Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 13-34
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Analytical Engineering for Data Stream

Rogério Rossi, Kechi Hirama

Digital and Computer Systems Engineering Department of Polytechnic School of the University of São Paulo, São Paulo, SP, Brazil
Email: rogeriorossi@usp.br, kechi.hirama@usp.br

How to cite this paper: Rossi, R. and Hirama, K. (2022) Analytical Engineering for Data Stream. *Journal of Computer and Communications*, **10**, 13-34.
https://doi.org/10.4236/jcc.2022.107002

Received: May 17, 2022
Accepted: July 15, 2022
Published: July 18, 2022

Abstract

The analytical capacity of massive data has become increasingly necessary, given the high volume of data that has been generated daily by different sources. The data sources are varied and can generate a huge amount of data, which can be processed in batch or stream settings. The stream setting corresponds to the treatment of a continuous sequence of data that arrives in real-time flow and needs to be processed in real-time. The models, tools, methods and algorithms for generating intelligence from data stream culminate in the approaches of Data Stream Mining and Data Stream Learning. The activities of such approaches can be organized and structured according to Engineering principles, thus allowing the principles of Analytical Engineering, or more specifically, Analytical Engineering for Data Stream (AEDS). Thus, this article presents the AEDS conceptual framework composed of four pillars (Data, Model, Tool, People) and three processes (Acquisition, Retention, Review). The definition of these pillars and processes is carried out based on the main components of data stream setting, corresponding to four pillars, and also on the necessity to operationalize the activities of an Analytical Organization (AO) in the use of AEDS four pillars, which determines the three proposed processes. The AEDS framework favors the projects carried out in an AO, that is, its Analytical Projects (AP), to favor the delivery of results, or Analytical Deliverables (AD), carried out by the Analytical Teams (AT) in order to provide intelligence from stream data.

Keywords

Analytical Engineering, Analytical Organization, Data Stream Analytics, Stream Mining

1. Introduction

The Big Data paradigm and the avalanche of massive data allow for improving decision-making capacity and increasing organizational efficiency [1]. Therefore, data analysis is presented as a fundamental activity in a Big Data Value Chain, as well it needs to be properly handled by organizations that seek to obtain results from data intelligence.

For [2], Big Data grows as more sectors of the globalized society are involved in the massive data era. The global economy, the management of society, scientific research, and even the security of nations can be treated as Big Data problems. Healthcare, public-sector administration, retail, a global manufacturing and personal data refer to other examples of massive data sources.

Big Data considers that massive data of many types (texts, geographic locations, human interactions) are a means for generating intelligence for organizations. Big Data comprises three main properties, generally the most cited: volume, variety and velocity [3]; many others are added, as an example, veracity [4]. However, just these three main properties of Big Data regarding data analysis activities have been determining many actions, investments and studies on the part of academia and the industry, imposing great demands on organizations that have the responsibility for offering value from data analysis.

The property regarding the high velocity of data generation and transmission presents a strict link to the concept of the data stream that can be observed in several applications, such as astronomical applications, scientific systems [5] [6], and systems that involve the Internet of Things paradigm, as well as financial, weather forecasting and telecommunication systems, among others.

For managing, manipulating and obtaining intelligence from the analysis of massive data considering the data stream setting, organizations have been adopting technologies, tools and models (computational, business, and data models) besides specific frameworks and techniques of data analysis. These vary according to their needs, capacities and computational conditions. This need for analyzing massive datasets establishes a new data paradigm called Modern Analytics, also known as Big Data Analytics [7], or just Analytics.

Analytics requires specific technological knowledge, considering specific tools, models and data analysis techniques in the Big Data domain. Its conceptual and practical approach has encouraged both academia and industry to provide better capacity to organizations, professionals and decision-makers. As an example, the Big - Data, Analytics, and Decisions Framework (B-DA), presented by [8], deals with a set of concepts, tools and technological solutions capable of providing a semantic capacity for Big Data and promoting better insights for users to make decisions.

The results observed in an organization, from the Analytics perspective, have to consider the data management lifecycle, which comprises the activities of obtaining data, storing data, analyzing data and visualizing the results. In this sense, it is possible to integrate engineering principles with Analytics activities to allow sufficient conditions for an organization to manage data, considering batch or stream setting approaches to generate the expected results.

Engineering per se is capable of providing solutions to open problems or those that have more than one solution; therefore, it can be presented as part of the solutions to the problems of Analytics, thus determining the proposal of Analytical Engineering (AE), capable of systematizing responses to the environment in which it operates under the aspects of (Big) Data.

As data streams have increasingly been present in different scenarios to generate knowledge from (massive) data, it is pertinent to explain/conceptualize Analytical Engineering from the data stream perspective, where data is generated at high velocity, requiring analysis to generate the results in real-time.

Therefore, some questions that permeate this research are: 1) How can the fundamentals of AE support organizations in their data stream analytical projects? 2) How should AE fundamentals and processes behave in relation to data stream? 3) How to handle AE processes in an organization that manages data in a data stream setting? The questions outlining this research work support the presentation of results that deal with Analytical Engineering for Data Stream (AEDS).

To present the results of this research, the article is organized into five sections: section two presents a detailed view of the main aspects of Analytical Engineering and its fundamental characteristics; section three highlights the fundamentals of the data stream, its concepts and applications; section four presents an integrated framework of AEDS, emphasizing the application of Analytical Engineering to data analytics projects that enable data stream; lastly, section five presents the final considerations and conclusions.

2. Literature Review

2.1. Analytical Engineering Principles

Big Data and its three main properties, or those most cited, concerning volume, variety and velocity [9], represent the amount of massive and varied data that is constantly generated by many sources. Other properties have been added to the paradigm, such as "veracity" [4] and "value" [3], to explain the meaning of this new paradigm of the digital information age.

The Big Data paradigm is commonly verified through data lifecycles [10], Big Data pipelines [11], Data Engineering and Machine Learning maturity models [12], and specific Big Data Value Chain [3] among other frameworks that favor data management explanation in organizations that seek data intelligence.

In certain value chains for Big Data, such as in [3] [13], the analysis of data can be considered in different formats: text, web data, data stream, structured data, etc. These are dealt with in a specific phase of the value chain related to Analytics, which incorporates technologies, techniques and specific models [14] to extract knowledge from massive datasets.

Given the complexity of the activities associated with Analytics, corporations that consider the Big Data paradigm to obtain better results from data analysis have created frameworks that may vary in a micro view; in a macro view, however, they consider two fundamental areas: Data Engineering and Data Analytics.

Data Engineering and Data Analytics deal with data and how it can be managed, being part of a single organizational structure or belonging to two different organizational structures, or areas. Considering a specific organizational structure for Data Engineering and another for Data Analytics, the organization focused on Data Engineering is generally managed by a Chief Data Officer (CDO) [15] who focuses on data lifecycle with activities related to data generation, data collection and data storage. In turn, the organization for Data Analytics, which can generally be directed by a Chief Analytics Officer (CAO) [16] deals with activities related to data analysis, considering the construction of mathematical, statistical, and computational models and their implementation to make the results available for visualization and decision making.

In this sense, a specific organization for data analytics can be nominated Analytical Organization (AO), being part of corporations of different domains and sizes, whether in the scientific or the business industry. The activities of an AO based on AE addresses a systematized method using data, model, tool and people to transform business problems using data analysis into assertive solutions for the business [17].

An AO is capable of dealing with data analysis and data mining in the specific domain of massive data, or Big Data, which according to [18], culminates in the term Big Data Mining presenting specific characteristics in relation to conventional Data Mining. Big Data Analytics introduces the need to manage data heterogeneity and incompleteness, scalability, integration, consistency, privacy and accuracy [19]. Therefore, an AO can consider aspects related to an engineering capability in its structure, culminating in its analytical engineering capability.

An AO can thus be managed according to the specific characteristics and principles of AE that must structure its activities and processes, as well as allow this organization to be able to deliver its results, or Analytical Deliverables (AD), through specific projects of data analysis, or Analytical Projects (AP).

The conceptual view of Analytical Engineering (AE) proposed by [17] presents a framework oriented to the integration of four pillars: Data, Model, Tool and People; and three processes: Acquisition, Retention and Review. The processes intersect the four pillars determining activities to be carried out by the Analytical Team (AT).

For the AE framework, the processes are used in a managerial and strategic view. The three processes that intersect with the four pillars of AE can be detailed and integrated with other management and operational processes of an AO. An integrated view of the AE pillars and processes is presented in **Table 1**, followed by its details.

The conceptual framework of AE as part of an AO allows the creation of a more extensive set of activities and possibly specific processes and procedures to meet the expectations of the organization itself and its customers. Thus, in this section, the presentation of this framework is not linked to any specific type of AO or to an AO that meets a specific business model (such as retail, healthcare,

Table 1. Analytical engineering pillars and processes.

Pillars / Processes	Data	Model	Tool	People
Acquisition	Data can be acquired, not necessarily/not all	Model – business should be internal. Computational not necessarily	Normally acquired; part of them is internal (algorithms)	Acquired/contracted with specific skills and experience
Retention	Data can be stored, not necessarily/not all, depending on the business	Models can be stored for future projects	Normally stored with configuration management	Should be retained for future work
Review	Depending on specific needs. Practically not or rarely.	Stored models can be reviewed and improved. Especially the business models.	Internally designed tools can be reused after improvements	Training and study are used to review and improve experience

logistics, etc.); the presented framework is not linked to specific mathematical or computational models, tools or technologies, either.

Data. Data correspond to the basic element for AE; they favor all stages of a possible cycle of AE, from its acquisition, considering preparation, mining and analysis activities to be made available for visualization to support decision making. Activities commonly related to AE deal with structured, semi-structured or unstructured data. Data is what determines the properties of Big Data, therefore determining the actions of Big Data Analytics, which refer to the main objective of AE. It corresponds to the main input of Analytics and can be obtained from different sources and in different ways, requiring specific and differentiated technological infrastructure in this domain (Big Data), with greater demands on tools and computational models if considering the fundamentals of Data Stream.

Model. The models deal with the means of transforming the data to be made available to meet the business intelligence. Various types of models considered in an AO can be referenced, such as business models, mathematical and statistical models, and computational models. The business model enables and underpins AO's actions, possibly arising from a business problem and designed by a specific business area. Mathematical and statistical models can allow conditions for the AT to carry out implementations of algorithmic modifications capable of enabling the results. In a systematic and methodological approach, there is an evolution between the models until reaching a computational model that, as a rule, uses specialized tools that support the analyses to present the results or the AD. According to [20], statistical and computational models are visualized in an integrated way, especially because specific tools provide this type of integration. (This AE pillar does not presuppose a methodological view regarding the use and application of models associated with AE, seeking to highlight the possible models associated with an organization).

Tool. Tools are fundamental instruments for supporting AO. They make the activities of the AO viable and, without them, the avalanche of massive data could not be interpreted. However, data and models are needed to determine which and how to use tools capable of delivering the results. From the AE perspective, tools correspond to technologies, frameworks, algorithms, integrated solutions or systems developed for data analytics. Various types of tools are capable of favoring data analysis, be it in the area of text analytics, web analytics, mobile analytics, social media analytics, or stream analytics, among others [14], [21] [22]. A particular example of a tool in the context of AE is an integrated solution for data analytics called Share AL presented by [23] which considers three main components: 1. A full stack web application, 2. A dashboard for analyzing streaming data, and 3. High-performance computing (HPC) for performing real time analysis.

People. They conduct the processes and activities of an AO developing and deploying AD. Starting from very specialized profiles and training, people are responsible for creating and maintaining an AO. They should have to have a specific and broad knowledge of three domains: business, mathematics and statistics, and computing (with an emphasis on Artificial Intelligence and Machine Learning). The multidisciplinary learning of AO professionals has become a relevant point for the expansion of these organizations to support the business through data intelligence.

The four pillars proposed for the AE determine the means for an AO to carry out its activities; however, the procedures and activities are specifically defined by its processes. According to [24], technological resources, people and processes are fundamental to facilitating the management of Big Data. Therefore, these are the processes that allow the AE to define the strategic and functional actions of an AO, which can be institutionalized according to the precepts of the AE and also be supported by other procedures and activities at the managerial and operational level.

The **Acquisition Process** deals with the possible acquisitions made by an AO for any of the pillars; that is, for all pillars, there may be components to be acquired by an AO, such as: data, specific models, tools or people can be acquired to form the AO. All Acquisition activities should be detailed so that an AO can execute (and re-execute) its acquisitions, which can be complex for each specific type of AP performed by the AO. Thus, the acquisitions for carrying out the AP are properly managed, possibly following specific standards and models defined by each AO. Specifically considering the data and tools for a given AP, the AO can define how to acquire external data and specific tools for analyzing them, especially considering that both are cost generators for an AO. People are a relevant intellectual asset for an AO; their "acquisitions" (or hires) must thus be treated according to specific criteria to meet the objectives of the AO.

Retention Process corresponds to another relevant and strategic process for an AO, as these organizations also depend on retention for all the pillars determined by AE. For a given AP, issues such as the retention of some data that are exclusively relevant to a given business may require substantial amounts of computational resources, such as data storage and manipulation, security criteria, among others. Tools that can also be produced internally for a specific AP reflect an important retention and versioning management action, linked to the data manipulated by a specific computational model and capable of generating value. Likewise, people also need to be "retained", as they represent an important intellectual asset for carrying out the ADS of an AO. The modeling of this process reflects all the activities and standards that favor the retention of AO data, models, tools and people, as well as the measurement activities that support this process.

The **Review Process** deals with the activities carried out by an AO and that can be essentially linked to the reuse of data, models, tools (specific algorithms developed for a specific AP) or their integration. With the high cost of maintaining massive databases with security and data privacy issues, the AO can establish periodic review activities of its assets. All the data, tools or models designed for a specific AP and that belong to an AO require periodic review actions to best support future activities and future AP. People must also be "reviewed" and their domain of knowledge, studies and research must thus be constantly evaluated so that they meet the projects carried out by the AO. Participation in specialized training, workshops or conferences specific to the AO area of expertise is part of an AO professional development program.

2.2. Data Stream Overview

The huge amount of data generated in a continuous stream is called a data stream. The data stream is part of data processing in various sectors, applications and systems, such as astronomical applications, critical scientific systems [5], traffic management systems [25], as well as retail and financial applications, forecasting, telecom systems, social media activities [26] [27] among others.

Data Stream has some specific characteristics for dealing with data intelligence in terms of its management, processing and mining. Therefore, this section addresses concepts that have been proposed for the data stream, presents an overview of the activities that involve Data Stream Managing, focuses on the CRISP-DM approach [28], besides considers the activities that involve Data Stream processing—stream with specific models, tools and algorithms and their applications to generate intelligence from Stream Mining and Stream Learning.

Data stream regards the fact that the amount of data is too large to be stored before being processed; no database nor memory capacity is sufficient or able to store the amount of data [29]. Data Stream corresponds to an infinite sequence of data records arriving continuously in real-time, which culminates in the need for online processing and analysis [30]. Thus, Data Stream usually corresponds to the processing of massive data that occur continuously and in real-time, without any control of the data that is received, challenging the storage, processing and analysis capacity.

Data Stream can be considered in two aspects: structured and unstructured. Structured data presented in stream format correspond to applications in various sectors, such as data processing of consumer goods (electricity, internet, gas); the processing of data generated by retail or financial systems (generation of purchase items by a retail network, finance application to support trading decisions); high capacity for generating data related to scientific experiments (genome, aerospace); data generated by sensors in the Internet of Things domain. Stream generation for unstructured data is related to audio and video stream generation.

Both structured and unstructured data should be managed, as also the environment and people involved in the context of data analysis. [29] presents some steps on the management and mining of Data Stream. Data management in stream setting differs from a conventional approach also known as batch setting, considering Standard Database System, whereby data is processed in batch format, or simply, batch processing. Data Stream and stream processing have an online processing feature; data is manipulated directly on streams and each item in the data stream set is processed only once. Therefore, the data stream management technology infrastructure is related to the Data Stream Management System (DSMS). A DSMS considers the storage structure for the structured stream, in a table format, being able to store datasets that can correspond to part of the received data.

Besides the specific technological and data storage infrastructure for the data stream, it is relevant to verify the phases that involve data stream managing, processing and mining through a methodological view that can be based on the CRISP-DM model [28]. The model considers six phases: 1) Business understanding, 2) Data understanding, 3) Data preparation, 4) Modeling, 5) Evaluation, and 6) Deployment.

Data Stream Management can be carried out in the CRISP-DM phases 1 and 2, considering that the beginning must be linked to the understanding of the business problem. How the data is generated and properly prepared culminates in the Data Stream Processing that corresponds to phase 3 of CRISP-DM. Solution modeling and analysis or Data Stream Mining are related to phases 4 and 5 involving learning methods and algorithms associated with Machine Learning and, more specifically, Stream Learning. Using specific tools, techniques and algorithms for stream analysis, culminating in the delivery of results is part of the phase 6 of CRISP-DM.

Specifically considering data stream processing, [3] point out that data stream processing analyzes data as soon as they arrive to derive its results. Data stream processing also considers some specific techniques described according to the observation of the instances to be analyzed: Moving Windows, Landmark Windows, and Damped Window, which involves types of processing related to data stream [31].

Data stream processing is usually preceded by data stream preprocessing, a relevant phase that can be allocated to any project involving a data stream. Data stream preprocessing presents a set of activities that deals with data obtained from different sources and, therefore, can present noisy, redundant values, duplicities and missing values [32]. [33] considers Data Stream Preprocessing as the set of activities linked to integration, normalization, cleaning and transformation.

Preprocessing for data stream needs fully automated methods and algorithms as new data continuously arrives. Hence, preprocessing models need to be updated automatically since data evolves in a stream fashion [32].

The data stream preprocessing phase also considers activities that correspond to dimensionality reduction, instance reduction and feature space simplification [33]. Preprocessing is a phase that precedes the data stream processing itself and that can somehow be interchangeable with data stream processing, depending on the models and approach.

Internet of Things (IoT) corresponds to an area where it is possible to evidence the need for processing in stream format considering smart building and data stream processing. IoT is presented as smart building platforms for cities, hospitals, energy networks and industries considering massive data generation in its applications, thus requiring data stream processing and preprocessing, and data stream mining activities for huge amounts of data [34].

Data stream mining is part of the activities involving data intelligence. Data Mining, specifically, corresponds to an interdisciplinary area of knowledge that involves Statistics, Machine Learning and Business Science to extract patterns from datasets as exemplified by association rule mining and clustering in [35]. Specifically for the data stream, mining concerns extracting knowledge from non-stopping streams of information as mentioned in [5].

According to [31], data stream mining has become crucial in many applications that demand the analysis of data in real-time (or near real-time), such as social networks analysis, sensor data analysis, etc., given that they can generate data from evolving distributions.

Data stream mining uses specific methods and tools to perform mining activities that cannot be processed by traditional Data Mining tools due to some constraints, such as limited memory, data speed and dynamic environment [27]. The use of learning models and specific algorithms for stream learning collaborate with Data Stream Mining, which considers the application and use of specific tools, such as SAMOA, StreamDM, Spark, Storm [30] [36] [37] [38] [39], for stream learning and mining activities.

References [31] [32] link Data Stream Mining activities to Big Data, presenting the term Big Data Stream. [32] explains the challenge to data stream mining as regards volume, velocity and volatility. Volatility emphatically corresponds to what can be observed in data received the stream, as it deals with dynamic data with changing patterns over time, which corresponds to concept drift. Concept drift is presented and discussed in more detail in [40] [41].

Machine Learning is the engine of data stream mining activities [42] favoring the conditions of stream learning that, using specific algorithms for data stream, are able to compute prediction on a stream of incoming data [43]. Concept drift and Sliding window correspond to some of the specific practices to data stream, verified in stream learning models.

Concept drift is a specific constraint for data stream mining that considers the changes observed in the stream, altering the distribution over time and impairing results and accuracy, especially for classification models [27] [41] [44] [45]. For [32], concept drift categories correspond to smoothness of concept transition, singular or recurring contexts, systematic or unsystematic, real or virtual.

Sliding Window is another practice related to stream learning that considers the analysis of the most recent data received in stream [5]. This method links the analysis of the most recently received instances, summarizing the analysis of previous versions (packages).

Above all, recognized and widely used models of Machine Learning make up stream learning, such as Naïve Bayes and Neural Networks [30] [46]. Hoeffding Trees [30] are incremental updates of decision tree models also used for data stream mining [43]. SAMOA, Storm, Spark Streaming and Flink are examples of tools that present the implementation of these algorithms to carry out stream mining activities [36] [39] [47].

To conclude this section, some challenges are posed to the data stream approach; for [31]—evolving data streams, running time, memory usage, high-dimensionality, concept drifts, delayed labeling and imbalanced labels. [5] considers that the continuous streams of data have challenged storage, computation and communication capabilities in information systems whereby new models, tools and techniques have been proposed to cover these needs. [48] also presents some relevant challenges for data stream and concept drift related to the construction of efficient online learning rules that adapt to models based on streaming data; how to handle preprocess data, a relevant factor for learning problems; and how to deal with concept drift for online learning.

3. Results

Data in the field of AE refers to the fundamental element for carrying out the analyses, followed by models that can cover business, statistical and computational aspects to provide the conditions for analysis. Tools stand out for being essential for these types of analyses. The tools vary in size and capabilities that are constantly changing, presented by academia and industry. With ample and greater emphasis on people, or human resources, AE depends on human knowledge about business, computing and statistics to deliver the proper results based on data, models and tools.

Therefore, the AE pillars and the three processes associated with the data stream paradigm, favor the activities of data stream analytics. In this sense, the conceptual framework of AE can promote a specific integrated framework for data stream analytics named Analytical Engineering for Data Stream (AEDS).

The pillars of AEDS represent its static capacity and for these pillars to become functional and achievable, three processes are proposed; they represent the dynamic capacity of AEDS—Acquisition, Retention and Review processes. Processes provide the operational and functional conditions for AOs, and their operationalization occurs in an integrated way with the proposed pillars. Based on these processes, specific procedures, activities and tasks that detail the operationalization can be derived for solving massive data analysis problems.

An organization that deals with data streams, as an example, can use the AEDS framework, considering the three processes to structure its activities, as well it could use the four pillars that are able to determine what should be addressed by the organization's processes. Supposedly the organization must maintain its other internal processes that are not considered in the AEDS.

Thus, the pillars and processes of AEDS are detailed below associated with the data stream paradigm. Their objectives, possible integration of the pillars and processes, and possible procedures for AEDS processes are presented in this section. An overview of an AO and its APs and results is exemplified to better understand AEDS framework.

Data is huge, massive, in a paradigm that can also be called Big Data. Data streams fundamentally correspond to the three main characteristics, or those most commonly associated with Big Data, volume, variety and velocity. In a data stream approach, it is possible to verify extensive and continuous data flows; high volumes of data can be observed in applications that consider data streams. Some examples are associated with the Internet of Things, astronomical applications, scientific systems, financial systems, weather forecasting and social networks. Variety in the data stream can also be considered, given the different stream data sources. Massive, varied data are observed and collected in stream format for online analysis and for responses that must be provided according to the same data reception speed, be it when considering structured and unstructured data applications. Velocity is an intrinsic part of the data stream, which presents itself at a speed that makes it difficult even for traditional Machine Learning models to offer answers about the data processing. Data stream does not refer to datasets that can be stored for future queries. Supposedly, there are repositories that can store specific portions of the continuous flows of data that are received for some kind of a posteriori analysis; however, it is not feasible to approach data stream with aspects of traditional storage. The unique data issues for AEDS projects should emphatically consider preprocessing aspects, that is, algorithms that deal with the problem of data that present noise, dirt, and various biases that can make analysis difficult. [33] presents some examples of algorithms and useful tools for preprocessing, expressively favoring preprocessing for the data stream, which, after collection, undergoes an analysis step dependent on specific data stream models for stream learning and stream mining. For anomaly detection, [49] proposes a new optimization function for evaluating data stream anomalies, *i.e.*, when some predictive values exceed predefined measures. However, for an AO that manages AEDS projects of the data stream, there is also the issue of data integration that allows them to be analyzed.

Model. They are responsible for transforming data. For data stream problems, AP should consider three types of models: business, mathematics and statistical models, and computational models. They are usually integrated into a unique model solution; although they have specific characteristics, they should be integrated to define a unique solution. A business model should cover aspects related to the business needs regarding data management, learning and mining. Different kinds of business problems can be defined in the Business model. As an example, it is possible to consider the part of a solution that uses the Internet of Things and its infrastructure, considering data collection based on sensors for some data integration for future analysis. This is part of the business model that should define how the data will be integrated with other data and how it should be processed. Specifically, considering data stream as part of the solution, the business model should define different aspects of the analysis, including data collection, preprocessing and cleaning, processing and, finally, results presentation (or visualization). Another part of the model for AP should think in an integrated way, since statistical and computational models for this type of solutions are highly integrated. Computational models treat algorithms and tools that offer the capacity to stream learning and mining based on the data stream. In this sense, the computational model is responsible for defining which types of Machine Learning methods could be applied to the solution. The methods are considered by many algorithms and are part of the implementation in different specific tools for data stream mining. The decision about its use can be part of a Computational model. In this case, the model can consider many methods, algorithms and tools and each of them should be tested and validated to define a better solution. Moreover, the computational model is fully integrated into the statistical one and the reflections regarding some statistical methods should thus be observed, even though these methods are part of the algorithms and tools. As an example, it is possible to observe the framework presented by [50] which represents the modeling of an integrated solution for data stream analytics defined as Big Data Stream Analytics for online Sentiment Analysis (BDSASA) with seven layers for data stream processing, mining, and learning, corresponding to 1) Data Stream, 2) Data pre-processing, 3) Data Mining, 4) Prediction, 5) Learning and Adaptation, 6) Presentation, and 7) Storage.

Tool. An extensive set of tools are usable for AEDS-based solutions; however, some restrictions are placed when it comes to data stream mining. For AP, some special features must be provided by these tools. According to [27] [51] [52] concept drift is a relevant practice linked to data stream solutions and systems. Considering the most useful tools for data stream learning and mining, it is possible to verify solutions based on tools called Data Stream Processing Engines (DSPE), such as SAMOA, Storm, Spark Streaming, Flink, Kafka Streams. These tools can be compared, according to [39], considering features such as programming model, data partition, state management, processing and fault tolerance. [38] also addresses some DSPEs, such as Spark, Flink and Storm, according to characteristics related to latency, throughput, fault tolerance, usability, resource expense and scalability. Other specific architectures involving data stream and analytics are found in [53], which presents a proposal for a multimodal architecture that deals with batch and streaming processing for knowledge extraction from multiple heterogeneous Big Data sources; or in [34], which presents a specific system called SenseRT, for managing and analyzing real-time streams of sensor data collecting streams and uses a range of network protocols.

People. In any of the AP, the knowledge acquired and developed by people is essential for managing and developing related activities for this type of project. As these are complex and multidisciplinary projects, the training and education of professionals in this area demand investments and efforts, whether personal or institutional. The courses and programs offered do not address the specificity of the data stream, and may be addressed in some way by some specific graduate programs. Thus, the training and instrumentation of professionals for AEDS projects for data stream become a crucial point of attention. People, in this context, have to have proper treatment and training; they need greater support and involvement with different means of instruction acquisition, either by participating in Workshops and Conferences in the Data Mining, Data Stream and Machine Learning subjects, by self-learning, or by education in postgraduate programs at the university or at specific technology institutions.

Possibly all organizations that need to manage data streams should consider these four pillars, as they are essential for managing the data cycle, from data capture, storage, analysis, ending with making the results available to decision-makers.

The four pillars of AEDS, as static components, require dynamization, which occurs, as aforementioned, by AEDS processes. The processes can be comprehensively considered as process areas, requiring that they be detailed by procedures, activities and tasks for real implementation.

Each of the proposed AEDS processes has a general objective and possibly some specific objectives, which may vary according to the AO and the type of AP to be carried out. The general objectives of each of the proposed processes for AEDS and their integration with the four pillars of AEDS are detailed below.

Acquisition Process. This process presents important activities for an AP for data stream since it highlights all the acquisitions necessary for the project to be carried out. Hence, in a macro view, these acquisitions are treated in accordance with what is identified for each of the pillars of the AEDS. Regarding the pillars referring to data, models, tools and people, what highlights can be made about the acquisition of an AP for the data stream? Even data can be acquired. How does data acquisition for an AP occur? This answer is linked to the way the AP should handle data acquisition according to a previously designed business model. The **data** and **models**’ pillars are intensively integrated at this point, yet models cannot be acquired, especially business models, as these are intrinsic to the organization that must use the AP results. However, the basis for data acquisition must be part of what is described in the business model. Possibly the other models (statistical and computational) cannot be acquired either, as they refer to the core of an AP. In this sense, modeling determines the business and technical-computational aspects to be addressed by the project. Acquiring **tools** is possibly the most necessary action in an AP for the data stream. No project can be materialized without the acquisition of specific tools, even if customizations have been performed or some algorithms slightly modified, others even implemented; an AP requires the acquisition of specific tools. The diversity of tools is extensive, requiring a proper capacity for the correct selection and subsequent use of the tools by the people involved in the AP. **People** must be "acquired", or rather hired, to meet the needs of the AP. As mentioned in the pillar referring to people, they need very specific training, which demands the contractors' important criteria and observations so that these people can properly perform their roles throughout the AP and provide the expected results.

Retention Process. The characteristics of this process are aimed at storing and managing the configuration of project deliverables or work products—developed or changed—during the AP. Retention of ADs, or part of what was acquired or produced, may be relevant for future APs. Retention or storage must be carried out with strict configuration management criteria. For example, possible models and algorithms designed and implemented in a given AP can be properly reused in future projects. The retention process is relevant for AEDS and can be applied to all AEDS-related pillars. For **data**, retention or storage is supposed to occur in a partial way, as it is not possible for an AEDS project to have all the data processed stored for reuse; there is no sense in retention or storage in a streaming view. However, possible mechanisms and databases for partial storage may occur in some specific business models. **Models**, in general, can be reusable, entirely or partially. The modeling performed by an AO for a given AP may undergo minor changes for future APs, supposedly considering the same problem or business area. **Tools** acquired or developed, can possibly be retained and, depending on new needs or adjustments, be reused. The tools acquired will certainly be part of an organization's assets and are reusable; however, specific algorithms or programs developed and stored under strict configuration criteria may undergo minor modifications to meet new projects. Supposedly, the intellectual capital of an AP, the **people**, considering the possibilities of the organization, must be properly "retained", that is, they must remain as part of the team of an AO; since they have acquired the knowledge and experience from a first AP, or the first APs carried out, they will be of total relevance for future APs.

Review Process. Reviews are periodic, according to pre-established periods, or, at a minimum, at the beginning of the end of each AP since they are essential to verify deliverables purchased or produced by an AO. Reviews should favor the reuse of work products developed for each project. For an AP for the data stream, the review is possible for models, tools and people, but not for the totality of previously processed data. **Data** is not reusable in its entirety for a data stream project. As mentioned in the Retention Process, for some business strategies, a subset of data obtained in a data stream project can be retained (stored), but not its entirety. Therefore, the review of the **models** is totally feasible for an AP for the data stream, and they can certainly be reused for new APs, whether the business models or the computational model, for example, whereby the strategies are linked to the data and the tools and possible algorithms developed specifically for a project. The review and subsequent reuse of these models can be common, since there are many similarities in data stream projects carried out for the same business area. **Tools**, acquired or internally developed for a given AP, specific algorithms developed by the AT, possibly retained and stored under rigorous configuration management, can be reused in many scenarios, even for different business areas or for computational models that demand small changes. **People** refer to the component that can undergo "review" in the sense that they can be reassigned to different APs, given skills and expertise, and can be evaluated in the sense that they are trained to improve knowledge or to be promoted to new roles and challenges. In general, this process must belong in the initial or final phase of an AP; reviews occur and favor the ATs to carry out their activities to properly produce the AD of the project.

An example can be posed by considering an Internet of Things area of an organization that deals with data stream; it can be supported by the AEDS framework, *i.e.*, consider the AEDS mapped processes and also the possible procedures highlighted in Table 2 and Table 3 to define part of data lifecycle management. Table 2 depicts the integration of the pillars and processes of AEDS, with the process's objectives and with possible procedures for each process. Table 2 also presents a description of the relationship of each pillar with the processes defined for AEDS. Table 3 specifically addresses some possible procedures related to each of the processes provided by AEDS.

Project Management principles can be applied by an AO to carry out its projects. Initiation, Planning, Execution, and Closing phases are capable of organizing the processes, procedures and activities of APs developed by the AO.

In the Initiation and Planning phases, processes and procedures associated with **Acquisition** and **Review** are possibly considered, as they favor the kick-off and the initial and planning activities that enable the execution of the project. In the Execution phase, the process and procedures of **Acquisition** are used for activities of acquiring the project components; project-specific work products are developed, such as new components (models, algorithms, databases, etc.), as well as final deliverables; the process and procedures of **Retention** can be started during this phase. For the Closing phase, the **Retention** process and its procedures are carried out since the components and the developed work products should be placed under configuration for future reuse. Table 4 summarizes the impact of each process on the projects' phases.

The project phases presented in Table 4 do not represent a linear solution,

Table 2. AEDS pillars and processes.

Process	Goals	Procedures	Data Stream Project			
			Data	Model	Tool	People
Acquisition	1. Acquire specific components, tools, etc. 2. Buy or develop components, tools, algorithms, etc.	See examples on Table 3	The acquisition depends on the business model. It can occur for data, tools and people.	Models are specific to each project, especially business ones. Computational models can be reused.	Tools can be acquired/purchased for specific projects.	People are needed for the Project and must be hired to form the Project teams.
Retention	1. Configuration Management. 2. Store Procedure.	See examples on Table 3	Data for stream projects cannot be retained, but some batches can be stored, depending on the business model.	Models can be retained for later reuse in similar projects.	Tools, acquired or developed for a specific project, may be retained.	People are necessarily retained, because with their acquired expertise, they should perform better in future projects.
Review	1. Review project work products, tools, algorithms, etc. 2. Reuse.	See examples on Table 3	There is no review for data.	Models can be revised, when starting new projects, to favor their reuse.	Tools and components developed for a specific project can be placed under configuration management to be reviewed at the start of a new project for reuse.	People should be used for future projects, receive new training and be promoted.

Table 3. Examples of procedures associated with each process of AEDS.

Process	Procedures
Acquisition	1. Define input acquisition criteria and mechanisms. 2. Evaluate possible suppliers. 3. Acquire the inputs for the project. 4. Validate/test purchased inputs (tools, for example). 5. Provide validated input.
Retention	1. Identify components (project work products, tools, software, algorithms, etc.) that can be reused. 2. Place components (developed or acquired) under a strict configuration management system. 3. Store components for possible reuse in future projects.
Review	1. Define review criteria (beginning and/or end of the project). 2. Define checklists for reviews. 3. Perform the appropriate project component reviews. 4. Reuse components in new projects.

Table 4. Processes' impact in the phases of an AP.

AP	Initiation	Planning	Execution	Closing
Acquisition	*Low impact*	*Medium impact*	*High impact*	*No impact*
Retention	*No impact*	*Medium Impact*	*Medium impact*	*High impact*
Review	*Low impact*	*Medium Impact*	*Low impact*	*Low impact*

since these phases can be overlapped or can be executed in a cyclical way, especially considering the Planning and Execution phases. The activities of each phase must meet the organizational processes, as well as the AEDS processes, according to their impacts highlighted for each phase. These phases can be applied to all APs carried out by an AO that supports data stream analytics which is represented (high level) in Figure 1.

A generic relation between APs, ATs and ADs is represented on the right side of Figure 1, meaning that, many APs can be performed by many ATs generating many ADs; the rest of Figure 1 exemplifies instantiations of APs being performed by ATs to deliver ADs. Based on Figure 1, supposing that AEDS is used by an internal organization of a particular corporation that carries out analytical activities in retail systems.

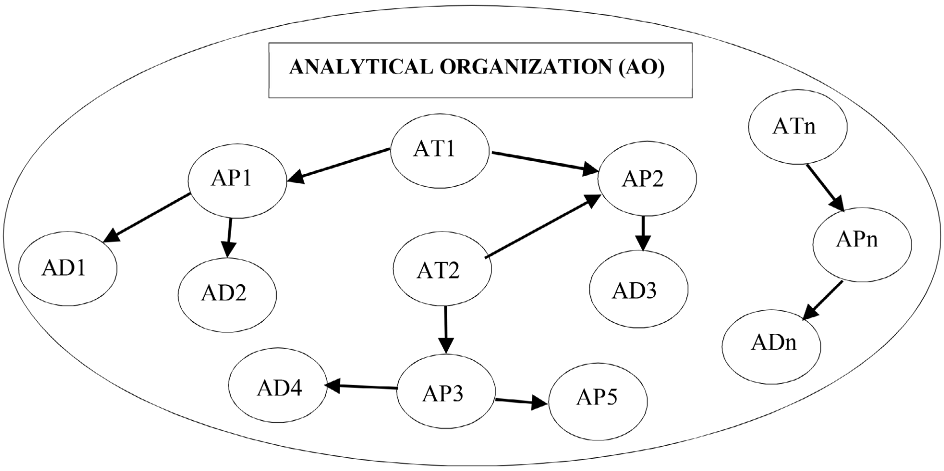

Figure 1. Analytical projects of an analytical organization.

This organization can realize several APs regarding data stream problems at the same time, considering any integration and disposition of ATs.

Specifically considering AP2, which is conducted by two teams (AT1 and AT2), since each of them has different specialties. AT1 acts on the acquisition process, data engineering and preprocessing activities; and AT2 acts on analysis and visualization activities, using the review and retention processes, both teams working according to the specifications of AP2 plan. Although the project follows the phases for its execution and also follows the AEDS processes and pillars to support the activities of this AO, it is not possible to determine linearity of activities to be performed by both ATs, given that the activities can be overlapped, depending on the project necessities. Regarding the delivery of the AP2, its representation as a single deliverable (AD3) does not underestimate the outcome of AP2 since AD3 can represent a specific algorithm, or function, to support an existing decision system, or it can represent the entire decision-making system constructed during AP2.

Similar situations to the aforementioned example of the project AP2 may occur for the other APs represented in **Figure 1**. Although this representation does not express the unique proposal of **Figure 1**, it also aims to highlight the interactions of APs and, by consequence, of ATs, that execute several APs in the organization to carry out its specific deliveries.

4. Conclusions

Of the three main properties defined for the Big Data paradigm, velocity can be highlighted with greater emphasis when related to the data stream. The data stream is strictly related to the fact that it refers to the high velocity of data transportation and also presents itself with a property of massive data.

Addressing data stream corresponds to a strong relation to the Big Data paradigm being possible to verify the proposition of the term Big Data Stream. Thus, when the data analysis activities are considered in the stream setting approach, it has also to observe specific characteristics of the Big Data paradigm, especially regarding data management that requires scalability, memory capacity, storage feasibility; in addition to computational and data intelligence models, Artificial Intelligence and Machine Learning algorithms.

Specific Machine Learning models, data mining methods and tools, specific techniques and concepts have been reflected in data analysis tasks for the data stream. Data stream setting involves complex preprocessing activities and new algorithms for stream mining. Some evolved from batch setting models, also requiring algorithms to deal with one of the most pertinent problems to stream setting related to concept drift. In any case, dealing with data analysis in a stream setting requires highly qualified professionals to orchestrate a set of complex tools and models to extract value from “stream” data, which can be supported by principles of engineering.

All the skills, tools and models involved in data stream activities for mining and learning require an engineering capability to make data stream analysis possible and achievable. Thus, AEDS can be presented as a relevant conceptual framework for projects in AO that deal specifically with AP for the data stream. The specificity concerning these projects favors the AEDS proposition, its four pillars (Data, Model, Tool, People) are possibly coherent and quite relevant to the problem of stream management and analysis. However, the processes (Acquisition, Review) may be insufficient if they are not treated as macro processes (or process area). The proposed processes for AEDS allow an AO to reinterpret them and create other processes or a set of procedures derived from them in a way that makes the APs and ADs of an AO viable.

The conceptual proposition associated with other processes can support an AO, considering that the four pillars proposed for AEDS are strictly related to the three processes; except for the pillar-process relationship referring to "data" and "retention", due to the fact that data cannot be retained in its entirety, due to obvious issues about volume. There is also a specificity regarding the process-pillar "acquisition" and "models", because some basic and commonly used models for data stream analysis can be acquired by an AO, but not all. Certainly, the core business or computational "model" (mining and learning) are not amenable to acquisition.

To conclude, the proposition of AEDS conceptual framework, possibly enables Analytical Organizations to perform their Analytical Projects and also deliver their Analytical results, favoring data intelligence based on data stream setting.

Acknowledgements

The authors would like to thank Programa de Educação Continuada em Engenharia (PECE) (Engineering Continuing Educational Program) of the Polytechnic School of University of São Paulo for supporting this work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Joseph, R.C. and Johnson, N.A. (2013) Big Data and Transformational Government. *IT Professional*, **15**, 43-48. https://doi.org/10.1109/MITP.2013.61

[2] Chen, C.P. and Zhang, C.Y. (2014) Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences*, **275**, 314-347. https://doi.org/10.1016/j.ins.2014.01.015

[3] Hu, H., Wen, Y., Chua, T.S. and Li, X. (2014) Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, **2**, 652-687. https://doi.org/10.1109/ACCESS.2014.2332453

[4] Gebara, F.H., Hofstee, H.P. and Nowka, K.J. (2015) Second-Generation Big Data Systems. *Computer*, **48**, 36-41. https://doi.org/10.1109/MC.2015.25

[5] Gaber, M.M., Zaslavsky, A. and Krishnaswamy, S. (2005) Mining Data Streams: A Review. *ACM SIGMOD Record*, **34**, 18-26. https://doi.org/10.1145/1083784.1083789

[6] Jansson, J. and Hakala, I. (2020) Managing Sensor Data Streams in a Smart Home Application. *International Journal of Sensor Networks*, **32**, 247-258. https://doi.org/10.1504/IJSNET.2020.106603

[7] Kohavi, R., Rothleder, N.J. and Simoudis, E. (2002) Emerging Trends in Business Analytics. *Communications of the ACM*, **45**, 45-48. https://doi.org/10.1145/545151.545177

[8] Elgendy, N. and Elragal, A. (2016) Big Data Analytics in Support of the Decision Making Process. *Procedia Computer Science*, **100**, 1071-1084.

https://doi.org/10.1016/j.procs.2016.09.251

[9] Reimsbach-Kounatze, C. (2015) The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis. OECD Digital Economy Papers, OECD Publications, Paris.

[10] Sinaeepourfard, A., Garcia, J., Masip-Bruin, X. and Marín-Tordera, E. (2016) A Comprehensive Scenario Agnostic Data Life Cycle Model for an Efficient Data Complexity Management. 2016 *IEEE* 12*th International Conference on e*-Science (*e-Science*), Baltimore, 23-27 October 2019, 276-281. https://doi.org/10.1109/eScience.2016.7870909

[11] Fisher, D., DeLine, R., Czerwinski, M. and Drucker, S. (2012) Interactions with Big Data Analytics. *Interactions*, **19**, 50-59. https://doi.org/10.1145/2168931.2168943

[12] Akkiraju, R., Sinha, V., Xu, A., Mahmud, J., Gundecha, P., Liu, Z., Liu, X. and Schumacher, J. (2020) Characterizing Machine Learning Processes: A Maturity Framework. 18*th International Conference on Business Process Management*, Seville, 13-18 September 2020, 17-31. https://doi.org/10.1007/978-3-030-58666-9_2

[13] Miller, H.G. and Mork, P. (2013) From Data to Decisions: A Value Chain for Big Data. *IT Professional*, **15**, 57-59. https://doi.org/10.1109/MITP.2013.11

[14] Chen, H., Chiang, R.H. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, **36**, 1165-1188. https://doi.org/10.2307/41703503

[15] Lee, Y., Madnick, S.E., Wang, R.Y., Wang, F. and Zhang, H. (2014) A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data. *MIS Quarterly Executive*, **13**, 1-13.

[16] Cao, L. (2017) Data Science: A Comprehensive Overview. *ACM Computing Surveys*, **50**, Article No. 43. https://doi.org/10.1145/3076253

[17] Rossi, R. and Hirama, K. (2020) Towards a Conceptual Approach of Analytical Engineering for Big Data. *International Journal of Engineering and Technology*, **12**, 9-17. https://doi.org/10.7763/IJET.2020.V12.1176

[18] Che, D., Safran, M. and Peng, Z. (2013) From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. 18*th International Conference on Database Systems for Advanced Applications*, Wuhan, 22-25 April 2013, 1-15. https://doi.org/10.1007/978-3-642-40270-8_1

[19] Kolajo, T., Daramola, O. and Adebiyi, A. (2019) Big Data Stream Analysis: A Systematic Literature Review. *Journal of Big Data*, **6**, Article No. 47. https://doi.org/10.1186/s40537-019-0210-7

[20] Louridas, P. and Ebert, C. (2013) Embedded Analytics and Statistics for Big Data. *IEEE Software*, **30**, 33-39. https://doi.org/10.1109/MS.2013.125

[21] Gandomi, A. and Haider, M. (2015) Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, **35**, 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[22] Sarker, I.H. (2018) Mobile Data Science: Towards Understanding Data-Driven Intelligent Mobile Applications. *EAI Endorsed Transactions on Scalable Information Systems*, **5**, Article No. e4. arXiv:1811.02491. https://doi.org/10.4108/eai.13-7-2018.155866

[23] Oesch, S., Gillen, R. and Karnowski, T. (2020) An Integrated Platform for Collaborative Data Analytics. 2020 *International Conferences on Internet of Things* (*iThings*) *and IEEE Green Computing and Communications* (*GreenCom*) *and IEEE Cyber*, *Physical and Social Computing* (*CPSCom*) *and IEEE Smart Data* (*SmartDa-*

ta) and IEEE Congress on Cybermatics (*Cybermatics*), Rhodes, 2-6 November 2020, 648-653. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00113

[24] Rossi, R. and Hirama, K. (2015) Characterizing Big Data Management. *Issues in Informing Science and Information Technology*, **12**, 165-180. https://doi.org/10.28945/2204

[25] Vera-Baquero, A. and Colomo-Palacios, R. (2018) A Big-Data Based and Process-Oriented Decision Support System for Traffic Management. *EAI Endorsed Transactions on Scalable Information Systems*, **5**, Article No. e5. arXiv:1806.05855. https://doi.org/10.4108/eai.29-5-2018.154810

[26] Chhabra, I. and Suri, G. (2019) Knowledge Discovery for Scalable Data Mining. *EAI Endorsed Transactions on Scalable Information Systems*, **6**, Article No. e3. https://doi.org/10.4108/eai.19-3-2019.158527

[27] Althabiti, M.S. and Abdullah, M. (2020) CDDM: Concept Drift Detection Model for Data Stream. *International Journal of Interactive Mobile Technologies*, **14**, 90-106. https://doi.org/10.3991/ijim.v14i10.14803

[28] Wirth, R. and Hipp, J. (2000) CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the* 4*th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, 11-13 April 2000, 29-40.

[29] Hébrail, G. (2008) Data Stream Management and Mining. In: Steinberger, R., Fogelman-Soulié, F., Perrotta, D. and Piskorski, J., Eds., *Mining Massive Data Sets for Security*, IOS Press, Amsterdam, 89-102.

[30] Bifet, A., Gavalda, R., Holmes, G. and Pfahringer, B. (2018) Machine Learning for Data Streams: With Practical Examples in MOA. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/10654.001.0001

[31] Bahri, M., Bifet, A., Gama, J., Gomes, H.M. and Maniu, S. (2021) Data Stream Analysis: Foundations, Major Tasks and Tools. *WIREs Data Mining Knowledge Discovery*, **11**, Article No. e1405. https://doi.org/10.1002/widm.1405

[32] Krempl, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., *et al.* (2014) Open Challenges for Data Stream Mining Research. *ACM SIGKDD Explorations Newsletter*, **16**, 1-10. https://doi.org/10.1145/2674026.2674028

[33] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M. and Herrera, F. (2017) A Survey on Data Preprocessing for Data Stream Mining: Current Status and Future Directions. *Neurocomputing*, **239**, 39-57. https://doi.org/10.1016/j.neucom.2017.01.078

[34] Verma, R., Brazauskas, J., Safronov, V., Danish, M., Merino, J., Xie, X., *et al.* (2021) SenseRT: A Streaming Architecture for Smart Building Sensors. arXiv:2103.09169.

[35] Agarwal, R. (2017) Decision Making with Association Rule Mining and Clustering in Supply Chains. *International Journal of Data and Network Science*, **1**, 11-18. https://doi.org/10.5267/j.ijdns.2017.1.003

[36] Morales, G.D. F. and Bifet, A. (2015) SAMOA: Scalable Advanced Massive Online Analysis. *Journal of Machine Learning Research*, **16**, 149-153.

[37] Tsai, C.W., Lai, C.F., Chao, H.C. and Vasilakos, A.V. (2015) Big Data Analytics: A Survey. *Journal of Big Data*, **2**, Article No. 21. https://doi.org/10.1186/s40537-015-0030-3

[38] Landset, S., Khoshgoftaar, T.M., Richter, A.N. and Hasanin, T. (2015) A Survey of

Open Source Tools for Machine Learning with Big Data in the Hadoop Ecosystem. *Journal of Big Data*, **2**, Article No 24. https://doi.org/10.1186/s40537-015-0032-1

[39] Isah, H., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F. and Khan, S. (2019) A Survey of Distributed Data Stream Processing Frameworks. *IEEE Access*, **7**, 154300-154316. https://doi.org/10.1109/ACCESS.2019.2946884

[40] Webb, G.I., Hyde, R., Cao, H., Nguyen, H.L. and Petitjean, F. (2016) Characterizing Concept Drift. *Data Mining and Knowledge Discovery*, **30**, 964-994. https://doi.org/10.1007/s10618-015-0448-4

[41] Iwashita, A.S. and Papa, J.P. (2018) An Overview on Concept Drift Learning. *IEEE Access*, **7**, 1532-1547. https://doi.org/10.1109/ACCESS.2018.2886026

[42] Gomes, H.M., Read, J., Bifet, A., Barddal, J.P. and Gama, J. (2019) Machine Learning for Streaming Data: State of the Art, Challenges, and Opportunities. *ACM SIGKDD Explorations Newsletter*, **21**, 6-22. https://doi.org/10.1145/3373464.3373470

[43] Baier, L., Reimold, J. and Kühl, N. (2020) Handling Concept Drift for Predictions in Business Process Mining. 2020 *IEEE 22nd Conference on Business Informatics* (*CBI*), Vol. 1, Antwerp, 22-24 June 2020, 76-83. https://doi.org/10.1109/CBI49978.2020.00016

[44] Tsymbal, A. (2004) The Problem of Concept Drift: Definitions and Related Work. Vol. 106, Computer Science Department, Trinity College, Dublin, 58.

[45] Hoens, T.R., Polikar, R. and Chawla, N.V. (2012) Learning from Streaming Data with Concept Drift and Imbalance: An Overview. *Progress in Artificial Intelligence*, **1**, 89-101. https://doi.org/10.1007/s13748-011-0008-0

[46] Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P. and Sheth, A.P. (2018) Machine Learning for Internet of Things Data Analysis: A Survey. *Digital Communications and Networks*, **4**, 161-175. https://doi.org/10.1016/j.dcan.2017.10.002

[47] L'heureux, A., Grolinger, K., Elyamany, H.F. and Capretz, M.A. (2017) Machine Learning with Big Data: Challenges and Approaches. *IEEE Access*, **5**, 7776-7797. https://doi.org/10.1109/ACCESS.2017.2696365

[48] Bifet, A., Hammer, B. and Schleif, F.M. (2019) Recent Trends in Streaming Data Analysis, Concept Drift and Analysis of Dynamic Data Sets. *ESANN* 2019: *European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning*, Bruges, 24-26 April 2009, 421-430.

[49] Hasani, Z. and Krrabaj, S. (2019) Survey and Proposal of an Adaptive Anomaly Detection Algorithm for Periodic Data Streams. *Journal of Computer and Communications*, **7**, 33-55. https://doi.org/10.4236/jcc.2019.78004

[50] Cheng, O.K.M. and Lau, R. (2015) Big Data Stream Analytics for Near Real-Time Sentiment Analysis. *Journal of Computer and Communications*, **3**, 189-195. https://doi.org/10.4236/jcc.2015.35024

[51] Kumar, A., Kaur, P. and Sharma, P. (2015) A Survey on Hoeffding Tree Stream Data Classification Algorithms. *CPUH-Research Journal*, **1**, 28-32.

[52] Rad, R.H. and Haeri, M.A. (2019) Hybrid Forest: A Concept Drift Aware Data Stream Mining Algorithm. arXiv:1902.03609.

[53] Yousfi, S., Rhanoui, M. and Chiadmi, D. (2019) Towards a Generic Multimodal Architecture for Batch and Streaming Big Data Integration. *Journal of Computer Science*, **15**, 207-220. arXiv:2108.04343. https://doi.org/10.3844/jcssp.2019.207.220