



**FP**Academy

# AZURE AS A BIGDATA PLATFORM

HDInsight

AUTOR:  
UTWORZONO:  
MODYFIKACJA :  
ODBIORCY:  
WERSJA:

**Sebastian Zęderowski**  
**28-12-2017**  
**28-12-2017**  
**Team FP**  
**1.0.0**

1. Tworzenia klastra HDInsight Spark z poziomu witryny Azure Portal oraz uruchomienie interaktywnego notetu Jupyter. .... 3
2. Tworzenia klastra HDInsight Spark za pomocą skryptów oraz uruchomienie zadania za pośrednictwem interfejsu Livy. 10

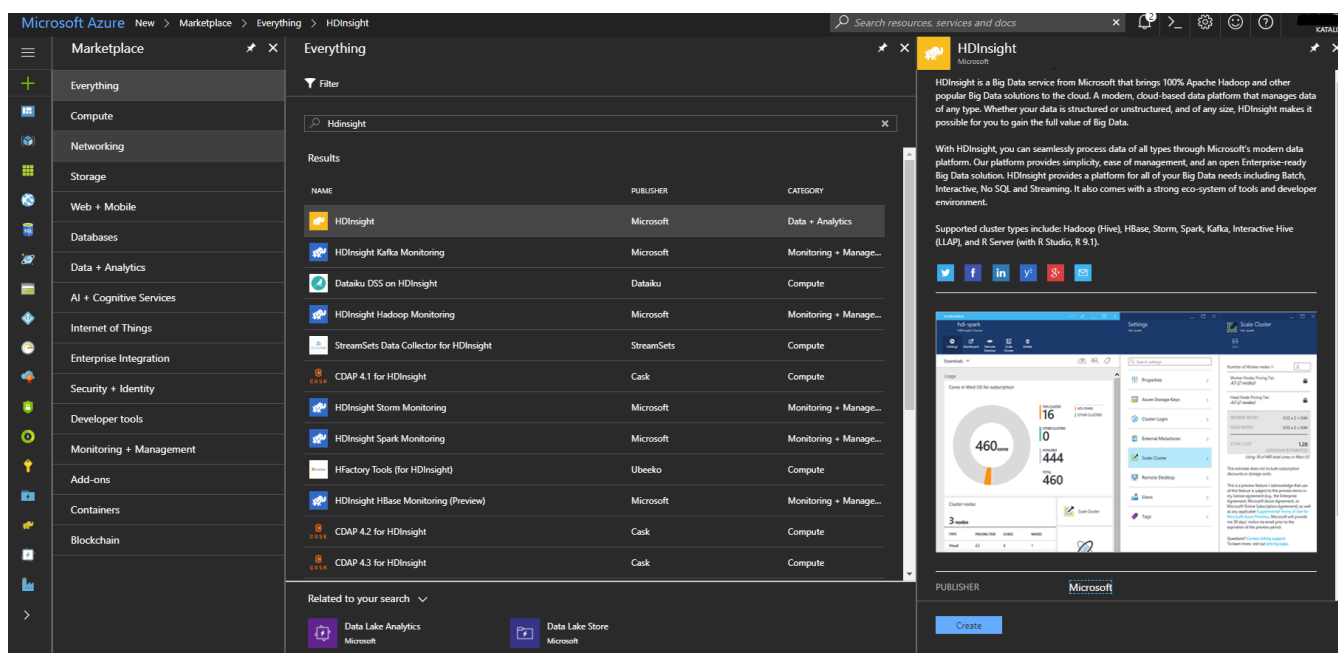
## 1. Wymagania

Posiadanie aktywnej subskrypcji Azure.

## 2. Tworzenia klastra HDInsight Spark z poziomu witryny Azure Portal oraz uruchomienie interaktywnego notetu Jupyter.

### 2.1 Zalogowanie się do portalu [Azure](#)

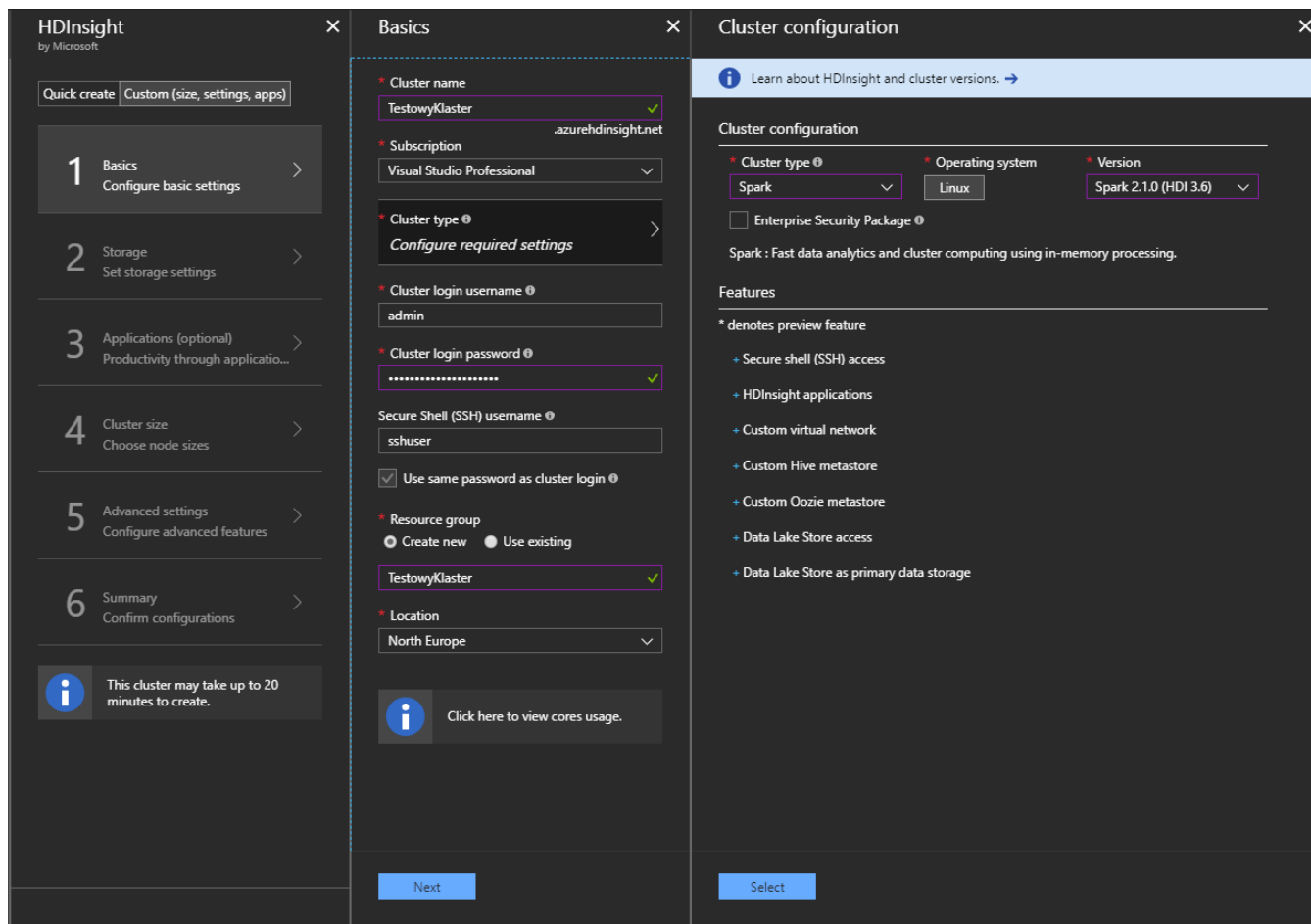
### 2.2 Utworzenie usługi HDInsight poprzez kliknięcie „Create a resource” i polu wyszukiwania wpisaniu HDInsight.



The screenshot shows the Microsoft Azure Marketplace interface. On the left, there's a navigation pane with categories like Compute, Networking, Storage, etc. The main area displays search results for 'HDInsight'. The results table lists various HDInsight-related services, including HDInsight, HDInsight Kafka Monitoring, Dataiku DSS on HDInsight, and others. The 'HDInsight' service is highlighted. To the right, there's a detailed view of the HDInsight service, including a description, supported cluster types, and a 'Create' button.

NAME	PUBLISHER	CATEGORY
HDInsight	Microsoft	Data + Analytics
HDInsight Kafka Monitoring	Microsoft	Monitoring + Manage...
Dataiku DSS on HDInsight	Dataiku	Compute
HDInsight Hadoop Monitoring	Microsoft	Monitoring + Manage...
StreamSets Data Collector for HDInsight	StreamSets	Compute
CDAP 4.1 for HDInsight	Cask	Compute
HDInsight Storm Monitoring	Microsoft	Monitoring + Manage...
HDInsight Spark Monitoring	Microsoft	Monitoring + Manage...
HFactory Tools (for HDInsight)	Ubeeko	Compute
HDInsight HBase Monitoring (Preview)	Microsoft	Monitoring + Manage...
CDAP 4.2 for HDInsight	Cask	Compute
CDAP 4.3 for HDInsight	Cask	Compute

2.3 Po kliknięciu „Create” uruchomi się wizzard, w pierwszym kroku wybieramy typ klastra oraz wprowadzamy dane takie jak nazwa klastra, użytkownik, hasło:



The screenshot displays the HDInsight cluster creation wizard with three panels: HDInsight by Microsoft, Basics, and Cluster configuration.

**HDInsight by Microsoft:**

- Buttons: Quick create, Custom (size, settings, apps)
- Progress steps:
  - 1 Basics: Configure basic settings
  - 2 Storage: Set storage settings
  - 3 Applications (optional): Productivity through application...
  - 4 Cluster size: Choose node sizes
  - 5 Advanced settings: Configure advanced features
  - 6 Summary: Confirm configurations
- Information: This cluster may take up to 20 minutes to create.

**Basics:**

- Cluster name: TestowyKlaster (with .azurehdinsight.net domain)
- Subscription: Visual Studio Professional
- Cluster type: Configure required settings
- Cluster login username: admin
- Cluster login password: (masked with dots)
- Secure Shell (SSH) username: sshuser
- Use same password as cluster login: ☒
- Resource group: Create new (selected), Use existing
  - TestowyKlaster
- Location: North Europe
- Buttons: Next, Click here to view cores usage.

**Cluster configuration:**

- Learn about HDInsight and cluster versions. →
- Cluster configuration:
  - Cluster type: Spark
  - Operating system: Linux
  - Version: Spark 2.1.0 (HDI 3.6)
  - Enterprise Security Package: ☐
  - Spark: Fast data analytics and cluster computing using in-memory processing.
- Features:
  - + Secure shell (SSH) access
  - + HDInsight applications
  - + Custom virtual network
  - + Custom Hive metastore
  - + Custom Oozie metastore
  - + Data Lake Store access
  - + Data Lake Store as primary data storage
- Buttons: Select

**2.4** W kolejnym kroku definiujemy *storage* dla naszego klastra, jako *primary storage* wybieramy *Azure Storage* (jeżeli nie mamy utworzonego *Azure Storage* możemy w tym kroku to zrobić). W tym kroku możemy również zdefiniować dodatkowe *Azure Storage* do których nasz klaster ma mieć dostęp oraz możemy zdefiniować dostęp do *Azure Data Lake Storage* (wymaga to utworzenia użytkownika w Azure AD oraz dodaniu mu uprawnień do naszego *Azure Data Lake Storage*).



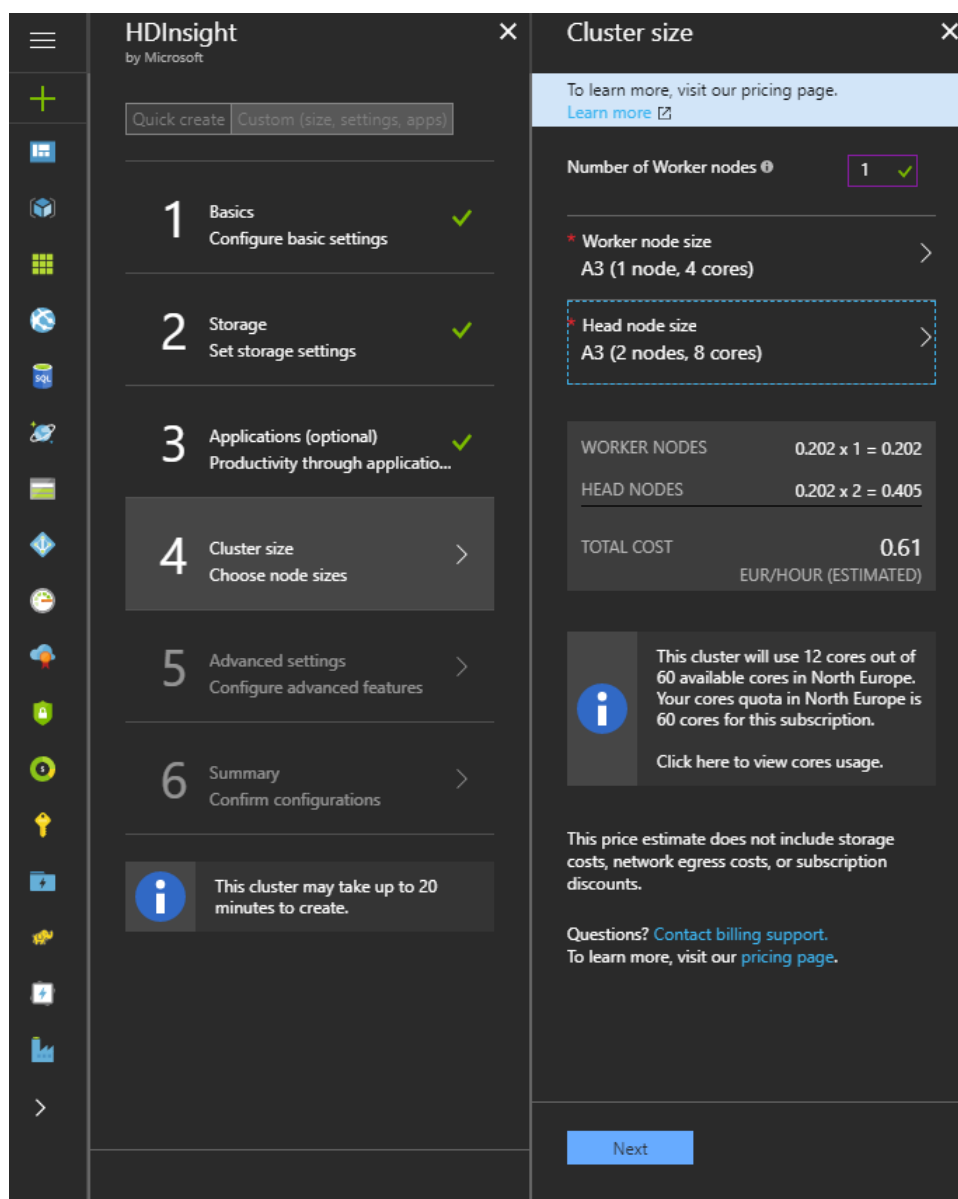
**2.5** W kroku 3 możemy wybrać dodatkowe aplikacje które będą zainstalowane.

2.6 W kolejnym kroku *Cluster Size* wybieramy rozmiar naszego klastra. Do przeprowadzenia ćwiczenia wystarczy najmniejszy klaster składający się z:

- 2x węzeł główny A3
- 1x węzeł roboczy A3

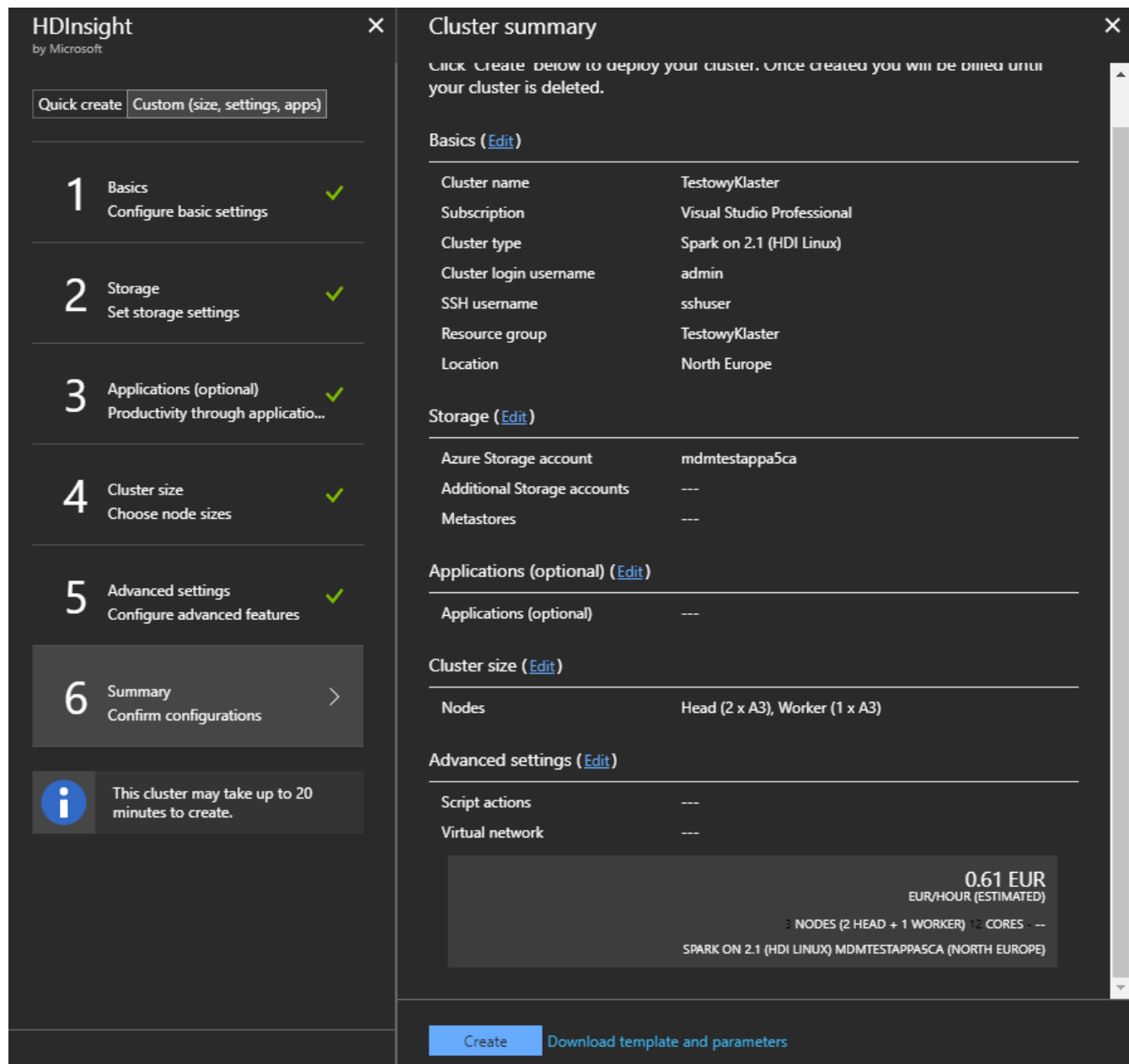
Domyślnie subskrypcje MSDN mają ograniczenie co do ilości rdzeni jakie możemy użyć w ramach klastrów HDInsight i wynosi ono 60, aby ścignąć powyższy limit należy się skontaktować z pomocą techniczną Azure.

[Limity subskrypcji i usługi Azure, przydziały i ograniczenia](#)



2.7 W następnym kroku możemy zdefiniować dodatkowe skrypty które mają zostać uruchomione oraz zdefiniować vnet. W celu przeprowadzenia ćwiczenia możemy pominąć ten krok.

2.8 Ostatni krok summary na którym widzimy podsumowanie informacji o naszym klastrze oraz możemy pobrać szablon tak zdefiniowanego klastra.



**HDInsight** by Microsoft

Quick create Custom (size, settings, apps)

- 1 Basics ✓  
Configure basic settings
- 2 Storage ✓  
Set storage settings
- 3 Applications (optional) ✓  
Productivity through applicatio...
- 4 Cluster size ✓  
Choose node sizes
- 5 Advanced settings ✓  
Configure advanced features
- 6 Summary >  
Confirm configurations

**Cluster summary**

Click **Create** below to deploy your cluster. Once created you will be billed until your cluster is deleted.

**Basics (Edit)**

Cluster name	TestowyKlaster
Subscription	Visual Studio Professional
Cluster type	Spark on 2.1 (HDI Linux)
Cluster login username	admin
SSH username	sshuser
Resource group	TestowyKlaster
Location	North Europe

**Storage (Edit)**

Azure Storage account	mdmtestappa5ca
Additional Storage accounts	---
Metastores	---

**Applications (optional) (Edit)**

Applications (optional)	---
-------------------------	-----

**Cluster size (Edit)**

Nodes	Head (2 x A3), Worker (1 x A3)
-------	--------------------------------

**Advanced settings (Edit)**

Script actions	---
Virtual network	---

**0.61 EUR**  
EUR/HOUR (ESTIMATED)

3 NODES (2 HEAD + 1 WORKER) 12 CORES ---

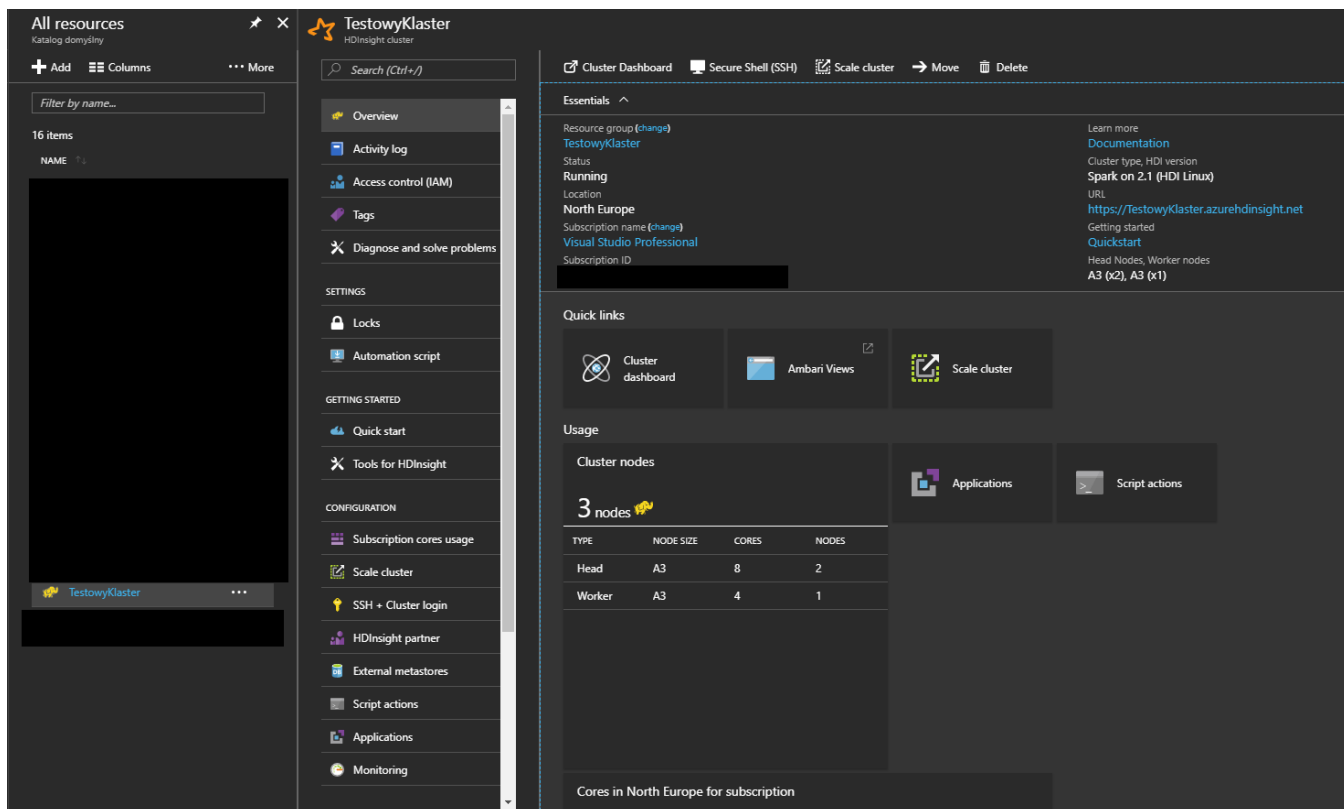
SPARK ON 2.1 (HDI LINUX) MDMTESTAPPASCA (NORTH EUROPE)

**Create** Download template and parameters

2.9 Klikamy *Create* i czekamy do 20 minut na utworzenie naszego klastra.



**2.10** Po utworzeniu klastra możemy uruchomić na nim pierwsze zadanie. W tym celu przechodzimy do naszego klastra w portalu Azure (np. z zakładki *All Resources*) i wybieramy z *Quick Links* pozycję *Cluster dashboard*.



The screenshot shows the Azure portal interface for an HDInsight cluster named 'TestowyKlastr'. The left sidebar contains navigation options like Overview, Activity log, Access control, and Settings. The main content area displays the 'Cluster Dashboard' with various metrics and links. The 'Usage' section includes a table for cluster nodes:

TYPE	NODE SIZE	CORES	NODES
Head	A3	8	2
Worker	A3	4	1

**2.11** Następnie wybieramy Jupyter Notebook, w oknie przeglądarki uruchomi się interaktywny notes jupyter:




The screenshot shows the Jupyter Notebook interface with the 'Running' tab selected. It displays a list of clusters: 'PySpark' and 'Scala'. The interface includes a search bar, a list of items, and buttons for 'Upload', 'New', and 'Refresh'.

**2.12** W katalogu PySpark znajdują się ćwiczenia w języku python, natomiast w katalogu Scala znajdują się ćwiczenia w języku scala.

**2.13** Po zakończeniu ćwiczeń należy usunąć klastrer aby nie zabierał nam środków.

### 3. Tworzenia klastra HDInsight Spark za pomocą skryptów oraz uruchomienie zadania za pośrednictwem interfejsu Livy.

#### 3.1 Wymagania do wykonania ćwiczenia.

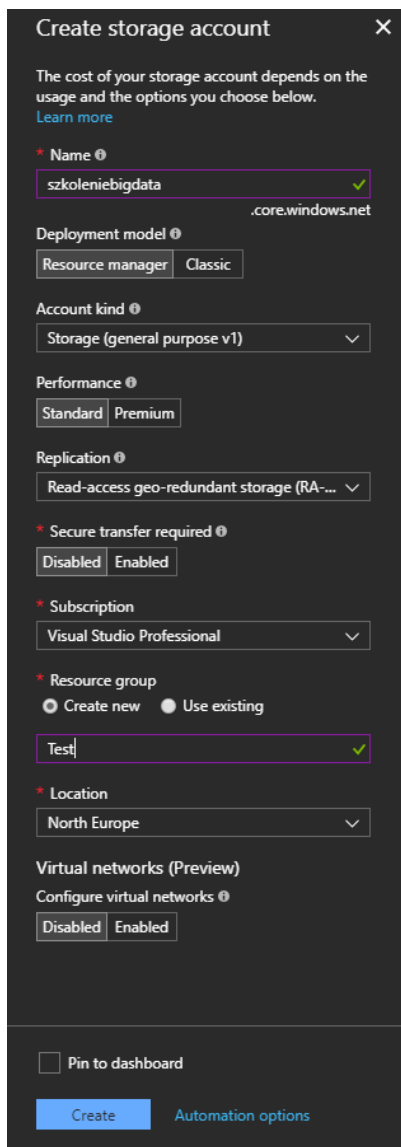
- Openssl (opcjonalnie do wygenerowania własnego certyfikatu)
- Azure Cmdlets Module [link](#)

#### 3.2 Do wykonanie ćwiczenia niezbędna jest utworzenie dwóch dodatkowych usług:

- Azure Storage Account
- Azure Data Lake Storage


#### 3.3 Tworzenie Azure Storage Account:


- Wybieramy „Create a resource” i wyszukiujemy usługi Storage account:





Create storage account


The cost of your storage account depends on the usage and the options you choose below. [Learn more](#)


\* Name   
 szkoleniebigdata ✓  
 .core.windows.net

Deployment model   
 Resource manager Classic

Account kind   
 Storage (general purpose v1) ▼

Performance   
 Standard Premium


Replication   
 Read-access geo-redundant storage (RA-... ▼

\* Secure transfer required   
 Disabled Enabled

\* Subscription  
 Visual Studio Professional ▼

\* Resource group  
☒ Create new ☐ Use existing  
 Test ✓

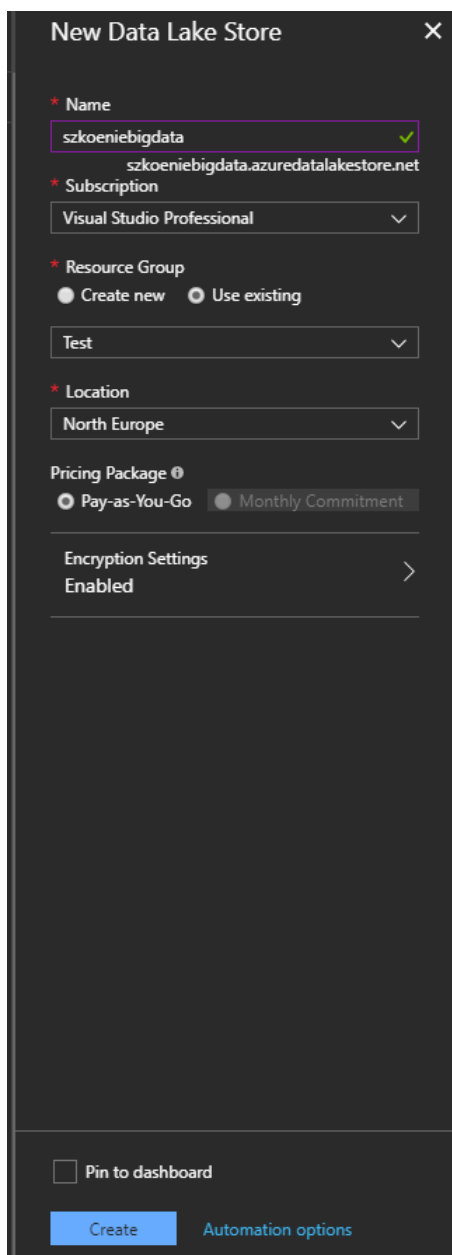
\* Location  
 North Europe ▼

Virtual networks (Preview)  
 Configure virtual networks   
 Disabled Enabled

☐ Pin to dashboard

Create Automation options

- Następnie wyszukiujemy i tworzymy usługę Data Lake Store:



- Aby nasz klaster miał dostęp do danych znajdujących się w Data Lake Store konieczne jest zdefiniowanie odpowiedniego użytkownika z odpowiednimi uprawnieniami. Dodatkowo do uwierzytelniania użytkownika musi być zastosowany certyfikat. W tym celu posłużymy się skryptyem scripts/CreateSecurityPrinciple.ps1:

- 

```
$scriptPath = split-path -parent $MyInvocation.MyCommand.Definition
$certificateFilePath = $scriptPath + "\mycert.pfx"
$password = "Szkolenie"

$certificatePFX = New-Object System.Security.Cryptography.X509Certificates.X509Certificate2($certificateFilePath,
$password)
$rawCertificateData = $certificatePFX.GetRawCertData()
```

```
$credential = [System.Convert]::ToBase64String($rawCertificateData)

Login-AzureRmAccount

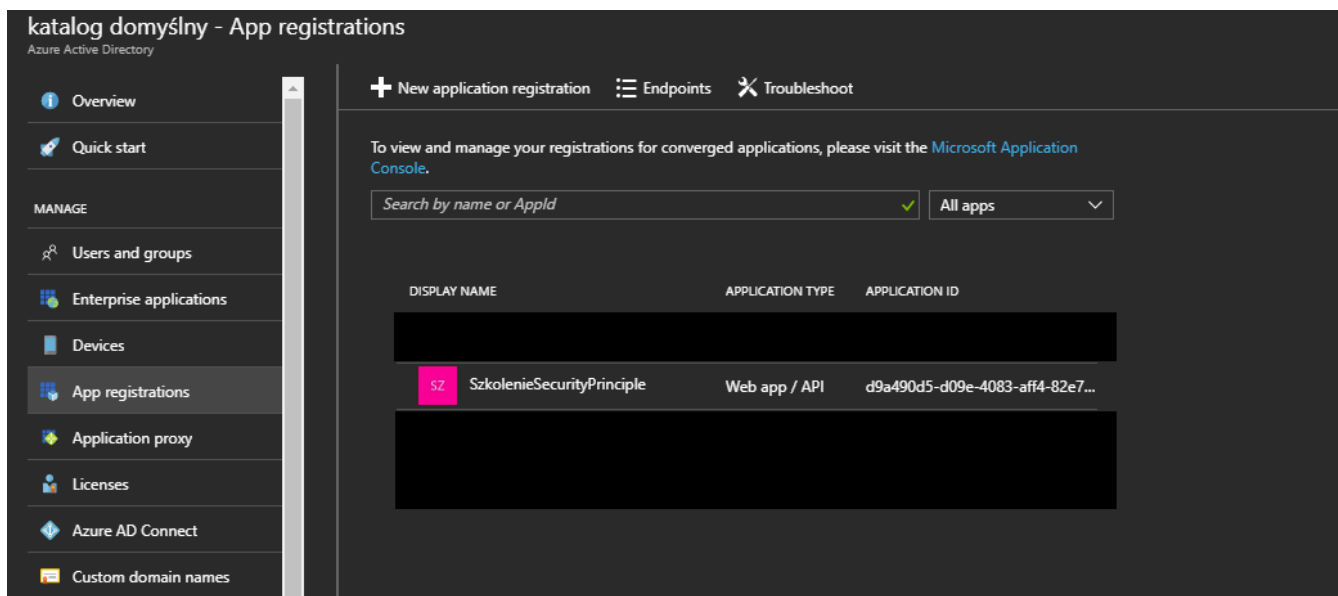
#Select-AzureRmSubscription -SubscriptionId "xxxxxxx-19a5-4f7a-8196-5f332c1bfaad"

$application = New-AzureRmADApplication `
  -DisplayName "SzkolenieSecurityPrinciple" `
  -HomePage "https://SzkolenieSecurityPrinciple" `
  -IdentifierUri "https://SzkolenieSecurityPrinciple" `
  -CertValue $credential `
  -StartDate $certificatePFX.NotBefore `
  -EndDate $certificatePFX.NotAfter

$applicationId = $application.ApplicationId

$servicePrincipal = New-AzureRmADServicePrincipal -ApplicationId $applicationId
$objectId = $servicePrincipal.Id
```

- Wywołanie polecenia `Login-AzureRmAccount` powoduje otwarcie okna logowania do portalu Azure. Jeżeli na naszym koncie jest więcej niż jedna subskrypcja możemy wybrać na której chcemy wykonać kolejne akcje za pomocą polecenia `Select-AzureRmSubscription`. Efektem wykonania powyższego skryptu jest zarejestrowanie aplikacji w AD uwierzytelnianej za pomocą certyfikatu.



katalog domyślny - App registrations

Azure Active Directory

Overview Quick start

MANAGE

- Users and groups
- Enterprise applications
- Devices
- App registrations
- Application proxy
- Licenses
- Azure AD Connect
- Custom domain names

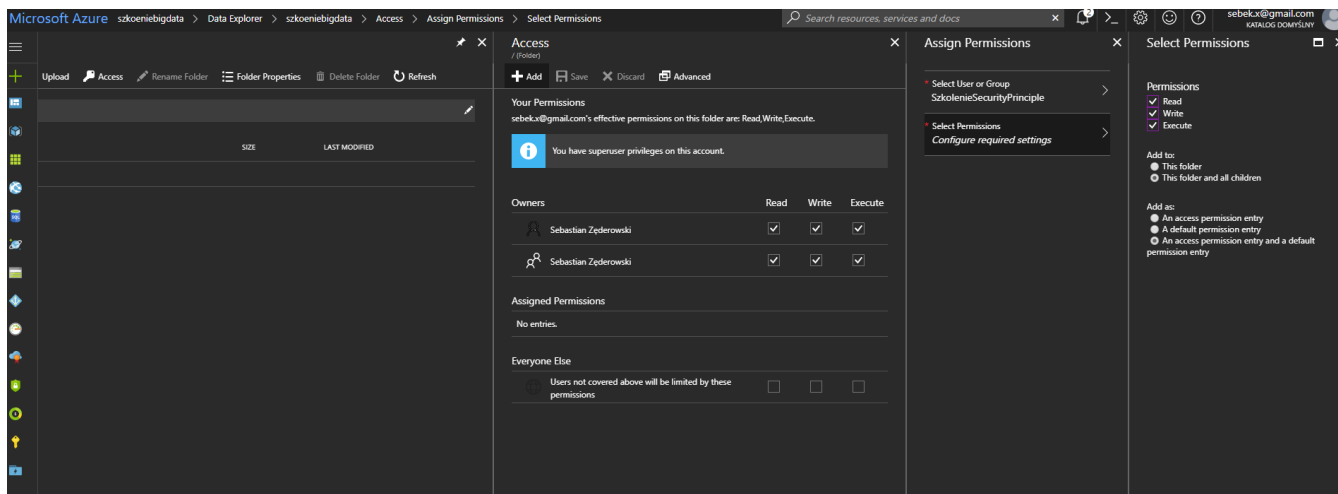
+ New application registration Endpoints Troubleshoot

To view and manage your registrations for converged applications, please visit the [Microsoft Application Console](#).

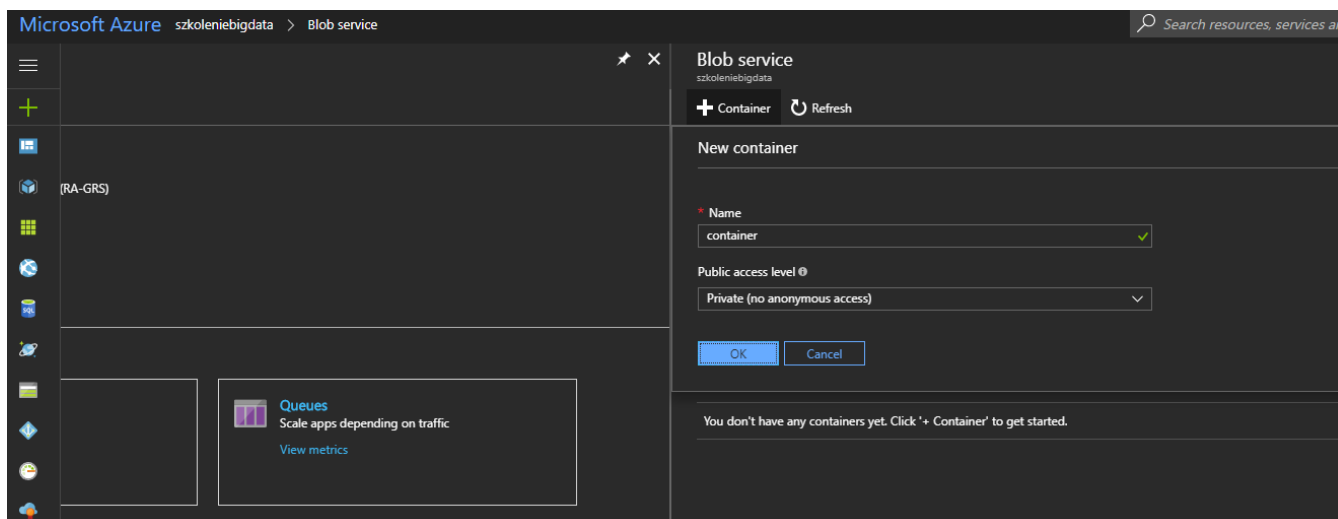
Search by name or Appld All apps

DISPLAY NAME	APPLICATION TYPE	APPLICATION ID
SzkolenieSecurityPrinciple	Web app / API	d9a490d5-d09e-4083-aff4-82e7...

- Kolejnym krokiem jest nadanie uprawnień do utworzonego wcześniej Data Lake Store w tym celu przechodzimy Data Lake Store na portalu azure a następnie otwieramy Data Explorer a następnie Access i nadajemy uprawnienia dla wcześniej utworzonego konta:



- Następnie przechodzimy do Storage Account utworzonego wcześniej i tworzymy Blob container:



- Następnym krokiem będzie przygotowanie danych, w tym celu umieścimy mały plik tekstowy z katalogu scripts davinci.txt we wcześniej utworzonym Data Lake Store. Najprościej można to zrobić za pośrednictwem portalu Azure. W tym celu otwieramy Data Lake Store i wybieramy Data Explorer a następnie Upload wybierając plik davinci.txt.
- Następnym etapem jest przygotowanie naszego zadania które będzie uruchomione na klastrze Spark. Zadanie będzie bardzo proste policzy wystąpienia wyrazów w pliku znajdującym się na Data Lake Store davinci.txt a następnie zapisze tam wyniki obliczeń. Kod powyższego zadania znajduje się w pliku TestJob.py należy go umieścić na wcześniej utworzonym Storage Account np. za pośrednictwem portalu Azure.
- Ostatnim etapem jest utworzenie klastra z uprawnieniami do Data Lake Store oraz uruchomienie Zadania TestJob.py umieszczonego na Azure Blob storage. W tym celu należy uruchomić skrypt scripts/CreateHDICluster.ps1. Skrypt ten podzielony jest na trzy części, w pierwszej tworzony jest klaster HDInsight Spark, w drugiej uruchamiane jest zadanie za pośrednictwem interfejsu LIVY polegające na uruchomieniu wcześniej utworzonego skryptu TestJob.py ostatnim krokiem skryptu jest usunięcie klastra po zakończeniu wykonywania zadania. Takie podejście jest często stosowane w celu optymalizacji kosztów.