

# Cluster Spark: Caso Microsoft Azure

Oscar Peredo A.  
Telefónica I+D Chile  
[oscar.peredo@telefonica.com](mailto:oscar.peredo@telefonica.com)

2 Junio 2016

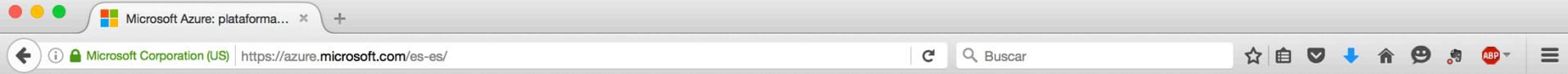


*ELEGIMOS TODO*

# Contenidos

- ¿Cómo levantar un cluster Spark?
  - Microsoft Azure: HDInsight (YARN) + Spark
- ¿Cómo veo la cola de trabajos en el cluster?
- ¿Cómo ejecuto trabajos Spark?

# ¿Cómo levantar un cluster Spark?



VENTAS 1-800-867-1389

MI CUENTA

PORTAL

Buscar



CUENTA GRATUITA >

Run your SAP HANA applications across dev/test and production >

## Introducción

Vea vídeos de tres minutos que le enseñarán a empezar a utilizar rápidamente Microsoft Azure.

Explorar más >

ELEGIMOS TODO\_



## Microsoft Azure

Por qué Azure    Productos    Documentación    Precios    Asociados    Blog    Recursos    Soporte técnico

# Cree su cuenta gratuita de Azure hoy mismo

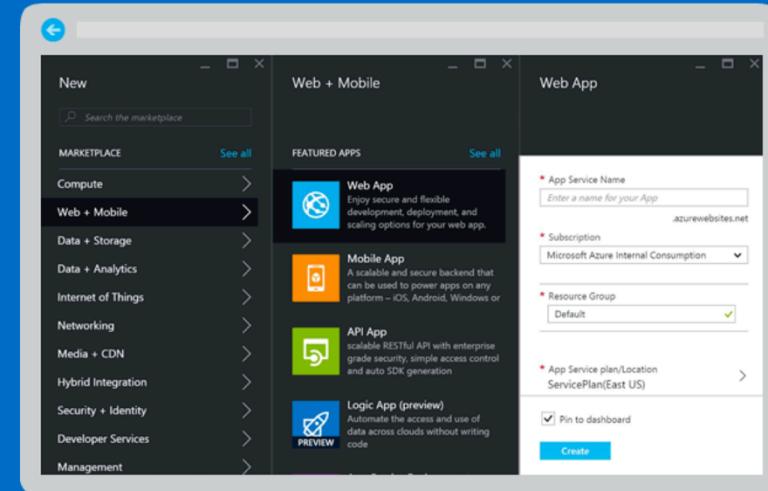
Comience con un crédito de €170 y continúe con opciones gratuitas

Empezar ahora >

O compre ahora ►

Preguntas más frecuentes ►

¿Tiene más preguntas? Call us 1-800-867-1389



Use su crédito de €170 y...

- ✓ Aprovisione hasta 14 máquinas virtuales, 40 bases de datos SQL u 8 TB de almacenamiento para un mes.
- ✓ Cree aplicaciones web, móviles y de API que usen los servicios Caché en Redis, Búsqueda o Red de entrega de contenido
- ✓ ...





## Iniciar sesión

Cuenta Microsoft [¿Qué es esto?](#)

Correo electrónico o teléfono

Contraseña

Mantener la sesión iniciada

**Iniciar sesión**

[¿No puedes acceder a tu cuenta?](#)

[Inicia sesión con un código de un solo uso](#)

[¿No dispones de una cuenta Microsoft?](#)

[Regístrate ahora](#)

Microsoft

[Condiciones de uso](#) [Privacidad y cookies](#) ©2016 Microsoft

ELEGIMOS TODO\_

## Versión de evaluación de un mes

\$200 Crédito de Azure

Sin ningún compromiso: la versión de evaluación no se actualiza automáticamente a una suscripción de pago

[Preguntas más frecuentes ▶](#)



1

### (-) Acerca de usted

\* País o región i

\* Nombre

\* Apellidos

\* Dirección de contacto i

\* Teléfono del trabajo

Empresa/escuela

[Siguiente](#)

2

### (+) Verificación de identidad mediante teléfono

i

3

### (+) Verificación de identidad mediante tarjeta

i

4

### (+) Acuerdo

[Suscribirse ➔](#)

Panel - Microsoft Azure

https://portal.azure.com/#

Microsoft Azure

Panel ▾ + Nuevo panel ⚙ Editar panel 🛡 Compartir ↗ Pantalla completa 🕒 Clonar 🗑 Eliminar

Grupos de recursos

Todos los recursos

Recientes

Servicios de aplicaciones

Máquinas virtuales (cl...)

Máquinas virtuales

Bases de datos SQL

Servicios en la nube (cl...)

Security Center

Suscripciones

Examinar >

Marketplace

Ayuda y soporte técnico

ELEGIMOS TODO\_



## Microsoft Azure

Nuevos &gt; Datos y análisis



## Nuevo

Grupos de recursos

Todos los recursos

Recientes

Servicios de aplicaciones

Máquinas virtuales (cl...

Máquinas virtuales

Bases de datos SQL

Servicios en la nube (cl...

Security Center

Suscripciones

Examinar &gt;

## Nuevos

Buscar en el Marketplace

## MARKETPLACE Ver todo

Máquinas virtuales

Web y móvil

Datos y almacenamiento

Datos y análisis

Internet de las cosas

Redes

Medios + CDN

Integración híbrida

Seguridad e identidad

Servicios de desarrollo

Administración

Intelligence

Contenedores

## RECIENTES

No creó nada recientemente.

## Datos y análisis

## APLICACIONES DESTACADAS Ver todo



## Power BI Embedded (versión preliminar)

Embed fully interactive, stunning data visualizations in your



## Cognitive Services APIs (versión preliminar)

Microsoft Cognitive Services lets you build apps with powerful algorithms



## Catálogo de datos

Detección de orígenes de datos para obtener más valor a partir de recursos de datos de empresa



## HDInsight

Servicio de Big Data basado en la nube de Microsoft. Apache Hadoop y otras soluciones populares de Big



## Data Lake Analytics (versión preliminar)

Big data analytics made easy



## Machine Learning

Build, deploy and share advanced analytics solutions



## Factoría de datos

Transforme datos en información de confianza



## Event Hub

Cloud-scale telemetry ingestion from websites, apps, and devices



## Trabajo de Análisis de



## Microsoft Azure

Nuevos &gt; Datos y análisis &gt; Nuevo clúster de HDInsight



Nuevo

Grupos de recursos

Todos los recursos

Recientes

Servicios de aplicaciones

Máquinas virtuales (cl...)

Máquinas virtuales

Bases de datos SQL

Servicios en la nube (cl...

Security Center

Suscripciones

Examinar &gt;

## Nuevo clúster de HDInsight

\* Nombre del clúster

Escribir nuevo nombre de clúster

.azurehdinsight.net

\* Suscripción

Select Cluster Type ⓘ

Configurar los valores obligatorios



\* Credenciales

Configurar los valores obligatorios



\* Origen de datos ⓘ

Configurar los valores obligatorios



\* Precios

Establezca la configuración neces...



Configuración opcional



\* Grupo de recursos ⓘ

 Nuevo  Use existing

Este clúster puede tardar hasta 20 minutos en crearse.



Corrija los errores de esta página antes de continuar.

 Anclar al panel

Crear



## Microsoft Azure

Nuevos &gt; Datos y análisis &gt; Nuevo clúster de HDInsight &gt; Cluster Type configuration



Nuevo

Grupos de recursos

Todos los recursos

Recientes

Servicios de aplicaciones

Máquinas virtuales (clá...)

Máquinas virtuales

Bases de datos SQL

Servicios en la nube (cl...

Security Center

Suscripciones

Examinar &gt;

## Nuevo clúster de HDInsight

\* Nombre del clúster  
ejemplo-spark-tid.azurehdinsight.net

\* Suscripción

Select Cluster Type ⓘ  
Configurar los valores obligatorios !>

\* Credenciales  
Configurar los valores obligatorios

\* Origen de datos ⓘ  
Configurar los valores obligatorios

\* Precios  
Establezca la configuración neces...

Configuración opcional

\* Grupo de recursos ⓘ  
 Nuevo  Use existing

Este clúster puede tardar hasta 20 minutos en crearse.

Corrija los errores de esta página antes de continuar.

Anclar al panel

Crear

## Cluster Type configuration

Learn about HDInsight and cluster versions. [Más información](#)

Tipo de clúster ⓘ  
Spark (Vista previa)

Sistema operativo  
Linux

Versión  
Spark 1.6.1 (HDI 3.4)

Cluster Tier ([more info](#))

STANDARD
Administration
Manage, monitor, connect
Scalability
On-demand node scaling
99.9% Uptime SLA
Automatic patching
<b>+ 0.00</b>
USD/CORE/HOUR

PREMIUM (VISTA PRE...★
Administration
Manage, monitor, connect
Scalability
On-demand node scaling
99.9% Uptime SLA
Automatic patching
Microsoft R Server for HDInsight
<b>+ 0.02</b>
USD/CORE/HOUR

Seleccionar

https://portal.azure.com/#create/Microsoft.HDInsightCluster

## Microsoft Azure

Nuevo clúster de HDInsight

Origen de datos

El clúster usará el orig. de datos como ubic. ppal. para acceder a datos, como la entr. de trabajos y salid. de reg.

Método de selección: Desde todas las suscripciones

\* Selección de cuenta de almacenamiento: cursostid (Centro-Norte de EE. UU.)

Nuevo

\* Elegir contenedor predeterminado: ejemplo-tid-spark

\* Ubicación: Centro-Norte de EE. UU.

Identidad de AAD del clúster: No configurado

Cuentas de almacenamiento

Opciones de suscripción:

Search to filter storage accounts ...

Nombre	Ubicación
cursostid	eastus
cursostid	northcentralus
	eastus
	southcentralus
	eastus
	eastus
	eastus

Este clúster puede tardar hasta 20 minutos en crearse.

Anclar al panel

Crear

Seleccionar

## Microsoft Azure

cursos-tid &gt; Todo &gt; HDInsight &gt; Nuevo clúster de HDInsight &gt; Precios &gt; Elegir el tamaño del nodo

Nuevo

Grupos de recursos

- Todos los recursos
- Recientes
- Servicios de aplicaciones
- Máquinas virtuales (clá...)
- Máquinas virtuales
- Bases de datos SQL
- Servicios en la nube (cl...)
- Security Center
- Suscripciones
- Examinar >

## Nuevo clúster de HDInsight

\* Nombre del clúster  
ejemplo-tid-spark ✓

\* Suscripción

\* Select Cluster Type i  
Estándar Spark en Linux (3.4) >

\* Credenciales  
Configurado >

\* Origen de datos i  
cursostid (Centro-Norte de EE. UU.) >

\* Precios  
Establezca la configuración necesaria... >

Configuración opcional >

\* Grupo de recursos i  
 Nuevo  Use existing  
cursos-tid >

i Este clúster puede tardar hasta 20 minutos en crearse.

Anclar al panel

Crear

## Precios

Para obtener más información, visite nuestra página de precios. [Más información](#)

Número de nodos de Trabajador i 4 ✓

\* Tamaño del nodo Trabajador  
D4 v2 (4 nodos, 32 núcleos) >

\* Tamaño del nodo Encabezado  
D4 v2 (2 nodos, 16 núcleos) >

TRABAJADOR NODOS  $1.24 \times 4 = 4.97$

ENCABEZADO NODOS  $1.24 \times 2 = 2.49$

COSTE TOTAL **7.46**  
USD/HORA (ESTIMADO)

Se usarán 48 de 60 núcleos en Centro-Norte de EE. UU.

Este presupuesto no incluye descuentos de suscripción, costos de almacenamiento ni costos de salida de red.

Esta es una característica de versión preliminar. Reconozco que el uso de esta característica está sujeto a los términos de versión preliminar de mi contrato de licencia (p. ej., el Contrato Enterprise, el Contrato de Microsoft Azure o el Contrato Microsoft Online Subscription), así como a cualquier [Término de uso complementario de las versiones preliminares de Microsoft Azure](#) aplicable. Microsoft me enviará un aviso por correo electrónico con 30 días de antelación antes de la expiración del período de versión preliminar.

¿Tiene preguntas? Póngase en contacto con el [soporte de facturación](#).

## Elegir el tamaño del nodo

Examine los tamaños disponibles del nodo y sus características. [Más información](#)

Los precios presentados a continuación son precios minoristas estimados que incluyen la infraestructura de Azure y costes de software de terceros aplicables.

★ Recomendado Ver todo

A3 Uso general	A4 Uso general	A6 Uso general
4 Núcleos	8 Núcleos	4 Núcleos
7 GB de RAM	14 GB de RAM	28 GB de RAM
8 Discos	16 Discos	8 Discos
<b>0,32</b> USD/HORA (ESTIMADO)	<b>0,64</b> USD/HORA (ESTIMADO)	<b>0,71</b> USD/HORA (ESTIMADO)
A7 Uso general	D3 Optimizado	D4 Optimizado
8 Núcleos	4 Núcleos	8 Núcleos
56 GB de RAM	14 GB de RAM	28 GB de RAM
16 Discos	8 Discos	16 Discos
<b>1,41</b> USD/HORA (ESTIMADO)	<b>0,62</b> USD/HORA (ESTIMADO)	<b>1,24</b> USD/HORA (ESTIMADO)
<span style="color: blue;">Seleccionar</span>	<span style="color: blue;">Seleccionar</span>	<span style="color: blue;">Seleccionar</span>

20 minutos después...

*ELEGIMOS TODO\_*

https://portal.azure.com/#resource/subscriptions/

## Microsoft Azure

ejemplo-tid-spark > Configuración

Nuevo

Grupos de recursos

Todos los recursos

Recientes

Servicios de aplicaciones

Máquinas virtuales (cl...)

Máquinas virtuales

Bases de datos SQL

Servicios en la nube (cl...)

Security Center

Suscripciones

Examinar >

### ejemplo-tid-spark

Clúster de HDInsight

Configur... Panel Shell seguro Escalar clúster Eliminar

#### Información esencial

Grupo de recursos: cursos-tid  
Estado: Ejecutando  
Ubicación: North Central US  
Nombre de la suscripción:   
Id. de suscripción:

Tienda de clúster: **estándar**  
URL: <https://ejemplo-tid-spark.azurehdinsight.net>  
Más información

Introducción: [Inicio rápido](#)  
Nodos Encabezado, Nodos Trabajador: A6 (x2), A6 (x4)

Agregar iconos:

#### Vínculos rápidos

Panels de clúster:  Vistas de Ambari:  Escalar clúster:

Agregar iconos:

#### Uso

Núcleos de North Central US para la suscripción

ESTE CLÚSTER: 24  
60 núcleos  
OTROS CLÚSTERES: 0

Nodos de clúster: 6 nodos

TIPO	TAMAÑO DEL ...	NÚCLEOS	NODOS
Encabezado	A6	8	2
Trabajador	A6	16	4

Configuración

ejemplo-tid-spark

Configuración del filtro

#### SOPORTE TÉCNICO Y SOLUCIÓN DE PROBLEMAS

Registros de auditoría >  
Nueva solicitud de soporte técnico >

#### INTRODUCCIÓN

Inicio rápido >  
Herramientas para HDInsight >

#### CONFIGURACIÓN

Inicio de sesión de clúster >  
Escalar clúster >  
Shell seguro >  
Asociado de HDInsight >  
Tiendas de metadatos externas >

#### PROPIEDADES

Propiedades >  
Claves de almacenamiento de Azure >  
Identidad de AAD del clúster >

#### ADMINISTRACIÓN DE RECURSOS

Etiquetas >  
Bloques >  
Usuarios >

Agregado por el usuario:

# Ambari (admin GUI)

Ambari - ejemplo-ti... + New Tab

https://ejemplo-tid-spark.azurehdinsight.net/#main/dashboard/metrics

Buscar

Dashboard Services Hosts Alerts Admin grid icon admin

Ambari ejemplo-ti... 0 ops 0 alerts

Metrics Heatmaps Config History

Metric Actions Last 1 hour

**HDFS Disk Usage** 8%

**DataNodes Live** 4/4

**HDFS Links**  
Active NameNode  
Standby NameNode  
4 DataNodes  
More...

**Memory Usage** 18.6 GB

**Network Usage** 488.2 KB

**CPU Usage** 50%

**Cluster Load**

**NameNode Heap** 5%

**NameNode RPC** 0.57 ms

**NameNode CPU WIO** n/a

**NameNode Uptime** 658.4 s

**ResourceManager Heap** 21%

**ResourceManager Uptime** 789.9 s

**NodeManagers Live** 4/4

**YARN Memory** 3%

**YARN Links**  
ResourceManager  
4 NodeManagers  
More...

Actions

¿Cómo veo la cola de  
trabajos?

# Primero, crear túnel ssh..

```
ssh -C2qTnNf -D 9876 profesor@ejemplo-tid-spark-ssh.azurehdinsight.net
```

# Después, usar proxy en browser...

The screenshot shows the Firefox preferences window on a Mac OS X system. The sidebar on the left has icons for General, Buscar, Contenido, Aplicaciones, Privacidad, Seguridad, Sync, and Avanzado, with 'Avanzado' selected. The main pane displays the 'Avanzado' settings. A modal dialog box titled 'Configurar proxies para el acceso a Internet' is open over the preferences window. The dialog contains several options for proxy configuration:

- Sin proxy
- Autodetectar configuración del proxy para esta red
- Usar la configuración del proxy del sistema
- Configuración manual del proxy:
  - Proxy HTTP: [ ] Puerto: 0
  - Usar el mismo proxy para todo
  - Proxy SSL: [ ] Puerto: 0
  - Proxy FTP: [ ] Puerto: 0
  - Servidor SOCKS: 127.0.0.1 Puerto: 9876
    - SOCKS v4
    - SOCKS v5
    - DNS remoto

No usar proxy para:  
localhost, 127.0.0.1

Ejemplo: .mozilla.org, .net.nz, 192.168.1.0/24

URL para la configuración automática del proxy:  
[ ] Recargar

No preguntar identificación si la contraseña está guardada

At the bottom of the dialog are 'Cancelar' and 'Aceptar' buttons.

# Opción 1: YARN > RM UI

Firefox Archivo Editar Ver Historial Marcadores Herramientas Ventana Ayuda

Ambari - ejemplo-ti... Preferencias

https://ejemplo-tid-spark.azurehdinsight.net/#/main/services/YARN/summary

Buscar

Dashboard Services Hosts Alerts Admin admin

Ambari ejempl... 0 ops 0 alerts

Summary Heatmaps Configs Quick Links Service Actions

YARN Clients 2 YARN Clients Installed

ResourceManager Uptime 17.89 mins

Memory Utilization CPU Utilization Container Failures App Failures Pending Apps

Cluster Memory Cluster Disk Cluster Network Cluster CPU

Actions Last 1 hour

hn0-ejempl.xwvqyxfcyle1bzjjii0wmr0zf.ex.internal.cloudapp.net (Standby)

hn1-ejempl.xwvqyxfcyle1bzjjii0wmr0zf.ex.internal.cloudapp.net (Active)

ResourceManager Heap 280.2 MB / 910.5 MB (30.8% used)

Containers 2 allocated / 0 pending / 0 reserved

Applications 3 submitted / 2 running / 0 pending / completed / 0 killed / 0 failed

Cluster Memory 3.0 GB used / 0 Bytes reserved / 97.0 GB available

Queues 2 Queues

ResourceManager logs ResourceManager JMX Thread Stacks

Actions

https://ejemplo-tid-spark.azurehdinsight.net/yarnui/hn/

The screenshot shows the Ambari Resource Manager (RM) User Interface (UI) for the YARN service. The left sidebar lists various services: HDFS, MapReduce2, YARN (selected), Tez, Hive, Pig, Sqoop, Oozie, ZooKeeper, Ambari Metrics, Jupyter, Livy, and Spark. The main content area has tabs for Summary, Heatmaps, and Configs, with Summary selected. The Summary section displays the status of the App Timeline Server (Started), Standby ResourceManager (Started), Active ResourceManager (Started), and NodeManagers (4/4 Started). It also shows the number of active, lost, unhealthy, rebooted, and decommissioned NodeManagers, along with the number of YARN Clients installed and the ResourceManager's uptime. A detailed view of the Active ResourceManager is shown, listing its host name (hn1-ejempl.xwvqyxfcyle1bzjjii0wmr0zf.ex.internal.cloudapp.net) as Active, its ResourceManager Heap usage (280.2 MB / 910.5 MB, 30.8% used), and links to its ResourceManager UI, logs, JMX, and Thread Stacks. Below the summary are five line charts under the Metrics tab: Memory Utilization, CPU Utilization, Container Failures, App Failures, and Pending Apps. The Pending Apps chart shows 1 Apps and 0.5 Apps. The bottom section contains four more charts: Cluster Memory, Cluster Disk, Cluster Network, and Cluster CPU. The Cluster Disk chart shows 20 Mbps and 10 Mbps. The Cluster Network chart shows 500. The Cluster CPU chart shows 50 %. A large plus sign icon is in the bottom right corner.

# Resource Manager UI

Ambari - ejemplo-ti... All Applications Preferencias

https://ejemplo-tid-spark.azurehdinsight.net/yarnui/hn/cluster Buscar

Logged in as: dr.who

 **hadoop**

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
3	0	2	1	2	3 GB	100 GB	0 B	2	28	0	4	0	0	0	0

Scheduler Metrics

Scheduler Type		Scheduling Resource Type		Minimum Allocation				Maximum Allocation								
Capacity Scheduler		[MEMORY]		<memory:512, vCores:1>				<memory:25600, vCores:7>								
Show 20 entries												Search:				
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1464884087522_0003	root	HIVE-c...	TEZ	default	Thu Jun 2 12:20:36 -0400 2016	Thu Jun 2 12:21:23 -0400 2016	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1464884087522_0002	hive	org.apache.spark.sql.hive.thriftserver.HiveThriftServer2	SPARK	thriftsvr	Thu Jun 2 12:19:07 -0400 2016	N/A	RUNNING	UNDEFINED	1	1	1536	3.0	1.5		ApplicationMaster	0
application_1464884087522_0001	hive	org.apache.spark.sql.hive.thriftserver.HiveThriftServer2	SPARK	thriftsvr	Thu Jun 2 12:18:01 -0400 2016	N/A	RUNNING	UNDEFINED	1	1	1536	3.0	1.5		ApplicationMaster	0

Showing 1 to 3 of 3 entries First Previous 1 Next Last

ELEGIMOS TODO\_

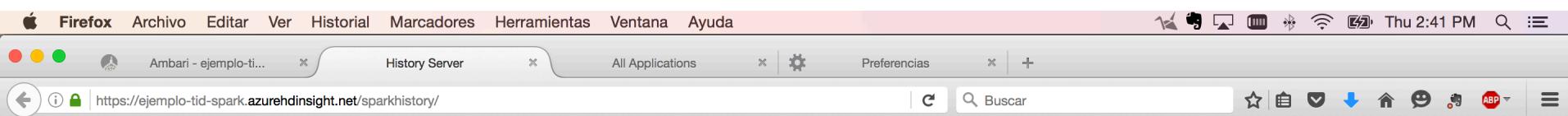
# Opción 2: Spark > Spark HS UI

The screenshot shows the Ambari interface running in a Firefox browser. The URL is <https://ejemplo-ti...azurehdinsight.net/#/main/services/SPARK/summary>. The sidebar on the left lists various Hadoop services: HDFS, MapReduce2, YARN, Tez, Hive, Pig, Sqoop, Oozie, ZooKeeper, Ambari Metrics, Jupyter, Livy, and Spark. The Spark service is currently selected. The main dashboard shows a summary of the Spark cluster status, including the Spark History Server (Started), 2/2 Spark Thrift Servers Live, and 6 Spark Clients Installed. A 'Quick Links' button is present, and a 'Service Actions' dropdown is available. The top navigation bar includes links for Archivo, Editar, Ver, Historial, Marcadores, Herramientas, Ventana, Ayuda, and user information for 'admin'.

Licensed under the Apache License, Version 2.0.

See third-party tools/resources that Ambari uses and their respective authors

# Spark History Server UI



Event log directory: wasb:///hdp/spark-events

Showing 1-3 of 3

1

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated
application_1464884087522_0006	PI-PySpark	2016/06/02 18:36:19	2016/06/02 18:37:17	58 s	profesor	2016/06/02 18:37:17
application_1464884087522_0005	PI-PySpark	2016/06/02 18:33:29	2016/06/02 18:34:15	46 s	profesor	2016/06/02 18:34:14
application_1464884087522_0004	PI-PySpark	2016/06/02 18:27:42	2016/06/02 18:28:34	52 s	profesor	2016/06/02 18:28:34

[Show incomplete applications](#)

# Job view

Firefox Archivo Editar Ver Historial Marcadores Herramientas Ventana Ayuda

Ambari - ejemplo-ti... PI-PySpark - Details for Job 0 All Applications Preferencias

https://ejemplo-tid-spark.azurehdinsight.net/sparkhistory/history/application\_1464884087522\_0006/jobs/job/?id=0 Buscar

Spark 1.6.1 Jobs Stages Storage Environment Executors PI-PySpark application UI

## Details for Job 0

Status: SUCCEEDED  
Completed Stages: 1

Event Timeline

Enable zooming

Executors

- Added
- Removed

Stages

- Completed
- Failed
- Active

Executor 1 added

Executor 3 added

Executor 2 added

reduce at /home/profesor/pi.py:12 (Stage 0.0)

55 0 5 10 15

2 June 14:36 2 June 14:37

DAG Visualization

Stage 0

parallelize

```
graph TD; Start(( )) --> Parallelize[parallelize]; Parallelize --> End(( ));
```

### Completed Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	reduce at /home/profesor/pi.py:12	+details 2016/06/02 18:36:56	20 s	2/2				

# Stage view

Firefox Archivo Editar Ver Historial Marcadores Herramientas Ventana Ayuda

Ambari - ejemplo-ti... PI-PySpark - Details for Stage 0 ... All Applications Preferencias

https://ejemplo-tid-spark.azurehdinsight.net/sparkhistory/history/application\_1464884087522\_0006/stages/stage/?id=0&attempt=0 Buscar

## Details for Stage 0 (Attempt 0)

Total Time Across All Tasks: 38 s  
Locality Level Summary: Process local: 2

DAG Visualization

```
graph TD; parallelize[parallelize] --> PythonRDD[1];
```

Show Additional Metrics

Event Timeline

Enable zooming

Scheduler Delay Task Deserialization Time Shuffle Read Time Executor Computing Time Shuffle Write Time Getting Result Time Result Serialization Time

1 / 10.0.0.7 57 58 59 0 1 2 June 14:36 2 June 14:37

### Summary Metrics for 2 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	19 s	19 s	19 s	19 s	19 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms

### Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks
1	10.0.0.7:42694	41 s	2	0	2

# Executor view

Firefox Archivo Editar Ver Historial Marcadores Herramientas Ventana Ayuda

Ambari - ejemplo-ti... PI-PySpark - Executors (4) All Applications Preferencias

https://ejemplo-tid-spark.azurehdinsight.net/sparkhistory/history/application\_1464884087522\_0006/executors/ Buscar

Spark 1.6.1 Jobs Stages Storage Environment Executors PI-PySpark application UI

## Executors (4)

Memory: 0.0 B Used (17.1 GB Total)  
Disk: 0.0 B Used

Executor ID	Address	RDD Blocks	Storage Memory	Disk Used	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time	Input	Shuffle Read	Shuffle Write	Logs
1	10.0.0.7:42694	0	0.0 B / 5.5 GB	0.0 B	0	0	2	2	40.6 s	0.0 B	0.0 B	0.0 B	stdout stderr
2	10.0.0.16:32944	0	0.0 B / 5.5 GB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	stdout stderr
3	10.0.0.8:46119	0	0.0 B / 5.5 GB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	stdout stderr
driver	10.0.0.22:46053	0	0.0 B / 511.5 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	

# ¿Cómo ejecuto trabajos Spark?

# Pi-PySpark

```
from pyspark import SparkConf, SparkContext, SQLContext
from random import random

def sample(p):
    x, y = random(), random()
    return 1 if x*x + y*y < 1 else 0

conf = SparkConf().setAppName("PI-PySpark")
sc = SparkContext(conf = conf)

NUM_SAMPLES=30000000
count = sc.parallelize(xrange(0, NUM_SAMPLES)).map(sample).reduce(lambda a, b: a + b)

print "Pi is roughly %f" % (4.0 * count / NUM_SAMPLES)
```

# Modo local (default)

```
profesor@hn0-ejempl:~$ vim pi.py
profesor@hn0-ejempl:~$ spark-submit pi.py
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.4.2.0-258/spark/lib/spark-assembly-1.6.1.2.4.2.0-2
SLF4J: Found binding in [jar:file:/usr/hdp/2.4.2.0-258/spark/lib/spark-examples-1.6.1.2.4.2.0-2
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/06/02 18:36:20 INFO SparkContext: Running Spark version 1.6.1
16/06/02 18:36:21 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/06/02 18:36:21 INFO SecurityManager: Changing view acls to: profesor
16/06/02 18:36:21 INFO SecurityManager: Changing modify acls to: profesor
16/06/02 18:36:21 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled
16/06/02 18:36:22 INFO Utils: Successfully started service 'sparkDriver' on port 45788.
16/06/02 18:36:23 INFO Slf4jLogger: Slf4jLogger started
16/06/02 18:36:23 INFO Remoting: Starting remoting
16/06/02 18:36:23 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDri...
16/06/02 18:36:23 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 4449.
16/06/02 18:36:24 INFO SparkEnv: Registering MapOutputTracker
16/06/02 18:36:24 INFO SparkEnv: Registering BlockManagerMaster
16/06/02 18:36:24 INFO DiskBlockManager: Created local directory at /var/tmp/spark/blockmgr-4f2...
16/06/02 18:36:24 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
```



# Modo YARN

```
profesor@hn0-ejempl:~$ vim pi.py
profesor@hn0-ejempl:~$ spark-submit --master yarn pi.py
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.4.2.0-258/spark/lib/spark-assembly-1.6.1.2.4.2.0-
SLF4J: Found binding in [jar:file:/usr/hdp/2.4.2.0-258/spark/lib/spark-examples-1.6.1.2.4.2.0-
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/06/02 20:21:58 INFO SparkContext: Running Spark version 1.6.1
16/06/02 20:21:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/06/02 20:22:00 INFO SecurityManager: Changing view acls to: profesor
16/06/02 20:22:00 INFO SecurityManager: Changing modify acls to: profesor
16/06/02 20:22:00 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled;
16/06/02 20:22:01 INFO Utils: Successfully started service 'sparkDriver' on port 46159.
16/06/02 20:22:02 INFO Slf4jLogger: Slf4jLogger started
16/06/02 20:22:02 INFO Remoting: Starting remoting
16/06/02 20:22:02 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkD...
16/06/02 20:22:02 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 47372.
16/06/02 20:22:02 INFO SparkEnv: Registering MapOutputTracker
16/06/02 20:22:02 INFO SparkEnv: Registering BlockManagerMaster
16/06/02 20:22:02 INFO DiskBlockManager: Created local directory at /var/tmp/spark/blockmgr-9d...
16/06/02 20:22:02 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
```



## Executors (4)

Memory: 0.0 B Used (17.1 GB Total)

Disk: 0.0 B Used

modo local

Executor ID	Address	RDD Blocks	Storage Memory	Disk Used	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time	Input	Shuffle Read	Shuffle Write	Logs
1	10.0.0.7:42694	0	0.0 B / 5.5 GB	0.0 B	0	0	2	2	40.6 s	0.0 B	0.0 B	0.0 B	stdout stderr
2	10.0.0.16:32944	0	0.0 B / 5.5 GB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	stdout stderr
3	10.0.0.8:46119	0	0.0 B / 5.5 GB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	stdout stderr
driver	10.0.0.22:46053	0	0.0 B / 511.5 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	

## Executors (4)

Memory: 0.0 B Used (17.1 GB Total)

Disk: 0.0 B Used

modo YARN

Executor ID	Address	RDD Blocks	Storage Memory	Disk Used	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time	Input	Shuffle Read	Shuffle Write	Logs
1	10.0.0.7:36058	0	0.0 B / 5.5 GB	0.0 B	0	0	1	1	20.6 s	0.0 B	0.0 B	0.0 B	stdout stderr
2	10.0.0.12:39501	0	0.0 B / 5.5 GB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	stdout stderr
3	10.0.0.8:44281	0	0.0 B / 5.5 GB	0.0 B	0	0	1	1	19.9 s	0.0 B	0.0 B	0.0 B	stdout stderr
driver	10.0.0.22:44299	0	0.0 B / 511.5 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B	

ELEGIMOS TODO\_

**¡Gracias!**

*ELEGIMOS TODO\_*