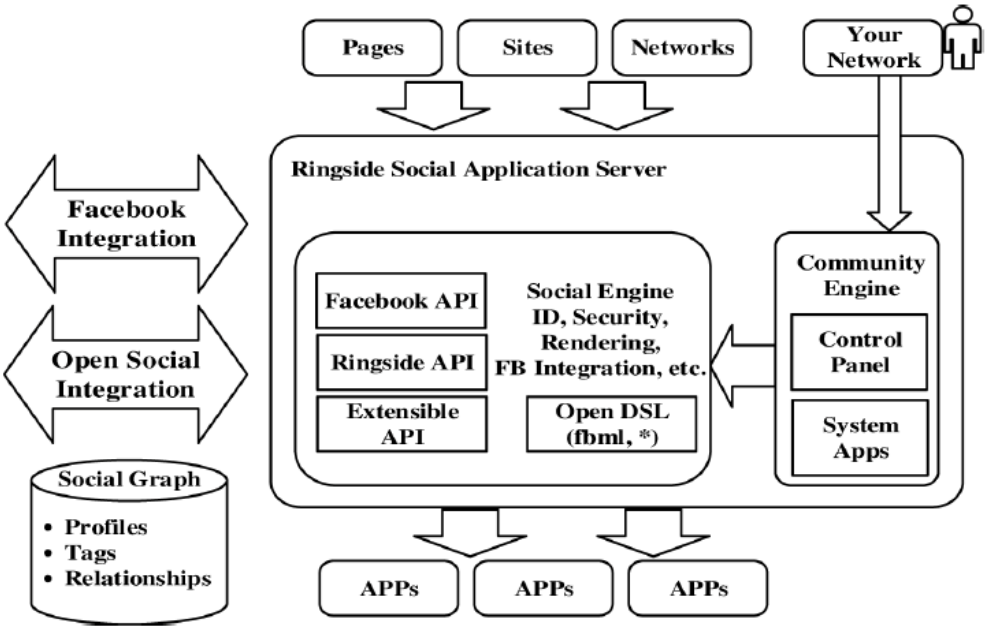


Figure: Infrastructure of the facebook platform

SASTRA
DEEMED TO BE UNIVERSITY

School of Computing
First CIA Exam – Aug 2024
Course Code: CSE409
Course Name: Parallel & Distributed Systems
Duration: 90 minutes Max Marks: 50

PART A

Answer ALL the Questions **5*10 = 50 marks**

1. Write CUDA kernel routines for matrix addition using both 2D grid with 2D blocks and 1D grid with 1D block. Analyse the performance of the approaches.
2. Elaborate the concept of CUDA memory models with required qualifiers and access scope.
3. (a)Explain the different access patterns that occur when access to shared memory is issued by a warp. (3)
   (b) Write a CUDA program to find the squares of each array element and multiply the result with the constant value 0.5(use constant memory). (7)
4. (a) Write the syntax and explain the functions used for (i) creating memory fence (ii) finding elapsed time (5)
5. (b) Draw and Illustrate the concept of false dependency and also explain how to reduce it in CUDA streams. (5)
6. Discuss the difference between pipelining and data parallelism using assemble widgets example and explain how speedup is achieved in both the methods.

****

SASTRA	School of Computing First CIA Exam – Aug 2024 Course Code: CSE409 Course Name: Parallel & Distributed Systems Duration: 90 minutes Max Marks: 50

Answer Key

1.The CUDA kernel for matrix addition using a 2D grid and 2D blocks can be written as follows: (4 marks)

```
__global__ void sumMatrixOnGPU2D(float *MatA, float *MatB, float *MatC, int nx, int ny) {
   unsigned int ix = threadIdx.x + blockIdx.x * blockDim.x;
   unsigned int iy = threadIdx.y + blockIdx.y * blockDim.y;
   unsigned int idx = iy * nx + ix;
   if (ix < nx && iy < ny)
      MatC[idx] = MatA[idx] + MatB[idx];
}
```

The CUDA kernel for matrix addition using a 1D grid and 1D blocks can be written as follows (4 marks)

```
__global__ void sumMatrixOnGPU1D(float *MatA, float *MatB, float *MatC, int nx, int ny) {
   unsigned int ix = threadIdx.x + blockIdx.x * blockDim.x;
   for (int iy = 0; iy < ny; iy++) {
      unsigned int idx = iy * nx + ix;
      if (ix < nx)
         MatC[idx] = MatA[idx] + MatB[idx];
   }
}
```

2D approach maximizes parallelism and often results in better performance due to coalesced memory access and optimal resource utilization. (2 marks)

2. CUDA memory model

- ➤ Registers
- ➤ Shared memory
- ➤ Local memory
- ➤ Constant memory
- ➤ Texture memory
- ➤ Global memory

3. (a) Three typical situations occur when a request to shared memory is issued by a warp:

- ➤ Parallel access: multiple addresses accessed across multiple banks
- ➤ Serial access: multiple addresses accessed within the same bank
- ➤ Broadcast access: a single address read in a single bank

(b)

```
#define N 1024
// Declare constant memory
__constant__ float constVal;
int main() {
   float h_in[N], h_out[N];
   float *d_in, *d_out;
   float h_constVal = 0.5f;
  // Initialize the input array
   for (int i = 0; i < N; i++) {
   // Allocate device memory
   cudaMalloc((void**)&d_in, N * sizeof(float));
   cudaMalloc((void**)&d_out, N * sizeof(float));

   // Copy input array from host to device
   cudaMemcpy(d_in, h_in, N * sizeof(float),
cudaMemcpyHostToDevice);
   // Copy constant value to constant memory
   cudaMemcpyToSymbol(constVal, &h_constVal, sizeof(float));
   // Define block size and grid size
   int blockSize = 256;
   int gridSize = (N + blockSize - 1) / blockSize;
   // Launch the kernel
   squareAndMultiply<<<gridSize, blockSize>>>(d_in, d_out);
   // Copy result from device to host
   cudaMemcpy(h_out, d_out, N * sizeof(float),
cudaMemcpyDeviceToHost);
  // Print first 10 results
   for (int i = 0; i < 10; i++) {
      std::cout << "h_out[" << i << "] = " << h_out[i] << std::endl;
   }
   // Free device memory
   cudaFree(d_in);
```

```
cudaFree(d_out);
    return 0; }
```

4(a) (i) void __threadfence_block();,void __threadfence(), void __thredfence_system()

(ii) cudaEventElapsedtime(ms,start,stop)

cudaEventRecord(…)

(b) False dependency ,HyperQ with diagram

5. difference between pipelining and data parallelism using assemble widgets (5 marks)

speedup is achieved by both methods.(5 marks)

SASTRA

DEEMED TO BE UNIVERSITY

School of Computing
Second CIA Exam – Sep 2024
Course Code: CSE409
Course Name: Parallel & Distributed Systems
Duration: 90 minutes Max Marks: 50

LTC-116/PPS-18

Answer any 5 questions (5 x 10 marks = 50 marks)

1. Demonstrate Enumeration sort and odd-even transposition sort to arrange the sequence: 5, 8, -3, 12, 4, 6, 4, -1, 10, 12.

2. Summarize and Illustrate the Hypercube SIMD algorithm to multiply the matrices A and B.
   A = {{2, 5}, {3, 6}} and B = {{1, 4}, {7, 9}}

3. Write the algorithm for merging two sorted lists and explain it with an example.

4a. Recall the steps involved in preorder tree traversal. (7)
4b. Discuss the staggering process with an example. (3)

5. How does the distributed system achieve a single system image? Discuss the factor and its types in detail.

6. Briefly describe the distributed system that can be used for high-performance computing tasks.

******

SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

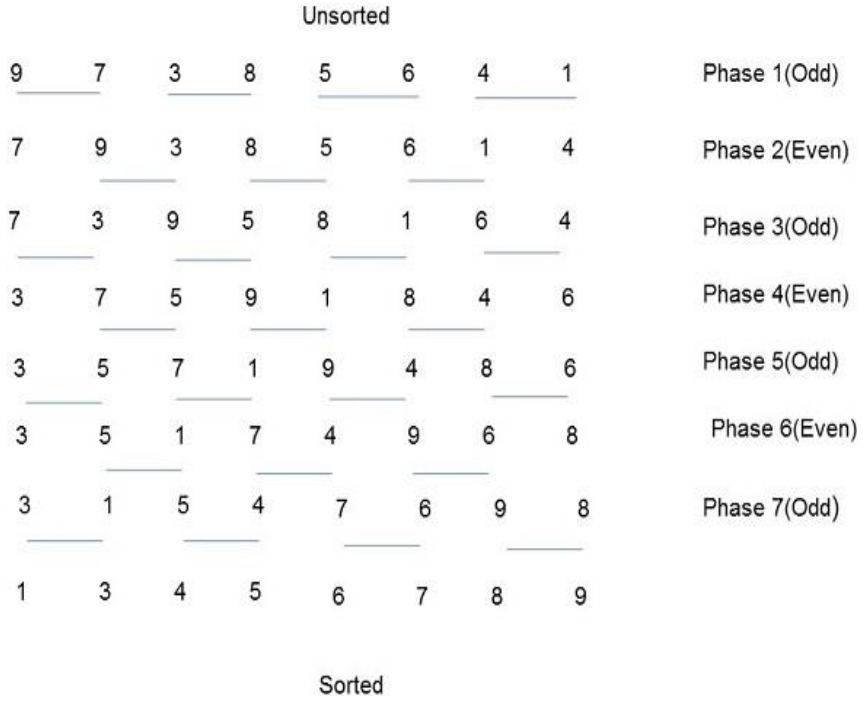
School of Computing
Second CIA Exam – Sep 2024
Course Code: CSE409
Course Name: Parallel & Distributed Systems
Duration: 90 minutes Max Marks: 50

Answer any 5 questions (5 x 10 marks = 50 marks)

1. Enumeration sort (5 marks)

It is a method of finding the exact position of each element in a sorted list by comparing and finding the frequency of elements having smaller values. If p elements are smaller than x, then x occupies the (p+1)th position in the sorted list.

Odd-even transposition sort (5 marks) – example



2. Hypercube SIMD Matrix Multiplication:
Algorithm – 5 marks
Illustration – 5 marks
Broadcast matrices to Hypercube nodes
Each node computes the product
Compute the sum of products – 3 marks

3. Merging Two Sorted Lists
Algorithm – 5 marks
Example – 5 marks

4a. Preorder tree traversal algorithm – 7 marks
4b. Stagger – 3 marks
Processors arranged in 2D Mesh
Move A matrix elements to east (one position leftward)
Move B matrix elements to south (one position upward)

5.Single System image
Factor – transparency
Types: Access, Location, Migration, Relocation, Replication, Concurrent, Failure

6.High-performance distributed system
- Cluster, Cloud, Grid computing

******

SASTRA
DEEMED TO BE UNIVERSITY

School of Computing
Third CIA Exam – Nov 2024
Course Code: CSE409
Course Name: Parallel & Distributed Systems
Duration: 90 minutes Max Marks: 50

PART A

Answer any four of the following questions **4x10=40**

1. Analyze how the hierarchy of GPU memory spaces impacts the performance of GPU computing. Provide a summary of the key characteristics of each memory type in a table format.

2. Explain the data-parallel Sieve of Eratosthenes approach, providing an example to illustrate its application.

3. Write the Hyperquicksort algorithm and compare the costs of different parallel sorting algorithms based on the number of processors utilized.

4. Explain the role of queue managers in message –queuing system and highlight the importance of message brokers.

5. Illustrate naïve flooding-based multicast communication and highlight its drawbacks. Explain how hypercube topology can enhance the efficiency.

6. Compare Vector Clocks with Lamport Timestamps in terms of their ability to capture causal relationships between events. Give an example that demonstrates the advantages of Vector Clocks over Lamport Timestamps.

PART B

Answer the following **1x10=10**

7. a. Explain 3PC protocol using a finite state machine diagram. (6)
   b. Illustrate the Ring algorithm for leader election in distributed systems, using an example with five processes (P0 to P5). Assume that both P2 and P5 initiate the election process. (4)

*****

SASTRA DEEMED TO BE UNIVERSITY (U/S 3 of the UGC Act, 1956) THINK MERIT | THINK TRANSPARENCY | THINK SASTRA	**School of Computing** **Third CIA Exam – Nov 2024** Course Code: CSE409 Course Name: Parallel & Distributed Systems Duration: 90 minutes Max Marks: 50

Answer Key

PART A

Answer any four of the following questions **4x10=40**

1.

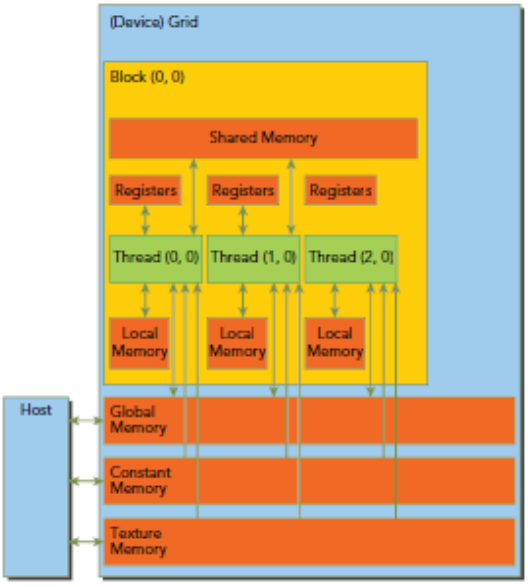


FIGURE 4-2

TABLE 4-2: Salient Features of Device Memory

MEMORY	ON/OFF CHIP	CACHED	ACCESS	SCOPE	LIFETIME
Register	On	n/a	R/W	1 thread	Thread
Local	Off	†	R/W	1 thread	Thread
Shared	On	n/a	R/W	All threads in block	Block
Global	Off	†	R/W	All threads + host	Host allocation
Constant	Off	Yes	R	All threads + host	Host allocation
Texture	Off	Yes	R	All threads + host	Host allocation

2. Explanation, Example (5+5)
   - Each processor will be responsible for a segment of the array representing the natural numbers.
   - All the processors perform the same operation (ie. Strikes off multiples of the same prime) on its own segment of data.

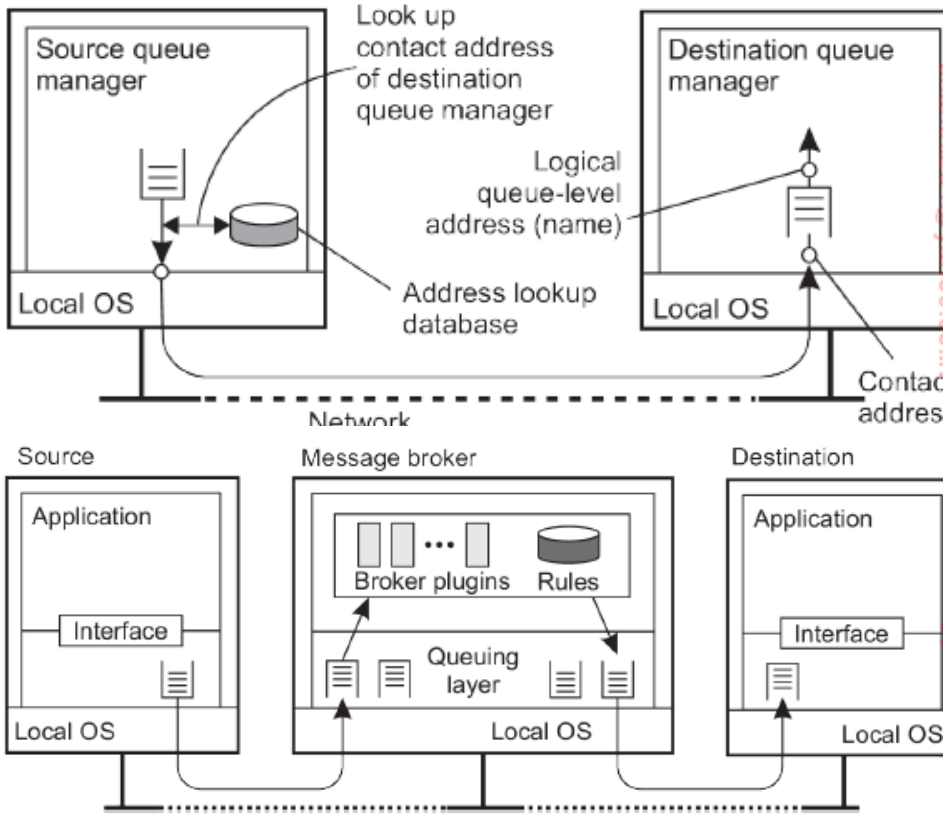
3. Algorithm (6)

```
HYPERQUICKSORT (HYPERCUBE MULTICOMPUTER):
  Global   n            (Initial number of elements per processor)
           d            (Dimension of hypercube)
           i            (Dimension number of current hypercube)
  Local    logical_num  (Unique processor number)
           partner      (Processor's partner in the exchange)
           root         (Root processor of current hypercube)
           splitter     (Median of root processor's sorted list)
begin
  for all P_j, where 0 ≤ j < 2^d do
    Sort n values using sequential quicksort algorithm
    if d > 0 then
      for i ← d downto 1 do
        root ← root of the binary i-cube containing processor logical_num
        if logical_num = root then
          splitter ← median of the sorted list held by processor logical_num
        endif
        Processor root broadcasts splitter to other processors in binary i-cube
        Use splitter to partition sorted values into low list, high list
        partner ← logical_num ⊗ 2^(i-1)     ( Bitwise exclusive "or" )
        if logical_num > partner then
          Send low list to processor partner
          Receive another high list from processor partner
        else (logical_num < partner)
          Send high list to processor partner
          Receive another low list from processor partner
        endif
        Merge two lists into a single sorted list of values
      endfor
    endif
  endfor
end
```

Processors (4)

Enumeration sort-n^2 , Odd-Even transposition sort-n, Quick sort- 0 to p, Hyperquicksort-2^d

4. Role of Queue manager(5)
   Importance of Message broker (5)

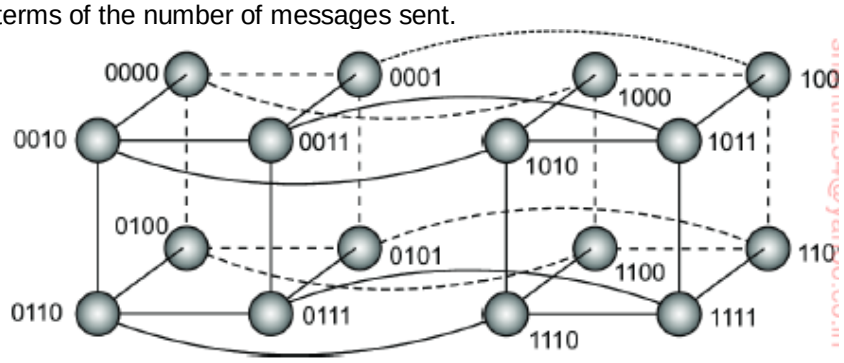


5.

Naive flooding: Each node simply forwards a message m to each of its neighbors, except to the one from which it received m. Furthermore, if a node keeps track of the messages it received and forwarded, it can simply ignore duplicates. We will roughly see twice as many messages being sent as there are links in the overlay network, making flooding quite inefficient. Only if G is a tree, will

flooding be optimal, for in that case, M = N - 1. In the worst case, when G is fully conneced, we will have to send out M = 1/2.N.(N -1).

Hybercube flooding: When a node receives a broadcast message, it will forward it only along edges that have a higher dimension. In other words, in our example, node 1101 will forward m only to nodes 1111 (joined to 1101 by an edge labeled “3”) and 1100 (joined by an edge with label “4”). Using this scheme, it can be shown that every broadcast requires precisely N - 1 messages, where $N = 2^n$, that is the number of nodes in a n-dimensional hypercube. This broadcasting scheme is therefore optimal in terms of the number of messages sent.



6. **Vector clock, Example** **(5+5)**

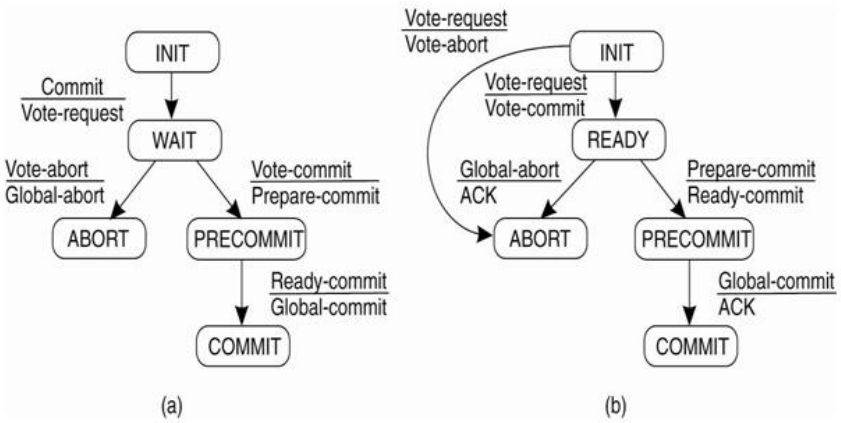
Causality can then be captured by means of **vector clocks**, which are constructed by letting each process P_i maintain a vector VC_i with the following two properties:

1. $VC_i[i]$ is the number of events that have occurred so far at P_i. In other words, $VC_i[i]$ is the local logical clock at process P_i.
2. If $VC_i[j] = k$ then P_i knows that k events have occurred at P_j. It is thus P_i’s knowledge of the local time at P_j.

PART B

Answer the following **1x10=10**

7. a. 3PC (6)



b. Ring algorithm illustration (4)

*****

SASTRA

School of Computing
First CIA Exam – Aug 2024
Course Code: INT404R01
Course Name: BIG DATA ANALYTICS
Duration: 90 minutes Max Marks: 50

Answer all FIVE questions. **5 X 10 = 50 Marks**

1. Describe the characteristics of big data and explain the layered architecture of big data technology stack and the interaction among the key components with a neat diagram.

2. Given the task of calculating bigram frequencies using the MapReduce algorithm, how would you design the Mapper function to process each document and emit bigrams? After the Mapper function emits bigrams and their counts, what role does the Reducer function play in the MapReduce algorithm? Suppose you have the following sample documents:
   a. Document 1: "Machine learning is a field of artificial intelligence."
   b. Document 2: "Artificial intelligence and machine learning are interrelated."
   c. Document 3: "Deep learning is a subset of machine learning."

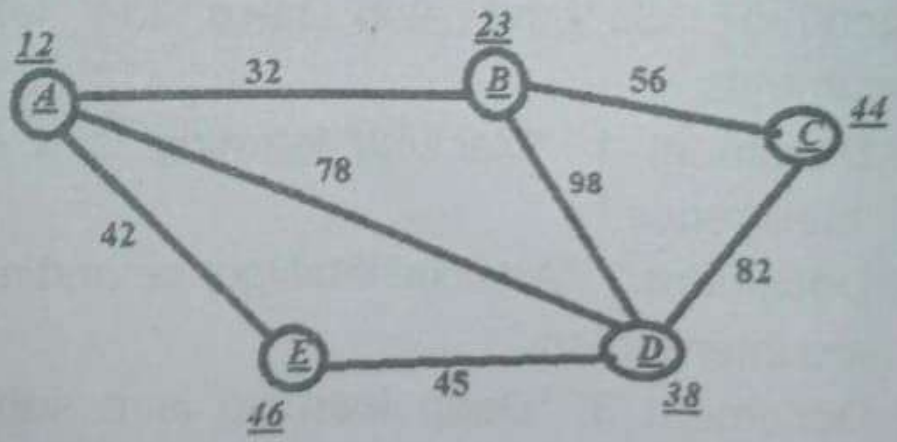
   What would be the output of the MapReduce job for these documents?

3. For the given set of m samples in S, where m = 10,
   S = {152, 173, 144, 182, 11, 22, 46, 85, 92, 210}
   with two hash functions, and a bit array of size 17 - bits
   i) $h(x) = (2x + 10) \% 17$ ii) $h(x) = 4x \% 17$

In the bit series, count the number of ones and find the error rate of false positives that the bloom filter produces.

4. Explain the following:
   i) Model of Data Stream management system 3 M
   ii) Execution of Relational algebra operations (Join and group) using map reduce algorithm. 7 M

5. The figure below shows the network connectivity between the systems. Here, the five separate systems are represented by nodes and are connected by a solid line. Construct the V matrix from the node value (i.e., A node value is 12) and the M matrix from the link value (i.e., A to B is 32 and B to A is 32). Perform a MapReduce based Matrix vector multiplication to determine the significance of each system and write an algorithm for the same.

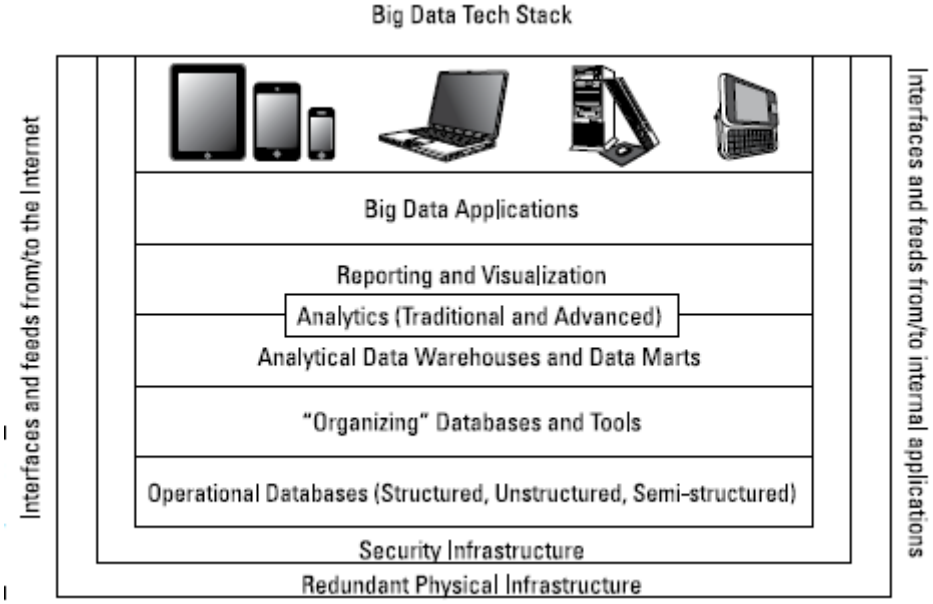


School of Computing
First CIA Exam – Aug 2024
Course Code: **INT404R01**
Course Name: **BIG DATA ANALYTICS**
Duration: **90** minutes Max Marks: **50**

ANSWER KEY

Answer all FIVE questions. **5 X 10 = 50 Marks**

1. Describe the characteristics of big data
(Velocity, Verity, Volume, Veracity). **(4 Marks)**
Layered architecture of big data technology: **(6 Marks)**



2. Mapper function (4 Marks)

```
def mapper(document_id, document_text):
    words = document_text.split()
    for i in range(len(words) - 1):
        bigram = (words[i], words[i + 1])
        emit(bigram, 1)
```

Reducer function (4 Marks)

```
def reducer (bigram, counts):
    # Sum up all counts for the given bigram
    total_count = sum(counts)
    # Emit the bigram with its total count
    emit(bigram, total_count)
```

Sample output (1 Mark)

Execution Step (1 Mark)

3. Demonstration of hash function 1: (3 Marks)

Demonstration of hash function 2: (3 Marks)

Number of 1's: (2 Marks)

False positive rate: (2 Marks)

h1(152) = 8
h2(152) = 13

h1(173) = 16
h2(173) = 12

h1(144) = 9
h2(144) = 15

h1(182) = 0
h2(182) = 14

h1(11) = 15 // collision
h2(11) = 10

h1(22) = 3
h2(22) = 3 // collision

h1(46) = 0 // collision
h2(46) = 14 // collision

h1(85) = 10 // collision
h2(85) = 0 // collision

h1(92) = 7
h2(92) = 11

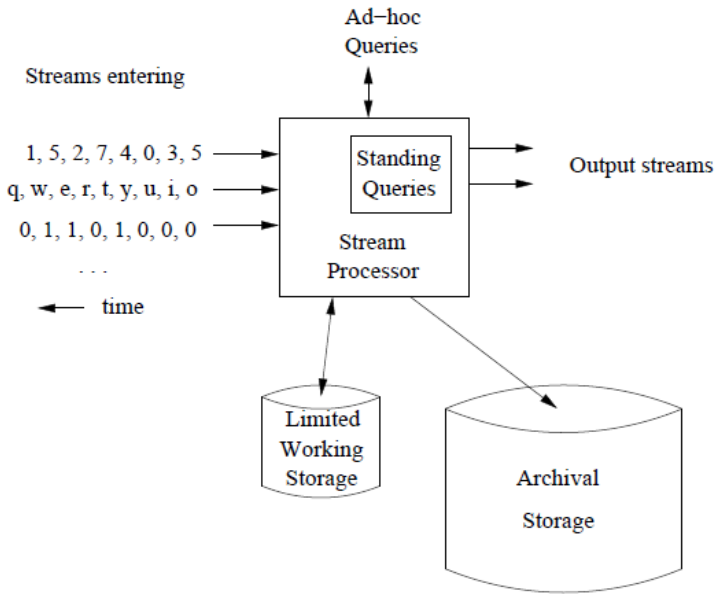
h1(210) = 5
h2(210) = 7 //collision

Number of ones = (number of elements * no. of hash func) - no. of collision
= (10 * 2) - 7
= 13

For k = 1,
Error rate of false positives = (1 - e^(-10/17)) = 0.4446

For k = 2,
Error rate of false positives = (1-e^(-2*10/17))^2 = 0.4783

4. A) Model of Data Stream management system (3 Marks)



B) Execution of Relational algebra operations (Join and group) using map reduce algorithm. (7 Marks)

Group:

Map: For a tuple *(a,b,c)* emit a key/value pair *(a, b)*

Reduce: Each key a represents *a* group, with values $[b_1, b_2, \ldots, b_n]$
Apply θ to the list $[b_1, b_2, \ldots, b_n]$
Emit the key/value pair (a,x), where $x = \theta([b_1, b_2, \ldots, b_n])$

Join:

Map: For a tuple *(a,b)* in R emit a key/value pair *(b, ('R',a))*
For a tuple *(b,c)* in S, emit a key/value pair *(b, ('S',c))*

Reduce: If key *b* has value list [('R',a),('S',c)], emit a key/value pair *(b, (a,b,c))*

5. Construction of the matrix **(2 Marks)**
Demonstration of Mapper task **(4 Marks)**
Demonstration of reducer task **(4 Marks)**

Matrix:

0	32	0	78	42	**5632**
32	0	56	98	0	**6572**
0	56	0	82	0	**4404**
78	98	82	0	45	**8868**
42	0	0	45	0	**2214**

SASTRA

DEEMED TO BE UNIVERSITY

School of Computing
Second CIA Test – September 2024
Course Code: INT404R01
Course Name: Big Data Analytics
Duration: 90 minutes
Max Marks: 50

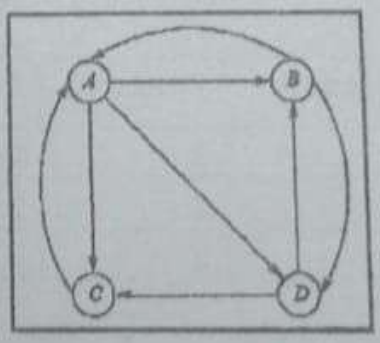
Answer any four questions PART -A 4 x 10 = 40 Marks

1. Using the Datar-Gionis-Indyk-Motwani algorithm, resolve the problem. Data streams: 1010110001011101100101 10, form the 5 window buckets by covering 1's. There should not exist more than two window sizes that are of the same size. Hence, write the basic principles of forming window sizes right from the data streams. [5]

   ii) Estimate the correctness of the value in DIGM, Data Streaming: 7 2 1 23 5 4, Make the window size as 4. Hence, find the corrected value and the estimated value. [5]

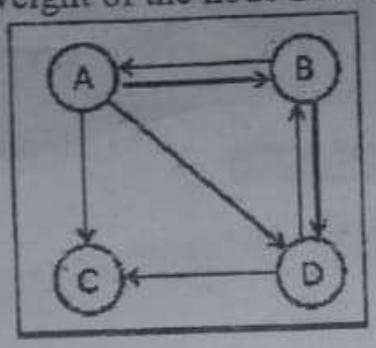
2. Flajolet-Martin algorithm with three hash functions to estimate the number of distinct user IDs in the stream. Stream: [102, 324, 512, 102, 876, 324, 512, 567, 876, 999, 567, 102]. hash1(x) = (x * 3 + 5) % 32, hash2(x) = (x * 7 + 11) % 32 and hash3(x) = (x * 5 + 13) % 32. [10]

3. Using the graph below, find the weight of the nodes by applying topic sensitive search algorithm. In the graph, treat B and D are sensitive nodes and find the weights accordingly. The damping factor is 0.80



[10]

4. Use the graph below and apply the avoiding dead end algorithm to find the weight of the nodes. If one outgoing link from node C to E, then find the weight of the node E. The damping factor is 0.80.



[10]

5. How can the AMS algorithm efficiently estimate the second moment of a large, continuous data stream in big data environments where storing and processing the entire dataset?

a, b, c, b, d, a, c, d, a, b, d, c, a, a, b
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
x x x

x is marked in the 3rd, 8th and 13th positions, respectively. Find the actual estimation (Second Moment) and AMS estimation, compare the results and provide an inference. [10]

Answer the following question PART-B 1 x 10=10

6. The below-mentioned utility matrix depicts how three users, U1, U2, and U3, graded five products, product 1 through product 5, on a 1–5 star scale. Calculate the following from the matrix's results.

(a) Compute the Jaccard and cosine distance between each pair of users using the utility matrix as a boolean value. [5]

(b) Treat 3, 4, and 5 as 1 and 1, 2, and a blank as 0. Calculate the distance between each pair of users using the Jaccard and cosine method. [5]

	Product 1	Product 2	Product 3	Product 4	Product 5
U1	5	3	4	4	
U2		5	1	3	5
U3	5	3	3		1

2024 September

CIA - II

INT 404 R01

Answer Key

Big data Analytics

1)

i) 1 0 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 1 1 0

2^2 2^2 2^1 2^1 2^0

Total 5 Buckets

→ Principals of forming Buckets. [5]

ii) 7 2 1 23 5 4

Make Window Size as 4

Size=4

7 → 0 0 1 1 1

2 → 0 0 0 1 0

1 → 0 0 0 1 1

23 → 1 0 1 1 1

5 ✗

4 ✗

C_4 C_3 C_2 C_1 C_0

1 0 2 3 3 ← ones in a cluster

Correct Value: 7 + 2 + 1 + 23 = 33 → ①

Estimation (E)

$E = 3 \times 2^0 + 3 \times 2^1 + 2 \times 2^2 + 0 \times 2^3 + 1 \times 2^4$

$E = 3 + 6 + 8 + 0 + 16$ [5]

$E = 33 \rightarrow ②$ Now ① = ② Hence proved

2) $h_1(x) = (x*3 + 5) \% 32$

$2^3 = 8$

$h_2(x) = (x*7 + 11) \% 32$

$2^1 = 2$

$h_3(x) = (x*5 + 13) \% 32$

$2^5 = 32$ [10]

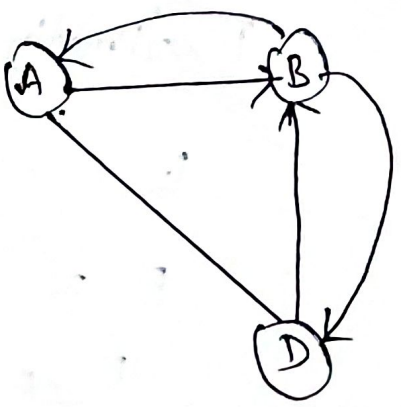
3) Topic Sensitive Search

Here B and D are topic sensitive

Damping factor is 0.80

Iteration 0	Iteration 1	Iteration 2	Iteration 3
0	1/4	7/32	31/128
1/2	3/8	15/32	63/128
0	1/4	7/32	31/128
1/2	3/8	15/32	63/128

4) $\begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 1/2 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix}$



Page Rank for c is $\frac{1}{3} \times \frac{2}{9} + \frac{1}{2} \times \frac{3}{9} = 13/54$

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \begin{bmatrix} \frac{1}{6} \\ \frac{3}{6} \\ \frac{2}{6} \end{bmatrix}, \begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix}, \begin{bmatrix} 5/24 \\ 11/24 \\ 8/24 \end{bmatrix} \cdots \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

[10]

5) Second Moment F_2

$$F_2 = 5^2 + 4^2 + 3^2 + 3^2$$

a occured 5 times c occured 3 times
b occured 4 times d occured 3 times

$$F_2 = 59$$ [3]

Estimation

$c = 3, d = 2, a = 2$

$E = N * (2 * \text{Value} - 1)$

For a : $15 * (2 \times 2 - 1) = 45$
For c : $15 * (2 \times 3 - 1) = 75$
For d : $15 * (2 \times 2 - 1) = 45$

$$\text{A.M} = \frac{45 + 75 + 45}{3} = 55$$ [7]

55 is close to 59

6)

Pair	Jaccard (a)	Cosine (a)	Jaccard (b)	Cosine (b)
u_1, u_2	0.4	0.25	0.6	0.423
u_1, u_3	0.4	0.25	0.4	0.134
u_2, u_3	0.4	0.25	0.6	0.334

i) ii)

[5] + [5] = [10]

SASTRA

School of Computing
THIRD CIA Exam – NOV 2024
Course Code: INT404R01
Course Name: BIG DATA ANALYTICS
Duration: 90 minutes
Max Marks: 50

Answer any Four questions PART A 4* 10 marks =50 marks

1. a) Multiply the given matrices using Map Reduce algorithm (6)

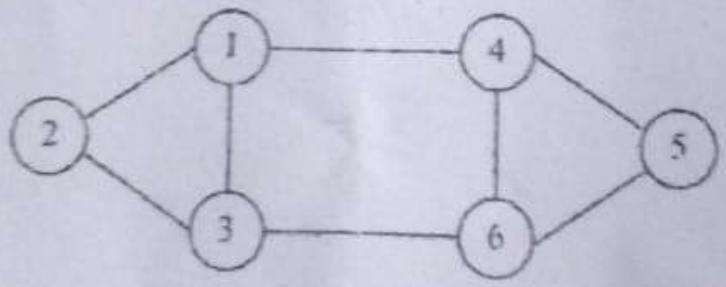
$$\begin{bmatrix} 10 & & 20 \\ & 30 & 40 \\ 50 & 60 & 30 \end{bmatrix} \quad \begin{bmatrix} -1 & \\ -2 & -3 \\ & -2 \end{bmatrix}$$

b) Describe Hadoop ecosystem with a suitable diagram. (4)

2.a) Find all possible (Friend, Friend) tuples that exist from the following dataset using MapReduce model. (5)

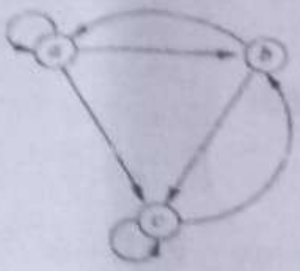
A->B, C, D
B->A, C, D
C->A, B, D, E
D->A , B, C, E
E->B, C, D

b) Find the Laplacian matrix for the graph (5)



3.Construct a graph with the vertices {1,2,3,4,5,6,7} and Edges {(1,2), (1,3), (1,4), (2,3), (2,5), (3,6), (4,5), (4,6), (5,7), (6,7)}. Find the maximum likelihood for generating the given graph.

4. Compute page rank for each page for 3 iterations, assume $\beta = 0.8$



5. Give the commands for the following using HIVE

1.Create Database called customer_*order_db* (if it doesnot exist)

2.Create customers table that stores customer details (*customer_id, customer_name, city*).

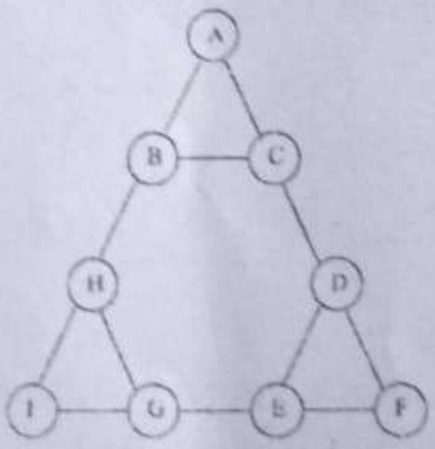
3.Create orders table that stores order information (*order_id, customer_id, order_date, amount*).

4.Insert sample data into customers and orders table.

5.Find the total amount spent by each customer using group by command

Compulsory Question **Part B** **1*10 marks =10 marks**

6. Use Girvan-Newman approach to find the number of shortest paths from each of the following nodes that pass through each of the edges. Using betweenness values determine reasonable candidates for communities

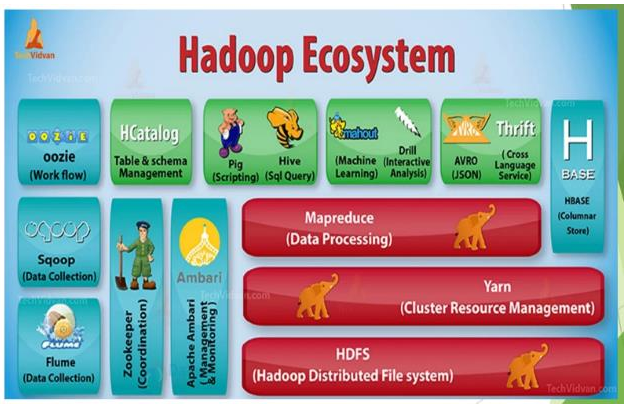


THIRD CIA Exam – NOV 2024
Course Code: INT404R01
Course Name: BIG DATA ANALYTICS
Answer Key

1. a) Map Reduce functions to perform matrix-vector multiplication process is as follows:
demonstration of mapper task – carries 5 marks
demonstration of reducer task – carries 5 marks

$$A = \begin{bmatrix} 10 & & 20 \\ & 30 & 40 \\ 50 & 60 & 30 \end{bmatrix} \quad B = \begin{bmatrix} -1 & \\ -2 & -3 \\ & -2 \end{bmatrix} \quad resultant\ matrix = \begin{bmatrix} -10 & -40 \\ -60 & -170 \\ -170 & -240 \end{bmatrix}$$

b) Description about each components



2. a) Finds of Friends key-value pairs

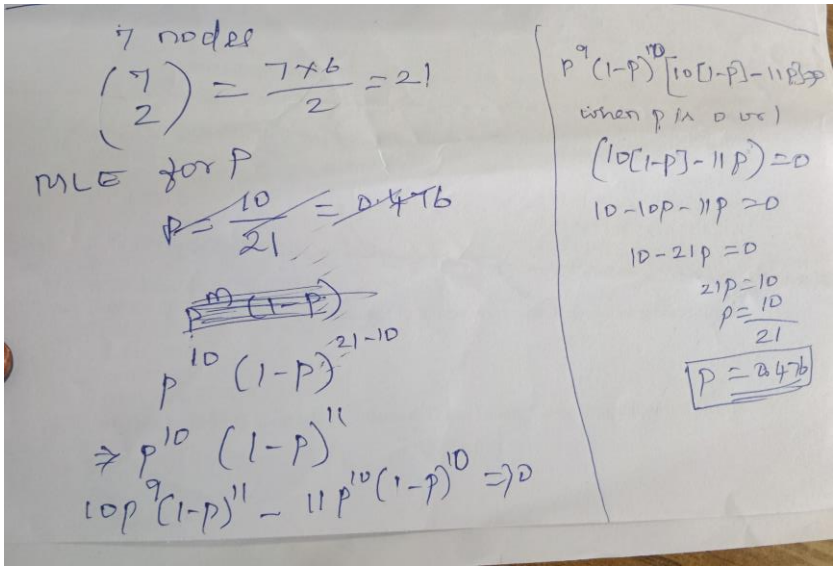
After shuffle and sort	Results after reduction
(A B) -> (A C D E) (B C D) (A C) -> (A B D E) (B C D) (A D) -> (A B C E) (B C D) (B C) -> (A B D E) (A C D E) (B D) -> (A B C E) (A C D E) (B E) -> (A C D E) (B C D) (C D) -> (A B C E) (A B D E) (C E) -> (A B D E) (B C D) (D E) -> (A B C E) (B C D)	(A B) -> (C D) (A C) -> (B D) (A D) -> (B C) (B C) -> (A D E) (B D) -> (A C E) (B E) -> (C D) (C D) -> (A B E) (C E) -> (B D) (D E) -> (B C)

2. b) Laplacian matrix for the given graph

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

$$L = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 & -1 \\ -1 & 0 & 0 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 3 \end{bmatrix}$$

3. Construction of graphs from the given nodes (2 marks)



4. $M = \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix}$ $v = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$

- Construction of transition matrix (2)
- Spider trap issue at node c (2)
- Three iteration (6)

$$I1 = \begin{bmatrix} 13/45 \\ 13/45 \\ 19/45 \end{bmatrix} \qquad I2 = \begin{bmatrix} 350/1350 \\ 422/1350 \\ 578/1350 \end{bmatrix} \qquad I3 = \begin{bmatrix} 10564/40500 \\ 12436/40500 \\ 17500/40500 \end{bmatrix}$$

5. CREATE DATABASE IF NOT EXISTS customer_order_db;
USE customer_order_db; (2)
CREATE TABLE IF NOT EXISTS customers (customer_id INT, customer_name STRING, city STRING) (2)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'

CREATE TABLE IF NOT EXISTS orders (order_id INT, customer_id INT, order_date STRING, amount FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' (2)
Inserting data carries two mark (2)

```
SELECT c.customer_id, c.customer_name, SUM(o.amount) AS total_amount_spent
FROM customers c
JOIN orders o ON c.customer_id = o.customer_id GROUP BY c.customer_id,
c.customer_name;
```
(2)

6. Root node A value 9
For threshold 3 = three communities
For threshold 4 = three communities

SASTRA
DEEMED TO BE UNIVERSITY

School of Computing
First CIA Exam – Aug 2024
Course Code: CSE415R02
Course Name: Internet of Things
Duration: 90 minutes Max Marks: 50

PART A

Answer all the questions **10*2 = 20 marks**

1. Why IoT middleware is required for developing an IoT system?
2. Narrate the different type of key enabling technologies for short range communications in IoT for a ubiquitous connectivity.
3. Discuss any two key differences between IoT and M2M.
4. Draw the diagram that depicts the hierarchy of needs of a thing in various Tiers based on Maslow's Hierarchy of Needs in Psychology.
5. Explain in what way Multi-channel MAC protocols are better than single-channel MAC protocols.
6. Mention any two OS for WSN.
7. Narrate the four messaging modes in CoAP.
8. Why LoRA is considered as the best option for environment monitoring and smart metering applications?
9. How IoT based smart road system can help in making the roads safer?
10. How CBR helps in the fields where it is hard to establish a quantitative mathematical model?

PART B

Answer all the questions **3*10 = 30 marks**

11. Illustrate the vision of IoT from four different perspectives with neat diagram.
12. Explain the major building blocks of M2M system with neat diagram.
13. a. Elaborate the architectural design of WSN. (5)

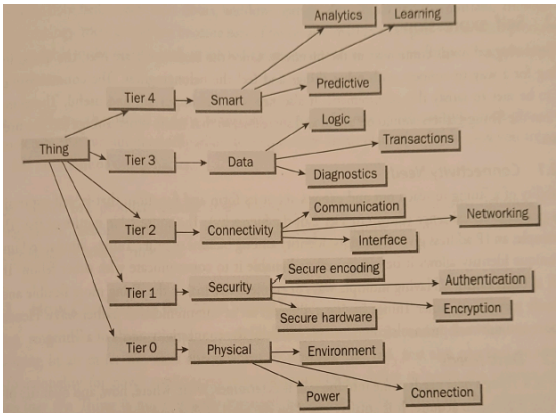
    b. Develop a use case scenario for air pollution monitoring; which shall describes the problem statement and solution using IoT sensors, communication technologies, storage and analysis. (5)

SASTRA DEEMED TO BE UNIVERSITY (U/S 3 of the UGC Act, 1956) THINK MERIT | THINK TRANSPARENCY | THINK SASTRA	**School of Computing** **First CIA Exam – Aug 2024** Course Code: CSE415R02 Course Name: Internet of Things Duration: 90 minutes Max Marks: 50

Answer Key

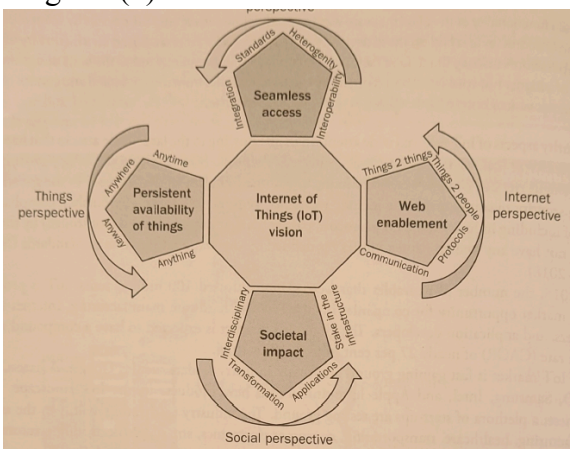
1. Collection, aggregation, filtering, and processing of data from heterogeneous devices over a variety of protocols and network toplogies
2. Wi-Fi, Z-Wave, Zigbee, IrDA, Bluetooth, BLE, RFID
3.

Machine-to-Machine (M2M)	Internet of Things (IoT)
M2M uses point-to-point connections between machines/devices of similar types using wireless or wired connections	IoT is based on IP network connections
M2M is all about machines	IoT is about Things
Older protocols and communication techniques	Varying new protocols specifically developed for IoT protocols such as 6LoWPAN for lower-powered IoT devices
Mainly hardware-based, i.e., the M2M technology is integrated into the hardware	IoT is both hardware and software based. Emphasis is on integration of a variety of Things
M2M has less built-in intelligence	More intelligent form of machine communications, cloud, context, collaboration, and insights
Ability for Integration and interoperability are limited due to lack of well-recognized standards and more proprietary systems are in vogue. Data is usually not shared	High levels of integration is possible. Plethora of standards has emerged for low-powered IoT devices. Data sharing from multiple assets is highly recommended for developing cross-domain applications
Business-to-Business (B2B)	Business-to-Business (B2B), Business-to-Consumer (B2C)
In majority of M2M systems, Internet connection is not required	Internet connection is required for majority of IoT systems
Limited scalability	Potential for high scalability due to Internet enablement and cloud services

4. 

5. Uses several orthogonal channels and divides the bandwidth into multiple channels. Improves network throughput and reduces collision

6. TinyOS, LiteOS, Contiki, RIOT, PreonVM
7. Confirmable, non-confirmable, piggyback, seperate
8. Long range, low power consumption
9. Provide information on driving conditions, travel time estimates, and alerts in case of poor driving conditions, traffic congestions and accidents
10. Based on past experience. Past experience is organized and represented as cases
11. Things, standards and semantics, internet, social (each 2)
    Diagram (2)



12. M2M device, M2M area network layer, M2M gateway, M2M communication network, service layer, applications (each 1.5)

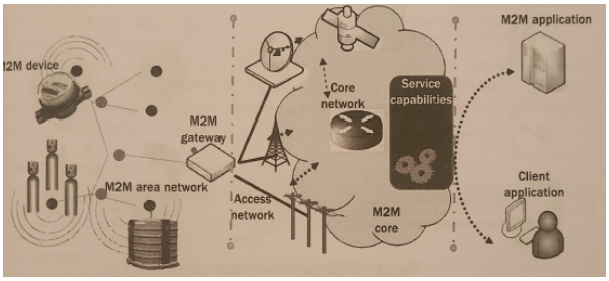
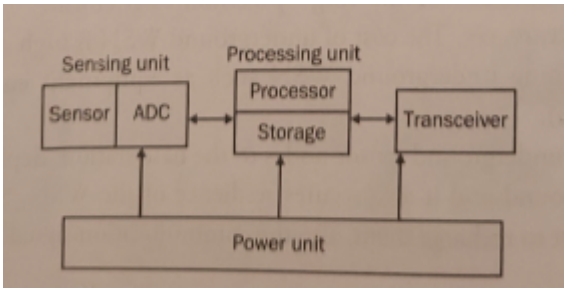
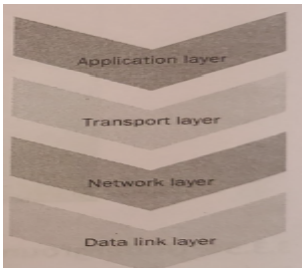


    Diagram (1 mark)
13. A. (5 marks)





    b. . harmful gases such as Co2, CO, etc. gaseous and meteorlogical sensors. Distributed monitoring stations. Communication via M2M. backend server (5 marks)

SASTRA
DEEMED TO BE UNIVERSITY
THANJAVUR KUMBAKONAM CHENNAI

School of Computing
Second CIA Test – September - 2024
Course Code: CSE415R02
Course Name: INTERNET OF THINGS
Duration: 90 minutes Max Marks: 50

PART - A 10 x 2 = 20 Marks

Answer all the questions

1. Illustrate the key elements of SDN with its functionalities.
2. Distinguish between SDN and NFV.
3. Whether SNMP is suitable for IoT systems? Justify your answer.
4. State the role of Rollback manager and Data model manager in NETCONF-YANG approach.
5. How do data collection and analysis approach differ in M2M and IoT?
6. Classify the relays based on the principle of operation and infer the role of relay as an actuator.
7. Organize the modules of IoT Device framework and state its functions.
8. Middleware is one of the driving force of IoT systems. Defend your answer.
9. Outline the functions of Lysimeter.
10. Summarize the Non-functional requirements of Middleware.

PART - B 3 x 10 = 30 Marks

Answer any 3 questions

1. Propose a Case study on design of an IoT system for Landslide monitoring systems based on IoT Design methodology.

2. Consider HAProxy load balancer for a commercial website. Build YANG modules and examine its visual representation.

3. Categorize the middleware based on their architecture.

4. a. Elaborate the functionalities of medical sensors. (5)
   b. State the working principle of Touch sensors with its types. (5)

* ALL THE BEST *

SASTRA DEEMED TO BE UNIVERSITY (U/S 3 of the UGC Act, 1956) THINK MERIT | THINK TRANSPARENCY | THINK SASTRA	**School of Computing** **Second CIA Exam – Sep 2024** Course Code: CSE415R02 Course Name: Internet of Things Duration: 90 minutes Max Marks: 50

Answer Key

1. Centralized Network Controller, Programmable Open API, Standard communication Interface (Openflow)

- **Centralized Network Controller**: With decoupled control and data planes and centralized network controller, the network administrators can rapidly configure the network. SDN applications can be deployed through programmable open APIs. This speeds up innovation as the network administrators no longer need to wait for the device vendors to embed new features in their proprietary hardware.
- **Programmable Open APIs**: SDN architecture supports programmable open APIs for interface between the SDN application and control layers (Northbound interface). With these open APIs various network services can be implemented, such as routing, quality of service (QoS), access control, etc.
- **Standard Communication Interface (OpenFlow)**: SDN architecture uses a standard communication interface between the control and infrastructure layers (Southbound interface). OpenFlow, which is defined by the Open Networking Foundation (ONF) is the broadly accepted SDN protocol for the Southbound interface. With OpenFlow, the forwarding plane of the network devices can be directly accessed and manipulated.

2. Differences between SDN and NFV:

S.No.	SDN	NFV
1	Focuses on Traffic flow management and control.	Focus on Virtualizing and Management of network functions
2	SDN abstracts the network infrastructure from the control plane.	NFV abstracts network functions from the underlying hardware.

3. **Limitations of SNMP:**

- SNMP was designed to provide a simple management interface between the management applications and the managed devices. SNMP is stateless in nature and each SNMP request contains all the information to process the request. The application needs to be intelligent to manage the device. For a sequence of SNMP interactions, the application needs to maintain state and also to be smart enough to roll back the device into a consistent state in case of errors or failures in configuration.
- SNMP is a connectionless protocol which uses UDP as the transport protocol, making it unreliable as there was no support for acknowledgement of requests.
- MIBs often lack writable objects without which device configuration is not possible using SNMP. With the absence of writable objects, SNMP can be used only for device monitoring and status polling.
- It is difficult to differentiate between configuration and state data in MIBs.
- Retrieving the current configuration from a device can be difficult with SNMP. SNMP does not support easy retrieval and playback of configurations.
- Earlier versions of SNMP did not have strong security features making the management information vulnerable to network intruders. Though security features were added in the later versions of SNMP, it increased the complexity a lot.

4.

- **Rollback Manager** : Rollback manager is responsible for generating all the transactions necessary to rollback a current configuration to its original state.
- **Data Model Manager**: The Data Model manager keeps track of all the YANG data models and the corresponding managed objects. The Data Model manager also keeps track of the applications which provide data for each part of a data model.

5.

- **Data Collection & Analysis:** M2M data is collected in point solutions and often in on-premises storage infrastructure. In contrast to M2M, the data in IoT is collected in the cloud (can be public, private or hybrid cloud). Figure 3.5 shows the various IoT-levels, and the IoT components deployed in the cloud. The analytics component analyzes the data and stores the results in the cloud database. The IoT data and analysis results are visualized with the cloud-based applications. The centralized controller is aware of the status of all the end nodes and sends control commands to the nodes. Observer nodes can process information and use it for various applications, however, observer nodes do not perform any control functions.

6. Types of Relay: (Based on the Principle of Operation)

1. Solid state relay
2. Hybrid relay
3. Electrothermal relay
4. Electromechanical relay

Role of relay as an actuator:

The relay device receives input signals from one side and controls the switching operations on the other side. When the relay is on, the armature moves and connects the normally opened contact pin

7. IoT Device Framework:

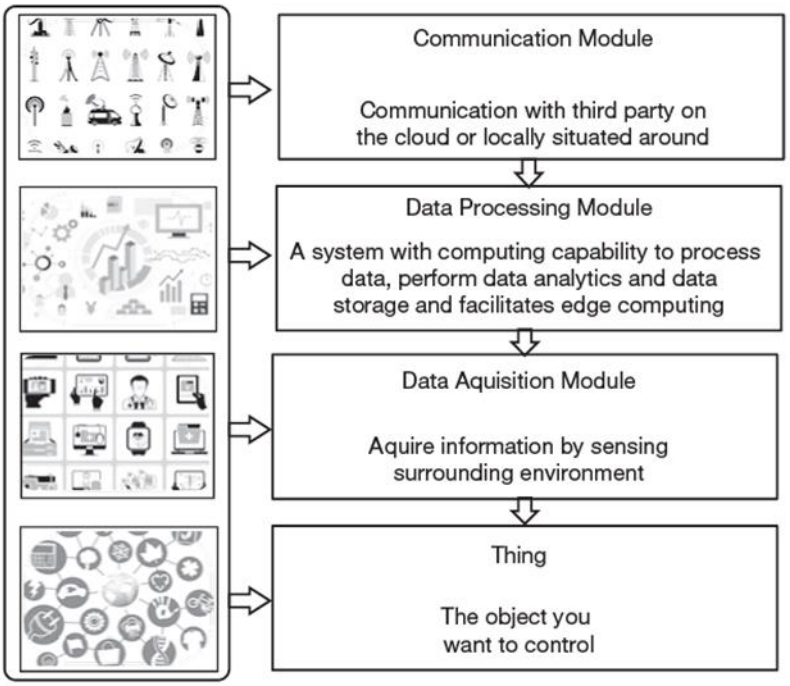


Fig. 6.2 IoT Device Framework

8. (i) It provides a layer of integrating software for performing all the required tasks rather than having a variety of tools from different vendors.

(ii) Further, the major needs for middleware in IoT are Device management, Data analytics, Cloud services and Security.

9. **Lysimeter:**

(i) It measures the amount of actual evapotranspiration which is released by plants. Also, the amount of precipitation that an area receives, the amount lost through the soil and amount of water lost to evapotranspiration can be calculated.

(ii) Two types of lysimeter are available namely weighing and non-weighing.

10. **Non- functional requirements:**

(i) It contains methods to evaluate the operation and performance of the middleware.

(ii) Ex: Scalability, Reliability, Availability, Real-Time / Timeliness

PART – B

11. **IoT Design Methodology includes:**

1. Purpose & Requirements Specification
2. Process Specification
3. Domain Model Specification
4. Information Model Specification
5. Service Specifications
6. IoT Level Specification
7.Functional View Specification
8.Operational View Specification
9.Device & Component Integration
10.Application Development

12. YANG Description – 2 marks

YANG Node types - 4 marks

Visual representation of HAProxy load balancer with explanation – 4 marks

13. (i) Component – based middleware (2)

(ii) Distributed middleware (2)

(iii) Service-Oriented middleware (2)

(iv) Cloud-based middleware (2)

(v) Node-based middleware (2)

14. **a) Medical sensors:**

(i) Heartbeat sensor (1)

(ii) Pulse sensor (1)

(iii) Blood glucose level sensor (1)

(iv) Blood pressure sensor (1)

(v) Body temperature sensor (1)

b) Touch sensor working principle (1)

Capacitive Touch sensor (2)

Resistive Touch sensor (2)

SASTRA

DEEMED TO BE UNIVERSITY

THINK SKILL | THINK TRANSPARENCY | THINK SASTRA

School of Computing
Third CIA Exam – Nov 2024
Course Code: CSE415R02
Course Name: Internet of Things
Duration: 90 minutes Max Marks: 50

PART A

Answer the following questions **5x2=10**

1. Identify and represent the key IoT enabling technologies in a diagrammatic way.
2. Why flooding algorithms are not useful for dense WSN?
3. List the commonly used M2M protocols.
4. Depict the SoCs at different levels for an IoT application, from end point to cloud, in a diagrammatic way.
5. List the different modules that create the Apache Hadoop framework.

PART B

Answer any three of the following questions **3x10=30**

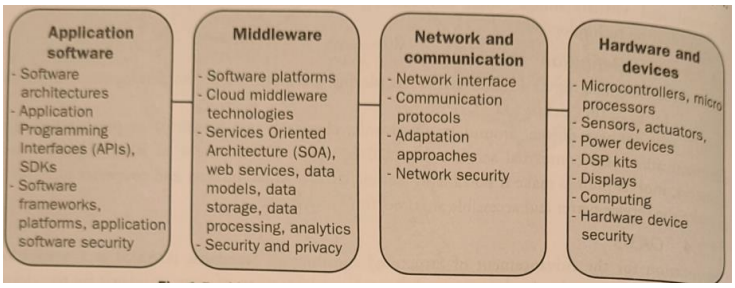
6. IoT application protocols are designed to address the constraints of IoT applications. Elaborate the various application layer messaging protocols that helps in addressing the constraints of IoT applications.
7. To address the limitations of existing network management protocols IETF listed a set of network operator requirements. Narrate these network operator requirements.
8. Various constraints for prototyping IoT product need to be well understood before the step of prototype design.
   a. Elaborate the common constraints for prototyping IoT product. (5)
   b. Narrate the important points to be considered while developing IoT products for different kinds of clients in consumer and industrial applications. (5)
9. IoT system security has to address basic security goals as defined by CIA. Elaborate the various attacks that are possible on IoT systems.

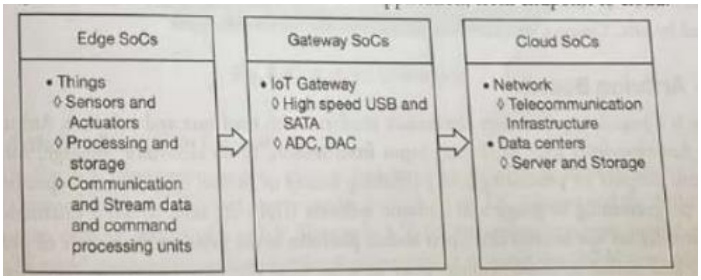
PART C

Answer the following question **1x10=10**

10. Designing IoT systems can be a complex and challenging tasks as these systems involve interactions between various components such as IoT devices and network resources, analytic components etc.
    a. Identify the steps involved in IoT system design methodology. (3)
    b. Depict the mapping of IoT layers to TCP/IP layers. (4)
    c. How Big Data analytics are useful in IoT environment. (3)

Part A

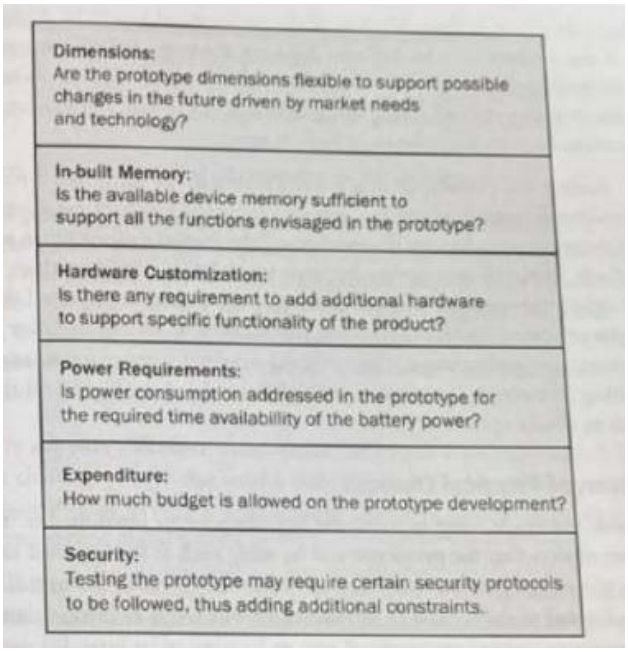
1. 

2. Unnecessary retransmission of messages and increased collision
3. Zigbee, Bluetooth, modbus, M-bus, wireless M-bus, PLC, 6LoWPAN, IEEE 802.15.4, Z-Wave
4. 

5. Hadoop common, Hadoop distributed file system, Hadoop MapReduce, Hadoop YARN

PART B

6. HTTP, FTP, CoAP, MQTT, SMQTT, XMPP, AMQP, DDS each (1.5)
7. Ease of use, distinction between configuration and state data, fetch configuration and state data separately, configuration of the network as a whole, configuration transactions across devices, configuration deltas, dump and restore configurations, configuration validation, configuration databse schemas, comparing configurations, role based access control, consistency of access control lists, multiple configuration sets, support for both data-oriented and task –oriented access control

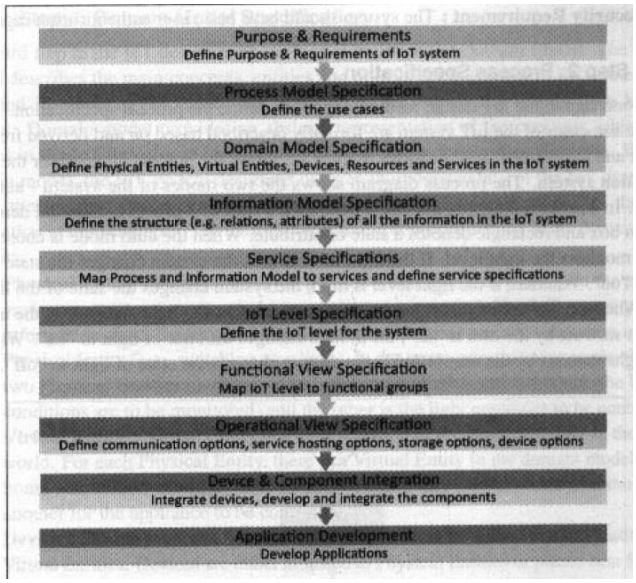
8. a. 

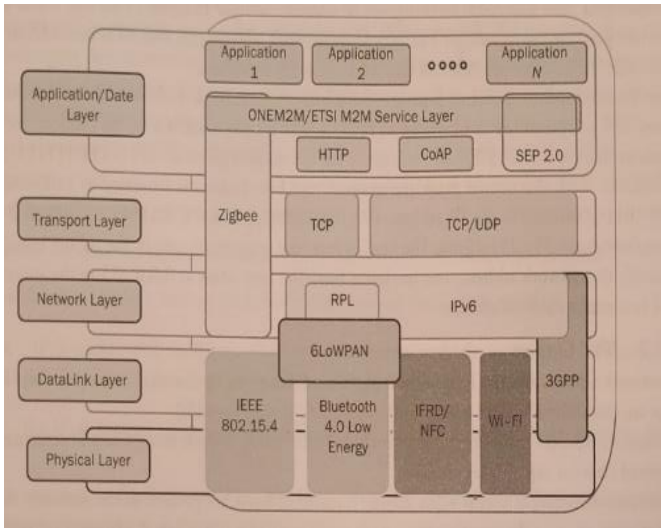
b. Unambiguous business strategy, optimizing design, selecting application specific power model, device security aspects, adapting to network technology evolution, support for multiple platforms, adopt well proven technologies

9. DoS in physical (each 2)
   DoS in Link layer
   DoS in Network
   DoS in transport
   DoS in application

PART C

10.

a. 

b. 

C.

IoT data analytics are useful for:

- Automating many decision-making processes so that human intervention is minimized and IoT devices and applications can autonomously perform actions.
- Increasing the efficiency with which processes can be executed. For example, supply chain operations can be made highly efficient by deploying IoT-based solutions.
- Condition-based monitoring and predictive maintenance of equipment, which is critical in many areas such as industries, manufacturing, healthcare, and transportation.
- Service efficiency that encompasses remote management, service chain, material management, etc.
- Analysis of the product usage by customers and accordingly customize the product thus enabling competitive advantage in the market.
- Reducing overall operational expenditure and increasing revenue.