

## Contents

What is Apache Spark? .....	1
1) Create an Apache Spark Cluster in HDInsight in the East US region.....	1
2) Create a Jupyter Notebook in the Spark cluster .....	4
3) Import the following dataset into the Jupyter notebook using Spark SQL.....	5
4) References : .....	8
5) Download Azure StoreExplorer.....	8

## What is Apache Spark?

Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. Apache Spark in Azure HDInsight is the Microsoft implementation of Apache Spark in the cloud.

you can use HDInsight Spark clusters to process your data stored in Azure([Azure Blob storage](#), [Azure Data Lake Storage Gen1](#), or [Azure Data Lake Storage Gen2](#))

### 1) Create an Apache Spark Cluster in HDInsight in the East US region

Create an HDInsight cluster to process massive amounts of data using popular open-source frameworks such as Hadoop, Spark, Hive, LLAP, Kafka, Storm, ML Services, and more.

Microsoft Azure

Search resources, services, and docs (G+/f)

All services

Filter services

Overview

Categories

All

General

Compute

Networking

Storage

Web

Mobile

Containers

Databases

Analytics

Blockchain

AI + machine learning

Internet of things

ANALYTICS (15)

Azure Synapse Analytics

Azure Databricks

Data factories

Stream Analytics jobs

Analysis Services

Event Hubs Clusters

Data Lake Storage Gen1

Power Platform

Azure Synapse Analytics (private link hubs)

HDInsight clusters

Power BI Embedded

Data Lake Analytics

Event Hubs

Log Analytics workspaces

Azure Data Explorer Clusters

All services >

HDInsight clusters

Default Directory (vocareumvocareum.onmicrosoft.com)

Create

Manage view

Refresh

Export to CSV

Open query

Assign tags

Delete

Feedback

Filter for any field...

Subscription == all

Resource group == all

Location == all

Add filter

Showing 0 to 0 of 0 records.

No grouping

List view

Name	Cluster type	Status	Resource group	Location	Cluster Version
------	--------------	--------	----------------	----------	-----------------

All services > HDInsight clusters >

Create HDInsight cluster

Select the subscription to manage, specify the resource group, and then create or manage your resources.

Subscription \*

Production 1

Resource group \*

Regroup\_1ktn

Create new

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name \*

sathihdinsightcluster

Region \*

East US

Cluster type \*

Spark

Change

Version \*

Spark 2.4 (HDI 4.0)

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Review + create

« Previous

Next: Storage »

Select cluster type

Hadoop

Petabyte-scale processing with Hadoop components like MapReduce, Hive (SQL on Hadoop), Pig, Sqoop and Oozie.

Select

Spark

Fast data analytics and cluster computing using in-memory processing.

Select

Kafka

Build a high throughput, low-latency, real-time streaming platform using a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system.

Select

HBase

Fast and scalable NoSQL database. Available with both standard and premium (SSD) storage options.

Select

Interactive Query

Build Enterprise Data Warehouse with in-memory analytics using Hive (SQL on Hadoop) and LLAP (Low Latency Analytical Processing). Note that this feature requires high memory instances.

Select

Storm

Select

2 | Sathisha NM Week5

### Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Cluster login username *	<input type="text" value="admin"/>
Cluster login password *	<input type="password" value="*****"/> ✓
Confirm cluster login password *	<input type="password" value="*****"/> ✓
Secure Shell (SSH) username *	<input type="text" value="sshuser"/>
Use cluster login password for SSH	<input checked="" type="checkbox"/>

[Review + create](#)

[« Previous](#)

[Next: Storage »](#)

[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 *	<input type="text" value="Azure Storage"/>
Selection method *	<input checked="" type="radio"/> Select from list <input type="radio"/> Use access key
Primary storage account *	<input type="text" value="(New) sathihdinsighhdistorage"/> <a href="#">Create new</a>
Container *	<input type="text" value="sathihdinsightcluster-2021-07-30t13-13-30-250z"/> ✓

### 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.

[Review + create](#)

[« Previous](#)

[Next: Security + networking »](#)

Microsoft Azure

Search resources, services, and docs (G+/I)

All services > HDInsight clusters >

Create HDInsight cluster

Cluster type

Spark 2.4 (HDI 4.0)

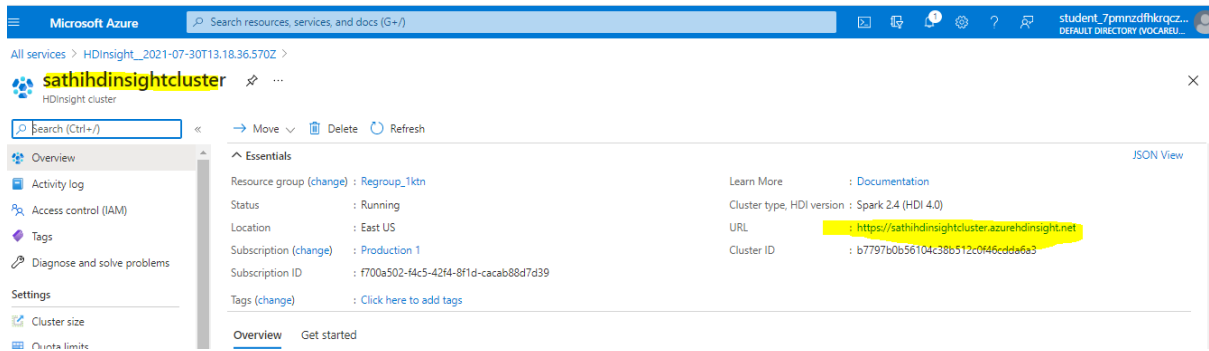
Cluster login username

admin

Submitting deployment...

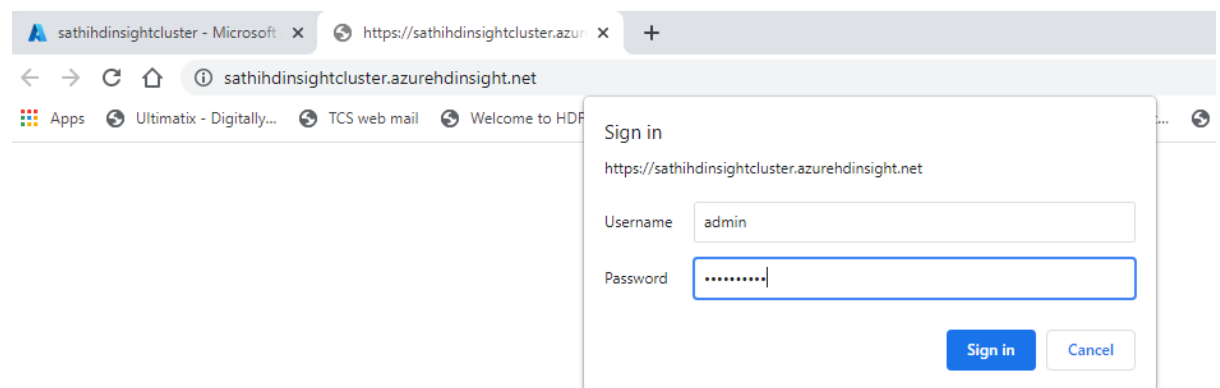
Submitting the deployment template for resource group 'Regroup\_1ktn'.

It takes about 20 minutes to create the cluster. The cluster must be created before you can proceed to the next session.

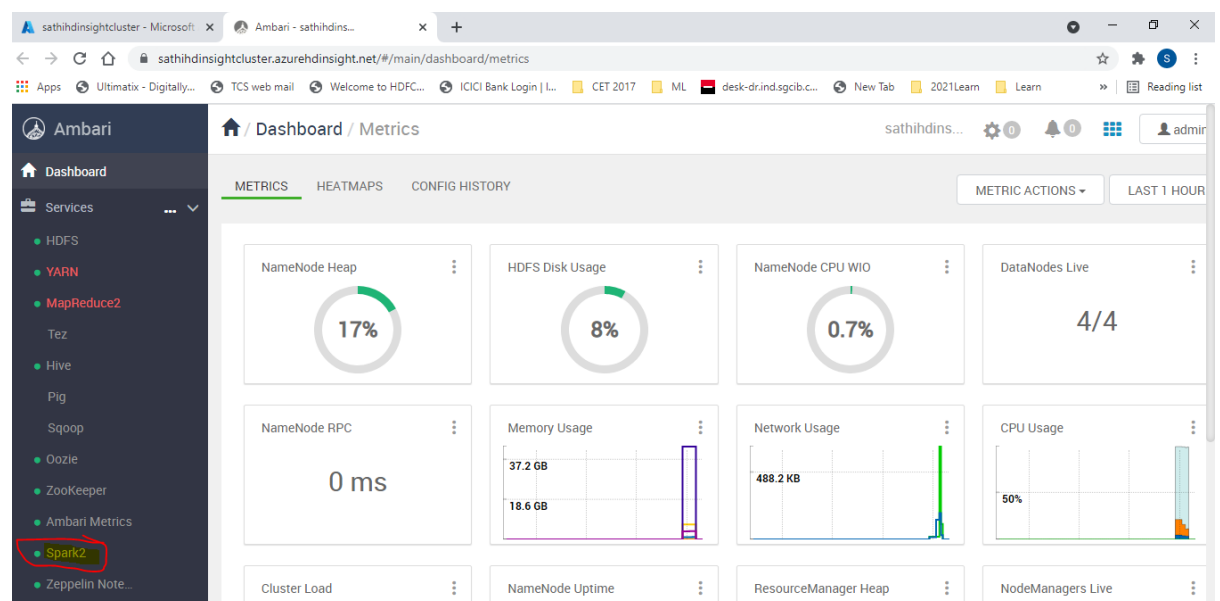


## 2) Create a Jupyter Notebook in the Spark cluster

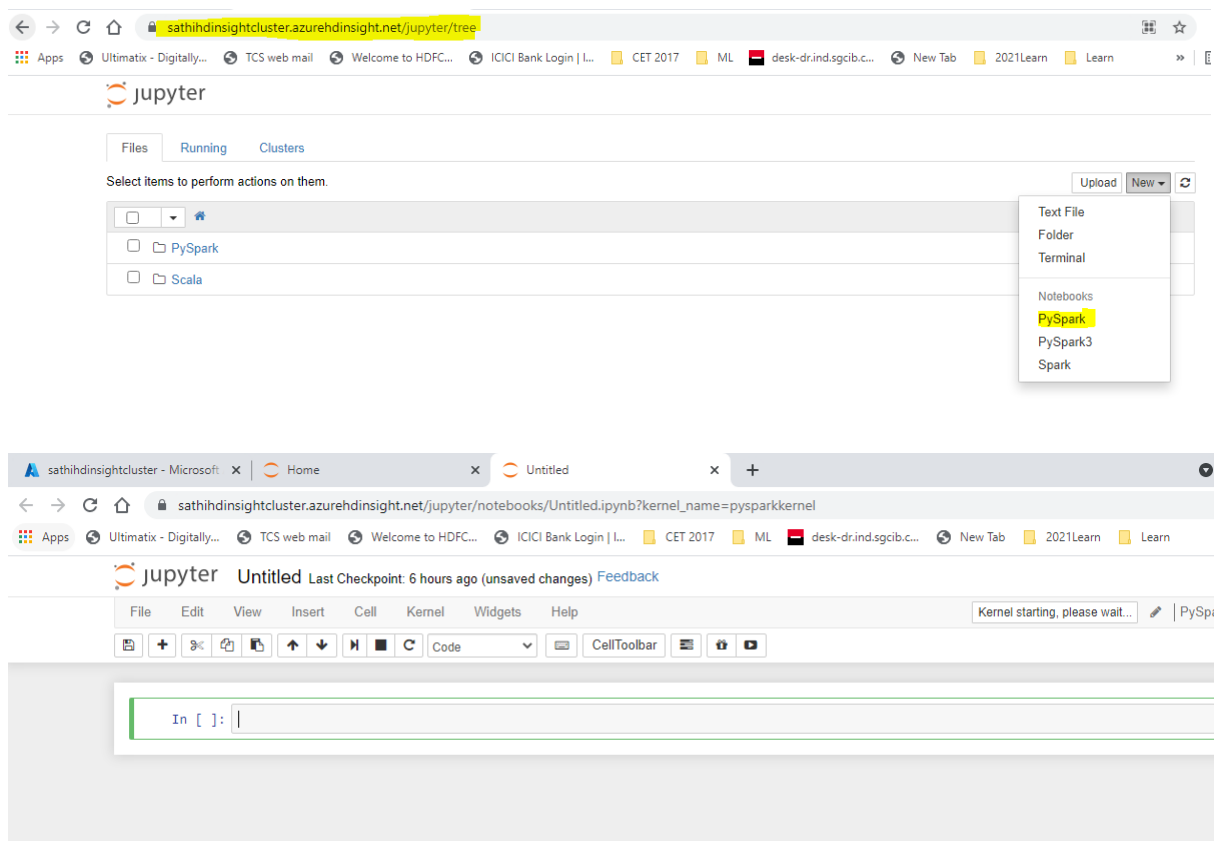
For Jupyter notebook web browser click on sathihdinsightcluster URL or navigate to <https://CLUSTERNAME.azurehdinsight.net/jupyter>, where CLUSTERNAME is the name of your cluster. If prompted, enter the cluster login credentials for the cluster



## Ambari View



<https://sathihdinsightcluster.azurehdinsight.net/jupyter>



## Run Apache Spark SQL statements and verify

In [1]: `%%sql`  
`SHOW TABLES`  
 Command finished at 07:30:2021 19:13:16.143 +05:30, execution took 31s 702ms

Jobs: 1 COMPLETED Spark: 6 EXECUTORS 15 CORES

Job ID	Job Name	Status	Stages	Tasks	Submission Time	Duration
0	runJob	COMPLETED	1/1		a few seconds ago	15s

Type: Table Pie Scatter Line Area Bar

database	tableName	isTemporary
default	hivesampletable	False

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1627651608052_0004	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

## 3) Import the following dataset into the Jupyter notebook using Spark SQL

[https://github.com/MicrosoftLearning/20775\\_Performing-Data-Engineering-on-Microsoft-HDInsight/blob/master/Allfiles/Demofiles/Mod04/clidata/hvac/HVAC.csv](https://github.com/MicrosoftLearning/20775_Performing-Data-Engineering-on-Microsoft-HDInsight/blob/master/Allfiles/Demofiles/Mod04/clidata/hvac/HVAC.csv)

```
from pyspark.sql import *
from pyspark.sql.types import *
```

```
In [1]: from pyspark.sql import *
from pyspark.sql.types import *
```

Command finished at 07-30-2021 20:07:12.563 +05:30, execution took 45ms

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
1	application_1627651608052_0005	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

SparkSession available as 'spark'.

/usr/bin/anaconda/lib/python2.7/site-packages/matplotlib/font\_manager.py:273: UserWarning: Matplotlib is building the font cache using fc-list. This may take a moment.  
warnings.warn('Matplotlib is building the font cache using fc-list. This may take a moment.')

```
In [2]: csvfile = spark.read.csv("/HdiSamples/HdiSamples/SensorSampleData/HVAC.csv")
csvfile.write.saveAsTable("hvac")
```

Command finished at 07-30-2021 20:07:47.780 +05:30, execution took 3s 286ms

u'Path does not exist: wasb://sathihdinsightcluster-2021-07-30t13-13-30-250z@sathihdinsighdistorage.blob.core.windows.net/HdiSamples/HdiSamples/SensorSampleData/HVAC.csv;'

Traceback (most recent call last):

File "/usr/hdp/current/spark2-client/python/pyspark/sql/readwriter.py", line 476, in csv  
return self.\_df(self.\_jreader.csv(self.\_spark.\_sc.\_jvm.PythonUtils.toSeq(path)))  
File "/usr/hdp/current/spark2-client/python/lib/py4j-0.10.7-src.zip/py4j/java\_gateway.py", line 1257, in \_\_call\_\_  
answer, self.gateway\_client, self.target\_id, self.name)  
File "/usr/hdp/current/spark2-client/python/pyspark/sql/utils.py", line 69, in deco

Navigate to sample document stored in azurehdinsightcluster storage

The screenshot shows the Azure portal interface for the storage account 'sathihdinsightcluster-2021-07-30t13-13-30-250z'. The breadcrumb navigation shows the path: Home > sathihdinsightcluster > sathihdinsighdistorage > sathihdinsightcluster-2021-07-30t13-13-30-250z. The main view displays the 'HdiSamples/HdiSamples/SensorSampleData/hvac' blob. The 'Overview' tab is selected, showing properties such as URL, LAST MODIFIED (7/30/2021, 7:01:09 PM), CREATION TIME (7/30/2021, 7:01:08 PM), VERSION ID, TYPE (Block blob), SIZE, ACCESS TIER (N/A), ACCESS TIER LAST MODIFIED (N/A), SERVER ENCRYPTED (true), and ETAG (0x8D9535E4A8ECB1F). A 'Copy to clipboard' button is visible next to the URL.

```
csvfile = spark.read.csv("/HdiSamples/HdiSamples/SensorSampleData/hvac/HVAC.csv")
ccsvfile.write.saveAsTable("hvac")
```

```
In [4]: csvfile = spark.read.csv("/HdiSamples/HdiSamples/SensorSampleData/hvac/HVAC.csv")
ccsvfile.write.saveAsTable("hvac")
```

Command finished at 07-30-2021 20:16:31.967 +05:30, execution took 21s 374ms

Job ID	Job Name	Status	Stages	Tasks	Submission Time	Duration
1	csv	COMPLETED	1/1		a minute ago	2s
2	saveAsTable	COMPLETED	1/1		a few seconds ago	4s

```
%%sql
SELECT * FROM hvac
```

2

saveAsTable

COMPLETED

1/1

9 minutes ago

4s

In [6]:

%%sql

SELECT \* FROM hvac

Command finished at 07-30-2021 20:24:29.154 +05:30, execution took 13s 574ms

Jobs: 1 COMPLETED Spark: 6 EXECUTORS 15 CORES

Job ID	Job Name	Status	Stages	Tasks	Submission Time	Duration
3	runJob	COMPLETED	1/1		a minute ago	10s

Type: Table Pie Scatter Line Area Bar

_c0	_c1	_c2	_c3	_c4	_c5	_c6
Date	Time	TargetTemp	ActualTemp	System	SystemAge	BuildingID
6/1/13	0:00:01	66	58	13	20	4
6/2/13	1:00:01	69	68	3	20	17
6/3/13	2:00:01	70	73	17	20	18
6/4/13	3:00:01	67	63	2	23	15

%%sql

SELECT \_C0 as date, (\_C2 - \_C3) AS temp\_diff FROM hvac WHERE \_C0 = \"6/1/13\"

In [13]:

%%sql

SELECT \_C0 as date, (\_C2 - \_C3) AS temp\_diff FROM hvac WHERE \_C0 = \"6/1/13\"

Command finished at 07-30-2021 20:37:46.175 +05:30, execution took 1s 332ms

Type: Table Pie Scatter Line Area Bar

2013-06-01	14.0
2013-06-01	6.0
2013-06-01	10.0
2013-06-01	4.0
2013-06-01	-7.0
2013-06-01	-2.0
2013-06-01	2.0
2013-06-01	5.0
2013-06-01	-8.0
2013-06-01	-10.0

267 rows x 2 columns



#### 4) References :

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-overview>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-jupyter-spark-sql-use-portal>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-use-with-data-lake-store>

<https://stackoverflow.com/questions/68508375/azure-hdinsight-sparksql-how-to-load-csv-file-from-github-in-dataframe>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-load-data-run-query>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-load-data-run-query>

[https://www.youtube.com/watch?v=pocW4ZHP\\_ng](https://www.youtube.com/watch?v=pocW4ZHP_ng)

#### 5) Dowload Azure StoreExplorer

<https://azure.microsoft.com/en-in/features/storage-explorer/#overview>