GSM – Hadoop 1 Day Workshop

Welcome back!

Running Hadoop on Azure

Running Hadoop on Azure

HDInsight (Windows)

Managed Hadoop framework using Microsoft Azure instances and storage

Pay-as-you-go

Running Hadoop on Azure

Pre-requisite:

Microsoft Azure Subscription

Running Hadoop on Azure

# PRACTICE 1 - Sign in

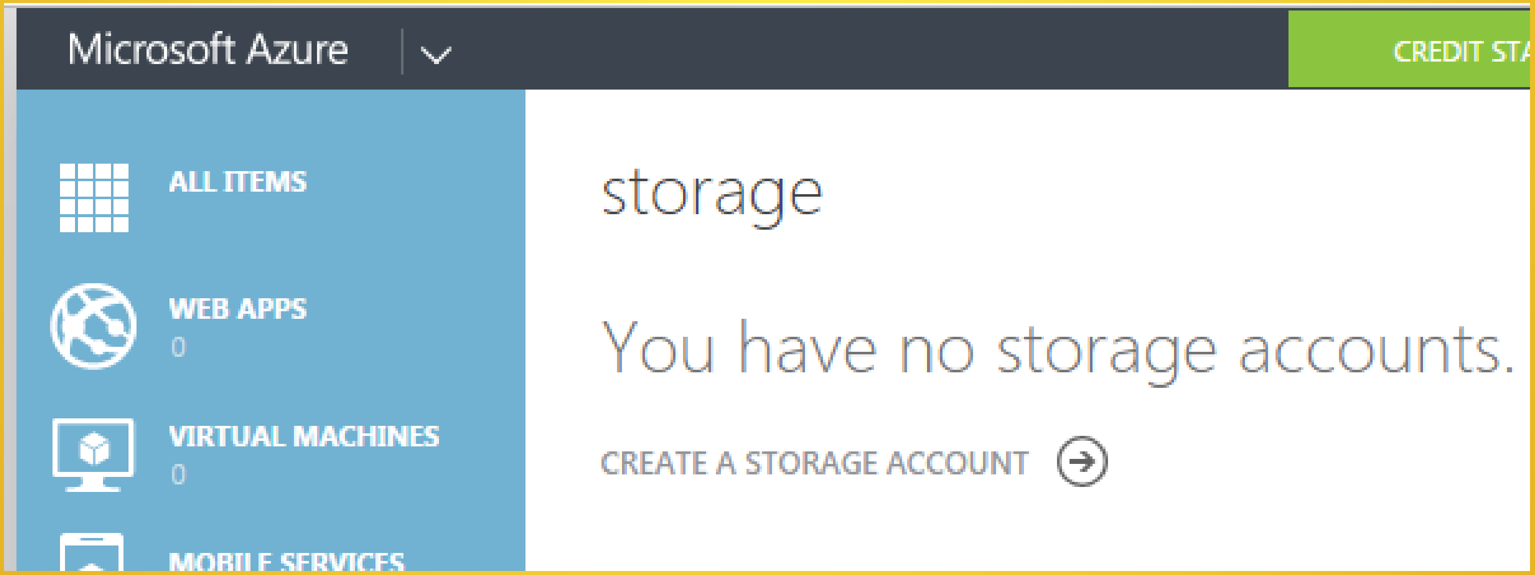
https://manage.windowsazure.com/

Sign in to the Portal

Running Hadoop on Azure

# PRACTICE 2 – Create a Storage Account

URL: gmshadoopworkshop
Location: East US
Replication: Locally Redundant

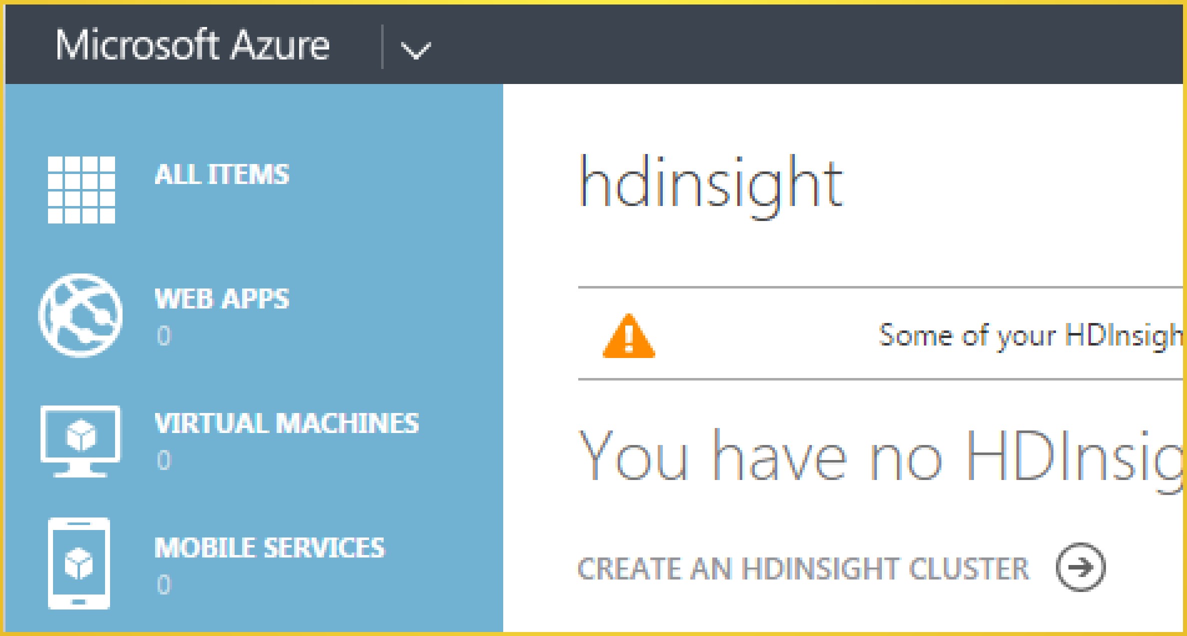


storage

You have no storage accounts.

Running Hadoop on Azure

# PRACTICE 3 - Create HDInsight Cluster
Cluster Name: Same as Storage Account
Cluster Size: 4 nodes
Password: (see rules)
Storage Account: (previously created)



hdinsight

Some of your HDInsigh

You have no HDInsig

Running Hadoop on Azure

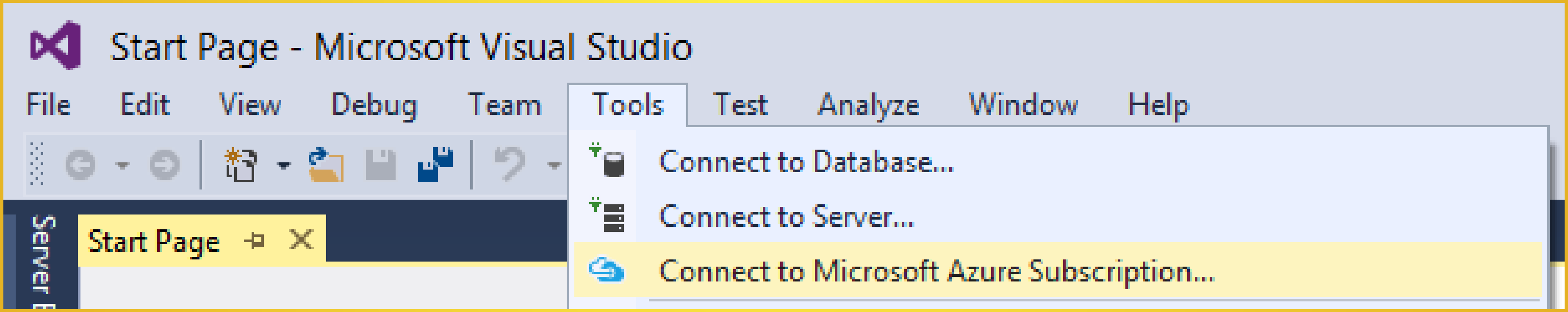
# PRACTICE 4 – Install Microsoft Tools

1. Visual Studio Community 2015 with Microsoft Azure SDK

Running Hadoop on Azure

# PRACTICE 5 – Connect to your subscription

Tools > Connect to Microsoft Azure Subscription



Running Hadoop on Azure

```
# PRACTICE 7 – Powershell Equivalent

> Add-AzureAccount
> Use-AzureHDInsightCluster dsikargmshadoopworkshop

# Check

> Get-AzureAccount
> Get-AzureSubscription
```

Running Hadoop on Azure

```
# PRACTICE 8 – Run Hive example

# Local
> Invoke-Hive "select country, state, count(*) as
records from hivesampletable group by country,
state order by records desc limit 5"
```

Running Hadoop on Azure

```
# PRACTICE 9 – Run Streaming example
# as per given example (error)

# Local
$jarFile = "/example/jars/hadoop-examples.jar"
$className = "wordcount"
$statusDirectory = "/samples/wordcount/status"
$outputDirectory = "/samples/wordcount/output"
$inputDirectory = "/example/data/gutenberg"
$wordCount = New-AzureHDInsightMapReduceJobDefinition \
-JarFile $jarFile -ClassName $className -Arguments \
$inputDirectory, $outputDirectory -StatusFolder $statusDirectory
```

Running Hadoop on Azure

# PRACTICE 10 – Debug Streaming example

# REMOTE

1. Test on shell
2. Write hadoop streamin command line equivalent

Running Hadoop on Azure

```
# PRACTICE 10 – Debug Streaming example
# solution (one of many)

> hadoop jar C:\apps\dist\hadoop-2.6.0.2.2.7.1-
34\share\hadoop\tools\lib\hadoop-streaming-
2.6.0.2.2.7.1-34.jar -input
/example/data/gutenberg -output
/example/data/output9 -mapper cat -reducer wc
-file cat.exe -file wc.exe
```

Running Hadoop on Azure

# PRACTICE 11 – Terminate Cluster and delete Storage