Implementing data masking/role-based security in configured HDInsight Cluster

Lab 4 – Role-based security in configured HDInsight cluster

Overview

In this lab, you will provision an HDInsight cluster. You will then run a sample MapReduce job on the cluster and view the results.

What You’ll Need

To complete the labs, you will need the following:

* A web browser
* A Microsoft account
* A Microsoft Azure subscription
* A Microsoft Windows, Linux, or Apple Mac OS X computer on which the Azure CLI has been installed.
* Azure HDInsight cluster with Enterprise Security Package(ESP)

The lab files for this course

**Note**: ESP is GA in HDI 3.6 for Apache Spark, Interactive, and Apache Hadoop. ESP for Apache HBase and Apache Kafka cluster types is in preview.

## Prerequisites:

* Enable Azure AD domain services by following this KB article: <https://docs.microsoft.com/en-us/azure/active-directory-domain-services/active-directory-ds-getting-started>
* While Azure AD-DS is enabled, next all users & objects can start synchronizing from Azure AD(AAD) to Azure AD-DS by default. The length of the sync operations should be dependant on the number of objects in Azure AD. The sync could take a few days for hundreds of thousands of objects.
* Customer can choose to sync only the groups that need access to the HDInsight clusters. This options for syncing only certain groups is called scoped synchronization. Here goes the link for configuration of Scoped synchronization: <https://docs.microsoft.com/azure/active-directory-domain-services/active-directory-ds-scoped-synchronization>
* Enablement of secure LDAP: You must create a self-signed certificate for secure LDAP using PowerShell. Here goes the step by step KB: <https://docs.microsoft.com/en-us/azure/active-directory-domain-services/active-directory-ds-admin-guide-configure-secure-ldap>
* Create & authorize a managed identity using Azure portal & provide HDI Domain service contributor role to the managed identity in Azure AD-DS access control. Link: <https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/how-to-manage-ua-identity-portal>
* You need to also make sure to configure network like configuration of Azure AD-DS VNET to use these custom DNS servers, locating the right IP address . Keep Azure AD-DS instance & HDInsight cluster in the same VNET. If they are in different VNET, configure VNET peering(<https://docs.microsoft.com/en-us/azure/virtual-network/virtual-network-peering-overview>)
* Provision a HDInsight cluster with ESP making enterprise security package as Enabled. You need to provide the following parameters like:
* **Cluster admin user**: Choose an admin for your cluster from your synced Azure AD-DS. This domain account must be already synced and available in Azure AD-DS.
* **Cluster access groups**: The security groups whose users you want to sync and have access to the cluster should be available in Azure AD-DS. For example, HiveUsers group. For more information, see [Create a group and add members in Azure Active Directory](https://docs.microsoft.com/en-us/azure/active-directory/fundamentals/active-directory-groups-create-azure-portal).
* **LDAPS URL**: An example is ldaps://contoso.com:636.

## Connecting to Apache Ranger Admin UI

1. You can connect to Ranger Admin UI using the following url : https://<YourHDIClusterName>.azurehdinsight.net/Ranger/
2. Login in using the cluster administrator domain user credentials. Ranger only works for yarn & hive. You need to create two users at least ‘hiveuser1’ & ‘hiveuser2’.
3. Under Ranger admin UI, you need to configure Ranger policies & user group permission like as the following. You need to save this policy.

* Policy name: read-hivesampletable-all
* Hive Database: default
* table: hivesampletable
* Hive column: \*
* Select User: hiveuser1
* Permissions: select

1. Example hive Ranger policy for Azure HDInsight cluster:

Policy name: read-hivesampletable-devicemake

Hive Database: default

table: hivesampletable

Hive column: clientid, devicemake

Select User: hiveuser2

Permissions: select

