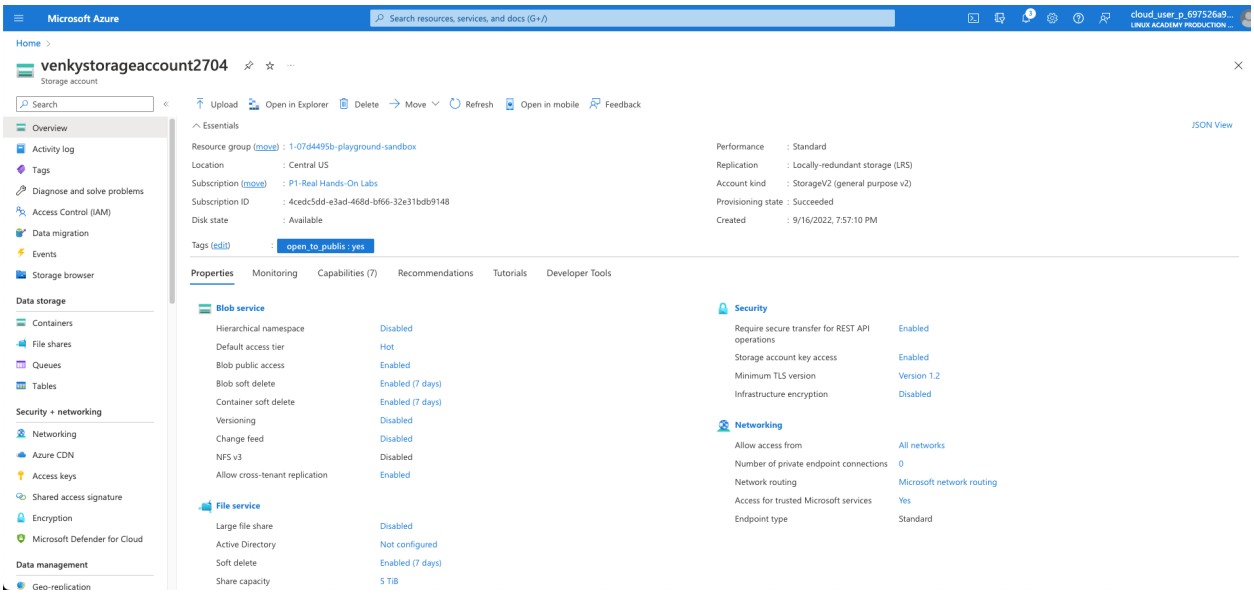
Azure Blob Storage & Azure Data Lake Storage Experiments

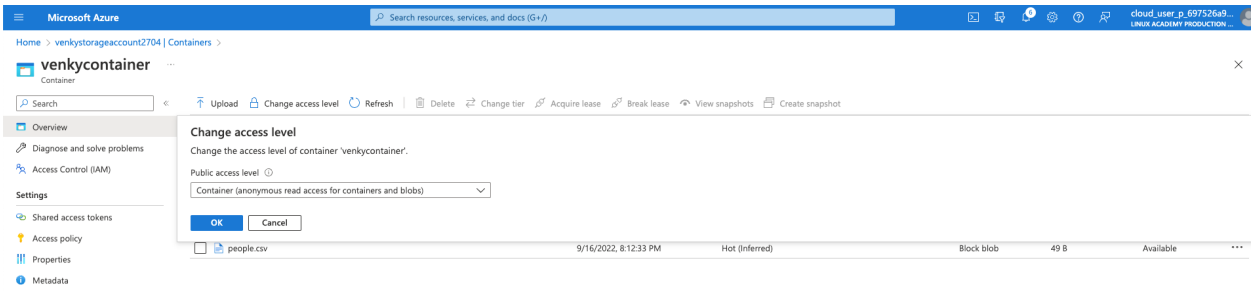
The idea of this document is to play with Azure storage and see whether we can access the data in there from a local spark/Hadoop cluster.

1: Install Azure CLI on MacOS via Homebrew.
brew install azure-cli

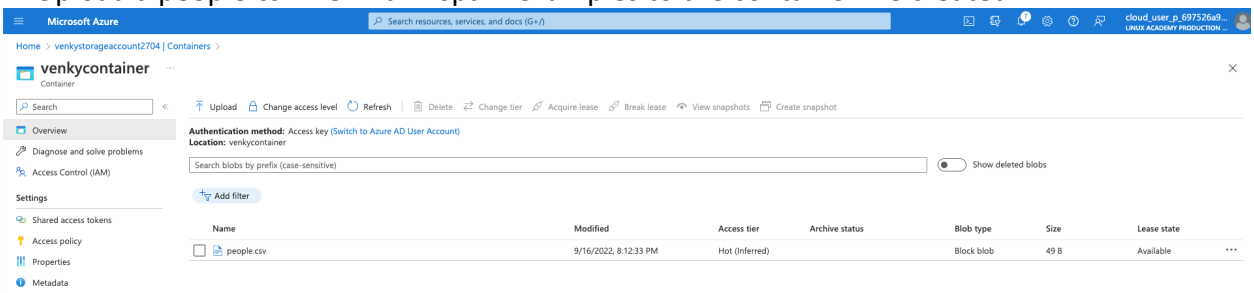
2. Login to the azure account and create a storage account with no hierarchical namespace enabled (regular blob storage). Public access is enabled.



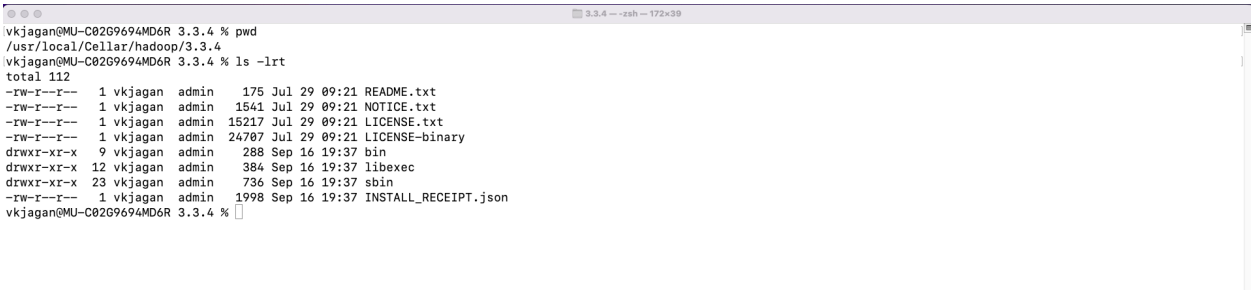
3. Create a container under the storage account with access level set at container anonymous access enabled.



4. Upload a people.csv we find in spark examples to the container we created.



5. Install Hadoop on mac via homebrew.
brew install hadoop.



Make sure we add the ADLS access jars to the classpath.

```
export HADOOP_OPTIONAL_TOOLS=hadoop-azure

hadoop fs -ls
wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
```

```
vkjagan@MU-C02G9694MD6R 3.3.4 % hadoop fs -ls wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
2022-09-16 20:29:09,941 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-09-16 20:29:10,228 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-09-16 20:29:10,292 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-09-16 20:29:10,292 INFO impl.MetricsSystemImpl: azure-file-system metrics system started
Found 1 items
-rwxrwxrwx   1         49 2022-09-16 20:12 wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/people.csv
2022-09-16 20:29:10,936 INFO impl.MetricsSystemImpl: Stopping azure-file-system metrics system...
2022-09-16 20:29:10,936 INFO impl.MetricsSystemImpl: azure-file-system metrics system stopped.
2022-09-16 20:29:10,936 INFO impl.MetricsSystemImpl: azure-file-system metrics system shutdown complete.
vkjagan@MU-C02G9694MD6R 3.3.4 %
```

As we can see, with container level anonymous access granted, we can use the Hadoop ls command to get to the data stored in the container.

Change access level to disallow anonymous access.

As we can see below, the list call now fails.

```
vkjagan@MU-C02G9694MD6R 3.3.4 % hadoop fs -ls wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
2022-09-16 20:33:02,930 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-09-16 20:33:03,224 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-09-16 20:33:03,294 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-09-16 20:33:03,294 INFO impl.MetricsSystemImpl: azure-file-system metrics system started
2022-09-16 20:33:03,809 WARN fs.FileSystem: Failed to initialize fileystem wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/: org.apache.hadoop.fs.azure
.AzureException: org.apache.hadoop.fs.azure.AzureException: No credentials found for account venkystorageaccount2704.blob.core.windows.net in the configuration, and its con
tainer venkycontainer is not accessible using anonymous credentials. Please check if the container exists first. If it is not publicly available, you have to provide accoun
t credentials.
2022-09-16 20:33:03,810 INFO impl.MetricsSystemImpl: Stopping azure-file-system metrics system...
2022-09-16 20:33:03,810 INFO impl.MetricsSystemImpl: azure-file-system metrics system stopped.
2022-09-16 20:33:03,810 INFO impl.MetricsSystemImpl: azure-file-system metrics system shutdown complete.
ls: org.apache.hadoop.fs.azure.AzureException: No credentials found for account venkystorageaccount2704.blob.core.windows.net in the configuration, and its container venkyc
ontainer is not accessible using anonymous credentials. Please check if the container exists first. If it is not publicly available, you have to provide account credentials
.
vkjagan@MU-C02G9694MD6R 3.3.4 %
```

To set credentials when we make the call, we need to follow the documentation presented here.
https://hadoop.apache.org/docs/stable/hadoop-azure/index.html

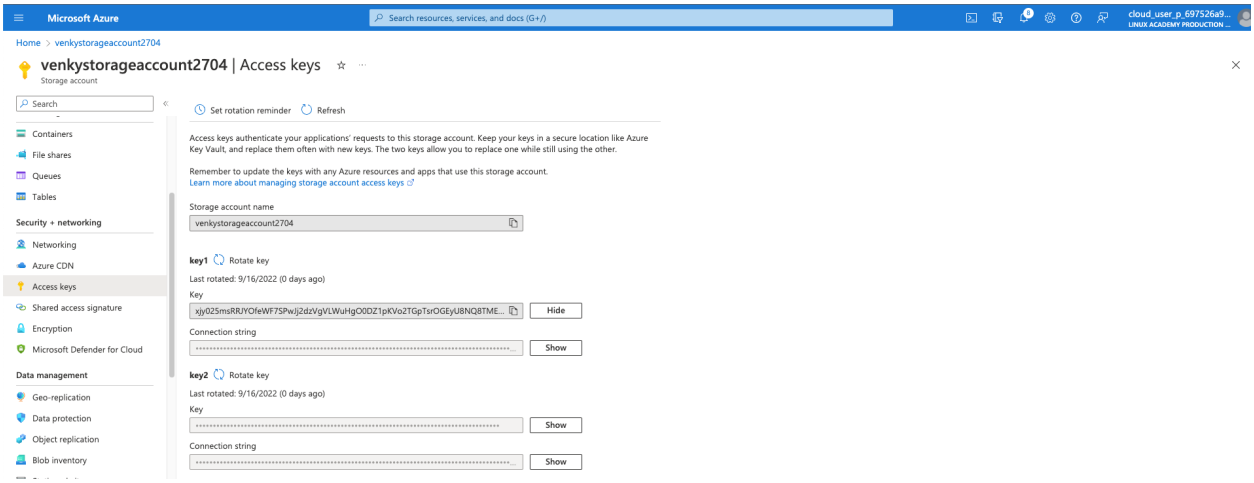
We must create a core-site.xml in the directory where brew installed Hadoop.
Change directory to the correct directory eg.
`/usr/local/Cellar/hadoop/3.3.4/libexec/hadoop`

REPLACE the file with the following contents:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>
  <name>fs.azure.account.key.venkystorageaccount2704.blob.core.windows.net</name>
  <value>xjy025msRRJYOfeWF7SPwJj2dzVgVLWuHgO0DZ1pKVo2TGpTsrOGEyU8NQ8TMEDNb9VDim9gKQi9+AStI0YUsg==</value>
</property>

</configuration>
~
```

The value of the access keys can be got from either the Azure portal, or we can get it from the command line.

```
vkjagan@MU-C02G9694MD6R 3.3.4 % az storage account show-connection-string --name
venkystorageaccount2704
{
  "connectionString":
"DefaultEndpointsProtocol=https;EndpointSuffix=core.windows.net;AccountName=venkystorageaccount27
04;AccountKey=xjy025msRRJYOfeWF7SPwJj2dzVgVLWuHgO0DZ1pKVo2TGpTsrOGEyU8NQ8TMEDNb9VDim9gKQi9+AStI0Y
Usg==;BlobEndpoint=https://venkystorageaccount2704.blob.core.windows.net/;FileEndpoint=https://ve
nkystorageaccount2704.file.core.windows.net/;QueueEndpoint=https://venkystorageaccount2704.queue.
core.windows.net/;TableEndpoint=https://venkystorageaccount2704.table.core.windows.net/"
}
```

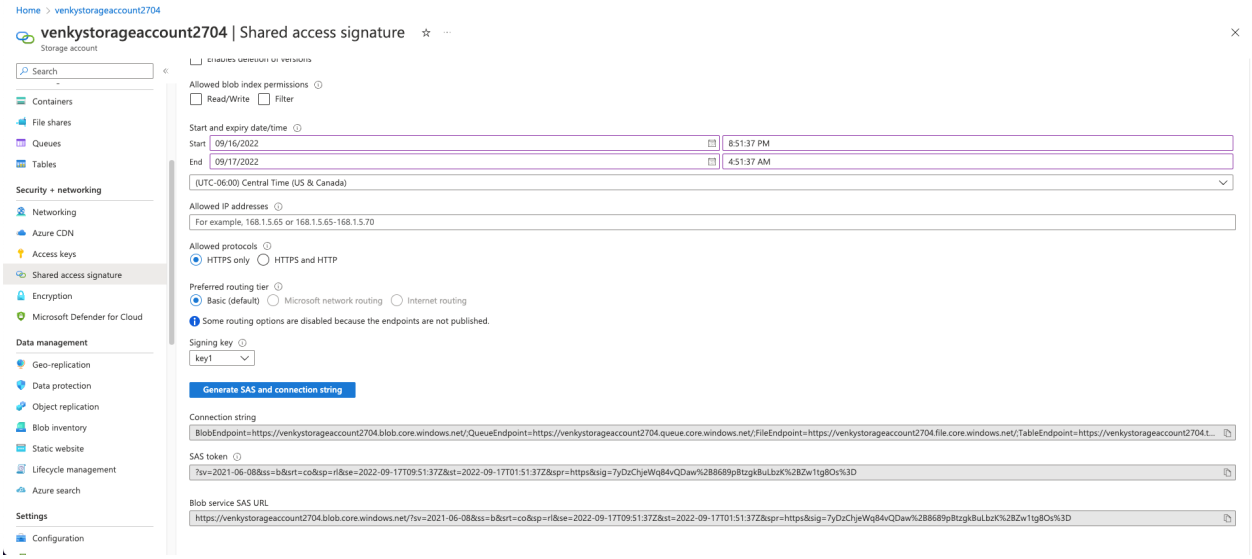


Once we make sure the put the access keys into the core-site.xml and save it in the correct location specified above, the Hadoop list command just works perfect!

```
vkjagan@MU-C02G9694MD6R 3.3.4 % ls
INSTALL_RECEIPT.json    LICENSE.txt             README.txt              core-site.xml           libexec
LICENSE-binary          NOTICE.txt              bin                     hadoop-env.sh           sbin
vkjagan@MU-C02G9694MD6R 3.3.4 % mv core-site.xml libexec/etc/hadoop
vkjagan@MU-C02G9694MD6R 3.3.4 % hadoop fs -ls wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
2022-09-16 20:48:39,235 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-09-16 20:48:39,521 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-09-16 20:48:39,589 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-09-16 20:48:39,589 INFO impl.MetricsSystemImpl: azure-file-system metrics system started
Found 1 items
-rwxrwxrwx   1         49 2022-09-16 20:12 wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/people.csv
2022-09-16 20:48:40,204 INFO impl.MetricsSystemImpl: Stopping azure-file-system metrics system...
2022-09-16 20:48:40,205 INFO impl.MetricsSystemImpl: azure-file-system metrics system stopped.
2022-09-16 20:48:40,205 INFO impl.MetricsSystemImpl: azure-file-system metrics system shutdown complete.
vkjagan@MU-C02G9694MD6R 3.3.4 %
```

Note that the core-site.xml has the ACCESS KEY in plain text! This is a serious violation of security principles.

Once way we can adjust this is to create a SAS string on the portal giving it exactly the access we want to give.



As we can see the SAS token is generated and the access is time-bound now.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

  <!-- <name>fs.azure.account.key.venkystorageaccount2704.blob.core.windows.net</name> -->
  <name>fs.azure.sas.venkycontainer.venkystorageaccount2704.blob.core.windows.net</name>

  <!-- <value>xjy025msRRJYOfeWF7SPwJj2dzVgVLWuHgO0DZ1pKVo2TGpTsrOGEyU8NQ8TMEDNb9VDim9gKQi9+AStI0YUsg==</value> -->
  <value>sv=2021-06-08&amp;ss=b&amp;srt=co&amp;sp=rl&amp;se=2022-09-17T09:51:37Z&amp;st=2022-09-17T01:51:37Z&amp;spr=https&amp;sig=7yDzChjeWq84vQDaw%2B8689pBtzgkBuLbzK%2BZw
1tg8Os%3D</value>
</property>

</configuration>
~
```

Note how the core-site.xml has changed. It has now used a new key for the <name> and the sas token value from the azure portal.

This should work right? NO!!!

```
vkjagan@MU-C02G9694MD6R hadoop % hadoop fs -ls wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
Exception in thread "main" [com.ctc.wstx.exc.WstxLazyException] com.ctc.wstx.exc.WstxUnexpectedCharException: Unexpected character '=' (code 61); expected a semi-colon afte
r the reference for entity 'ss'
 at [row,col,system-id]: [14,26,"file:/usr/local/Cellar/hadoop/3.3.4/libexec/etc/hadoop/core-site.xml"]
        at com.ctc.wstx.exc.WstxLazyException.throwLazily(WstxLazyException.java:40)
        at com.ctc.wstx.sr.StreamScanner.throwLazyError(StreamScanner.java:737)
        at com.ctc.wstx.sr.BasicStreamReader.safeFinishToken(BasicStreamReader.java:3745)
        at com.ctc.wstx.sr.BasicStreamReader.getTextCharacters(BasicStreamReader.java:914)
        at org.apache.hadoop.conf.Configuration$Parser.parseNext(Configuration.java:3403)
        at org.apache.hadoop.conf.Configuration$Parser.parse(Configuration.java:3182)
        at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:3075)
        at org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:3036)
        at org.apache.hadoop.conf.Configuration.loadProps(Configuration.java:2914)
        at org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2896)
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1412)
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1384)
        at org.apache.hadoop.conf.Configuration.setBoolean(Configuration.java:1726)
        at org.apache.hadoop.util.GenericOptionsParser.processGeneralOptions(GenericOptionsParser.java:340)
        at org.apache.hadoop.util.GenericOptionsParser.parseGeneralOptions(GenericOptionsParser.java:573)
        at org.apache.hadoop.util.GenericOptionsParser.<init>(GenericOptionsParser.java:175)
        at org.apache.hadoop.util.GenericOptionsParser.<init>(GenericOptionsParser.java:157)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:75)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:95)
        at org.apache.hadoop.fs.FsShell.main(FsShell.java:390)
Caused by: com.ctc.wstx.exc.WstxUnexpectedCharException: Unexpected character '=' (code 61); expected a semi-colon after the reference for entity 'ss'
 at [row,col,system-id]: [14,26,"file:/usr/local/Cellar/hadoop/3.3.4/libexec/etc/hadoop/core-site.xml"]
        at com.ctc.wstx.sr.StreamScanner.throwUnexpectedChar(StreamScanner.java:666)
        at com.ctc.wstx.sr.StreamScanner.parseEntityName(StreamScanner.java:2080)
        at com.ctc.wstx.sr.StreamScanner.fullyResolveEntity(StreamScanner.java:1538)
        at com.ctc.wstx.sr.BasicStreamReader.readTextSecondary(BasicStreamReader.java:4765)
        at com.ctc.wstx.sr.BasicStreamReader.finishToken(BasicStreamReader.java:3789)
        at com.ctc.wstx.sr.BasicStreamReader.safeFinishToken(BasicStreamReader.java:3743)
        ... 17 more
vkjagan@MU-C02G9694MD6R hadoop %
```

The characters we get from the portal are not compatible with XML format. So we need to go to this site https://www.freeformatter.com/xml-escape.html#before-output and fix the characters.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

  <!-- <name>fs.azure.account.key.venkystorageaccount2704.blob.core.windows.net</name> -->
  <name>fs.azure.sas.venkycontainer.venkystorageaccount2704.blob.core.windows.net</name>

  <!-- <value>xjy025msRRJYOfeWF7SPwJj2dzVgVLWuHgO0DZ1pKVo2TGpTsrOGEyU8NQ8TMEDNb9VDim9gKQi9+AStI0YUsg==</value> -->
  <value>sv=2021-06-08&amp;ss=b&amp;srt=co&amp;sp=rl&amp;se=2022-09-17T09:51:37Z&amp;st=2022-09-17T01:51:37Z&amp;spr=https&amp;sig=7yDzChjeWq84vQDaw%2B8689pBtzgkBuLbzK%2BZw
1tg8Os%3D</value>
</property>

</configuration>
~
```

With this escaping in place, we are able to access the data.

```
vkjagan@MU-C02G9694MD6R hadoop % hadoop fs -ls wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/
2022-09-16 21:18:09,276 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-09-16 21:18:09,617 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-09-16 21:18:09,718 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-09-16 21:18:09,718 INFO impl.MetricsSystemImpl: azure-file-system metrics system started
Found 1 items
-rwxrwxrwx   1         49 2022-09-16 20:12 wasbs://venkycontainer@venkystorageaccount2704.blob.core.windows.net/people.csv
2022-09-16 21:18:10,339 INFO impl.MetricsSystemImpl: Stopping azure-file-system metrics system...
2022-09-16 21:18:10,340 INFO impl.MetricsSystemImpl: azure-file-system metrics system stopped.
2022-09-16 21:18:10,340 INFO impl.MetricsSystemImpl: azure-file-system metrics system shutdown complete.
vkjagan@MU-C02G9694MD6R hadoop %
```

Next if we do not want to access these using oAuth, we need to follow this: https://hadoop.apache.org/docs/stable/hadoop-azure-datalake/index.html#Configuring_Credentials_and_FileSystem

We need to create an application registration into AAD.