AUTO SCALABLE BIG DATA AS-A-SERVICE IN THE CLOUD: A LITERATURE REVIEW

Naseer Ahmad Shinwari[1], Dr. Neeta Sharma[2]

M.Tech Pursuing[1], **Noida International University, Greater Noida, (U.P.)**

Assistant Professor[2], **Dept of CSE, Noida International University, Greater Noida, (U.P.)**

Abstract: cloud computing is the widely adopted technology which comes into mind when talking about elasticity and limitless scalability. Most of the heaving tasks and complex calculations which require big machines are directed to the cloud services for cost minimization, and on demand resource provisioning. But if we are leveraging the Big data as-a-service services on the cloud, which of course requires lots of resources and quite big data stores in order to properly manage and process big data we would defiantly mark the scalability as the red line since velocity (data generation at an alarming rate) is the one of the main Vs of big data. This paper reviews many scalability models, frameworks and algorithms for cloud scalability, unpredictable load balancing and resource prediction for efficient scalable system which serves best in case of auto scalable Big data as-a-service services over the cloud.

Keywords: auto-scalability, unpredictable load balancing, big data as-a-service, cloud scalability.

I. INTRODUCTION:

Every step that we take or action we perform now a days is been recorded in some digital format known as data which is been generated at an alarming rate and in various forms and formats either structured, semi structured or unstructured, that eventually become a massive store of data called big data. Big data is managed, processed and analyzed for many purposes in many fields, as in weather forecasting we analyze the big historical data which can be life saving if done accurately and on time, in the other hand most online retailers use huge big data sets of online surfers for building recommendation systems based on their interest. These days everyone has big data in hand and want to properly manage it and draw conclusion from it for timely decision making. Hadoop is the tool which is widely used for big data management and many other tools accompanying it are used for the analysis of this big data which require big machines to operate on these huge data sets and draw insights from it. That's why people turned to cloud platform for their big data analysis and management where services are provided on pay-per-use bases over different as-a-service model as Hadoop as-a-service, Big data as-a-service and many others. We have concerns in the scalability portion of the services provided by the cloud infrastructure of handling big data and providing Big data as-a-service services, to address the issue the subsequent sections of the paper presents a literature review on various scalability models, frameworks and algorithms, chasing for best solution over the cloud for an automatic scalable big data as-a-service.

II. Literature review:

Che-Lun Hung et. al. proposed a cloud auto-scaling architecture and an auto-scaling algorithm in the paper "Auto-scaling model for cloud computing system (2012)" [1]. They presented two scaling scenarios to address the automatic scalability of web applications and distributed computing jobs in a virtual cluster on the virtualized cloud computing environment. Their proposed auto-scaling architecture contains three main components: front-end load balancer, virtual cluster monitor system and auto-provisioning system with an auto-scaling algorithms as demonstrated in fig 1.

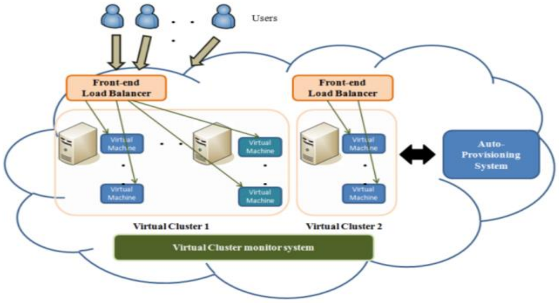

Figure 1 Architecture of the Auto-scaling in the cloud

In the paper "Modeling the auto-scaling operation in cloud with time series data (2015)" [2] Mehran N. A. H. Khan et. al. conducted an overview of 'the auto scaling operations' on various commercial cloud service providers namely, AWS auto-scaling architecture, Open stack auto-scaling architecture, Azure auto-scaling application block and Google cloud auto-scaling architecture with the purpose of identifying common features and entities from these operations. Their observation of the study states that these operations mainly adopt the rule-based approach by using a set of metrics collected at the infrastructure and platform level and thereby they presented a model that consist of parameters whose values calibrate the auto-scaling workflow in time series data collected for these metrics. Their model explores time series data and workload prediction technique to capture the dynamics of the auto-scaling workflow. They used Google trace data as sample data to calibrate the model and presented the model in another analytical model of Petri Net to estimate the end-to-end delay of the auto-scaling workflow given what-if analysis as shown in fig 2.

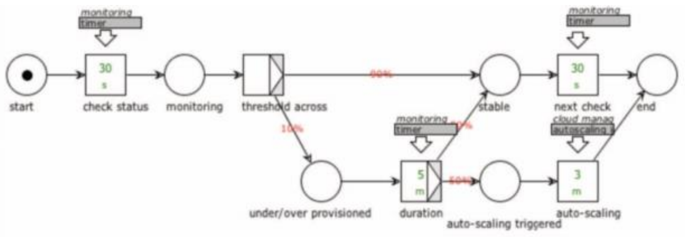

Figure 2 Petri Net model of the auto-scaling workflow

Yang Xia et. al. in the paper "A scalable framework for cloud powered workflow execution (2013)" [3] proposed a scalable framework enabling cloud powered execution for scientific workflows. The framework supports automatic cluster provisioning and scaling. Their work further discusses strategies to parallelize workflow execution to improve execution speed and they also proposed and algorithm to translate workflows so that operations within a flow are also parallelized. Fig 3 shows an overview of the system architecture depicting various modules of the framework amongst which the cluster provisioning module plays a vital role in taking care of the cluster provisioning and preparing the execution environment for each workflow. It supports provisioning for of clusters in both private and public clouds. When the remaining resources in the private cloud are insufficient for executing new workflows, it will automatically scale out and provision clusters in public cloud.

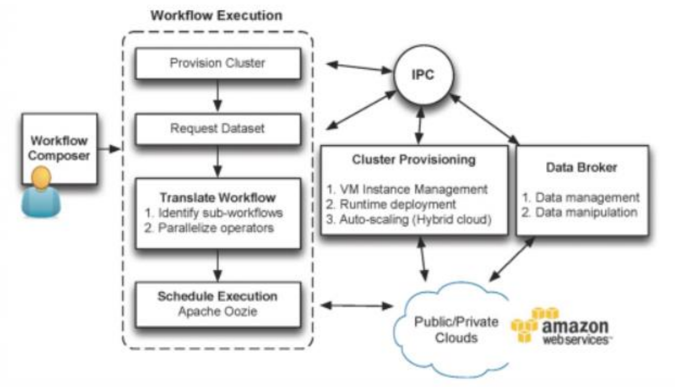

Figure 3 An overview of the system architecture

In the paper "Cloud resource scaling for big data streaming applications using a layered multi-dimensional hidden markov model (2017)" [4] Olubisi Runsewe et. al. proposed a layered multi-dimensional hidden markov Model (LMD-HMM) for facilitating the management of resource auto-scaling for big data streaming applications in the cloud. Their experimental evaluation shows that layered multi-dimensional hidden markov model performs best with an accuracy of 98%, outperforming the single-layer hidden markov model. Their architecture includes a data ingestion layer, a processing layer and a storage layer. To support resource scaling, present within the architecture is the resource controller that is made up of a resource monitor, a predictor and a resource allocator as shown in fig 04. For the robustness of the resource prediction model in order to represent the typical variations in the underlying streaming jobs they made use of two layered multi-dimensional HMM as shown in fig 5. The lower-layer allows for the abstraction of individual streaming jobs' observations at a particular times and also ensures independence from others while the upper-layer is used to predict the resource usage pattern for the group streaming applications running on a cluster.

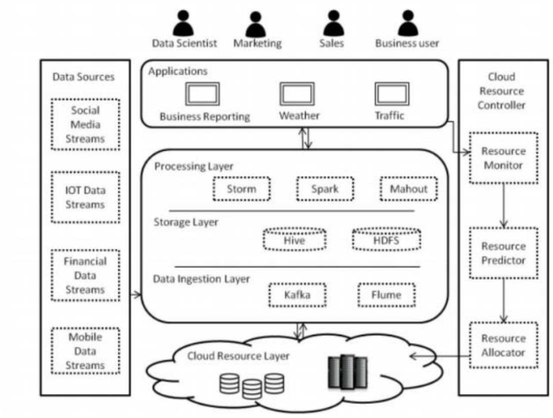

Figure 4 LMD-HMM proposed architecture

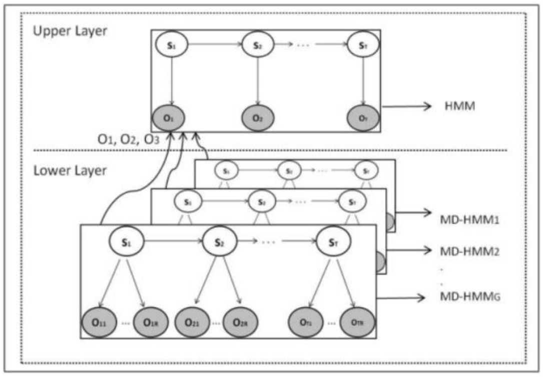

Figure 5 A layered multi-dimensional HMM

Under the title “Autoscaling for hadoop clusters” [5] Anshul Gandhi et. al. presented the design and implementation of a model-driven autoscaling solution for Hadoop clusters for the concern of unforeseen events such as node failures and resource contention. For dynamically adding or removing nodes from Hadoop cluster, they created a customized VM image preloaded with Hadoop. A new slave can dynamically be added by booting a new VM using the customized image, and then starting the TaskTracker and DataNode services on it. And for removing a slave node, they update the exclude list on the Master and dynamically refresh the node configuration. The Master node migrates that data from the node before excluding it if any. Fig 06 demonstrates their ability to dynamically scale up capacity as needed to meet execution time SLAs.

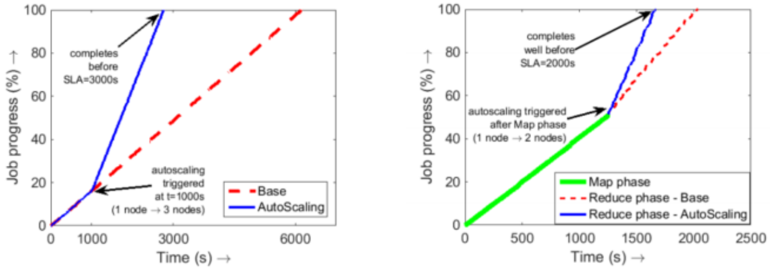

Figure 6 Dynamically scaling up to meet execution time SLA

Zhenlong Li’s et. al. paper “Automatic scaling Hadoop in the cloud for efficient process of big geospatial data (2016)” [6] proposes an auto-scaling framework to automatically scale cloud computing resources for Hadoop cluster based on the dynamic processing workload, with the aim of improving the efficiency and performance of big geospatial data processing. Their framework contains following components (1) Cloud computing platform, (2) CoveringHDFS enabled Hadoop cluster, (3) Auto-scaler, and (4) Cluster monitor, as illustrated in fig 06.

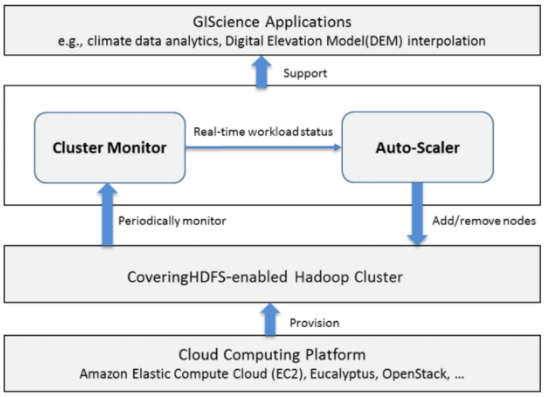
GIScience Applications
e.g., climate data analytics, Digital Elevation Model(DEM) interpolation
Support
Cluster Monitor
Real-time workload status
Auto-Scaler
Periodically monitor
Add/remove nodes
CoveringHDFS-enabled Hadoop Cluster
Provision
Cloud Computing Platform
Amazon Elastic Compute Cloud (EC2), Eucalyptus, OpenStack, ...

Figure 7 Auto-scaling Framework

Among which the most important components of the framework are the predictive scaling-up algorithm which accurately determines the number of slaves to be added by monitoring the white-box-based metrics considering the proposed scale-up time; and a CoveringHDFS mechanism that scales down the cluster promptly to avoid unnecessary resource consumption. According to their evaluation the proposed framework is able to (1) significantly reduce the computing resource utilization for about 80% while delivering similar performance as a full-powered cluster by dynamically adjusting the cluster size based on changing workloads; and (2) effectively handle the processing workload by increasing the computing resources to ensure that the processing is finished within an acceptable time.

To overcome the general lack of effective techniques for workload forecasting and optimal resources allocation, Nilabja Roy et. al. made three contributions in the paper "Effective autoscaling in the cloud using predictive models for workload forecasting (2011)" [7]. Their work mainly describes a look-ahead resource allocation algorithm based on model predictive techniques which allocates or deallocates machines to the application based upon optimizing the utility of the application over a limited prediction horizon.

Under "Autoscaling prediction models for cloud resource provisioning (2016)" [8] Yazhou Hu et. al. proposed a prediction framework for virtual machines provisioning which includes three main modules: monitor, filter and predictor, with the aim of predicting upcoming workload in order to overcome virtual machine provision latency. Moreover for processing raw data they proposed the Kalman filter method and as based predictor they presented five different prediction models, namely: moving average (MA), auto regression (AR), auto regression integrated moving average (ARIMA), neural networks (NN) and support vector machine (SVM). In order to evaluate the performance of prediction framework, four evaluation metrics are proposed, including the prediction error, the time saving, under-prediction resource and over-prediction resource.
Virtual machines provisioning is based on analyzing the historical workload to prepare virtual machines ahead of time as demonstrated in fig 07.

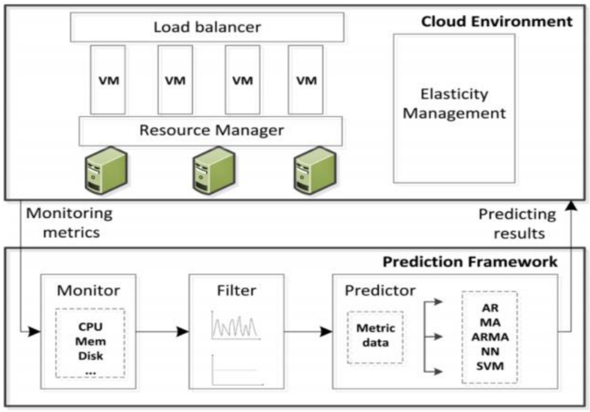

Figure 8 Overview of prediction framework

In "A proactive cloud scaling model based on fuzzy time series and SLA awareness (2017)" [9] paper, Dang Tran et. al. worked on a proactive autoscaling model for cloud computing which consists of two major component, namely: prediction and scaling decision. For the prediction module they used fuzzy approach to process multivariate monitoring resources, genetic algorithm, back-propagation, and neural network with the purpose of efficient and precise forecasting; and for the decision module they proposed a formula to calculate SLA violations, then the SLA-aware data is sent back to system in order to integrate with predicted values to adapt their autoscaling model, as shown in fig 08.

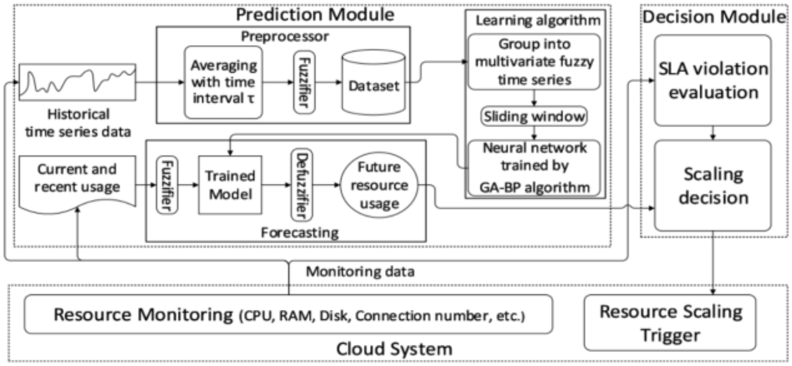

Figure 9 proposed proactive scaling model for clouds

III. Conclusion:

Keeping automatic scalability for Big data as-a-service in mind, the literature survey is carried out in this paper to understand available solutions and proposed models, frameworks and algorithms over the cloud. Our findings from the literature by looking at each model and proposed work are: Cloud is already elastic and provides scalability to a great level, but for the big data as-a-service the resource provisioning time and data stores charging are the things which may not be bearable in most cases. A better model should be presented for quick resource provisioning and resource prediction ahead of time, the computational resources (CPU & memory) and disk space billing methods should be separately and properly considered keeping the resource wastage and unmet demand in mind. Which we will do in our future work.

IV. References:

[1] Che-Lun Hung et. al. "Auto-scaling model for cloud computing system", international journal of hybrid information technology Vol. 5, No. 2, April, 2012.

[2] Mehran N. A. H. Khan et. al. "Modeling the auto-scaling operation in cloud with time series data", 2015 IEEE 34th symposium on reliable distributed systems workshop, DOI 10.1109/SRDSW.2015.20.

[3] Yang Xia et. al. "A scalable framework for cloud powered workflow execution", Globcom 2013 workshop – cloud computing systems, networks and applications.

[4] Olubisi Runsewe et. al. "Cloud resource scaling for big data streaming applications using a layered multi-dimensional hidden markov model", 2017 17th IEEE/ACM international symposium on cluster, cloud and grid computing, DOI 10.1109/CCGRID.2017.147.

[5] Anshul Gandhi et. al. "Autoscaling for hadoop clusters".

[6] Zhenlong Li et. al. "Automatic scaling Hadoop in the cloud for efficient process of big geospatial data", (2016) International journal of Geo-Information, DOI 10.3390/ijgi5100173.

[7] Nilabja Roy et. al. "Effective autoscaling in the cloud using predictive models for workload forecasting", 2011 IEEE 4th international conference on cloud computing. DOI 10.1109/CLOUD.2011.42.

[8] Yazhou Hu et. al. "Autoscaling prediction models for cloud resource provisioning". 2016 2nd IEEE international conference on computer and communications.

[9] Dang Tran et. al. "A proactive cloud scaling model based on fuzzy time series and SLA awareness", international conference on computational sciences, ICCS 2017.