

Aplicación de Link Analysis sobre los artículos de Wikipedia

Diana Salazar Báez dsalaz539@gmail.com

Ignacio Gómez Rodriguez jignaciogomezr@gmail.com

Iván Morales Cotes ibanivan33@gmail.com

Universidad Distrital Francisco José De Caldas
Maestría en Ciencias de la Información y las Comunicaciones
Bogotá, Colombia

Tendencias En Ingeniería de Software
Noviembre de 2017

RESUMEN

EXPLICACIÓN GENERAL DEL PROYECTO

DESARROLLO

CONCLUSIONES

RESUMEN

Este trabajo describe el análisis de enlaces de las páginas de artículos de Wikipedia utilizando el algoritmo PageRank junto con Hadoop y MapReduce, alojando la aplicación en una maquina virtual de Google Cloud.

EXPLICACIÓN GENERAL DEL PROYECTO

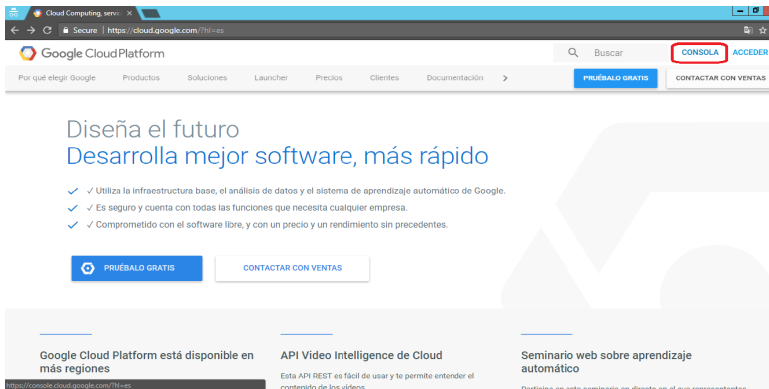
El objetivo principal del proyecto es aplicar las técnicas de “Link Analysis” sobre un conjunto de datos obtenidos de la página de Wikipedia. El proyecto se enfocará en encontrar el ranking para determinar cuáles son las más referenciadas. Se utilizará Hadoop, Google Cloud Platform y R entre otros. Se estudiarán temas como Big Data, Inteligencia De Negocios y Gestión Del Conocimiento en búsqueda de soluciones que permitan mejorar la extracción de conocimiento a partir del análisis de los datos.

DESARROLLO

Para el desarrollo de este proyecto se realizaron diferentes actividades como:

- Asociar cuenta de Google a Google Cloud Platform
- Crear máquina virtual en Google Cloud Platform
- Instalaciones en máquina virtual
- Trabajo sobre la máquina virtual
- Utilización de bucket en GCP
- Utilización de Google Cloud Shell
- Utilización de Sql Server y Sql Server Data Tools
- Utilización de R

Asociar cuenta de Google a Google Cloud Platform Ingresar a la página principal de GCP

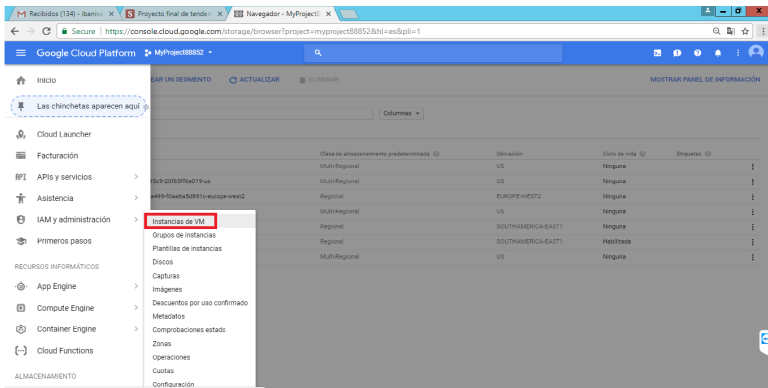


Asociar cuenta de Google a Google Cloud Platform

Se debe seleccionar la opción "Gratis"

The screenshot shows the Google Cloud Platform console interface. At the top, there's a navigation bar with the Google Cloud Platform logo and a search bar. Below the navigation bar, there's a section titled 'Introducción'. The main content area displays several cards for getting started with Google Cloud Platform. The first card, 'Prueba Google Cloud Platform de forma gratuita', is highlighted in blue and contains the text: 'Regístrate y recibirás un crédito de 300 USD y 12 meses para explorar la solución.' Below this text is a button labeled 'Empezar'. Other cards include 'Aprende a usar Cloud Storage', 'Aprende a usar Google Cloud Platform', 'Prueba Compute Engine', 'Crea una instancia de Cloud SQL', 'Usa APIs de Google', 'Prueba App Engine', and 'Documentación'. Each card has a brief description and an 'Empezar' button.

Crear máquina virtual en Google Cloud Platform



Crear máquina virtual en Google Cloud Platform

Seleccionar las características de la Máquina

The screenshot shows the Google Cloud Platform console interface for creating a new instance. The browser address bar shows the URL: <https://console.cloud.google.com/compute/instancesAdd?project=myproject8852&hl=es#preconfigured-image-debian-9-stretch-v20171025>.

The page title is "Crear una instancia". The left sidebar contains navigation icons for various cloud services.

The main configuration area includes the following sections:

- Nombre:** A text input field containing "instance-2".
- Zona:** A dropdown menu showing "us-east1-b".
- Tipo de máquina:** A dropdown menu showing "16 vCPU" and "60 GB de memoria". A link "Personalizar" is available.
- Disco de arranque:** A section showing a new persistent SSD disk of 1000 GB. The image selected is "Windows Server 2016". A "Cambiar" button is present.

Below the disk section, there is a note about Microsoft software licensing and a link to "Más información sobre los requisitos de movilidad de licencias de Microsoft".

The **Identidad y acceso de API** section shows the "Cuenta de servicio" dropdown set to "Compute Engine default service account". Under "Alcance del acceso", the option "Permitir el acceso predeterminado" is selected.

The **Costos/precios** section at the bottom left mentions adding firewall rules and network traffic tags.

On the right side, the estimated costs are displayed:

- 1.025,56 \$ al mes (estimación)**
- Tarifa por horas efectiva: 1,405 \$ (730 horas al mes)

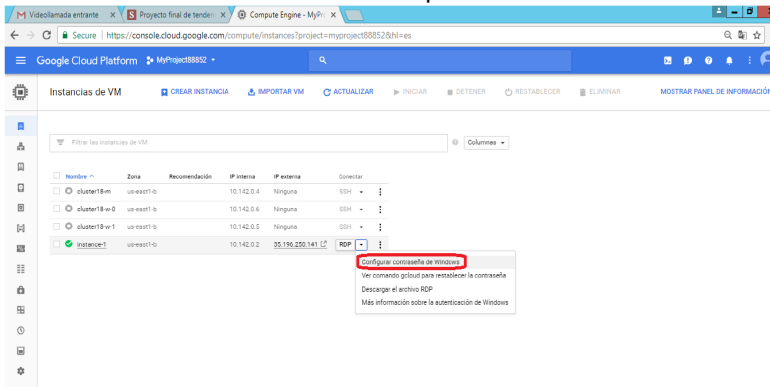
Elemento	Costos estimados
16 vCPUs con 60 GB de memoria	554,80 \$/mes
Disco persistente SSD de 1000 GB	170,00 \$/mes
Tarifa por uso de Windows Server 2016*	467,20 \$/mes
Descuento por uso continuado	- 166,44 \$/mes
Total	1.025,56 \$/mes

Footnote: * Microsoft cobra la tarifa por uso de imagen y Google la factura.

Links: [Precios de Compute Engine](#) and [Menos](#).

Crear máquina virtual en Google Cloud Platform

Establecer una contraseña para inicio de sesión.



The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud Platform logo and the project name 'MyProject88852'. The main content area is titled 'Instancias de VM' and contains a table of virtual machine instances. The table has columns for 'Nombre', 'Zona', 'Recomendación', 'IP interna', 'IP externa', and 'Conectar'. The instance 'instance-1' is highlighted, and the 'RDP' button is clicked, opening a dropdown menu with the following options:

- Configurar contraseña de Windows
- Ver comando gcloud para restablecer la contraseña
- Descargar el archivo RDP
- Más información sobre la autenticación de Windows

Crear máquina virtual en Google Cloud Platform

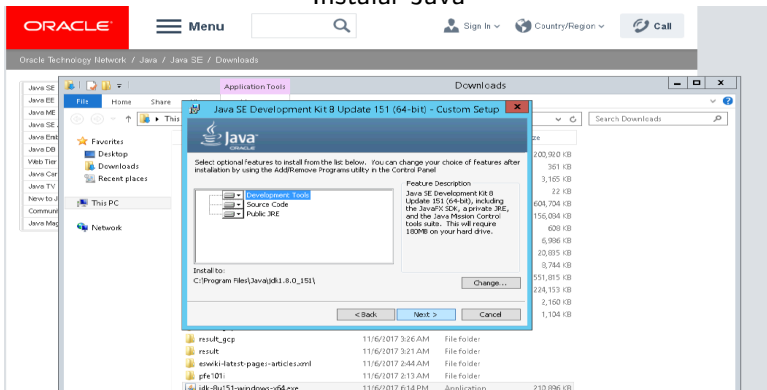
Iniciar la conexión Remota sobre la IP Asignada.

The screenshot shows the Google Cloud Platform console for project 'MyProject88852'. The 'Instancias de VM' (VM Instances) page is active, displaying a table of instances. The instance 'instance-1' is highlighted with a green checkmark. A red box highlights the 'IP externa' (35.196.250.141) for 'instance-1'. A red arrow points from this IP to the 'Computer' field of a 'Remote Desktop Connection' dialog box, which also has a red box around the IP address. The dialog box shows the user name 'grupotendencias' and a 'Connect' button.

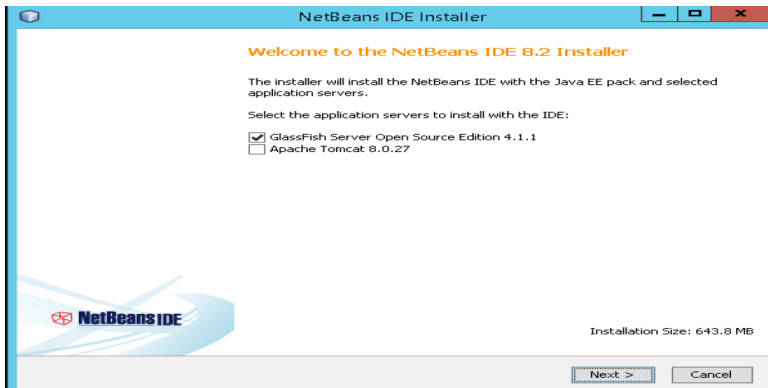
Nombre	Zona	Recomendación	IP interna	IP externa	Conectar
cluster18-m	us-east1-b		10.142.0.4	Ninguna	SSH
cluster18-w-0	us-east1-b		10.142.0.6	Ninguna	SSH
cluster18-w-1	us-east1-b		10.142.0.5	Ninguna	SSH
instance-1	us-east1-b		10.142.0.2	35.196.250.141	RDP

Instalaciones en máquina virtual

Instalar Java

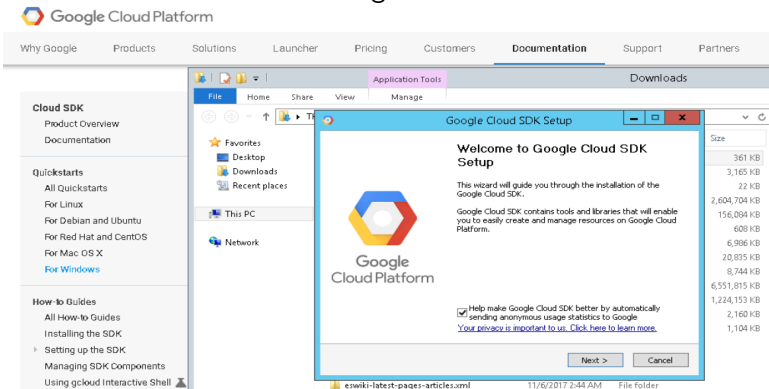


Instalaciones en máquina virtual NetBeans



Instalaciones en máquina virtual

Instalar Google Cloud SDK



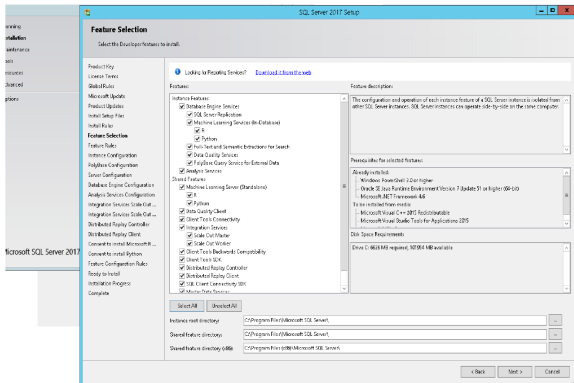
Instalaciones en máquina virtual

Instalar Google Cloud SDK (Seleccionar Componentes)

The image shows a screenshot of the Google Cloud Platform website and the Google Cloud SDK Setup window. The website is in the background, displaying the Google Cloud Platform logo and navigation links. The setup window is in the foreground, showing the 'Google Cloud SDK Setup' dialog. The window has a title bar with 'Google Cloud SDK Setup' and standard window controls. The main content area is titled 'Google Cloud Platform' and contains the following text: 'Check the components you want to install and uncheck the components you don't want to install. Click Install to start the installation.' Below this, there is a section 'Select components to install:' with a list of components and checkboxes. The components are: 'Cloud SDK Core Libraries and Tools' (checked), 'Bundled Python' (checked), 'Cloud Tools for PowerShell' (checked), and 'Beta Commands' (checked). Below the list, there is a 'Description' section with the text: 'Position your mouse over a component to see its description.' At the bottom of the window, there is a 'Space required: 89.1MB' label and three buttons: '< Back', 'Install', and 'Cancel'. In the background, the Google Cloud Platform website is visible, showing the 'Documentation' tab selected in the top navigation bar. The left sidebar of the website lists 'Cloud SDK' with links to 'Product Overview' and 'Documentation', and 'Quickstarts' with links for Linux, Debian and Ubuntu, Red Hat and CentOS, Mac OS X, and Windows. The right sidebar of the website shows a list of files with their sizes.

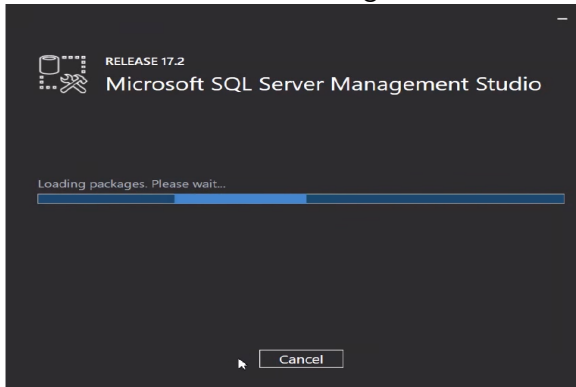
Instalaciones en máquina virtual

Instalar SQL Server 2017



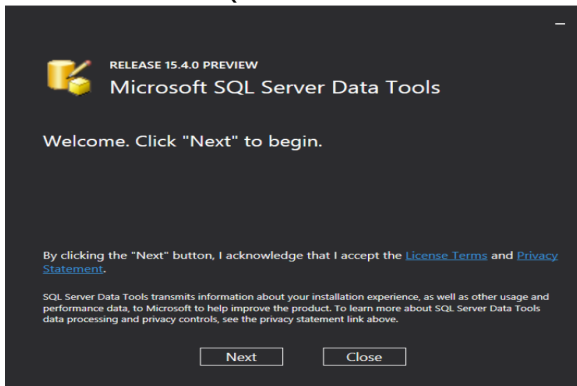
Instalaciones en máquina virtual

Instalar SQL Studio Management Studio



Instalaciones en máquina virtual

Instalar SQL Server Data Tools



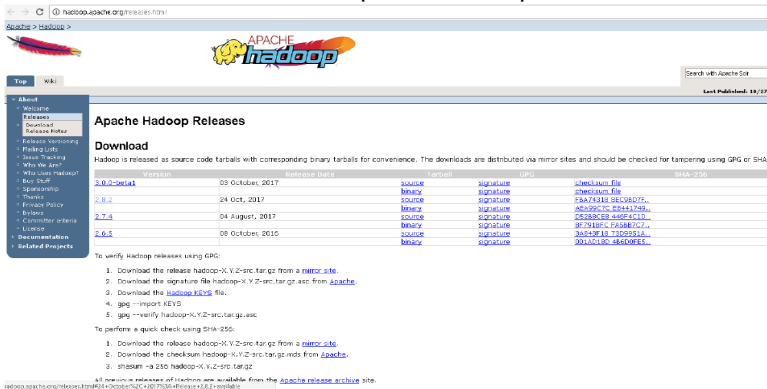
Instalaciones en máquina virtual

Instalar SQL Server Data Tools (Seleccionar Herramientas)



Instalaciones en máquina virtual

Instalar Apache Hadoop



The screenshot shows the Apache Hadoop Releases page. The browser address bar displays `hadoop.apache.org/releases.html`. The page features the Apache Hadoop logo and a search bar. A left sidebar contains navigation links such as 'About', 'Releases', 'Download', 'Release Notes', 'Release Versioning', 'Hacking Lists', 'Issue Tracking', 'Who We Are', 'Who Uses Hadoop?', 'Buy Stuff', 'Sponsorship', 'Thanks', 'Privacy Policy', 'Blogs', 'Committer Criteria', 'License', 'Documentation', and 'Related Projects'. The main content area is titled 'Apache Hadoop Releases' and includes a 'Download' section. This section explains that Hadoop is released as source code tarballs with corresponding binary tarballs for convenience, distributed via mirror sites and checked for tampering using GPG or SHA-256. A table lists the releases, with columns for Version, Release Date, Tarball, GPG, and SHA-256. The table includes entries for 3.0.0-beta1, 2.8.2, 2.7.4, and 2.6.5, each with links to source and binary tarballs, GPG signatures, and SHA-256 checksum files. Below the table, instructions are provided for verifying releases using GPG and performing a quick check using SHA-256. At the bottom, a note states that all previous releases of Hadoop are available from the Apache release archive site.

hadoop.apache.org/releases.html

Apache > Hadoop >

Top Wiki

Apache Hadoop Releases

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release Date	Tarball	GPG	SHA-256
3.0.0-beta1	03 October, 2017	source	signature	checksum file
		binary	signature	checksum file
2.8.2	24 Oct, 2017	source	signature	FBA753A8-9E59ED75...
		binary	signature	AC499C7C-82441749...
2.7.4	04 August, 2017	source	signature	25238CE8-446F4C10...
		binary	signature	9F7910F1-F4587C7...
2.6.5	08 October, 2016	source	signature	3A549F18-7102934...
		binary	signature	D016C18C-486D0FE...

To verify Hadoop releases using GPG:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the signature file `hadoop-X.Y.Z-src.tar.gz.asc` from [apache](#).
3. Download the [hadoop-X.Y.Z](#) file.
4. `gpg --import KEYS`
5. `gpg --verify hadoop-X.Y.Z-src.tar.gz.asc`

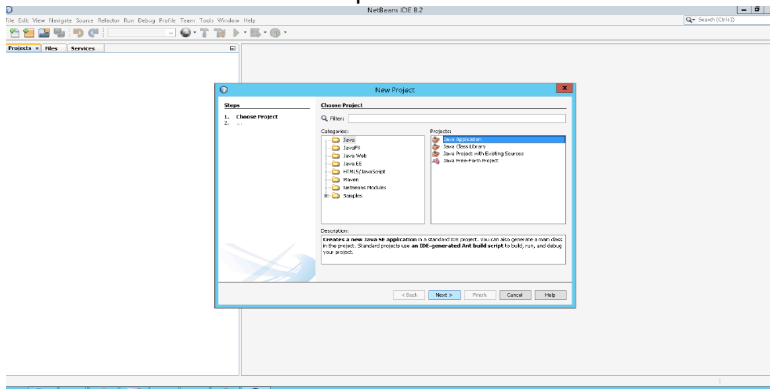
To perform a quick check using SHA-256:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the checksum `hadoop-X.Y.Z-src.tar.gz.md5` from [apache](#).
3. `shsum -h 256 hadoop-X.Y.Z-src.tar.gz`

All previous releases of Hadoop are available from the [Apache release archive](#) site: `hadoop.apache.org/releases.html#20170304+20170304+Release+2.8.2+available`

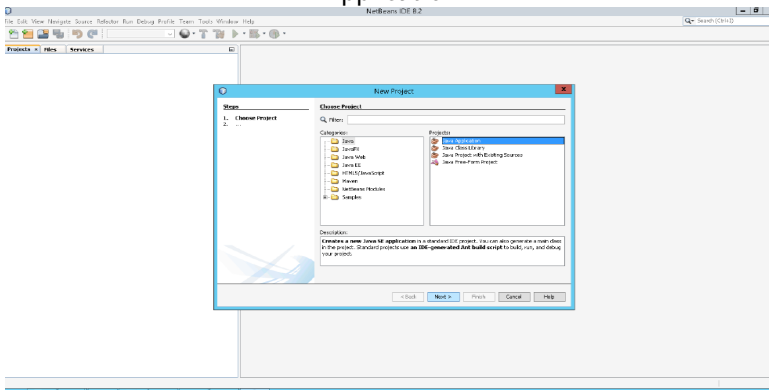
Crear máquina virtual en Google Cloud Platform

Se selecciona la opción "Crear Instancia"



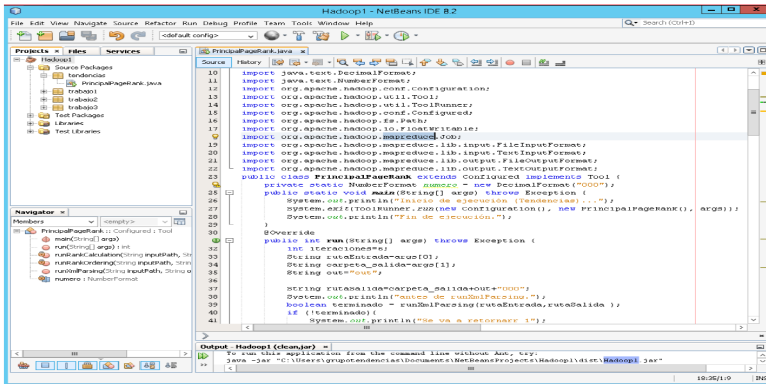
Trabajo en la máquina virtual

Se realiza la creación de un nuevo proyecto sobre NetBeans de tipo Java Application



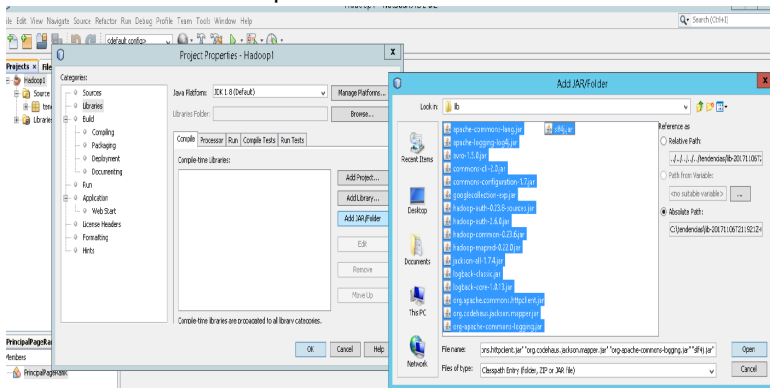
Trabajo en la máquina virtual

Se realiza la codificación de la solución. (Código Fuente Disponible en GitHub)



Trabajo en la máquina virtual

Se realiza la importación de la librerías necesarias.



















Trabajo en la máquina virtual

Se realiza la ejecución del .jar sobre Hadoop, donde se genera un error.

[illegible]

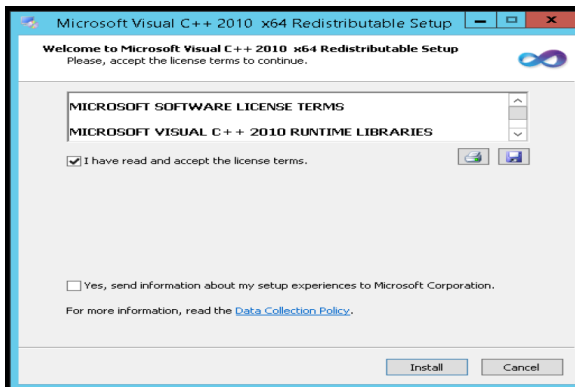
Trabajo en la máquina virtual

Para solucionar los errores se utiliza winutils.exe

GitHub, Inc. [US] https://github.com/steveloughran/winutils/tree/master/hadoop-2.8.1		Latest commit 2878787 on Aug 29
steveloughran sign Hadoop artifacts		
..		
 hadoop.dll	Added Hadoop 2.8.1 libraries	2 months ago
 hadoop.dll.asc	sign Hadoop artifacts	2 months ago
 hadoop.exp	Added Hadoop 2.8.1 libraries	2 months ago
 hadoop.exp.asc	sign Hadoop artifacts	2 months ago
 hadoop.lib	Added Hadoop 2.8.1 libraries	2 months ago
 hadoop.lib.asc	sign Hadoop artifacts	2 months ago
 hdfs.dll	Added Hadoop 2.8.1 libraries	2 months ago
 hdfs.dll.asc	sign Hadoop artifacts	2 months ago
 hdfs.exp	Added Hadoop 2.8.1 libraries	2 months ago
 hdfs.exp.asc	sign Hadoop artifacts	2 months ago
 hdfs.lib	Added Hadoop 2.8.1 libraries	2 months ago
 hdfs.lib.asc	sign Hadoop artifacts	2 months ago
 libwinutils.lib	Added Hadoop 2.8.1 libraries	2 months ago
 libwinutils.lib.asc	sign Hadoop artifacts	2 months ago
 winutils.exe	Added Hadoop 2.8.1 libraries	2 months ago
 winutils.exe.asc	sign Hadoop artifacts	2 months ago

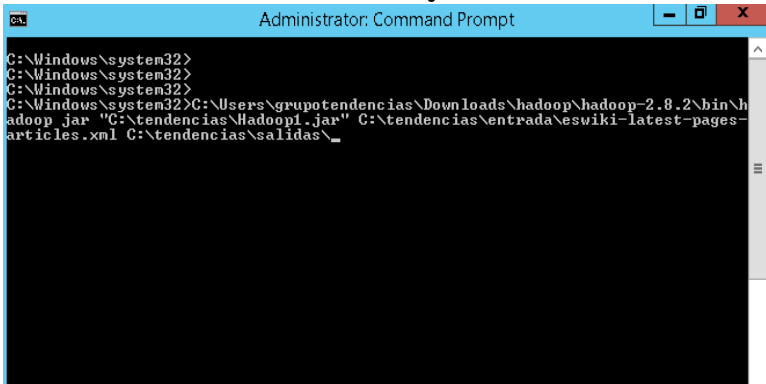
Trabajo en la máquina virtual

Para solucionar otro error se realiza la instalación de Visual C++



Trabajo en la máquina virtual

Se realiza nuevamente la ejecución sin errores.



```
Administrator: Command Prompt
C:\Windows\system32>
C:\Windows\system32>
C:\Windows\system32>
C:\Windows\system32>C:\Users\grupotendencias\Downloads\hadoop\hadoop-2.8.2\bin\h
adoop jar "C:\tendencias\Hadoop1.jar" C:\tendencias\entrada\eswiki-latest-pages-
articles.xml C:\tendencias\salidas\
```

Utilización de bucket en GCP:

Crear Bucket (segmento) en GCP



Google Cloud Platform MyProject88852

Crear un segmento

Nombre ⓘ
Debe ser único en todo Cloud Storage. Privacidad: no incluyas información confidencial en los nombres de segmento. Los demás pueden descubrir el nombre de este si coincide con un nombre que intentan usar.

bucket-tendencias

Clase de almacenamiento predeterminada ⓘ

- ☒ **Multi-Regional**
Se utiliza para emitir videos y alojar contenido web muy solicitado.
Es la mejor opción para los datos a los que se accede con frecuencia y desde cualquier lugar del mundo.
- ☐ **Regional**
Se utiliza para almacenar datos y ejecutar análisis de datos.
Es la mejor opción para los datos a los que se accede con frecuencia y desde una zona concreta.
- ☐ **Nearline**
Se utiliza para almacenar documentos a los que no se suele acceder.
Es la opción perfecta para los datos a los que se accede menos de una vez al mes.
- ☐ **Coldline**
Se utiliza para almacenar documentos a los que casi nunca se accede.
Es la opción perfecta para los datos a los que se accede menos de una vez al año.

Ubicación de Multi-Regional
Redundante en más de 2 regiones en la ubicación seleccionada.

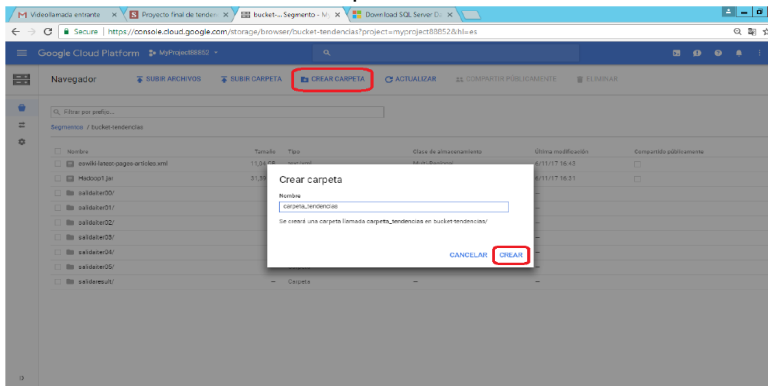
Estados Unidos

☒ Especificar etiquetas

Crear Cancelar

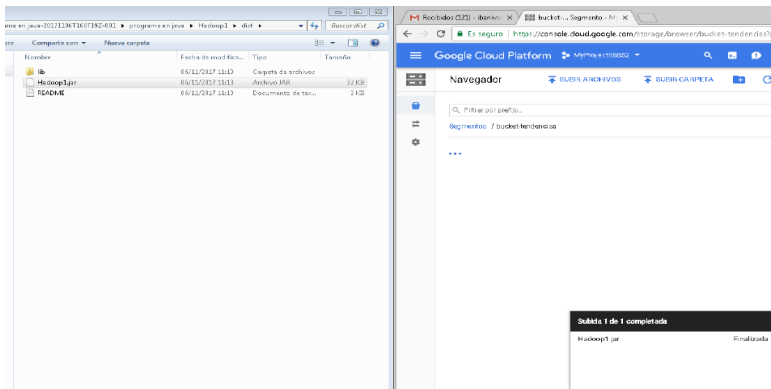
Utilización de bucket en GCP

Crear carpeta en Bucket



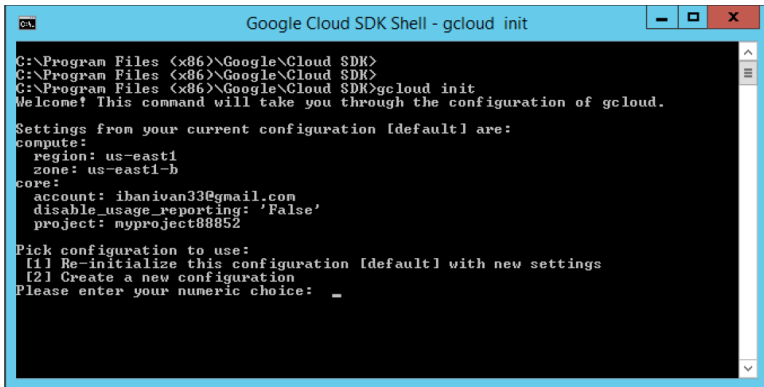
Utilización de bucket en GCP

Cargar Archivos .jar y archivos de artículos



Utilización de Google Cloud Shell y Google Cloud

Crear cluster (dataproc) en Google Cloud Platform



```
Google Cloud SDK Shell - gcloud init

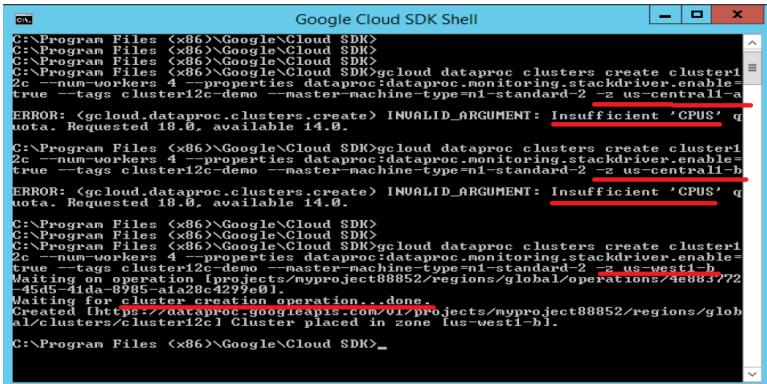
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud init
Welcome! This command will take you through the configuration of gcloud.

Settings from your current configuration [default] are:
compute:
  region: us-east1
  zone: us-east1-b
core:
  account: ibanivan33@gmail.com
  disable_usage_reporting: 'False'
  project: myproject88852

Pick configuration to use:
[1] Re-initialize this configuration [default] with new settings
[2] Create a new configuration
Please enter your numeric choice: _
```


Utilización de Google Cloud Shell y Google Cloud

Cambio de región, por falta de CPUs para procesamiento, y creación exitosa.



```

C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster1
2c --num-workers 4 --properties dataproc:dataproc.monitoring.stackdriver.enable=
true --tags cluster12c-demo --master-machine-type=n1-standard-2 -z us-central1-a
ERROR: (gcloud.dataproc.clusters.create) INVALID_ARGUMENT: Insufficient 'CPUS' q
uota. Requested 18.0, available 14.0.
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster1
2c --num-workers 4 --properties dataproc:dataproc.monitoring.stackdriver.enable=
true --tags cluster12c-demo --master-machine-type=n1-standard-2 -z us-central1-b
ERROR: (gcloud.dataproc.clusters.create) INVALID_ARGUMENT: Insufficient 'CPUS' q
uota. Requested 18.0, available 14.0.
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster1
2c --num-workers 4 --properties dataproc:dataproc.monitoring.stackdriver.enable=
true --tags cluster12c-demo --master-machine-type=n1-standard-2 -z us-west1-b
Waiting on operation [projects/myproject88852/regions/global/operations/4e883772
-45d5-41da-8985-a1a28c4299e0].
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/myproject88852/regions/glob
al/clusters/cluster12c1 Cluster placed in zone us-west1-b1.
C:\Program Files (x86)\Google\Cloud SDK>_
  
```

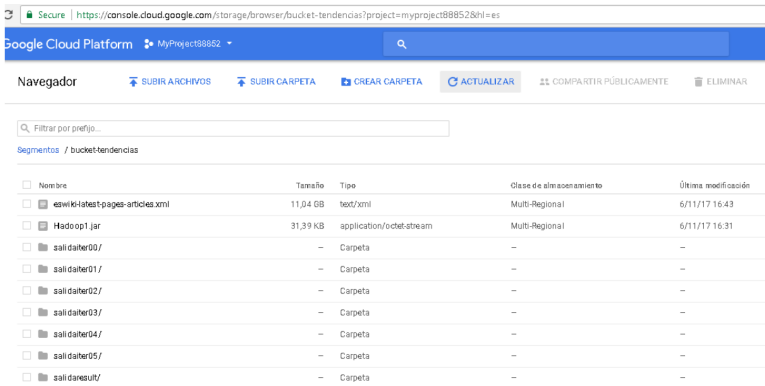
Utilización de Google Cloud Shell y Google Cloud

Se realiza la ejecución del .jar.



```
Cloud Shell
myproject88652 x
Welcome to Cloud Shell! Type "help" to get started.
ibnani@myproject88652:~$
ibnani@myproject88652:~$ gcloud dataproc jobs submit hadoop --cluster cluster18 --jar "gs://bucket-tendencias/Esdoop1.jar" -- gs://bucket-tendencias/eswiki-latest-pages-articles.xml gs://bucket-tendencias/salida
```

Utilización de Google Cloud Shell y Google Cloud Visualización de los archivos de salida.

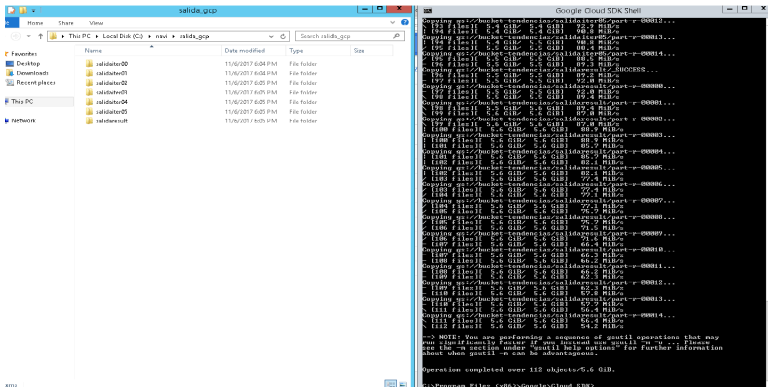


The screenshot shows the Google Cloud Platform console interface. The address bar displays the URL: <https://console.cloud.google.com/storage/browser/bucket-tendencias?project=myproject88852&hl=es>. The page title is "Google Cloud Platform" with a dropdown menu showing "MyProject88852". Below the header, there is a navigation bar with buttons: "Navegador", "SUBIR ARCHIVOS", "SUBIR CARPETA", "CREAR CARPETA", "ACTUALIZAR", "COMPARTIR PÚBLICAMENTE", and "ELIMINAR". A search bar is present with the placeholder text "Filtrar por prefijo...". Below the search bar, the breadcrumb "Segmentos / bucket-tendencias" is visible. The main content area displays a table of files and folders in the bucket.

<input type="checkbox"/>	Nombre	Tamaño	Tipo	Clase de almacenamiento	Última modificación
<input type="checkbox"/>	eswiki-latest-pages-articles.xml	11,04 GB	text/xml	Multi-Regional	6/11/17 16:43
<input type="checkbox"/>	Hadoop1.jar	31,39 KB	application/octet-stream	Multi-Regional	6/11/17 16:31
<input type="checkbox"/>	salidaiter00/	-	Carpeta	-	-
<input type="checkbox"/>	salidaiter01/	-	Carpeta	-	-
<input type="checkbox"/>	salidaiter02/	-	Carpeta	-	-
<input type="checkbox"/>	salidaiter03/	-	Carpeta	-	-
<input type="checkbox"/>	salidaiter04/	-	Carpeta	-	-
<input type="checkbox"/>	salidaiter05/	-	Carpeta	-	-
<input type="checkbox"/>	salidaresult/	-	Carpeta	-	-

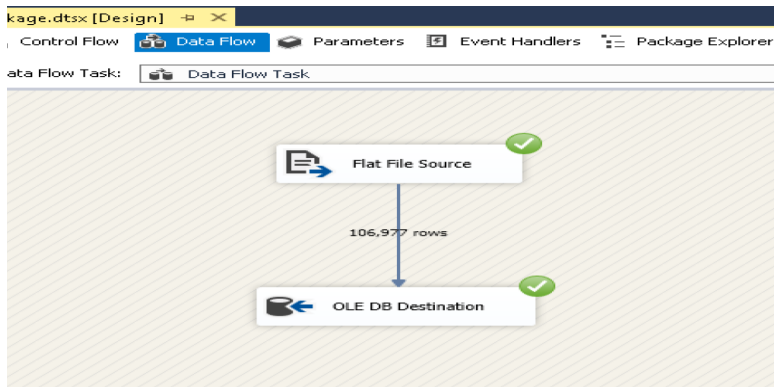
Utilización de Google Cloud Shell y Google Cloud

Descarga de los archivos de salida hacia archivos locales.



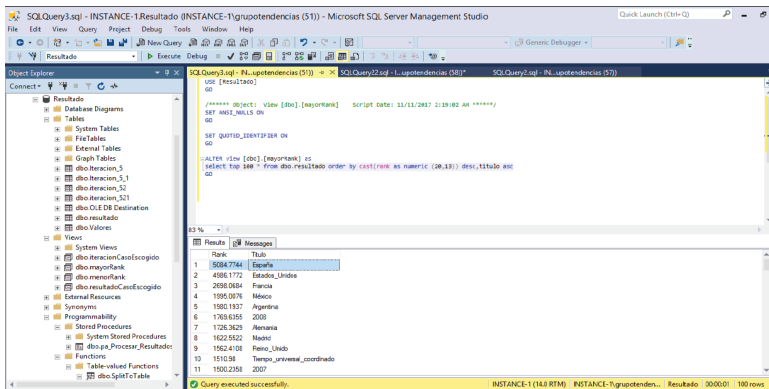
Utilización de Sql Server y Sql Server Data Tools

Creación y ejecución de la DTS, para el cargue de los archivos de resultado.



Utilización de Sql Server y Sql Server Data Tools

Creación de la base de datos y consultas realizadas.



The screenshot shows the Microsoft SQL Server Enterprise Manager interface. The Object Explorer on the left displays the database structure, including tables, views, and functions. The central pane shows the SQL query being executed, and the bottom pane displays the results of the query.

SQL Query:

```
USE [Resultado]
GO

/***** Object: View [dbo].[mayorRank]    Script date: 11/11/2017 21:39:02 AM *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

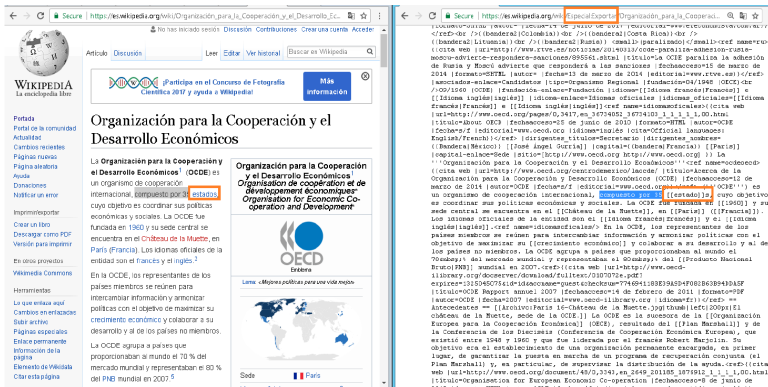
--ALTER view [dbo].[mayorRank] as
select top 10 * from dbo.Resultado order by cast(rank as numeric (20,13)) desc,titulo asc
GO
```

Query Results:

Rank	Titulo
1	5084.7744 España
2	4586.1772 Estados_Unidos
3	2698.0684 Francia
4	1995.0076 México
5	1980.1937 Argentina
6	1769.6355 2008
7	1726.3429 Alemania
8	1622.5822 Madrid
9	1562.4108 Reino_Unido
10	1510.98 Tiempo_universal_coordinado
11	1500.2558 2007

Query executed successfully. 100 rows.

Análisis funcional de una iteración Especial:Export para ver XML de la página



Wikipedia la enciclopedia libre

Portal de la comunidad
Actualidad
Cambios recientes
Páginas nuevas
Páginas especiales
Ayuda
Donaciones
Notificar un error
Imprimir/exportar
Crear un libro
Descargar como PDF
Vender para imprimir
En otros proyectos
Wikimedia Commons
Herramientas
La que estás aquí
Cambios en enlaces
Subir archivo
Páginas especiales
Enlace permanente
Información de la página
El historial de Wikidata
Citar esta página

Organización para la Cooperación y el Desarrollo Económicos

La **Organización para la Cooperación y el Desarrollo Económicos** (**OCDE**) es un organismo de cooperación internacional, compuesto por 34 **estados** cuyo objetivo es coordinar sus políticas económicas y sociales. La OCDE fue fundada en 1960 y su sede central se encuentra en el **Château de la Muette**, en París (Francia). Los idiomas oficiales de la entidad son el francés y el inglés.¹

En la OCDE, los representantes de los países miembros se reúnen para intercambiar información y armonizar políticas con el objetivo de maximizar su crecimiento económico y colaborar a su desarrollo y al de los países no miembros. La OCDE agrupa a países que **proponen** al mundo el "modelo" del mercado mundial y representan el 80% del PIB mundial en 2007.²

Organización para la Cooperación y el Desarrollo Económicos
Organisation de coopération et de développement économiques
Organisation for Economic Co-operation and Development

Logo de la OCDE

Lema: «Mejora política para una vida mejor»

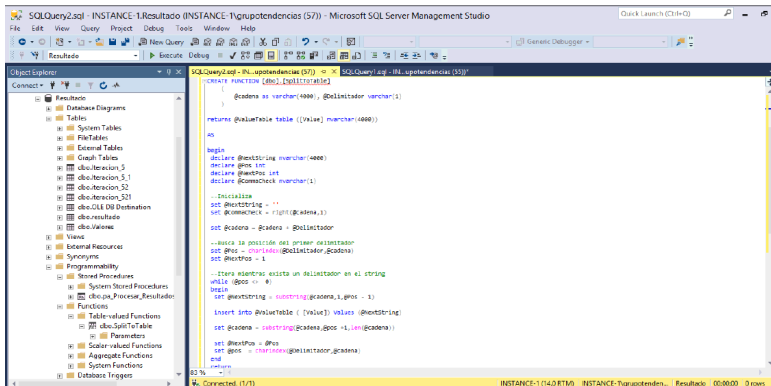
Sede: París

El **Château de la Muette**, sede de la OCDE. El **Château de la Muette** es la sede de la Organización Europea para la Cooperación Económica (OECE) (1962), resultado del (Plan Marshall) y de la Conferencia de los Diecisiete (Conferencia de Cooperación Económica Europea), que existió entre 1948 y 1960 y que fue liderada por el francés Robert Marjolin. Su objetivo era el establecimiento de una organización permanente encargada, en primer lugar, de garantizar la puesta en marcha de un programa de recuperación conjunta (el Plan Marshall) y, en particular, de supervisar la distribución de la ayuda. **Export para ver XML de la página**

Enlaces hacia pagina seleccionada para ejemplo

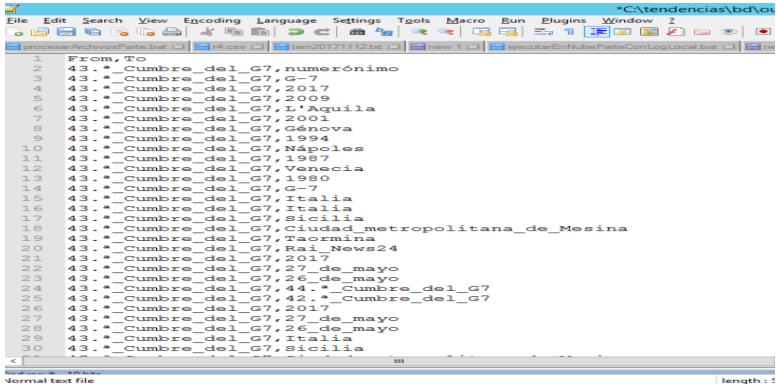
[illegible]

Creación de Función y Procedimiento Almacenado para generación de Entrada de R



Análisis funcional de una iteración

Visualización de archivo de Entrada de R

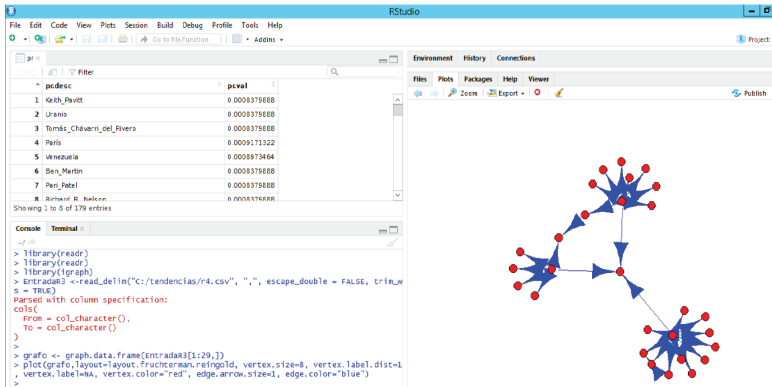


The screenshot shows a text editor window with a menu bar (File, Edit, Search, View, Encoding, Language, Settings, Tools, Macro, Run, Plugins, Window) and a toolbar. The file path in the title bar is *C:\tendencias\bd\ot. The editor displays a list of entries, each starting with a line number and a tab character, followed by the text 'From, To' and a list of items. The items are: Cumbre_del_G7, numerónimo; Cumbre_del_G7, G-7; Cumbre_del_G7, 2017; Cumbre_del_G7, 2009; Cumbre_del_G7, L'Aquila; Cumbre_del_G7, 2001; Cumbre_del_G7, Génova; Cumbre_del_G7, 1994; Cumbre_del_G7, Nápoles; Cumbre_del_G7, 1987; Cumbre_del_G7, Venecia; Cumbre_del_G7, 1980; Cumbre_del_G7, G-7; Cumbre_del_G7, Italia; Cumbre_del_G7, Italia; Cumbre_del_G7, Sicilia; Cumbre_del_G7, Ciudad metropolitana de Mesina; Cumbre_del_G7, Taormina; Cumbre_del_G7, Rai_News24; Cumbre_del_G7, 2017; Cumbre_del_G7, 27 de mayo; Cumbre_del_G7, 26 de mayo; Cumbre_del_G7, 44.*_Cumbre_del_G7; Cumbre_del_G7, 42.*_Cumbre_del_G7; Cumbre_del_G7, 2017; Cumbre_del_G7, 27 de mayo; Cumbre_del_G7, 26 de mayo; Cumbre_del_G7, Italia; Cumbre_del_G7, Sicilia. The status bar at the bottom indicates 'length : 1'.

```
1 From, To
2 43.*_Cumbre_del_G7, numerónimo
3 43.*_Cumbre_del_G7, G-7
4 43.*_Cumbre_del_G7, 2017
5 43.*_Cumbre_del_G7, 2009
6 43.*_Cumbre_del_G7, L'Aquila
7 43.*_Cumbre_del_G7, 2001
8 43.*_Cumbre_del_G7, Génova
9 43.*_Cumbre_del_G7, 1994
10 43.*_Cumbre_del_G7, Nápoles
11 43.*_Cumbre_del_G7, 1987
12 43.*_Cumbre_del_G7, Venecia
13 43.*_Cumbre_del_G7, 1980
14 43.*_Cumbre_del_G7, G-7
15 43.*_Cumbre_del_G7, Italia
16 43.*_Cumbre_del_G7, Italia
17 43.*_Cumbre_del_G7, Sicilia
18 43.*_Cumbre_del_G7, Ciudad metropolitana de Mesina
19 43.*_Cumbre_del_G7, Taormina
20 43.*_Cumbre_del_G7, Rai_News24
21 43.*_Cumbre_del_G7, 2017
22 43.*_Cumbre_del_G7, 27 de mayo
23 43.*_Cumbre_del_G7, 26 de mayo
24 43.*_Cumbre_del_G7, 44.*_Cumbre_del_G7
25 43.*_Cumbre_del_G7, 42.*_Cumbre_del_G7
26 43.*_Cumbre_del_G7, 2017
27 43.*_Cumbre_del_G7, 27 de mayo
28 43.*_Cumbre_del_G7, 26 de mayo
29 43.*_Cumbre_del_G7, Italia
30 43.*_Cumbre_del_G7, Sicilia
```

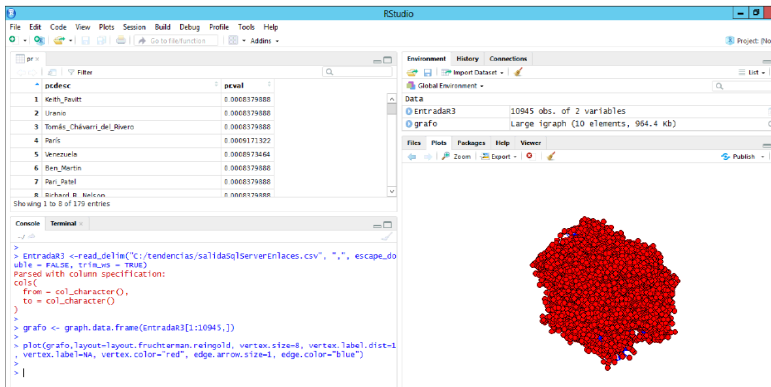
Análisis En R

Generación de gráfico con pocos datos



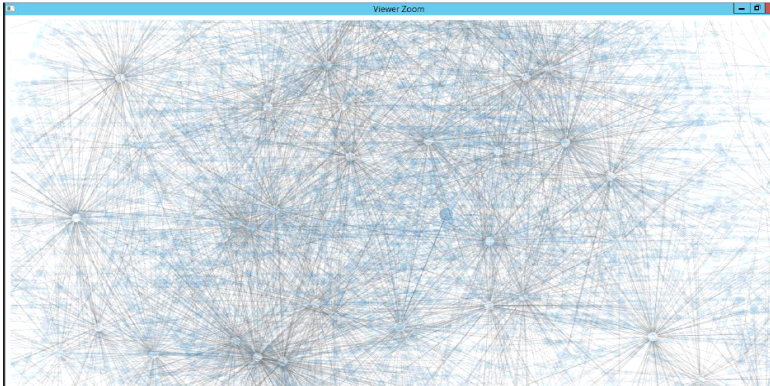
Análisis En R

Generación de gráfico con todos los datos de la página de ejemplo



Análisis En R

Generación de gráfico con visualización en 3D y todo el conjunto de páginas



CONCLUSIONES

Conclusiones

- Se realizó un proceso de link analysis sobre una muestra de todos los artículos de la base de datos de wikipedia.
- Se utilizó Google Cloud Platform para realizar el procesamiento de la información en un cluster; dicho procesamiento consistió en utilizar la técnica de MapReduce con una librería de Hadoop 2.8.2.
- Se logro evidenciar que gracias a la aplicación de MapReduce y la librería de Hadoop se pudo realizar procesamiento paralelo en varios servidores pertenecientes a un cluster de tal manera que se redujo el tiempo de procesamiento significativamente.
- El utilizar Google Cloud Platform nos permite de manera dinámica utilizar infraestructura como servicio en la nube, mejorando nuestro hardware en los momentos críticos (Procesamiento de la información con mayor demanda de recursos).

- Se aplicaron los conceptos adquiridos durante el curso, como fueron el manejo de ETLs, Link Analysis, Conceptos de Big Data, MapReduce, KDD, programación literaria y Bases de Datos.
- Aplicando el algoritmo de PageRank se determinaron cuáles páginas resultaban ser mas relevantes.
- El proyecto nos aportó conocimientos y experiencia para aplicar a futuros proyectos académicos y laborales.
- Las páginas con mayor PageRank son las relacionadas con los nombres de países. Lo anterior puede deberse al hecho de que el contenido de dichas páginas es bastante formal y completo, y además sabemos que es muy frecuente que en diferentes artículos se haga referencia a algún país relacionado.
- Una de las acciones que se puede tomar ante los resultados obtenidos, es analizar las páginas con mayor PageRank para saber si existe o no algún "atacante" que esté logrando aumentar su PageRank aplicando alguna técnica como SpamFarming.

BIBLIOGRAFÍA

- Introducing the Azure services platform: Chappell, David and others, White paper, Oct 2008
- Service Orientation. (2017). Service Orientation. [online] Available at:<http://soapprinciples.com/p2esp.php> Accessed 17 Agost. 2017
- Sommerville, I. (2004). Software Engineering. International computer science series. ed: Addison Wesley.