Scheduling and Optimization of Big Data Stream Processing

Miss. A. K. Nachankar, Dr. R.V. Dharaskar, Dr. V.M. Thakare, Ashwini Nachankar

SGBAU, Amravati, Maharashtra, India.

ABSTRACT- Big data stram processing is rapidly growing day by day due to the immediate demand of many application. This growth compels industries to leverage scheduling order to optimally allocate resource to big data streams which requires data-driven big data analysis.Optimal scheduling of big data stream process should guarantee the QoS requirnment of computing task. This paper proposes a framework "online energy-efficient scheduler". The proposed frameworkmaximizes the QoS of Big-Data streaming applications under energy and resources constraints.The scheduler uses online adaptive reinforcement learning techniques and requires no offline information. Moreover, the scheduler is able to detect concept drifts and to smoothly adapt the scheduling strategy.

Index Terms— *Big data, scheduling, optimization, stream processing,*

I) INTRODUCTION

Scheduling and optimization plays an important role in big data stream processing. Different scheduling and optimization techniques have been designed such as Ensemble learning algorithm,update the aggregation rule in order to adapt to the underlying data dynamics. Rigorously determine an upper bound for the worst-case mis-classification probability of algorithm, which tends asymptotically to 0 if the misclassificationprobability of the best static aggregation rule is 0[1].Staged multi-armed bandits scheduler, is able to learn online which processing method to assign to each stream and how to allocate its resources over time in order to maximize the performance on the fly, at run-time, without having access to any offline information[2]. Index Queueing theory approach, is used for modeling of the streams as a collection of sequential and parallel tasks .Then, with the queueing theory, an optimization problem is defined to minimize the total number of resources required to serve the big data streamswhile guaranteeing the QoS requirements of their tasks [3].Proactive architecture is used to exploits historical data using machine learning for prediction in conjunction with complex event processing.. Analysis of such models has been carried out to calculate the performance evaluation [4]. Big data streaming applications are now widely used in several domains such as social media analysis, financial analysis, video annotation, surveillance and medical services [5].These applications are characterized with stringent delay constraints, increasing parallel computation requirement and a highly variable stochastic input data stream which have direct impact on the application complexity and the final Quality of Service (QoS).

This paper, introduce different frameworks which uses the concept scheduling and optimization of big data. There are some drawbacks in previous methodology. To overcome such problems this paper proposed new framework i.e**"Online Energy-efficient Scheduler".**The proposed framework maximizes the QoS of Big-Data streaming applications under energy and resources constraints. The scheduler uses online adaptive reinforcement learning techniques and requires no offline information.

II) BACKGROUND

Many studies on scheduling and optimization have been done to develop the processing of big data stream in recent past years. Previous methodologies are as follows. Load-shedding technique is used to determine how much data to discard given the desired QoS requirnment and the available resources [1].G/M/c Model used to evaluate the performance of the cloud threads in the data process The analysis is then extended to systems with general arrival and service time distributions. With the objective of minimizing the number of resources required to serve the stream, an optimization problem is defined. Finally, a heuristic algorithm is proposed to mitigate the complexity of the optimization problem.[2]. Selection of regression algorithmfor dynamic IoT data and evaluated it using a real-world use case with an accuracy of over 96%. It can perform accurate predictions in near real-time due to reduced complexity and can work along CEP in our architecture [3]. Finding optimal training window size scheme The choice of the optimum training window size for ML models is an open research issue. In general, the accuracy of prediction model increases as the size of training data increases which reflects to have large historical data for training prediction models so that it covers all possible patterns spanning time series.[4]. Size of prediction horizon is an adaptive size for prediction window or more commonly known as prediction horizon in order to ensure a certain level of accuracy. The intuition behind it is to increase the size of prediction window if the accuracy of model is high and decrease it if the performance of the prediction model decreases [5].

This paper is organized as follows:

Section I containsIntroduction of this paper. In **Section II** discussed Background. **Section III** introduced previous work done. **Section IV** explains existing methodologies. In **SectionV** discussed existing framework and analysed it. **SectionVI** presents the overview of Online Energy Efficient scheduler framework. Its outcome possible results are analysed in **Section VII**. **Section VIII** concludes this paper. Finally **Section IX** presents future scope.

III) PREVIOUS WORK DONE

Karim Kanoun et al. (2017) [1] have proposed load-shedding scheme tolearn online which processing method, to assign to each stream and how to allocate its resources over time in order to maximize the performance on the fly, at run-time, without having access to any offline information.

Shahin Vakilinia et al. (2016) [2] has proposes an effective G/M/c model to minimize the stream processing resources in terms of threads with constraints over the task waiting time of the application tasks.

Adnan Akbar et al.(2016) [3] have proposed selection of regression algorithm for for time series regression (prediction) ranging from statistical to pure ML domain. Traditionally, statistical methods like auto regressive moving average and autoregressive integrated moving average were used for time series regression.

Luca Canzian et al.(2016)[4] has proposed Finding optimal training window size scheme for accuracy of prediction model increasesas the size of training data increases which reflects to have large historical data for training prediction models so that it covers all possible patterns spanning time series.

Salvador Garcia et al.(2015) [5] has proposed the Size of the prediction horizon techniquein order to ensure a certain level of accuracy.

IV) EXISTING METHODOLOGIES

A. Load Shedding Scheme:

Load shedding scheme where designed to determine when, where, what, and how much data to discard given the desired QoS requirements and the available resources. The impact of load shedding is known a-priori and the load shedder was decoupled from the scheduler assuming that an external scheduler will handle the assignment of freed resources. A load shedding scheme ensures that dropped load has minimal impact on the benefits of mining and dynamically learns a Markov model to predict feature values of unseen data. Instead of deciding on what fraction of the data to process, as in load shedding, the second set of approaches determine how the available data should be processed given the underlying resource allocation [1]. In these works, individual tasks operate at a differentperformance level given the resources allocated to them. They assume a fixed model complexity for eachclassifier and the variation of the output quality is known a-priori. The problem was formulated as a network optimization problem and solved with sequential quadratic programming.

B.G/M/cqueue model:

The G/M/cqueue model can be appliedto investigate the delay of the stream tasks in the system. Assuming general distribution for the task arrival rate, average waiting time of the tasks in the system is given by,

$$W_{ij} = \frac{K\zeta^{c_{ij}}}{c_{ij}\zeta\,(1-\zeta)^2}$$

The G/M/c model is applied to evaluate the performance of the cloud threads in the data process. The analysis is then extended to systems with general arrival and service time distributions. The objective is minimizing the number of resources required to serve the stream, an optimization problem is defined. Finally, a heuristic algorithm is proposed to mitigate the complexity of the optimization problem [2]. In the most complicated situation, when arrival rate and service time follow general distributions, the analysis has to be done with general assumption on arrival rate and service time. However due to the computational complexity, it isl not be able to find the closed form waiting delay for this model.

C. Selection of Regression Algorithm:

Selection of regression algorithm, There are several algorithms available for time series regressionranging from statistical to pure ML domain. Traditionally, statistical methods like auto regressive moving average and auto regressive integrated moving average were used for time series regression. However, recently the trend is shifted toward more sophisticated ML models such as different variants of support vector regression (SVR) and artificial neural networks because of their robustness and ability to provide more accurate solutions [3]. Approach is implemented using SVR due to its ability to model nonlinear data using kernel functions. SVR is an extension of SVM which is widely used for regression analysis. The main idea is the same as in SVM, it maps the training data into higher feature space using kernel functions and find the optimum function which fits the training data using hyper-plane in higher dimension. Methods based on SVR often provide more accurate models as their counterpart regression algorithms at the expense of additional complexity. However, algorithm is propose to use a small training window, the added complexity is almost negligible for such small datasets.

D. Finding Optimal Training Window Size:

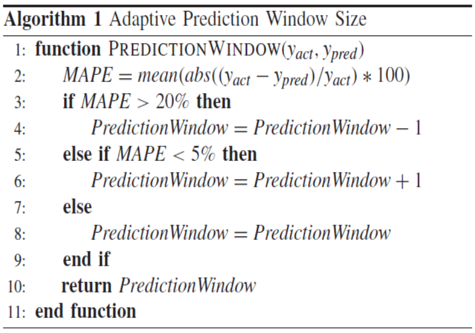
The choice of the optimum training window size for ML models is an open research issue. In general, the accuracy of prediction model increases as the size of training data increases which reflects to have large historical data for training prediction models so that it covers all possible patterns spanning time series[4]. In contrast to this approach, have proposed to use the moving window for training the ML model in which most recent data is fed to the models. The size of the optimum window is a challenging task with no generic solution. Least squares spectral analysis (LSSA) or more commonly known as Lomb–Scargle can be used to find the highest periodic component in a time series data. Lomb first proposed the method while studying variable stars in astronomy and is definedby the following equations:

$$P_X(f) = \frac{1}{2\sigma^2}\left\{\frac{\left[\sum_{n=1}^{N}(x(t_n)-\bar{x})\cos(2\pi f(t_n-\tau))\right]^2}{\sum_{n=1}^{N}\cos^2(2\pi f(t_n-\tau))} + \frac{\left[\sum_{n=1}^{N}(x(t_n)-\bar{x})\sin(2\pi f(t_n-\tau))\right]^2}{\sum_{n=1}^{N}\sin^2(2\pi f(t_n-\tau))}\right\}$$

A large window size can have more accurate results but it increases the complexity of the model making it unsuitable for realtime applications whereas a small window size can result into an increased error and hence effecting the reliability of the system.

E. Adaptive Prediction Window:

To have an adaptive size for prediction window or more commonly known as prediction horizon in order to ensure a certain level of accuracy. The intuition behind it is to increase the size of prediction window if the accuracy of model is high and decrease it if the performance of the prediction model decreases [5].



Algorithm 1 Adaptive Prediction Window Size

```
function PREDICTIONWINDOW($y_{act}, y_{pred}$)
    $MAPE = mean(abs((y_{act} - y_{pred})/y_{act}) * 100)$
    if $MAPE > 20\%$ then
        $PredictionWindow = PredictionWindow - 1$
    else if $MAPE < 5\%$ then
        $PredictionWindow = PredictionWindow + 1$
    else
        $PredictionWindow = PredictionWindow$
    end if
    return $PredictionWindow$
end function
```

The performance of the model is evaluated by comparing the predicted data with actual data when it arrives.

V) ANALYSIS AND DISCUSSION

The load shedding schemeensures that dropped load has minimal impact on the benefits of mining and dynamically learns a Markov model to predict feature values of unseen data [1]. It maximizes the performance o at run-time, without having access to any offline information.

G/M/c model is applied to evaluate the performance of the cloud threads in the data process [2]. The analysis is then extended to systems with general arrival and service time distributions.G/M/cqueue can also be appliedto investigate the delay of the stream tasks in the system .

Selection of regression algorithm, perform accurate prediction in near real time due to reduced complexity and can work along CEP in architecture [3]. Several algorithms available for time series regressionranging from statistical to pure ML domain

Finding optimal training window size scheme shows theaccuracy of prediction model increases as the size of training data increases [4] which reflects to have large historical data for training prediction models so that it covers all possible patterns spanning time series.

Adaptive prediction window scheme shows that to increase the size of prediction window if the accuracy of model is high and decrease it if the performance of the prediction model decreases [5].to have an adaptive size for prediction window or more commonly known as prediction horizon in order to ensure a certain level of accuracy.

Sheduling and Optimization scheme	Advantages	Disadvantages
Load shedding scheme	Complexity grows linearly in the size of the action space. It maximizes the performance o at run-time	All action are selected in sequencehence, no feedback is available between the actions.
G/M/c Model	provides the feature of parallelism in an easy way for the user to pre-configure it in the cluster.	devote most of their execution time in disk I/O for processing a larger volume of data.
Selection of regression algorithm	It uses for time series regression. Provide more accurate solution.	Data collected from different sources can vary in quality and format.
Finding optimal training window size	It finds optimal window size and validate result using real world dat.a	The only drawback is that it increases complexity of mode.l
Size of prediction horizon	This method improves the adaptive size of training window	The only drawback is that the data could be incomplete.

TABLE 1: Comparisons between different scheduling and optimization scheme.

VI) PROPOSED METHODOLOGY

Online Energy Efficient Scheduler

An online energy efficient scheduler that adopts reinforcement learning techniques to learn the environment dynamics in order to maximizethe Quality of Service QoS of dynamic Big-Data streaming applications given energy constraints. The key contributions of this work such as,model the scheduling problem as a Stochastic Shortest Path problem (SSP) and propose a reinforcement learning algorithm to learn the environment dynamics to solve this problem even in the presence of concept drift. The exploration phase of reinforcement learning algorithm guarantees fast convergence to the targeted QoS. Scheduler learns online the dynamics of the environment and provide full control of the quality, throughput and resource constraints without any offline information.

Acronyms:

Learned Sheduler Policy – LSP

Greedy Selection – GS

Resource Check – RC

Quality Check – QC

Update State Transition – UST

Generate New Sheduling Policy – GNSP

Online Energy Efficient Scheduler Algorithm:

STEP 1: State Identificatiom

STEP 2: IF known state is observed

Set the input to LSP

ELSE

Do not set LSP

STEP 3: IF unknown state is observed

Set the input to GS

ELSE

Do not set to GS

STEP 4: RC and QC

A. IF concept drift observed

Then policy update required i.e UST

ELSE

B. Does not required UST

STEP 5: Generate new scheduling policy

In this way the above algorithm improves the throughput and quality using online energy efficient schedule

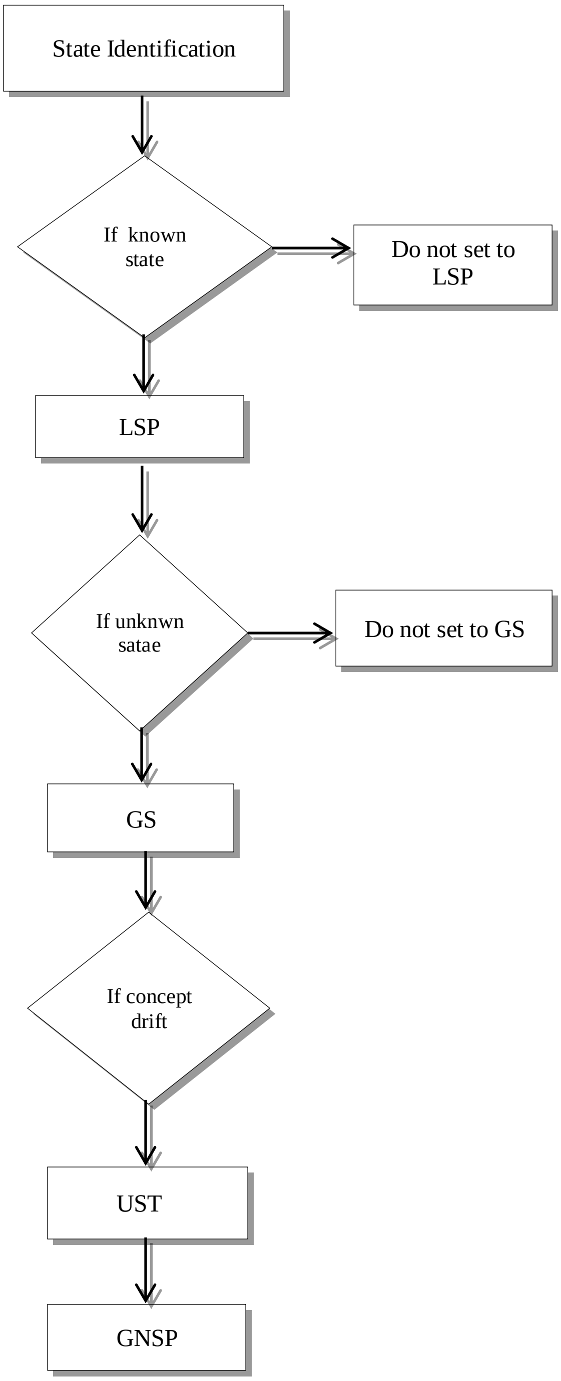


Fig 1: Online Energy Efficient Scheduler Algorithm

VII) OUTCOME POSSIBLE RESULT

The proposed framework “Online Energy Efficient Scheduler”is able to learn the scheduling policy and to adapt it such that it maximizes the targeted QoS given energy constraint as the Big-Data characteristics are dynamically changing. The proposed framework successfully improves the throughput and quality.

VIII) CONCLUSION

This paper focused on various scheduling and optimization framework and introduced problemspresent in previous methodologies. To overcome these

drawbacks this paper proposed a novel “Online Energy Efficient Scheduler” framework that adopts reinforcement learning techniques to learn theenvironment dynamics in order to maximize the QoS of dynamic Big-Data streaming applications given resource usage constraints. It has timecomplexity of O(1) which is always in best case soproposed framework saves time and it is very beneficial. It also saves searching time so throughput is increased.

IX) FUTURE SCOPE

From observation, the scope is planned to be studied in future work that include the new improved scheduler whichmaximizes the targeted QoSgiven resources constraints as the Big-Data characteristics are dynamically changing.

REFERENCES

[1] Karim Kanoun, Cem Tekin, David Atienza, “Big-Data Streaming Application Scheduling Based on Staged Multi-Armed Bandits”, IEEE TRANSACTIONS ON COMPUTERS,Vol .65, No. 12, DECEMBER 2017.

[2] Shahin Vakilinia, Xinyao Zhang, Dongyu Qiu, “Analysis and Optimization of Big-Data Stream Processing”,IEEE TRANSACTION ON INFORMATION PROCESSING OVER NETWORKS,NOVEMBER 2016.

[3] Adnan Akbar, Abdullah Khan, Francois Carrez ,“ Predictive Analytics for Complex IoT Data Streams”, IEEE TRANSACTIONS ON THINGS JOURNAL,Vol. 4, No. 5,OCTOBER 2016.