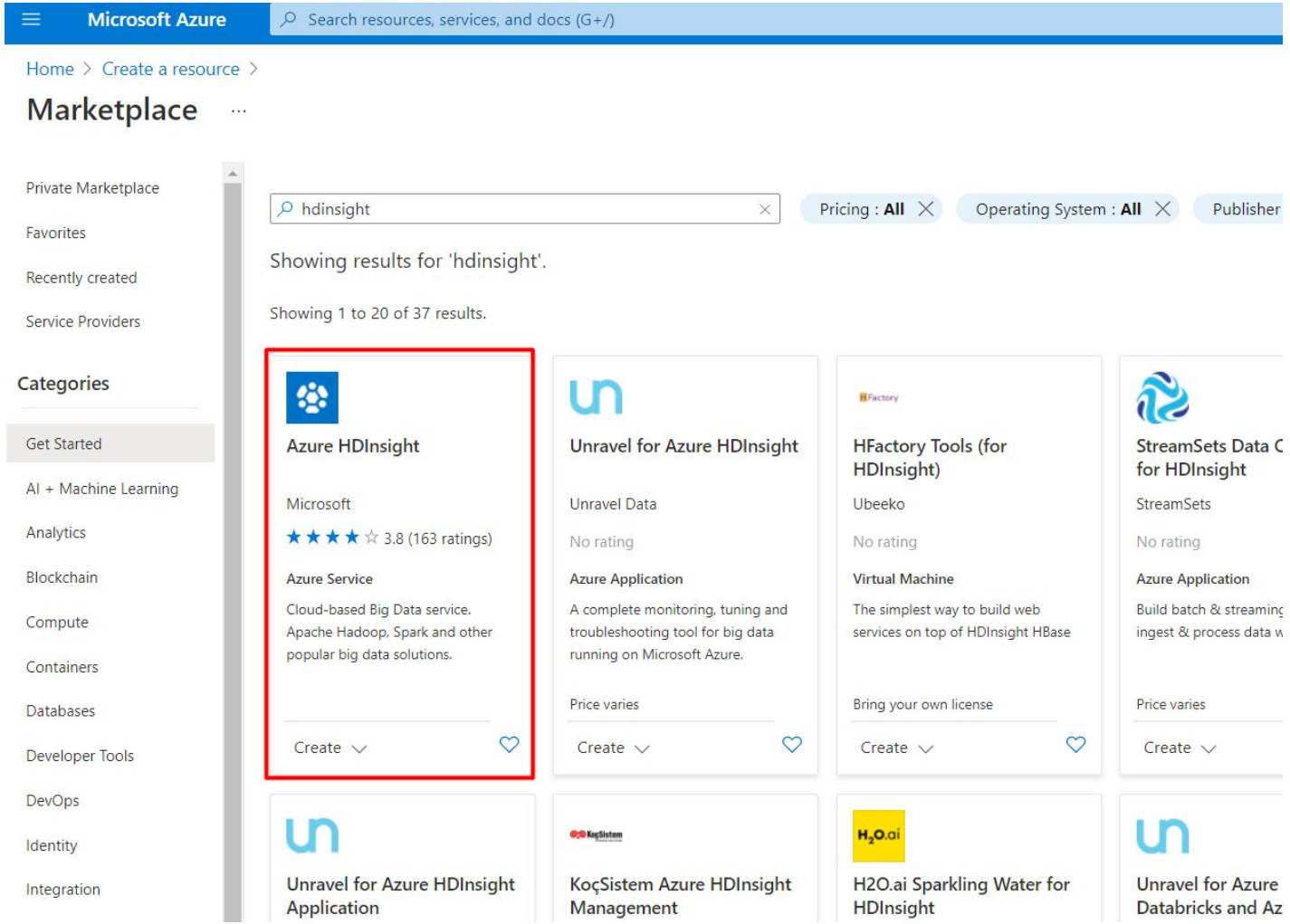
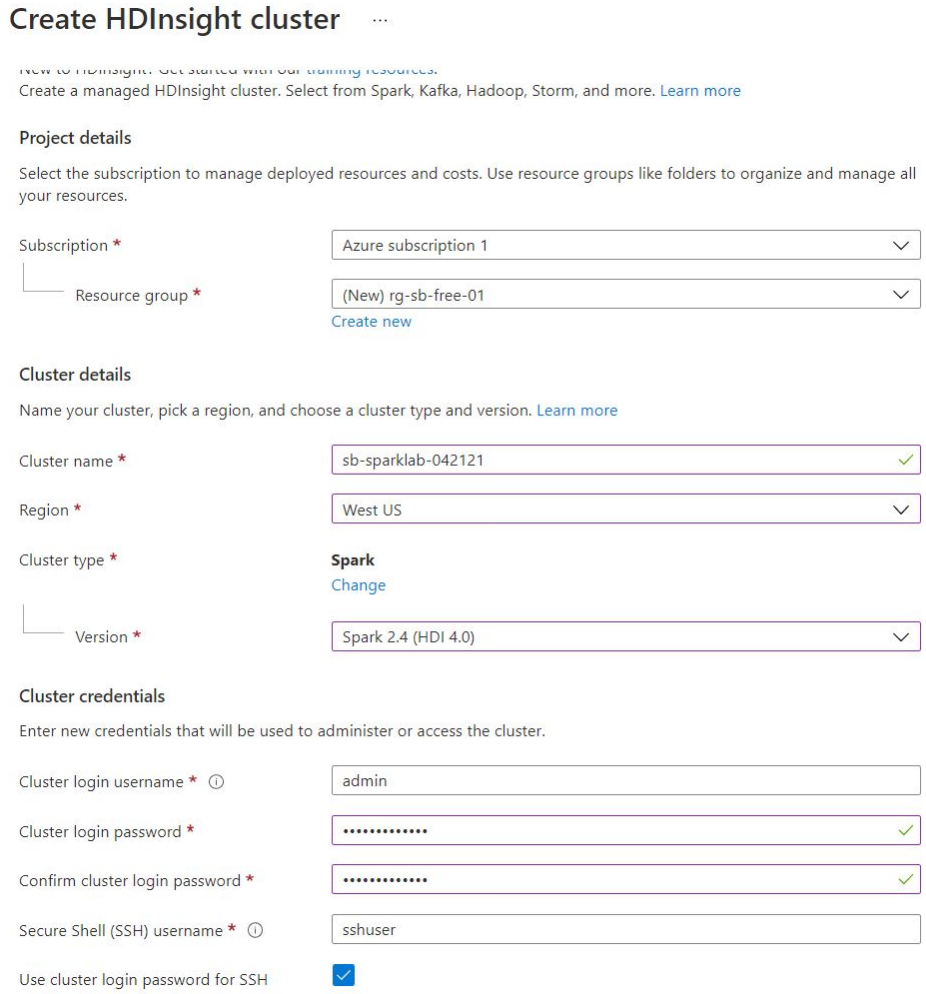
After logging into your Azure Portal, click “Create a resource”. This will take you to the Marketplace where you can search for Azure HDInsight.



Fill in the required fields and make sure to select Spark as the Cluster type and pick the latest version.



Create HDInsight cluster ...

Basics | Storage | Security + networking | Configuration + pricing | Tags | Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type * : Azure Storage

Selection method * : (•) Select from list ( ) Use access key

Primary storage account * : (New) sbsparklab042hdistorage

Create new

Container * : sb-sparklab-042121-2021-04-21t18-50-37-763z

Data Lake Storage Gen1

Provide details for the cluster to access Data Lake Storage Gen1. The cluster will be able to access any Data Lake Storage Gen1 accounts that the chosen service principal has access to.

Data Lake Storage Gen1 access — Configure access settings

Additional Azure Storage

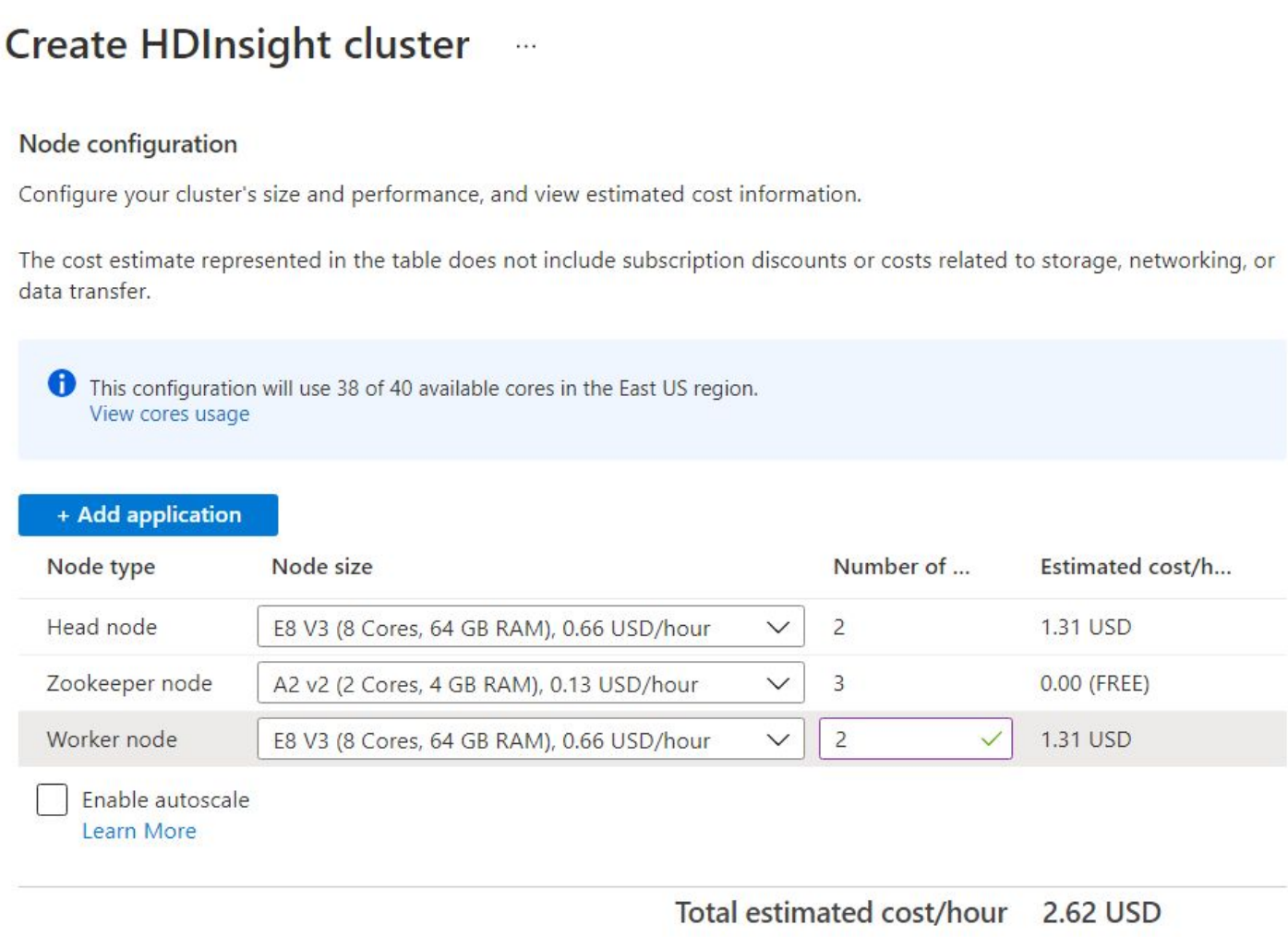
Link additional Azure Storage accounts to the cluster.

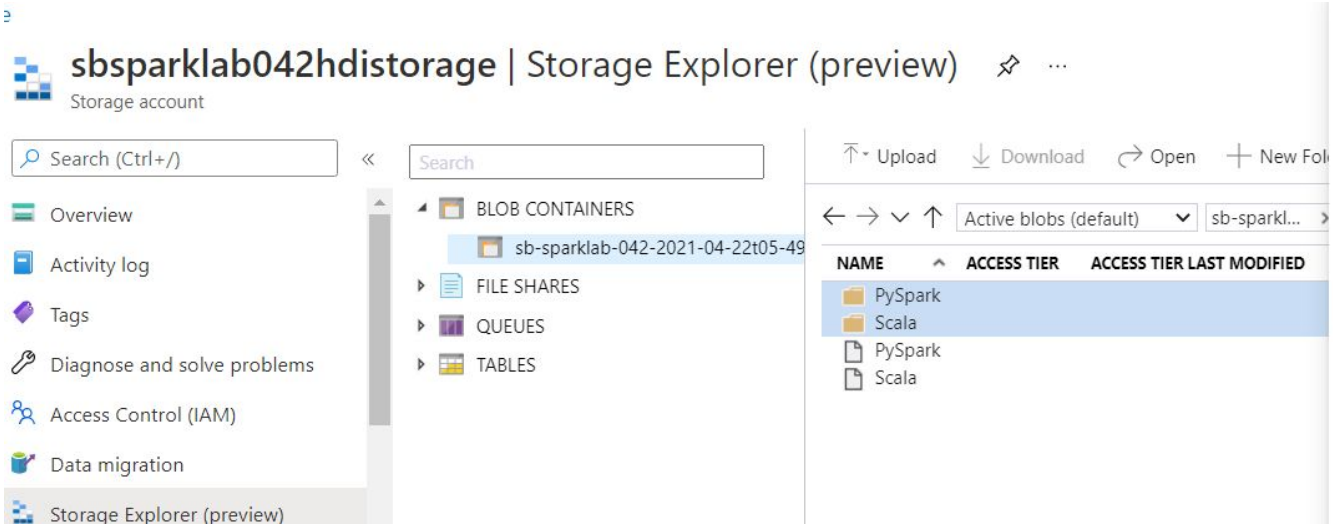
Add Azure Storage

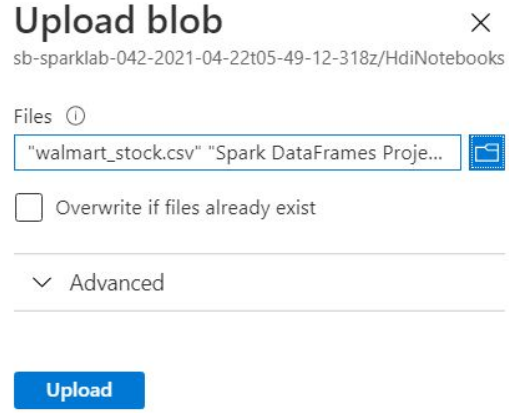
Custom Ambari DB

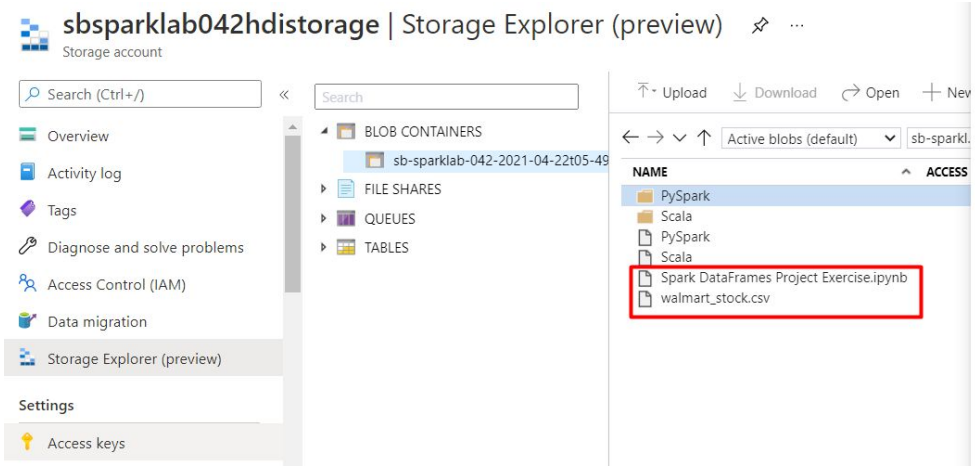
Use an external Ambari database for greater flexibility, control, and customization. Learn More

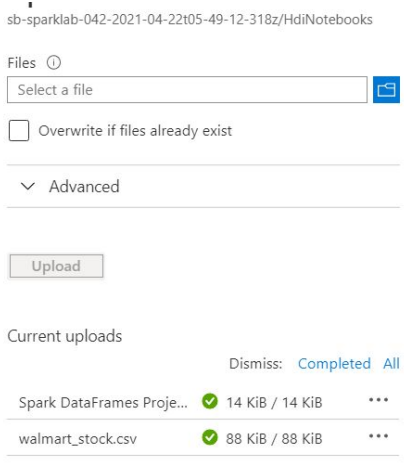
Since there is a maximum of 40 available cores, I tweaked the number of worker nodes to fit the limit. The data we are processing wasn't large so there weren't any issues. After configuring the cluster size, hit the create button and wait for completion.



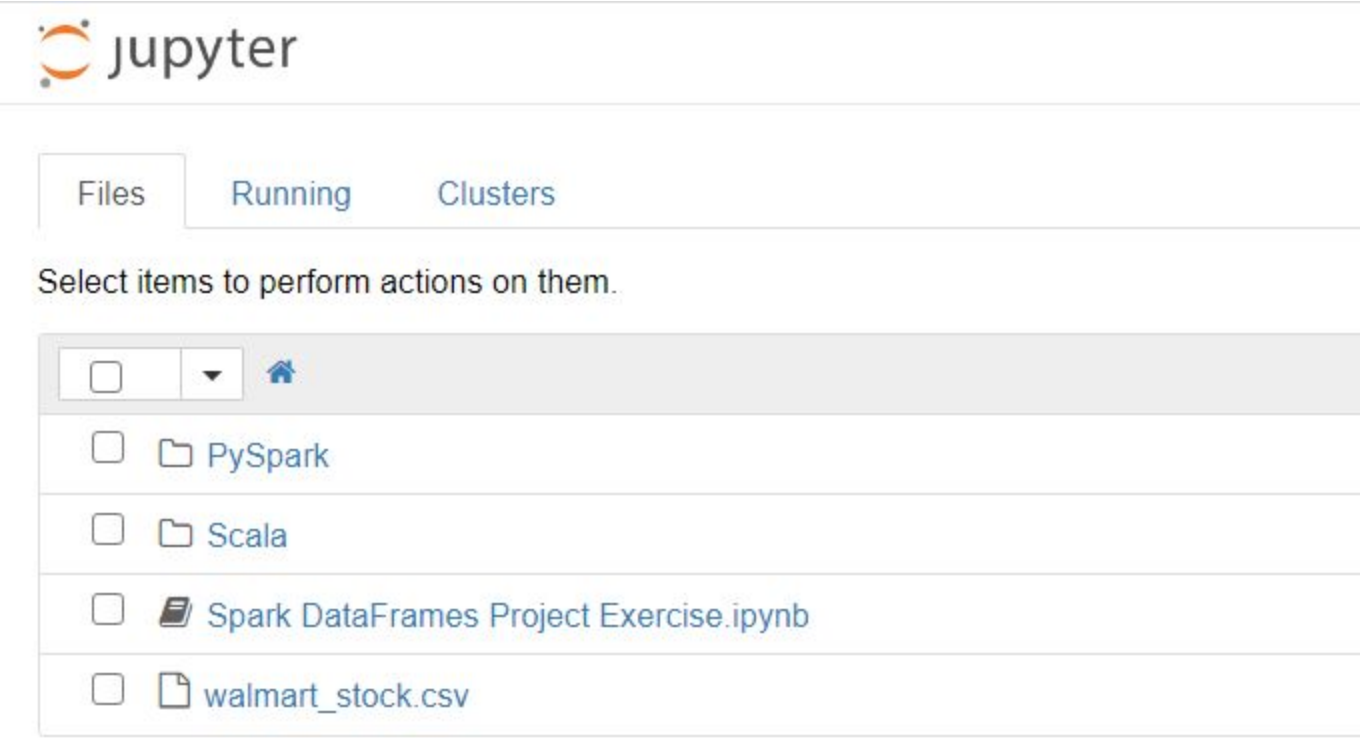




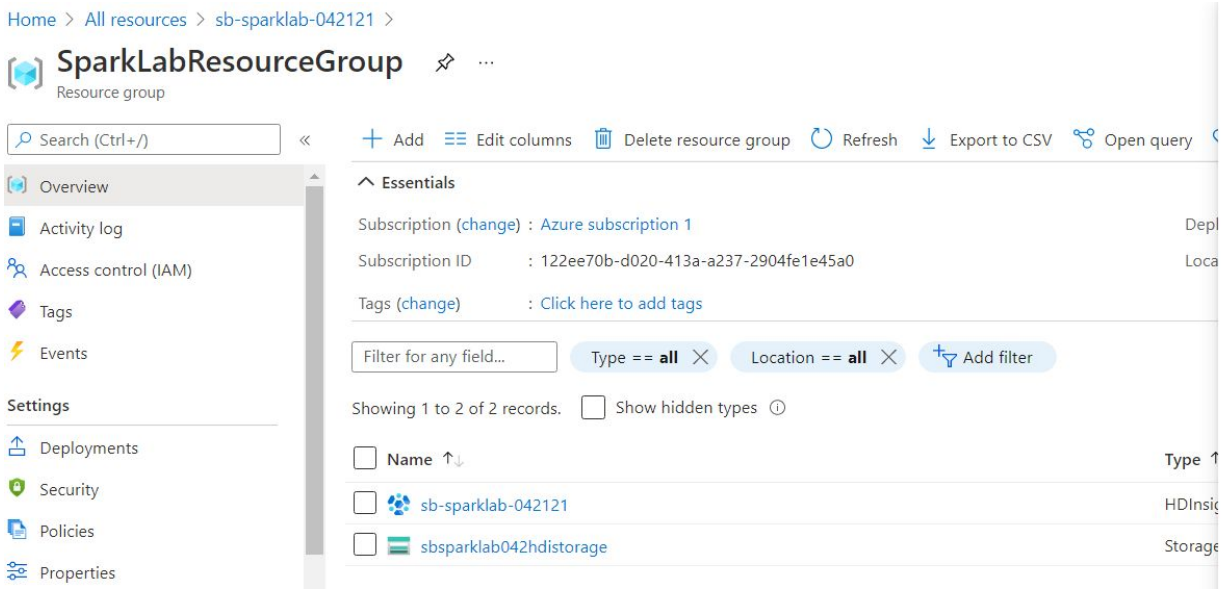


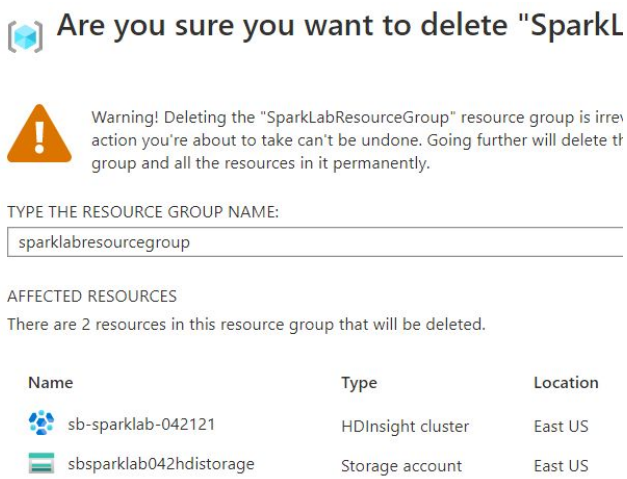


Insert the Jupyter Notebook and csv files into the “HdiNotebooks” folder.



Return to the HDInsight Cluster and find Jupyter Notebook under the dashboards header in “Overview”. This is launch Jupyter Notebook in your browser where can see the files you uploaded earlier.





Be sure to delete the cluster once the tasks are completed to avoid unnecessary charges.