

This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

Public Domain Mark 1.0

<https://creativecommons.org/publicdomain/mark/1.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.



An edge–cloud integrated framework for flexible and dynamic stream analytics

Xin Wang^{a,b}, Azim Khan^{a,b}, Jianwu Wang^{a,b,*}, Aryya Gangopadhyay^{a,b}, Carl Busart^c, Jade Freeman^c

^a Department of Information Systems, University of Maryland, Baltimore County, MD, United States

^b Center for Real-time Distributed Sensing and Autonomy, University of Maryland, Baltimore County, MD, United States

^c DEVCOM Army Research Laboratory, Adelphi, MD, United States

ARTICLE INFO

Article history:

Received 20 May 2022

Received in revised form 23 July 2022

Accepted 29 July 2022

Available online 5 August 2022

Keywords:

Edge computing

Internet of Things (IoT)

Cloud computing

Edge–cloud integration

Stream data analytics

Concept drift

Hybrid learning

Long short-term memory (LSTM)

ABSTRACT

With the popularity of Internet of Things (IoT), edge computing and cloud computing, more and more stream analytics applications are being developed including real-time trend prediction and object detection on top of IoT sensing data. One popular type of stream analytics is the recurrent neural network (RNN) deep learning model based time series or sequence data prediction and forecasting. Different from traditional analytics that assumes data are available ahead of time and will not change, stream analytics deals with data that are being generated continuously and data trend/distribution could change (a.k.a. concept drift), which will cause prediction/forecasting accuracy to drop over time. One other challenge is to find the best resource provisioning for stream analytics to achieve good overall latency. In this paper, we study how to best leverage edge and cloud resources to achieve better accuracy and latency for stream analytics using a type of RNN model called long short-term memory (LSTM). We propose a novel edge–cloud integrated framework for hybrid stream analytics that supports low latency inference on the edge and high capacity training on the cloud. To achieve flexible deployment, we study different approaches of deploying our hybrid learning framework including edge-centric, cloud-centric and edge–cloud integrated. Further, our hybrid learning framework can dynamically combine inference results from an LSTM model pre-trained based on historical data and another LSTM model re-trained periodically based on the most recent data. Using real-world and simulated stream datasets, our experiments show the proposed edge–cloud deployment is the best among all three deployment types in terms of latency. For accuracy, the experiments show our dynamic learning approach performs the best among all learning approaches for all three concept drift scenarios.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Stream analytics has become a major data analytics area due to hardware and software advances in Internet of Things (IoT), edge computing and cloud computing. It is now much easier to obtain sensing data from IoT devices, which leads to more and more stream analytics applications including real-time trend prediction [1–3] and real-time object detection [4,5] on top of IoT sensing data. Different from traditional analytics that assumes data to be processed are available ahead of time and will not change, stream analytics processes data that are being generated on the fly and continuously. A well-known challenge of stream analytics is concept drift [6,7] which describes changes in the concept or distribution of stream data.

There is a growing number of studies on how to conduct stream analytics by leveraging IoT, edge and cloud resources. Edge computing in an IoT environment brings computation and data storage closer to data sources. It operates on “instant data” that is usually time sensitive. Besides the latency benefit, edge computing is normally designed for remote locations, where there is limited or no connectivity to a centralized computation location. However, resources on edges are constrained and limited in their capacity/capability and can only support relatively simple data processing like inference/prediction based on a pre-trained model. So it often relies on additional resources, such as storage or memory optimized devices, for more complex processing. Cloud computing [8] provides on-demand computational resources for data analytics over the Internet and becomes a major approach for supporting complex and high-performance computation. Besides IoT environments, streaming data could also be delivered directly to the cloud and be computed with enough computing power and storage capacity. However, considering its distance to the data source, it is hard to have a

* Corresponding author.

E-mail addresses: xinwang11@umbc.edu (X. Wang), azimkhan22@umbc.edu (A. Khan), jianwu@umbc.edu (J. Wang), gangopad@umbc.edu (A. Gangopadhyay), carl.e.busart.civ@army.mil (C. Busart), jade.l.freeman2.civ@army.mil (J. Freeman).

quick response when data injection for some time-sensitive applications like earthquake warnings and automatic driving. Since both edge and cloud resources have their advantages and disadvantages, a related computing paradigm like Edge-to-Cloud Continuum [9] has been proposed to integrate edge with cloud. In an edge–cloud integrated framework, the computation involves both front-end on-premise edge resources like Raspberry Pi and NVIDIA Jetson Nano, and back-end computing resources like big data and GPU clusters in cloud.

Deep learning has been widely used in stream analytics in IoT, edge or cloud environments. As a recent survey paper [10] shows, about one third of studies surveyed in the paper employs recurrent neural network (RNN) based deep learning models for time series or sequence data prediction and forecasting. RNN models can help learn temporal dependence and structures like trends and seasonality. Most existing studies and systems, such as [5,10], only support deep learning based inference on IoT/edge devices. A new research area is how to best integrate both edge resources and cloud resources for deep learning applications. Several researchers [9,11–17] have proposed solutions and frameworks for streaming data analytics that leverage the capabilities of cloud services. However, to integrate edge with cloud, we need to achieve a proper trade-off between latency and accuracy for stream analytics between edge and cloud resources.

Accuracy and latency are two common metrics in stream analytics and many studies have how to balance them or make trade-offs. In the paper, we focus on how to achieve good accuracy and latency for the RNN-based deep learning model in an edge–cloud integrated environment by addressing the following two challenges. First, while the existing studies like [9,11,12] provided promising direction, it is still not clear how to best deploy RNN-based deep learning models in edge and cloud resources for stream analytics to achieve better latency. Second, even though there have been many studies [1,18] on how to deal with unknown or changing data distributions in stream data, a.k.a. concept drift, it is still an open question how to balance accuracy and latency for RNN based stream analytics in an edge–cloud environment.

To tackle the above two challenges, we propose a novel edge–cloud integrated framework and its corresponding open-source modules [19] for stream analytics. To the best of our knowledge, our work is the first to achieve hybrid RNN-based deep learning for stream data in an edge–cloud integrated environment. Our contributions are summarized as follows.

- We propose a novel edge–cloud integrated framework for stream analytics that supports low latency inference on the edge and high capacity training on the cloud. Tasks like data injection, model inference and synchronization are encapsulated as modules and can be flexibly deployed on either an edge device like Raspberry Pi or a cloud resource like AWS.
- Based on users' preferences, we propose three flexible deployment modalities for our hybrid learning framework: edge-centric, cloud-centric and edge–cloud integrated. Based on a modular design, the hybrid learning framework can still work even if parts of the cloud services or edge analytics are unavailable. We further measured the latency differences between the three deployments using a real-world stream analytics application. Our experiments show the proposed edge–cloud deployment is among the best in terms of latency for inference, also will not run into capacity limitation for training.
- To adapt the concept drift challenge of stream data in edge–cloud integrated environments, we propose an adaptive hybrid learning framework that combines and benefits from both cloud resources' high capacity and edge resources'

low latency. Our hybrid learning framework contains batch learning by employing a pre-trained RNN model from large historical data, speed learning by periodically re-training an RNN model from most recent data and hybrid learning by combining predictions from batch and speed learning. We also study a new hybrid learning algorithm that can combine results dynamically. Our experiments show our hybrid learning approaches can have better RMSE than cloud-based batch learning and edge-based speed learning in most cases and our dynamic learning approach performs the best among all learning approaches for all three concept drift scenarios.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the related background our work is built on. Section 3 provides an overview of the proposed edge–cloud integrated hybrid learning framework. Section 4 introduces our three flexible deployment modalities of hybrid learning framework, including edge-centric, cloud-centric and edge–cloud integrated deployments. The adaptive hybrid stream analytics and its two weight combinations, namely static and dynamic weighting algorithms, are explained in Section 5. Evaluations and benchmarking results are next discussed in Section 6. We summarize related studies and compare them with our work in 7 and conclude in Section 8.

2. Background

2.1. Edge computing and edge–cloud integration

Applications that utilize IoT devices are increasing day by day and data volumes produced by IoT edge devices could be enormous. In order to alleviate the heavy load of data transfer, edge devices can pre-process, analyze and quickly react to the time-sensitive application near data sources, and only deliver the processed data or inference results to back-end computation centers. So, when data is handled by an edge device that is close to data generation source, we could achieve faster response time, higher computing efficiency and lower network traffic in comparison to the case where IoT data is processed in a centralized computation location. However, the capacity of edge devices limits their capability to handle complex heterogeneous data and even could lead to unacceptable and unpredictable performance. To deal with the challenge, work at [13] extended the idea of computing continuum, and proposed an edge-to-cloud integration to support dynamic and data-driven application workflows, which are capable of reacting to unpredictable and heterogeneous real-time data. Pilot-Edge [9] proposed its abstraction to support data and machine learning (ML) applications in the edge-to-cloud continuum, which was designed to address the challenge of computation performance in heterogeneous edge environments.

2.2. RNN models on edge devices

In recent years deep learning (DL) has gained attention due to its ability to facilitate analytics in the IoT domain [10]. Sequence DL models like RNN and LSTM are useful for streaming data prediction since these models can learn hidden features from a sequence of records. Hermans et al. [20] state that considering the architecture and the functionality of RNNs, the hidden layers in RNNs are supposed to provide a memory instead of hierarchical processing of features. LSTM, as a special form of RNN, uses the concept of gates to actively control the memory cell and prevent perturbation from irrelevant inputs. The work by Chung et al. [21] show that LSTM models perform better than RNN models when data is characterized by a long dependency like the observations

from IoT applications. Tao et al. [22] use LSTM architecture and mobile phone sensor data for human activity recognition.

More advanced sequence models have also been proposed for stream analytics. Zhang et al. [23] propose a multi-head convolutional neural network with multi-path attention to detect human activity signals received from the wearable sensors. These experiments are carried out on a local computer rather than the edge device and the authors mention their attention models are computationally expensive.

The above studies only support a pre-trained RNN model. As an initial work that supports RNN model update based on more recent data, in this paper, we study a lighter weight LSTM model on the edge device. In future works, we will explore more complicated sequence models which use attention and study how to best enable model inference and updates with limited resources at the edge.

2.3. Concept drifts in real-world IoT data streams

In real-world data-driven applications, analytics of IoT streaming data often encounters the change in the data distribution while extracting different features from stream sources. These hidden changes in the concept or distribution of streaming data, which are unknown to the learning algorithms, are termed as concept drift [6,7] or nonstationary data. Mathematically, if we denote X as an input vector and y as an output vector, then (X, y) will be an infinite sequence of data streams. Concept drifts between time point t_i and time point t_j can be defined as

$$p_{ti}(X, y) \neq p_{tj}(X, y) \quad (1)$$

where p_{ti} and p_{tj} denote joint probability distribution at time t_i and t_j , respectively.

Changes in streaming data distribution over time might appear in various ways such as gradual drift and abrupt drift. Abrupt drift happens suddenly by switching from one concept to another in any time period [7]. Gradual or incremental drift does not change abruptly, instead happens over a long period and therefore can be expected. It defines a continuous change that happens from one underlying process behavior to another one. In this paper, simulated datasets consisting of gradual and abrupt drift were used to know how hybrid stream analytics reacts in the context of different types of drifts.

2.4. Adaptive learning and Lambda architecture

Because underlying concepts of real-world stream data could evolve over time, adaptive learning algorithms have been proposed to address concept drift by adapting new instances and forgetting old ones in order to naturally follow drifts in the stream. It can also be considered as improved incremental learning algorithms that are able to integrate fresh data during their operation to react to concept drifts [7]. Mentioned by [24], concept drift detector, sliding windows, online learner and ensemble learners are the most common adaptive learning approaches. One challenge is, the estimation of performance feedback is difficult for any adaptive learning system due to the absence of ground truth in stream data. Besides, the anomalies of the algorithm can readily be confused for changes in the stream data. In our paper, we infer and evaluate our adaptive learning method by analyzing earlier historical data and the data in the past time windows.

Lambda architecture is a data-processing design pattern which is usually used in data-driven applications by taking advantage of both batch and stream processing methods [1]. The lambda architecture has three layers, batch layer for batch processing based on historical data, speed layer for real-time stream processing, and serving/hybrid layer for combining outputs from both

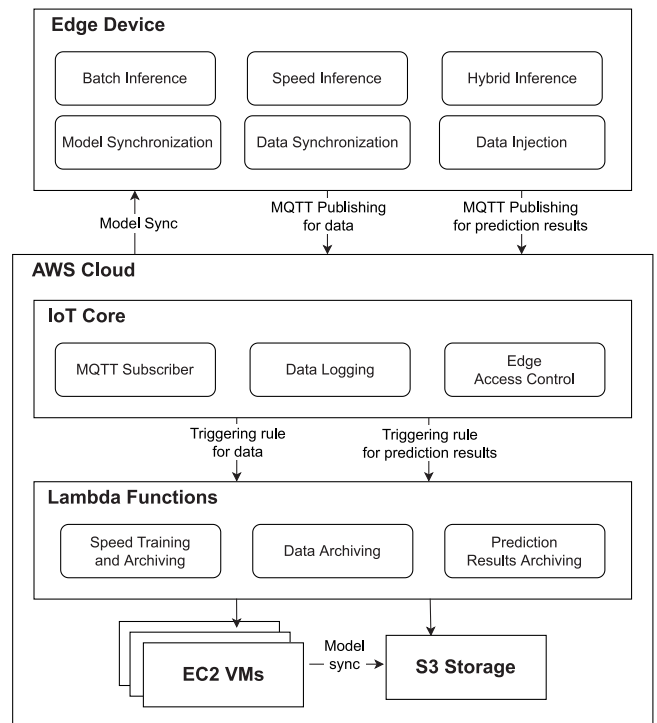


Fig. 1. The overview of our proposed hybrid stream analytics framework.

batch and speed layer. The goal for lambda architecture design pattern is to abstract and balance both the accuracy by using batch processing to provide comprehensive knowledge from historical data, and the latency by using stream processing to learn the recent changes from real-time data. Inspired by this design pattern, we propose a hybrid stream learning model to achieve adaptive learning.

3. Overview of hybrid stream analytics framework

In this section, we briefly introduce our proposed hybrid stream analytics framework from a high-level view. By combining edge resources' lower communication latency with cloud resources' higher computational power, we propose a novel hybrid learning framework that can achieve good latency and accuracy for stream analytics.

We summarize our hybrid stream analytics framework in Fig. 1. In our design, different functionalities are wrapped into multiple modules, which can be deployed on either edge or cloud. Within the edge side, six modules are designed for flexible and dynamic stream analytics. The first three modules, namely *batch inference*, *speed inference* and *hybrid inference*, are the main functionality for the inference task of stream analytics. When stream data is injected, batch inference provides batch predictions based on a pre-trained model from historical data; speed inference enables predictions based on the latest model trained from the previous time window; and hybrid inference combines their inferred values to get a new prediction value. We will explain in detail how we leverage our hybrid learning model to achieve adaptive prediction further in Section 5. Next, we introduce the last three modules on the edge side. For *model synchronization*, it synchronizes the models for speed inference from cloud to edge periodically. For *data synchronization*, it synchronizes the streaming raw data and all inference results to the cloud storage. All the synchronizations are achieved through the edge–cloud MQTT messaging [25] based on specific topics. Module *data injection* acts

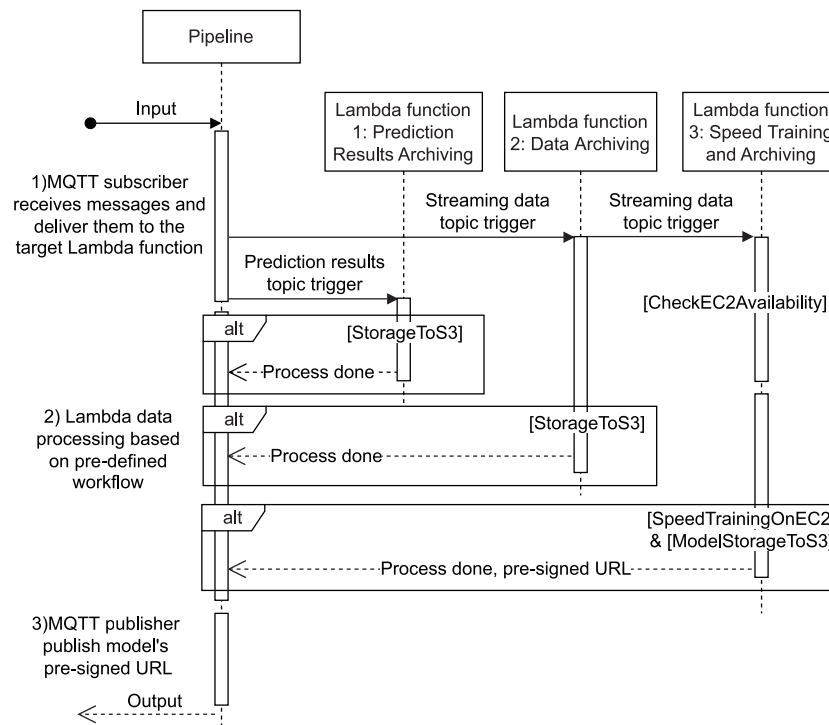


Fig. 2. The system sequence diagram for pipelines of Lambda functions for back-end processing in the cloud.

as a transfer station to throttle the amount of streaming data in each time window and control them to the target modules. All these modularized functionalities can work both independently and cooperatively based on usage. With this modular design, our hybrid stream analytics framework can still work even if some modules are unavailable. Details of how our framework achieves flexibility with different deployment modalities including edge-centric, cloud-centric and edge-cloud integrated scenarios will be explained in Section 4.

Within the cloud side, two resources are used as the back-end of the hybrid stream analytics framework: (1) AWS IoT Core manages the edge-cloud communication and accession, and (2) AWS Lambda Function implements the pipeline of complex data processing. AWS IoT Core provides resources and services that help users achieve edge-cloud computing with AWS IoT-based solutions. Within AWS open source IoT edge runtime, called Greengrass [26], our pre-built modules can be deployed, communicated and managed on the edge through AWS web console or command line. Specifically, defined by AWS access control, all permitted edge-to-edge and edge-to-cloud communications are achieved by MQTT publishing and subscribing protocol. Besides, IoT Core enables triggering rules that provide a SQL-based language to filter MQTT payloads and deliver them to the target services like Lambda Function. As shown in Fig. 2, filtering by the Lambda triggering rules, the incoming MQTT payloads will be delivered to different target Lambda functions as Lambda events (Step 1). In Step 2, Lambda functions will execute these triggered events asynchronously based on their pre-defined pipeline. *Prediction Results Archiving* function only receives events from the inference results topic and directly stores the payloads to AWS S3 object storage. *Data Archiving* and *Speed Training and Archiving* functions both receive events from the streaming data topic. For *Data Archiving* function, just like the first function, it stores the payloads to S3 directly. For *Speed Training and Archiving* function, it will first check AWS EC2's availability, deliver streaming data to an EC2 virtual machine for model training, and then upload the latest model to S3 when training finished. In the meanwhile, in

Step 3, this Lambda function will also publish a one-time pre-signed S3 URL to the edge. This S3 URL is signed with cloud credentials, which grants temporary access to the edge's model synchronization.

4. Flexible deployments of hybrid stream analytics framework

To achieve the flexibility of the proposed hybrid stream analytics framework, we use a modular design for all framework components, which achieves a proper trade-off between latency and accuracy for stream analytics. Based on different scenarios, we design three types of deployments for the hybrid stream analytics: edge-centric, cloud-centric and edge-cloud integrated deployments. The summary of the three deployment modalities is shown in [Fig. 3](#). We also summarize the advantages and limitations of the proposed deployments in [Table 1](#).

4.1. Edge-centric stream analytics deployment

Because stream analytics needs to process incoming data continuously, it is common that the back-end cloud service will be temporally unavailable due to network disconnection or resource overload problems. For this, we design an edge-centric deployment modality, which allows the edge to execute stream analytics autonomously with local events, as shown in Fig. 3a.

We can summarize the unavailability of the cloud into two scenarios: part of cloud computational resources (like EC2) is unavailable and the whole cloud service is unavailable. For the first scenario, IoT Core and other Lambda functions still work well except for the *Speed Training and Archiving*. If the Lambda function cannot connect to any virtual machine in EC2, it will put the process event into its waiting queue and wait for the available resources. At this time, although other services still work fine, the performance of speed inference may not have good accuracy since it still uses an “out-of-date” model trained by the data from the time window before the unavailability of EC2.

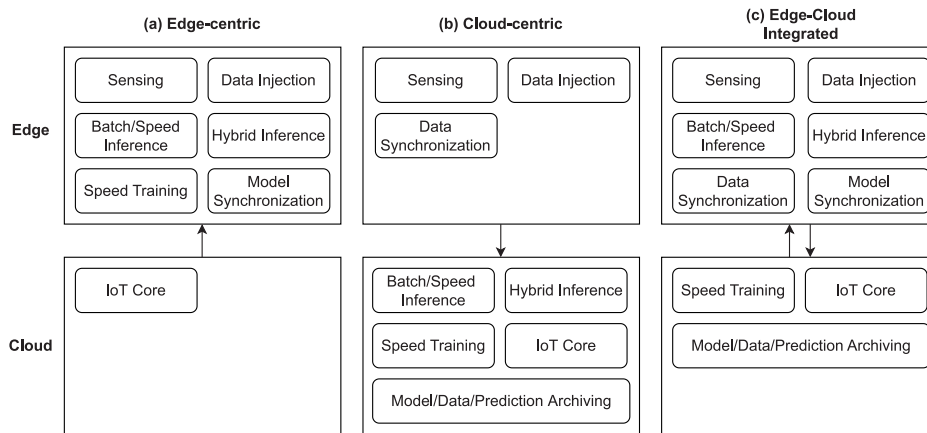


Fig. 3. Flexible deployments of our hybrid stream analytics framework.

Table 1

Advantages and limitations of the proposed three deployment types for stream analytics.

	Advantages	Limitations
Edge centric	Quick respond since the computation is near to the source of data	Capacity and capability shortage for edge device
Cloud centric	Enough capacity and capability for high accuracy computation	High communication overhead between edge and cloud
Edge–cloud integrated	Quick respond for inference and high accuracy for training	Complex coordination between edge and cloud

When the whole cloud service is unavailable, that means the edge cannot get any update from the cloud. In this scenario, all the flexible modules and the functionalities of Lambda Function are wrapped into AWS Greengrass runtime and deployed to the edge side. Based on the usage, if the stream analytics framework is only assigned to do model inference, the batch, speed and hybrid inference modules will wait for the data from data injection and output the results separately. If the stream analytics framework also requires to do model training, the speed training module can be deployed to the edge device, subscribe to the data injection, fulfill speed training and synchronize the new model for next time window inference. To achieve this usage, MQTT publishes and subscribes messages locally within the edge device. Specifically, the speed training module is in a containerized Spark [27] based design. By default, it will initiate the training environment from a pulled docker image and allocate available edge resources to train the model in a Spark standalone mode. Our future work will study how this Spark-based speed training module can be extended to different edge devices as a distributed master-worker computing. If there are idle edge devices, Spark will initiate these edges as workers and auto-scale the training tasks to them. This Spark-based parallel design of the speed training module also avoids the issue of limited computational capability of individual edge devices.

4.2. Cloud-centric stream analytics deployment

To achieve the flexibility of the hybrid stream analytics framework, we also provide a cloud-centric deployment for stream analytics. In this scenario, as shown in Fig. 3b, the edge device is only used to sense the streaming data and synchronize them to the cloud.

When the edge device cannot perform any data processing, the batch, speed, hybrid inference and model synchronization modules should be deployed on cloud computational resources like EC2. At this time, IoT Core service will mark the EC2 virtual machine (VM) as the substituted edge computational capability and subscribe to the MQTT payloads from EC2.

Leveraged by Lambda functions, our cloud-centric deployment achieves automatic data processing for each time window of stream analytics in the back-end cloud. Like the introduction in Section 3, for the hybrid stream analytics framework, when an incoming MQTT payload triggers a filtering rule, IoT Core invokes corresponding Lambda functions asynchronously and passes the data payload from edge to the specific function. After that, based on the pre-defined functions of data processing pipeline, Lambda will check the availability of AWS EC2 virtual machines, train the streaming data in EC2 and archive the model to the S3 object storage. Later, Lambda will reply with a one-time pre-signed S3 URL to edge, which grants the model synchronization module temporary access to synchronize the latest model from cloud.

4.3. Edge–cloud integrated stream analytics deployment

We first make a short summary about the advantages and limitations of the first two proposed deployments, as shown in Table 1. For the edge-centric deployment, although all the hybrid stream analytics can run closer to the data sources, the limitation is that the weak computational capability and capacity of edge devices may cause process congestion or even crash during stream analytics. Whereas for the cloud-centric deployment, all the data cannot be pre-processed before it arrivals to cloud. In another word, the edge-centric deployment focus on the quick on-site response of the stream analytics and the cloud-centric deployment mainly benefits of the computing power and storage capacity of cloud.

In order to achieve the proper trade-off between these two deployment modalities, we propose a third deployment modality, namely edge–cloud integrated deployment. As shown in Fig. 3c, with edge–cloud integrated deployment, all the inference and synchronization modules are developed on the edge, while speed training and all the archiving are developed on the cloud. With this edge-centric deployment solution, hybrid stream analytics can enjoy not only the computing power and storage capacity of cloud, but also the low latency for edge resources.

Table 2
Cloud service mapping for hybrid stream analytics.

Service category	Service description	Amazon AWS	Microsoft Azure	Google cloud
Virtual machine	Virtual instance that enables to host speed training.	EC2	Virtual machines	Compute engine
IoT platform	Manage the edge–cloud communication and accession.	IoT Core	IoT Hub	Cloud IoT
IoT runtime	Help edges to build, deploy and manage the application.	Greengrass	IoT edge	Cloud edge
Container service	Store, manage, and secure container images in private or public.	ECR	Azure Container Registry	Artifact Registry
Object storage	Store, manage, and secure any amount of data in storage.	S3	Blob storage	Firebase
Serverless	Run and manage the application with zero server management.	Lambda functions	Azure functions	Cloud functions
Cloud Python SDK	Easy-to-use interface to access cloud services.	Boto/Boto3	.NET Core	Cloud SDK

4.4. Flexible deployment of our framework

In our hybrid stream analytics framework, we have six modules implemented as Python functions. For the flexible deployment, we use different ways to wrap these modules to achieve proper coordination and trigger their invocations based on incoming stream data. Specifically, we use AWS IoT Component with its Greengrass runtime for edge-based deployment and AWS Lambda function for cloud-based deployment. To deploy a module on edge, AWS IoT Component is required with an update interval configuration so the modules can be triggered by an AWS IoT event and the records of the time windows periodically. To deploy a module on AWS, it can be encapsulated into the docker container with its software environment. In this way, the same modules and implementations can be reused when switching from one deployment to another.

4.5. Extensibility of our framework

In this paper we implement the hybrid stream analytics framework on AWS cloud, however the proposed framework can be easily extended to other cloud providers. Most services from different cloud providers can be mapped to each other. Table 2 lists related cloud services provided by Amazon AWS, Microsoft Azure and Google Cloud for hybrid stream analytics. In order to achieve extension to Microsoft Azure and Google Cloud, the user needs to wrap the flexible modules into its corresponding IoT runtime for each cloud. All functionalities can be wrapped into the Greengrass runtime for AWS, the IoT Edge runtime for Azure, and IoT Cloud Edge for Google. Additionally, the serverless functions are needed to adapt to the specific structure and format of each cloud.

5. Adaptive and dynamic hybrid learning model for stream analytics

In order to tackle the challenge of concept drift, we propose a hybrid learning model, which can adapt to the changes in stream analytics by weighted combining the results from batch and speed inference. Like the design pattern of Lambda architecture, hybrid stream analytics should contain a batch layer, a speed layer and a serving/hybrid layer. In this section, we will first provide an overview of our hybrid learning model. Then, we introduce the orchestration of the hybrid stream analytics, and its two weight combination algorithms, namely static weighting algorithm and dynamic weighting algorithm.

5.1. Overview of adaptive hybrid stream analytics

Leveraging the lambda architecture, our hybrid stream analytics achieves adaptability of stream data concept drift. We first introduce problem statements of our hybrid stream analytics. In our hybrid learning model, we separate the inference tasks of stream analytics into three layers: batch layer, speed layer and hybrid layer.

Batch layer tasks. For the batch layer, our hybrid learning model only trains the model once and reuses it for inference all received stream data. Its model is defined as

$$\hat{y}^i = f(y^{i-1}, y^{i-2}, \dots, y^{i-n}) \quad (2)$$

We call the training in batch layer as the batch training and its inference as batch inference.

Speed layer tasks. For the speed layer, there is no pre-trained model before the stream analytics begins. Instead, the speed layer re-trains a new model for every time window and uses it to infer the next time window data. We define the inference task as follows. For each time window t , the stream analytics trains a model f_t and uses it to make predictions for the new time window $t + 1$. For each timestep i within time window $t + 1$, the prediction value \hat{y}_t can be defined as

$$\hat{y}_{t+1}^i = f_t(y_{t+1}^{i-1}, y_{t+1}^{i-2}, \dots, y_{t+1}^{i-n}) \quad (3)$$

We call the training in speed layer as the speed training and its inference as speed inference.

Hybrid layer task. For hybrid layer, in order to aggregate the inference results from both consistent patterns of historical data distribution and the hidden changes of streaming data distribution, its model works based on formula

$$Pred_{hybrid} = W^s * Pred_{speed} + W^b * Pred_{batch} \quad (4)$$

where the weights $W^s + W^b = 1$. The hybrid layer only has inference (no training), so that we call it as hybrid inference.

Because model training happens only once for the batch layer, referred to as Fig. 4, the latency of batch training is not part of the latency occurred for incoming streaming data. Instead, we only focus on the latency for batch inference, speed training, speed inference and hybrid inference for each time window in the paper. Also, we run each module asynchronously to lower overall latency.

5.2. Orchestration of hybrid stream analytics

We explain how the modules of the hybrid stream analytics orchestrate, which is illustrated in Fig. 4. There exist a one-time batch training before the stream analytics start. After that, the data injection module acts as a transfer station to throttle the amount of sensing streaming data into a payload for each time window, for example, catching streaming data every 30 s. With data injection throttling, incoming streaming data can be temporarily stored in a buffer queue which avoids the receiver from the crash when absorbing the peaks of incoming data for a very short time lapse. Then, the data injection module delivers data based on the usages of stream analytics, which contains two asynchronous phases: training phase and inference phase.

In the training phase, stream analytics executes the speed training based on the stream data in each time window. After receiving raw streaming data, the speed training module trains a new model based on the current payload batch. Then this new trained model will be synchronized to the speed inference module for the prediction of the later stream analytics.

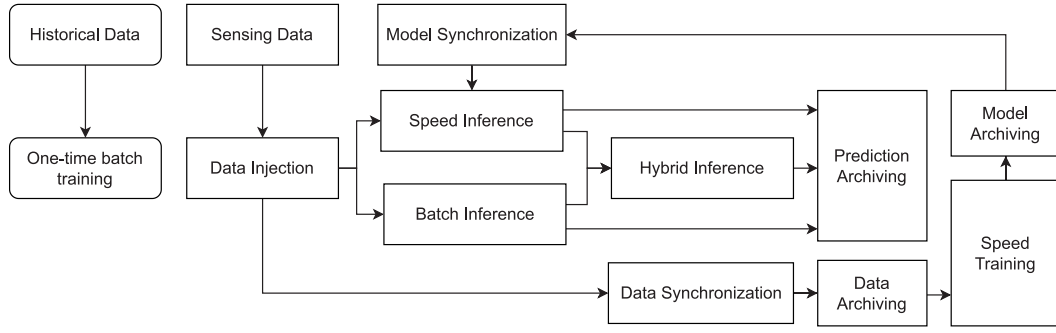


Fig. 4. Module orchestration of our hybrid stream analytics (rectangular boxes denote periodic operations and round boxes denote one-time operations).

In the inference phase, the stream analytics pipeline requires batch and speed inference, and the data injection module will deliver the streaming payload to the inference modules in each time window. With the design pattern of lambda architecture, batch inference module predicts results using a pre-trained model as the batch layer, which is learned from the historical dataset. As the speed layer, speed inference module updates its model for each time window, which uses a model learned from the previous time window and tests it in the current time window. And hybrid layer will aggregate the inference results both from the consistent patterns of historical data distribution and the hidden changes in streaming data distribution in both batch and speed layers. The hybrid inference results will later be published by the MQTT publisher for archiving and further notification.

5.3. Hybrid learning model with weight combination

In this section, we introduce the two weight combination algorithms for hybrid inference.

Static Weighting Algorithm. The weight combination algorithm can be defined as the static weighting algorithm if the weights W^s and W^b (in Eq. (4)) are been set as a fixed value for every time window of the stream analytics. Because of the fixed weights, it is obvious that the hybrid learning model with the static weighting algorithm is hard to adapt to the dynamic changes of streaming data. To solve this problem, we provide an optimized approach to dynamically learn the weights during stream analytics, namely dynamic weighting algorithm.

Dynamic Weighting Algorithm. In theory, finding the dynamic weights is a mathematical optimization problem that is used to find the best solution from all feasible solutions. Shifting it to a machine learning problem, stacking ensemble methods [28,29] combine multiple machine learning algorithms to obtain a better predictive performance than that could be obtained from any of the constituent learning algorithms alone. Based on these ideas, we propose our dynamic weighting algorithm.

As shown in Algorithm 1, for each time window t , taking the inputs of batch layer model M^b , speed layer model M_{t-1}^s at time window $t - 1$ with the test dataset X_{t-1}^{test} , the dynamic weighting algorithm stacks the provided models and collects their predictions using the test dataset as the serving layer. By listing the constraints and bounds, like limiting the sum of weights to equal 1 ($W_t^b + W_t^s = 1$) and limiting the weights W_t^b, W_t^s in range $[0, 1]$, the optimization solver will find the optimum values that can minimize the objective loss function, starting from an initial guess W^{init} (we choose 0.5 as our initial weights). In our paper, we use Sequential Least Squares Programming (SLSQP) [30] as the optimization solver *Solver*, which is always used to solve nonlinear programming (NLP) problems. We also use Root Mean

Algorithm 1: Dynamic Weighting Algorithm (DWA)

Input: $M^b, M_{t-1}^s, X_{t-1}^{test}, Y_{t-1}^{test}$

Output: W_t^b, W_t^s

function DWA():

$EnsembleModels \leftarrow []$

$Pred \leftarrow []$

$EnsembleModels.append(M^b, M_{t-1}^s)$

for $model$ **in** $EnsembleModels$ **do**
 | $Pred.append(model.predict(X_{t-1}^{test}))$

end

$W^{init} = [0.5] * len(Pred)$

$cons = lambda W : 1 - sum(W)$

$bounds = [(0, 1)] * len(Pred)$

$loss = LossFunc(Y_{t-1}^{test}, Pred)$

$W_t^b, W_t^s \leftarrow minimize(loss, W^{init}, bounds, cons, Solver)$

return W_t^b, W_t^s

Squared Error [31] regression loss as our loss function *LossFunc* that can be defined as

$$L_{rmse}(y) = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (5)$$

which is the square root of the average of squared differences between prediction \hat{y}_j and actual observation y_j .

In our algorithm implementation, at each time window t , we stack two pre-trained models in the serving layer, which include one speed layer model at time window $t - 1$ and the batch layer model. Since the ensemble method does not require a constant pattern for stacking models from batch or speed layer, the dynamic weighting algorithm also has its variants like stacking the most recent n speed layer models or stacking speed layer models continuously. We will study these variants as part of our future work.

6. Evaluation

This section conducts the evaluation of our proposed flexible and dynamic hybrid stream analytics framework. We implemented our framework and open-sourced it on GitHub at [19]. One real-world and two synthetic datasets are applied in our experiments and the metric includes latency and accuracy. The evaluation compares the difference in two aspects: (1) different types of hybrid learning framework deployment and (2) different types of hybrid stream analytics approaches.

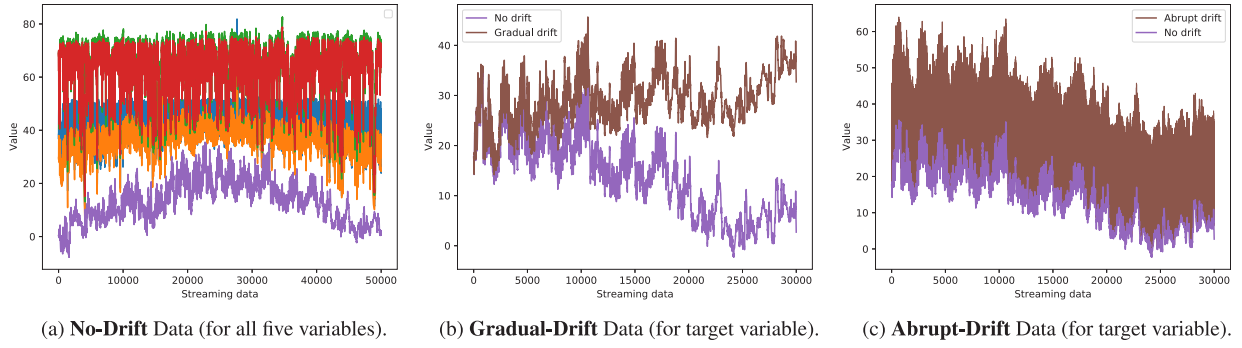


Fig. 5. Data distribution for one real-world and two synthetic time-series of wind turbine temperature.

We note that our experiments have implemented approaches from several related works as baselines for specific capabilities: (1) For hybrid time series forecasting, we implemented the static approach in reference [1] as a baseline to evaluate the advantage of our proposed dynamic capability (see Section 6.3.2 and Fig. 8); (2) For hybrid inference, we implemented reference [2] as the baseline for batch inference modality to evaluate the advantages of our dynamic hybrid inference modality (see Section 6.2 and Table 3).

6.1. Datasets and evaluation settings

6.1.1. Dataset description

We use one real-world dataset and two simulated datasets for gradual drifts and abrupt drifts, in order to evaluate our proposed edge-cloud integrated framework. The data distributions of each dataset are shown in Fig. 5.

One real-world dataset. Our application is designed for the real-world prediction of wind turbine temperature, based on the ENGIE's open wind farm data [32]. For the actual data distribution of wind turbine temperature, as shown in Fig. 5(a), we use one turbine time series (from five temperature sensors) from January to December in 2017, recorded every 10 min, which has around 50,000 observations in total. In order to check concept drifts and data stationary for each variable in actual time-series, we perform the augmented Dickey-Fuller test [33], which is used to determine how strongly a series is defined by a trend by calculating the corresponding p -value [34]. The null hypothesis of the test is that the tested series have a certain time-dependent structure (namely not stationary). The p -values of the five variables, namely Db1t_avg, Db2t_avg, Gb1t_avg, Gb2t_avg and Ot_avg, turn out to be 1.82×10^{-22} , 3.34×10^{-17} , 3.44×10^{-20} , 2.38×10^{-17} and 4×10^{-6} , respectively. Since these values are less than 0.05, we can reject our null hypothesis and conclude that the time series is stationary without concept drifts. We use this actual dataset in our no drift scenario.

Two synthetic datasets. In order to evaluate the adaptiveness of our hybrid learning dealing with streaming data concept drifts, we synthetically generate two datasets and simulate gradual drifts and abrupt drifts on each of them, as shown in Figs. 5(b) and 5(c).

Let $GD_i(t)$ and $AD_i(t)$ be the generated gradual and abrupt drift value of target variable at timestamp t , and $Y_i(t)$ be the true value of input feature, where $i \in [0 \dots n]$. For gradual drift scenario and abrupt drift scenario, the simulation rule for all n variables is specified as Eqs. (6) and (7) separately, where α_i is the drift value for variable i , ε is an invariant noise and λ is the random abrupt parameter.

$$GD_i(t) = \alpha_i t + Y_i(t) + \varepsilon \quad (6)$$

$$AD_i(t) = \alpha_i t \lambda + Y_i(t) + \varepsilon \quad (7)$$

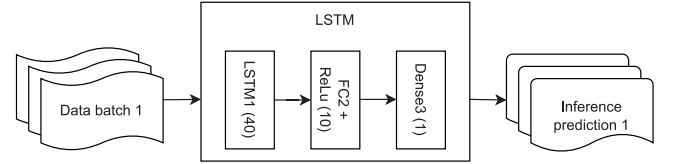


Fig. 6. LSTM network architecture.

6.1.2. Machine learning setting

For the evaluation of no-drift, gradual drift and abrupt drift scenario, we split the modeling dataset into training and testing subsets with the ratio of 4 : 6. We use 20,000 observations to produce a pre-trained model for batch inference, and send 30,000 observations as streaming data in each time window to test our hybrid learning analytics. The data from all five variables are been normalized using Min-Max Scaling to the range of [0, 1] during computation.

Settings for model training. We run a multilayer-perception long short-term memory (LSTM) network shown in Fig. 6, which has one long short-term memory layer with 40 units, one fully connected layer with 10 units and ReLu activation, and one final output layer (10,981 total parameters). For the pre-train model used in batch inference, we train the model using 50 epochs and 512 batch sizes with a 0.001 learning rate. For the speed model in each time window, we use 100 epochs and 64 batch sizes with a 0.001 learning rate. Because this study focuses on hybrid learning and its deployment on edge-cloud resources, we did not employ more complicated RNN deep learning models, which can be easily evaluated in future work.

Settings for model inference. Our batch inference loads the pre-trained model in each time window and makes predictions for the records in the current time window. We set the time window size equal to 30 s in all experiments and throttle no less than 200 records in each time window. For speed inference, we have three parallel processes, which include (1) fetching the latest pre-trained model from S3 and saving it to the edge disk; (2) loading the latest pre-trained model from the edge disk and making a prediction for the current time window; and (3) training a new LSTM model based on the current time window and uploading the model to S3. Because the three processes run in parallel, we cannot guarantee to use the model trained from the previous time window to infer the current time window. But this approach can improve the latency greatly. For the hybrid inference, the predicted value of each record is calculated from its batch and speed inference prediction.

As the problem statement in Section 5.1, assuming values for y_{t+1}^{i-1} , y_{t+1}^{i-2} , ..., y_{t+1}^{i-n} are known when making predictions (same with batch inference without the time window t). We evaluate

Table 3

The latency of inference phase for different stream analytics with three deployment modalities (unit: second).

	Speed Inference			Batch Inference			Serving/Hybrid Inference		
	Computation	Communication	Total	Computation	Communication	Total	Computation	Communication	Total
Cloud centric	8.82	13.88	22.70	8.49	14.52	23.01	23.65	16.47	40.12
Edge centric	10.25	6.83	17.08	10.65	6.61	17.26	27.19	8.52	35.71
Edge-cloud integrated	9.89	6.75	16.64	10.83	7.20	18.03	26.71	8.93	35.64

the prediction performance by calculating RMSE between predictions \hat{y}_{t+1}^i and actual observations y_{t+1}^i . In the paper, we set the time lag to be 5, namely $n = 5$.

6.1.3. Hardware and software setting

Hardware settings. For our experiments, we use a Raspberry Pi 4 as our front-end on-premise edge device, which is attached with the 32 GB MicroSD memory card and 4 GB RAM. We use Amazon Web Services as our back-end cloud platform. A data analytics server is deployed on AWS EC2, which allocate to a compute-optimized c5.4xlarge instance with 16 virtual CPUs (vCPUs) and 32 GB of memory.

Software settings. For the software environment on the edge, we use Debian 11 Bullseye OS with Python 3.8. The dependencies on edge include Tensorflow-lite 2.5, Spark 3.0 and Pandas for inference learning, Kafka 3.1 for data injection, and AWS SDK Boto3 for edge-cloud data and model synchronization. The Kafka data injection bandwidth is around 7 records/second in our experiments. Meanwhile, the software environment on the cloud is encapsulated in our public Docker image, which contains Tensorflow 2.2, Spark 3.0 and Pandas for model training and also Boto3 for synchronization. Both our software environments support the Spark big data analytics engine, which enables parallel computation on two sides.

6.2. Latency evaluation for different deployment modalities

We first evaluate the performance of hybrid learning framework with the three deployment types explained in Section 4. Since the deployment modalities, including edge-centric, cloud-centric and edge-cloud integrated, only change the resources where the modules are deployed in, the stream analytics still executes based on the same logic which results in the same accuracy performance. Therefore, we only evaluate their latency.

We separate the pipeline of stream analytics into two phases: inference phase and training phase. These two phases work asynchronously as illustrated in Fig. 4. In inference phase, starting from data injection to prediction archiving, we record both computation latency and communication latency for each time window, then we calculate their averages over all time windows and show the results in Table 3. The table shows edge-centric and edge-cloud integrated deployment are more efficient than cloud-centric deployment because of their small communication overheads. For edge-centric and edge-cloud integrated deployment, referred to as Figs. 3a and b, they have roughly the same latency in inference phase since their module deployments are exactly the same, except the speed training. Next in the training phase, starting from data injection to model synchronization, we only measure the average computation and the communication latency for speed layer, since batch layer only trains a model once and hybrid layer does not have the training phase. For cloud-centric deployment, the average latency of speed layer are 14.73 s for computation, 14.47 s for communication, and 29.20 s in total. For the edge-cloud integrated deployment, the average latency are 15.69, 14.04 and 29.73 s, respectively. These two deployment modalities perform in the same trend since their speed training modules are both deployed in cloud. For edge-centric deployment, referred to as Fig. 3a, the speed training

module should be deployed in edge resource. We also evaluated the edge-centric deployment with our Raspberry Pi edge device, but the experiment failed with out-of-memory error. It shows the edge device with a limited capacity cannot support this type of deployment. So, if we compare the total latency (inference and training), edge-cloud integrated deployment is the best.

In summary, for three deployment modalities, edge-cloud integrated deployment works best, as its efficiency in inference phase and the sufficient capacity in training phase. Specially, comparing with the other two deployments, the edge-cloud deployment can achieve similar latency performance as edge-centric deployment without worrying about capacity limitations. Therefore, for the rest of our evaluation, we only conduct experiments with the edge-cloud integrated deployment.

6.3. Latency and accuracy evaluation for different stream analytics approaches

We focus on both latency and accuracy aspects when evaluating our adaptive hybrid stream analytics approaches. For latency, we measure overhead created by stream analytics. For accuracy, we compare its performance with the proposed dynamic weighting algorithm in different streaming concept drift scenarios. For the dynamic weighting algorithm, we evaluate the stacking of two pre-trained models (one latest speed model and one batch model) as explained in Section 5.3.

6.3.1. Latency evaluation

We first discuss the latency of hybrid stream analytics with static weighting and dynamic weighting. As shown in Fig. 7, we record the latency of execution in every streaming time window which includes around 200 streaming observations. The latencies of speed and batch inference are in the same trend and they are both lower than the latency of hybrid inference. Since speed and batch inference are executing in parallel and their latencies have overlap, we also evaluate the total latency for the whole hybrid stream analytics.

With static weighting in Fig. 7(a), the latencies averaged from all time windows are: 10.43 s for speed inference, 9.93 for batch, 15.81 for hybrid and 26.63 for total, respectively. And with dynamic weighting in Fig. 7(b), the average latencies are 10.25, 10.63, 18.34 and 29.19 s, respectively. Since the weight combination algorithm is only applied in hybrid inference, the latencies of speed and batch inference are roughly the same in the two evaluations. For the hybrid inference and the overall hybrid stream analytics, the percentage of average latency increment of dynamic weighting turn out to be 14.82% and 9.54%, compared with static weighting. The increment is because our dynamic approach requires time to find the best weights.

6.3.2. Accuracy evaluation

To evaluate accuracy performance of hybrid stream analytics, we use Root Mean Squared Error (RMSE) metric to measure how far the predicted values \hat{y}_j are from the ground-truth values y_j , as mentioned in Eq. (5). We compare the performance of hybrid stream analytics in three data drifting scenarios with two weight combination algorithms. We record the RMSE of inference results for each streaming time window (100 time windows in total),

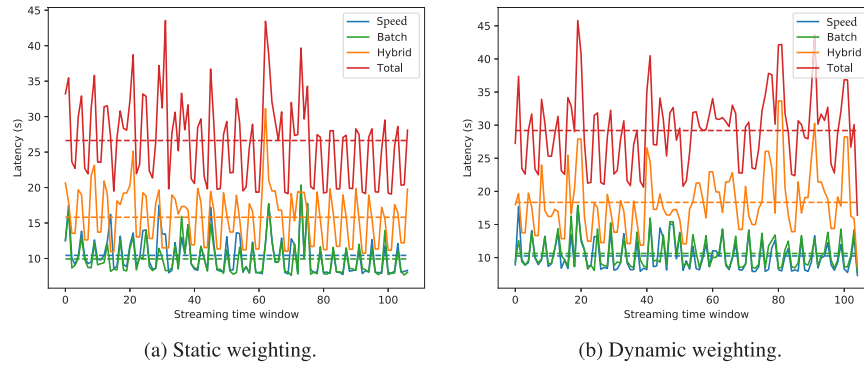


Fig. 7. Inference and total latency of hybrid stream analytics for edge-cloud integrated deployment.

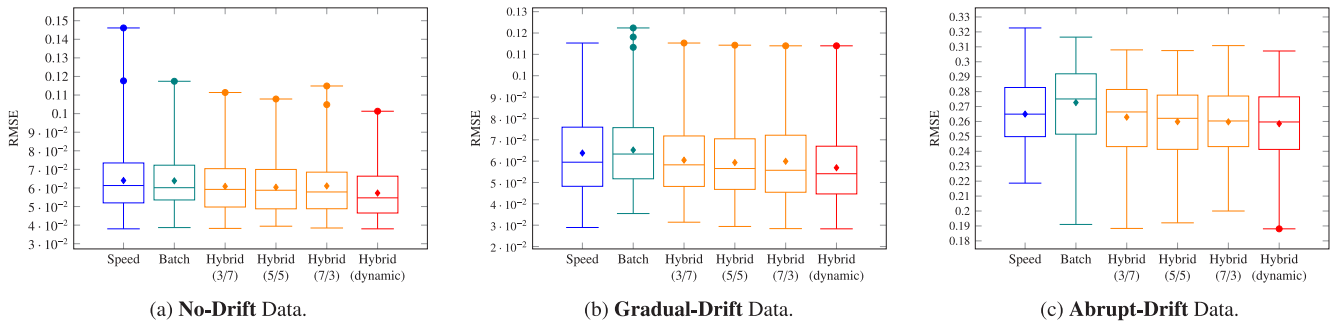


Fig. 8. RMSE box-plots for different inference approaches.

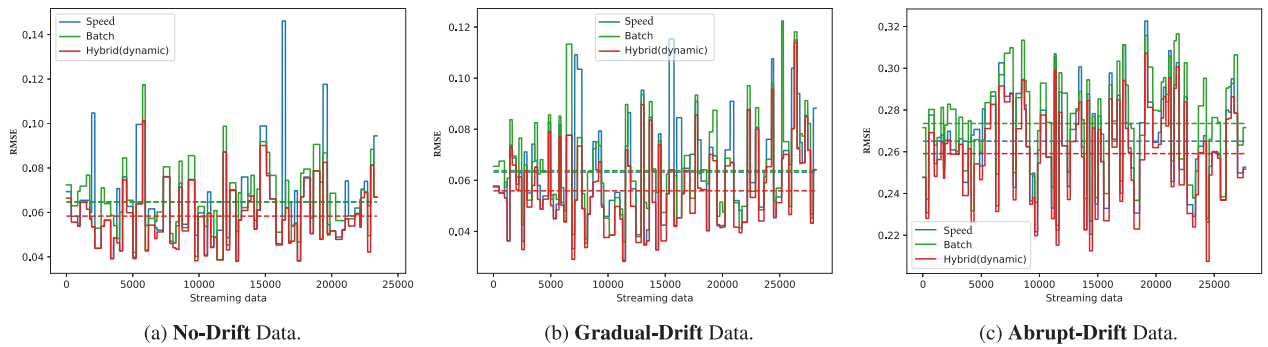


Fig. 9. RMSE of each time window for batch, speed and dynamic hybrid inference.

and convert them into the boxplots, as shown in Fig. 8. For the static weighting algorithm, we also measure the different weights in 3:7, 5:5 and 7:3 (speed:batch) of the accuracy performance evaluation.

Fig. 8 assesses the accuracy of hybrid stream analytics and baseline approaches. In summary, for hybrid stream analytics, both static and dynamic weighting algorithm achieve better RMSE values than speed and batch inference. For no-drift scenario, the batch and speed inference get roughly the same RMSE since there are no unexpected changes in the concept or distribution or the streaming data. For both gradual-drift and abrupt-drift scenarios, the speed inference works better than the batch inference since the latter cannot detect the changes in data distribution based on the model from historical data. On the contrary, speed inference updates the model for each time window, using a model trained from the previous time window to test the streaming data in the current time window, which can catch the drifting in time. Besides, this evaluation also shows the hybrid stream analytics with dynamic weighting algorithm achieves the best average RMSE in all three scenarios, and the improved percentages of RMSE are 10.73%, 12.73% and 5.20% respectively. We also record

Table 4

Time percentage of each inference being the best in terms of RMSE for no-drift data.

	Static (3:7)	Static (5:5)	Static (7:3)	Dynamic
Speed	0.5460	0.4757	0.3311	0.1648
Batch	0.1172	0.1967	0.2578	0.088
Hybrid	0.3368	0.3275	0.4111	0.7472

the RMSE for each time window in all three drifting scenarios, as shown in Fig. 9. It also shows our dynamic hybrid approach achieves the best RMSE for most time windows.

In order to verify above conclusions, we further draw Tables 4–6, which show the percentage of each inference being the best with no-drift, gradual-drift and abrupt-drift data separately. For no-drift scenario, hybrid inference is the best approach only in dynamic weights and one static weighting (7:3) algorithm. For both the gradual-drift and abrupt-drift scenarios, however, hybrid inference is always the best approach for every weight combination algorithm. As a result, our hybrid stream analytics can adapt the gradual and abrupt concept drift effectively.

Table 5

Time percentage of each inference being the best in terms of RMSE for gradual-drift data.

	Static (3:7)	Static (5:5)	Static (7:3)	Dynamic
Speed	0.3472	0.2830	0.1702	0.0513
Batch	0.1093	0.1332	0.2230	0.0252
Hybrid	0.5436	0.5838	0.6068	0.9235

Table 6

Time percentage of each inference being the best in terms of RMSE for abrupt-drift data.

	Static (3:7)	Static (5:5)	Static (7:3)	Dynamic
Speed	0.4839	0.2106	0.1129	0.0290
Batch	0.0000	0.0091	0.0183	0.0000
Hybrid	0.5161	0.7802	0.8687	0.971

7. Related work

There have been many studies related to the topics we discussed in the paper. Based on their framework capabilities and main focused challenges, as shown in Table 7, we categorize them into different groups. Next, we describe the details of these related work from the three aspects we listed in the table.

7.1. Machine learning based IoT stream data analytics

When dealing with concept drift in stream data analytics, we study how to optimally integrate batch learning and stream learning in our paper. Paper [1] proposes a framework that generates improved time series forecasting by supporting batch-based, stream-based and hybrid time series forecasting, to tackle the adaptability challenges. Several papers [2,35–39,43,44] study how to update the model based on streaming data and propose their solutions. Shao et al. [35] propose an adaptive strategy in conjunction with ensemble learning for the task of concept drift detection, while Puschmann et al. [36] use an online clustering mechanism to cluster the streaming data, which remains adaptive to drifts by adjusting itself as the data changes. Yang et al. [37] propose their drift adaptation method algorithm based on the combination of sliding and adaptive window-based methods, as well as performance-based methods. All these studies focus on the algorithm design for training and updating one identical ML/DL model to deal with streaming concept drift with the best performance. Some of these [36,38,39] did not consider the computational power of the edge computing environment, so the algorithms need to be further deployed in additional resources (like storage or computation optimized device). In contrast, the adaptive hybrid stream analytics we proposed is a weight combination solution from two (batch and stream) inferences, which does not re-evaluate or adjust the layers of the neural network based on the results. By periodically applying knowledge from two trained and compressed models, our work is more lightweight and portable for edge devices in real-time concept drift adaptation.

7.2. Edge–cloud integration

There are also several studies [13–17] for workload management on edge–cloud integrated resources, more targeting an optimized cyber-infrastructure design but not much for machine learning related workload. Luckow et al. [9] study how to manage distributed edge and cloud resources and studies the performance of machine learning models for the outlier detection. However, the edge devices in the paper are simulated so it is yet to be seen whether the findings are the same in real-world. Also, the

machine learning models are only deployed on the cloud side, not on the edge side. Osia et al. [11] deploy deep learning models to predict images collected by the edge, where sensitive information is first pre-processed on the edge and its representation is sent to the cloud for complex inferences. Since the edge devices are only used for data pre-processing, not for actual machine learning based inference, the total latency of their framework will be higher than inferring directly on edge. Abdulla et al. [12] argue that using adaptive learning for streaming data processing could solve concept drift problems, and the proposed cooperative fog–cloud architecture shows updating machine learning models periodically can help reduce RMSE by about 20%. However, their experiments did not use portable edge devices such as Raspberry Pi, instead they used a local computer. The edge–cloud integrated framework we proposed is a general and flexible design, whose docker-based modules can be developed in either cloud or edge side even with different types of edge devices.

7.3. Complete system or toolkit for deep learning based inference

There have been many systems or toolkits that support deep learning based inference on IoT/edge devices. NVIDIA's DeepStream [5] is a streaming analytic toolkit that helps the user build and deploy video analytics applications on-premises, on the edge, and in the cloud. DeepStream features hardware-accelerated building blocks [45] that bring deep neural networks and other complex stream data processing tasks into GStreamer processing pipelines and maximize the computation using GPUs. Based on its design, DeepStream can be highly optimized to run on NVIDIA series or GPU-enabled edge devices like Jetson Xavier NX and Jetson Nano. However, while DeepStream has multiple examples that are provided as source code, its SDK is not released as open-source software. An alternative toolkit for deep learning based inference is Google Coral [40]. Coral is a complete toolkit for building intelligent devices with fast deep neural network inferencing. Same with DeepStream, Coral can enable its peak capability with the proposed hardware and software solutions like Edge TPU coprocessor [41].

More focused on deployment rather than inference learning, some Cloud platforms like AWS and Azure also provide their general-purpose solution for edge devices inference even offline from the cloud. AWS IoT Greengrass [26] is an open-source edge runtime and cloud service for building, deploying, and managing edge devices. Greengrass manages and operates multiple edge devices in the field locally or remotely using MQTT or other protocols. With the solution, inference can be deployed across edges using any language, packaging technology, or runtime. Our hybrid learning framework is based on the Greengrass runtime. We further study how to deploy stream analytics among edge and cloud resources and improve their accuracies. Microsoft also provides Azure IoT Edge [42] service to scale out inference learning by packaging the logic into standard containers, deploying these containers to any of the edge devices and monitoring it all from the cloud. Different from Greengrass, the applications like inference learning in Azure IoT Edge need to be developed in one of the supported programming languages.

8. Conclusions

Stream analytics aims to analyze and process high volumes of streaming data continuously. In this paper, we study how to best leverage edge and cloud resources to achieve better accuracy and latency for RNN-based stream analytics and better adapt concept drift in stream data. We propose three flexible deployments for the hybrid stream analytics framework in order to achieve the proper trade-off between latency and accuracy for

Table 7

Related works that support different of capabilities. ○No ●Yes ◐Some approaches support it.

Approaches	Inference on edge	Periodic model training on cloud	Model update from cloud to edge	Multi-model analytics for adaptability	Flexible edge-cloud deployment
Machine learning based IoT stream data analytics [1,35,36,36–39]	◐	○	○	●	○
Edge-cloud integrated framework [9,11–17]	◐	●	◐	○	○
System or toolkit for deep learning based inference [5,26,40–42]	●	○	○	○	●
Edge-cloud integrated framework	●	●	●	●	●

stream analytics. We also propose an adaptive and dynamic hybrid learning model with two weight combination algorithms for solving the concept drift during stream analytics. The evaluation with real-world stream datasets shows the proposed edge-cloud deployment can archive similar latency performance as edge-centric deployment without worrying about capacity limitations, and our dynamic weighting algorithm performs the best among other hybrid learning model approaches for all three concept drift scenarios in terms of accuracy.

For future work, we will mainly focus on the following three aspects of the hybrid stream analytics framework. First, the Spark-based speed training module can be extended to multiple edge devices as a distributed master-worker computing. Second, we will study more variants of the proposed dynamic weighting algorithm, like stacking the most recent n speed layer models or stacking speed layer models continuously. Last, we will try more advanced RNN deep learning models like attention models with our framework.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is public and its information is included in the paper text.

Acknowledgments

This work is supported by the National Science Foundation (NSF), United States Grant No. OAC-1942714 and U.S. Army, United States Grant No. W911NF2120076.

References

- [1] A. Pandya, O. Odunsi, C. Liu, A. Cuzzocrea, J. Wang, Adaptive and efficient streaming time series forecasting with lambda architecture and spark, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 5182–5190.
- [2] E. Uchiteleva, S. Primak, M. Luccini, A. Refaey, A. Shami, The TrILS approach for drift-aware time-series prediction in IIoT environment, IEEE Trans. Ind. Inf. (2021).
- [3] S.B. Qaisar, M. Usman, Fog networking for machine health prognosis: A deep learning perspective, in: International Conference on Computational Science and Its Applications, Springer, 2017, pp. 212–219.
- [4] D. Li, T. Salonidis, N.V. Desai, M.C. Chuah, Deepcham: Collaborative edge-mediated adaptive deep learning for mobile object recognition, in: 2016 IEEE/ACM Symposium on Edge Computing (SEC), IEEE, 2016, pp. 64–76.
- [5] K. Purandare, An Introduction to Deepstream SDK, Nvidia, 2018, Available at GStreamer Conference: <https://on-demand.gputechconf.com/gtc/2018/presentation/s81047-introduction-to-deep-stream-sdk.pdf>.
- [6] I. Žliobaitė, M. Pechenizkiy, J. Gama, An overview of concept drift applications, in: Big Data Analysis: New Algorithms for a New Society, Springer, 2016, pp. 91–114.
- [7] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 1–37.
- [8] Amazon Web Services, Inc, What is cloud computing? 2021, <https://aws.amazon.com/what-is-cloud-computing/>.
- [9] A. Luckow, K. Rattan, S. Jha, Pilot-edge: Distributed resource management along the edge-to-cloud continuum, in: 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), IEEE, 2021, pp. 874–878.
- [10] M. Mohammadi, A. Al-Fuqaha, S. Sorour, M. Guizani, Deep learning for IoT big data and streaming analytics: A survey, IEEE Commun. Surv. Tutor. 20 (4) (2018) 2923–2960.
- [11] S.A. Osia, A.S. Shamsabadi, A. Taheri, H.R. Rabiee, H. Haddadi, Private and scalable personal data analytics using hybrid edge-to-cloud deep learning, Computer 51 (5) (2018) 42–49.
- [12] N. Abdulla, M. Demirci, S. Özdemir, Adaptive learning on fog-cloud collaborative architecture for stream data processing, in: 2021 International Symposium on Networks, Computers and Communications (ISNCC), 2021, pp. 1–6, <http://dx.doi.org/10.1109/ISNCC52172.2021.9615824>.
- [13] D. Balouek-Thomert, E.G. Renart, A.R. Zamani, A. Simonet, M. Parashar, Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows, Int. J. High Perform. Comput. Appl. 33 (6) (2019) 1159–1174.
- [14] Z. Nezami, K. Zamanifar, K. Djemame, E. Pournaras, Decentralized edge-to-cloud load balancing: Service placement for the Internet of Things, IEEE Access 9 (2021) 64983–65000.
- [15] R.K. Naha, S. Garg, A. Chan, S.K. Battula, Deadline-based dynamic resource allocation and provisioning algorithms in fog-cloud environment, Future Gener. Comput. Syst. 104 (2020) 131–141.
- [16] M. AbdelBaky, M. Zou, A.R. Zamani, E. Renart, J. Diaz-Montes, M. Parashar, Computing in the continuum: Combining pervasive devices and services to support data-driven applications, in: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), IEEE, 2017, pp. 1815–1824.
- [17] I.M. Murwantara, P. Yugopusito, An adaptive IoT architecture using combination of concept-drift and dynamic software product line engineering, TELKOMNIKA 19 (4) (2021) 1226–1233.
- [18] W. Zhang, J. Wang, A hybrid learning framework for imbalanced stream classification, in: 2017 IEEE International Congress on Big Data (BigData Congress), IEEE, 2017, pp. 480–487.
- [19] Hybrid stream analytics on edge-cloud, 2022, <https://github.com/big-data-lab-umbc/Hybrid-Streaming-Analytics-on-Edge-Cloud>. Accessed: 2022-05-01.
- [20] M. Hermans, B. Schrauwen, Training and analysing deep recurrent neural networks, Adv. Neural Inf. Process. Syst. 26 (2013).
- [21] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: NIPS 2014 Workshop on Deep Learning, December 2014, 2014.
- [22] D. Tao, Y. Wen, R. Hong, Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition, IEEE Internet Things J. 3 (6) (2016) 1124–1134.
- [23] H. Zhang, Z. Xiao, J. Wang, F. Li, E. Szczerbicki, A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention, IEEE Internet Things J. 7 (2) (2019) 1072–1080.
- [24] B. Krawczyk, A. Cano, Online ensemble learning with abstaining classifiers for drifting and noisy data streams, Appl. Soft Comput. 68 (2018) 677–692.
- [25] Andrew Banks and Rahul Gupta, MQTT Version 3.1.1, OASIS Standard, 2014, <http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/mqtt-v3.1.1.html>.
- [26] Amazon Web Services, Inc, AWS IoT greengrass, 2022, <https://aws.amazon.com/greengrass/>.

- [27] Apache spark project, 2021, <http://spark.apache.org>. Accessed: 2021-5-28.
- [28] R. Polikar, Ensemble based systems in decision making, *IEEE Circuits Syst. Mag.* 6 (3) (2006) 21–45, <http://dx.doi.org/10.1109/MCAS.2006.1688199>.
- [29] L. Rokach, Ensemble-based classifiers, *Artif. Intell. Rev.* 33 (1) (2010) 1–39.
- [30] D. Kraft, et al., A software package for sequential quadratic programming, 1988.
- [31] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.* 7 (3) (2014) 1247–1250.
- [32] ENGIE's first open data windfarm: La Haute Borne, 2017, <https://opendata-renewables.engie.com/>.
- [33] R. Mushtaq, Augmented dickey fuller test, 2011, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.1911068>.
- [34] T. Dahiru, P-value, a true test of statistical significance? A cautionary note, *Ann. Ibadan Postgrad. Med.* 6 (1) (2008) 21–26.
- [35] Z. Shao, S. Yuan, Y. Wang, Adaptive online learning for IoT botnet detection, *Inform. Sci.* 574 (2021) 84–95.
- [36] D. Puschmann, P. Barnaghi, R. Tafazolli, Adaptive clustering for dynamic IoT data streams, *IEEE Internet Things J.* 4 (1) (2016) 64–74.
- [37] L. Yang, A. Shami, A lightweight concept drift detection and adaptation framework for IoT data streams, *IEEE Internet Things Mag.* 4 (2) (2021) 96–101.
- [38] L. Yang, D.M. Manias, A. Shami, PWPAE: An ensemble framework for concept drift adaptation in IoT data streams, 2021, arXiv preprint [arXiv: 2109.05013](https://arxiv.org/abs/2109.05013).
- [39] H. Mehmood, P. Kostakos, M. Cortes, T. Anagnostopoulos, S. Pirttikangas, E. Gilman, Concept drift adaptation techniques in distributed environment for real-world data streams, *Smart Cities* 4 (1) (2021) 349–371.
- [40] Google LLC., Google coral, 2020, <https://coral.ai/>.
- [41] Google LLC., Google coral edge TPU, 2020, <https://cloud.google.com/edge-tpu>.
- [42] Microsoft Azure, Azure IoT edge, 2022, <https://azure.microsoft.com/en-us/services/iot-edge/>.
- [43] G. Rocher, S. Lavirotte, J.-Y. Tigli, G. Cotte, F. Dechavanne, An IOHMM-based framework to investigate drift in effectiveness of IoT-based systems, *Sensors* 21 (2) (2021) 527.
- [44] W. Zhang, M. Zhang, J. Zhang, Z. Liu, Z. Chen, J. Wang, E. Raff, E. Messina, Flexible and adaptive fairness-aware learning in non-stationary data streams, in: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2020, pp. 399–406.
- [45] NVIDIA Corporation, DeepStream SDK, 2021, <https://developer.nvidia.com/deepstream-sdk>.



Xin Wang is a Ph.D. candidate at the Big Data Analytics Lab in the Department of Information Systems, University of Maryland, Baltimore County. She also works as a Research Assistant in the Center for Real-time Distributed Sensing and Autonomy (CARDS). Her research interests include distributed computing (systems), blockchains, big data analytics, federated learning, cloud computing and reproducibility.



Azim Khan is a Ph.D. student at the Big Data Analytics Lab in the Department of Information Systems, University of Maryland, Baltimore County. He also works as a Research Assistant in the Center for Real-time Distributed Sensing and Autonomy (CARDS) at UMBC. His research interests include data science, IoT, cloud computing and deep learning.



Jianwu Wang is an Associate Professor of Data Science and the Director of the Big Data Analytics Lab at University of Maryland, Baltimore County. His research interests include big data analytics, distributed computing and service oriented computing. He has published 110+ papers with more than 2000 citations (h-index: 23).



Aryya Gangopadhyay is a Professor in the Department of Information Systems and the Director of the Center for Real-time Distributed Sensing and Autonomy (cards.umbc.edu) at University of Maryland, Baltimore County. His research interests include Machine Learning and cybersecurity. His research has been funded by NSF, ARL, IBM, and the US Department of Education.



Carl Busart received the B.S. and M.S. degrees from Johns Hopkins University, an MBA from the University of Maryland, College Park, and the D.Eng. degree from George Washington University. He is a branch chief at the U.S. Army Research Laboratory and a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). His research interests include artificial intelligence/machine learning (AI/ML) and secure design.



Jade Freeman is the Chief of Battlefield Information Systems Branch at DEVCOM Army Research Laboratory. Her research interests include information systems for decision support, human-information interactions, and information theory.