***Review Article***

# Streaming Data Processing


## ABSTRACT

**Aims:** The data which is continuously being produced by hundreds of thousands of data sources is recognized as streamed data. The data which is processed via this kind of source is relatively smaller in size and is being sent at the same time it is generated.
**Study design:** In streaming data, the data range is so wide like the telemetry from interconnected devices or other such forms of data with the inclusion of certain web applications. This information should be handled consecutively and steadily on a record-by-record premise or throughout sliding time windows and utilized for a wide assortment of examinations including relationships, totals, separating, and inspecting.
**Place and Duration of Study:** Service usage (for metering and billing), server activity, website clicks, and the geo-location of devices, people, and physical goods are just a few of the many aspects of a company's business and customer activity that can be seen through this type of analysis. It also enables companies to respond quickly to new arising situations.
**Methodology:** The research methodology being used for the current research work is the qualitative method through which the research studies of a similar domain are studied thoroughly. It has been analyzed that the specified changes in the large volumes of data can better be managed through stream data processing.
**Results:** The flaws of batch data processing are better dealt with through the usage of streaming data processing agenda. Real-time monitoring as well as response functionality are the keys to success in the given method of data processing.
**Conclusion:** Stream data processing connects analytics and applications. Because multiple systems can be constructed using the same architecture, this makes the construction of the infrastructure as a similar architecture. It additionally allows designers to fabricate applications that utilize scientific outcomes to straightforwardly answer information experiences and make a move.



*Keywords: Streaming data, processing, qualitative, real-time, monitoring*


## 1. INTRODUCTION

Stream data processing is the process of acting related to a sequence of data, simultaneously at the time the data is created. In the past, data professionals talked about "real-time processing" as a general term for data is found to be processed just at the time it was needed for a particular use-case. However, the term "stream processing" is now used in a more specific sense due to the emergence and widespread use of stream processing frameworks and technologies, as well as the decreasing costs of RAM. The agenda of stream processing frequently involves numerous and varying tasks concerning the

approaching series of information (i.e., the "information stream"), which can be performed in parallel, sequentially, or both. Not only the generation of stream data but also, it's processing, as well as the delivery to a final location, are all parts of this workflow. The underlying agenda is recognized as a stream-processing pipeline.

Applications can respond immediately to new data events based on the agenda of stream processing. The stream processing engine processes the input data pipeline in real-time in this simplified example. A streaming analytics application receives the output data and incorporates it into the output stream. Analytics (like predicting the future based on certain past events), transformations (like the alteration of a specified number to the date format after performing certain action), enrichment (i.e., a combination of the data point with other data sources so may the useful & meaningful outcomes can be generated), as well as ingestion (like the data insertion into a database) are some of the actions that stream processing takes on data [1].

In the past, data has been processed typically in the form of batches according to a schedule or a predetermined threshold (e.g., every hundred rows, every two megabytes, or every night at one in the morning). However, because of the increased speed and volume of data, batch processing is no longer sufficient for many different kinds of use cases. Stream processing is now required by certain modern applications. The business entities have adopted ways that can better help them to deal with their data effectively i.e., along with the consideration of the specified use cases. The new data events are better recognized and responded to through the stream data processing agenda. Unlike the process of stream data processing, the phenomenon of batch processing considered grouping and the gathering of data at a predetermined level. While the agenda of stream data processing considers gathering and processing the data at the time the data is being generated.

On the other hand, is the concern of talking about the use cases that tend to effectively support the agenda of stream data processing. Most of the time, use cases involve event data that comes from some action and should have some action done right away. Examples of common applications for real-time stream data processing are as given below:

- The real-time detection of fraud and anomalies: One of the largest credit card issuers in the world has been able to cut their annual fraud write-downs by $800 million just due to stream data processing-powered fraud and anomaly detection. Both the store trying to process the credit card and the customer (along with any other customers in line) suffer from delays in credit card processing. Post-exchange, Visa organizations used to play out their tedious misrepresentation identification methodology in bunches. The said system has greater strengths for running complex algorithms so may the fraudulent charges can better be identified & blocked as soon as you swipe your card because they use stream data processing. They are also able to send alerts for unusual charges that require further investigation without making their (non-fraudulent) clients wait.
- Personalization, advertising, and marketing in real-time. Companies can provide customers with contextually relevant, personalized experiences through the use of real-time stream data processing. This could include a discount related to something that you have added to your cart on a website but didn't buy right away. Also, you might have given a suggestion to connect with a friend you just registered on a social media site or an advertisement for a similar product.
- Edge analytics for the Internet of Things (IoT). Stream processing is used by businesses in manufacturing, oil & gas, transportation, and the design of smart cities and buildings to keep up with data from billions of "things." An illustration of IoT information examination is identifying the issues in assembling that demonstrate the need for knowledge about the anomalies to sort out to further develop tasks and increment yields. With constant stream handling, a maker might perceive that a creation line is turning out an excessive number of irregularities as it is happening (instead of tracking down a whole terrible group after the day's shift). By stopping the line for immediate repairs, they can identify significant savings and avoid significant waste.

## 2. RELATED WORKS

One remarkable component of business entities is data collection. Businesses in the modern economy simply cannot hold to wait for the specified data to be processed in the form of batches. Instead, real-time event streams are used by ride-sharing apps, e-commerce websites, platforms for the stock market, fraud detection, and other applications. Applications evolve to process, filter, analyze, and react to events as they occur in real time when paired with streaming data. This opens a new set of applications, including Netflix recommendations, and real-time fraud detection, along with the consideration of a continuous shopping experience that you choose to update as you shop across multiple devices. In a nutshell, platforms for continuous, real-time event stream processing can be beneficial to any sector that deals with significant volumes of real-time data [2].

Stream data processing is a term that is turning out to be progressively pertinent to the specialized side of any organization yet is still somewhat ambiguous on the business side. Data streams, also known as data, that continuously enters a business, are referred to by the term "Stream." Computer systems have been receiving data for decades; the only difference is that the volume and speed are now unprecedented. Onboarding, analysis, as well as the integration of the data in a way that gives technology users insight, i.e., preferably in real-time, is referred to as stream processing. With time, firms are developing, and thus the quantity of workers is expanding with the upgrade in the topographical area of business limits. In the end, this leads to more data for the businesses. which businesses must benefit from the most recent technology. Stream data processing is very helpful when it comes to securing the data for a variety of uses, such as forecasting business demands, planning, making decisions, having precise values for the financials of the company, reporting data, or analyzing the data to understand the level of demand from customers. The user can store a large amount of data according to their needs. Users can choose any kind of data warehouse that meets their needs [3].

Stream information processing design refers to the production of an information handling and storage-related framework for a specified business entity because for information to be useful, it needs to be properly sorted, cleaned, and arranged. Finding the most efficient method for organizing information from a direct set into a simple structure that yields useful BI insights is the goal of stream data processing. The majority of vendors and IT manufacturers are constantly making investments in developing energy-efficient computing devices and working to reduce hazardous and harmful materials. Numerous producers are likewise attempting to empower the recyclability of these computerized gadgets. Green computing methods gained prominence in 1992. The Environmental Protection Agency launched the Energy Star program at that time.

There are billions of cell phones, possibly trillions of IoT gadgets, every one of them streaming information into frameworks that are hustling to stay aware of volumes that are expanding dramatically. You are a part of a stream that is dealing with this, as are all of the businesses you interact with on a daily (or even multiple daily) basis. It is a possibility that a billion people might be doing the same thing at the same time whenever you upload something to Facebook, tweet, or complete an online transaction. Whether we are aware of it or not, we are all a part of the streaming ecosystem. We all need to have a mechanism that can effectively help us to determine the direction in which our data is being gathered as well as assembled. As the data keeps on growing so we cannot evaluate the grounds to segregate the data as well as the anomalies [4].

The better inquiry is, how significant as well as valuable is it for you to know immediately how your business is getting along? Consider commodity trading in real time; A fraction of a second can mean the difference between making or losing millions of dollars. E-commerce, financial, healthcare, and security transactions all require immediate stream data processing responses. The problem is that businesses need to be able to 1) in a blizzard, identification

of the relevant snowflake, and 2) take meaningful and immediate action when something has happened. Immediacy is perceived to be remarkable because the majority of data is highly perishable with a shelf life that can be effectively measured in microseconds [5].
Stream data processing has emerged as an urgent requirement for every organization due to the rapid growth and development of technology. The following are the primary reasons why it is required:

- The stream data processing method provides improved business intelligence;
- It saves time;
- It improves the consistency and quality of the data;
- The data warehouse also provides historical intelligence;
- It offers a high return on investment.

Six main components are included in the stream data processing architecture:

- The data warehouse's database: The entirety of the data is managed and stored in the database section for reporting purposes. The data set can be chosen by the prerequisites out of the ordinary social data set, insightful data set, information distribution center application and cloud-based data set.
- Instruments for extraction, the change of information, and stacking: Because they are used to collect data from various sources, transform it, and load it into the warehouse, these are the central component of the architecture.
- Metadata: This framework of data includes both technical and business metadata.
- Access methods to the data warehouse: This architecture includes tools for application development, data mining, OLAP, query-related tools, and more.
- Bus for a data warehouse: The data warehouse bus specifies the data flow within a data warehouse. A data mart is included in this.
- Layer that reports: The reporting layer grants the end user access to the BI interface for reporting purposes [6].

Because it is difficult for a number of business applications to locate these actual statistics, they require enormous amounts of formless data. Companies have acknowledged issues like the difficulty of effectively utilizing data, particularly in light of the rise of novel data resources, the demand for fresh data, and the need to increase processing speed [7].

## 3. METHODOLOGY

The theoretical and methodical examination of the specific approaches typically taken in the research area is referred to as the methodology. In addition, it covers the fundamentals and provides a theoretical analysis of the body methods associated with the particular field of knowledge. In most cases, it includes particular ideas like; the theoretical model, paradigms, varying phases, and quantitative and qualitative techniques, among others. The methodology's terms have not been designed to provide study-related solutions. As a result, it is impossible to claim that all kinds of research employ the same approaches.
The topic and type of the study are always taken into consideration when designing the study. It provides theoretical support for determining which method or set of methods can be demonstrated to be the most effective for a particular case; for example, in order to extract the various results. The systematic techniques that are utilized in the research are also referred to as the process of the research design. Simply, the research design can investigate the research guide and the procedure for conducting the research. The current research work considers the agenda of qualitative basis as the observation method has better helped to collect the data from already existing research studies in a similar context [8].
It is one of the most important and well-known methods for the underlying study. The most significant and remarkable attribute of the qualitative method is that it is the most significant and notable procedure for small samples in the meantime its outcomes and results can be effectively quantifiable and measurable. It can easily provide a comprehensive description

and analysis of the research topic without taking into account the responses of participants. In addition to examining the study's scope, it can also examine the study's procedure (figure 1). Consequently, the various abilities and skills of the researchers are the foundation of qualitative research and its effectiveness, despite the fact that the results cannot be considered reliable. Interpretations and personal judgments provide the qualitative research method's data (figure 2) [9].

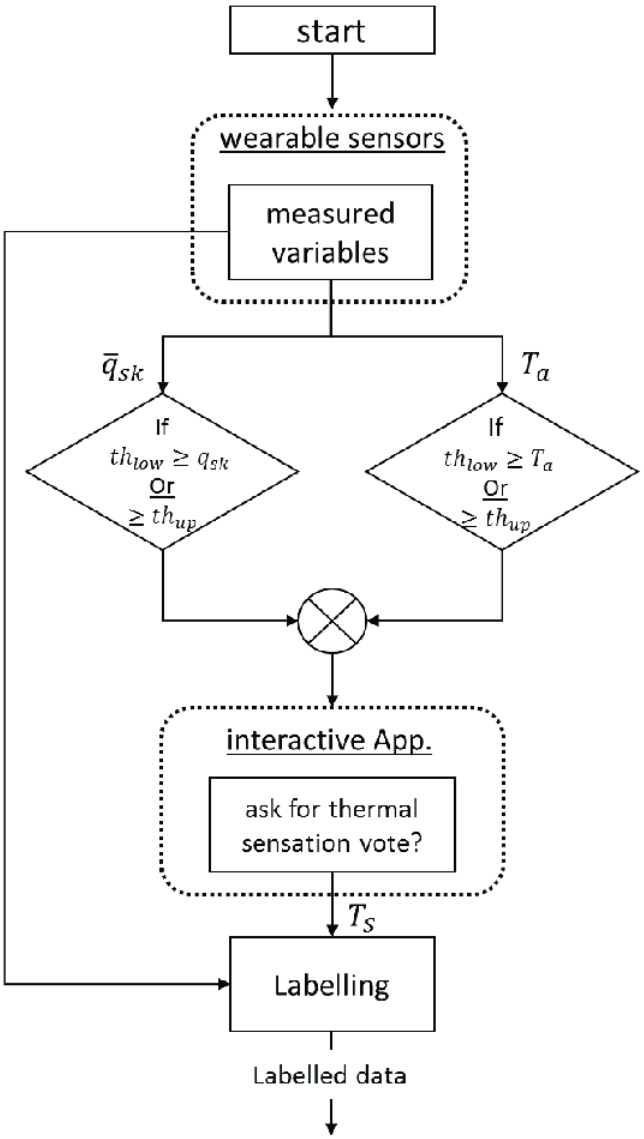


**Fig. 1. Workflow of stream data processing**

```
Given:
    GDB, // geographic database
     ❄, // geographic ontology
     T, // target feature type
     S, // set of relevant feature types O
     R; // set of all topological relationships
Variables:
      D; // relationships to compute for Data mining
Find: a dataset Ψ without geographic dependences between T and S;

Method:
Dependence_Elimination
Begin
    Ψ = T – geometry column;
    For (i=1; i=#O in S, i++) do
    Begin
      Find T in ❄;
          If (T has a one-one or one-many property with Oi in ❄)
              Remove Oi from S;   // dependence elimination
          Else
              If (T has prohibited properties P with Oi in ❄)
                      D = R – P; // possible relationships to compute
              Else
                      D = R // all topological relationships
          Ψ = Ψ + Spatial_Join (D,T,Oi);//computes spatial relationships D between T and O
     End;
End;
Transformation (Ψ) // transforms the resultant dataset into the data mining algorithm
                   // format preserving the non-spatial attributes of T;
```

**Fig. 2. Pseudo Code for Streaming data processin**

Two layers are needed for processing streaming data: a layer for processing and a layer for storage. In order to enable fast, inexpensive, and replayable reads and writes of large streams of data, the storage layer needs to be able to support record ordering and strong consistency. The handling layer is answerable for consuming information from the capacity layer, running calculations on that information, and afterward advising the capacity layer to erase information that is not generally required. Scalability, data durability, and fault tolerance in both the storage and processing layers must also be planned for [10].

## 4. RESULTS AND DISCUSSION

The ability to sort based on the noise to locate a useful signal is where streaming technologies really come into their own. In other words, how do you gain insight from such a vast and varied stream of data? Stream processing truly adds value in this area while having the ability to query a continuous data stream and identify data anomalies quickly enough to take action is a significant competitive advantage. For example, the sensors based on bioinformatics that can remotely report changes in a patient's condition in real-time, tend to allow medical professionals to respond quickly and accurately. We can better consider the volume of information that circulates throughout a large hospital at any given time. Streaming can identify the one crucial data point that has the potential to make a real difference for the patient. The businesses that are driving their initiatives are the ones that have sorted this out and applied it to their data. Consider Facebook as an earlier illustration. There are literally billions of individuals sharing everything they know about ecosystem vendors. When all a potential customer needs are a little push to turn toward you, how can you, as a vendor, entice them? Millions of people will simultaneously order the next must-have item from major electronic device manufacturers and receive a confirmation within seconds, where the manufacturer has checked inventory, location, shipping options, taxes, previous purchase history, service contracts, trade-in value, and other factors. all of this took place in a split second. The broad volumes of the information must be overseen actually on

the off chance that a legitimate arrangement exists to deal with this information and make the predefined moves of streaming something similar [11]. Stream processors are the systems that execute the application or analytics logic while also receiving and sending data streams. A stream processor's primary duties are to ensure that the computation scales are fault-tolerant while also ensuring that data flows smoothly. These problems can be solved by Apache Flink, a robust open-source stream processing framework. The stream data handling worldview normally addresses many difficulties that designers of constant information examination and occasion-driven applications face today:

• Analytics and applications respond immediately to events: There is no delay between "event occurs," "insight derived," and "action taken." The data is reflected in the most recent actions and analytics, when it is still relevant, valuable, and fresh.

• In comparison to other data processing systems, stream processing is able to handle significantly larger data volumes: Only a small portion of the data that is meaningful is saved after the event streams are directly processed.

• The continuous and timely nature of the majority of data is easily and naturally modeled by stream processing: Scheduled (batch) queries and analytics on static or resting data are different from this. The stream processing model is naturally compatible with incremental computation of updates as opposed to periodic re-computation of all data.

• The infrastructure is separated and decentralized through stream processing: Shared databases, which are large and costly, are less necessary with the streaming paradigm. The stream-processing framework, on the other hand, makes it simple for each stream-processing application to manage its own data and state [12].

Stateful stream data processing is a type of data stream processing where the computation keeps track of the context. This state is utilized to store data which is gathered from the currently happening occasions. Stateful stream processing is required by virtually all non-trivial stream processing applications. The most recent transactions for each state credit card would be kept by a fraud prevention application. The state is updated whenever a new transaction is compared to existing ones and classified as legitimate or fraudulent. The user's preferences would be described by parameters in an online recommender application. Every item association produces an occasion that refreshes these boundaries. Every time a user interacts with a song playlist or an e-commerce shopping cart, events are sent to a microservice (figure 3). The list of all added items is kept by the state [13].

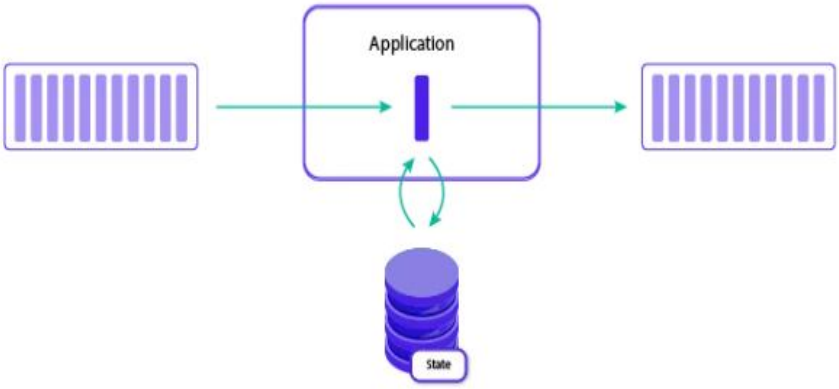


**Fig. 3. Stateful stream data processing**

Conceptually, stream data processing integrates the event-driven/reactive application or analytics logic with the database or key/value store tables into a single entity. High performance, scalability, data consistency, and ease of use are all achieved through the

application/analytics logic's close integration of state and execution. A stream processor that supports state management is required for stateful stream processing (figure 4).

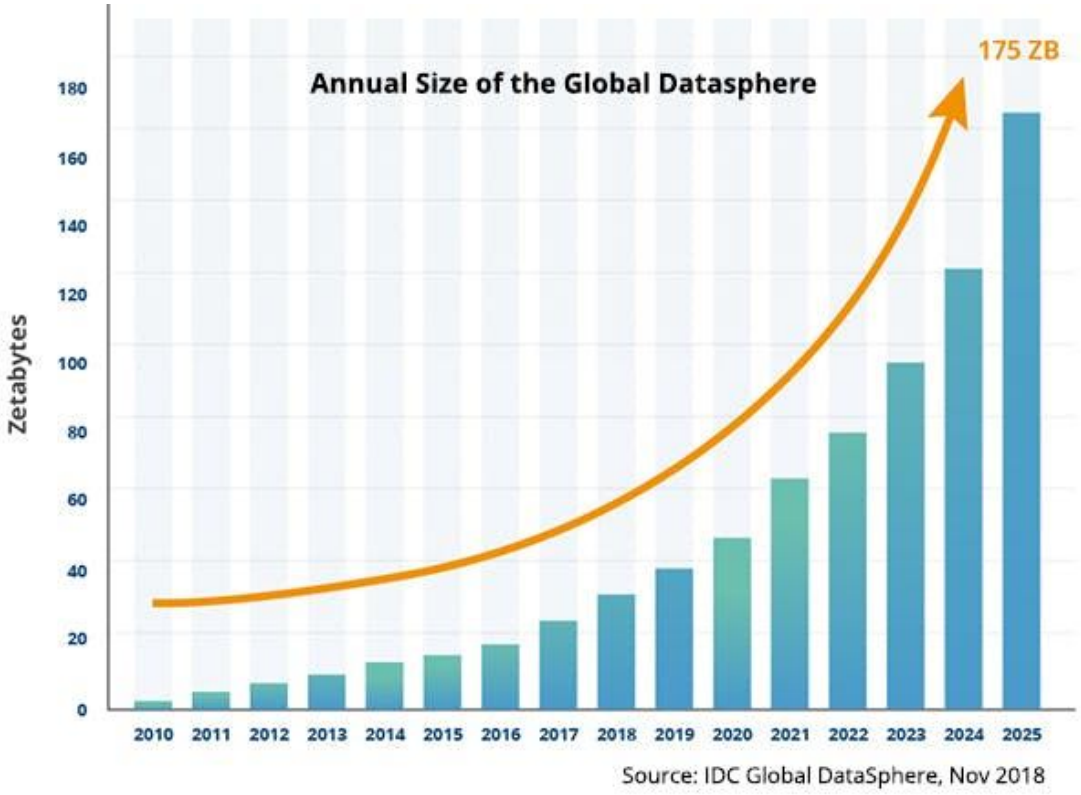


**Fig. 4. Stream data processing of global data**

## 4. CONCLUSION

Stream data processing connects analytics and applications. Because multiple systems can be constructed using the same architecture, this makes the construction of the infrastructure as a similar architecture. It additionally allows designers to fabricate applications that utilize scientific outcomes to straightforwardly answer information experiences and make a move. Some extraordinary examples include sending push notifications to users based on models of their behavior, adjusting a machine's parameters based on the results of real-time sensor data analysis, or using an analytical model to automatically block a fraudulent banking transaction.

Stream data processing involves a small number of components. The staging zone is where information from various operational frameworks is extracted, modified, and stacked. Here, it goes through various processes like profiling and standardizing. The information that has been "scrubbed" or "cleansed" is combined into a single structure that can be sent to the data warehouse for storage. The integration layer is where this job is finished.

The data is broken up into subsets, which are then moved into a variety of different data marts. Analytical processing is used to finish this job. These data marts are made to meet the needs of the user so that they can use the particular data for reporting. The main concern is the fact that how well the underlying agenda is being used by the business entities.

COMPETING INTERESTS DISCLAIMER:

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

## REFERENCES


1. Alwaisi SSA, Abbood MN, Jalil LF, Kasim S, Mohd Fudzee MF, Hadi R, et al. A review on big data stream processing applications: Contributions, benefits, and limitations. JOIV Int J Inform Vis [Internet]. 2021;5(4):456. Available from: http://dx.doi.org/10.30630/joiv.5.4.737
2. Amazon.com. [cited 2022 Dec 27]. Available from: https://aws.amazon.com/streaming-data/
3. Fatima N. Data warehouse architecture: Types, components, & concepts [Internet]. Astera. 2019 [cited 2022 Dec 27]. Available from: https://www.astera.com/type/blog/data-warehouse-architecture/
4. Burri M. Understanding the implications of big data and big data analytics for competition law: An attempt for a primer. In: New Developments in Competition Law and Economics. Cham: Springer International Publishing; 2019. p. 241–63.
5. Lmu.de. [cited 2022 Dec 27]. Available from: https://www.dbs.ifi.lmu.de/Lehre/BigData-Management&Analytics/WS15-16/Chapter-5_Stream_Processing_part1.pdf
6. 5 steps to a green data center [Internet]. Facilitiesnet. [cited 2022 Dec 27]. Available from: https://www.facilitiesnet.com/datacenters/contributed/5-Steps-to-a-Green-Data-Center--40635
7. Gupta PK, Ören T, Singh M. Predictive intelligence using big data and the internet of things. Gupta PK, OEren T, Singh M, editors. Hershey, PA: IGI Global; 2018.
8. Kolajo T, Daramola O, Adebiyi A. Big data stream analysis: a systematic literature review. J Big Data [Internet]. 2019;6(1). Available from: http://dx.doi.org/10.1186/s40537-019-0210-7
9. Ullah N. Lecture Notes for Big Data Analytics [Internet]. Edu.pk. 2019 [cited 2022 Dec 27]. Available from: https://web.lums.edu.pk/~imdad/pdfs/CS5312_Notes/CS5312_Notes-15-Data-Streams.pdf
10. Muthukrishnan S. Data streams: Algorithms and applications. Found Trends Theor Comput Sci [Internet]. 2005 [cited 2022 Dec 27];1(2):117–236. Available from: https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/Muthu-Survey.pdf
11. Lakey E. Top 5 benefits of a data warehouse for your data-driven organization [Internet]. The TIBCO Blog. TIBCO Software Inc.; 2022 [cited 2022 Dec 27]. Available from: https://www.tibco.com/blog/2011/08/10/how-to-sell-the-business-on-a-data-warehousing-project/
12. Walther T. Stream Processing beyond streaming data -Batch, Streaming, and Applications [Internet]. Datacouncil.ai. [cited 2022 Dec 27]. Available from: https://www.datacouncil.ai/hubfs/Data%20Council/slides/bcn19/Stream%20Proc%20Beyond%20Streaming%20Data.pdf

13. Zheng T, Chen G, Wang X, Chen C, Wang X, Luo S. Real-time intelligent big data processing: technology, platform, and applications. Sci China Inf Sci [Internet]. 2019;62(8). Available from: http://dx.doi.org/10.1007/s11432-018-9834-8т