情報システム実験

ガイダンス(K-4)

リアルタイムデータの処理と分析

天笠俊之教授、Savong Bou助教
計算科学研究センター
筑波大学

議題

- ガイダンス
- データストリームの基本
- 課題１の説明

担当教員・TA

- **教員：**
  - **天笠俊之**教授（計算科学研究センター）
  - **Savong Bou**助教（計算科学研究センター）

- **TA:**
  - **山﨑昂輔** (M2)

実験目的

- 基本的なリアルタイム処理と分析を行う方法

- リアルタイムで視覚化できるアプリケーション

Query OLAP Help

StreamingCube

OLAP MAP

KDE Lab

Operation AVG Time Grain Second

OLAP Operations

Rollup

Time Grain

Live Stats (lol)

Result Timestamp:

1536551723

Total QTY of PM2.5:

59.9363

Highest-QTY Dim:

辻堂新町３丁目,15365517:

QTY of PM2.5 at the Highest-QTY Dim:

44.0327

Least-QTY Dim:

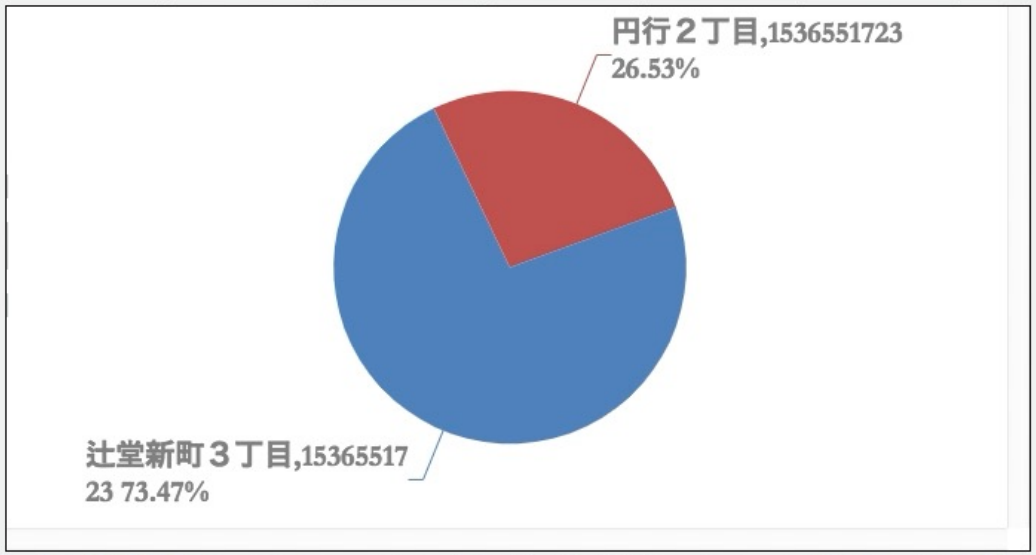
円行２丁目,1536551723

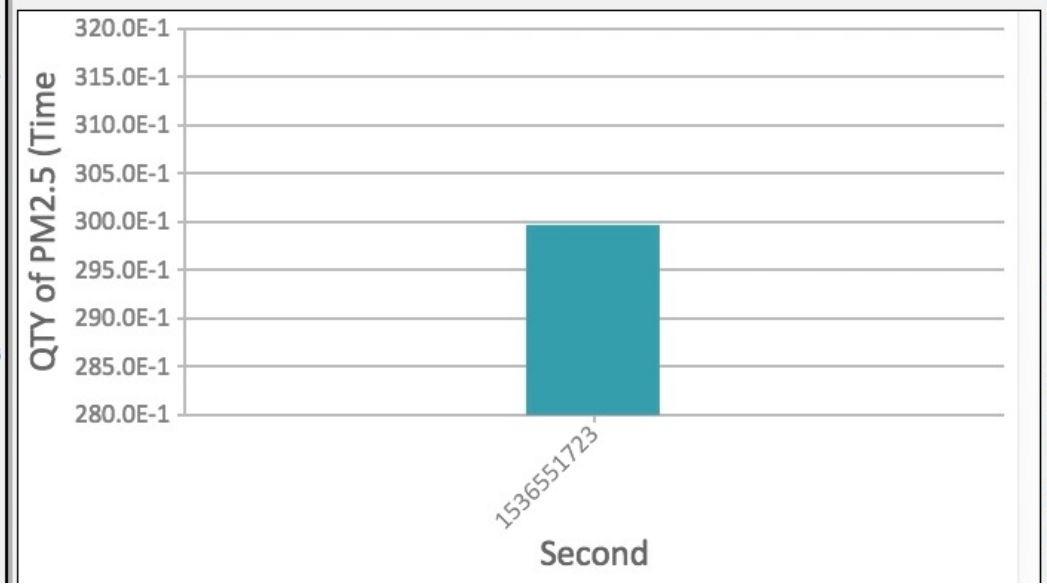
QTY of PM2.5 at the Least-QTY Dim:

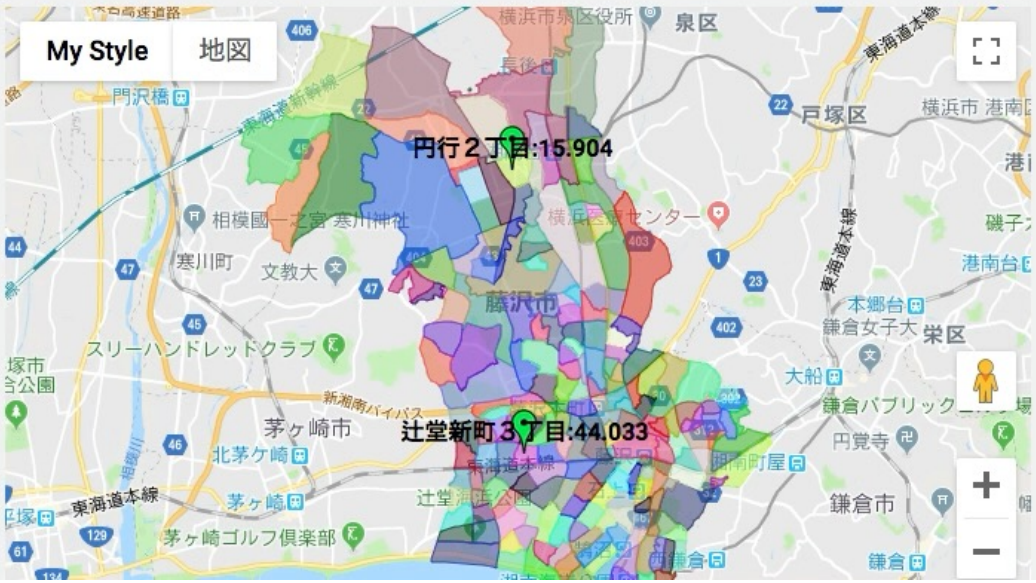
15.9036

chomeID	timestamp	value
辻堂新町３丁目	1536551723	44.0327
円行２丁目	1536551723	15.9036

Bar Column Area Spline Pie Doughnut







実験スケジュール・教室

- **春学期ABC：**
  - 水３・４限
  - 金５・６限

基本的にオンライン

ソフトウェアサイエンス実験 A／情報システム実験 A／知能情報メディア実験 A（春学期）

	4 月				5 月				
水曜(3・4 時限)	(17)	24			1	8	15	22	29
金曜(5・6 時限)		26				10	17	24	31
注意事項	4 月 17 日（水）12:15～ 実験ガイダンス（オンライン Zoom） 4 月 17 日（水）～19 日（金）17:00 テーマ選択・希望登録。 4 月 19 日（金） カリキュラム上の取り決めで休み。 4 月 23 日（火）テーマ選択結果発表 4 月 24 日（水）実験開始 5 月 1 日（水）金曜日の授業を実施								
	6 月				7 月				
水曜(3・4 時限)	5	12	19	26	3	10	17	24	31
金曜(5・6 時限)	7	14	21	28	5	12	19	26	
	8 月								
水曜(3・4 時限)									
金曜(5・6 時限)									

全日数 28 日（6 日以上欠席不可）

レポート提出締切： 8 月 7 日（水）17:00

最終発表

- 日程：
  - 7月31日水3・4限
- 現地：
  - 計算科学研究センター、会議室C

成績評価

- 全課題の提出するのは必須になります。一つの課題でも未提出場合は、成績はDになります。
  - 課題1：5%、
  - 課題2：10%、
  - 課題3：10%、
  - 課題4：15%、
  - 課題5：15%、
  - 課題6（最終発表・レポート）：45%

出席

- 毎回TAさんに進捗を報告
- 出席できない場合は、事前に連絡
- 欠席多くと、成績に影響があるので、気をつけてください

データストリームの基礎

Savong Bou
計算科学研究センター
筑波大学

Content

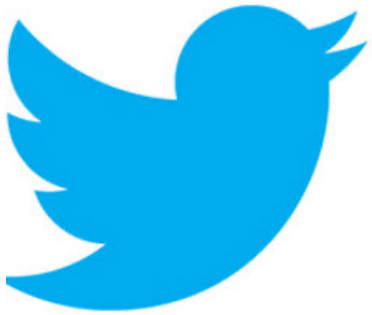
- データストリームの概要
- データストリームモデル
- ストリーム処理
  - Windowの種類
- ストリーム処理フレームワーク
  - Storm, Triden, Spark, Samza, Flink

Icons: https://www.flaticon.com/

データストリーム

- リアルタイムデータ、ライブデータ

ソーシャルメディア



IoTデバイス

データストリーム

Commodity	Month	Last	High	Low	Chg.	Chg. %	Time
Gold	Dec 22	1,736.95	1,737.95	1,732.00	+0.65	+0.04%	01:33:14
XAU/USD		1,725.45	1,726.61	1,720.68	+1.13	+0.07%	01:33:17
Silver	Sep 22	18.240	18.258	18.150	+0.081	+0.45%	01:33:14
XAG/USD		18.529	18.529	18.377	+0.093	+0.50%	01:33:17
Copper	Dec 22	3.5873	3.5880	3.5505	+0.0298	+0.84%	01:33:18
Platinum	Oct 22	841.10	841.40	830.80	+8.30	+1.00%	01:33:14
Palladium	Dec 22	2,125.77	2,126.27	2,073.77	+37.50	+1.80%	01:33:02
Crude Oil WTI	Oct 22	92.72	92.73	91.84	+1.08	+1.18%	01:33:17
Brent Oil	Nov 22	98.98	99.00	98.11	+1.14	+1.17%	01:33:15
Natural Gas	Oct 22	9.106	9.122	9.020	+0.039	+0.43%	01:32:13
Heating Oil	Oct 22	3.8186	3.8209	3.7827	+0.0332	+0.88%	01:33:06
Gasoline RBOB	Oct 22	2.5586	2.5642	2.5316	+0.0283	+1.12%	01:33:06
London Gas Oil	Sep 22	1,150.12	1,151.00	1,138.25	+10.62	+0.93%	01:31:59
Aluminium		2,390.00	2,443.00	2,387.00	-105.50	-4.23%	13:30:02
Zinc		3,475.50	3,546.50	3,469.50	-86.00	-2.41%	30/08
Nickel		21,416.50	21,784.50	20,800.50	-279.00	-1.29%	30/08
Copper		7,862.00	7,966.00	7,838.00	-275.50	-3.39%	30/08
US Wheat	Dec 22	827.00	827.12	814.25	+7.00	+0.85%	01:33:06
Rough Rice	Nov 22	17.783	17.790	17.743	+0.010	+0.06%	21:46:16
US Corn	Dec 22	680.38	680.38	674.62	+1.88	+0.28%	01:33:11
US Soybeans	Nov 22	1,440.50	1,441.00	1,422.50	+6.25	+0.44%	01:33:18
US Soybean Oil	Dec 22	67.23	67.25	66.22	+0.75	+1.13%	01:33:18
US Soybean Meal	Dec 22	424.30	424.60	421.15	+0.20	+0.05%	01:33:18

リアルタイムの商品価格

https://www.investing.com/commodities/real-time-futures

データストリーム

MARKET		DAILY % CHANGE	SENTIMENT
EUR/USD	1.00430 / 1.00436	+0.29%	BULLISH
USD/JPY	138.443 / 138.453	-0.25%	BULLISH
AUD/USD	0.68869 / 0.68878	+0.49%	BEARISH
GBP/USD	1.16886 / 1.16901	+0.29%	BEARISH
USD/CAD	1.30678 / 1.30699	-0.18%	BULLISH
USD/HKD	7.84769 / 7.84819	-0.01%	
NZD/USD	0.61477 / 0.61495	+0.32%	MIXED
AUD/CAD	0.89991 / 0.90031	+0.31%	
AUD/JPY	95.343 / 95.363	+0.24%	BEARISH
CAD/JPY	105.917 / 105.957	-0.07%	

外国為替
レートの監視

Forex Rates:
https://www.dailyfx.com/forex-rates#currencies

データストリーム

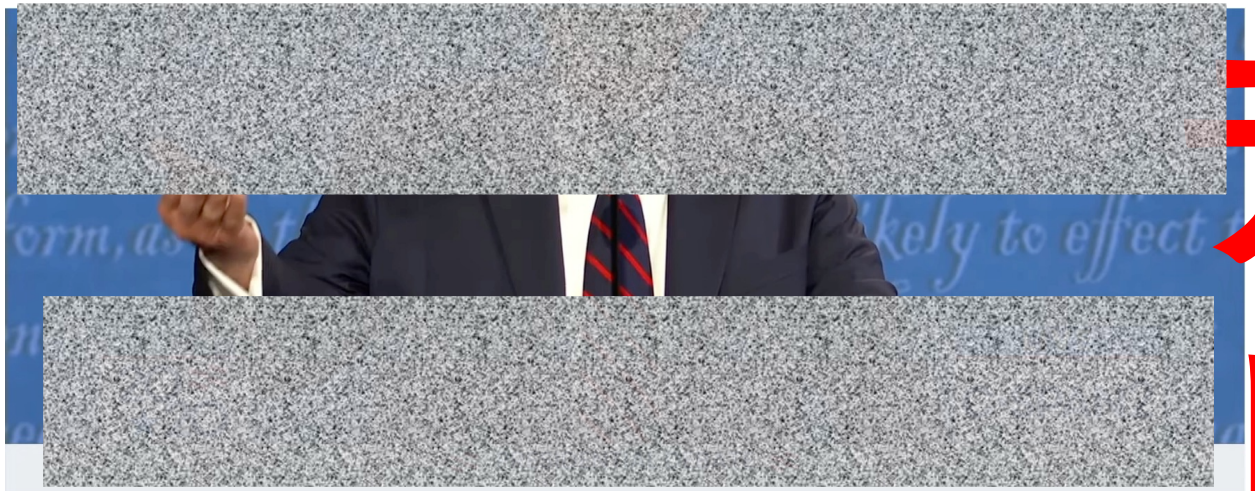
ビデオストリーム

LIVE

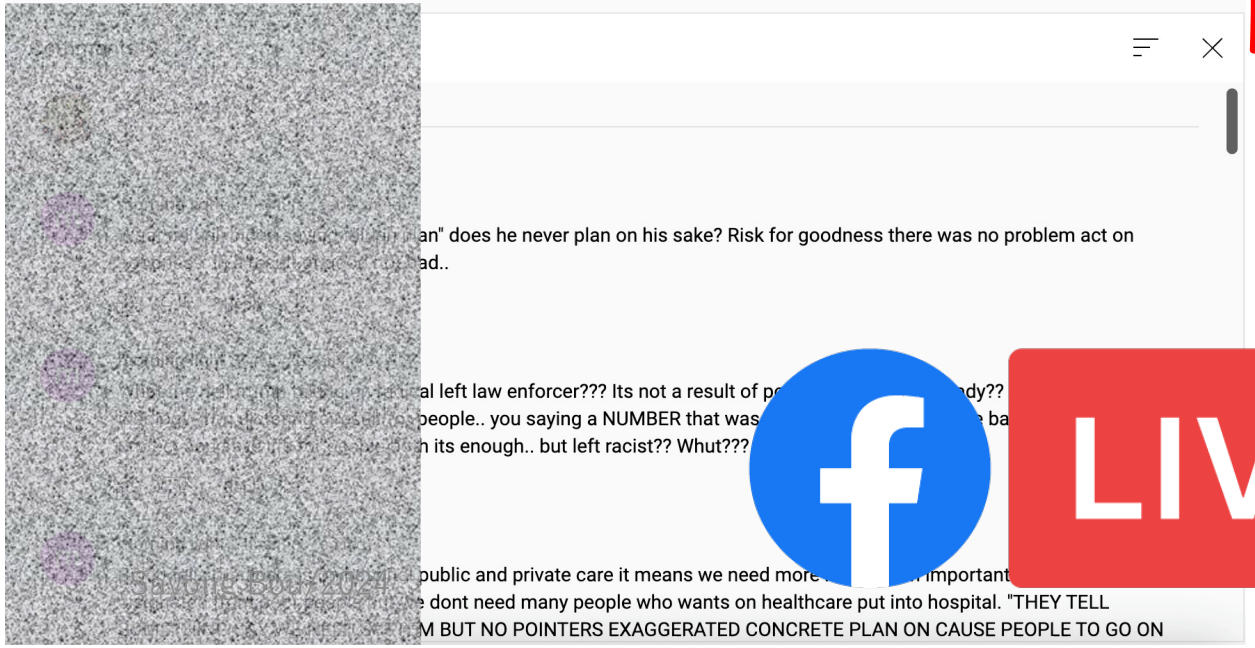
YouTube LIVE

Skynews: https://www.youtube.com/watch?v=9Auq

データストリーム



テキストストリーム



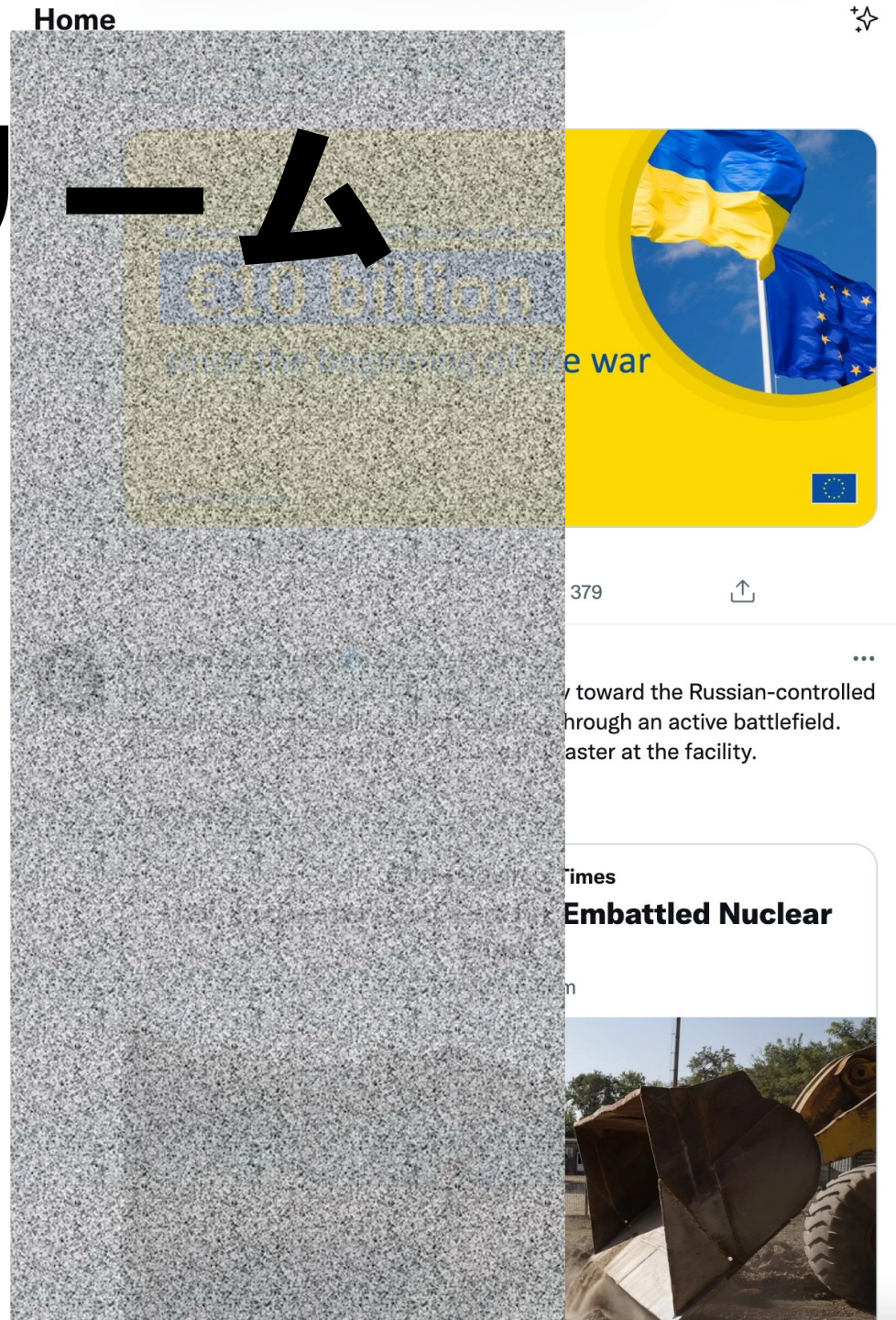
First 2020 Presidential Debate between Donald Trump and Joe Biden
https://www.youtube.com/watch?v=wW1lY5jFNcQ&t=20s

LIVE

YouTube LIVE

データストリーム

ツイート
ストリーム



Icons: https://www.flaticon.com/

データストリームの応用

- リアルタイムの結果が必要なアプリケーション

交通管制システム

防災システム

天気予報

Twitterのトレンド検出

株価予測

Icons: https://www.flaticon.com/

ストリーム処理モデル

時間

クエリ

, 23, 45, 345, 56, 1, 2, 3, 54

, o, a, s, s, c, l, a ,y ,g ,k ,u

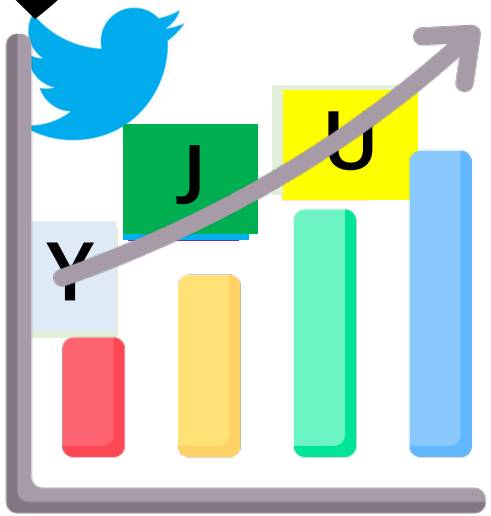
, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1

データストリーム

プロセッサー

限られたメモリ

アーカイブストレージ



142.87

出力

データストリームモデル

- 定義:
  - $X_i = (V_i, t_i)$
    - t_i:タイムスタンプ
    - $V_i = (v_i^1, v_i^2, \ldots, v_i^d)$:
      - 多次元のレコード

Example

V_i t_i

(ミルク, 卵, 380円)$_{00:00:01}$
(砂糖, 豚肉, 905円)$_{00:00:02}$

Icons: https://www.flaticon.com/

ストリーム処理モデル

時間

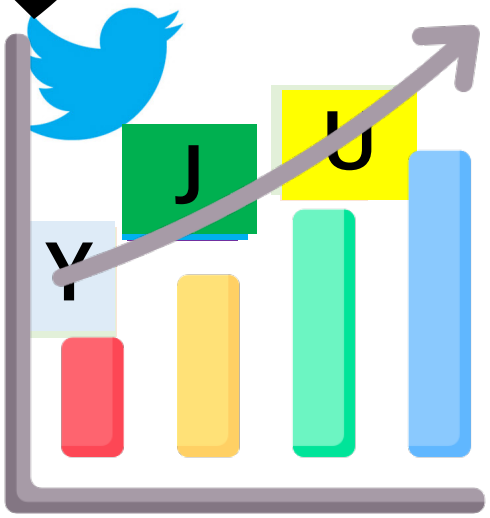
クエリ

, 23, 45, 345, 56, 1, 2, 3, 54

, o, a, s, s, c, l, a ,y ,g ,k ,u

, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1

プロセッサー



142.87

デ

ス

リアルタイム入・出力

全体データ保存できない

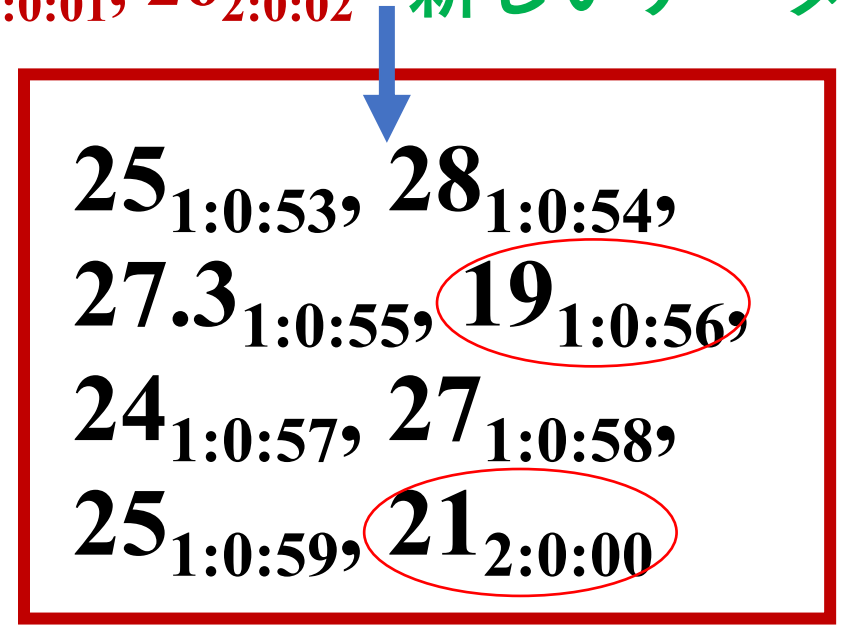
ストリーム処理

- 限られたメモリで無限のストリームに対処しますか?
  - サンプリング:
    - 古い要素を置き換える確率
  - Window:
    - 無限のストリームを有限のバッチに分割

サンプリングアプローチ

- 確率関数によって利用可能なメモリを適合させる
  - 既存の要素を新しい要素で置き換える

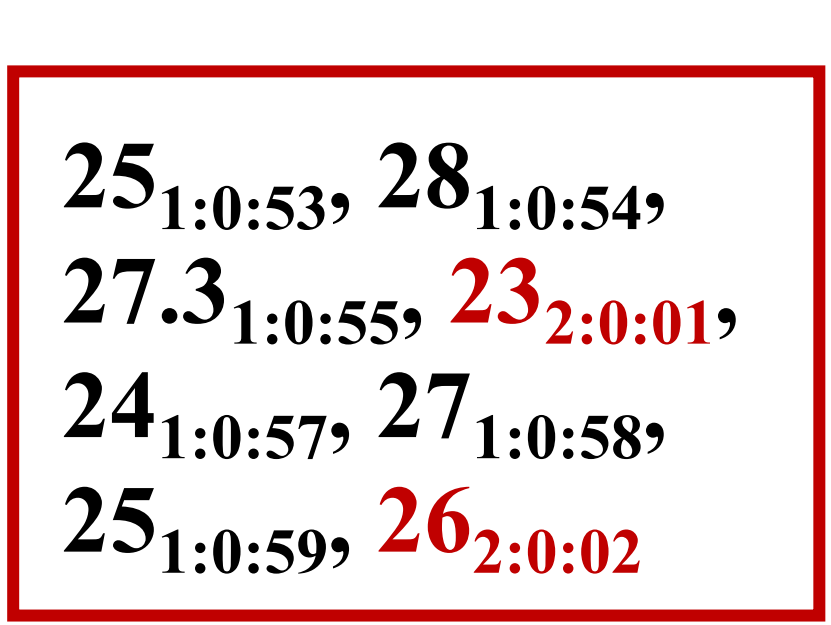
$23_{2:0:01}$, $26_{2:0:02}$ 新しいデータストリーム



使用可能なメモリ

確率関数

i.e.,最小の要素を置き換える



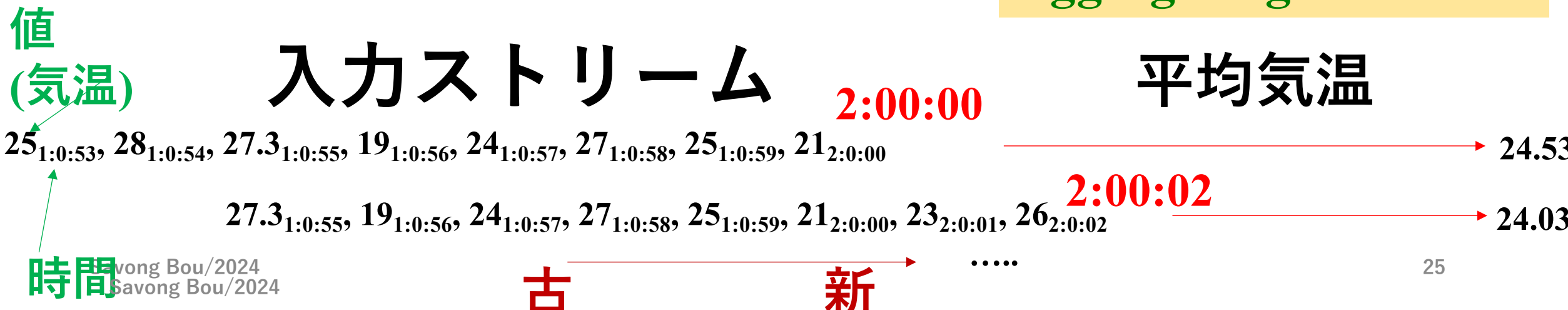
使用可能なメモリ

Windowアプローチ

- Window: 無限ストリームを計算が実行される有限バッチに分割します。

過去 8 秒間の 2 秒ごとのリアルタイム平均気温

Window: 8s
Slide: *2s*
Aggregating: *AVG*



Savong Bou/2024

Window in Apache Flink: https://medium.com/big-data-processing/windowing-in-apache-flink-b12d220cbd9a

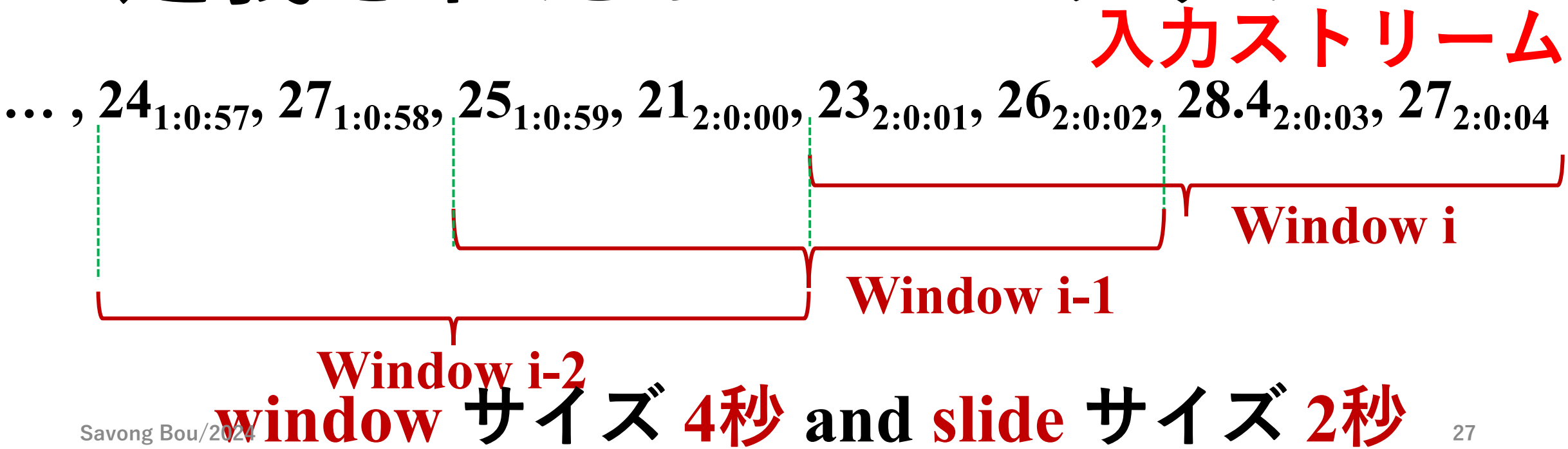
Window種類

- Tumbling Window
- Sliding Window
- Session Window
- Global Window
- User-defined Window

Window in Apache Flink: https://medium.com/big-data-processing/windowing-in-apache-flink-b12d220cbd9a

Sliding Window

- WindowとSlideのサイズで定義されたオーバーラップ



Window・Slide サイズ仕様

- Count-based window（数）
- Time-based window（時間）

新しく到着した 2 つのレコードごとに、最後に確認された 8 つのレコードにわたるリアルタイムの平均気温

過去 8 秒間の 2 秒ごとのリアルタイム平均気温

Window: *8 records*
Slide: *2 records*
Aggregating: *AVG*

Window: *8 seconds*
Slide: *2 seconds*
Aggregating: *AVG*

Window・Slide サイズ仕様

入力ストリーム

現在の時刻

2:00:04

… , $25_{1:0:57}$, $28_{1:0:57}$, $27.3_{1:0:58}$, $19_{1:0:58}$, $24_{1:0:59}$, $27_{1:0:59}$, $25_{1:0:59}$, $21_{2:0:00}$, $23_{2:0:01}$, $26_{2:0:02}$, $28.4_{2:0:03}$, $27_{2:0:04}$

現在の Window

Count-based window：window サイズ 8件のレコード

現在の時刻

2:00:04

… , $25_{1:0:57}$, $28_{1:0:57}$, $27.3_{1:0:58}$, $19_{1:0:58}$, $24_{1:0:59}$, $27_{1:0:59}$, $25_{1:0:59}$, $21_{2:0:00}$, $23_{2:0:01}$, $26_{2:0:02}$, $28.4_{2:0:03}$, $27_{2:0:04}$

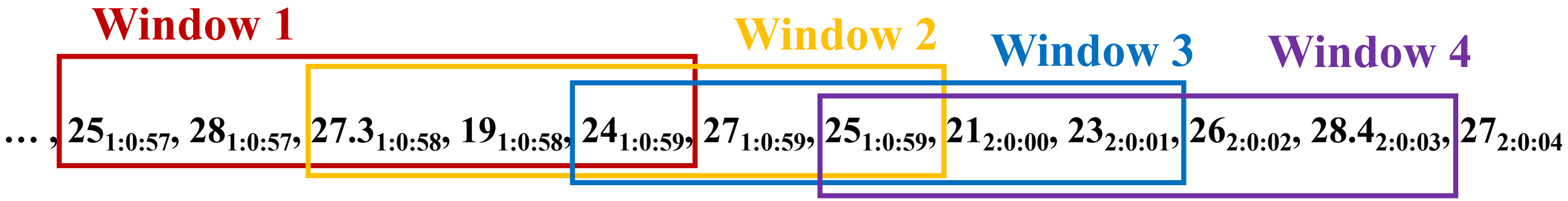
現在の Window

Time-based window：window サイズ 8秒

Count-based window

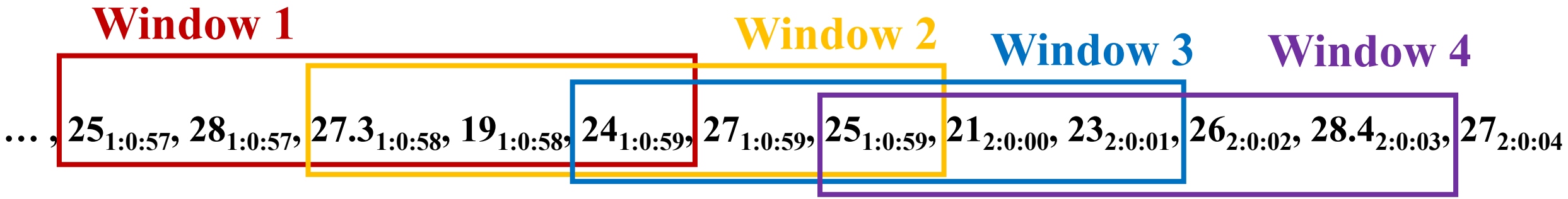
- WindowとSlideのサイズは両方ともレコード数によって定義

Window: *5 records*
Slide: *2 records*
Aggregating: *AVG*



Count-based window

Window: *5 records*
Slide: *2 records*
Aggregating: *AVG*



At time 1:0:59

$25_{1:0:57}, 28_{1:0:57}, 27.3_{1:0:58}, 19_{1:0:58}, 24_{1:0:59}$ → **Result = avg (25, 28, 27.3, 19, 24) = 24.66**

At time 1:0:59

$27.3_{1:0:58}, 19_{1:0:58}, 24_{1:0:59}, 27_{1:0:59}, 25_{1:0:59}$ → **Result = avg (27.3, 19, 24, 27, 25) = 24.46**

At time 2:0:01

$24_{1:0:59}, 27_{1:0:59}, 25_{1:0:59}, 21_{2:0:00}, 23_{2:0:01}$ → **Result = avg (24, 27, 25, 21, 23) = 24**

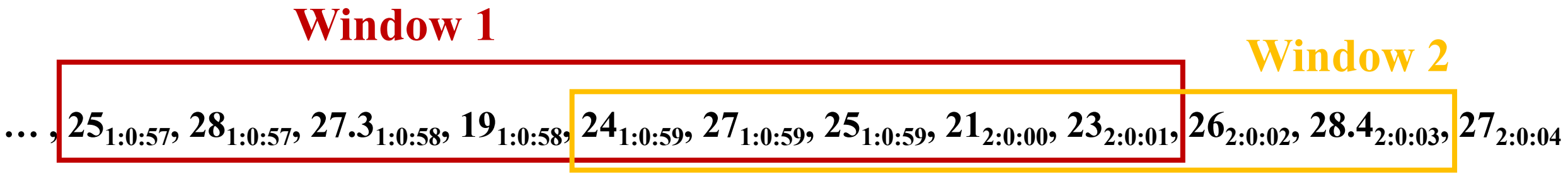
集計はリアルタイムで行われます。

……..

Time-based window

- WindowとSlideのサイズは両方とも時間によって定義

Window: *5 秒*
Slide: *2 秒*
Aggregating: *AVG*



Time-based window

Window: 5 秒
Slide: 2 秒
Aggregating: *AVG*

Window 1 **Window 2**

… , $25_{1:0:57}$, $28_{1:0:57}$, $27.3_{1:0:58}$, $19_{1:0:58}$, $24_{1:0:59}$, $27_{1:0:59}$, $25_{1:0:59}$, $21_{2:0:00}$, $23_{2:0:01}$, $26_{2:0:02}$, $28.4_{2:0:03}$, $27_{2:0:04}$

At time 2:0:01

$25_{1:0:57}$, $28_{1:0:57}$, $27.3_{1:0:58}$, $19_{1:0:58}$, $24_{1:0:59}$, $27_{1:0:59}$, $25_{1:0:59}$, $21_{2:0:00}$, $23_{2:0:01}$ → **Result =**

avg (25, 28, 27.3, 19, 24, 27, 25, 21, 23) = 24.36

集計はリアルタイムで行われます。

At time 2:0:03

$24_{1:0:59}$, $27_{1:0:59}$, $25_{1:0:59}$, $21_{2:0:00}$, $23_{2:0:01}$, $26_{2:0:02}$, $28.4_{2:0:03}$ →

Result = avg (24, 27, 25, 21, 23, 26, 28.4) = 24.91

……..

ストリーム処理フレームワーク

Spark Streaming

Apache Flink

Apache Storm

Apache TRIDEN

samza

結論

- データストリームの概要
- データストリームモデル
- ストリーム処理の基本概念
  - Window種類
  - 時間の概念
- ストリーム処理フレームワーク
  - Storm, Triden, Spark, Samza, Flink