AKTU B.Tech 3rd Year

CS & CS Allied

Data Analytics & Visualization

Unit-3

in ONE SHOT

Part-I

AKTU PYQs Covered



By Shubhangi Ma'am

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

SYLLABUS

UNIT-III: Mining Data Streams

Introduction to Streams Concepts, Stream Data Model and Architecture, Stream Computing, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Counting Oneness in a Window, Decaying Window, Real-Time Analytics Platform (RTAP) Applications, Case Studies: Real-time Sentiment Analysis, Stock Market Prediction.

Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 Define data streams with one real – time example. **AKTU(2025-26)**

Q.2 Describe stream data model and architecture with neat diagram. **AKTU(2025-26)**

Q.3 Explain sampling and filtering techniques in data streams . **AKTU(2025-26)**

Q.4 Analyze challenges in mining high speed data streams . **AKTU(2025-26)**

Q.5 Analyze the challenges of real – time sentiment analysis . **AKTU(2024-25)**

Q.6 Describe the concept of decaying windows in stream analytics. How does it help in managing memory & computational resources . **AKTU(2024-25)**

Q.7 Discuss the role of Real – Time analytics Platforms (RTAP) in modern enterprises . Provide a case study to study to support your explanation. **AKTU(2024-25)**

Q.8 Describe the steps involved in building a real time stock market prediction system using stream analytics . What are the key considerations. **AKTU(2024-25)**

GATEWAY CLASSES
Gateway Classes

Q.9 What is real time analysis ? **AKTU(2023-24)**

Q.10 Define data model . **AKTU(2023-24)**

Q.11 Discuss Sentiment Analysis . **AKTU(2023-24)**

Q.12 Discuss RTAP applications in detail. **AKTU(2023-24)**

Q.13 Demonstrate any algorithm for counting oneness in a window. **AKTU(2023-24)**

Q.14 Discuss as a Case study – Real time sentiment analysis . **AKTU(2023-24)**

Q.15 Elaborate on the methods used for filtering streams in real –time analytics .**AKTU(2024-25)**

Q.16 Provide an overview of Real-time Analytics Platforms applications , emphasizing their role in processing continuous data streams . How do these platforms support the development of real – time analytics solutions ? **AKTU(2024-25)**

Q.17 Discuss the concept of filtering in Data Stream Processing. Explain Bloom Filtering in detail. **AKTU(2024-25)**

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Q.18 Design a Real time analytics platform for social media sentiment Analysis .**AKTU(2025-26)**

Q.19 Explore the term – market Based Analysis . Is it supervised or unsupervised ? Determine how would a company use market basket analysis to improve its marketing strategies. **AKTU(2023-24)**

Gateway Classes

INTRODUCTION TO DATA STREAM AKTU(2025-26)

A data stream is a continuous sequence of data items that are generated and received over time. The data items arrive one after another and are usually processed as they arrive.

Unlike traditional data, where the complete dataset is available for processing, data in a stream keeps arriving continuously. Therefore, it is generally processed in real time or near real time.

Example:

Stock market prices are continuously generated throughout the day:

₹100 → ₹102 → ₹101 → ₹105 → ₹103 → ...

This continuously arriving sequence of stock prices is called a **DATA STREAM.**

Examples of Data Streams

Sensor readings from IoT devices , Online banking transactions , Social media posts, Network traffic data

Gateway Classes

CHARACTERISTICS OF DATA STREAM

Characteristics	Description
Continuous Data generation	Data streams are generated continuously over time. New data items keep arriving while the system is operating.
High Data Rate	A data stream may generate a large number of data items in a short period of time. The rate of data arrival can be very high.
Unbounded data	A data stream is generally unlimited or unbounded because new data can continue to arrive indefinitely
Real time processing	Data streams often need to be processed immediately or within a short period of time. This is important when quick decisions are required.
Time – varying data	The characteristics or values of data may change with time.

GATEWAY CLASSES
Gateway Classes

APPLICATION OF DATA STREAM

- Online Shopping – Processes orders, searches, and payments in real time.
- Social Media – Analyses posts, likes, comments, and trends.
- Banking – Detects unusual transactions and fraud.
- Stock Market – Monitors continuously changing prices and trades.
- Healthcare – Monitors patient data from medical devices.
- IoT – Processes continuous sensor data from smart devices.
- Network Monitoring – Detects unusual network traffic and security problems.
- Transportation – Analyses live traffic and vehicle location data.

Gateway Classes

STREAM DATA MODEL AKTU(2025-26)

A stream data model represents data as a continuous sequence of data items that arrive over time.

The data items are processed as they arrive rather than waiting for the complete dataset.

For example,

a stream may contain:

A B C D E F ...or 1 2 3 4 5 6 ...or 1010101010...

Each value or record arriving in the sequence is called a data item.

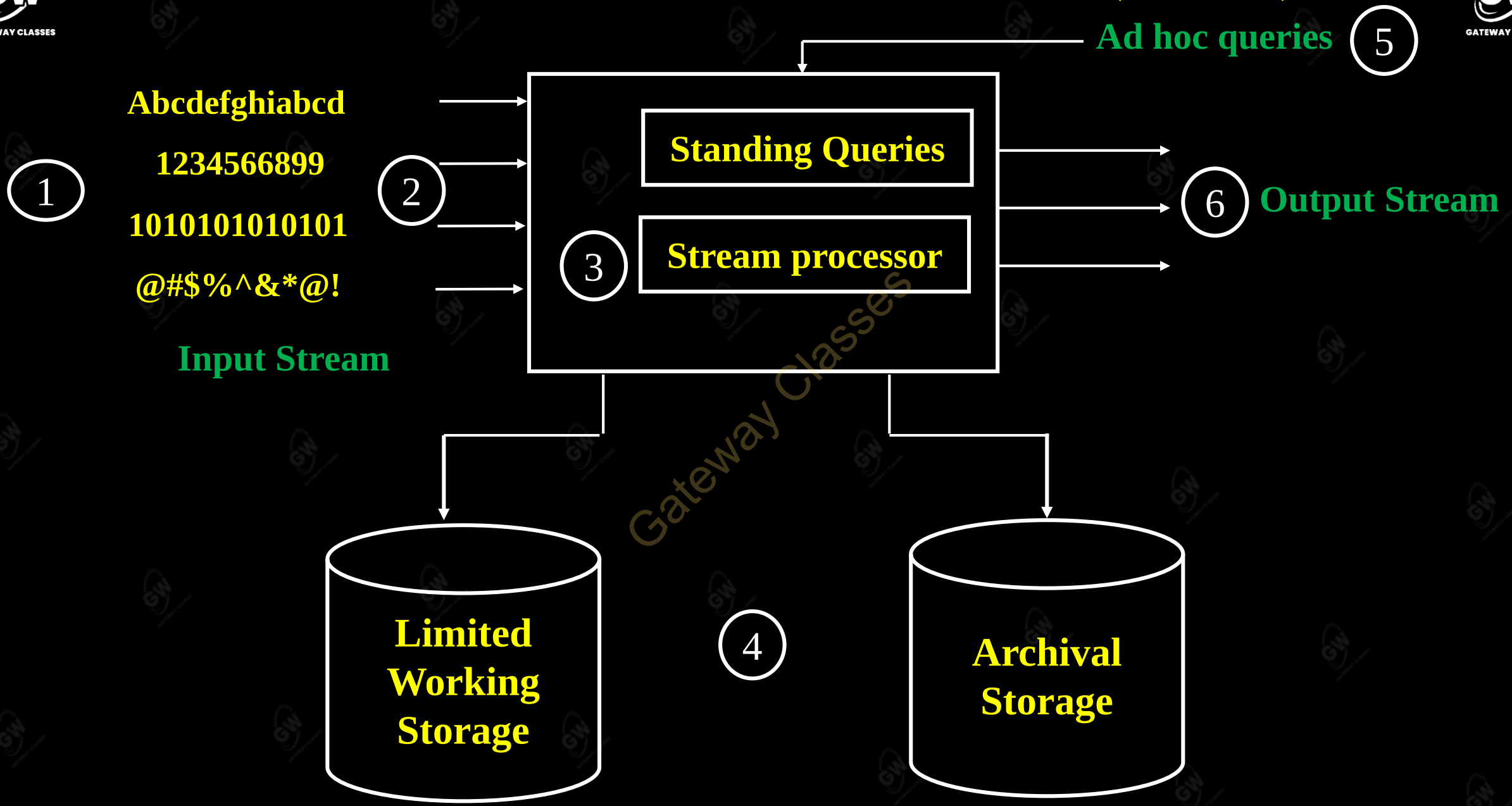
A data stream can be represented as:

$S = \{a_1, a_2, a_3, a_4, \ldots\}$

where $a_1, a_2, a_3, \ldots$ are the data items arriving over time.

Gateway Classes

STREAM DATA ARCHITECTURE AKTU(2025-26)



Gateway Classes

STREAM DATA ARCHITECTURE

The architecture shown in the diagram describes how a data stream is received , processed , stored and converted into an output stream.

Components of Stream Data Architecture :

1. **Input :** The input consists of continuously arriving data items. The data may come from different such as : Sensors , online transactions, social media etc.
2. **Input Data to the Stream Processor :** The incoming data stream is passed to the stream processor . Since data keeps arriving continuously , the processor handles the data items as they arrive.
3. **Stream processor :** It is the main processing component of the architecture. It continuously processes the incoming data stream & perform operations according to the required queries.

It can perform operations such as : filtering data, counting data items sampling data , finding patterns.

Gateway Classes

The stream processor can also use temporary working storage and archival storage during processing.

4. Limited working Storage : It is used to temporary store a limited amount of data required during stream processing. Since a data stream can be very large or unlimited, it is generally not possible to keep the entire stream in memory. Therefore, only the required or recent data is kept in working storage.

5. Archival Storage : It is used to store data for long term use . It can be accessed later for : Historical analysis, Reporting , Future processing etc.

6. Standing Queries : It is predefined query that continuously runs on the incoming data stream. It remains active & checks new data as it arrives .

7. Ad Hoc Queries : An ad hoc query is a query created by a user according to a specific requirement. Unlike a standing query, it is not continuously running by default. It is issued when the user needs particular information.

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

8. Output Stream

After processing the incoming data, the stream processor produces the required results as an output stream.

The output may contain : Filtered data , Calculated values , Alerts ,Statistical results , Detected patterns

Gateway Classes

STREAM COMPUTING

Stream computing is a method of processing continuously arriving data in real time or near real time.

It processes data items as they arrive instead of waiting for the complete dataset to be collected.

Example:

Processing live stock prices continuously to identify changes in the market.

Gateway Classes

WORKING OF STREAM COMPUTING

Working	Description
Data Generation	Data is Continuously generated from various sources
Data Ingestion	The incoming data is received as a continuous stream
Stream Processing	The data is processed immediately using required operations such as filtering, counting or aggregation.
Analysis	The processed data is analyzed to obtain useful information
Output	Results are generated in real time and used for quick decisions or actions

Gateway Classes

CHARACTERISTICS OF STREAM COMPUTING

1. Continuous Processing	**Data is processed continuously as new data items arrive in the system.**
2. Real-Time Processing	Stream computing provides results quickly, which makes it suitable for applications that require immediate responses.
3. High Data Rate	It is designed to handle data that arrives at a high speed and in large quantities.
4. Continuous Data	The input is not a fixed dataset. New data keeps arriving over time and may continue indefinitely.
5. Limited Storage	Since the stream can be very large or unlimited, it is usually not possible to store the complete stream. Only required or summarized data may be stored.
6. Quick Decision Making	Stream computing helps organizations analyze current data and make decisions quickly.

Gateway Classes

STREAM COMPUTING OPERATIONS

Operations	Description
Filtering	Selects only the required data items from the stream
Sampling	Selects a representative items from a large stream
Counting	Counts the required elements in the data stream
Aggregation	Calculates summary values such as sum, average, minimum, or maximum
Pattern detection	Identifies useful patterns or activities unusual in the stream

Gateway Classes

SAMPLING DATA IN A STREAM AKTU(2025-26)

Sampling data in a stream is the process of selecting a smaller set of data items from a continuously arriving data stream for analysis.

Since a data stream may be very large or unlimited, it is not always possible to process or store every data item. Sampling helps in analyzing a representative portion of the stream while using less memory and processing time.

Example

Suppose a website receives 1,00,000 user clicks in an hour. Instead of analyzing every click, a sample of 1,000 clicks can be selected for analysis.

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Need for Sampling

- Data streams can contain a very large number of data items.
- Storing the complete stream may require large storage.
- Processing every item can be time-consuming.
- Sampling reduces the amount of data that needs to be processed.
- It helps in performing analysis efficiently.

Gateway Classes

TECHNIQUES OF SAMPLING

1. Reservoir Sampling

- Reservoir sampling is a technique used to select and maintain a fixed-size random sample from a continuously arriving data stream. It is useful when the total size of the data stream is unknown or very large.

Example: A stream contains millions of posts, but we want to maintain only 100 randomly selected posts.

2. Bernoulli Sampling

- Bernoulli sampling is a technique in which each data item in a data stream is independently selected or rejected based on a fixed probability. The sample size is not fixed and may vary depending on the random selection.

Example: If the probability is 10%, every incoming post has an independent 10% chance of being selected.

Gateway Classes

GATEWAY CLASSES

Systematic sampling

It is a technique in which data items are selected from a data stream at a fixed interval. After choosing a starting point, every kth item is selected according to the given interval.

Example: If the interval is 5, then every 5th item is selected: 5, 10, 15, 20....

4. Sliding Window Sampling

Sliding window sampling is a technique in which a fixed-size window is maintained over a data stream to process recent data. As new data arrives, older data is removed and the window moves forward.

5. Priority Sampling

- Priority sampling is a technique in which data items in a stream are assigned priority values, and items with higher priorities are selected or maintained in the sample.

GatewayClasses

FILTERING STREAMS AKTU(2025-26)

Filtering a stream is the process of selecting only the required data items from a continuously arriving data stream and ignoring the unwanted items.

Working

- A filter condition is defined according to the requirement.
- Each incoming data item is checked against the condition.
- Data items that satisfy the condition are selected and passed to the output stream.
- Other data items are discarded.

Example:

If a stream contains transaction amounts and we want transactions above ₹10,000, only those transactions will be selected.

Gateway Classes

TECHNIQUES OF FILTERING STREAMS

Techniques	Description
Value based filtering	Data items are selected based on their specific values
Range based filtering	Data items within a specified range are selected
Keyword based filtering	Data items containing a particular keyword are selected
Pattern based filtering	Data items matching a specified pattern are selected
Probabilistic filtering	Data items are selected or filtered using a probabilistic based method

GATEWAY CLASSES
Gateway Classes

BLOOM FILTERING

Bloom filtering is a space-efficient technique used to determine whether a particular data item may be present or is definitely not present in a data set or data stream.

It uses a Bloom Filter, which is a probabilistic data structure based on a bit array and multiple hash functions.

Example : Suppose an online shopping website has millions of registered product IDs:

P101, P102, P103, P104, P105, ...

Now a customer searches for product ID P150.

Instead of searching the complete database, the website first checks the Bloom Filter.

Bloom Filter says “Definitely Not Present”

→ P150 is definitely not in the database. No further search is needed.

Bloom Filter says “Maybe Present”

→ P150 may be in the database. Now the website checks the actual database to confirm.

Gateway Classes

COUNTING DISTINCT ELEMENTS IN A STREAM

AKTU(2024-25)

Counting distinct elements is the process of finding the number of different or unique elements present in a continuously arriving data stream.

Since a data stream can be very large, storing and checking every element may require a large amount of memory. Therefore, approximate techniques are often used.

Example

Consider the data stream:

A, B, A, C, B, D, C

The distinct elements are:

A, B, C, D

Therefore, the number of distinct elements is 4.

Gateway Classes

TECHNIQUES TO COUNT DISTINCT ELEMENT

1. Exact Counting

Stores the elements that have already appeared.

When a new element arrives, it is checked against the stored elements.

If it is new, the distinct count is increased.

Gives the exact answer, but requires more memory.

Example:

A, B, A, C, B → unique elements = {A, B, C} → count = 3

Gateway Classes

2. FLAJOLET – MARTIN ALGORITHM

- It is a probabilistic technique for estimating the number of distinct elements.
- It uses a hash function on incoming elements.
- The hash values are examined to estimate how many different elements are present.
- It uses very little memory compared with exact counting.
- The result is an estimate, not necessarily the exact count.

3. HyperLogLog (HLL)

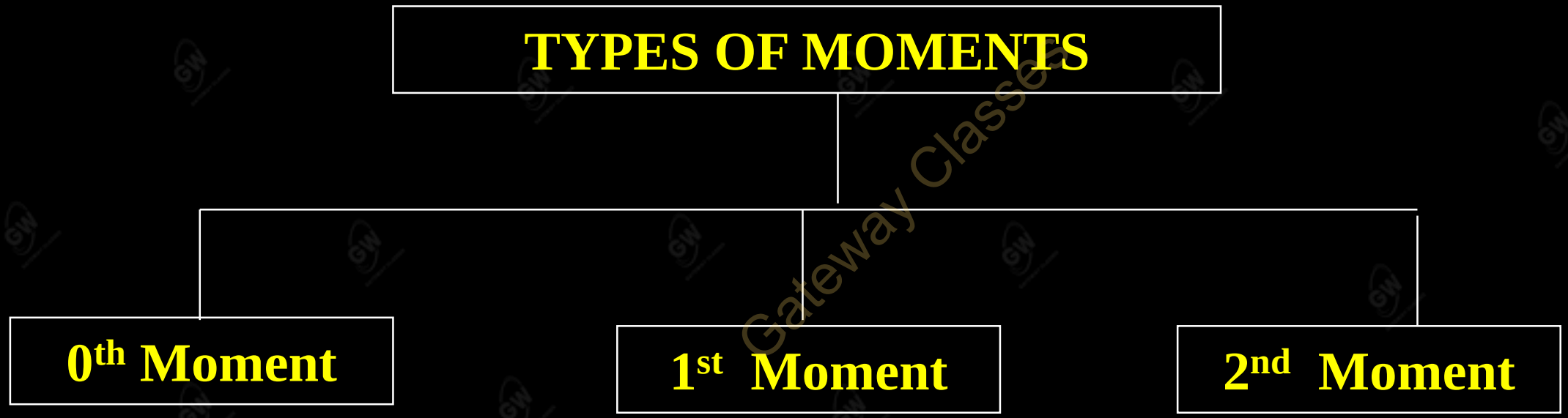
- It is an improved probabilistic technique for estimating distinct elements.
- It uses multiple registers to store information obtained from hash values.
- The information from these registers is combined to estimate the number of unique elements.
- It provides good accuracy with very low memory.

GATEWAY CLASSES
Gateway Classes

ESTIMATING MOMENTS IN A DATA STREAM

Estimating moments is the process of calculating statistical properties of a data stream using limited memory, without storing the complete stream.

In a data stream, moments are used to understand the frequency and distribution of data items.



Gateway Classes

1. ZEROTH MOMENT (F_0)

The zeroth moment represents the number of distinct (different) elements present in a data stream.

Example:

Data stream: A, B, A, C, B, A

Different elements are A, B, C.

Therefore, $F_0 = 3$.

2. FIRST MOMENT (F_1)

The first moment represents the total number of elements in a data stream.

Example:

Data stream: A, B, A, C, B, A

Total number of elements = 6

Therefore, $F_1 = 6$.

Gateway Classes

3. SECOND MOMENT (F_2)

The second moment measures the concentration of frequencies in a data stream. It shows whether some elements occur much more frequently than others.

Formula:

$F_2 = \Sigma(f_i)^2$

Example:

Suppose:

A = 3 times

B = 2 times

C = 1 time

Then:

$F_2 = 3^2 + 2^2 + 1^2 = 9 + 4 + 1 = 14$

Gateway Classes

4. HIGHER-ORDER MOMENTS (F_3, F_4, ...)

Higher-order moments measure the dominance of frequently occurring elements more strongly than the second moment.

Formula:

$F_k = \Sigma(f_i)^k$

Example:

If:

A = 3 times

B = 2 times

C = 1 time

Then:

$F_3 = 3^3 + 2^3 + 1^3$

$= 27 + 8 + 1 = 36$

Here, the high-frequency element A has a much larger effect.

Gateway Classes

TECHNIQUES FOR ESTIMATING MOMENTS

1. Exact Computation

- All required data or frequency information is maintained.
- It can provide an exact result.
- It may require a large amount of memory for very large data streams.

2. Approximate Algorithms

- Probabilistic or randomized techniques are used.
- They estimate the required moment using limited memory.
- They are suitable for very large or continuous data streams.

Gateway Classes

COUNTING ONENESS IN A WINDOW AKTU(2024-25)

Counting oneness in a window is the process of counting the number of 1s in a recent portion (window) of a data stream.

Since a data stream is continuous, it is not practical to store the complete stream. Therefore, a window is used to focus on recent data.

Window

A window is a selected portion of the most recent data items in a stream.

For example:

1 0 1 1 0 1 0 1

If the last 5 bits are considered, the window is:

1 0 1 0 1

The number of 1s in this window is 3.

Gateway Classes

TYPES OF WINDOWS IN DATA STREAMS

A window is a selected portion of a continuously arriving data stream that is considered for processing or analysis.

1. Landmark Window : A landmark window contains all data items from a fixed starting point up to the current time. The starting point remains fixed, so the window keeps growing.

Example :

A company wants to analyze all customer orders received since the beginning of the day. At 10 AM, it analyzes orders from 12 AM–10 AM; at 2 PM, it analyzes orders from 12 AM–2 PM.

2. Sliding Window: A sliding window contains only the most recent fixed number of data items or a fixed period of time. As new data arrives, old data is removed.

Example:

A website monitors the last 10 minutes of user activity. When a new minute of data arrives, the oldest minute is removed.

GatewayClasses

GATEWAY CLASSES

GATEWAY CLASSES

3. Tumbling Window

A tumbling window divides a data stream into fixed-size, non-overlapping windows. Once one window ends, the next window starts.

Example:

A shopping website counts orders every 10 minutes:

10:00–10:10 → Window 1

10:10–10:20 → Window 2

10:20–10:30 → Window 3

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

4. Session Window

A session window groups data according to continuous user activity and ends when the user remains inactive for a specified period.

Example:

A user continuously browses an online shopping website. All activities during that visit belong to one session. If the user remains inactive for 30 minutes, the session ends; activity after that starts a new session.

Gateway Classes

GATEWAY CLASSES

DECAYING WINDOW

A decaying window is a window in which recent data items are given more importance than older data items. The importance or weight of an item decreases gradually as the item becomes older.

Example

In stock market analysis, the latest stock prices may be given more importance than prices from several hours or days ago.

Gateway Classes

TYPES OF DECAY FUNCTIONS

Decay functions determine how the importance of data decreases as it becomes older.

1. Exponential Decay

- The importance of data decreases rapidly with time.
- Recent data gets much higher weight than older data.
- It is commonly used in real-time data analysis.

2. Polynomial Decay

- The importance of data decreases according to a polynomial function.
- The decrease is generally slower than exponential decay.
- It gives older data some influence for a longer period.

GATEWAY CLASSES
Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

3. Linear Decay

- The importance of data decreases at a constant rate with time.
- After a specified time, its weight can become zero.
- It is simple and easy to implement.

Gateway Classes

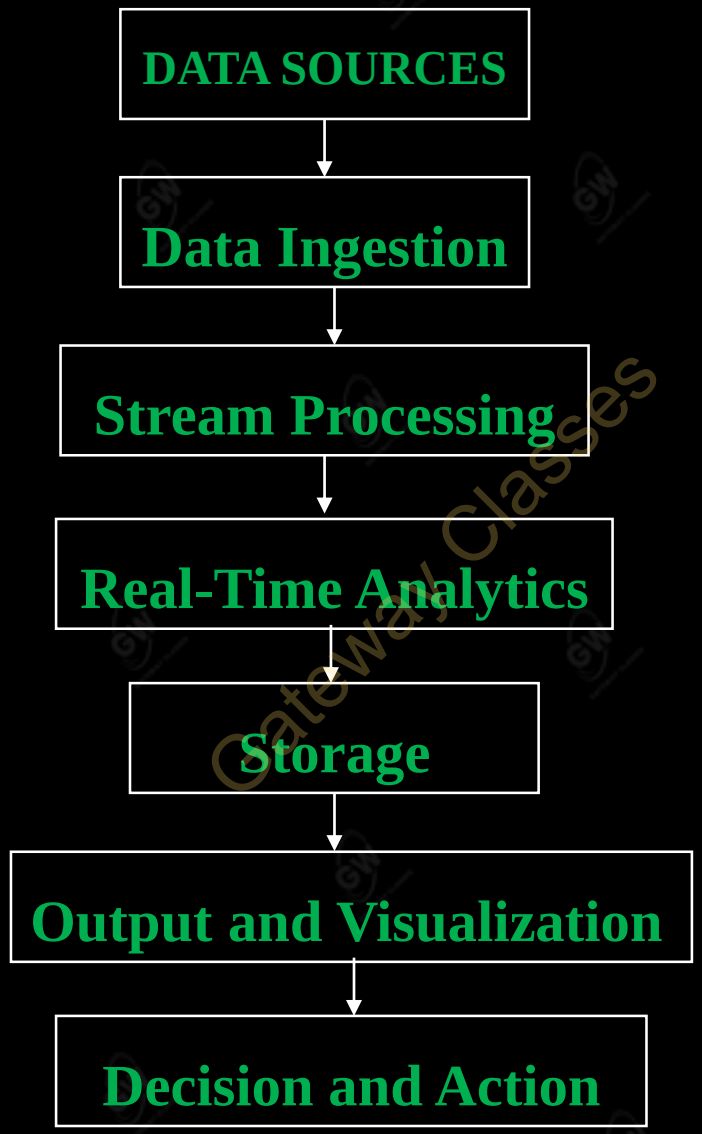
REAL-TIME ANALYTICS PLATFORM (RTAP) AKTU(2024-25)

A Real-Time Analytics Platform (RTAP) is a system that collects, processes, and analyzes continuously generated data in real time or near real time. It provides quick results so that organizations can monitor situations and take immediate decisions.

Unlike traditional analytics, where data is often stored first and analyzed later, RTAP processes data as it arrives.

Gateway Classes

WORKING OF RTAP



GATEWAY CLASSES
Gateway Classes

MAIN COMPONENTS OF RTAP

1. Data Sources

Data is continuously generated from different sources such as:

Sensors and IoT devices , Websites and applications , Social media , Online transactions , Stock markets

2. Data Ingestion

The incoming data is collected and transferred to the analytics platform in the form of a continuous data stream.

3. Stream Processing

The incoming data is processed immediately. Operations such as filtering, sampling, counting, and aggregation can be performed.

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

4. Real-Time Analytics

The processed data is analyzed to identify patterns, trends, events, or unusual activities.

5. Storage

Important data or results can be stored for historical analysis and future reference.

6. Output and Visualization

The results can be presented through dashboards, reports, notifications, or alerts so that users can understand the current situation.

7. Decision and Action

The results are used to take quick decisions or automatic actions.

Gateway Classes

APPLICATIONS OF RTAP

1. Fraud Detection

Financial transactions can be analyzed immediately to detect suspicious activities.

2. Stock Market Analysis

Live market data can be analyzed to monitor price changes and market trends.

3. Real-Time Sentiment Analysis

Social media posts can be continuously analyzed to identify people's opinions and sentiments.

4. IoT and Sensor Monitoring

Sensor data can be analyzed continuously to detect changes or abnormal conditions.

5. Network Monitoring

Network traffic can be monitored to detect unusual activities and possible security threats.

6. Recommendation Systems

User activities can be analyzed in real time to provide relevant recommendations.

GATEWAY CLASSES
Gateway Classes

CHARACTERISTICS OF RTAP

Real-Time Processing: Processes data as it arrives.

Low Latency: Produces results with very little delay.

Continuous Analysis: Continuously analyzes incoming data.

Scalability: Can handle large volumes of streaming data.

Quick Decision Making: Helps in taking immediate decisions.

Continuous Monitoring: Allows systems to be monitored continuously.

Fault Tolerance: Can continue processing even when some components fail.

GATEWAY CLASSES
Gateway Classes

1. Real-Time Sentiment Analysis

Real-time sentiment analysis is the process of continuously analyzing incoming text data to identify the sentiment or opinion expressed in it.

The sentiment is generally classified as:

- Positive
- Negative
- Neutral

Gateway Classes

Example

A company can analyze continuously arriving social media posts about its new product to determine whether customers are responding positively or negatively.

GATEWAY CLASSES
GATEWAY CLASSES

Working

- Text data is continuously collected from sources such as social media, reviews, and online platforms.
- The incoming text is processed as a data stream.
- The text is analyzed using sentiment analysis techniques.
- Each text item is classified as positive, negative, or neutral.
- The results are continuously updated to monitor the current public opinion.

Applications

- Monitoring customer opinions
- Social media monitoring
- Customer feedback analysis

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

2. Stock Market Prediction

- Stock market prediction is the process of analyzing stock-related data to estimate future price movements or market trends.
- In stream analytics, stock data is continuously received and analyzed in real time.

Example

If the price and trading volume of a stock are continuously increasing, the system may identify an upward market trend.

Gateway Classes

GATEWAY CLASSES

Working

- Live stock prices and transaction data are continuously collected.
- The data enters the system as a data stream.
- Stream processing techniques are used to process the incoming data.
- Current prices, trading volume, and market trends are analyzed.
- The analyzed information is used to identify possible future price movements.
- Results are provided quickly to support investment decisions.

Applications

- Monitoring stock prices
- Identifying market trends
- Risk analysis
- Real-time market monitoring

Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 Define data streams with one real – time example. **AKTU(2025-26)**

Q.2 Describe stream data model and architecture with neat diagram. **AKTU(2025-26)**

Q.3 Explain sampling and filtering techniques in data streams . **AKTU(2025-26)**

Q.4 Analyze challenges in mining high speed data streams . **AKTU(2025-26)**

Q.5 Analyze the challenges of real – time sentiment analysis . **AKTU(2024-25)**

Q.6 Describe the concept of decaying windows in stream analytics. How does it help in managing memory & computational resources . **AKTU(2024-25)**

Q.7 Discuss the role of Real – Time analytics Platforms (RTAP) in modern enterprises . Provide a case study to study to support your explanation. **AKTU(2024-25)**

Q.8 Describe the steps involved in building a real time stock market prediction system using stream analytics . What are the key considerations. **AKTU(2024-25)**

GATEWAY CLASSES
Gateway Classes

Q.9 What is real time analysis ? **AKTU(2023-24)**

Q.10 Define data model . **AKTU(2023-24)**

Q.11 Discuss Sentiment Analysis . **AKTU(2023-24)**

Q.12 Discuss RTAP applications in detail. **AKTU(2023-24)**

Q.13 Demonstrate any algorithm for counting oneness in a window. **AKTU(2023-24)**

Q.14 Discuss as a Case study – Real time sentiment analysis . **AKTU(2023-24)**

Q.15 Elaborate on the methods used for filtering streams in real –time analytics .**AKTU(2024-25)**

Q.16 Provide an overview of Real-time Analytics Platforms applications , emphasizing their role in processing continuous data streams . How do these platforms support the development of real – time analytics solutions ? **AKTU(2024-25)**

Q.17 Discuss the concept of filtering in Data Stream Processing. Explain Bloom Filtering in detail. **AKTU(2024-25)**

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

Q.18 Design a Real time analytics platform for social media sentiment Analysis .**AKTU(2025-26)**

Q.19 Explore the term – market Based Analysis . Is it supervised or unsupervised ? Determine how would a company use market basket analysis to improve its marketing strategies. **AKTU(2023-24)**

Gateway Classes

Gateway Classes

GW
GATEWAY CLASSES

Empowering Learners, Transforming Futures

THANK YOU

Gateway Classes

GET IT ON Google Play

Download on the App Store

Helpline No- 7819 0058 53 , 9259833200

AKTU B.Tech 3rd Year

CS & CS Allied

Data Analytics & Visualization

Unit-3

in ONE SHOT

Part-II (Numerical)

AKTU PYQs Covered



By Shubhangi Ma'am

GATEWAY CLASSES

Gateway Classes

GATEWAY CLASSES

GATEWAY CLASSES

SYLLABUS

UNIT-III: Mining Data Streams

Introduction to Streams Concepts, Stream Data Model and Architecture, Stream Computing, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Counting Oneness in a Window, Decaying Window, Real-Time Analytics Platform (RTAP) Applications, Case Studies: Real-time Sentiment Analysis, Stock Market Prediction.

Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 A data streams generates 120,000 records per minute. If only 35% data is retained using a decaying window , calculate the number of records retained per minute & justify the use of windowing . **AKTU(2025-26)**

Q.2 A sliding window of size 200 slides by 50 elements . Calculate the umber of windows formed after processing 1,000 elements and explain its impact on latency . **AKTU(2025-26)**

Q.3 Explain Alon – Matias – Szegedy algorithm for second moments on the given stream{ a, b, b, d , a ,c ,d ,a , b , d , c , a , a , b}. **AKTU(2022-23)**

Q.4 Explain and apply flajolet- Martin Algorithm on the following stream of data to identify unique elements in the stream. **AKTU(2024 - 25)**

S = 1,3,2,1,2,3,4,3,1,2,3,1

Given : h(x) = (6x + 1) mod 5

Gateway Classes

SLIDING WINDOW – NUMERICALS

Q.1 : A sliding window has a size of 200 elements and slides by 50 elements at a time. Calculate the number of windows formed after processing 1000 elements. AKTU(2025-26)

Solution

Given:

Window size , W = 200

Slide size S = 50

Total elements , N = 1000

First window (1st) = 200

After that, every new window adds 50 new elements.

Number of windows: [N – W]/ S + 1

Put values : [1000 – 200]/50 + 1 = 800/50 + 1 = 16 + 1 = 17

No. of windows = N – W/S + 1

Final Answer = 17

Gateway Classes

GATEWAY CLASSES

Q.2 A data stream contains 20 elements. A sliding window has a size of 5 elements and slides by 2 elements at a time. If the final partial window is also counted, determine the total number of windows formed.

Solution

Given:

Total number of elements, N = 20

Window size, W = 5

Slide size, S = 2

The windows will start at:

1, 3, 5, 7, 9, 11, 13, 15, 17

Gateway Classes

Windows	Elements	Type
1	1-5	Complete
2	3-7	Complete
3	5-9	Complete
4	7-11	Complete
5	9-13	Complete
6	11-15	Complete
7	13-17	Complete
8	15-19	Complete
9	17-20	partial

Gateway Classes

For the 9th window, the required range would be 17–21, but element 21 does not exist. Therefore, the available elements 17–20 form the final partial window.

So, **Complete windows = 8 & partial windows = 1**

Complete windows = 9

Q.3 : A stream contains 1000 elements. A sliding window of size 100 slides by 50 elements. Determine the starting and ending positions of the 5th window.

Solution :

1st window = 1 - 100

2nd window : It slides by 50 = 51 - 150

3rd window = 101 - 200

4th window = 151 - 250

5th window = 201 – 300

Answer : 5^{th} window = 201 to 300

Gateway Classes

Final Answer ,
5^{th} window = 201 to 300

DECAYING WINDOW – NUMERICALS

Q.4 : A data stream generates 120,000 records per minute. A decaying window retains 35% of the data. Calculate the number of records retained per minute. **AKTU(2025-26)**

Solution

Given:

Total records = 120,000

Retained data = 35%

Retained Records = 35/100 * 120000 = 0.35 * 120000 = 42, 000

Final Answer
Retained records = 42,000

Gateway Classes

Q.5 : A decaying window uses a decay factor of 0.5. If the current data has weight 1, find the weight of data that is 1, 2 and 3 time units old.

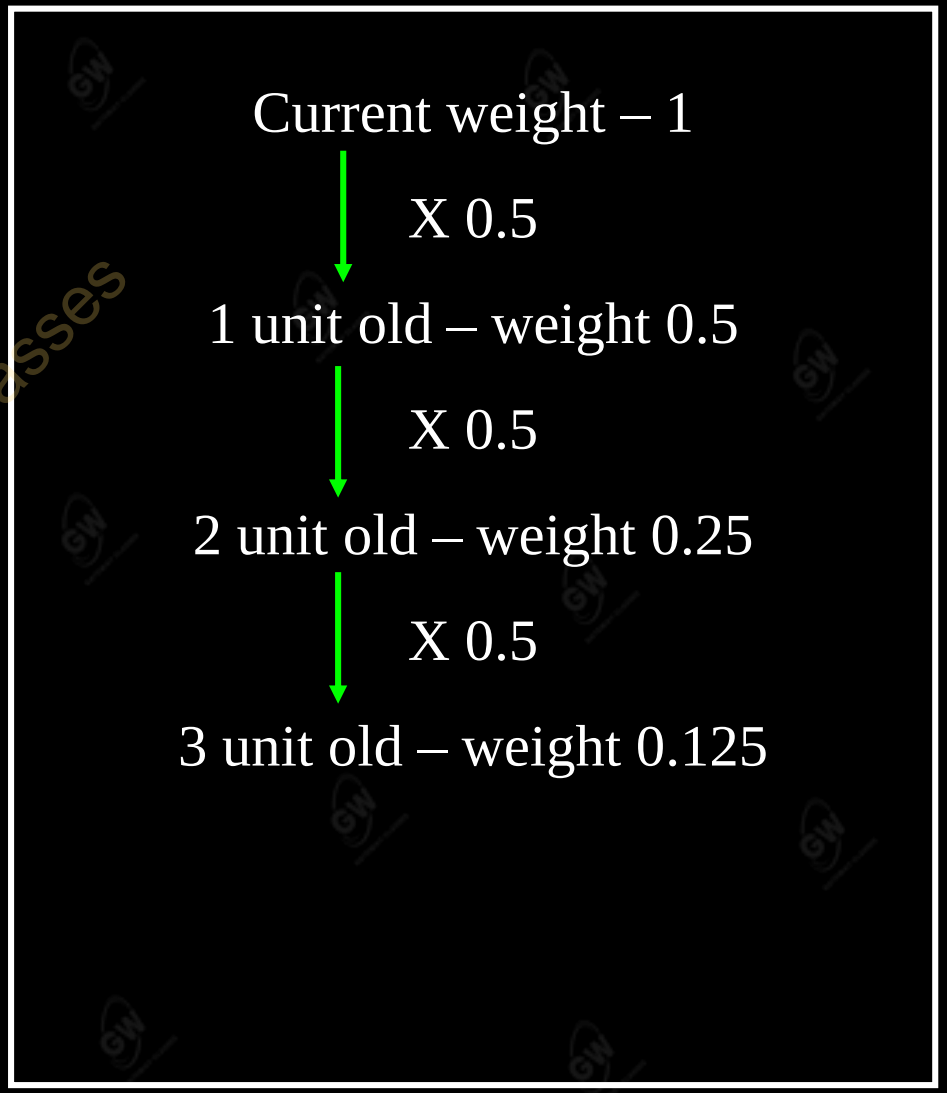
Solution

Formula : $w = (0.5)^t$

For 1 time unit : $w = (0.5)^1 = 0.5$

For 2 time unit : $w = (0.5)^2 = 0.25$

For 3 time unit : $w = (0.5)^3$ 0.125



Gateway Classes

COUNTINF DISTINCT ELEMENTS
FLAJOLET – MARTIN (FM)– NUMERICALS

Q.6 : Consider the following data stream : A , B , C , A , D , B

Using the Flajolet–Martin (FM) algorithm, estimate the number of distinct elements. Assume the following hash values:

ELEMENT	HASH VALUE
A	5
B	10
C	12
D	8

Gateway Classes

Step 1 : convert hash values into binary values

Element	Hash	Binary
A	5	101
B	10	1010
C	12	1100
D	8	1000

Step 2 :Count tailing zeros

Trailing Zeros :

Element	Hash	Trailing zeros
A	5	0
B	10	1
C	12	2
D	8	3

Gateway Classes

Maximum trailing zeros ,r = 3

Step 3 : Estimating Distinct elements

FM Algorithm : Estimated distinct elements = 2^n

Therefore , R = $2^r = 2^3 = 8$

Final Answer :
R = 8

Gateway Classes

Q.7 : Explain and apply flajolet- Martin Algorithm on the following stream of data to identify unique elements in the stream. AKTU(2024 - 25)

S = 1,3,2,1,2,3,4,3,1,2,3,1

Given : h(x) = (6x + 1) mod 5

Solution :

Step 1 : Calculating hash function

h(x)	Given	Hash value
h(1)	(6 * 1 + 1) mod 5	2
h(3)	(6 * 3 + 1) mod 5	4
h(2)	(6 * 2 + 1) mod 5	3
h(1)	(6 * 1 + 1) mod 5	2
h(2)	(6 * 2 + 1) mod 5	3
h(3)	(6 * 3 + 1) mod 5	4
h(4)	(6 * 4 + 1) mod 5	0
h(3)	(6 * 3 + 1) mod 5	4
h(1)	(6 * 1 + 1) mod 5	2
h(2)	(6 * 2 + 1) mod 5	3
h(3)	(6 * 3 + 1) mod 5	4
h(1)	(6 * 1 + 1) mod 5	2

Gateway Classes

Step 2 : Calculating Binary Equivalent :

h(x)	Hash value	Equivalent Binary
h(1)	2	010
h(3)	4	100
h(2)	3	011
h(1)	2	010
h(2)	3	011
h(3)	4	100
h(4)	0	000
h(3)	4	100
h(1)	2	010
h(2)	3	011
h(3)	4	100

Gateway Classes

R = 2

The Distinct value ,

$R = 2^r = R = 2^2 = 4$

Final Answer :
R = 4

Q.8 : For a data stream, the maximum number of trailing zeros obtained from the hash values is 4. Using the Flajolet–Martin algorithm, estimate the number of distinct elements.

Solution

Given : $\mathbf{r} = 4$

FM formula :

Estimated distinct elements = $2^r = 2^4 = 16$

Final Answer :
R = 16

Gateway Classes

ESTIMATING MOMENTS

Formulas:

$$\text{Zero moment } (F_0) = \sum_{i \in A} I(f_i > 0)$$

where $I(\text{condition}) = 1$ if condition is true
$= 0$ otherwise

$$\text{First moment } (F_1) = \sum_{i \in A} f_i$$

where f_i is the frequency of element i

$$\text{Second moment } (F_2) = \sum_{i \in A} f_i^2$$

Example:

Consider the data stream:

$\{ a, b, a, c, a, d, b, a, e, b, c, a \}$

Find the zero moment, first moment and second moment of the stream.

Gateway Classes

Solution:

Frequencies of elements:

a: 5, b: 3, c: 2, d: 1, e: 1

$$\text{Zero moment } (F_0) = \sum_{i \in A} I(f_i > 0) = 5$$

[Elements with non-zero frequency: a, b, c, d, e $\rightarrow$ 5 elements]

$$\text{First moment } (F_1) = \sum_{i \in A} f_i = 5 + 3 + 2 + 1 + 1 = 12$$

[Length of the stream]

$$\text{Second moment } (F_2) = \sum_{i \in A} f_i^2 = 5^2 + 3^2 + 2^2 + 1^2 + 1^2 = 25 + 9 + 4 + 1 + 1 = 40$$

[Surprise number]

ESTIMATING MOMENTS – AMS ALGORITHM – NUMERICALS

Q.9 : Consider the data stream : A, B, A, C, B, A, C , B

Estimate the second frequency moment F_2 of the stream .

Step 1 :

ELEMENT	FREQUENCY
A	3
B	3
C	2

Gateway Classes

Step 2 : formula for second moment =

$$F_2 = (3)^2 + (3)^2 + (2)^2 = 9 + 9 + 4 = 22$$

$$F_2 = \sum f_i^2$$

Final Answer :
$F_2 = 22$

Q.10 : In a data stream , an element x is selected at random .Its frequency in the stream is $f_x = 5$, and the stream length is n = 20 Using the AMS estimator X = n ($2f_x$ -1). Calculate the estimate of f_2

Solution :

Given : n = 20 $f_x = 5$

AMS estimator , $X = n(2f_x - 1)$

Substitute : X = 20(2(5) -1) = 20(10-1) = 20 * 9 = 180

Estimated F_2 = 180

AMS estimator , $X = n(2f_x - 1)$

Final Answer :
$F_2 = 180$

Gateway Classes

GATEWAY CLASSES

Q.11 : Suppose AMS algorithm gives three estimates of the second moment . X_1 = 180 , X_2 = 220 X_3 = 200 . Find the final estimated F_2 using their average .

Solution : $F_2 = X_1 + X_2 + X_3 / 3 = 180 + 220 + 200 / 3$

$= 600/3 = 200$

Final Answer :
F_2 = 200

Gateway Classes

Q.12 : Explain Alon – Matias – Szegedy algorithm for second moments on the given stream{ a, b , c, b, d , a ,c ,d ,a , b , d , c , a , a , b}. AKTU(2022-23)

Solution :

Step 1 : There are 14 elements , n = 14

Element	Frequency
a	5
b	4
c	3
d	3

Gateway Classes

Apply the AMS estimator , The AMS estimator for the second frequency moment is :

$$X = n(2f_x - 1)$$

Let us select 4 random positions : 1 , 2, 3, 4

Position	Element	Frequency	AMS Estimate
1	a	5	15(2*5-1) = 135
2	b	4	15(2*4-1) = 105
3	c	3	15(2*3-1) = 75
4	d	3	15(2*4-1) = 105

Step 2 : Average

$f_2 \approx 135 + 105 + 75 + 105 / 4 = 105$

Gateway Classes

Final Answer :
$F_2 \approx 105$

COUNTING ONENESS IN A WINDOW – DGIM ALGORITHM NUMERICALS

Q.13 Consider the following binary stream : 1 0 1 1 0 1 0 1

Using the DGIM algorithm, construct the buckets for the stream and estimate the number of 1s in the complete window.

Step 1: Find Timestamps

TIMESTAMPS	BIT
1	1
2	0
3	1
4	1
5	0
6	1
7	0
8	1

Gateway Classes

GATEWAY CLASSES

Step 2: formation of buckets , Maintain Bucket Size Maximum 2 buckets.

Initially, At timestamp 1 → [1]

At timestamp 3 → [1][1]

At timestamp 4 → three size-1 buckets occur, so the two oldest are merged → [2][1]

At timestamp 6 → [2][1][1]

At timestamp 8 → two oldest size-1 buckets are merged → [2][2][1]

So, the final buckets are : [2][2][1]

Step 3 : For the complete window , all buckets are included ,

Estimated number of 1's = 2 + 2 + 1 = 5

Final Answer : 5 ones

GatewayClasses

RESERVOIR SAMPLING : ALGORITHM

Suppose the reservoir size is k.

Step 1: Store the first k elements in the reservoir.

Step 2: For the $(k + 1)^{th}$ element, generate a random number between 1 and k + 1.

Step 3: If the generated number is within the reservoir positions, replace that position with the new element.

Otherwise, do not replace any element.

Step 4: Repeat this process for every remaining element in the stream.

Step 5: The final reservoir is the required random sample.

Gateway Classes

GATEWAY CLASSES

Q. 14 : Consider the following data stream : { 10,20,30,40,50} A reservoir of size k = 2 is maintained.

Suppose the following decisions are given:

30 replaces 10

40 is not selected

50 replaces 20

Find the final reservoir.

Solution : Initially, the first two elements are stored : R = [10,20]

When 30 arrives : [30,20]

When 40 arrives, it is not selected

When 50 arrives, it replaces 20 : R = [30,50]

Final Answer :
R = [30 , 50]

Gateway Classes

GATEWAY CLASSES

Q. 15 : A data stream contains 100 elements, and a reservoir of size 10 is maintained.

Find the probability that a particular element is present in the final reservoir.

Solution :

Given : n = 100 , k = 10

Using , P = k/n = 10/100 = 0.1

Therefore , P = 10%

The probability that any particular element is present in the final reservoir is : 10%

Gateway Classes

BLOOM FILTER : NUMERICAL

Q. 16 : A Bloom Filter uses a bit array of size 10, initially containing all zeros.

It uses two hash functions : $h_1(x) = x \bmod 10$ & $h_2(x) = (2x + 1) \bmod 10$

Insert the elements 5 and 7 into the Bloom Filter.

Then determine whether the element 6 is definitely not present or possibly present.

Solution :

Given : Step 1 : Initial bit Array , The bit array has 10 positions

Position	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Bit	0	0	0	0	0	0	0	0	0	0

Step 2 : Insert element 5

Calculate both hash values : $h_1(5) = 5 \bmod 10 = 5$

$h_2(5) = (2(5) + 1) \bmod 10 = 11 \bmod 10 = 1$

Gateway Classes

So, set positions 5 and 1 to 1

Position	0	1	2	3	4	5	6	7	8	9
Bit	0	1	0	0	0	1	0	0	0	0

Step 3 : Insert element 7

$h_1(5) = 7 \bmod 10 = 5$

$h_2(5) = (2(7) + 1) \bmod 10 = 15 \bmod 10 = 1$

So, set positions 7 and 5 to 1

Position 5 is already 1 , so it remains 1

Gateway Classes

Position	0	1	2	3	4	5	6	7	8	9
Bit	0	1	0	0	0	1	0	1	0	0

Therefore , 0100010100

Step 4 : Query element 6

$h_1(6) = 6 \bmod 10 = 6$

$h_2(6) = (2(6) + 1) \bmod 10 = 13 \bmod 10 = 3$

Now, check position 6 and 3 , Position 6 = 0 & Position 3 = 0

Since at least one required bit is 0, element 6 is definitely not present.

GATEWAY CLASSES
GATEWAY CLASSES
Gateway Classes

PREVIOUS YEAR IMPORTANT QUESTIONS

Q.1 A data streams generates 120,000 records per minute. If only 35% data is retained using a decaying window , calculate the number of records retained per minute & justify the use of windowing . **AKTU(2025-26)**

Q.2 A sliding window of size 200 slides by 50 elements . Calculate the umber of windows formed after processing 1,000 elements and explain its impact on latency . **AKTU(2025-26)**

Q.3 Explain Alon – Matias – Szegedy algorithm for second moments on the given stream{ a, b, b, d , a ,c ,d ,a , b , d , c , a , a , b}. **AKTU(2022-23)**

Q.4 Explain and apply flajolet- Martin Algorithm on the following stream of data to identify unique elements in the stream. **AKTU(2024 - 25)**

S = 1,3,2,1,2,3,4,3,1,2,3,1

Given : h(x) = (6x + 1) mod 5

Gateway Classes

Gateway Classes

GW
GATEWAY CLASSES

Empowering Learners, Transforming Futures

THANK YOU

Gateway Classes

GET IT ON Google Play

Download on the App Store

Helpline No- 7819 0058 53 , 9259833200