Supporting intra and inter cloud mining by using SWMaaS

A Manohara S Vani
KMM Institute Post Graduate Studies

***Abstract*- Computing clouds became the platform of alternative for the preparation and execution of scientific workflows. owing to the uncertainty and unpredictability of scientific exploration, scientific workflows can't be absolutely specific at the modeling stage. it's so of nice significance to be able to discover actual workflows from the execution history (event logs) so as to breed experimental results and to ascertain place of origin. However, most existing method mining techniques target discovering management flow-oriented business processes during a centralized surroundings, thus, they're largely irrelevant to discovering information flow-oriented, unstructured scientific workflows in distributed cloud environments. during this paper, we gift SWMaaS (Scientific progress Mining as a Service) to support each intra-cloud and inter-cloud scientific progress mining. The approach is enforced as a promenade plug-in and is evaluated on event logs derived from real-world scientific workflows. Through experimental results, we demonstrate the effectiveness and potency of our approach.**

***Index Terms*- Cloud computing, Scientific Workflow, inter-cloud scientific workflow mining, intra-cloud scientific workflow mining**

INTRODUCTION

Logical tests in fields, for example, bioinformatics, cosmology, rudimentary molecule material science, and life science require the putting away and handling of huge information. The appropriation of distributed computing to process these logical information has expanded as of late as cloud empowers information connections, design mining, information expectations, and information examination in a cost-proficient way. Cloud gives assets, for example, systems, stockpiling, applications, and servers can be dispensed from a mutual asset pool with negligible administration or association. By and large, distributed computing gives Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) in a compensation for every utilization demonstrate. Of these administrations, IaaS is the most as often as possible used as it furnishes clients with a versatile office to arrangement or discharge virtual machines (VMs) in the cloud framework. The versatile idea of cloud encourages changing of asset amounts and attributes to differ at runtime, in this manner powerfully scaling up when there is a more prominent requirement for extra assets and downsizing when the request is low. Because of its versatile asset provisioning component, the cloud is generally utilized as a part of a few regions including vast scale logical applications. With enhanced administration support and cloud blasting innovations, clients don't have to arrangement their assets for the most pessimistic scenario situations. With regards to distributed computing, an application's execution fetched implies the fiscal cost of leasing assets from the cloud specialist co-op. Tending to the issues of limiting the asset use with the best execution is an imperative research region in an IaaS cloud.

Logical applications comprise of thousands of errands requiring substantial calculation and information exchange. Work process errands additionally have certain conditions amid execution. Displaying of these logical applications is fundamental for compelling handling. The most broadly utilized portrayal display is as coordinated non-cyclic chart (DAG) in which the structure of the work process demonstrates the request of execution of errands. Our enhancement system manages logical work process as DAG amid advancement experimentations. Likewise, it is expected that the fiscal cost for work process execution in the cloud depends on the measure of assets utilized, that is, assignment execution cost connected to the aggregate number of CPU cycles for all undertakings. By

utilizing cloud framework, clients can likewise diminish the handling expense of work process applications with the guide of the cloud's compensation per-utilize display. In addition, an errand that offers a similar time interim with a past undertaking facilitated in a similar occasion won't not deliver additional cost, diminishing the general work process execution cost.

For between cloud work process mining, occasion logs are vertically parceled on various cloud stages (physical machines). For protection and security reasons and because of the innate conveyance, these occasion logs are normally not put away on a solitary server. Since there is no worldwide physical clock, it is even difficult to consolidate conveyed follows into one worldwide follow as per the nearby occasion log timestamps. To address this test, we initially get immediate priorities autonomously from the individual occasion sign on each cloud stage, and after that join them to get the general logical work process in light of between cloud message logs. Logical work process mining can be executed as an esteem included administration SWMaaS (Scientific Workflow Mining as a Service) gave by the cloud stage. We execute our approach (SWMaaS) in the instrument SWMC. The device is executed in Java and furthermore embodied as a module to the ProM system. The device includes two functionalities: intra-cloud and between cloud logical work process mining. To assess our approach, we perform broad analyses on occasion logs of true logical work processes from my Experiment1 . The trial comes about show the adequacy and productivity of our approach for both intra-cloud and between cloud logical work process mining. The exploratory assessment likewise demonstrates that our approach is more suitable for unstructured logical work process mining than the cutting edge process mining strategy. To whole up, the commitments of our work are recorded as takes after.

We introduce an intra-cloud logical work process mining approach in view of direct priority relations between occasions in the log. The approach ensures that given any logical work process W, a traceequivalent logical work process can be extricated from halfway occasion log of W.

- On the premise of intra-cloud mining, we additionally introduce a between cloud logical work process mining approach. This approach use coordinate priorities inside disseminated occasion and message logs to determine the general work process. To the best of our insight, we are the first to examine logical work process mining in circulated cloud conditions.
- We actualize our intra-cloud and between cloud logical work process mining calculations in a device SWMC which is additionally distributed as a ProM module.
- We use SWMC to perform broad trials on occasion logs of certifiable logical work processes to assess the adequacy and effectiveness of our approach practically speaking.

ALGORITHM

Scientific workflows in cloud environments can be deployed and executed on a central cloud or multi-clouds. If workflows have loose deadlines, they can be executed in one cloud (e.g., a PM) in order to reduce cost and energy consumption. For the sake of efficiency, scientific workflows are usually scheduled and executed in multiple cloud platforms, i.e., distributed PMs. In the following, we will discuss how to mine scientific workflows in both situations.

Intra-Cloud Mining

Without loss of generality, we first assume that the event log of a scientific workflow is available in a centralized PM. While workflows can be mined from event logs, the mining results rely heavily on the completeness of event logs. First of all, global completeness is introduced.

The number of traces (topological sorts) of a scientific workflow W grows exponentially with the increase of the number of activities in W. On the other hand, scientific workflows are usually employed to solve big data problems, and thus they are long-running. Due to the above reasons, global completeness of an event log can be hardly satisfied within a short time. Here, the crux lies in whether we can rediscover W from a part of its traces. To rediscover W, all activities and edges should be derived from the event log L. The completeness of activities is easy to satisfy because all activities can be obtained from any trace. For the edge (called causal relation), say (A, B), there must exist at least one trace in which A directly precedes B. However, it

is not established conversely, for A and B can be parallel activities (A and B are with an interleaving relation). Fortunately, causal relation and interleaving relation can be differentiated as follows: there is a causal relation between A and B if there is a trace such that A directly precedes B and there is no trace such that B directly precedes A; there is an interleaving relation between A and B if there are two traces σ1 and σ2 in L such that A directly precedes B in σ1 and B directly precedes A in σ2. In this way, we can rediscover W from L which contains all direct precedences.

Based on the relation of direct precedence, both the causal relation →L and the interleaving relation ∥L can be formally defined: A→LB iff A>LB and B≯LA, A∥LB iff A>LB and B>LA. Besides the causal relation and interleaving relation, there is another relation transitive causal relation L: ALB iff ∃V1, V2, ..., Vn such that A→LV1, V1→LV2, ..., Vn→LB. Since causal relation corresponds to edge set in the workflow model, based on the causal relation, we are able to rediscover the scientific workflow W from the locally-complete event log L, which is formulated in Algorithm 1, where Ev(L) in Line 9 summarizes all activities in L. If l and n represent the length and the number of traces in L, respectively, the time complexity of Algorithm 1 is O(l×n).

Algorithm 1 Intra-cloud scientific workflow mining

Input: Event log L
Output: Workflow W
1: $>_L \leftarrow \emptyset$
2: $\rightarrow_L \leftarrow \emptyset$
3: **for** each $\sigma \in L$ **do**
4: **for** $i \leftarrow 1$ to $|\sigma| - 1$ **do**
5: $>_L \leftarrow >_L \cup \{\sigma[i] >_L \sigma[i+1]\}$
6: **for** each $A >_L B \in >_L$ **do**
7: **if** $B >_L A \notin >_L$ **then**
8: $\rightarrow_L \leftarrow \rightarrow_L \cup \{A \rightarrow_L B\}$
9: $W \leftarrow (Ev(L), \rightarrow_L)$
10: **return** W

Inter-Cloud Mining

In this subsection, we elaborate on scientific workflow mining across cloud platforms. The scientific workflow is partitioned into several subworkflows each of which is executed on an independent PM and communicates with one another through communicating activities, whereas the other activities are referred to as internal computing activities. Communicating activities are further divided into two categories: sending activities and receiving activities.

Suppose that a scientific workflow W = (N, E) is scheduled and executed on m distributed PMs, each of which executes a sub-workflow SWi = (SNi , SEi) (16i6m) of W such that N = ∪m i=1SNi , E = ∪m i=1(SEi∪Mi). Traces of SWi (e.g., σ∈SNi ∗) are recorded in the event log Li on PM pmi , where SNi ∗ represents the set of all sequences over alphabet SNi . The distribution and partition of event logs L1, L2, ..., Lm on m different PMs bring about a great challenge for scientific workflow mining.

Algorithm 2 Inter-cloud scientific workflow mining

Input: Event logs $\bigcup_{i=1}^{m} L_i$ and message logs $\bigcup_{i=1}^{m} ML_i$
Output: Workflow W
1: $N \leftarrow Ev(L_1) \cup Ev(L_2) \ldots \cup Ev(L_m)$
2: $E \leftarrow (\rightarrow_{L_1} \cup \rightarrow_{L_2} \ldots \cup \rightarrow_{L_m}) \cup (\rightarrow_{ML_1} \cup \rightarrow_{ML_2} \ldots \cup \rightarrow_{ML_m})$
3: $W' \leftarrow (N, E)$
4: $W \leftarrow r(W')$
5: **return** W

CONCLUSION

Mists are promising stages for the execution of logical work processes in the enormous information time. A logical work process mining approach is proposed in this paper, which encourages logical work process reuse and provenance investigation. Our approach bolsters both intra-cloud and between cloud logical work process mining, which is actualized and distributed as a ProM module SWMC. The broad tests on occasion logs of genuine logical work processes show both the adequacy and proficiency of our approach.

REFERENCES

[1] J. Zhang, D. Kuc, and S. Lu, "Confucius: A tool supporting collaborative scientific workflow composition," IEEE Trans. Services Computing, vol. 7, no. 1, pp. 2–17, 2014.

[2] R. N. Calheiros and R. Buyya, "Meeting deadlines of scientific workflows in public clouds with tasks replication," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 7, pp. 1787–1796, 2014.

[3] R. Duan, R. Prodan, and X. Li, "Multi-objective game theoretic schedulingof bag-of-tasks workflows on hybrid clouds," IEEE Trans. Cloud Computing, vol. 2, no. 1, pp. 29–42, 2014.

[4] Y. Zhao, Y. Li, I. Raicu, S. Lu, C. Lin, Y. Zhang, W. Tian, and R. Xue, "A service framework for scientific workflow management in the cloud," IEEE Trans. Services Computing, vol. 8, no. 6, pp. 930–940, 2015.

[5] G. Cordasco, R. D. Chiara, and A. L. Rosenberg, "An area-oriented heuristic for scheduling DAGs on volatile computing platforms," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 8, pp. 2164–2177, 2015.

[6] H. Wu, X. Hua, Z. Li, and S. Ren, "Resource and instance hour minimization for deadline constrained DAG applications using computer clouds," IEEE Trans. Parallel Distrib. Syst., vol. 27, no. 3, pp. 885–899, 2016.

[7] S. B. Davidson and J. Freire, "Provenance and scientific workflows: Challenges and opportunities," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '08, 2008, pp. 1345–1350.

[8] I. D. Santos, J. Dias, D. de Oliveira, E. S. Ogasawara, K. A. C. S. Ocana, and M. Mattoso, "Runtime dynamic structural changes of ˜ scientific workflows in clouds," in IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC'13, Dresden, Germany, December 9-12, 2013, pp. 417–422.

[9] A. C. Zhou and B. He, "Transformation-based monetary costoptimizations for workflows in the cloud," IEEE Trans. Cloud Computing, vol. 2, no. 1, pp. 85–98, 2014.

[10] M. A. Rodriguez and R. Buyya, "Deadline based resource provisioningand scheduling algorithm for scientific workflows on clouds," IEEE Trans. Cloud Computing, vol. 2, no. 2, pp. 222–235, 2014.

[11] Z. Li, J. Ge, H. Hu, W. Song, H. Hu, and B. Luo, "Cost and energy aware scheduling algorithm for scientific workflows with deadline constraint in clouds," IEEE Trans. Services Computing, PrePrints, doi:10.1109/TSC.2015.2466545.

[12] X. Xu, W. Dou, X. Zhang, and J. Chen, "Enreal: An energy-aware resource allocation method for scientific workflow executions in cloud environment," IEEE Trans. Cloud Computing, vol. 4, no. 2, pp. 166–179, 2016.

[13] W. M. P. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," IEEE Trans. Knowl. Data Eng., vol. 16, no. 9, pp. 1128–1142, 2004.

[14] A. Weijters, W. Aalst, and A. Medeiros, "Process Mining with the Heuristics Miner-algorithm," BETA Working Paper Series, WP 166, Eindhoven University of Technology, Eindhoven, 2006.

[15] R. Bergenthum, J. Desel, R. Lorenz, and S. Mauser, "Process mining based on regions of languages," in Business Process Management, 5th International Conference, BPM'07, Brisbane, Australia, September 24-28, Proceedings, 2007, pp. 375–383.

[16] W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, and C. W. Gunther, "Process mining: a two- ¨ step approach to balance between underfitting and overfitting," Software and System Modeling, vol. 9, no. 1, pp. 87–111, 2010.

[17] R. Lorenz, M. Huber, C. Etzel, and D. Zecha, "SYNOPS - generation of partial languages and synthesis of petri nets," in Proceedings of the International Workshop on Petri Nets and Software Engineering, Hamburg, Germany, June 25-26, 2012, pp. 237–252.

[18] M. Sole and J. Carmona, "Region-based foldings in process dis- ´ covery," IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 192–205, 2013.

[19] W. Song, H. Jacobsen, C. Ye, and X. Ma, "Process discovery from dependence-complete event logs," IEEE Trans. Services Computing, vol. 9, no. 5, pp. 714–727, 2016.

[20] M. Sadoghi, M. Jergler, H. Jacobsen, R. Hull, and R. Vacul´ın, "Safe distribution and parallel execution of data-centric workflows over the publish/subscribe abstraction," IEEE Trans. Knowl. Data Eng., vol. 27, no. 10, pp. 2824–2838, 2015.