# Big Data Technologies

Editor: Gregor von Laszewski

March 16, 2017

## 0.1 Contributors

| Name | HID | Title | Pages |
|---|---|---|---|
| Avadhoot Agasti | SL-IO-3000 | TBD | 1 |

# Apache Ranger

AVADHOOT AGASTI[1,*,+]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: aagasti@indiana.edu
[+] HID - SL-IO-3000

paper2, March 10, 2017

**Apache Hadoop provides various data storage, data access and data processing services. Apache Ranger is part of the Hadoop eco-system. Apache Ranger provides capability to perform security administration tasks for storage, access and processing of data in Hadoop. Using Ranger, Hadoop administrator can perform security administration tasks using a central UI or Restful web services. He can define policies which enable users/user-groups to perform specific action using Hadoop components and tools. Ranger provides role based access control for datasets on Hadoop at column and row level. Ranger also provides centralized auditing of user access and security related administrative actions.**



**Keywords:** Apache Ranger, LDAP, Active Directory, Apache Knox, Apache Atlas, Apache Hive, Apache Hadoop, Yarn, Apache HBase, Apache Storm, Apache Kafka, Data Lake, Apache Sentry

https://github.com/avadhoot-agasti/sp17-i524/tree/master/paper2/S17-IO-3000/report.pdf

## 1. INTRODUCTION

Apache Ranger is open source software project designed to provide centralized security services to various components of Apache Hadoop. Apache Hadoop provides various mechanism to store, process and access the data. Each Apache tool has its own security mechanism. This increases administrative overhead and is also error prone. Apache Ranger fills this gap to provide a central security and auditing mechanism for various Hadoop components. Using Ranger, Hadoop administrator can perform security administration tasks using a central UI or Restful web services. He can define policies which enable users/user-groups to perform specific action using Hadoop components and tools. Ranger provides role based access control for datasets on Hadoop at column and row level. Ranger also provides centralized auditing of user access and security related administrative actions.

## 2. ARCHITECTURE OVERVIEW

[1] describes the important components of Ranger as explained below:

### 2.1. Ranger Admin Portal

Ranger Admin Portal is the main interaction point for the user. Using Admin Portal, user can define policies. The policies are stored in Policy Database. The Policies are polled by various plugins. The Admin Portal also collects the audit data from plugins and stores in HDFS or in a relational database.

### 2.2. Ranger Plugins

Plugins are Java Programs, which are invoked as part of the cluster component. For example, the ranger-hive plugin is embedded as part of Hive Server2. The plugins cache the policies, and intercept the user request and evaluates it against the policies. Plugins also collect the audit data for that specific component and send to Admin Portal.

### 2.3. User group sync

While Ranger provides authorization or access control mechanism, it needs to know the users and the groups. Ranger integrates with Unix users management or LDAP or Active Directory to fetch the users and groups. The User group sync component is responsible for this integration.

## 3. HADOOP COMPONENTS SUPPORTED BY RANGER

Ranger supports auditing and authorization for following Hadoop components [2].

### 3.1. Apache Hadoop and HDFS

Apache Ranger provides plugin for Hadoop, which helps in enforcing data access policies. The HDFS plugin works with Name Node to check if the user's access request to a file on HDFS is valid or not.

### 3.2. Apache Hive

Apache Hive provides SQL interface on top of the data stored in HDFS. Apache Hive supports two types of authorization -

Storage based authorization and SQL standard authorization. Ranger provides centralized authorization interface for Hive which provides granular access control at table and column level. Ranger implements a plugin which is part of Hive Server2.

### 3.3. Apache HBase

Apache HBase is NoSQL database implemented on top of Hadoop and HDFS. Ranger provides coprocessor plugin for HBase, which performs authorization checks and audit log collections.

### 3.4. Apache Storm

Ranger provides plugin to Nimbus server which helps in performing the security authorization on Apache Storm.

### 3.5. Apache Knox

Apache Knox provides service level authorization for users and groups. Ranger provides plugin for Knox using which, administration of policies can be supported. The audit over Knox data enables user to perform detailed analysis of who and when accessed Knox.

### 3.6. Apache Solr

Solr provides free text search capabilities on top of Hadoop. Ranger is useful to protect Solr collections from unauthorized usage.

### 3.7. Apache kafka

Ranger can manager access control on Kafka topics. Policies can be implemented to control which users can write to a Kafka topic and which users can read from a Kafka topic.

### 3.8. Yarn

Yarn is resource management layer for Hadoop. Administrators can setup queues in Yarn and then allocate users and resources per queue basis. Policies can be defined in Ranger to define who can write to various Yarn queues.

## 4. IMPORTANT FEATURES OF RANGER

The blog article [3] explains the 2 important features of Apache Ranger.

### 4.1. Dynamic Column Masking

Dynamic data masking at column level is an important feature of Apache Ranger . Using this feature, the administrator can setup data masking policy. The data masking makes sure that only authorized users can see the actual data while other users will see the masked data. Since the masked data is format preserving, they can continue their work without getting access to the actual sensitive data. For example, the application developers can use masked data to develop the application whereas when the application is actually deployed, it will show actual data to the authorized user. Similarly, a security administrator may chose to mask credit card number when it is displayed to a service agent.

### 4.2. Row Level Filtering

The data authorization is typically required at column level as well as at row level. For example, in an organization which is geographically distributed in many locations, the security administrator may want to give access of a data from a specific location to the specific user. In other example, a hospital data security administrator may want to allow doctors to see only his or her patients. Using Ranger, such row level access control can be specified and implemented.

## 5. HADOOP DISTRIBUTION SUPPORT

Ranger can be deployed on top of Apache Hadoop. [4] provides detailed steps of building and deploying Ranger on top of Apache Hadoop.

Hortonwork Distribution of Hadoop (HDP) supports Ranger deployment using Ambari. [5] provides installation, deployment and configuration steps for Ranger as part of HDP deployment.

Cloudera Hadoop Distribution (CDH) does not support Ranger. According to [6], Ranger is not recommended on CDH and instead Apache Sentry should be used as central security and audit tool on top of CDH.

## 6. USE CASES

Apache Ranger provides centralized security framework which can be useful in many use cases as explained below.

### 6.1. Data Lake

[7] explains that storing many types of data in the same repository is one of the most important feature of Data Lake. With multiple datasets, the ownership, security and access control of the data becomes primary concern. Using Apache Ranger, the security administrator can define fine grain control on the data access.

### 6.2. Multi-tenant Deployment of Hadoop

Hadoop provides ability to store and process data from multiple tenants. The security framework provided by Apache Ranger can be utilized to protect the data and resources from un-authorized access.

## 7. APACHE RANGER AND APACHE SENTRY

According to [8], Apache Sentry and Apache Ranger have many features in common. Apache Sentry ([9]) provides role based authorization to data and metadata stored in Hadoop.

## 8. EDUCATIONAL MATERIAL

[10] provides tutorial on topics like A)Security resources B)Auditing C)Securing HDFS, Hive and HBase with Knox and Ranger D) Using Apache Atlas' Tag based policies with Ranger. [11] provides step by step guidance on getting latest code base of Apache Ranger, building and deploying it.

## 9. LICENSING

Apache Ranger is available under Apache 2.0 License.

## 10. CONCLUSION

Apache Ranger is useful to Hadoop Security Administrators since it enables the granular authorization and access control. It also provides central security framework to different data storage and access mechanism like Hive, HBase and Storm. Apache Ranger also provides audit mechanism. With Apache Ranger, the security can be enhanced for complex Hadoop use cases like Data Lake.

## ACKNOWLEDGEMENTS

The authors thank Prof. Gregor von Laszewski for his technical guidance.

## REFERENCES

[1] Hortonworks, “Apache ranger - overview,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: https://hortonworks.com/apache/ranger/#section_2

[2] A. S. Foundation, “Apache ranger - frequenty asked questions,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: http://ranger.apache.org/faq.html#How_does_it_work_over_Hadoop_and_related_components

[3] S. Mahmood and S. Venkat, “For your eyes only: Dynamic column masking & row-level filtering in hdp2.5,” Web Page, Sep. 2016, online; accessed 9-Mar-2017. [Online]. Available: https://hortonworks.com/blog/eyes-dynamic-column-masking-row-level-filtering-hdp2-5/

[4] A. S. Foundation, “Apache ranger 0.5.0 installation,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: https://cwiki.apache.org/confluence/display/RANGER/Apache+Ranger+0.5.0+Installation

[5] Horto, “Installing apache rang,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.6/bk_installing_manually_book/content/ch_installing_ranger_chapter.html

[6] Cloudera, “Configuring authorization,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: https://www.cloudera.com/documentation/enterprise/5-6-x/topics/sg_authorization.html

[7] Teradata and Hortonworks, “Putting the data lake to work - a guide to best practices,” Web Page, Apr. 2014, online; accessed 9-Mar-2017. [Online]. Available: https://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf

[8] S. Neumann, “5 hadoop security projects,” Web Page, Nov. 2014, online; accessed 9-Mar-2017. [Online]. Available: https://www.xplenty.com/blog/2014/11/5-hadoop-security-projects/

[9] A. S. Foundation, “Apache sentry,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: https://sentry.apache.org/

[10] Hortonworks, “Apache ranger overview,” Web, online; accessed 9-Mar-2017. [Online]. Available: https://hortonworks.com/apache/ranger/#tutorials

[11] A. S. Foundation, “Apache ranger - quick start guide,” Web Page, online; accessed 9-Mar-2017. [Online]. Available: http://ranger.apache.org/quick_start_guide.html