International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

# Self-Adjusting Slot Configurations For Hadoop Clusters Using Data Security In Cloud

**M.K.Mohamed Faizal[1], P.Christopher[2], A.Joshua Issac[3]**

[1,2,3] *Assistant Professor, Department of Computer Science and Engineering, M.I.E.T Engineering College, Trichy, Tamilnadu,*

---

**Abstract**—*The Cloud ability to share information, substance and provides certain services to the people connected through the network. The Map Reduce framework in a cloud is an open source implementation, Hadoop have become the defect platform for scalable analysis on large data sets. this paper the major perspectives is to provide security to one's data which is stored remotely from the user's location. Motivated by this, propose The data security model includes generation of onetime password (OTP) using HMAC (Hash based message authentication code) for user authentication process. This model best suits for any of the layers in cloud e.g. PaaS, SaaS and IaaS, to achieve this we use certain encryption algorithms. This paper also describes how to minimize the completion length (i.e:makespan) of a set of Map Reduce jobs in hadoop. The current Hadoop only allows static slot configuration, i.e., fixed numbers of map -slots and reduce slots throughout the lifetime of a cluster. Found that such a static configuration may lead to low system resource utilizations as well as long completion length .Motivated by this, propose simple yet effective schemes which use slot ratio between map and reduce tasks as a tunable knob for reducing the make span of a given set. By leveraging the workload information of recently completed jobs, our schemes dynamically allocates resources (or slots) to map and reduce tasks.*

***Keywords— Authentication, Cloud Computing,OTP, HMAC, Mapreduce, Makespan, Static Configuration, Dynamic allocation resources.***

## Introduction

Cloud computing is the unfold trend in the recent times and also having rapid development over the past few years. There are certain problems present with the cloud services, as the resources are put in the hands of another provider; the user has no idea over regarding that environment. We are generally unaware, how our data is stored in the cloud, how security is provided for data. Security in cloud computing is the major part and one of the most important aspect for any organization where they move their substance to the cloud. They need to have confidence that their data is safe, both at the provider's site and during conveyance between the cloud user and owner. Furthermore, to protect the data we need the best authentication procedure that uses encryption algorithms, that offers better security. Many cloud providers are still using the same old login forms which do not offer any security and there is need to tighten up their security to ensure that data present in the cloud database is trustworthy and risk free. The authentication procedure in cloud computing must be comfortable to the user, but at the same time it should be very secure to protect the data that it Stored in the cloud. An encryption method should be used during a conveyance that offers security and that algorithm should not consume much computer power and processing time. Authentication in cloud computing is achieved by using the static passwords that do not offer any security to the user's present in the cloud .Static password can be easily cracked by the hackers as they are non complex passwords preferred by the users for their convenience. So static passwords must be replaced by the dynamic password Schemes that provide two way authentications in the cloud environment, and that should be cost effective both for the user and the provider as users cannot afford the device for the authentication. So cloud provides initiated the one time password schemes as a factor of two way authentication that sends a code to users mobile for every login session of the user.

Map reduce has become the leading paradigm in recent years for parallel big data processing. Its open source implementation Apache Hadoop has also emerged as a popular platform for daily data processing and information analysis. With the rise of cloud computing, Map Reduce is no longer just for internal data process in big companies. It is now convenient for a regular user to launch a Map Reduce cluster on the cloud, e.g., AWS Map Reduce, for data-intensive applications. When more and more applications are adopting the Map Reduce framework, how to improve the performance of a Map Reduce cluster becomes a focus of research and development. Both academia and industry have put tremendous efforts on job scheduling, resource management, and Hadoop applications . As a complex system, Hadoop is configured with a large set of system parameters. While it provides the flexibility to customize the cluster for different applications, it is challenging for users to understand and set the optimal values for those parameters. In this paper, we aim to develop algorithms for adjusting a basic system parameter with the goal to improve the performance (i.e., reduce the makespan) of a batch of Map Reduce jobs. The key idea of this new mechanism, named TuMM, is to automate the slot assignment ratio between map and reduce tasks in a cluster as a tunable knob for reducing the make span of Map Reduce jobs. The Workload Monitor (WM) and the Slot Assigner (SA) are the two major components introduced by TuMM. The WM that resides in the Job Tracker periodically collects the execution time information of recently finished tasks and estimates the present map and reduce workloads in the

cluster. The SA module takes the estimation to decide and adjust the slot ratio between map and reduce tasks for each slave node.

### 1.1 Problem Statement

Authentication in cloud computing is achieved by using the static passwords that do not offer any security to the User's present in the cloud. Static password can be easily cracked by the hackers as they are non complex Passwords preferred by the users for their convenience. So static passwords must be replaced by the dynamic password schemes that provide two way authentications in the cloud environment, and that should be cost effective both for the user and the provider as users cannot afford the device for the authentication. So cloud provides initiated the one time password schemes as a factor of two way authentication that sends a code to users mobile for every login session of the user. Finally data availability can be considered as a major concern which is viewed as Threat associated with the cloud environment. To overcome this problem we generally replicate our data and store in various locations.

## Related Work

### Optimizing the schedule of map reduce jobs to minimize Their makespan

Large-scale Hadoop clusters with their data intensive, Map Reduce style applications, that routinely process pet bytes of unstructured and semi-structured data, represent a new entity in the changing landscape of modern data centers.

A key challenge is to increase the utilization of these Map Reduce clusters. For a set of production jobs that are executed periodically on a new data, we can perform an offline analysis for evaluating performance benefits of different optimization techniques. In this work, we consider a subset of the production workload that consists of Map Reduce jobs with no dependencies. We observe that the order in which these jobs are executed can have a significant impact on their overall completion time and the cluster resource utilization. Our goal is to automate the design of a job schedule that minimizes the completion time (makespan) of such a set of Map Reduce jobs.

We introduce a simple abstraction where each Map Reduce job is represented as a pair of map and reduce stage durations. This representation enables us to apply the classic Johnson algorithm that was designed for building an optimal two-stage job schedule. Simulations performed over a set of realistic workloads demonstrate that 10%-25% makespan improvements are achievable by simply processing the jobs in the right order the simplified abstraction assumed by Johnson's algorithm may lead to a suboptimal job schedule. We design a novel heuristic, called Balanced Pools, that significantly improves Johnson's schedule results (up to 15%-38%), exactly in the situations when it produces suboptimal makespan. The results of our simulation study are validated through experiments on a 66- node Hadoop cluster. We apply the classic Johnson algorithm to construct an optimized schedule for a set of independent Map Reduce jobs.

We represent each Map Reduce job Ji by a pair of computed durations (mi , ri) of its map and reduce stages. This representation enables us to apply Johnson's algorithm that has been proposed for building the optimal two-stage job schedule. Since the set of production jobs is executed periodically, it permits their automated profiling from past executions. When jobs in a batch need to process new datasets, we use the knowledge of extracted job profiles to pre-compute new estimates of jobs' map and reduce stage durations, and then construct an optimized schedule for future executions.

### A One-Time Password System

Entity authentication is a process in which an entity proves his identity and his presence to another entity. Authentication requires both an identity guarantee, which is usually connected to the presence of a secret (for example a password) and a time guarantee which will be made by some time variant parameters - to ensure that this authentication did not happened before. Authentications are usually challenge response protocols in which an entity sends a random challenge to another entity who wishes to prove his identity. In this paper we will use the term user to denote the entity which needs to authenticate and the term system to denote the entity to which identity is be proven. Password authentication is the most commonly authentication.

## Proposed method

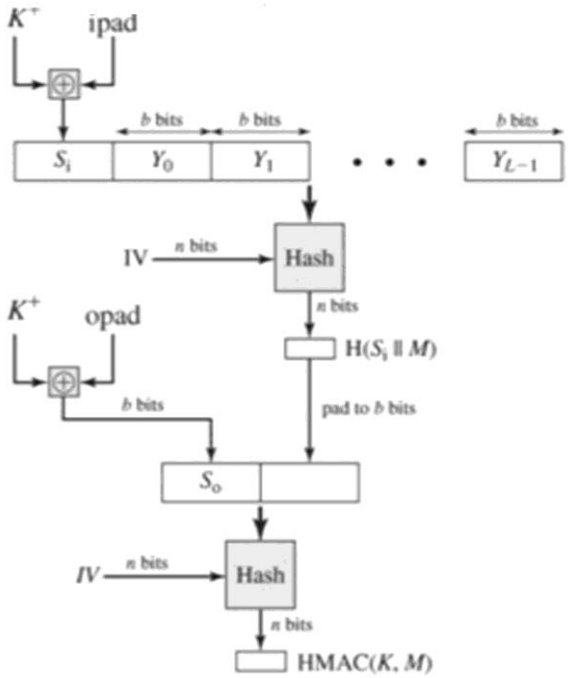
- Provide better password solution for login procedures than the insecure method of static passwords.
- Use an encryption algorithm that is secure but also fast, to be able to serve the vast amount of cloud users.
- Offer a solution that is free of charge in order to attract more customers to the cloud services.
- The solution presented here will be free of charge for both the users and the provider, and at the same time easy and flexible for the clients to download, install and use.
- So static passwords must be replaced by the dynamic password schemes that provide two way authentications in the cloud environment, and that should be cost effective both for the user and the provider as users cannot afford the device for the authentication.
- Propose and implement a new mechanism to dynamically allocate slots for map and reduce tasks.
- The primary goal of the new mechanism is to improve the completion time (i.e., the makespan).
- The key idea of this new mechanism, named TuMM, is to automate the slot assignment ratio between map and reduce tasks in a Cluster as a tunable knob for reducing the makespan of Map Reduce jobs.

- The Workload Monitor (WM) and the Slot Assigner (SA) are the two major components introduced By TuMM.
- The WM that resides in the Job Tracker periodically We further investigate
- The dynamic slot assignments in heterogeneous environments
- And propose a new version of TuMM, named H_TuMM, which sets the slot configurations for each individual node to reduce the makespan of a batch of jobs

## System Design

### HMAC Algorithm

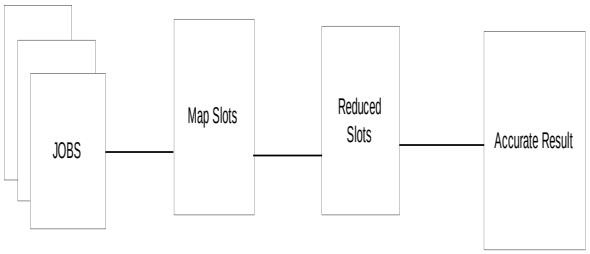
In cryptography, an HMAC involving a specific type of cryptographic hash functions and a secret cryptographic key . HMAC may used to simultaneously verify the both data integrity and authentication of a message . HMAC uses two passes of hash computation. The secret key is first used to derive two keys – inner and outer. The first phase is to pass the algorithm produces an internal hash derived from the message and the inner key. The second phase which passes the final HMAC code derived from the inner hash result and the outer key. Thus the algorithm provides better immunity against length extension attacks.




### Maximize the completion length

Map Reduce framework, how to improve the performance of a Map Reduce cluster becomes a focus of research and development as a complex system, Hadoop is configured with a large set of system parameters. While it provides the flexibility to customize the cluster for different applications, it is challenging for users to understand and set the optimal values for those parameters.

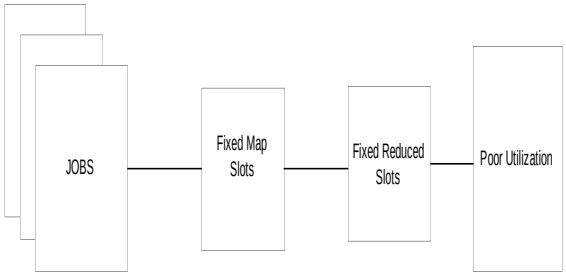
Minimizing Makespan




### Static slot configuration

The Hadoop framework, however, uses fixed numbers of map slots and reduce slots at each node as the default setting throughout the lifetime of a cluster.

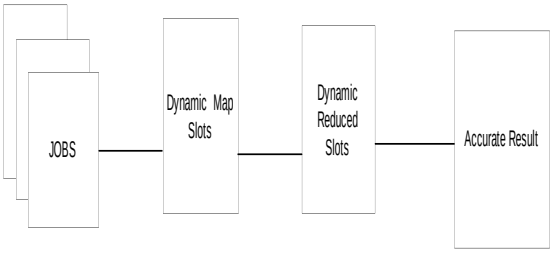
The values in this fixed configuration are usually heuristic numbers without considering job characteristics therefore this static setting is not well optimized and may hinder the performance improvement of the entire cluster.




### Dynamic Slots Configuation:

Dynamically allocates resources (or slots) to map and reduce tasks. We implemented the presented schemes in Hadoop evaluated them with representative Map Reduce. The primary goal of the new mechanism is to improve the completion time (i.e., the makespan) of a batch of Map Reduce jobs while retain the simplicity in implementation and management of the slot-based Hadoop design.

Slot assignments in heterogeneous environments, and propose a new version of TuMM, named H_TuMM, which sets the slot configurations for each individual node to reduce the makespan of a batch of jobs.



### Ratio between map and reduce tasks.

Dynamically allocate slots for map and reduce tasks the new mechanism, named TuMM, is to automate the slot assignment ratio between map and reduce tasks in a cluster as a tunable knob for reducing the makespan of Map Reduce jobs.

## Conculsion

Certainly cloud computing will be a boon in enhancing information systems as its benefits out number its shortcomings. Cloud computing offers deployment architecture, with the ability to address vulnerabilities recognized in traditional IS but its dynamic Characteristics are able to prevent the effectiveness of Traditional counter measures. Here, we have identified generic design principles of a Cloud environment which stem from the necessity to Control relevant vulnerabilities and threats. So, for this scenario we have proposed to make use of Dynamic one time password with two factor authentication as a Strong authentication technique. We presented a novel slot management scheme, named TuMM, to enable dynamic slot configuration in Hadoop. The main objective of TuMM is to improve resource utilization and reduce the makespan of multiple jobs. To meet this goal, the presented scheme introduces two main components: Workload Monitor periodically tracks the execution information of recently completed tasks and estimates the present workloads of map and reduce tasks and Slot Assigner dynamically allocates the slots to map and reduce tasks by leveraging the estimated workload information. We further extended our scheme to manage resources (slots) for heterogeneous clusters. The new version of our scheme, named H_TuMM, reduces the makespan of multiple jobs by separately setting the slot assignments for the node in a heterogeneous cluster. We implemented TuMM and H_TuMM on the top of Hadoop v0.20.2 and evaluated both schemes by running representative Map Reduce benchmarks and TPC-H query sets in Amazon EC2 clusters. The experimental results demonstrate up to 28 percent reduction in the makespan and 20 percent increase in resource utilizations. The effectiveness and the robustness of our new slot management schemes are validated under both homogeneous and heterogeneous cluster environments. In the future, we will further investigate the optimal total slot number configuration in the slot based Hadoop platform as well as the resource management policy in next generation Hadoop YARN platforms.

## References

[1] J. Dean and S. Ghemawat, “Map Reduce: Simplified data processing on large clusters,” Common. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[2] Apache Hadoop. [Online]. Available: http://hadoop.apache.org/,2015.

[3] M. Zaharia, D. Borthakur, J. S. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, “Delay scheduling: A simple technique for achieving locality and fairness in cluster scheduling,” in Proc. 5th Eur. Conf. Comput. Syst., 2010, pp. 265–278.

[4] A. Verma, L. Cherkasova, and R. H. Campbell, “Two sides of a coin: Optimizing the schedule of mapreduce jobs to minimize their makespan and improve cluster performance,” in Proc. Proc. IEEE 20th Int. Symp. Model., Anal., Simul. Comput. Telecommun., Aug. 2012, pp. 11–18.

[5] M. Isard, Vijayan Prabhakaran, J. Currey, U. Wieder, K. Talwar,and A. Goldberg, “Quincy: Fair scheduling for distributed computing clusters,” in Proc. ACM SIGOPS 22nd Symp. Oper. Syst. Principles, 2009, pp. 261–276.

[6] Dr. Mark D. Bedworth PhD BSc FSS. February 2008. A Theory of Probabilistic One-Time Password. Computer Science Computer Engineering and Applied Computing, Security and Management.

[7] Kiddo. 2010. Hacking Website: Menemukan Celah Keamanan & MelindungiWebsite dariSerangan Hacker. Mediakita

[8] Rivest, Ronald L. 1992.The MD5 Message Digest Algorithm.

[9] Myung-Jun Kim, “Korea's Cloud Computing Strategy," IT21 Global Conference, 2009

[10] Hyun-Seong Kim, Choon-Sik Park, “Cloud computing and the personal authenticationservice.” Journal of the Information Security, vol. 20