KUMARAGURU
college of technology
character is life

M.TECH DEGREE EXAMINATIONS: NOV/DEC 2023

(Regulation 2018)

Third Semester

DATA SCIENCE

P18ITE0012: Streaming Analytics

COURSE OUTCOMES

CO1: Explain the need for stream computing.

CO2: Comprehend the architecture of stream analytics.

CO3: Build data flow management pipelines for streams.

CO4: Process the streaming data.

CO5: Deliver the results of streaming analytics.

Time: Three Hours   **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1. Which of the following statements about data streaming is true?   CO1   [K$_2$]

   a) Stream data is always unstructured data.

   b) Stream data often has a high velocity.

   c) Stream elements cannot be stored on disk.

   d) Stream data is always structured data.

2. Which of the stream processing engines is used in Twitter?   CO1   [K$_2$]

   a) Heron

   b) Apache Spark

   c) Apache Flume

   d) Google dataflow

3. Kafka is comparable to traditional messaging systems such as.   CO2   [K$_2$]

   a) Impala

   b) ActiveMQ

   c) BigTop

   d) Zookeeper

4. Matching type item with multiple choice code   CO2   [K$_2$]

List I	List II
A. Messaging system	i. Key-Value Pair
B. Clients and the servers	ii. Configuration information
C. NoSQL databases	iii. Kafka
D. ZooKeeper	iv. TCP

	A	B	C	D
a)	ii	i	iii	iv
b)	iii	iv	i	ii
c)	iv	i	ii	iii
d)	iii	i	ii	iv

5. The following items consists of two statements, one labeled as the "Assertion (A)" and the other as "Reason (R). You are to examine those two statements carefully and select the answers to these items using the codes given below: CO3   [K₃]

   Assertion (A) : secureClinentPort to listen on for secure client connections using SSL

   Reason (R) : *dataDir* is the location where ZooKeeper will store the in-memory

   database snapshots

   a) Both A and R are Individually true and R is the correct explanation of A   b) Both A and R are Individually true but R is not the correct explanation of A

   c) A is true but R is false   d) A is false but R is true

6. In the analysis phase which is the correct order of execution after forming unresolved logical plan _____.   CO3   [K₂]
   1) Search relation BY NAME FROM CATALOG
   2) Determine which attributes match to the same value to give them unique ID
   3) Map the name attribute
   4) Propagate and push type through expressions   c

   a)   1-3-2-4   b)   4-1-2-3

   c)   4-3-1-2   d)   2-1-3-4

7. Which of the following statements about sampling are incorrect?   CO4   [K₂]
   1. Sampling reduces the amount of data fed to a subsequent data mining algorithm.
   2. Sampling reduces the diversity of the data stream.
   3. Sampling aims to keep the statistical properties of the data intact.
   4. Sampling algorithms often need multiple passes over the data.
   5. Sampling increases the amount of data fed to a data mining algorithm

   a)   1,2,3,4,5   b)   1,3

   c)   2,3,4   d)   2,3,4,5

8. The following items consist of two statements, one labeled as the "Assertion (A)" and the other as "Reason (R). You are to examine those two statements carefully and select the answers to these items using the codes given below: CO5   [K₂]

   Assertion (A): Video streaming was among the first media applications to find widespread use on the Internet.

   Reason (R): Broadband connections and low-cost infrastructure have made video streaming services widely available

   a)   Both A and R are Individually true and   b)   Both A and R are Individually true but

	R is the correct explanation of A	R is not the correct explanation of A
c)	A is true but R is false	d) A is false but R is true

9. You have the following process: CO4   [K₂]

i.A CSV file is ingested by Microsoft Azure Stream Analytics.

Scoring is performed by Azure Machine Learning.

ii.Stream Analytics returns sentiment scoring through a web service endpoint.

iii.Stream Analytics creates an output blob.

You need to view the output of the scoring operation and to evaluate the throughput to

the Machine Learning models.

Which monitoring data should you evaluate from the Azure portal

a)	the request count of Stream Analytics	b)	the request count of Machine Learning
c)	the event count of Stream Analytics	d)	the event count of Machine Learning

10. Point out the correct statement: CO5   [K₂]

a) Video streaming was among the first media applications to find widespread use on

the Internet

b) VoIP is a form of audio streaming with a set of specialized protocols

c) Broadband connections and low-cost infrastructure have made video streaming

services widely available

d) Multi-streaming in mobile is called Simulcast and Simulcasting

a)	b,c,d	b)	a,b,c
c)	b and c	d)	a only

PART B (10 x 2 = 20 Marks)

11.	List the challenges in streaming data architecture.	CO1	[K₂]
12.	Differentiate vertical scaling and horizontal scaling.	CO1	[K₂]
13.	What is the role of sliding windows in the analysis of streaming data?	CO2	[K₂]
14.	Write short notes on high availability in streaming analytics.	CO2	[K₂]
15.	List out the different types of channel selectors in the flume agent. Which is the default channel selector in the flume agent?	CO3	[K₂]
16.	State some of the major methodologies for the data streaming process.	CO3	[K₂]
17.	Illustrate the leader election using Zookeeper with an example.	CO4	[K₂]
18.	Difference between streaming data bases and traditional database.	CO4	[K₂]
19.	Discuss the stochastic optimization in the context of streaming of data.	CO5	[K₂]
20.	Explain about the Zero MQ messages and Events.	CO5	[K₂]

PART C (10 x 5 = 50 Marks)

21.	Differentiate between DBMS and DSMS?	CO1	[K₂]
22.	What is Real Time Analytics Platform (RTAP)?	CO1	[K₂]
23.	Explain the components of streaming architecture with an example.	CO2	[K₂]
24.	Describe about Zookeeper's strengths and limitations to consider for streaming analytics.	CO2	[K₂]
25.	Discuss about dataflow programming model.	CO3	[K₂]
26.	Discuss the role of bloom filters in duplication detection task.	CO3	[K₂]
27.	Explain the distributed streaming processing operations like coordination, partition, and merging of the transactions.	CO4	[K₂]
28.	Discuss the Flume agent and its usage in streaming analytics.	CO4	[K₂]
29.	Describe various visualizing methods and techniques for the streaming of data.	CO5	[K₂]
30.	Difference between time series database and streaming analytics engine like Spark streaming.	CO5	[K₂]

Answer any TWO Questions
PART D (2 x 10 = 20 Marks)

31.	Discuss about to produce and consume records/messages with Kafka brokers.	CO1	[K₃]
32.	Describe the Zookeeper's architecture. Summarize the strengths, and limitations of the Zookeeper that are to be considered for streaming analytics.	CO2	[K₂]
33.	Explain the architecture of Apache Spark. Illustrate the role of Apache Spark to handle the data processing with streams of data.	CO4	[K₂]

****