



**UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS**

Aplicación de Link Analysis sobre los artículos de Wikipedia

Diana Salazar Báez

COD. 20171495019

Ignacio Gómez Rodriguez

COD. 20171495011

Iván Morales Cotes

COD. 20171495013

Maestría En Ciencias De La Información Y Las Comunicaciones

Tendencias En Ingeniería De Software

Universidad Distrital Francisco José de Caldas

Bogotá, Agosto 11 de 2017

Aplicación de Link Analysis sobre los artículos de Wikipedia

Datos de contacto:
Diana Salazar Báez
COD. 20171495019
dsalaz539@gmail.com
Cel: 3106785270
Ignacio Gómez Rodriguez
jignaciogomezr@gmail.com
COD. 20171495011
Cel: 3007075686
Iván Morales Cotes
ibanivan33@gmail.com
COD. 20171495013
3006291217

Maestría En Ciencias De La Información Y Las Comunicaciones
Tendencias En Ingeniería De Software
Universidad Distrital Francisco José de Caldas
Bogotá, Agosto 11 de 2017

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 máquina virtual de Google Cloud.

Palabras clave: (Wikipedia, PageRank, Hadoop, MapReduce, Google Cloud).

Contenido

Resumen	III
1 Introducción	2
2 Capítulo 1. Planteamiento del problema	3
2.1 Explicación General del proyecto	3
2.2 Resultados Esperados	3
2.3 Recursos requeridos	3
2.4 Requerimientos	4
2.4.1 Requerimientos funcionales	4
2.4.2 Requerimientos no funcionales	5
2.5 Valor del Proyecto	5
2.6 Selección de datos	6
3 Capítulo 2. Marco de referencia	7
3.1 Análisis de enlaces.	7
3.2 PageRank.	8
3.3 Hadoop.	9
3.4 MapReduce.	10
3.5 Google Cloud Platform	11
4 Capítulo 3. Desarrollo	12
4.1 Asociar cuenta de google a Google Cloud Platform	12
4.2 Crear máquina virtual en Google Cloud Platform	15
4.3 Instalaciones en máquina virtual	21
4.3.1 Instalar Java	21
4.3.2 Instalar Netbeans	22
4.3.3 Instalar Google SDK	23
4.3.4 Instalar SQL Server	25
4.3.5 Instalar SQL Server Management Studio	27
4.3.6 Instalar SQL Server Data Tools (SSDT)	30
4.3.7 Descargar y descomprimir Hadoop	33
4.4 Trabajo en la máquina virtual	35
4.4.1 Realizar programa en Java con Netbeans	35

4.4.2	Generar archivo Jar con Netbeans	37
4.4.3	Agregar librerías con Netbeans	37
4.4.4	Ejecutar Jar con Hadoop	38
4.4.5	Agregar winutils a Hadoop	39
4.5	Utilización de bucket en GCP:	43
4.5.1	Crear Bucket (segmento) en Gcp	43
4.5.2	Crear carpeta en Bucket	44
4.5.3	Cargar en Bucket el programa y los datos	44
4.6	Utilización de Google Cloud Shell y Google Cloud SDK Shell	47
4.6.1	Crear cluster (dataproc) en Google Cloud Platform	47
4.6.2	Ejecutar en cluster el programa utilizando Bucket (Cloud Shell) . . .	51
4.6.3	Descargar del bucket en la máquina virtual el resultado del programa (Sdk Shell)	54
4.7	Utilización de Sql Server y Sql Server Data Tools	58
4.7.1	Crear base de datos en Sql Server	58
4.7.2	Crear ETL	59
4.7.3	Usar ETL	63
4.7.4	Consultar tabla de resultado de PageRank	63
4.7.5	Análisis funcional del resultado	64
4.7.6	Análisis funcional de una iteración	66
4.7.7	Crear y Ejecutar Procedimiento almacenado y Función	73
4.7.8	Exportar muestra de iteración resultante de procedimiento almacenado	75
4.8	Utilización de R	77
4.8.1	Importar csv de muestra de datos de iteración en R	77
4.8.2	Procesar datos en R	77
4.8.3	Graficar datos en R	78
5	Conclusiones	81
Bibliografía		82

Lista de Figuras

2-1. Torrent of Wikimedia Commons files torrents	6
3-1. Ejemplo de Page Rank[3]	8
3-2. Página principal de Hadoop[1]	9
3-3. Panorama conceptual de Map Reduce[4]	10
3-4. Página de productos y servicios de Google Cloud Platform[5]	11
4-1. Página inicial de Google Cloud Platform	12
4-2. Inicio de sesión con la cuenta de Google	13
4-3. Bienvenida a Google Cloud Platform	13
4-4. Selección de Prueba Gratuita de Google Cloud Platform	14
4-5. Selección de País y Aceptación de condiciones	14
4-6. Ingreso de datos personales	15
4-7. Selección de opción de instancias de máquinas virtuales	15
4-8. Click para crear máquina virtual	16
4-9. Selección de sistema operativo para la máquina virtual	16
4-10. Selección de CPUs y memoria RAM	17
4-11. Política de disponibilidad y reinicio automático	17
4-12. Configurar contraseña de Windows	18
4-13. Establecer usuario sobre la máquina virtual	18
4-14. Conexión de escritorio remoto de Windows	19
4-15. Ip externa para acceder remotamente	20
4-16. Instalación del JDK de Java	21
4-17. Instalación del IDE Netbeans V.8.2	22
4-18. Progreso de instalación de Netbeans	22
4-19. Página de descarga en Google Cloud Platform	23
4-20. Instalación del Google Cloud SDK	23
4-21. Selección de componentes de Google Cloud SDK	24
4-22. Archivos resultado de la instalación	24
4-23. Arrancar la interacción con Google Cloud	25
4-24. Instalación de SQL Server 2017	25
4-25. Configuración del servidor	26
4-26. Proceso de la instalación de SQL Server 2017	26
4-27. Instalación completa de SQL Server	27

4-28. Ingreso a la página de SSMS y descarga del instalador	27
4-29. Ejecución del instalador de SSMS	28
4-30. Cargue de paquetes de SSMS	28
4-31. Progreso de la instalación de SSMS	29
4-32. Instalación exitosa	29
4-33. Página principal y Descarga de SSDT	30
4-34. Inicio de Instalación de SSDT	30
4-35. Instalación de herramientas de SSDT	31
4-36. Progreso de la instalación de SSDT	31
4-37. Instalación exitosa de SSDT	32
4-38. Icono de acceso a SSDT	32
4-39. Descarga de Apache Hadoop	33
4-40. Carpeta de Hadoop	33
4-41. Asignación de la variable JAVA HOME	34
4-42. Creación de una aplicación java	35
4-43. Asignación del nombre y ubicación del proyecto	36
4-44. Nombre y ubicación de la clase	36
4-45. Generación de código en Java y Compilación	37
4-46. Agregar Librerías del Proyecto	37
4-47. Error generado en hadoop	38
4-48. Descarga de winutils desde Github	39
4-49. Paso de archivos descomprimidos de winutils a la carpeta de hadoop	40
4-50. Error generado en hadoop	40
4-51. Instalación de Visual C++	41
4-52. Ejecución correcta de hadoop	42
4-53. Creación del Segmento	43
4-54. Crear Carpeta en Bucket	44
4-55. Paso de archivos .jar al Bucket	44
4-56. Carga del archivo xml	45
4-57. Progreso de la carga del archivo xml	45
4-58. Carga completa de los archivos	46
4-59. Ejecución del comando gcloud init	47
4-60. Escoger zona de gcloud	48
4-61. Comando para creación del dataproc cluster	48
4-62. Error por falta de espacio	49
4-63. Solución a falta de espacio	49
4-64. Error por falta de CPU	50
4-65. Lista de zonas para creación del cluster	50
4-66. Error, solución y creación exitosa del Cluster Dataproc	51
4-67. Comando para el cluster	51

4-68. Primera Iteración	52
4-69. Progreso de iteraciones	52
4-70. Finalización de las iteraciones	53
4-71. Resultado en el Bucket	53
4-72. Iniciar la interacción con Google Cloud	54
4-73. Asignación de permisos a la cuenta personal	55
4-74. Código para la aplicación	55
4-75. Ingreso de Código	56
4-76. Ventana de Shell inicializando gcloud	56
4-77. Ejecución del comando para traer los archivos de Google Cloud	57
4-78. Archivos copiados provenientes de Google Cloud Platform	57
4-79. Conexión al servidor INSTANCE -1	58
4-80. Seleccionar base de datos nueva	58
4-81. Crear base de datos	59
4-82. Crear un nuevo proyecto	59
4-83. Proyecto de tipo Integration Service Project	60
4-84. Creación del Data Flow	60
4-85. Creación del Data Flow	61
4-86. Creación de la conexión hacia el archivo plano	61
4-87. Conexión hacia la base de datos	62
4-88. Selección de tabla de la base de datos	62
4-89. Ejecución de la ETL	63
4-90. Consulta del resultado final ordenado	63
4-91. Página de España en Wikipedia	64
4-92. América con un enlace hacia España en Wikipedia	65
4-93. Título y Special:Export	66
4-94. Visualización del título en el tag title	67
4-95. Pagerank de la página seleccionada para analizar	67
4-96. Parte de resultado de iteración 04	68
4-97. Resultado de iteración 04; relación con la página de Gerardo Jiménez Sánchez	69
4-98. Referencia desde página de Gerardo Jiménez Sánchez	70
4-99. Referencia desde página Biblioteca depositaria	71
4-100. Referencia a otras páginas	71
4-101. Comparación de enlaces hacia la página de José Ángel Gurría	72
4-102. Comparación de enlaces hacia la página de StatLinks	72
4-103. rocedimiento almacenado en SQL Server	73
4-104. La función en SQL Server	74
4-105. Ejecución del procedimiento almacenado	74
4-106. Ejecutar la consulta para generar la muestra	75
4-107. Guardar el resultado de la consulta en un archivo .csv	75

4-108	Muestra final de iteración	76
4-109	Vizualización de pequeña muestra de resultados	77
4-110	Vizualización del grafo	78
4-111	Visualización aplicando PageRank	79
4-112	Visualización 3D aplicando PageRank	79
4-113	Visualización 3D aplicando PageRank con menor zoom	80

1 Introducción

El análisis de enlaces brinda un aporte significativo a la búsqueda de páginas web que contengan información confiable, facilitando clasificar las páginas por su importancia teniendo en cuenta el rango en que se encuentran y graficar las diferentes relaciones entre los enlaces de las páginas; para determinar el rango Google desarrolló el algoritmo PageRank el cual analiza los enlaces de salida y de entrada de un determinado sitio web y le asigna un puntaje de importancia.

Wikipedia es una herramienta frecuentemente utilizada para realizar consultas de información y por tanto se hace relevante identificar las páginas que contengan información confiable; esto se puede hacer realizando un análisis de enlaces en las páginas de Wikipedia.

Este trabajo describe como se puede realizar el análisis de enlaces de Wikipedia aplicando el algoritmo PageRank conjuntamente con Hadoop y MapReduce. Para su desarrollo se abordan 4 capítulos principales los cuales consisten en: En el capítulo uno tenemos la descripción de lo relacionado con el planteamiento del problema, el cronograma y el presupuesto; en el capítulo dos el marco de referencia; en el capítulo tres el desarrollo; en el capítulo cuatro análisis de resultados y finalmente el capítulo de conclusiones y recomendaciones.

2 Capítulo 1. Planteamiento del problema

2.1. 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. Se analizarán principalmente los enlaces que contiene cada entrada direcccionando a otras entradas. Así como se realiza Link Analysis sobre internet para determinar el nivel de confianza e importancia de una página web teniendo en cuenta diferentes variables como la cantidad de usuarios que la visitan, la cantidad de links que le apuntan en diferentes sitios web, y la cantidad de links que tiene dicha página hacia otras de alta relevancia; el proyecto se enfocará en encontrar el PageRank para determinar cuáles son las mas referenciadas. Se utilizarán herramientas reconocidas para el análisis y el tratamiento de los datos; se utilizará Hadoop, Google Cloud Platform y R entre otros.

2.2. Resultados Esperados

Se espera que a partir del Link Analysis se pueda determinar cuáles son las páginas más citadas en Wikipedia. Se espera que las herramientas utilizadas (software) nos faciliten el tratamiento de un gran volumen de datos, el análisis de los mismos y la obtención de resultados (reportes gráficos que le ayuden a identificar que páginas tienen un mayor rango de citación). Se espera que durante el proyecto se adquieran conocimientos relevantes para el contexto laboral y profesional, dado que el Link Analysis puede llegar a ser utilizado en diferentes áreas y procesos de las organizaciones que pretenden ser más eficientes.

2.3. Recursos requeridos

- Datos de la página web de Wikipedia.
- 144 horas de trabajo correspondientes a 3 ingenieros.

- 5 computadores.
- 1 Cluster con 4 workers de Procesamiento.
- Google Cloud Platform, Microsoft Sql Server, Netbeans, Hadoop.
- Referencias bibliográficas.
- Revisión y asesoría por parte del Profesor.

2.4. Requerimientos

- Se utilizarán herramientas reconocidas para el análisis y el tratamiento de los datos como Hadoop y Dataproc.
- Se desarrollará un programa con Java que permitirá utilizar la técnica de MapReduce con el fin de procesar todos los datos de los artículos previamente descargados de Wikipedia.
- Se utilizará Google Cloud Platform para tener en la nube los datos a analizar, el programa a ejecutar y las máquinas (clusters) que se encargarán de ejecutar el programa con el fin de realizar el procesamiento de los datos.
- Se utilizarán Sql Server y ETLs (SSDT) con el fin de cargar en una base de datos los archivos resultantes del programa previamente ejecutado con Hadoop, para poder realizar diferentes consultas y generar lo que será la entrada de R (csv).
- Se utilizará R con el fin de verificar el PageRank previamente calculado y graficar una muestra de los resultados de tal forma que se puedan visualizar los diferentes nodos y sus enlaces.
- 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.
- Se debe aplicar Link Analysis sobre un conjunto de datos grande y de interés, con el fin de llegar a conclusiones que permitan establecer relaciones entre los diferentes nodos y la importancia de los mismos.

2.4.1. Requerimientos funcionales

- Se debe determinar cuáles artículos de WikiPedia son los que tienen mejor calificación (PageRank).

2.4.2. Requerimientos no funcionales

- Utilizando Hadoop, en la nube (ejemplo: Google Cloud Platform), se debe ejecutar un proceso de Link Analysis que permita finalmente graficar y mostrar resultados cuantitativos sobre al análisis de los nodos del respectivo grafo.
- Se debe utilizar un lenguaje como Java, y se debe tener una base de datos con los resultados (ejemplo: SqlServer).

2.5. Valor del Proyecto

Link Analysis es una técnica de análisis de datos usada para evaluar las relaciones (conexiones) entre nodos. Las relaciones pueden ser identificadas entre diferentes tipos de nodos (objetos), incluyendo organizaciones, personas y transacciones entre otros. Link Analysis ha sido usado en investigaciones de actividades criminales (detección de fraude, lucha antiterorista e inteligencia), análisis de seguridad informática, optimización de motores de búsqueda, investigación de mercado, investigación médica y arte.

El aporte principal del proyecto será la clasificación de las páginas más citadas en Wikipedia con la finalidad de identificar la calidad de la información en los diferentes artículos.

2.6. Selección de datos

Se tomó como referencia la base de datos de Wikipedia de 23 TB (Figura 2.1); esta base de datos contiene imágenes, videos, etc.

Se utilizó una maquina virtual en Google Cloud, para realizar la descarga (gracias a la velocidad que presentan estas máquinas en la nube) y extraer únicamente los artículos completos en formato XML correspondientes a las páginas de diferentes países alrededor del mundo.

Commons [edit]
[Torrent of Wikimedia Commons files torrents](#) (over 23 TB in total as of early 2013).

English Wikipedia [edit]

- [enwiki-20170820-pages-meta-current.xml.bz2](#) (25.86 GiB) on itorrents.org
- [enwiki-20170820-pages-articles.xml.bz2](#) (13.27 GiB) on itorrents.org
- [enwiki-20170820-pages-articles-multistream.xml.bz2](#) (14.1 GiB) on itorrents.org
- [enwiki-20170720-pages-meta-current.xml.bz2](#) (25.71 GiB) on itorrents.org
- [enwiki-20170720-pages-articles.xml.bz2](#) (13.19 GiB) on itorrents.org
- [enwiki-20170620-pages-meta-current.xml.bz2](#) (25.55 GiB) on itorrents.org
- [enwiki-20170620-pages-articles.xml.bz2](#) (13.11 GiB) on itorrents.org
- [enwiki-20170420-pages-meta-current.xml.bz2](#) (25.23 GiB) on itorrents.org
- [enwiki-20170420-pages-articles.xml.bz2](#) (12.95 GiB) on itorrents.org
- [enwiki-20170401-pages-meta-current.xml.bz2](#) (25.1 GiB) on itorrents.org
- [enwiki-20170401-pages-articles.xml.bz2](#) (12.9 GiB) on itorrents.org
- [enwiki-20170320-pages-meta-current.xml.bz2](#) (25 GiB) on itorrents.org
- [enwiki-20170320-pages-articles.xml.bz2](#) (12.8 GiB) on itorrents.org
- [enwiki-20170301-pages-meta-current.xml.bz2](#) (24.9 GiB) on itorrents.org
- [enwiki-20170301-pages-articles.xml.bz2](#) (12.8 GiB) on itorrents.org
- [enwiki-20170220-pages-meta-current.xml.bz2](#) (24.9 GiB) on itorrents.org
- [enwiki-20170220-pages-articles.xml.bz2](#) (12.7 GiB) on itorrents.org
- [enwiki-20170220-all-files](#) (226 GiB) on itorrents.org

Figura 2-1: Torrent of Wikimedia Commons files torrents

3 Capítulo 2. Marco de referencia

En este capítulo se describe el marco teórico del proyecto.

3.1. Análisis de enlaces.

Análisis de enlaces es una técnica de análisis de datos utilizada para evaluar relaciones (conexiones) entre nodos. Las relaciones pueden identificarse entre varios tipos de nodos (objetos), incluidas organizaciones, personas y transacciones. El análisis de enlaces se ha utilizado para la investigación de actividades delictivas (detección de fraude, antiterrorismo e inteligencia), análisis de seguridad informática , optimización de motores de búsqueda , estudios de mercado, investigación médica y arte.

3.2. PageRank.

Es un conjunto de algoritmos para asignar puntajes a un sitio web y determinar su importancia o relevancia.

Formula:

$$\text{PR} (A) = (1-d) / N + d (\text{PR} (T_1) / C (T_1) + \dots + \text{PR} (T_n) / C (T_n))$$

$T_1, T_2 \dots T_n$ = son páginas que apuntan a A. d = Factor de amortiguación que se establece entre 0 y 1; el mas usado es 0.85. $C (T_i)$ es el número de citas de T_i .

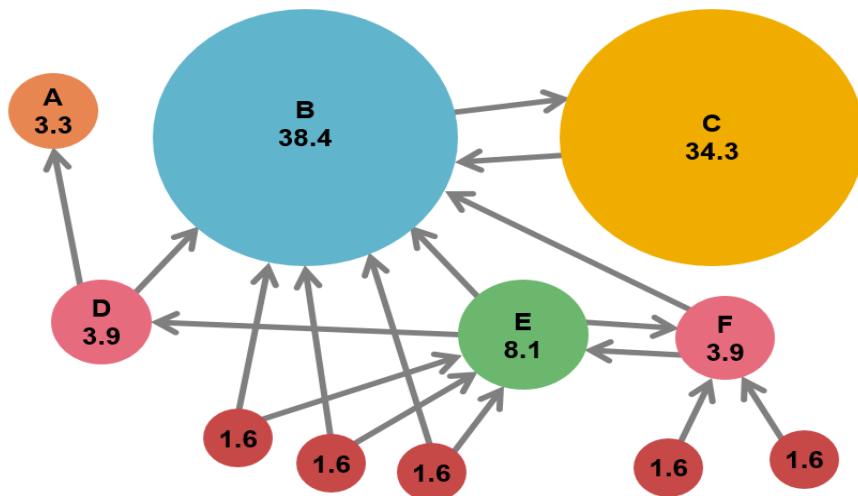
Si un conjunto de datos tiene N documentos T_1, T_2, \dots, T_N , entonces la suma de todos los rangos de página es:

$$\text{PR} (T_1) + \text{PR} (T_2) + \text{PR} (T_3) + \dots + \text{PR} (T_N) = 1.$$

El concepto básico de esta fórmula es que una página distribuye su puntaje (PageRank) igualmente a todos sus enlaces de salida.

El PageRank máximo para cualquier página dada no puede exceder 1.

Example: PageRank Scores



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

Figura 3-1: Ejemplo de Page Rank[3]

3.3. Hadoop.

Hadoop es un framework de software open-source licenciado bajo Apache License 2.0 para procesamiento a gran escala. Hadoop es un software abierto el cual permite realizar procesamiento distribuido de grandes conjuntos de datos; esto se realiza en varios clusters de servidores.[1].

Hadoop incluye los siguientes módulos [1]:

- Hadoop Common : Contiene las utilidades comunes que son compatibles con los otros módulos de Hadoop.
- HDFS: sistema de archivos distribuido que proporciona acceso de alto rendimiento a los datos de las aplicaciones.
- HADOOP YARN : un marco para la programación de trabajos y la administración de recursos de clusters.
- Hadoop MapReduce : un sistema basado en YARN para el procesamiento paralelo de grandes conjuntos de datos.

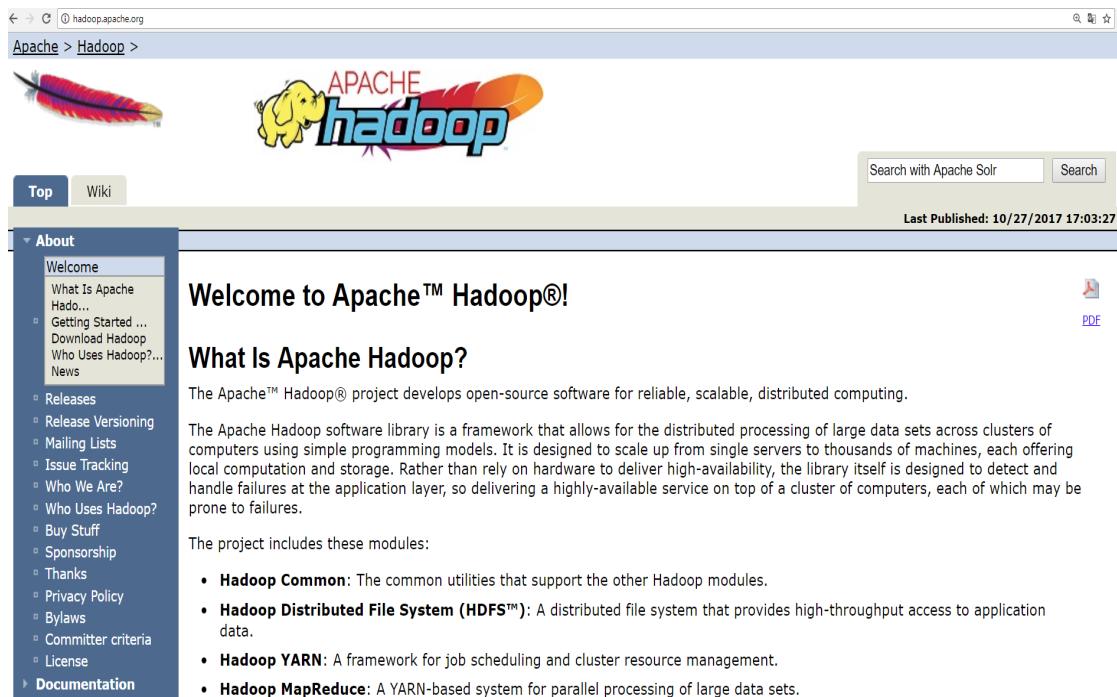


Figura 3-2: Página principal de Hadoop[1]

3.4. MapReduce.

El concepto de MapReduce es fundamental para entender la arquitectura Hadoop; este concepto se basa en que MapReduce es un paradigma de programación que permite manejar grandes cantidades de datos.[2],

MapReduce utiliza un algoritmo paralelo y distribuido en N clusters.

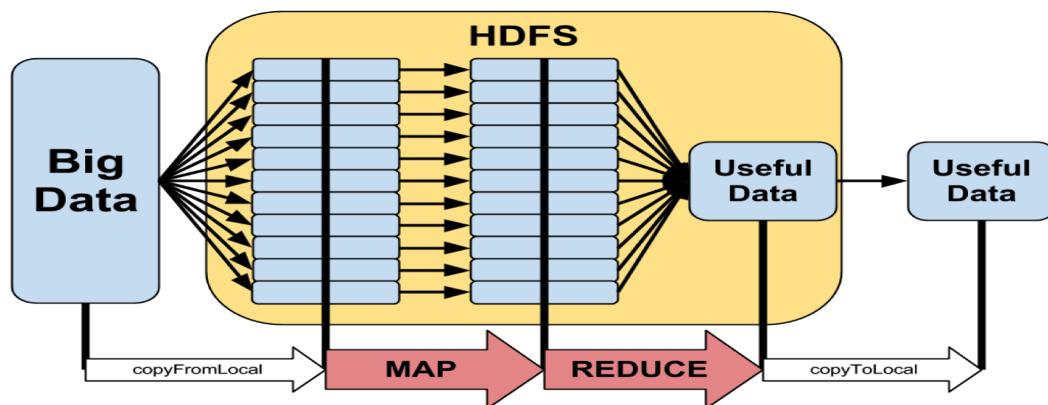


Figura 3-3: Panorama conceptual de Map Reduce[4]

3.5. Google Cloud Platform

Google Cloud Platform consiste en una plataforma que reúne todos las aplicaciones web de Google, la cual permite a desarrolladores diseñar, implementar y probar diversas aplicaciones. Ventajas de Google Cloud Platform:

- Escalabilidad
- Rapidez
- Libera de los gastos operativos relacionados con la administración de la infraestructura.
- Datos y analítica de gran alcance.
- Búsquedas en milisegundos.

The screenshot shows the Google Cloud Platform website at https://cloud.google.com/products/?hl=es. The top navigation bar includes links for 'Seguro', 'https://cloud.google.com/products/?hl=es', 'Google Cloud Platform' logo, 'Productos' (selected), 'Soluciones', 'Launcher', 'Precios', 'Clientes', 'Documentación', 'As', a search bar, and user account options. Below the navigation is a blue button labeled 'PRUEBALO GRATIS'. The main content area is titled 'PRODUCTOS Y SERVICIOS' and features a sub-header 'Ejecuta tu aplicación con la misma tecnología y herramientas que se utilizan en Google'. It lists several product categories: 'Recursos informáticos' (Compute Engine, App Engine, Container Engine), 'Almacenamiento y bases de datos' (Cloud Storage, Cloud SQL, Cloud Bigtable), and 'Redes' (Virtual Private Cloud (VPC), Cloud Load Balancing, Cloud CDN). Each category has a detailed description and a right-pointing arrow indicating more options.

Figura 3-4: Página de productos y servicios de Google Cloud Platform[5]

4 Capítulo 3. Desarrollo

En este capítulo se describe de forma detallada el proceso técnico del desarrollo del proyecto.

Para este proyecto se realizó el desarrollo de una aplicación en Java utilizando Netbeans 8.2, con las librerías de Apache Hadoop, con el fin de realizar un proceso de MapReduce, y así poder encontrar el PageRank de las páginas de Wikipedia; a continuación se explican los pasos necesarios para realizarlo:

4.1. Asociar cuenta de google a Google Cloud Platform

Se ingresa a la página principal de Google Cloud Platform y se hace click en la opción de Consola.

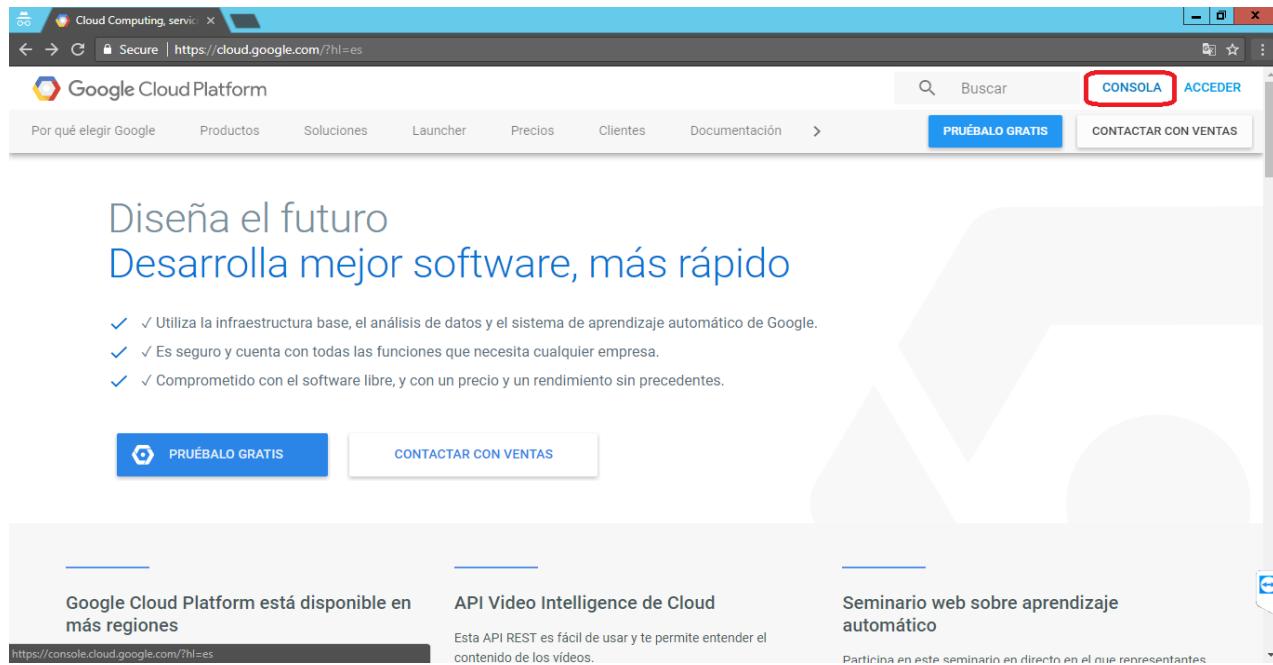


Figura 4-1: Página inicial de Google Cloud Platform

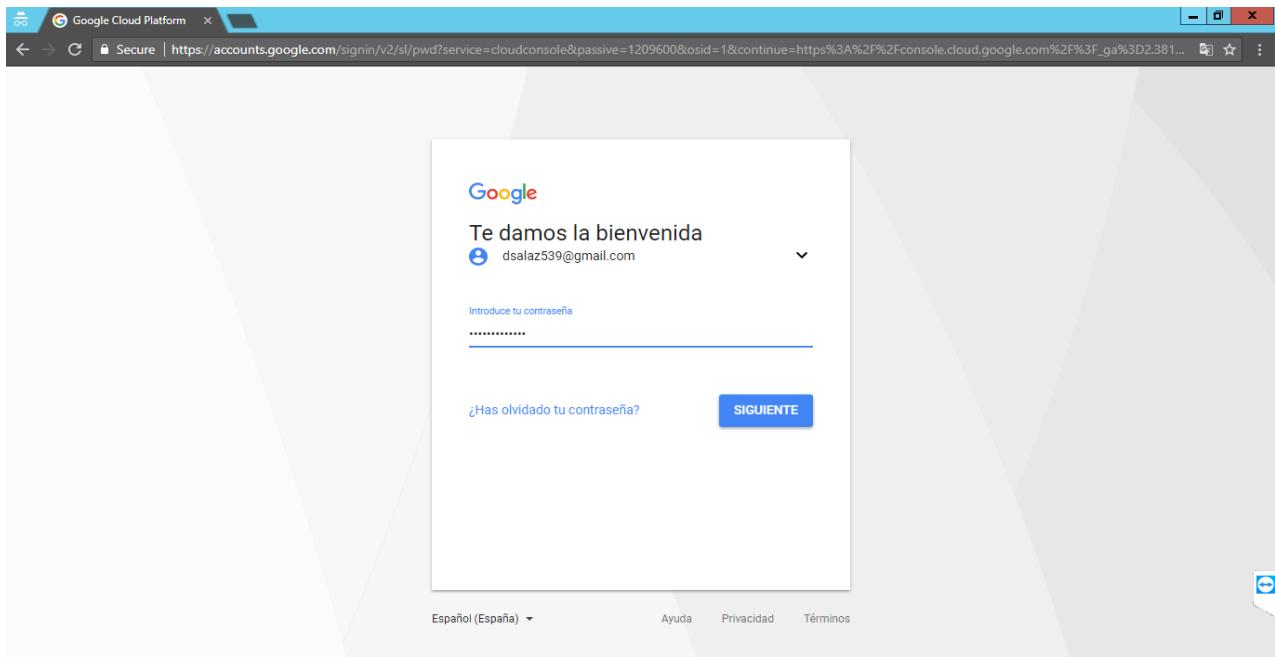


Figura 4-2: Inicio de sesión con la cuenta de Google

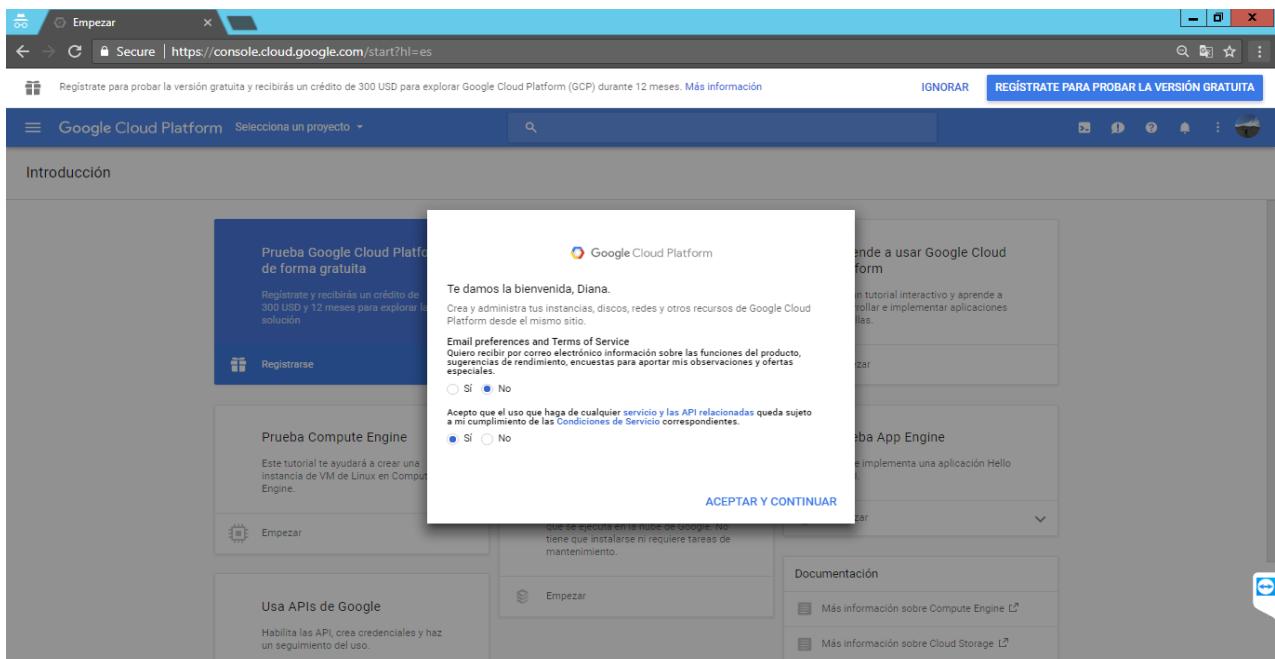


Figura 4-3: Bienvenida a Google Cloud Platform

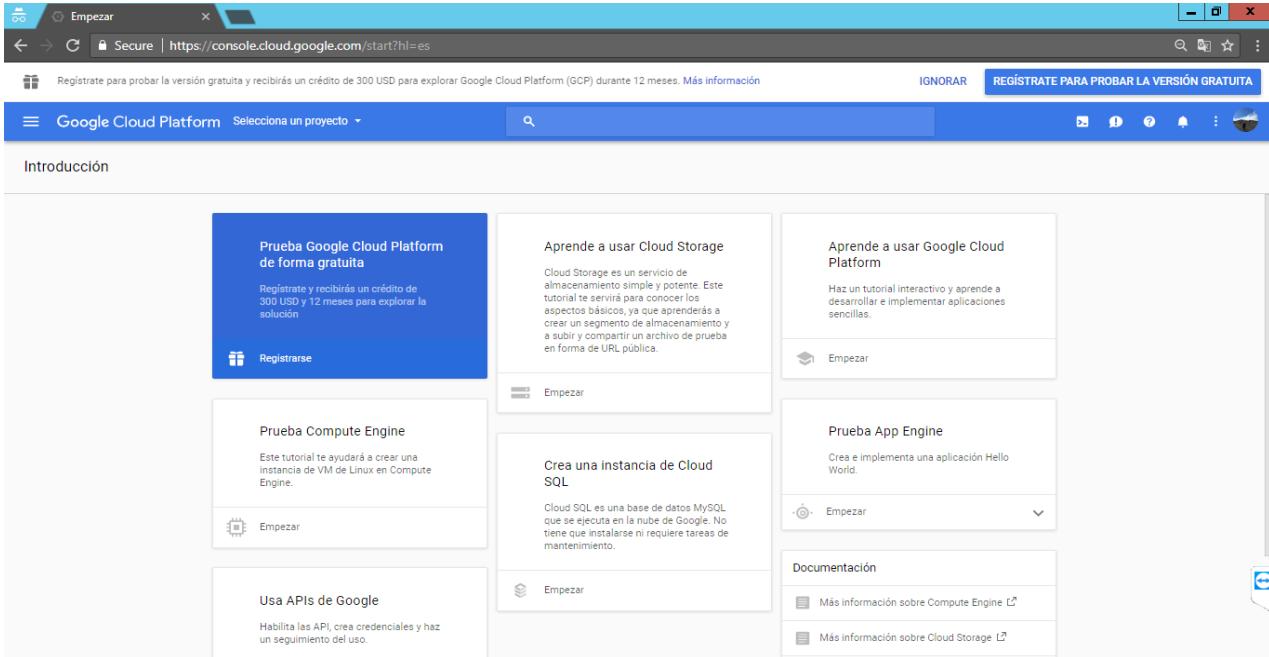


Figura 4-4: Selección de Prueba Gratuita de Google Cloud Platform

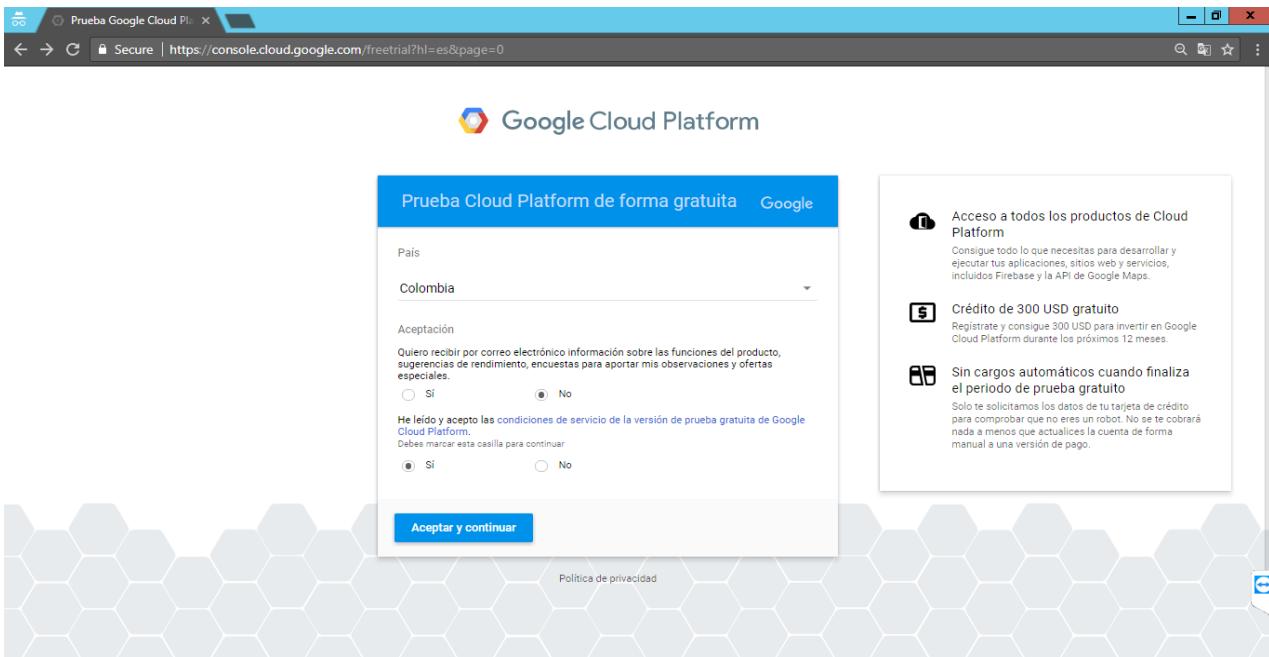


Figura 4-5: Selección de País y Aceptación de condiciones

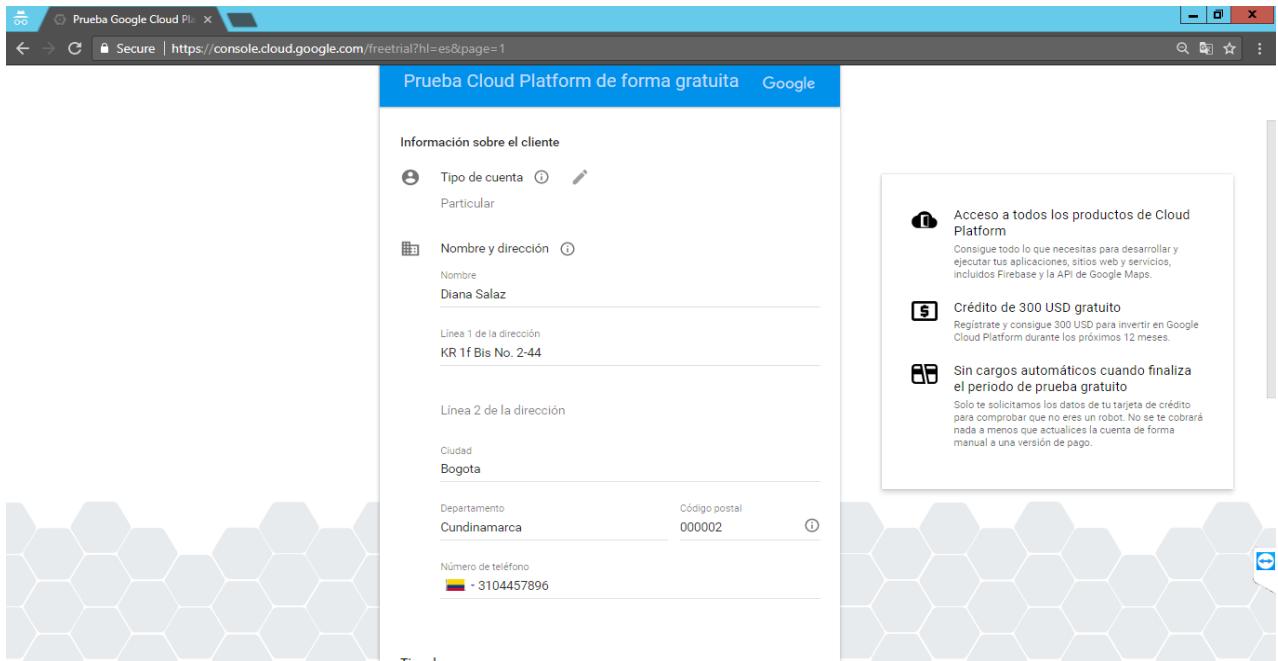


Figura 4-6: Ingreso de datos personales

4.2. Crear máquina virtual en Google Cloud Platform

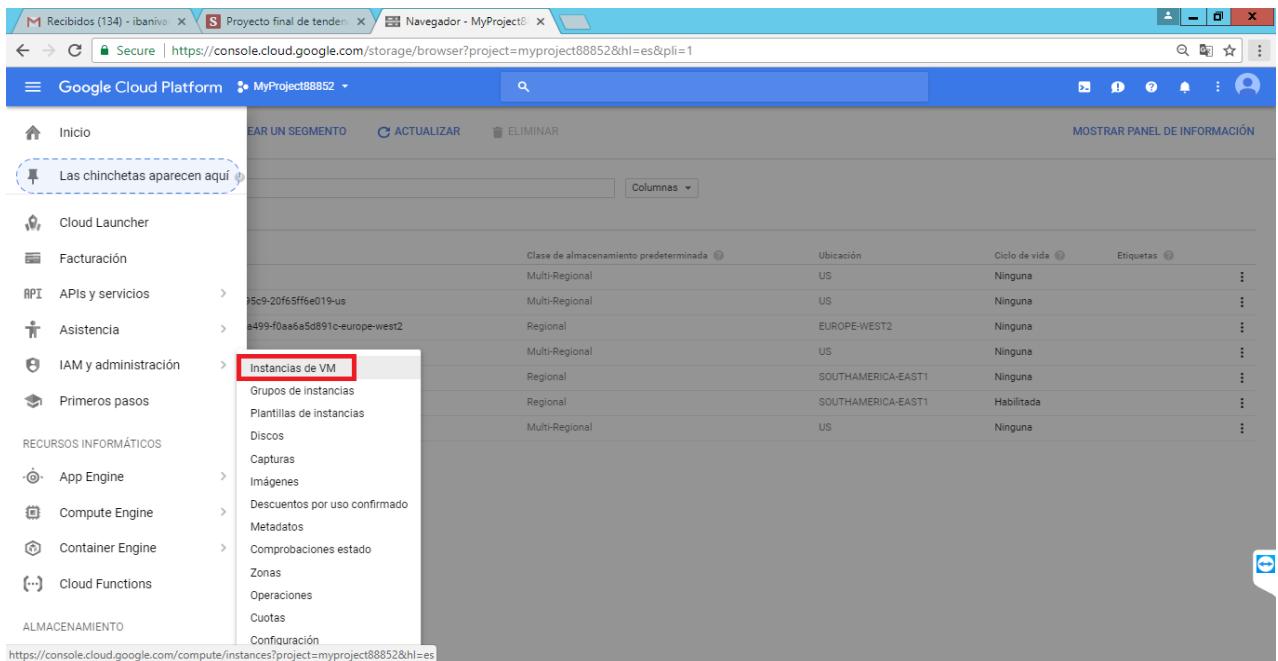


Figura 4-7: Selección de opción de instancias de máquinas virtuales

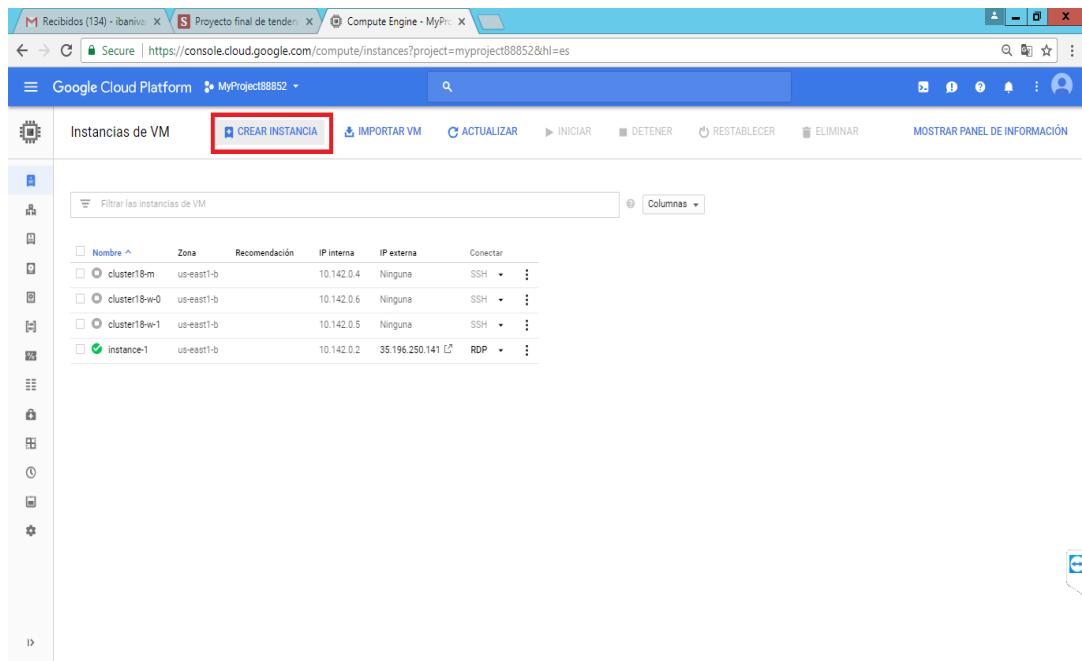


Figura 4-8: Click para crear máquina virtual

Se selecciona el sistema operativo Windows Server 2012 R2, dado que es una versión bastante estable y tenemos experiencia sobre la misma.

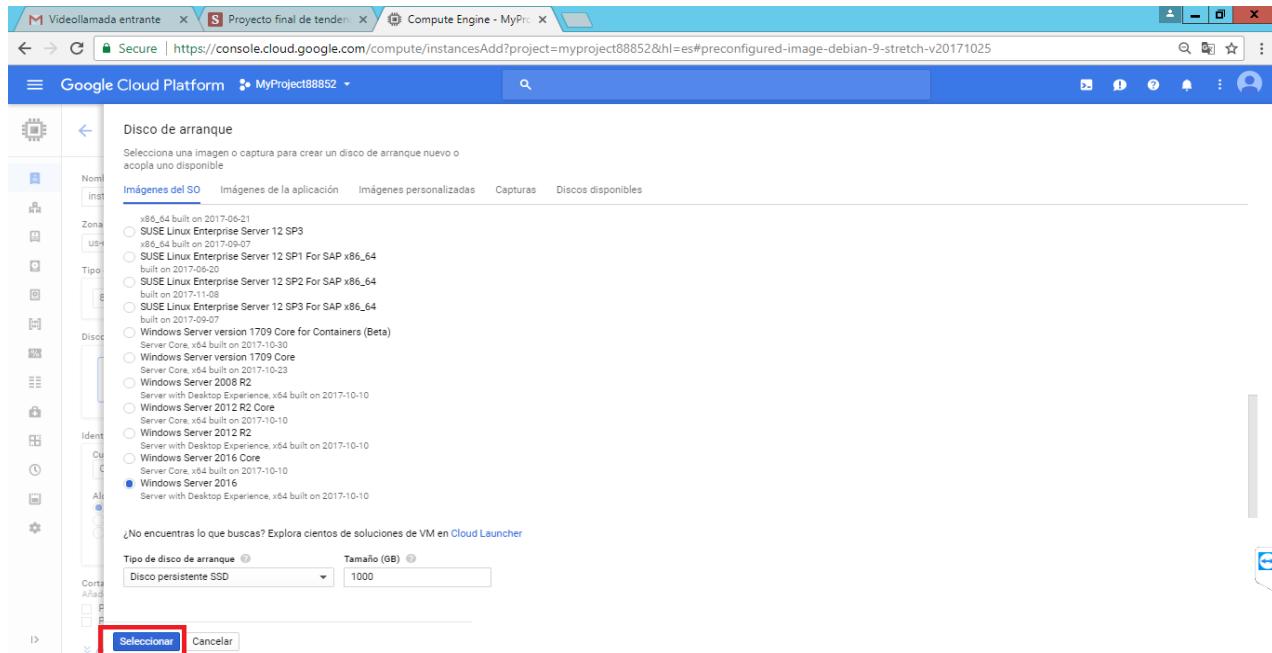


Figura 4-9: Selección de sistema operativo para la máquina virtual

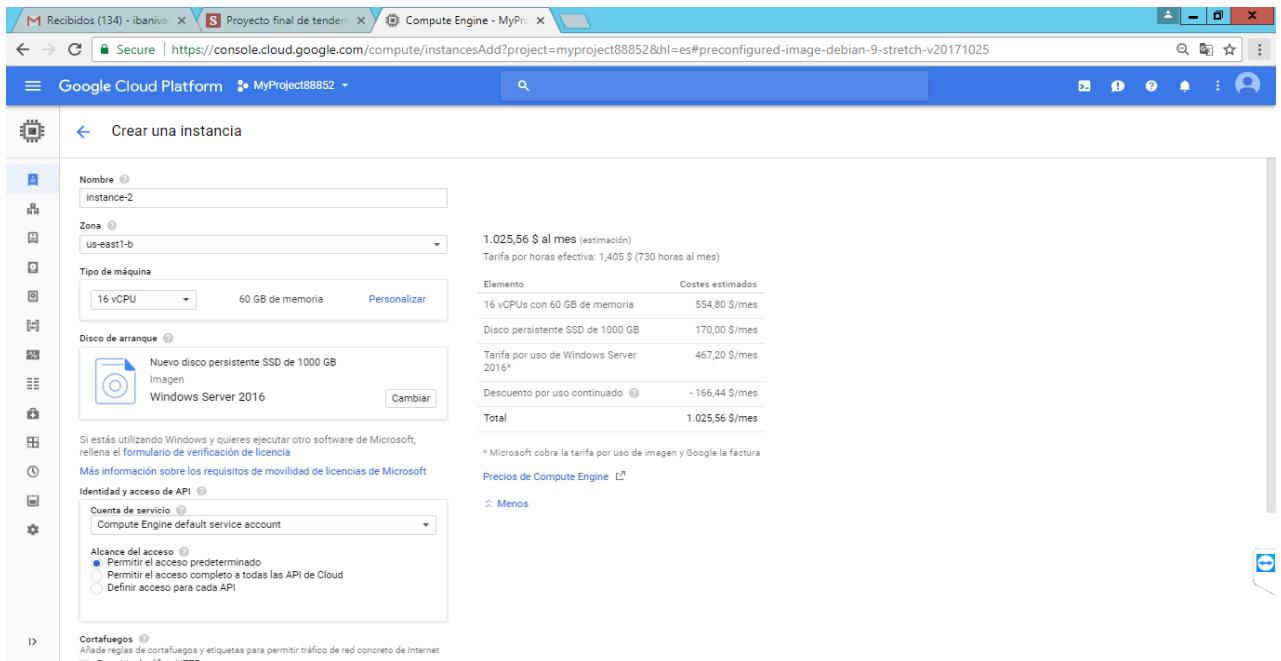


Figura 4-10: Selección de CPUs y memoria RAM

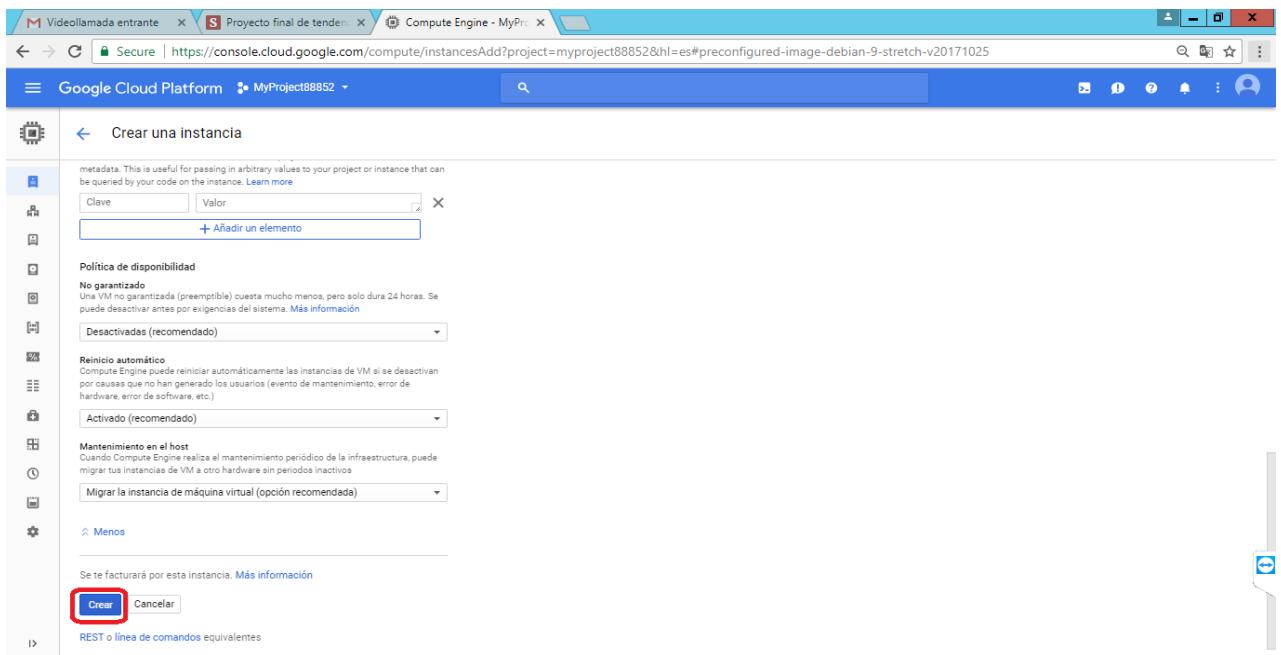


Figura 4-11: Política de disponibilidad y reinicio automático

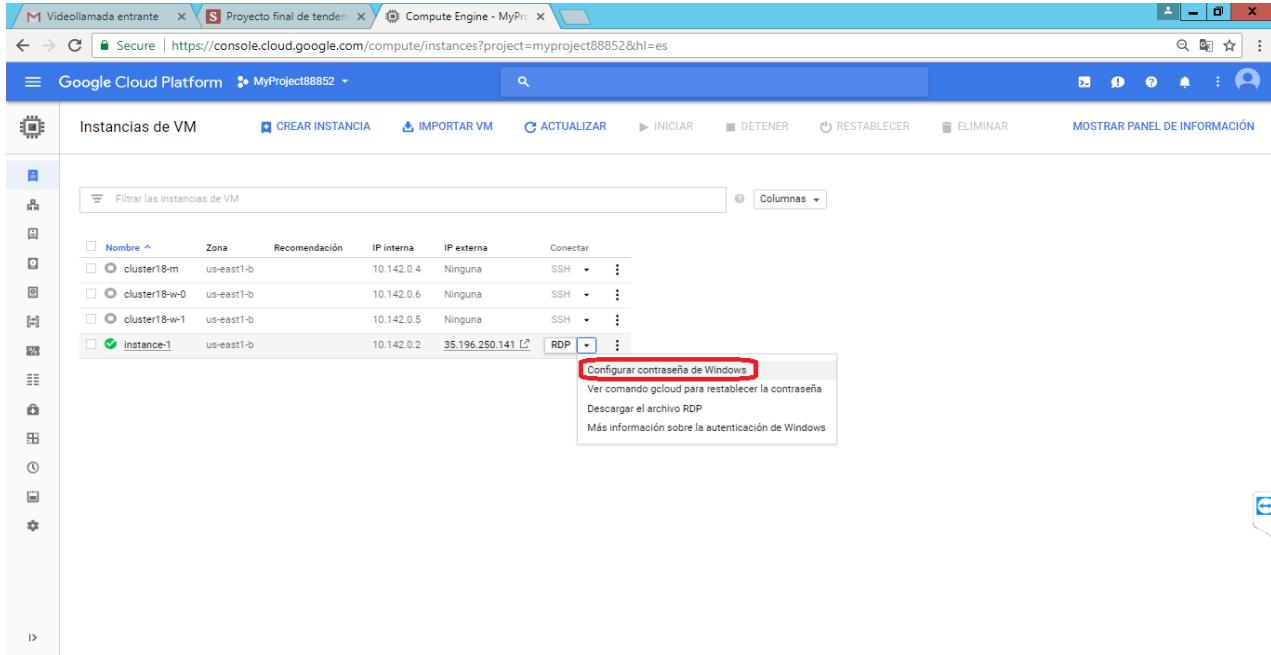


Figura 4-12: Configurar contraseña de Windows

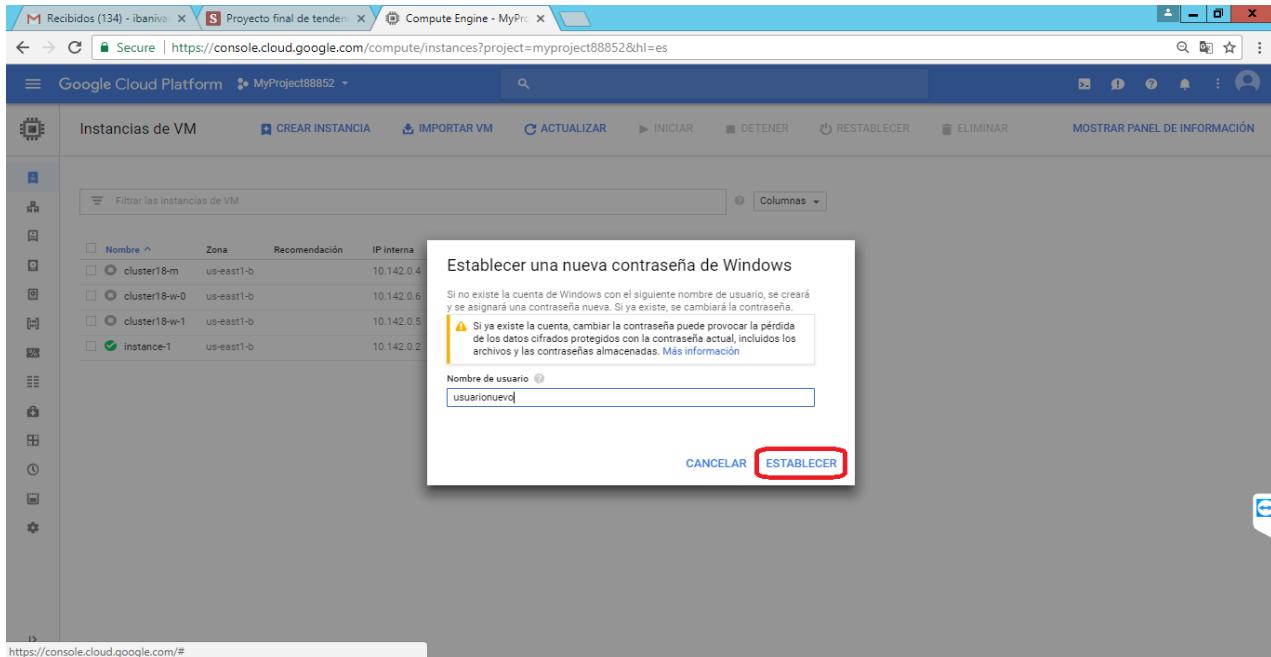


Figura 4-13: Establecer usuario sobre la máquina virtual

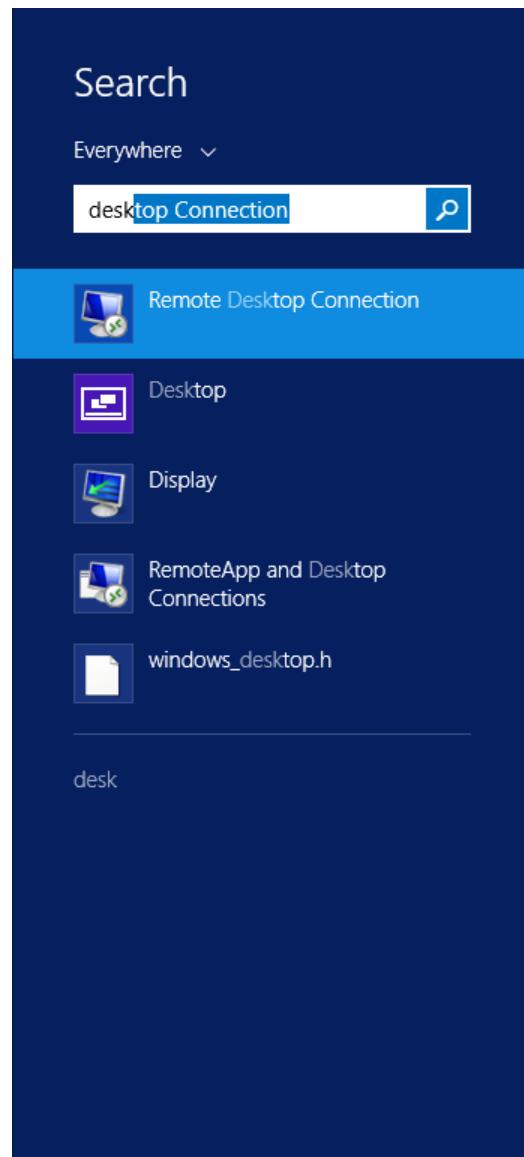


Figura 4-14: Conexión de escritorio remoto de Windows

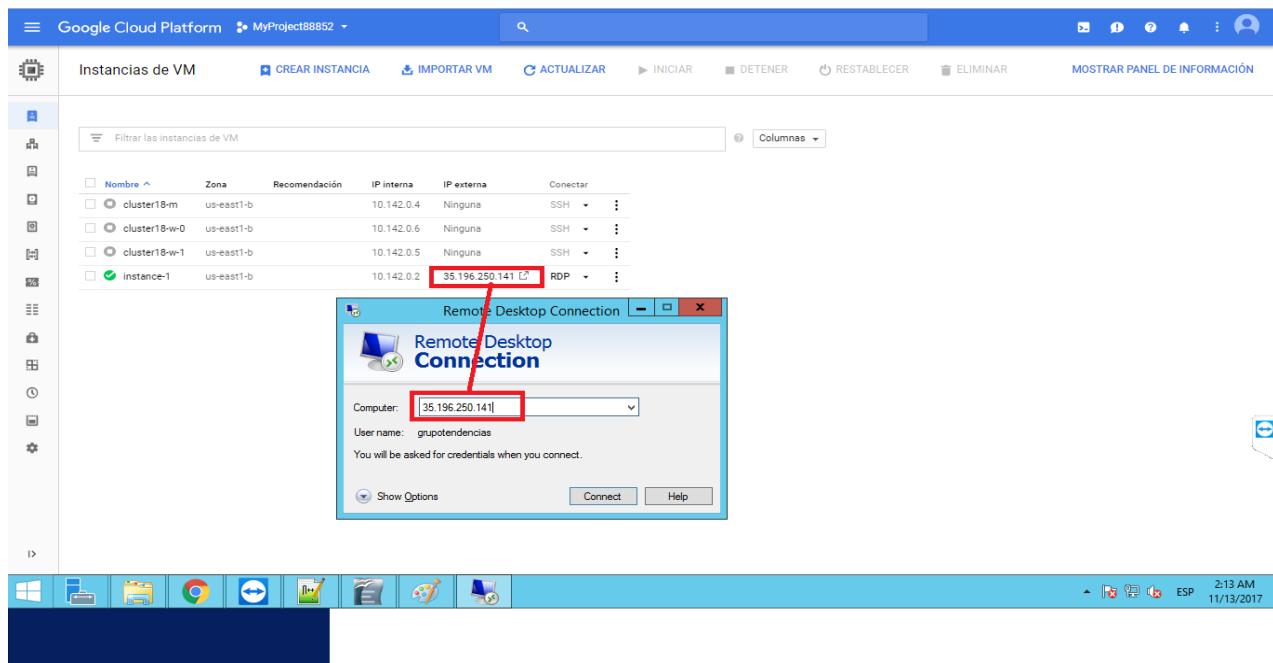


Figura 4-15: Ip externa para acceder remotamente

4.3. Instalaciones en máquina virtual

4.3.1. Instalar Java

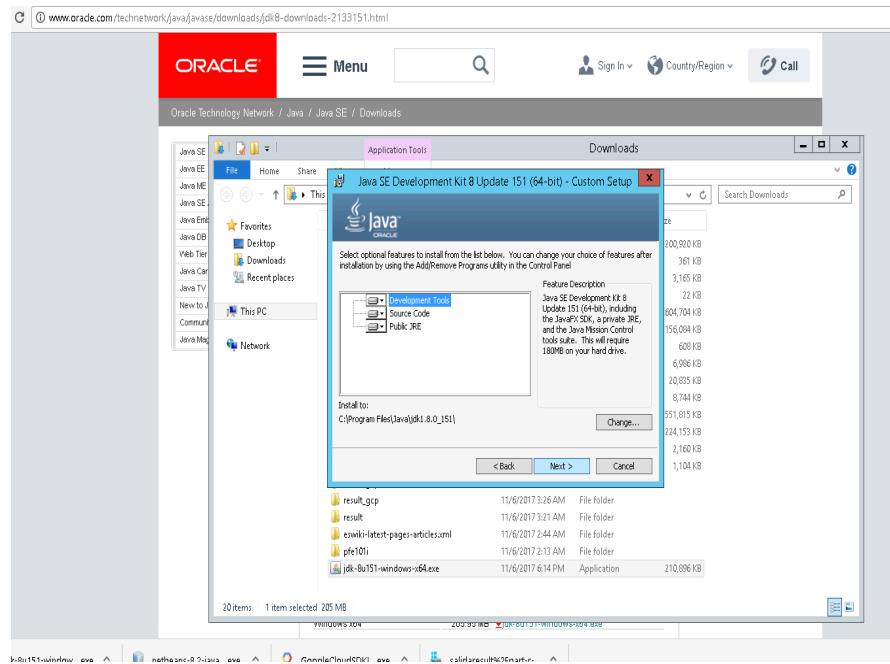


Figura 4-16: Instalación del JDK de Java

4.3.2. Instalar Netbeans

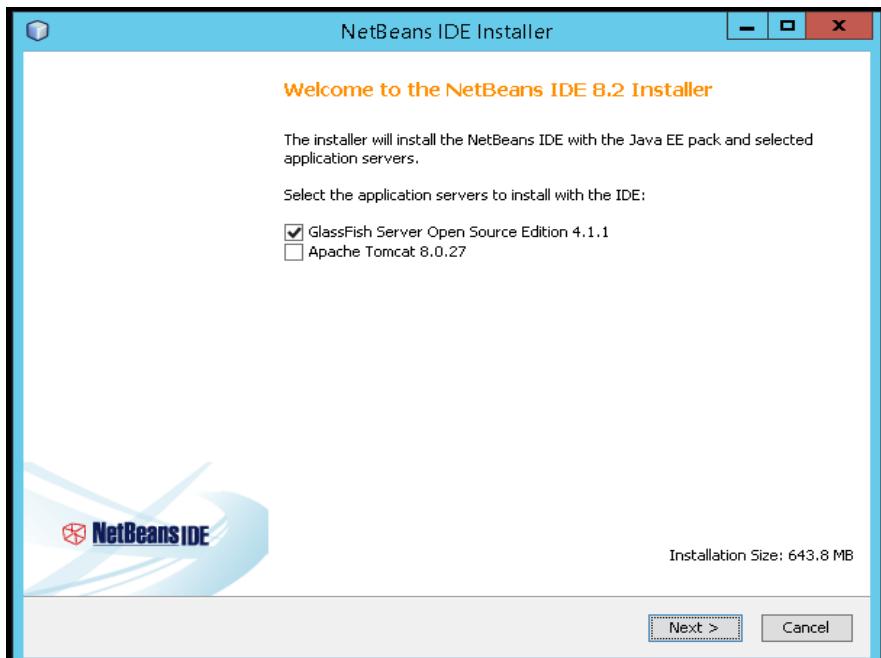


Figura 4-17: Instalación del IDE Netbeans V.8.2

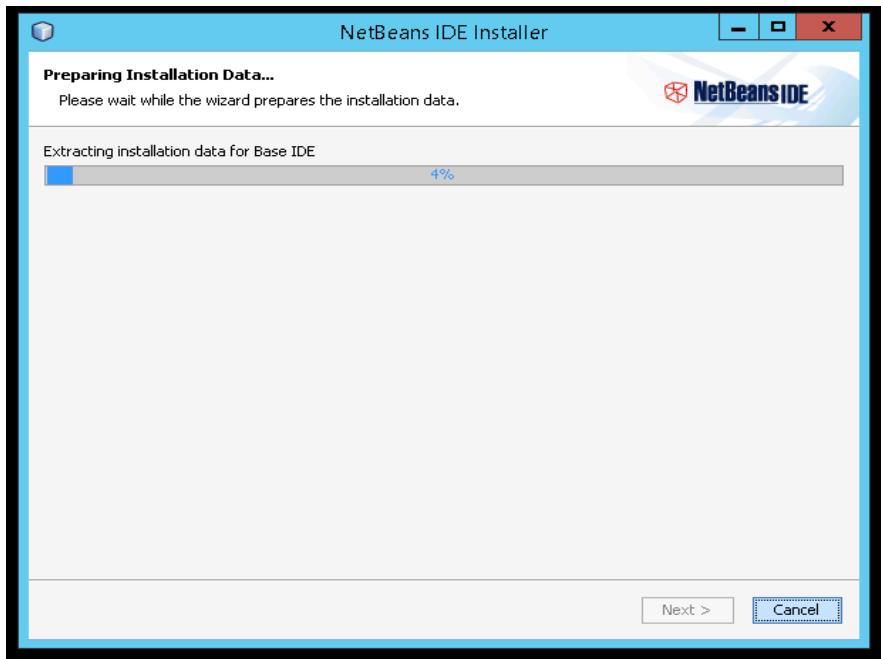


Figura 4-18: Progreso de instalación de Netbeans

4.3.3. Instalar Google SDK

1. Ingresar a la página de Google Cloud Platform y descargar el Google Cloud SDK (figura 4.19)

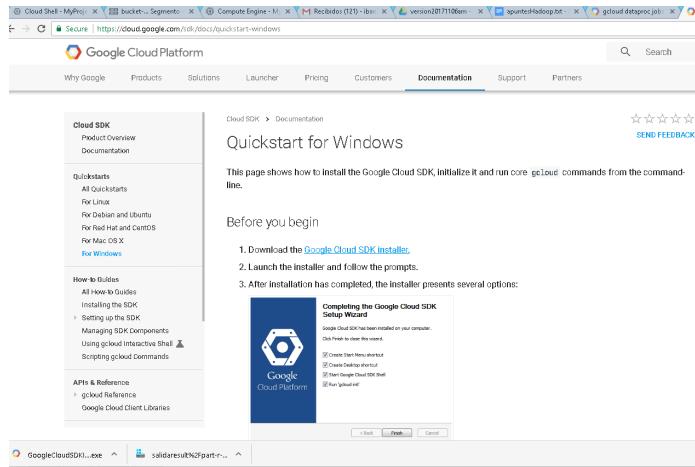


Figura 4-19: Página de descarga en Google Cloud Platform

2. Instalar el Google Cloud SDK (Figura 4.20), el cual contiene herramientas, librerías y administración de recursos sobre Google Cloud (Permite tener localmente un Shell que interactúa con los recursos en la nube).

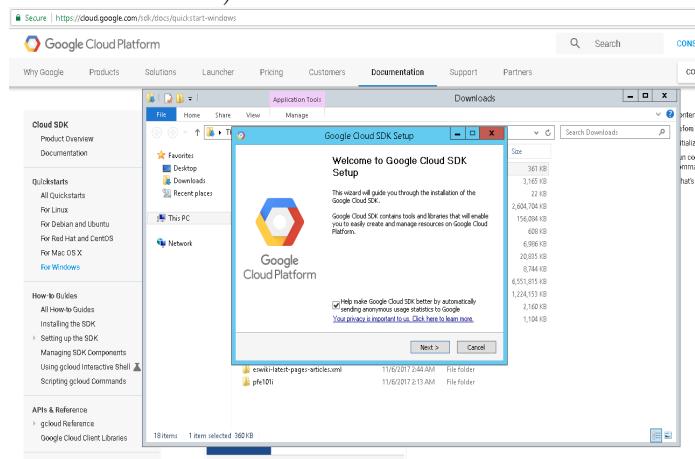


Figura 4-20: Instalación del Google Cloud SDK

3. En el proceso de instalación seleccionar las características y los comandos beta (Figura 4.21).

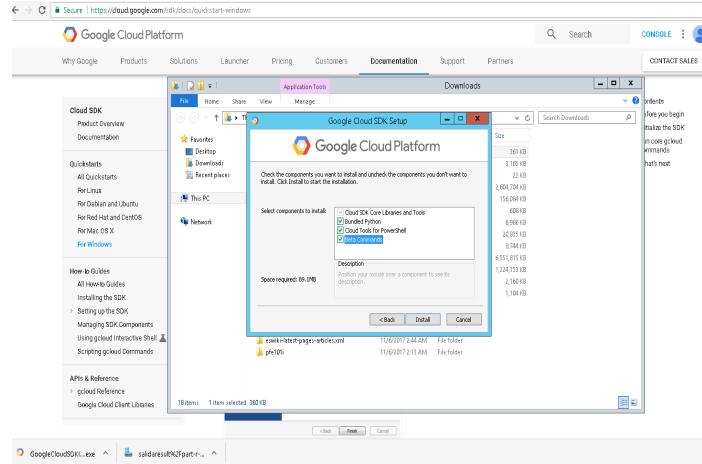


Figura 4-21: Selección de componentes de Google Cloud SDK

En la figura 4.22 se muestran los archivos resultado de la instalación.

The screenshot shows a Windows File Explorer window displaying the contents of the 'google-cloud-sdk' folder located at 'C:\Program Files (x86)\Google\Cloud SDK\google-cloud-sdk'. The table below lists the files and their details.

Name	Date modified	Type	Size
.install	10/10/2017 5:48 PM	File folder	
bin	10/10/2017 5:48 PM	File folder	
deb	10/10/2017 5:48 PM	File folder	
lib	10/10/2017 5:48 PM	File folder	
platform	10/10/2017 5:48 PM	File folder	
rpm	10/10/2017 5:48 PM	File folder	
completion.bash.inc	10/10/2017 5:47 PM	INC File	3 KB
completion.zsh.inc	10/10/2017 5:47 PM	INC File	3 KB
install.bat	10/10/2017 5:47 PM	Windows Batch File	2 KB
install.sh	10/10/2017 5:47 PM	SH File	4 KB
LICENSE	10/10/2017 5:47 PM	File	1 KB
path.bash.inc	10/10/2017 5:47 PM	INC File	1 KB
path.fish.inc	10/10/2017 5:47 PM	INC File	2 KB
path.zsh.inc	10/10/2017 5:47 PM	INC File	1 KB
properties	10/10/2017 5:48 PM	File	1 KB
README	10/10/2017 5:47 PM	File	1 KB
RELEASE_NOTES	10/10/2017 5:47 PM	File	199 KB
VERSION	10/10/2017 5:47 PM	File	1 KB

Figura 4-22: Archivos resultado de la instalación

4. Es necesario arrancar la interacción con Google Cloud; para esto se debe seleccionar la cuenta.



```
Google Cloud SDK Shell
Welcome to the Google Cloud SDK! Run "gcloud -h" to get the list of available commands.

:::Program Files <x86>\Google\Cloud SDK>
:::Program Files <x86>\Google\Cloud SDK>
:::\Program Files <x86>\Google\Cloud SDK>gcloud init_
```

Figura 4-23: Arrancar la interacción con Google Cloud

4.3.4. Instalar SQL Server

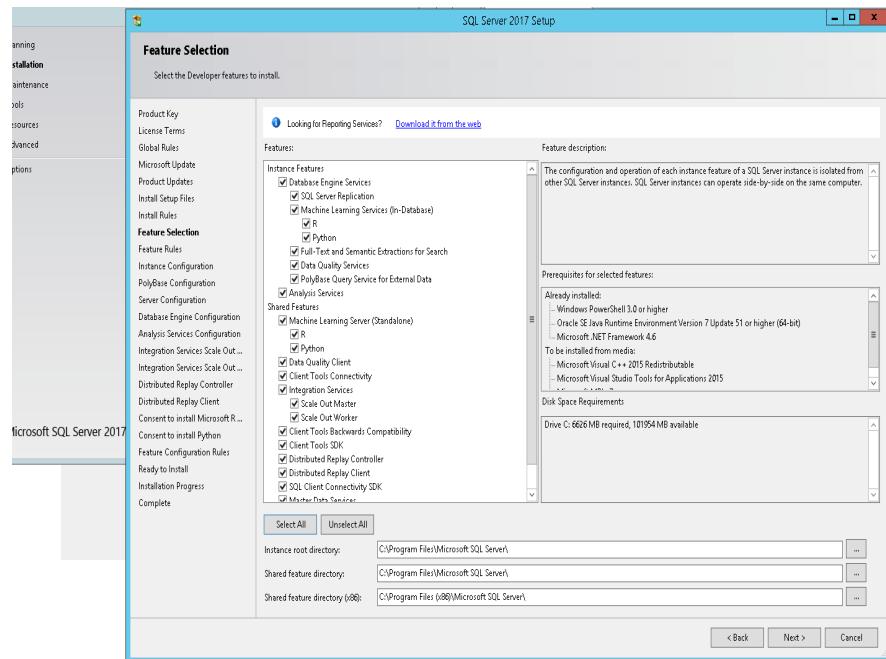


Figura 4-24: Instalación de SQL Server 2017

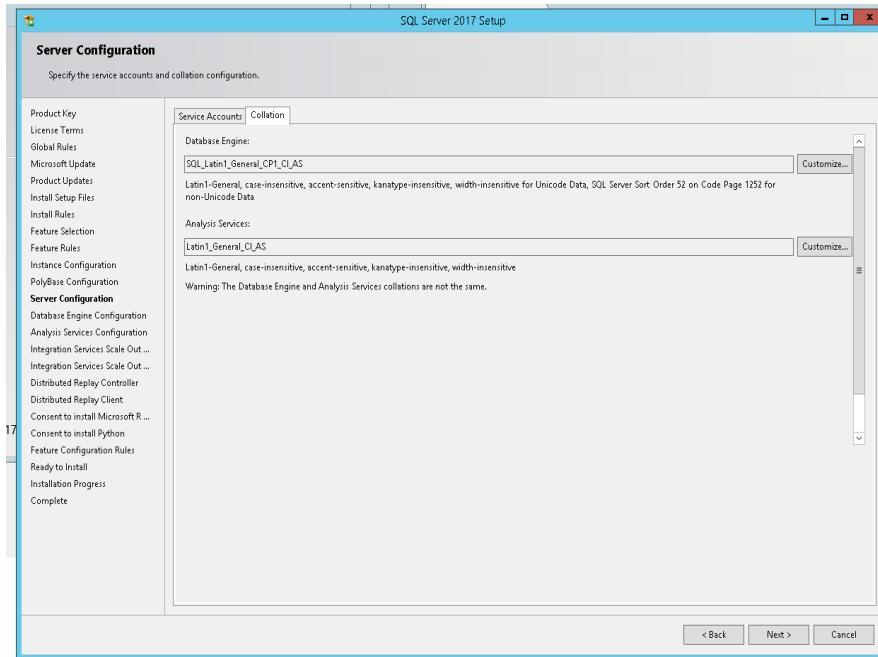


Figura 4-25: Configuración del servidor

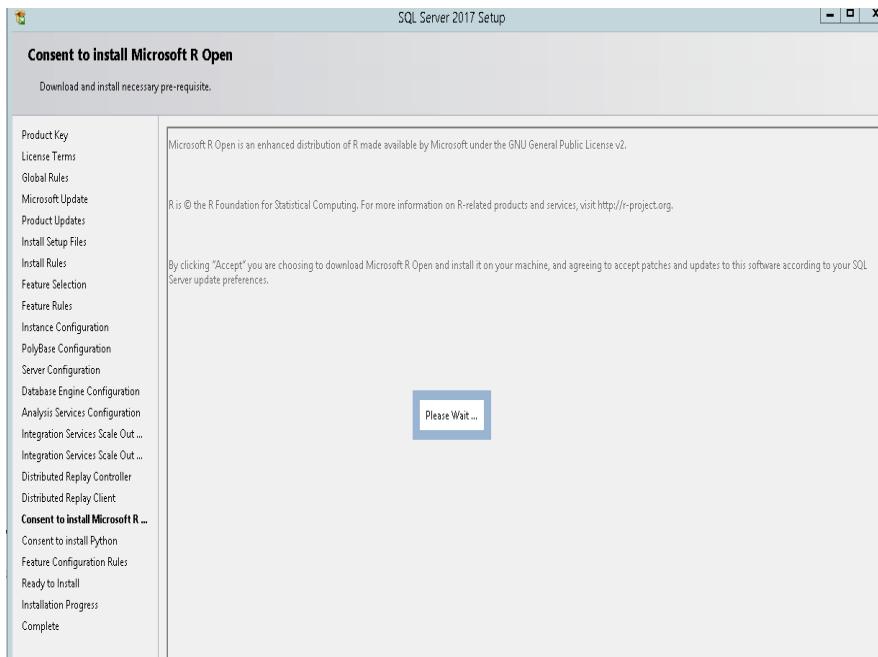


Figura 4-26: Proceso de la instalación de SQL Server 2017

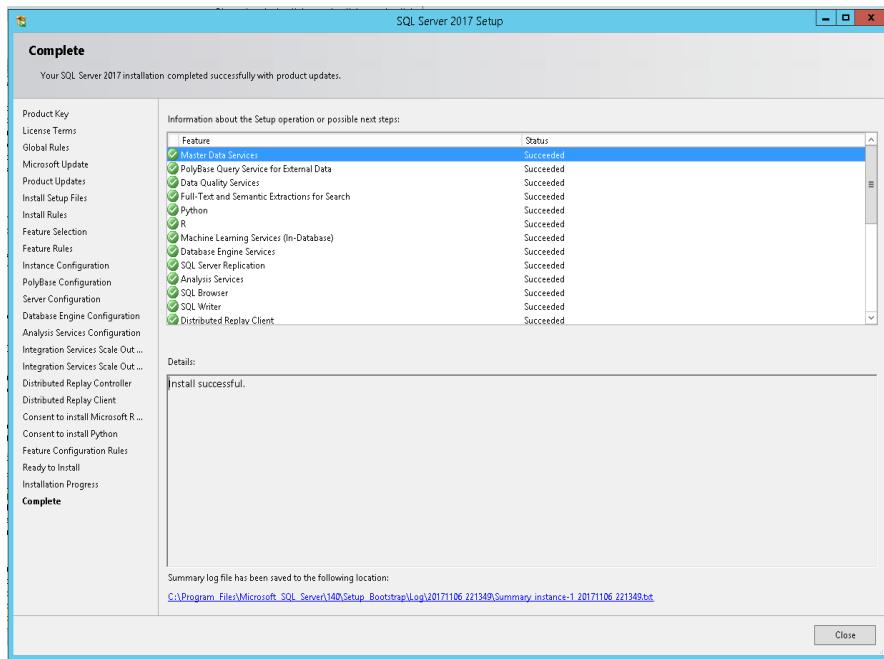


Figura 4-27: Instalación completa de SQL Server

4.3.5. Instalar SQL Server Management Studio

The screenshot shows the Microsoft Docs website for the 'Download SQL Server Management Studio (SSMS)' page. The page includes a sidebar with links to SSMS components like 'What is SSMS?', 'What's New', and 'Download SQL Server Power Shell Module'. The main content area describes SSMS as an integrated environment for managing SQL infrastructure and provides a link to download 'SQL Server Management Studio 17.3'. The right sidebar contains sections for 'Comments', 'Edit', 'Share', 'Theme' (set to 'Light'), and 'In this article' with links to 'SQL Server Management Studio', 'New in this Release', 'Supported SQL offerings', 'Supported Operating systems', 'SSMS installation tips and issues', 'Available Languages', 'Release Notes', 'Previous releases', and 'Feedback'. A 'See Also' section is also present. At the bottom right, there is a 'Is this page helpful?' poll with 'YES' and 'NO' options.

Figura 4-28: Ingreso a la página de SSMS y descarga del instalador

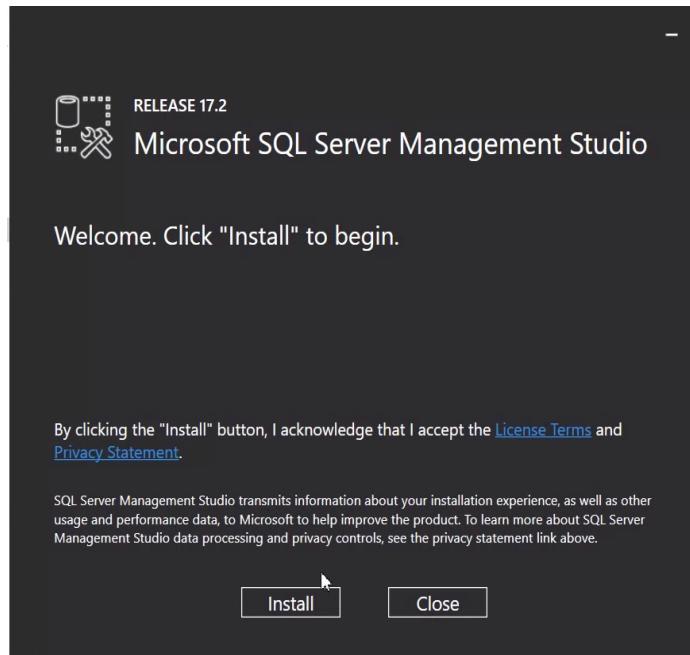


Figura 4-29: Ejecución del instalador de SSMS

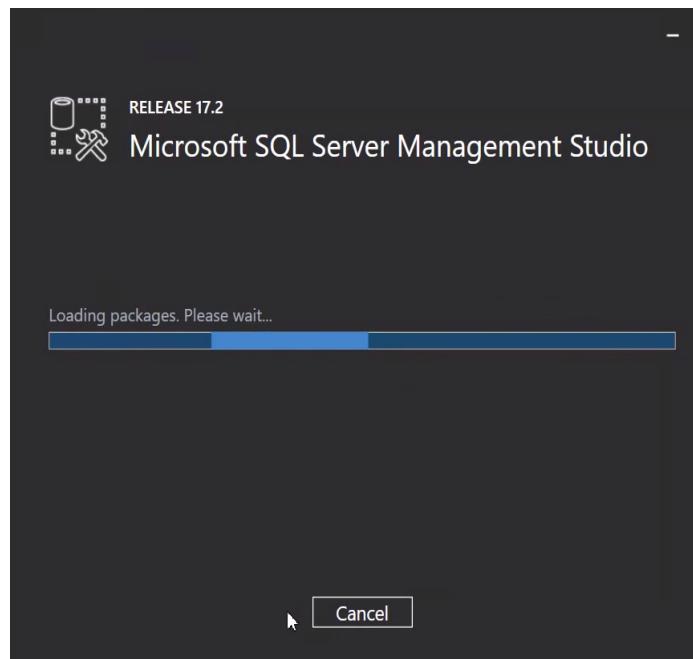


Figura 4-30: Cargue de paquetes de SSMS

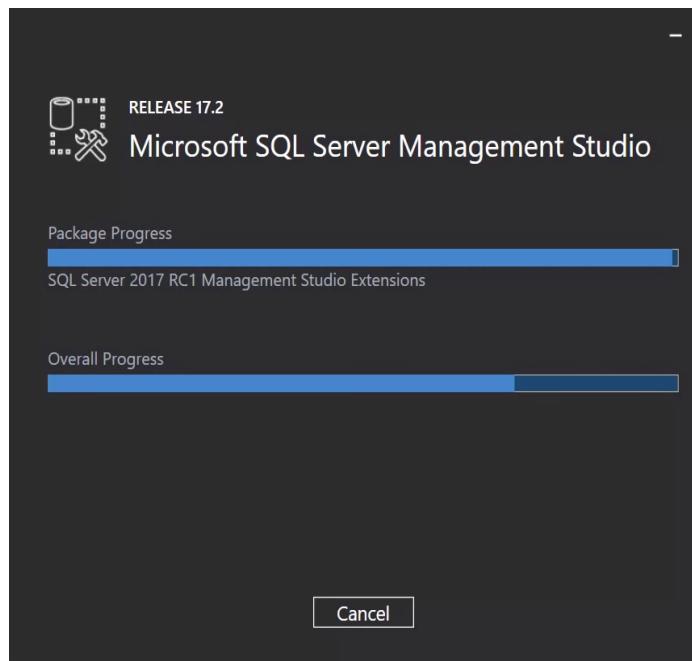


Figura 4-31: Progreso de la instalación de SSMS

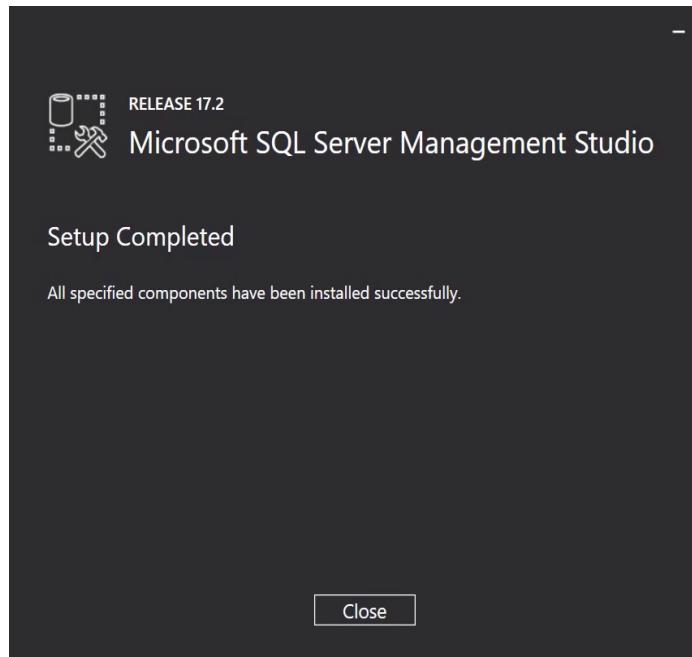


Figura 4-32: Instalación exitosa

4.3.6. Instalar SQL Server Data Tools (SSDT)

Download SQL Server Data Tools (SSDT)

10/19/2017 • 3 minutes to read • Contributors all

SQL Server Data Tools is a modern development tool that you can download for free to build SQL Server relational databases, Azure SQL databases, Integration Services packages, Analysis Services data models, and Reporting Services reports. With SSDT, you can design and deploy any SQL Server content type with the same ease as you would develop an application in Visual Studio.

SSDT for Visual Studio 2017 (15.4.0 preview) is now available. This release introduces a standalone web installation experience for SQL Server Database, Analysis Services, Reporting Services, and Integration Services projects in Visual Studio 2017 15.4 or later.

[SSDT for Visual Studio 2017 \(preview\)](#) [SSDT for Visual Studio 2015](#)

[Download SSDT for Visual Studio 2017 \(15.4.0 preview\)](#) [Download SSDT for Visual Studio 2015 \(17.3\)](#)

Important
Before installing SSDT for Visual Studio 2017 (15.4.0 preview), close all VS instances, and 7
[uninstall the "Microsoft Analysis Services Projects" and "Microsoft Reporting Services"](#)

Comments
Edit
Share
Theme
Light

In this article
[SSDT for Visual Studio 2017](#)
[SSDT for Visual Studio 2015](#)
[Download Visual Studio](#)
[Installing SSDT without Visual Studio pre-installed](#)
[Supported SQL versions](#)
[Next steps](#)
[See Also](#)

Is this page helpful?
YES NO

Figura 4-33: Página principal y Descarga de SSDT

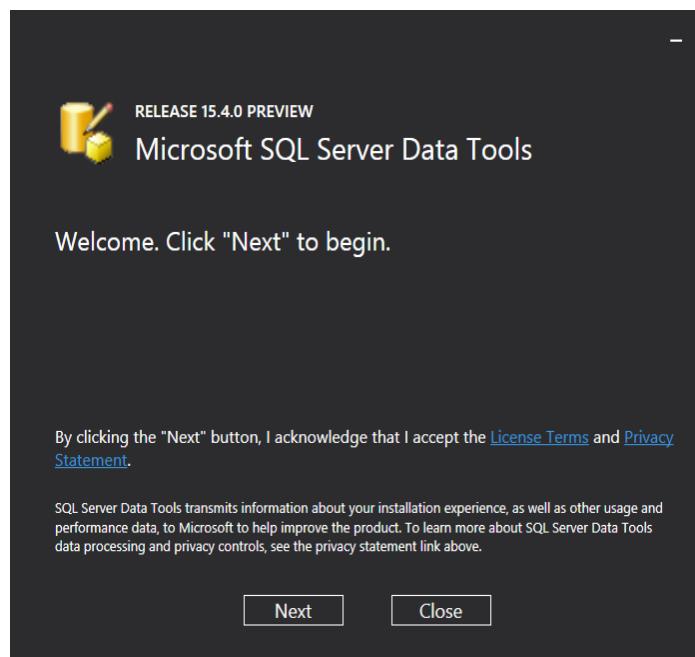


Figura 4-34: Inicio de Instalación de SSDT

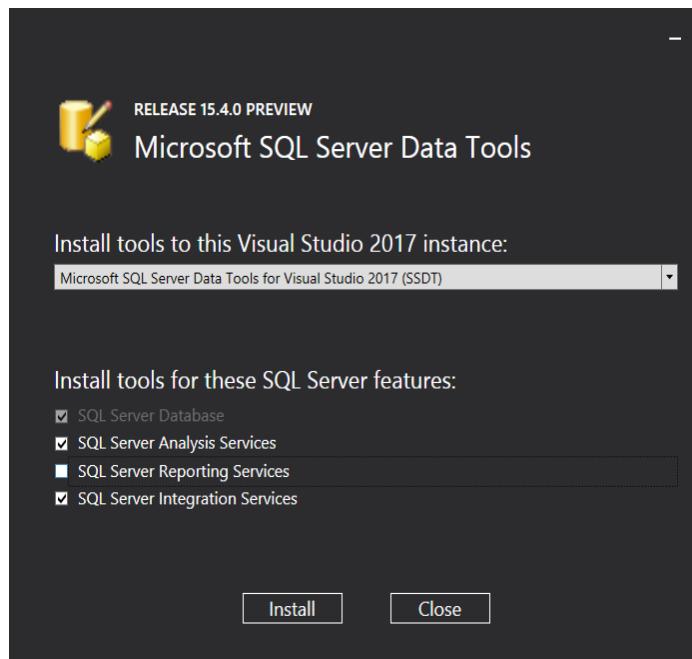


Figura 4-35: Instalación de herramientas de SSDT

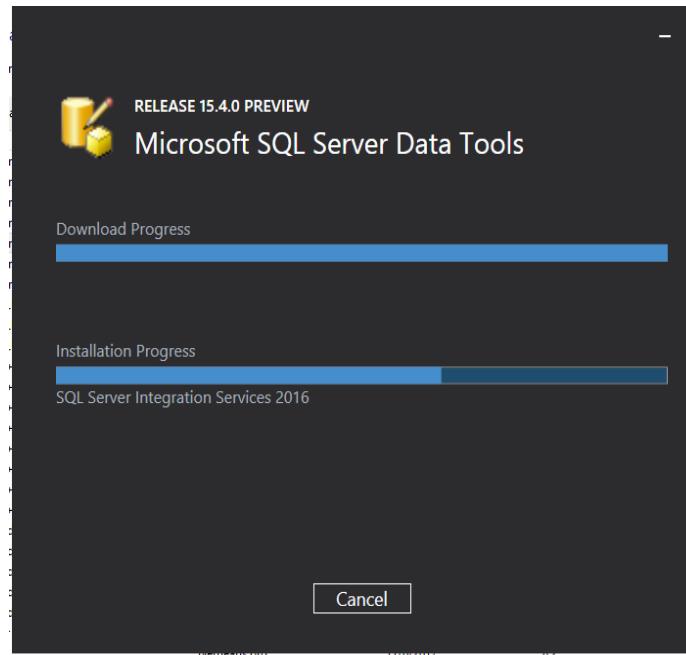


Figura 4-36: Progreso de la instalación de SSDT

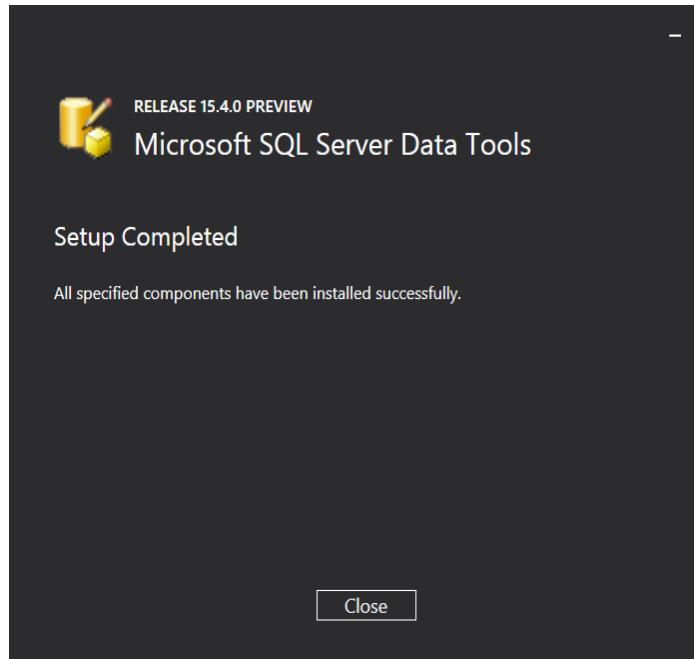


Figura 4-37: Instalación exitosa de SSDT

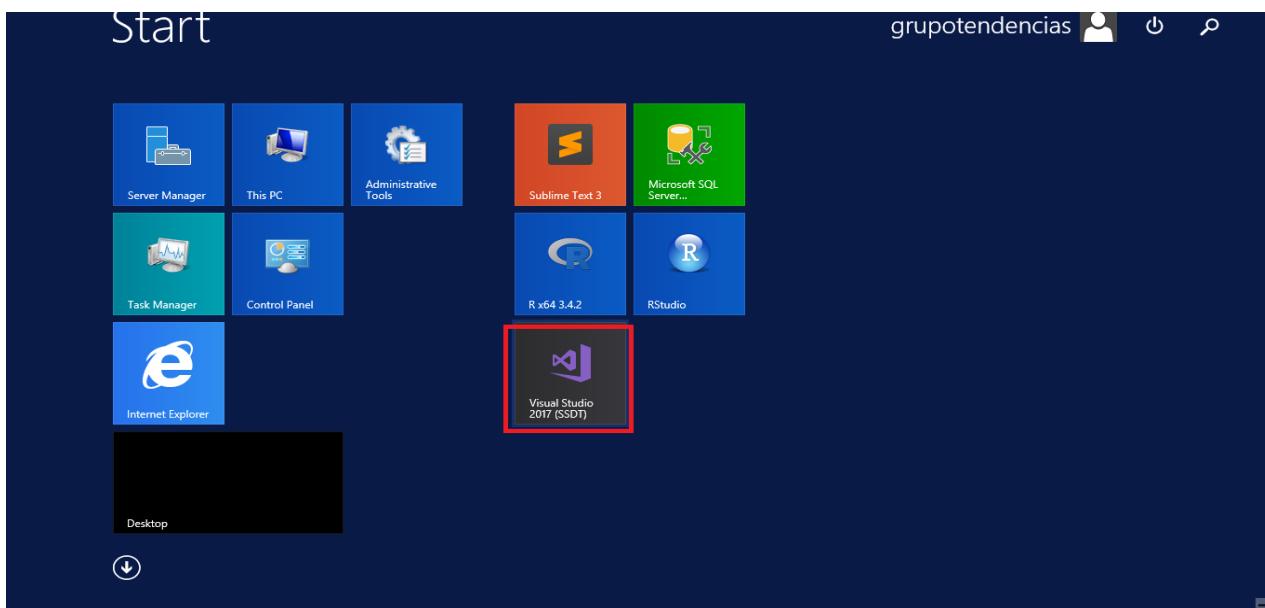


Figura 4-38: Icono de acceso a SSDT

4.3.7. Descargar y descomprimir Hadoop

Se realiza la descarga de Hadoop 2.8.2 desde la página oficial, tal como se muestra en la siguiente imagen.

The screenshot shows the Apache Hadoop Releases page. On the left, there's a sidebar with links like About, Release Versioning, and Documentation. The main content area is titled "Apache Hadoop Releases" and "Download". It states that Hadoop is released as source code tarballs and binary tarballs. A table lists releases from 3.0.0-beta1 to 2.6.5. For each release, it shows the version, release date, tarball type (source or binary), GPG signature, checksum file, and SHA-256 hash. Below the table, instructions for verifying releases using GPG are provided, along with a note about performing a quick check using SHA-256. At the bottom, a link to the Apache release archive is mentioned.

Version	Release Date	Tarball	GPG	SHA-256
3.0.0-beta1	03 October, 2017	source binary	signature signature	checksum file checksum file
2.8.2	24 Oct, 2017	source binary	signature signature	FBA7431B8EC98D07F.. AE499C7C6B441749..
2.7.4	04 August, 2017	source binary	signature signature	DS280CE8446FC10.. 8F791BFCFA3887C7..
2.6.5	08 October, 2016	source binary	signature signature	3A943F1873D9951A.. 0014D1804B6D0FEE..

Figura 4-39: Descarga de Apache Hadoop

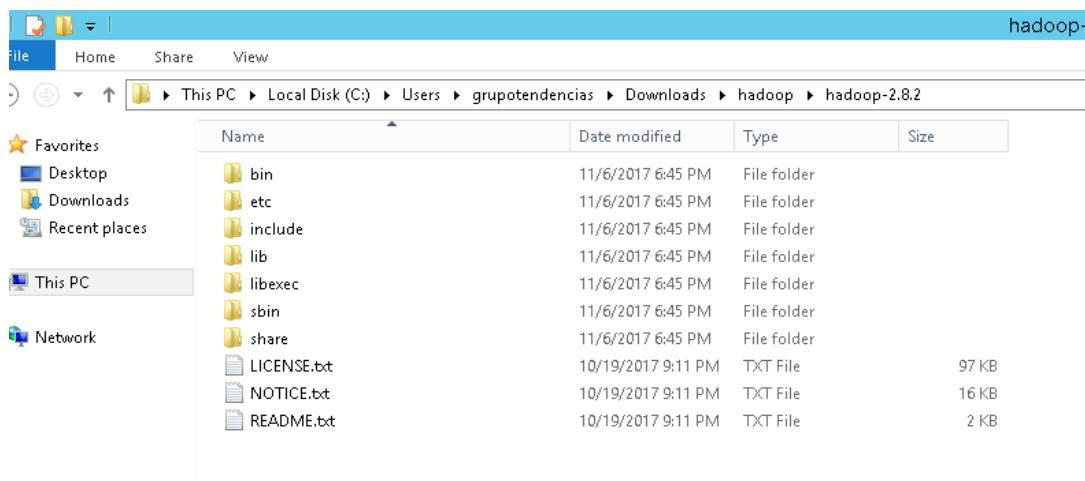


Figura 4-40: Carpeta de Hadoop

En la figura 4.41 se describe cómo se indica a Hadoop dónde está Java sin tener la necesidad de colocar en Windows la variable de entorno; En primer lugar entramos al archivo "hadoop-env.cmd" de la ruta:

C:/Users/grupotendencias/Downloads/hadoop/hadoop-2.8.2/etc/hadoop
y posteriormente se cambia la línea:

set JAVA_HOME= "%JAVA_HOME%"

por la línea:

set JAVA_HOME=C:\ Progra~1\Java\jdk1.8.0_151

```

1  @echo off
2  @rem Licensed to the Apache Software Foundation (ASF) under one or more
3  @rem contributor license agreements. See the NOTICE file distributed with
4  @rem this work for additional information regarding copyright ownership,
5  @rem ASF's ASL license. This file is subject to the terms of the license, Version 2.0
6  @rem (the "License"); you may not use this file except in compliance with
7  @rem the License. You may obtain a copy of the License at
8  @rem
9  @rem     http://www.apache.org/licenses/LICENSE-2.0
10 @rem
11 @rem Unless required by applicable law or agreed to in writing, software
12 @rem distributed under the License is distributed on an "AS IS" BASIS,
13 @rem WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
14 @rem See the License for the specific language governing permissions and
15 @rem limitations under the License.
16
17 @rem Set Hadoop-specific environment variables here.
18
19 @rem The only required environment variable is JAVA_HOME. All others are
20 @rem optional. When running a distributed configuration it is best to
21 @rem set JAVA_HOME in this file, so that it is correctly defined on
22 @rem remote nodes.
23
24 @rem The java implementation to use. Required.
25 set JAVA_HOME=C:\Progra~1\Java\jdk1.8.0_151
26
27 @rem The java implementation to use. Jsvc is required to run secure datanodes.
28 @rem set JSVC_HOME=%JAVA_HOME%
29
30 @rem set HADOOP_CONF_DIR=
31
32 @rem Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
33 if exist %HADOOP_HOME%\contrib\capacity-scheduler (
34   if not defined HADOOP_CLASSPATH (
35     set HADOOP_CLASSPATH=%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
36   ) else (
37     set HADOOP_CLASSPATH=%HADOOP_CLASSPATH%;%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
38   )
39 )

```

Figura 4-41: Asignación de la variable JAVA HOME

4.4. Trabajo en la máquina virtual

4.4.1. Realizar programa en Java con Netbeans

Se procede a crear un nuevo proyecto en Netbeans de tipo Java Application

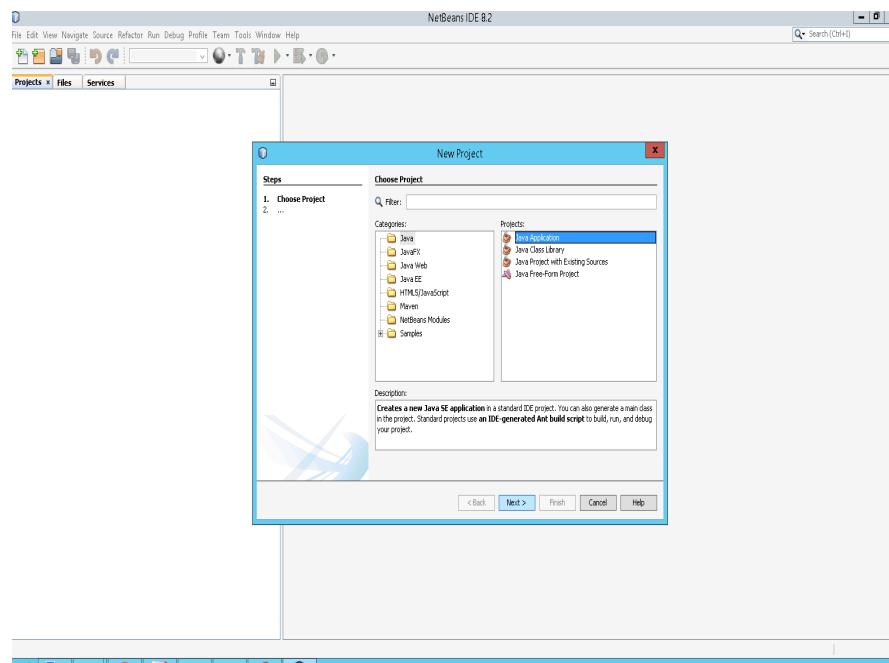


Figura 4-42: Creación de una aplicación java

Se crea un proyecto llamado Hadoop1.

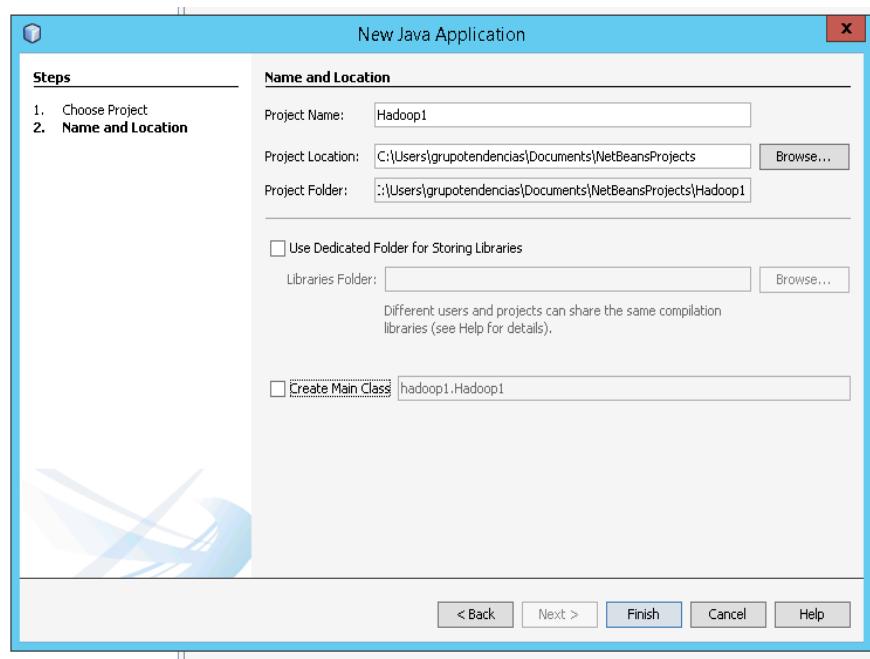


Figura 4-43: Asignación del nombre y ubicación del proyecto

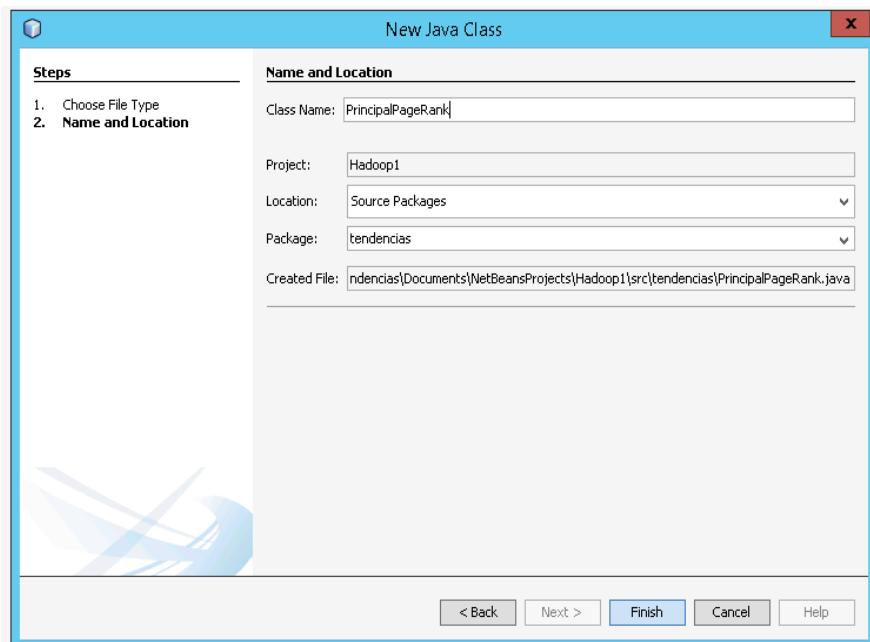


Figura 4-44: Nombre y ubicación de la clase

4.4.2. Generar archivo Jar con Netbeans

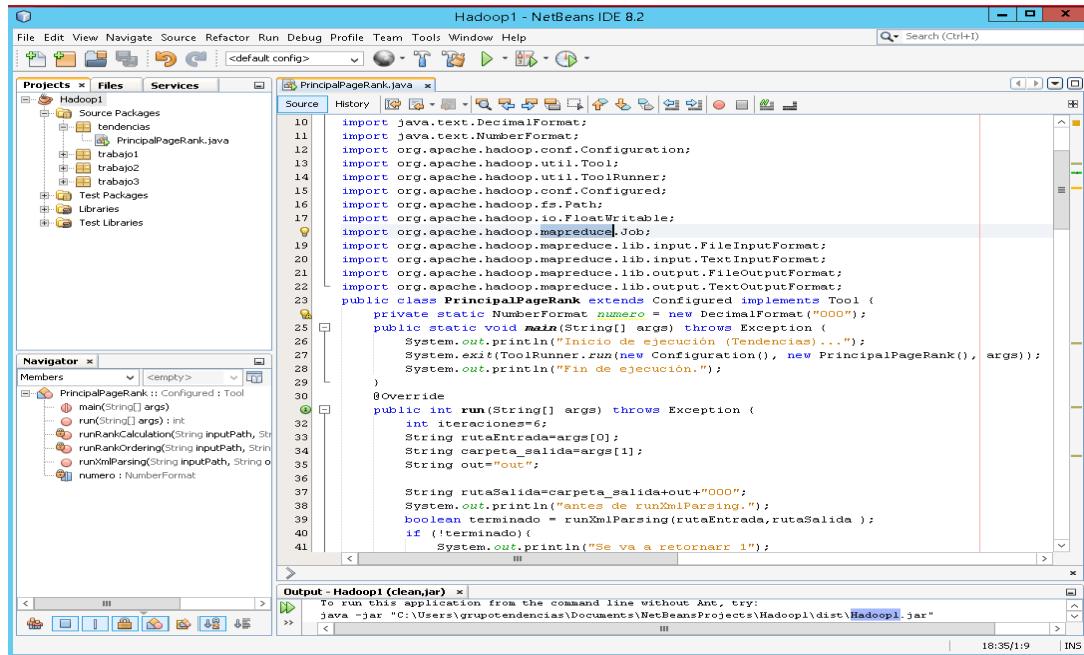


Figura 4-45: Generación de código en Java y Compilación

4.4.3. Agregar librerías con Netbeans

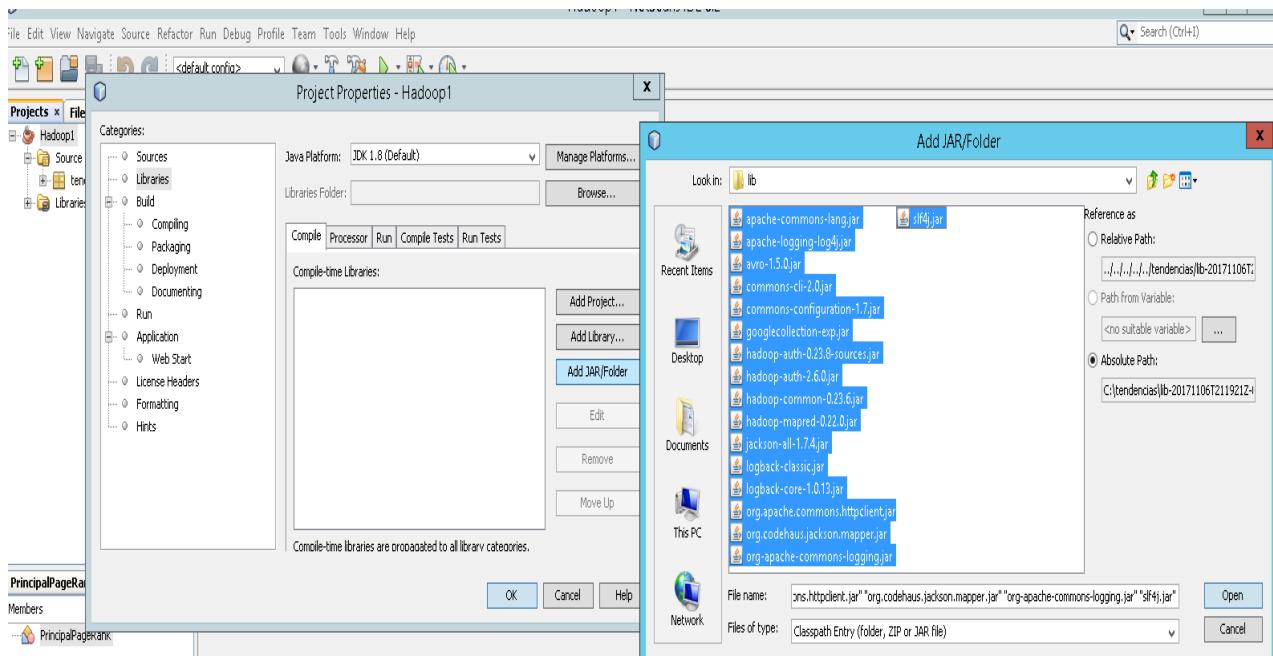
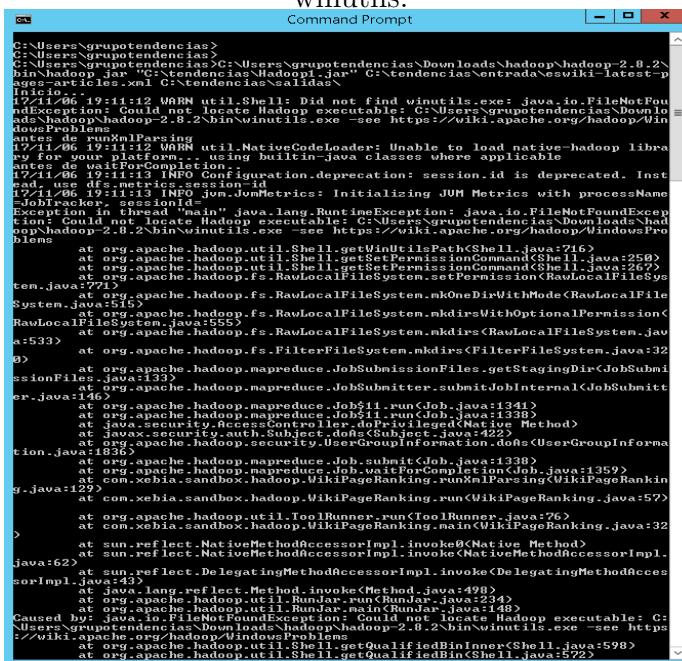


Figura 4-46: Agregar Librerías del Proyecto

4.4.4. Ejecutar Jar con Hadoop

La figura 4.47 muestra un error generado en la ejecución de hadoop relacionado con winutils.



```

C:\Users\grupotendencias>
C:\Users\grupotendencias>C:\Users\grupotendencias\Downloads\hadoop\hadoop-2.8.2\bin\hadoop jar C:\Users\grupotendencias\Downloads\hadoop-2.8.2\bin\hadoop.jar com.xebia.hadoop.WikiPageRanking C:\tendencias\corredateswiki\latest\pages-articles.xml C:\tendencias\salidas\

17/11/06 19:11:12 WARN util.Shell: Did not find winutils.exe: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Users\grupotendencias\Downloads\hadoop\hadoop-2.8.2\bin\winutils.exe - see https://wiki.apache.org/hadoop/WindowsProblems
antes de runXmlParsing
17/11/06 19:11:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
antes de waitForCompletion
17/11/06 19:11:13 INFO mapreduce.JobSubmissionHandler$JobSubmission.deprecation: session_id is deprecated. Instead use dfs.metrics.session_id
17/11/06 19:11:13 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=hadoop-task
Exception in thread "main" java.lang.RuntimeException: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Users\grupotendencias\Downloads\hadoop\hadoop-2.8.2\bin\winutils.exe - see https://wiki.apache.org/hadoop/WindowsProblems
        at org.apache.hadoop.util.Shell.getWinUtilsPath(Shell.java:716)
        at org.apache.hadoop.util.Shell.(Shell.java:259)
        at org.apache.hadoop.util.Shell.setPermissionCommand(Shell.java:267)
        at org.apache.hadoop.fs.RawLocalFileSystem.setPermission(RawLocalFileSystem.java:146)
        at org.apache.hadoop.fs.RawLocalFileSystem.mkOneDirWithMode(RawLocalFile
System.java:515)
        at org.apache.hadoop.fs.RawLocalFileSystem.mkdirsWithOptionalPermission(RawLocalFileSystem.java:555)
        at org.apache.hadoop.fs.RawLocalFileSystem.mkdirs(RawLocalFileSystem.java:533)
        at org.apache.hadoop.fs.FilterFileSystem.mkdirs(FilterFileSystem.java:32
0)
        at org.apache.hadoop.mapreduce.JobSubmissionFiles.getStagingDir(JobSubmissionFiles.java:133)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:146)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1341)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1338)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:744)
Caused by: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Users\grupotendencias\Downloads\hadoop\hadoop-2.8.2\bin\winutils.exe - see https://wiki.apache.org/hadoop/WindowsProblems
        at org.apache.hadoop.util.Shell.getQualifiedBinInner$Shell.java:598>
        at org.apache.hadoop.util.Shell.getQualifiedBin$Shell.java:572>

```

Figura 4-47: Error generado en hadoop

4.4.5. Agregar winutils a Hadoop

Para resolver el inconveniente con winutils se descargó de la página de GitHub (Figura 4.48)

steveloughran sign Hadoop artifacts		
Latest commit 2878787 on Aug 29		
..		
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

Figura 4-48: Descarga de winutils desde Github

Se copiaron los archivos en la carpeta de Hadoop

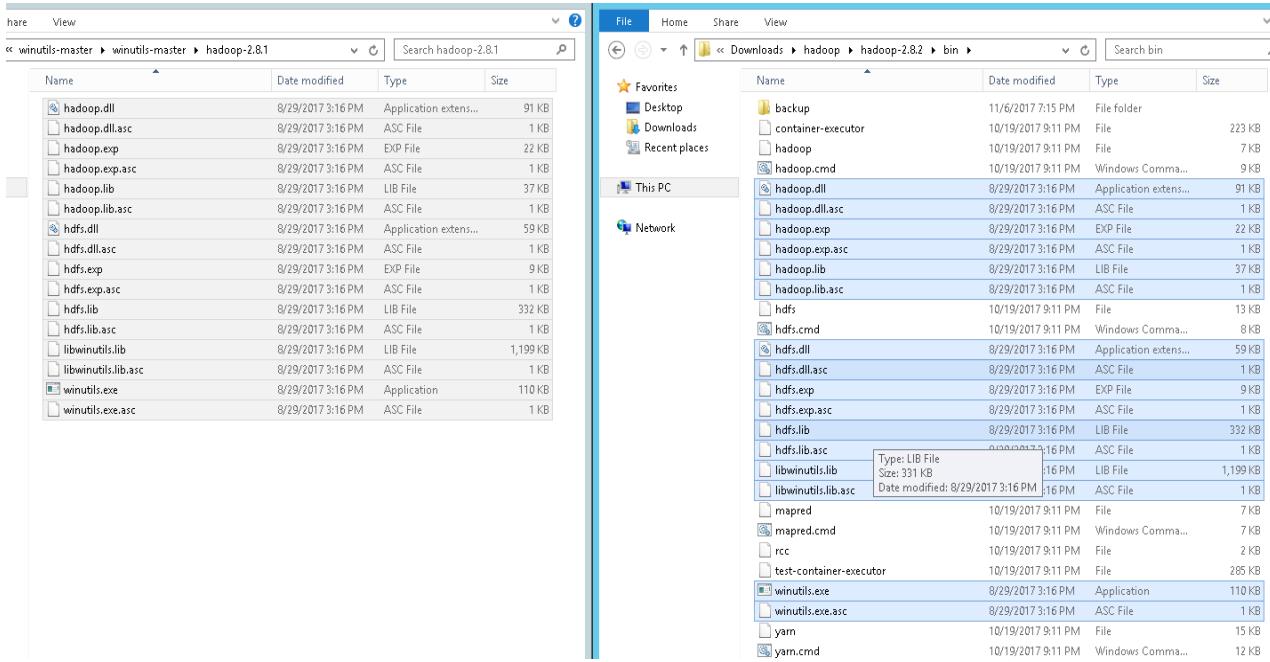


Figura 4-49: Paso de archivos descomprimidos de winutils a la carpeta de hadoop

Se genera un nuevo error.

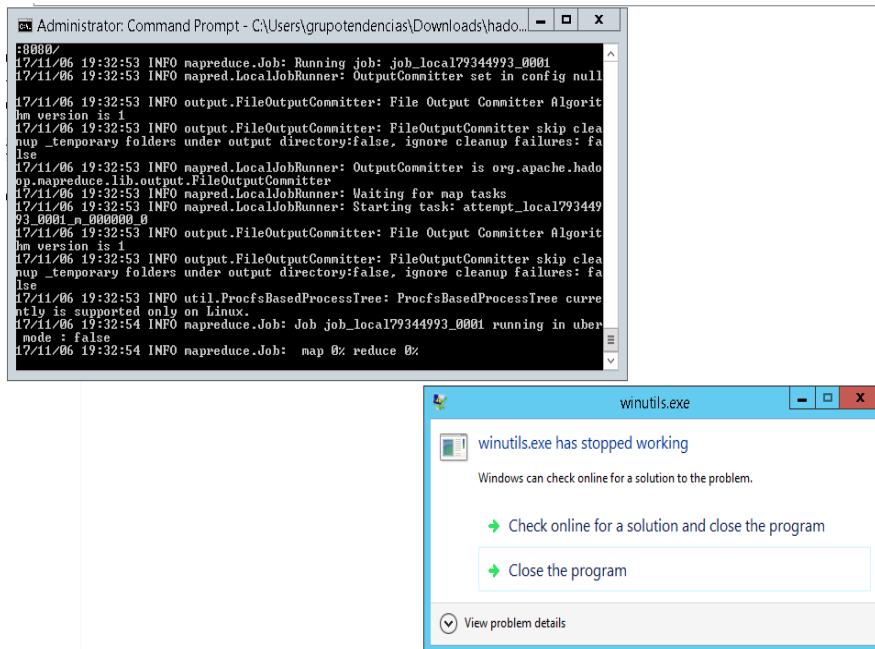


Figura 4-50: Error generado en hadoop

Se instala Microsoft Visual C++ para corregir el error.

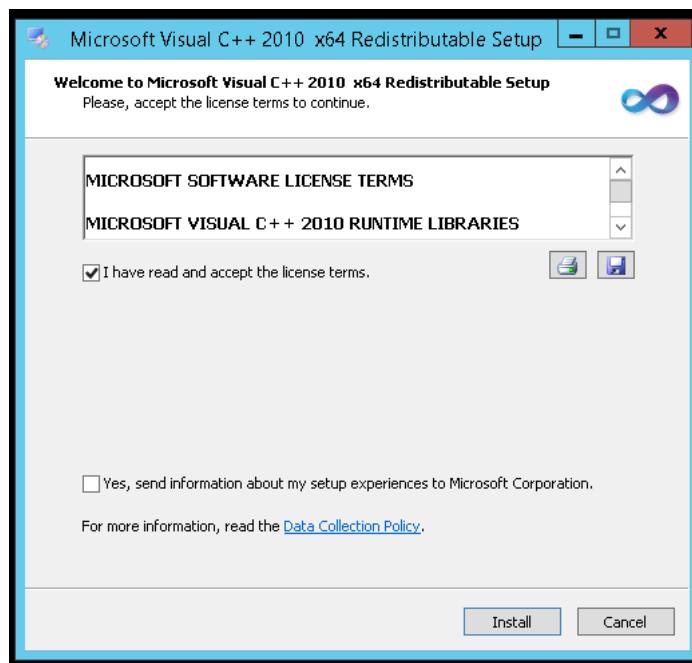
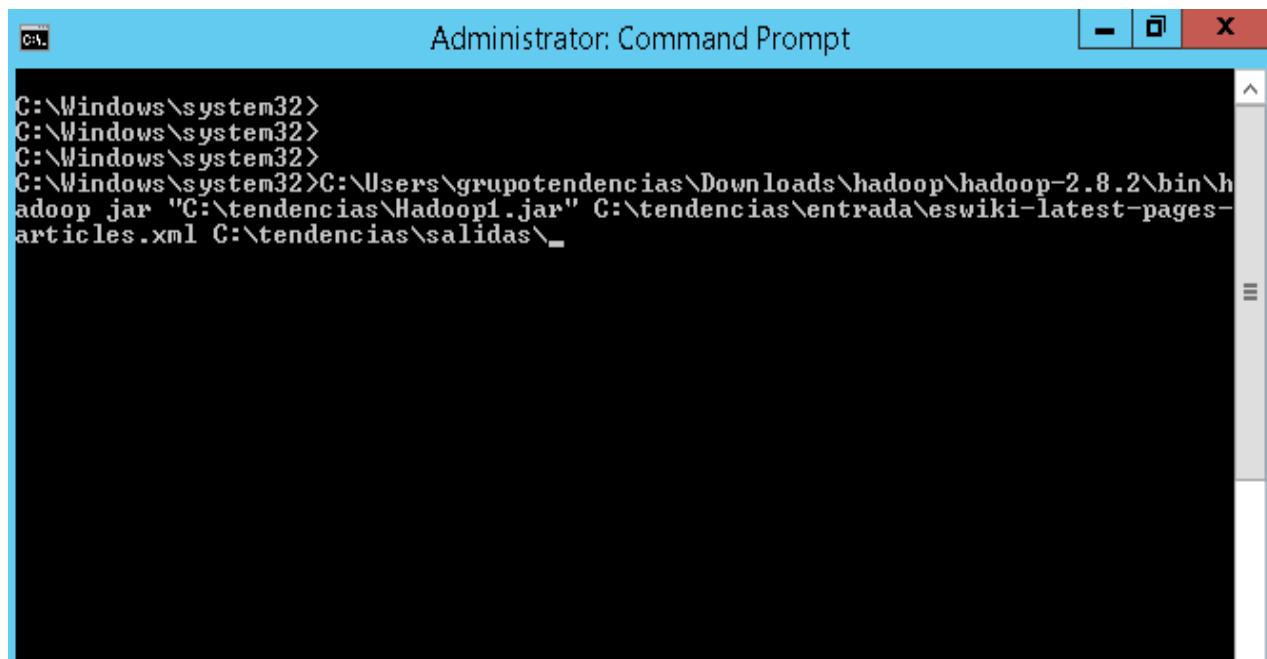


Figura 4-51: Instalación de Visual C++

Se ejecuta el jar, con Hadoop, con el fin de hacer procesamiento distribuido en los diferentes nodos simulados.



The screenshot shows a Windows Command Prompt window titled "Administrator: Command Prompt". The window has a blue header bar with the title and standard window controls (minimize, maximize, close). The main area of the window is black, representing the command line interface. The text displayed in the window is as follows:

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

Figura 4-52: Ejecución correcta de hadoop

4.5. Utilización de bucket en GCP:

4.5.1. Crear Bucket (segmento) en Gcp

En la figura 4.53 se realiza la creación del segmento (bucket tendencias).

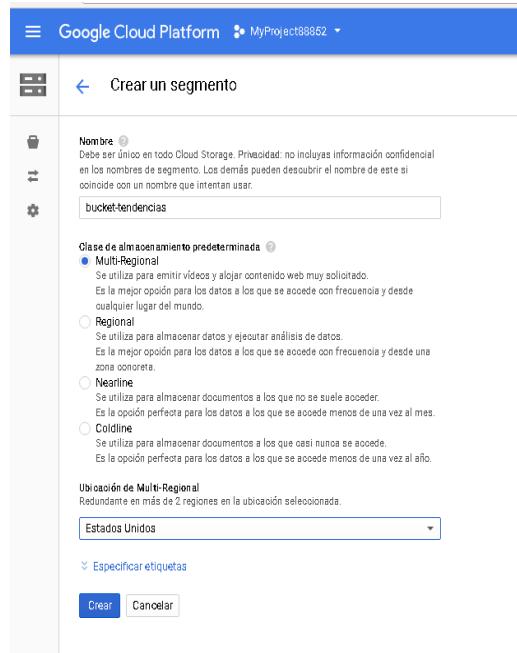


Figura 4-53: Creación del Segmento

4.5.2. Crear carpeta en Bucket

Luego se procede a subir el archivo hadoop1.jar al Bucket como lo muestra la figura 4.54.

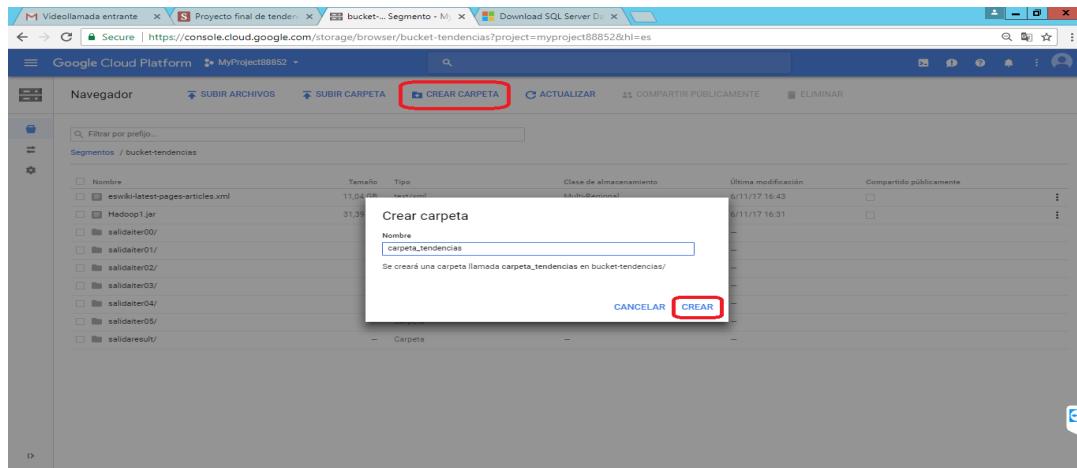


Figura 4-54: Crear Carpeta en Bucket

4.5.3. Cargar en Bucket el programa y los datos

Luego se procede a subir el archivo hadoop1.jar al Bucket como lo muestra la figura 4.55.

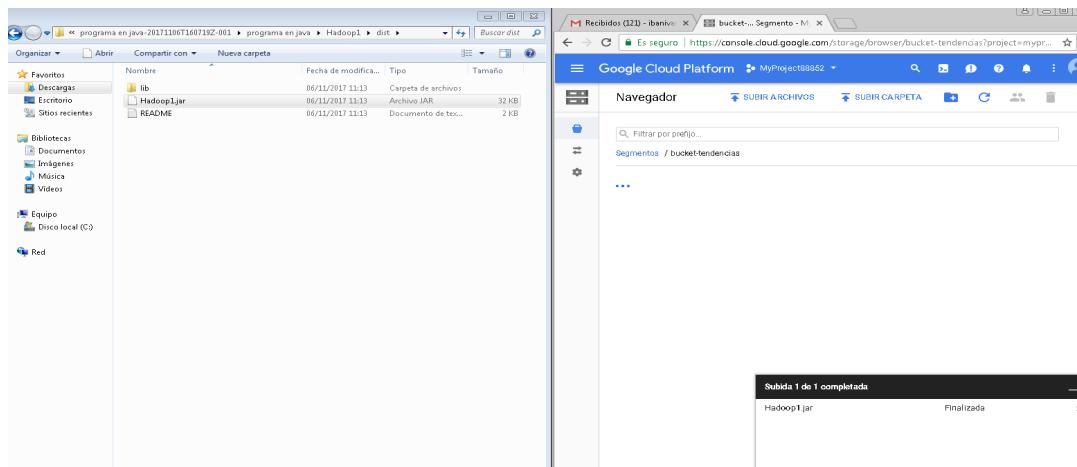


Figura 4-55: Paso de archivos .jar al Bucket

Se realiza la carga del archivo xml arrastrándolo al Bucket (Figura 4.56).

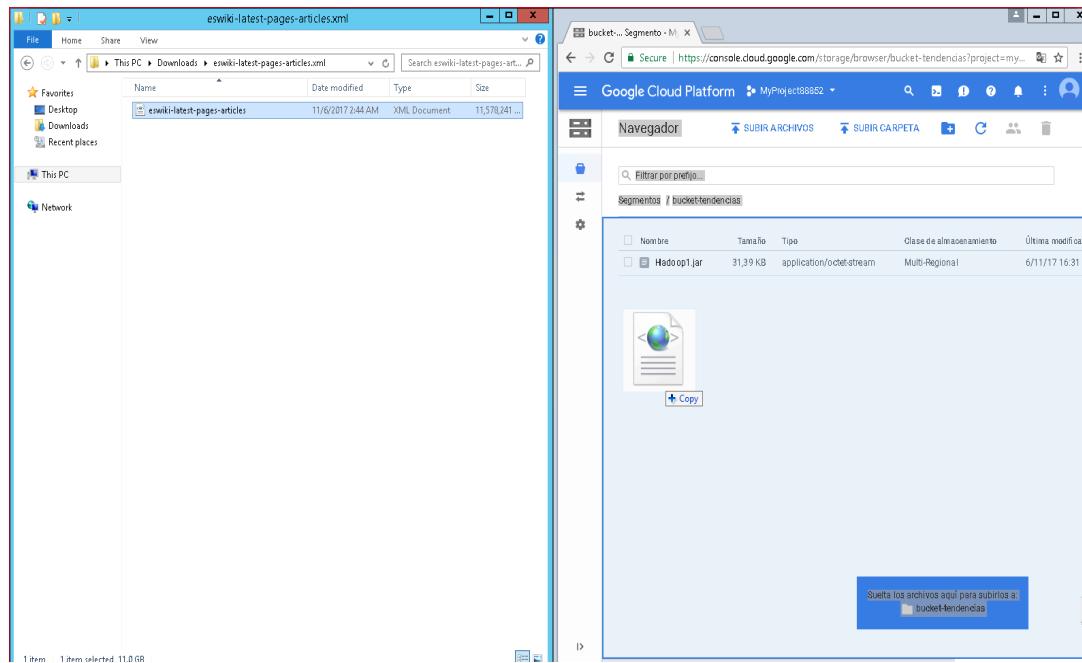


Figura 4-56: Carga del archivo xml

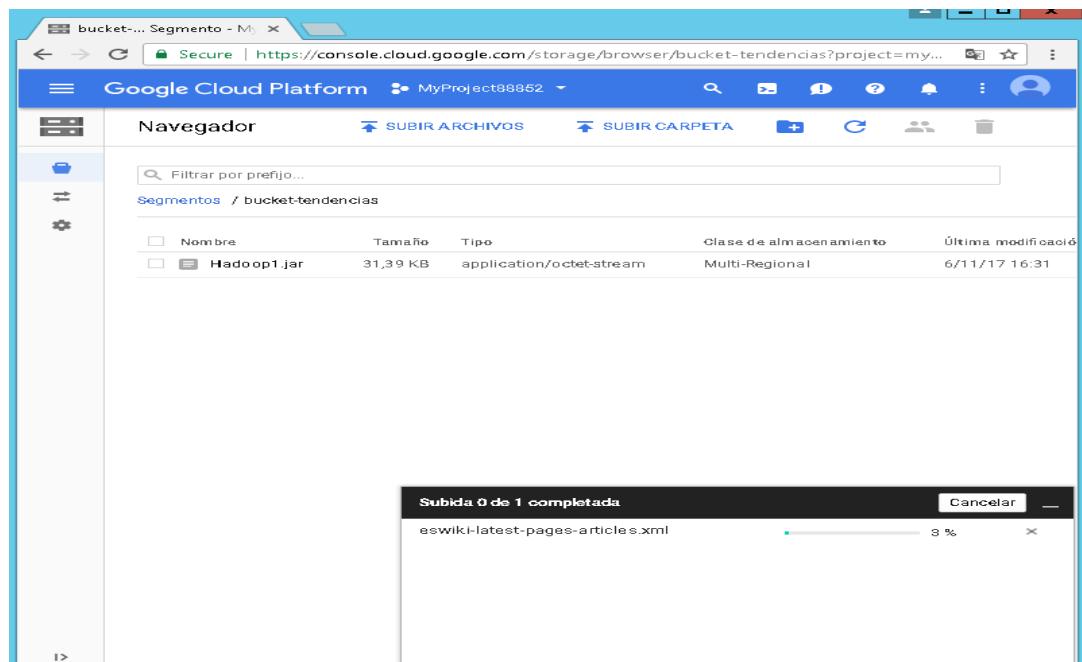


Figura 4-57: Progreso de la carga del archivo xml

En la Figura 4.58 se observa que los archivos se cargaron correctamente en el Bucket.

The screenshot shows the Google Cloud Platform Storage Browser interface. At the top, there is a navigation bar with the project name 'MyProject88852'. Below the navigation bar, there are several buttons: 'SUBIR ARCHIVOS', 'SUBIR CARPETA', 'CREAR CARPETA', 'ACTUALIZAR', 'COMPARTIR PÚBLICAMENTE', and 'ELIMINAR'. There is also a search bar and a filter input field labeled 'Filtrar por prefijo...'. The main area displays a list of files under the path 'Segmentos / bucket/tendencias'. The table has columns for Nombre (Name), Tamaño (Size), Tipo (Type), Clase de almacenamiento (Storage class), Última modificación (Last modified), and Compartido públicamente (Shared publicly). Two files are listed:

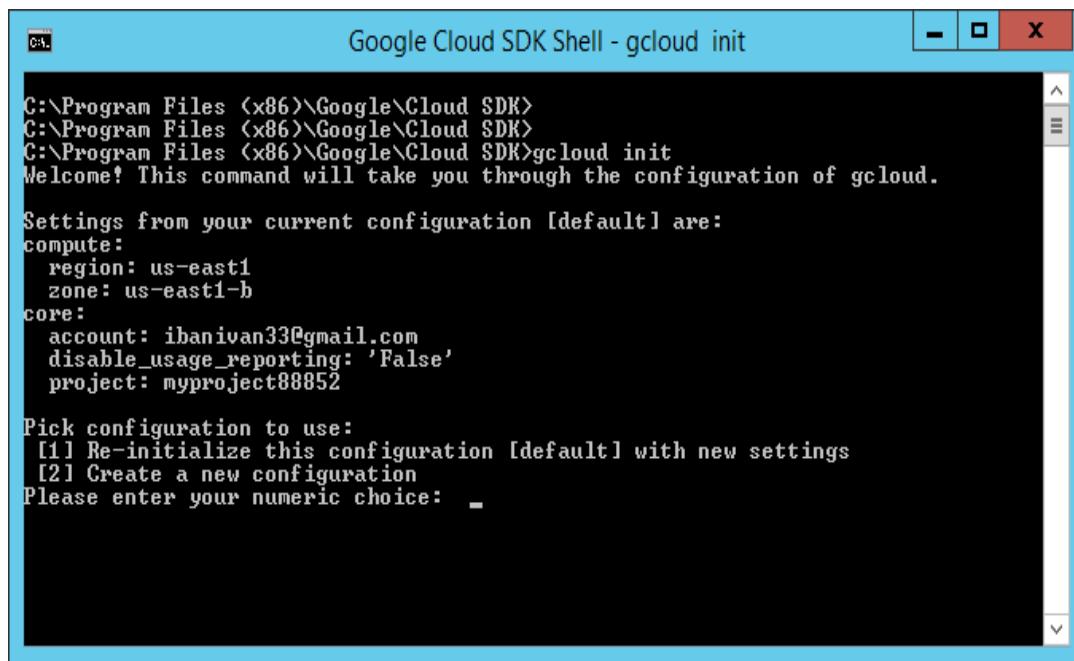
Nombre	Tamaño	Tipo	Clase de almacenamiento	Última modificación	Compartido públicamente
eswiki-latest-pages-articles.xml	11,04 GB	text/xml	Multi-Regional	6/11/17 16:43	<input type="checkbox"/>
Hadoop1.jar	31,39 KB	application/octet-stream	Multi-Regional	6/11/17 16:31	<input type="checkbox"/>

Figura 4-58: Carga completa de los archivos

4.6. Utilización de Google Cloud Shell y Google Cloud SDK Shell

4.6.1. Crear cluster (dataproc) en Google Cloud Platform

En Google Cloud Shell ejecutar el comando gcloud init.



The screenshot shows a terminal window titled "Google Cloud SDK Shell - gcloud init". The window contains the following text:

```
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: _
```

Figura 4-59: Ejecución del comando gcloud init

Igualmente se selecciona la zona de gcloud

```
[18] europe-west2-c
[19] europe-west2-a
[20] europe-west2-b
[21] europe-west3-a
[22] europe-west3-c
[23] europe-west3-b
[24] southamerica-east1-a
[25] southamerica-east1-c
[26] southamerica-east1-b
[27] us-central1-a
[28] us-central1-c
[29] us-central1-f
[30] us-central1-b
[31] us-east1-b
[32] us-east1-d
[33] us-east1-c
[34] us-east4-c
[35] us-east4-a
[36] us-east4-b
[37] us-west1-c
[38] us-west1-a
[39] us-west1-b
[40] Do not set default zone
Please enter numeric choice or text value (must exactly match list item):
```

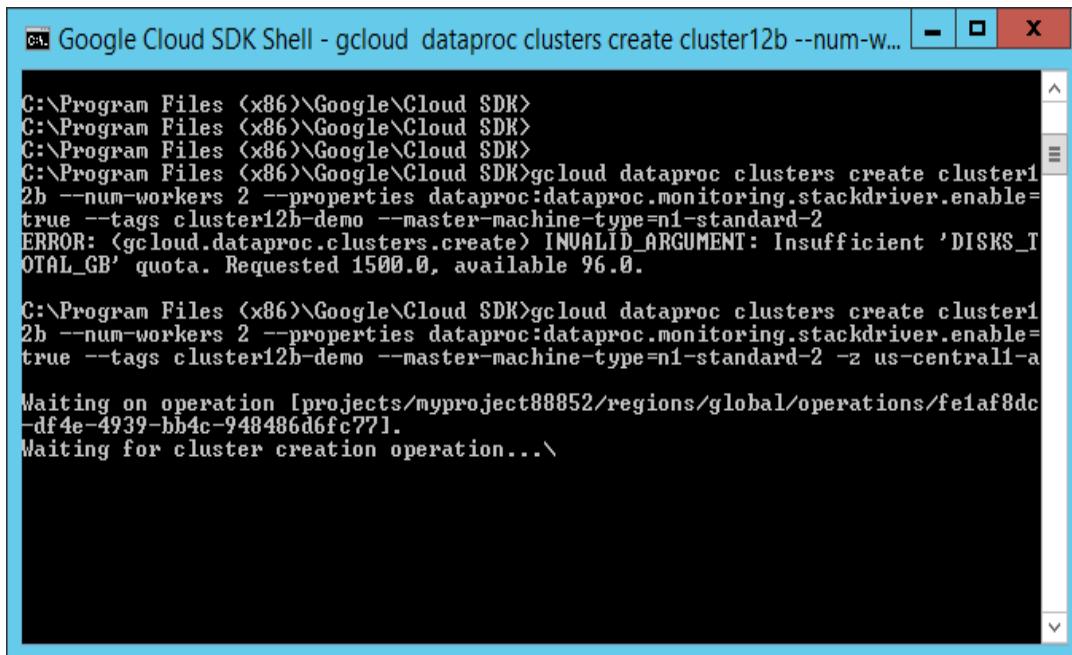
Figura 4-60: Escoger zona de gcloud

Luego se ejecuta el comando para crear el dataproc cluster

```
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 cluster12 --num-workers 2 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12-demo --master-machine-type=n1-standard-2
Waiting on operation [projects/myproject88852/regions/global/operations/9da7475c-2864-4133-8cf8-05e8c40423f3].
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/myproject88852/regions/global/clusters/cluster12] Cluster placed in zone [us-east1-b].
C:\Program Files (x86)\Google\Cloud SDK>
```

Figura 4-61: Comando para creación del dataproc cluster

Se genera un error relacionado con la falta de espacio.

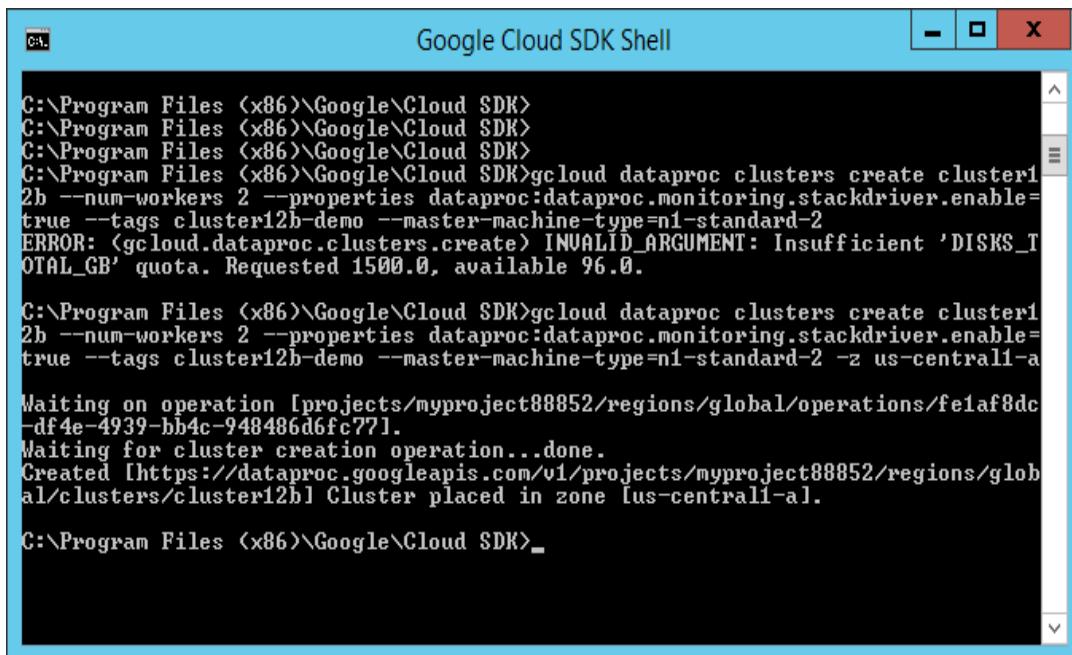


```
Google Cloud SDK Shell - gcloud dataproc clusters create cluster12b --num-w...
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12b --num-workers 2 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12b-demo --master-machine-type=n1-standard-2
ERROR: <gcloud.dataproc.clusters.create> INVALID_ARGUMENT: Insufficient 'DISKS_TOTAL_GB' quota. Requested 1500.0, available 96.0.

C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12b --num-workers 2 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12b-demo --master-machine-type=n1-standard-2 -z us-central1-a
Waiting on operation [projects/myproject88852/regions/global/operations/fe1af8dc-df4e-4939-bb4c-948486d6fc77].
Waiting for cluster creation operation...<
```

Figura 4-62: Error por falta de espacio

Para el error de falta de espacio se procede a seleccionar otra zona

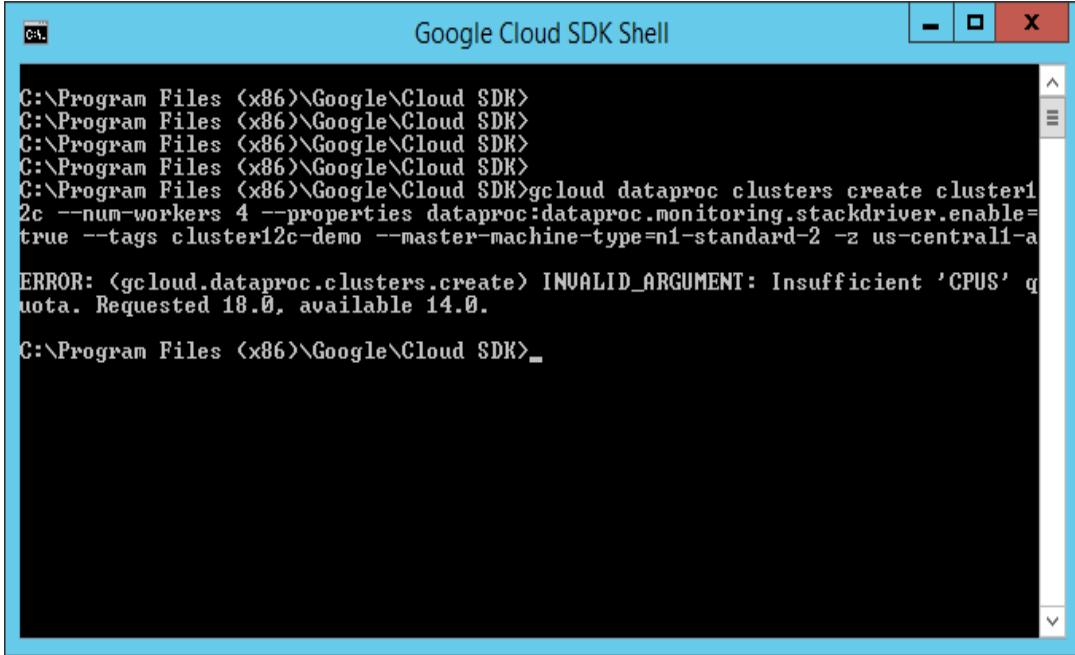


```
Google Cloud SDK Shell
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12b --num-workers 2 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12b-demo --master-machine-type=n1-standard-2
ERROR: <gcloud.dataproc.clusters.create> INVALID_ARGUMENT: Insufficient 'DISKS_TOTAL_GB' quota. Requested 1500.0, available 96.0.

C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12b --num-workers 2 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12b-demo --master-machine-type=n1-standard-2 -z us-central1-a
Waiting on operation [projects/myproject88852/regions/global/operations/fe1af8dc-df4e-4939-bb4c-948486d6fc77].
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/myproject88852/regions/global/clusters/cluster12b] Cluster placed in zone [us-central1-a].
C:\Program Files (x86)\Google\Cloud SDK>_
```

Figura 4-63: Solución a falta de espacio

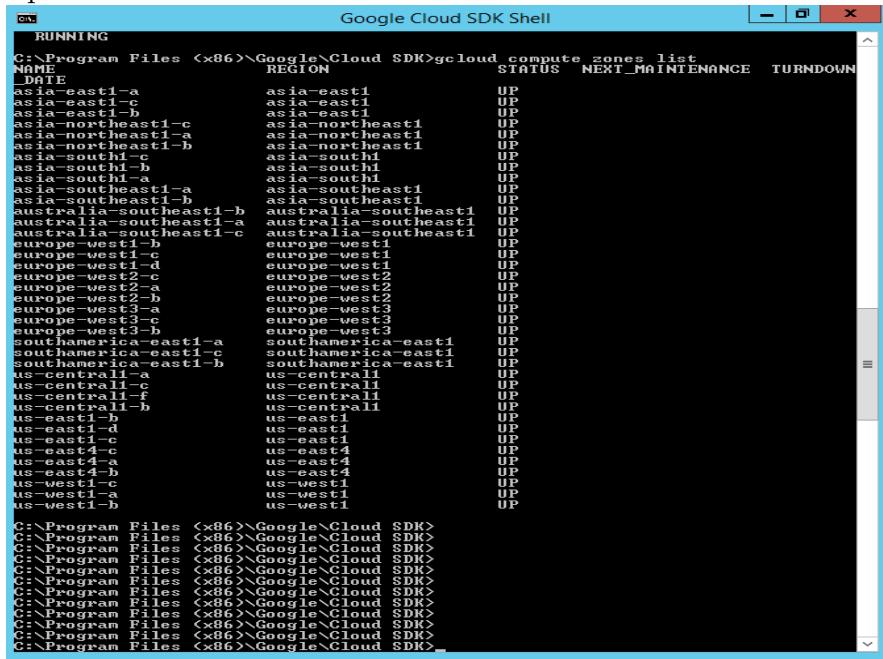
Posteriormente se presenta otro error relacionado con la falta de CPU



```
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>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12c --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' quota. Requested 18.0, available 14.0.
C:\Program Files (x86)\Google\Cloud SDK>
```

Figura 4-64: Error por falta de CPU

Para resolver el problema de falta de CPU procedemos a consultar la lista de zonas para elegir una que permita crear el cluster.



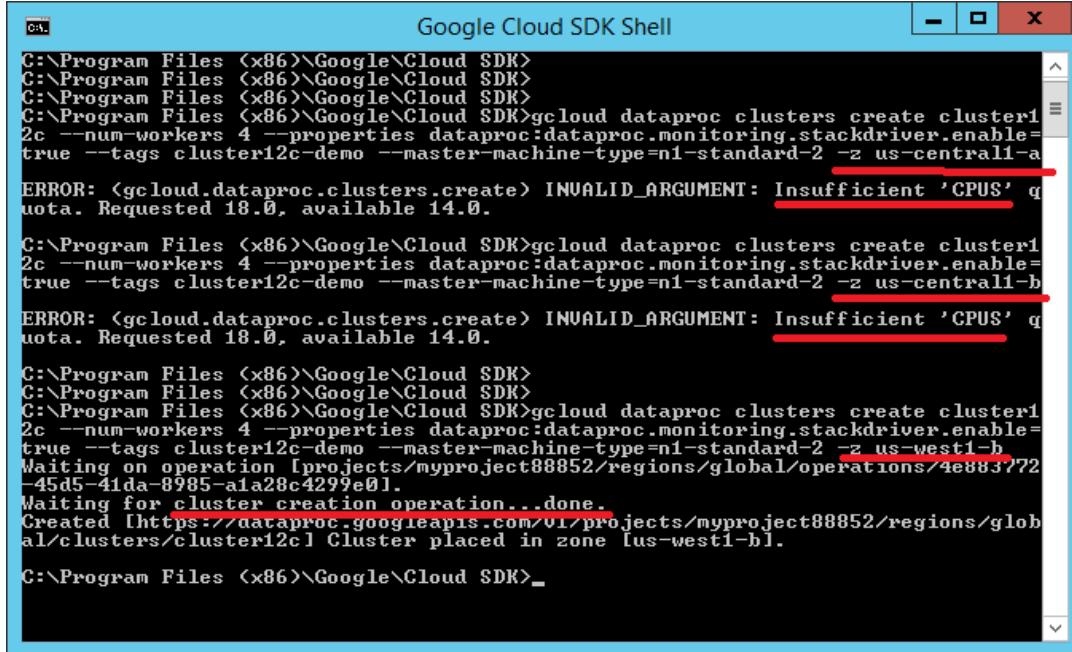
NAME	REGION	STATUS	NEXT_MAINTENANCE	TURNDOWN
asia-east1-a	asia-east1	UP		
asia-east1-b	asia-east1	UP		
asia-northeast1-c	asia-northeast1	UP		
asia-northeast1-a	asia-northeast1	UP		
asia-northeast1-b	asia-northeast1	UP		
asia-south1-a	asia-south1	UP		
asia-south1-b	asia-south1	UP		
asia-southeast1-a	asia-southeast1	UP		
asia-southeast1-b	asia-southeast1	UP		
australia-southeast1-b	australia-southeast1	UP		
australia-southeast1-a	australia-southeast1	UP		
australia-southeast1-c	australia-southeast1	UP		
europe-west1-b	europe-west1	UP		
europe-west1-c	europe-west1	UP		
europe-west1-d	europe-west1	UP		
europe-west2-c	europe-west2	UP		
europe-west2-a	europe-west2	UP		
europe-west2-b	europe-west2	UP		
europe-west3-a	europe-west3	UP		
europe-west3-c	europe-west3	UP		
europe-west3-b	europe-west3	UP		
southamerica-east1-a	southamerica-east1	UP		
southamerica-east1-c	southamerica-east1	UP		
southamerica-east1-b	southamerica-east1	UP		
us-central1-a	us-central1	UP		
us-central1-c	us-central1	UP		
us-central1-f	us-central1	UP		
us-central1-b	us-central1	UP		
us-east1-b	us-east1	UP		
us-east1-d	us-east1	UP		
us-east1-c	us-east1	UP		
us-east4-c	us-east4	UP		
us-east4-a	us-east4	UP		
us-east4-b	us-east4	UP		
us-west1-c	us-west1	UP		
us-west1-f	us-west1	UP		
us-west1-b	us-west1	UP		

```
C:\Program Files (x86)\Google\Cloud SDK>
```

Figura 4-65: Lista de zonas para creación del cluster

Después de seleccionar la zona se procede a incluirla en el código de creación del cluster; posteriormente se muestra la creación correcta del cluster.

Si el cambio de zona no funciona, se aconseja revisar y activar el proceso de facturación de Google Cloud Platform, el cual le permitirá utilizar mas recursos.



```

Google Cloud SDK Shell
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 cluster12c
--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' quota. Requested 18.0, available 14.0.

C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12c
--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' quota. Requested 18.0, available 14.0.

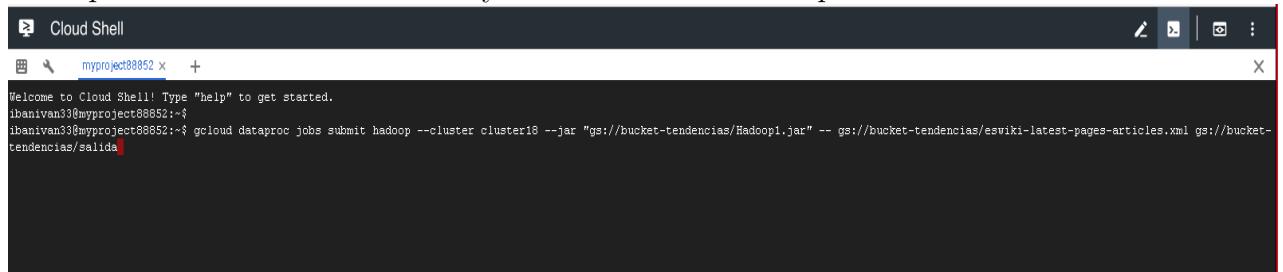
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>
C:\Program Files (x86)\Google\Cloud SDK>gcloud dataproc clusters create cluster12c
--num-workers 4 --properties dataproc:dataproc.monitoring.stackdriver.enable=true --tags cluster12c-demo --master-machine-type=n1-standard-2 -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/global/clusters/cluster12c] Cluster placed in zone [us-west1-b].
C:\Program Files (x86)\Google\Cloud SDK>

```

Figura 4-66: Error, solución y creación exitosa del Cluster Dataproc

4.6.2. Ejecutar en cluster el programa utilizando Bucket (Cloud Shell)

El lanzamiento del comando para enviar un trabajo al cluster se detalla en la figura 4.67; el archivo .jar corresponde al trabajo que se va a ejecutar con Hadoop; los parámetros corresponden a la ruta de entrada y a la ruta de salida del proceso.



```

Cloud Shell
myproject88852 x +
Welcome to Cloud Shell! Type "help" to get started.
ibanivan38@myproject88852:~$ gcloud dataproc jobs submit hadoop --cluster cluster18 --jar "gs://bucket-tendencias/Hadoop1.jar" -- gs://bucket-tendencias/eswiki-latest-pages-articles.xml gs://bucket-tendencias/salida

```

Figura 4-67: Comando para el cluster

En la figura 4.68 podemos observar el avance de la primera iteración después de ejecutar el comando.

Figura 4-68: Primera Iteración

Igualmente en la figura 4.69 se muestra el progreso de las iteraciones.

```
Cloud Shell
mpio@ec38862 x +
```

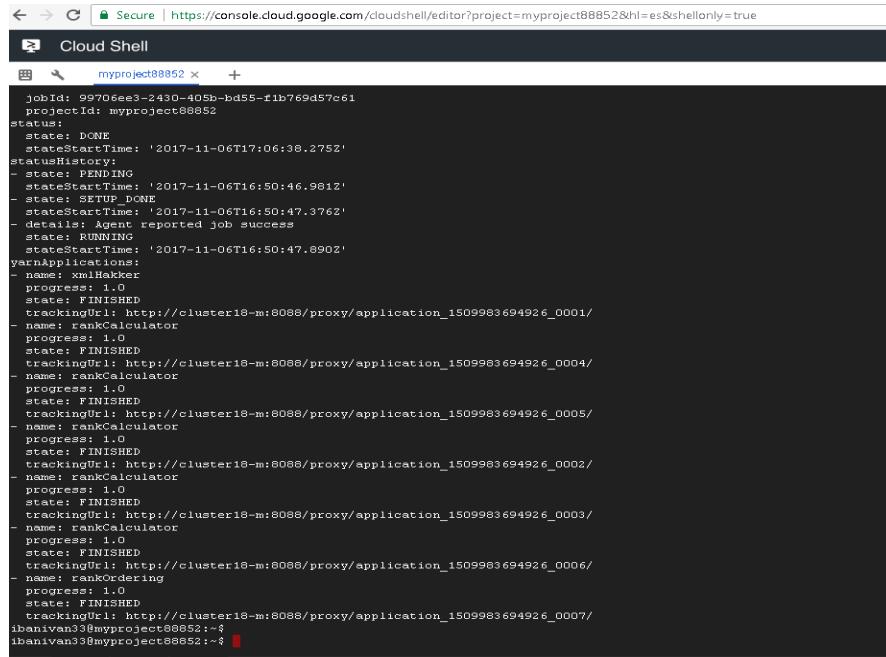
```
Input split bytes=1695
Combine input records=0
Combine output records=0
Reduce input groups=437857
Reduce shuffle bytes=37459813781
Reduce input records=61283741
Reduce output records=2928713
Spilled Records=163851223
Shuffled Maps =225
Failed Shuffles=0
Merged Map outputs=225
GC time elapsed (ms)=17309
CPU time spent (ms)=705940
Physical memory (bytes) snapshot=31742756912
Virtual memory (bytes) snapshot=131978018816
Total committed heap usage (bytes)=30867980288

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_PARTITION=0
  WRONG_REDUCE=0

File Input Reader Counters
  Bytes Read=77415914
File Output Format Counters
  Bytes Written=930842468
dentro del ciclo en el iteracion 1
---inPathss://Bucket-tendencias/maldaiter01
-----laeResultPath _gs://bucket-tendencias/maldaiter02
antes de otro waitforcompletion
17/11/06 17:00:23 INFO client.JNFCopy: Connecting to ResourceManager at cluster18-m-10.142.0.3:8802
17/11/06 17:00:23 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/11/06 17:00:23 INFO input.FileInputFormat: Input splits files to process : 15
17/11/06 17:00:23 INFO mapreduce.JobSubmitter: number of splits:15
17/11/06 17:00:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1509983694926_0003
17/11/06 17:00:23 INFO impl.YarnClientImpl: Submitted application application_1509983694926_0003
17/11/06 17:00:23 INFO mapreduce.Job: The url to track the job: http://cluster18-mr8088/proxy/application_1509983694926_0003/
17/11/06 17:00:30 INFO mapreduce.Job: Job: Running job: job_1509983694926_0003
17/11/06 17:00:30 INFO mapreduce.Job:   map 0% reduce 0%
17/11/06 17:00:48 INFO mapreduce.Job:   map 10% reduce 0%
17/11/06 17:00:49 INFO mapreduce.Job:   map 55% reduce 0%
```

Figura 4-69: Progreso de iteraciones

Posteriormente se observa como finalizan las iteraciones en la figura 4.70.



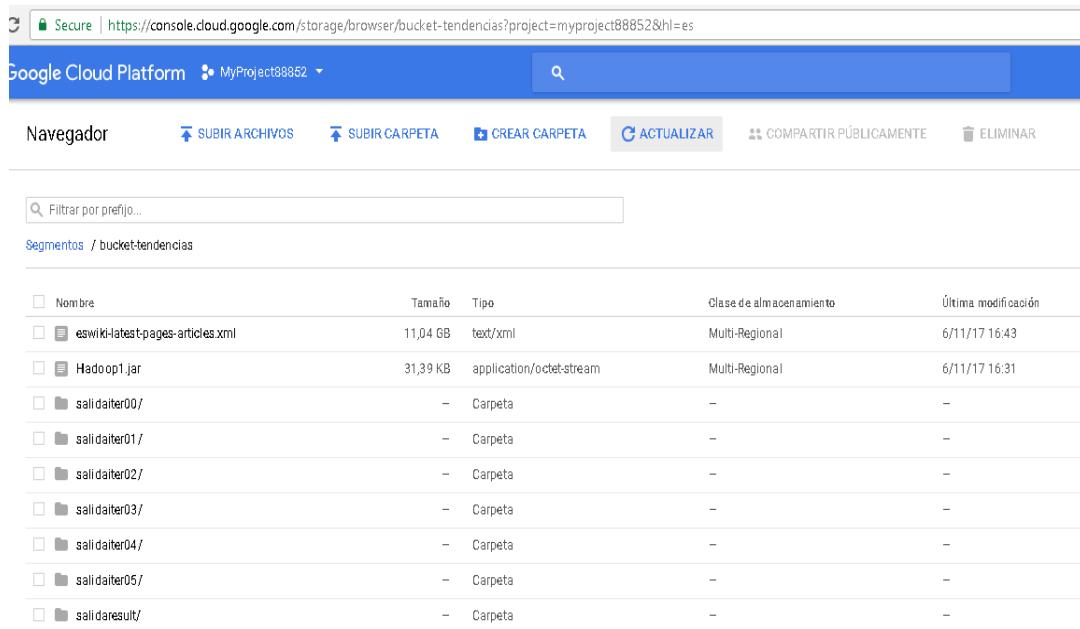
```

← → ⌂ Secure | https://console.cloud.google.com/cloudshell/editor?project=myproject88852&hl=es&shellonly=true
Cloud Shell
myproject88852 × +
 jobId: 99706ee3-2430-405b-bd55-f1b769d57c61
 projectId: myproject88852
status:
- state: DONE
  stateStartTime: '2017-11-06T17:06:38.275Z'
statusHistory:
- state: PENDING
  stateStartTime: '2017-11-06T16:50:46.981Z'
- state: DONE
  stateStartTime: '2017-11-06T16:50:47.376Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2017-11-06T16:50:47.890Z'
yarnApplications:
- name: xmlHakker
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster18-m:8088/proxy/application_1509983694926_0001/
- name: rankCalculator
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster18-m:8088/proxy/application_1509983694926_0004/
- name: rankCalculator
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster18-m:8088/proxy/application_1509983694926_0005/
- name: rankCalculator
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster18-m:8088/proxy/application_1509983694926_0006/
- name: rankCalculator
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster18-m:8088/proxy/application_1509983694926_0007/
ibanivan3@myproject88852:~$ ibanivan3@myproject88852:~$ ibanivan3@myproject88852:~$ 

```

Figura 4-70: Finalización de las iteraciones

Finalmente se muestra el resultado en el Bucket.



Navegador	SUBIR ARCHIVOS	SUBIR CARPETA	CREAR CARPETA	ACTUALIZAR	COMPARTE PUBLICAMENTE	ELIMINAR
<input type="text"/> Filtrar por prefijo...						
Segmentos / bucket-tendencias						
Nombre	Tamaño	Tipo	Clase de almacenamiento	Última modificación		
eswiki-latest-pages-articles.xml	11,04 GB	text/xml	Multi-Regional	6/11/17 16:43		
Hadoop1.jar	31,39 KB	application/octet-stream	Multi-Regional	6/11/17 16:31		
salidaiter00/	-	Carpeta	-	-		
salidaiter01/	-	Carpeta	-	-		
salidaiter02/	-	Carpeta	-	-		
salidaiter03/	-	Carpeta	-	-		
salidaiter04/	-	Carpeta	-	-		
salidaiter05/	-	Carpeta	-	-		
salidaresult/	-	Carpeta	-	-		

Figura 4-71: Resultado en el Bucket

4.6.3. Descargar del bucket en la máquina virtual el resultado del programa (Sdk Shell)

1. Es necesario iniciar la interacción con Google Cloud; para esto se debe seleccionar la cuenta.



The screenshot shows a terminal window titled "Google Cloud SDK Shell". The window has a blue header bar with the title and standard window controls (minimize, maximize, close). The main area of the terminal is black and contains white text. It displays the following output:

```
Welcome to the Google Cloud SDK! Run "gcloud -h" to get the list of available commands.  
--  
C:\Program Files (x86)\Google\Cloud SDK>  
C:\Program Files (x86)\Google\Cloud SDK>  
C:\Program Files (x86)\Google\Cloud SDK>gcloud init
```

Figura 4-72: Iniciar la interacción con Google Cloud

2. Se conceden permisos a la cuenta personal para poder interactuar a través del Shell local (Google Sdk) con los recursos previamente establecidos en Google Cloud.

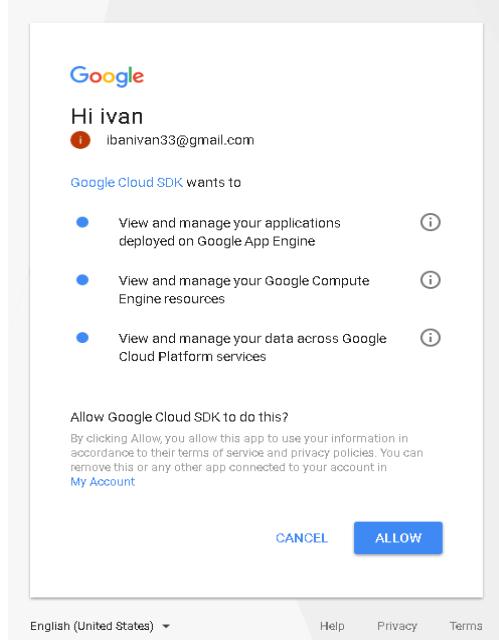


Figura 4-73: Asignación de permisos a la cuenta personal

3. Se realiza la generación del código para colocarlo luego en el Shell con el fin de realizar la verificación de seguridad.

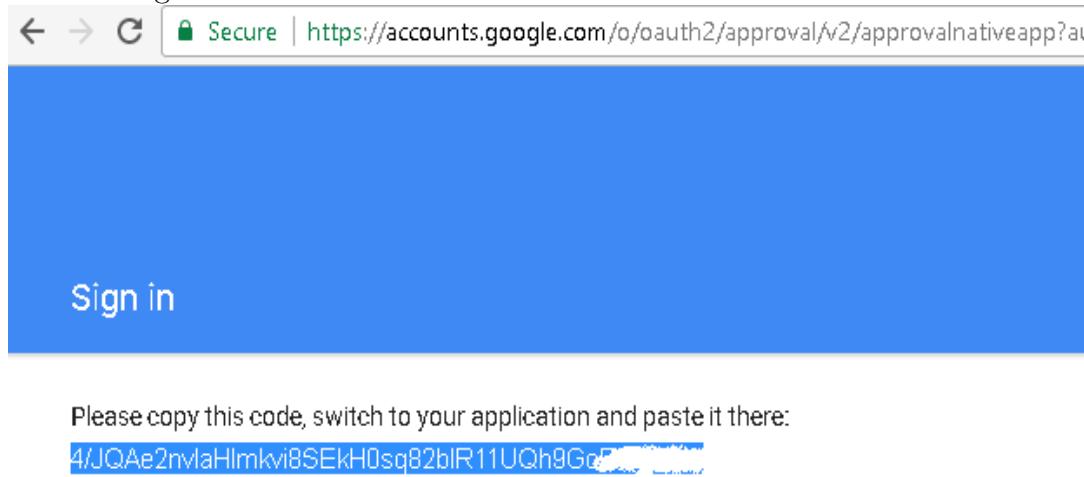


Figura 4-74: Código para la aplicación

4. Se ingresa el código de verificación en el Shell

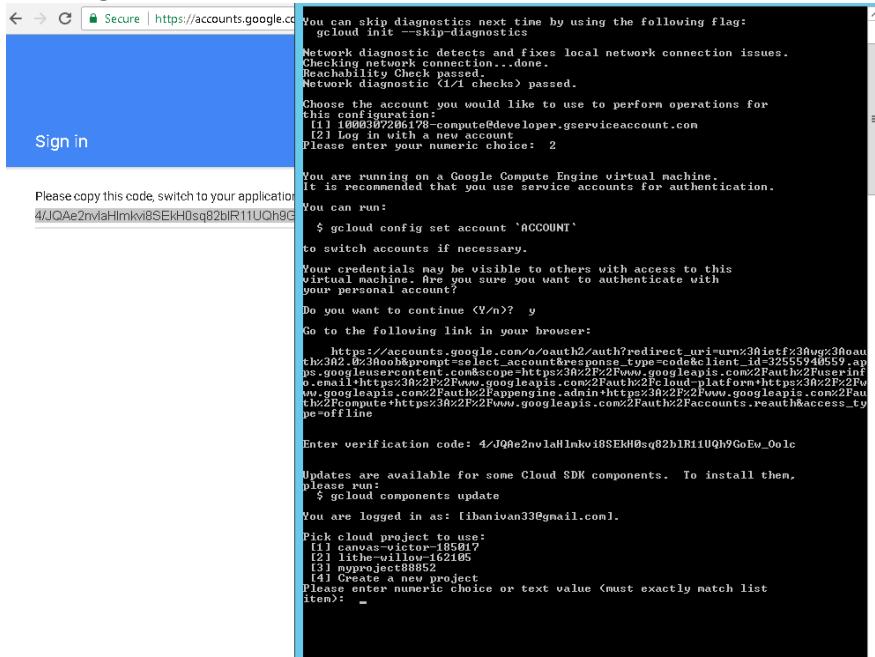


Figura 4-75: Ingreso de Código

5. Estando en el Shell se inicializa gcloud y se selecciona la zona

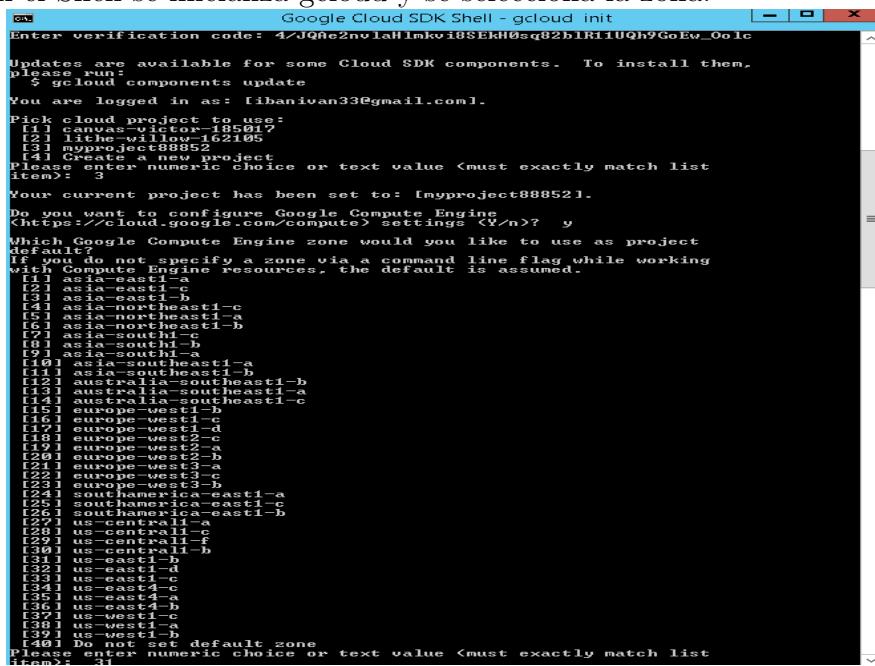


Figura 4-76: Ventana de Shell inicializando gcloud

6. Con el fin de copiar el resultado que está en la nube en la máquina virtual, se ingresa el siguiente comando en el Shell:

```
gsutil cp -r gs://bucket-tendencias/salida* C:\navi\salida_gcp
```



Figura 4-77: Ejecución del comando para traer los archivos de Google Cloud

7. Finalmente se muestran los archivos copiados desde Google Cloud Platform en la carpeta local, gracias al Google Cloud SDK. Dichos archivos se analizarán posteriormente.

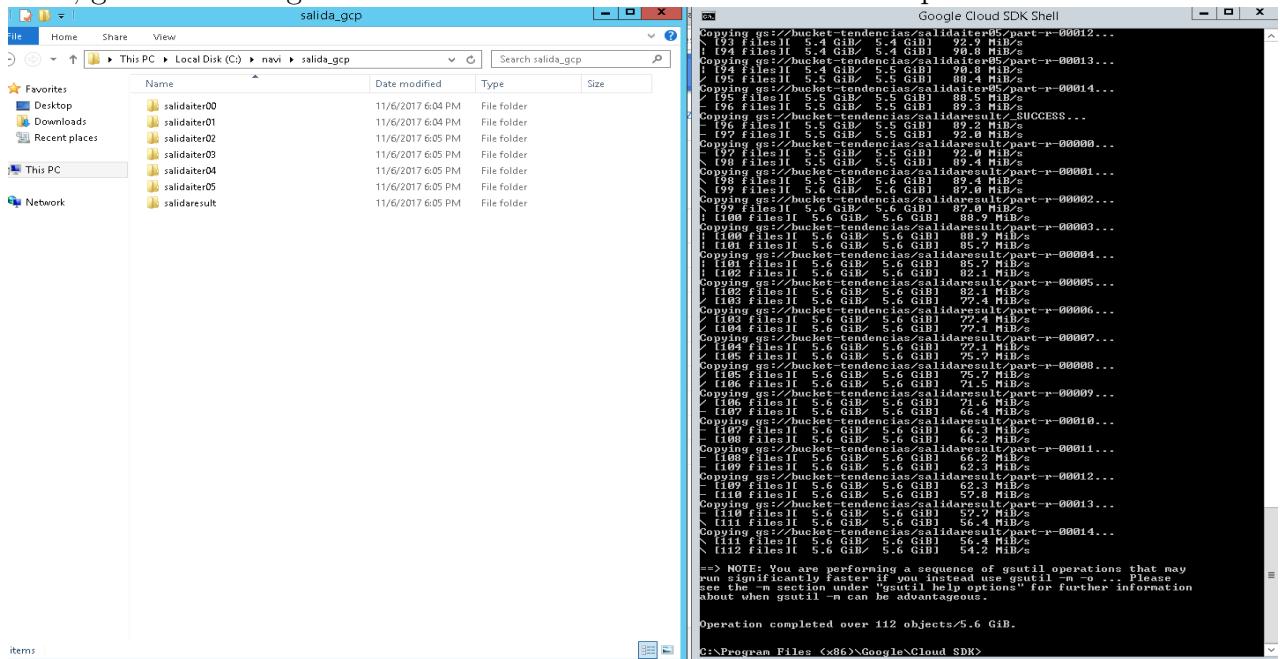


Figura 4-78: Archivos copiados provenientes de Google Cloud Platform

4.7. Utilización de Sql Server y Sql Server Data Tools

4.7.1. Crear base de datos en Sql Server

1. Se realiza la conexión al servidor INSTANCE -1, utilizando la autenticación de Windows.

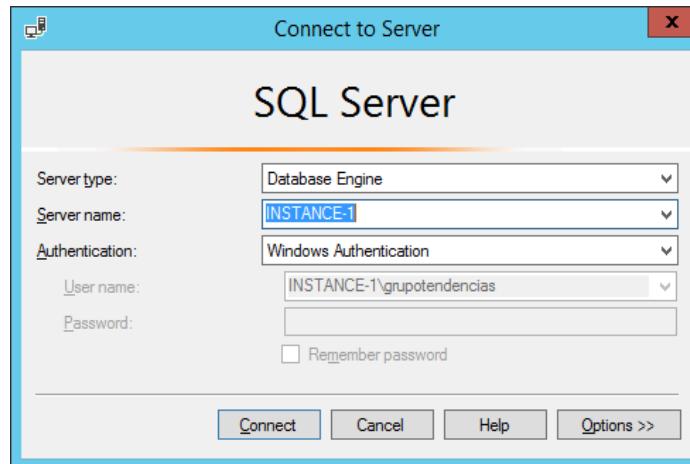


Figura 4-79: Conexión al servidor INSTANCE -1

2. En el explorador de objetos en bases de datos con click derecho se selecciona nueva base de datos

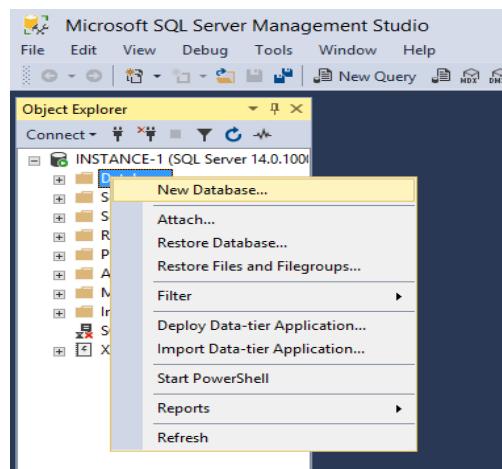


Figura 4-80: Seleccionar base de datos nueva

3. Se asigna el nombre de resultado a la base de datos.

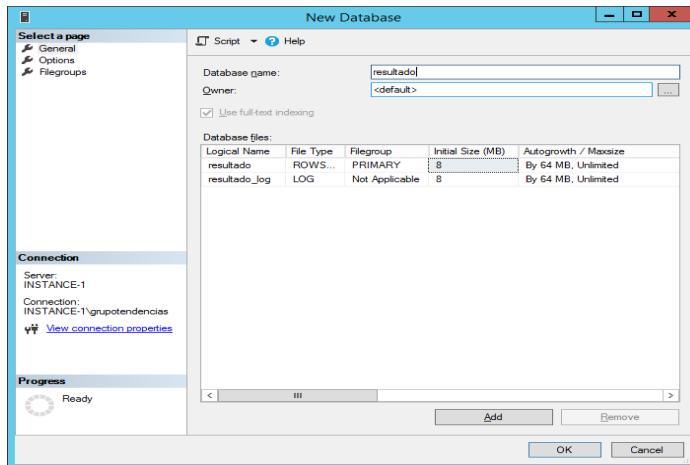


Figura 4-81: Crear base de datos

4.7.2. Crear ETL

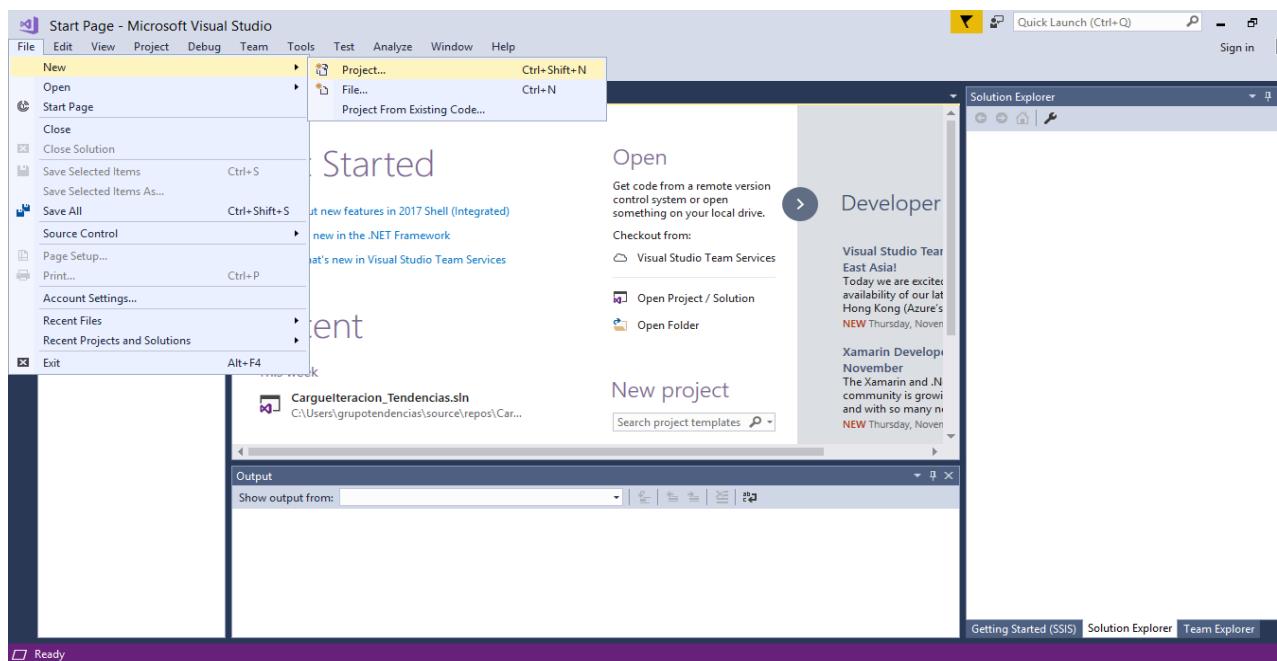


Figura 4-82: Crear un nuevo proyecto

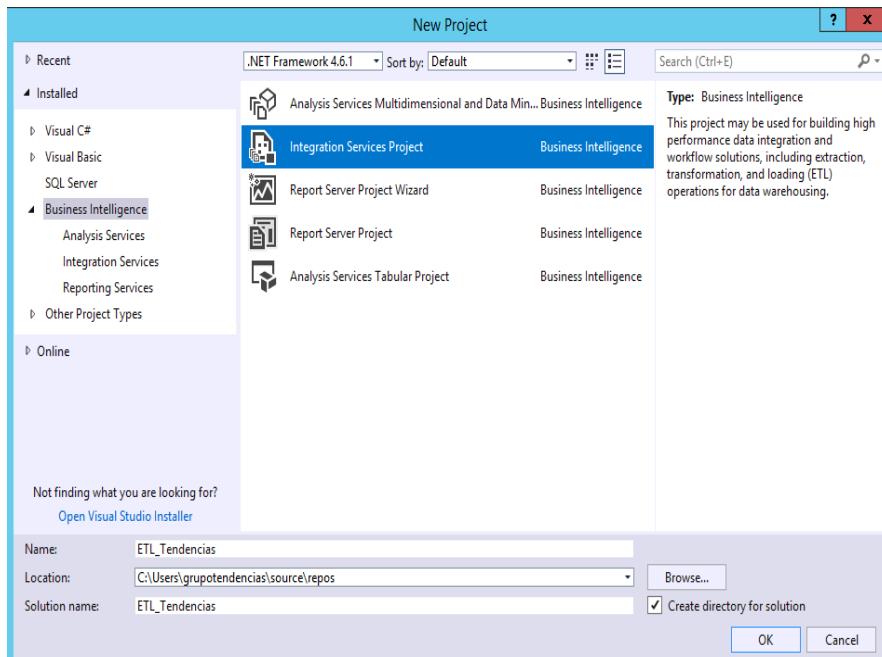


Figura 4-83: Proyecto de tipo Integration Service Project

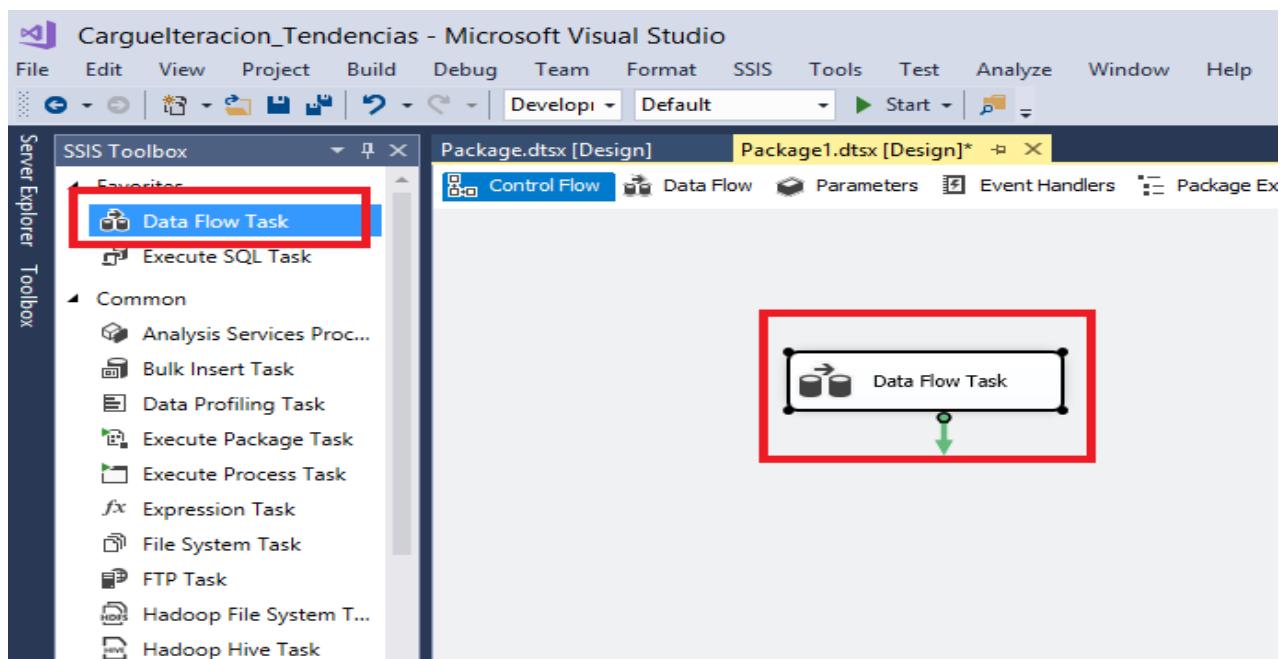


Figura 4-84: Creación del Data Flow

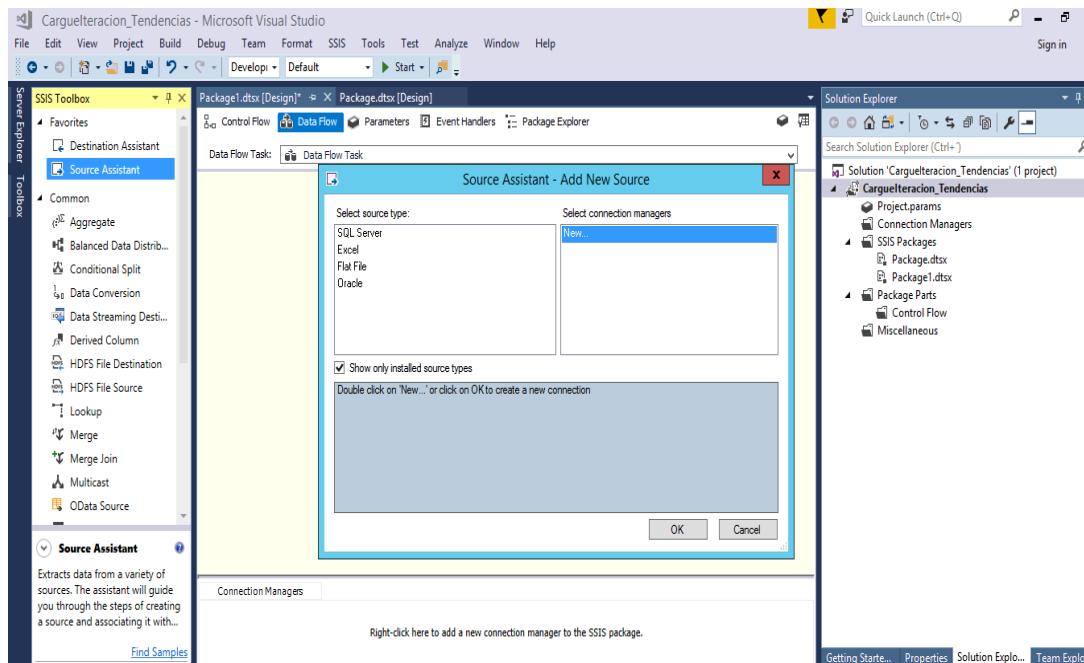


Figura 4-85: Creación del Data Flow

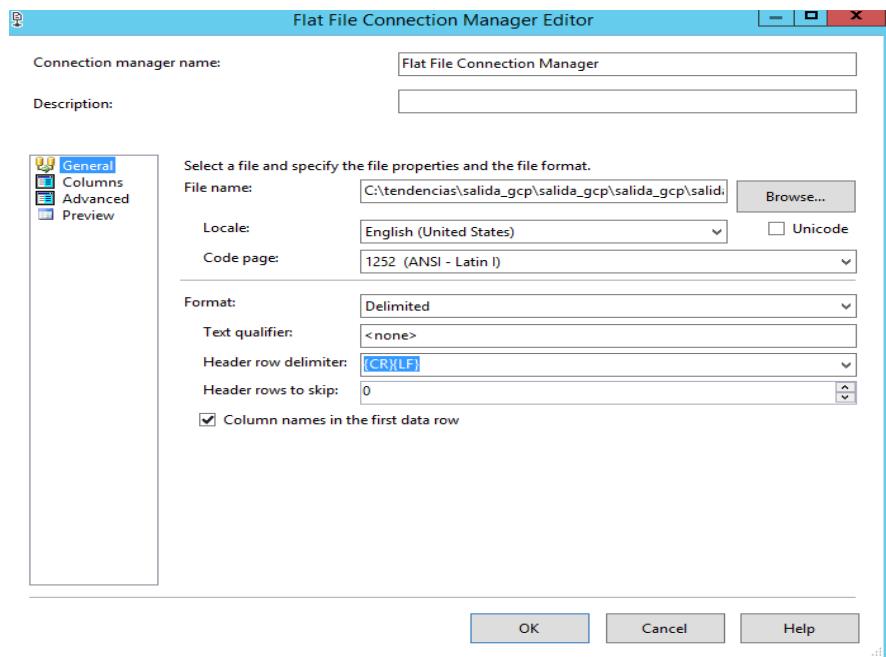


Figura 4-86: Creación de la conexión hacia el archivo plano

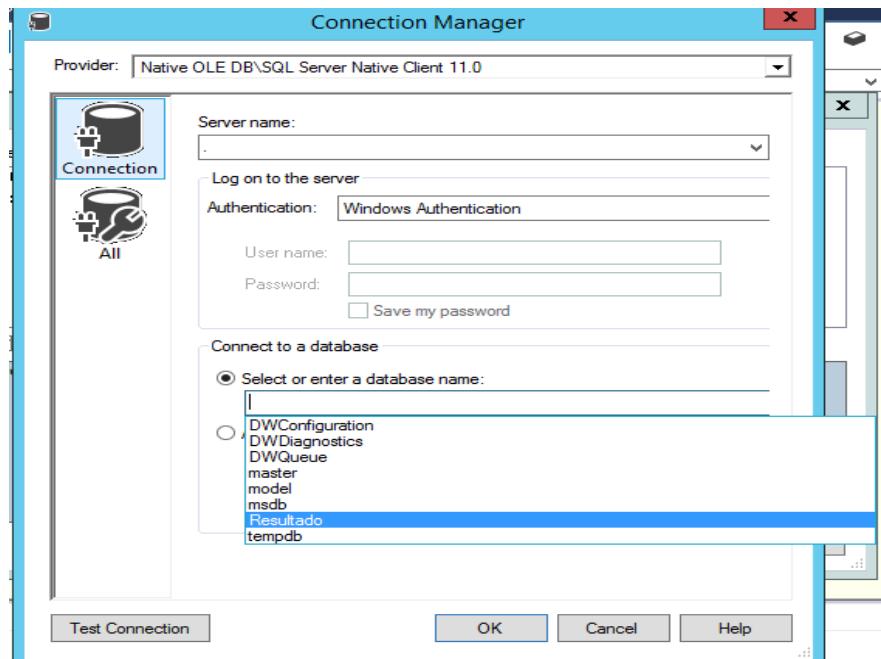


Figura 4-87: Conexión hacia la base de datos

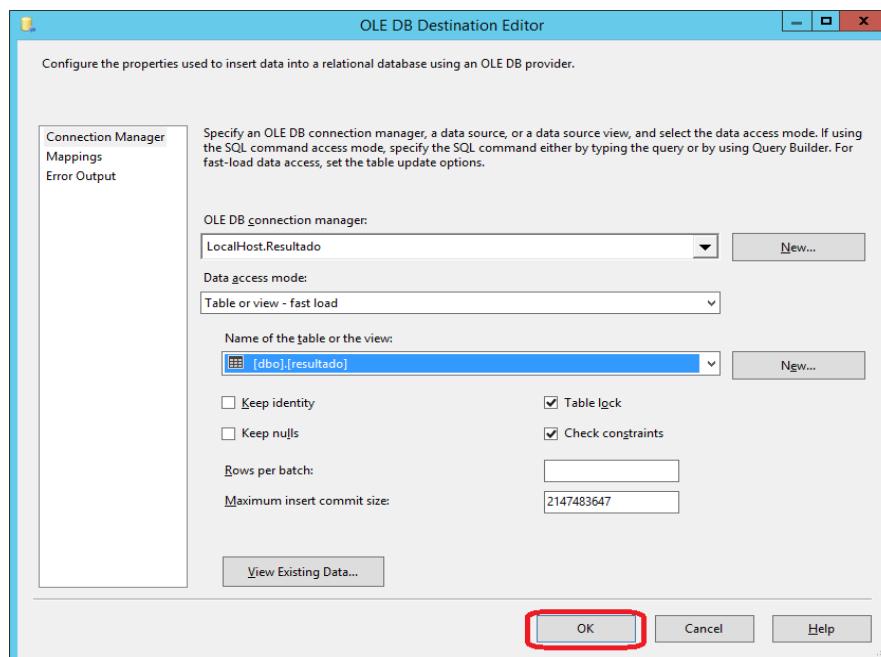


Figura 4-88: Selección de tabla de la base de datos

4.7.3. Usar ETL

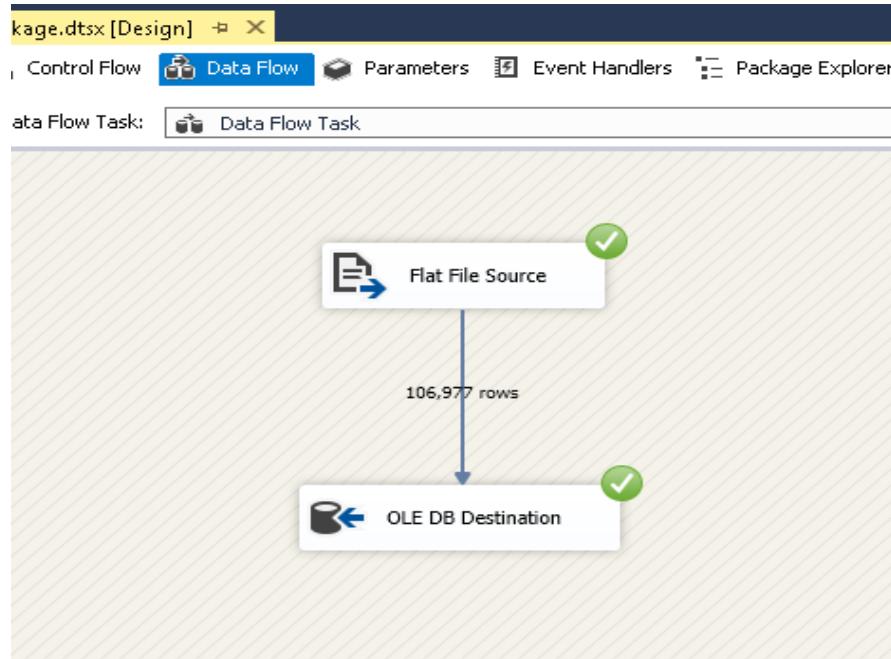


Figura 4-89: Ejecución de la ETL

4.7.4. Consultar tabla de resultado de PageRank

The screenshot shows the SSMS interface with three tabs open: 'SQLQuery3.sql - INSTANCE-1.Resultado (INSTANCE-1\grupotendencias (51))' (selected), 'SQLQuery2.sql - ...upotencias (58)'*, and 'SQLQuery2.sql - IN...upotencias (57)'. The Object Explorer on the left shows various database objects like tables, views, and stored procedures. The results pane displays the top 10 rows of the 'mayorRank' view, ordered by rank. The status bar at the bottom indicates the query was executed successfully.

Rank	Título
1	5084.7744 España
2	4986.1772 Estados_ Unidos
3	2698.0884 Francia
4	1995.0076 México
5	1980.1937 Argentina
6	1769.6356 2008
7	1726.3629 Alemania
8	1622.5522 Madrid
9	1562.4108 Reino_ Unido
10	1510.98 Tiempo_universal_coordinado
11	1500.2358 2007

Figura 4-90: Consulta del resultado final ordenado

4.7.5. Análisis funcional del resultado

Visualización de la página de España en Wikipedia

Figura 4-91: Página de España en Wikipedia

Visualización de la página de América en Wikipedia

The screenshot shows the Spanish Wikipedia page for "América". The page content discusses the history of the continent, mentioning diseases like viruela, African slaves, and independence movements. It includes images of cave paintings and historical documents. A red box highlights a link to the "España" page.

Pinturas rupestres de la Cueva de las Manos, Santa Cruz, Argentina. Estas pinturas rupestres, fechadas en el 7350 a.C., se encuentran entre las expresiones artísticas más antiguas de América.

También se instalaron en América del Sur repúblicas de pueblos de origen africano que lograron huir de la esclavitud a la que habían sido reducidos por los portugueses, como el Quilombo de los Palmares o el Quilombo de Macaco.

Luego de tres siglos de dominio colonial, los pueblos americanos comenzaron a declarar su independencia reclamando su derecho para organizarse como estados nacionales, enfrentando militarmente a las potencias europeas, abriendo de ese modo el proceso mundial de descolonización. Las primeras en hacerlo fueron las Trece Colonias británicas mediante la Revolución estadounidense que dio origen a los Estados Unidos de América, en 1776, organizando un nuevo tipo de sociedad a partir de conceptos políticos novedosos como independencia, constitución, federalismo y derechos del Hombre.

En 1804, los esclavos de origen africano de Haití se sublevaron contra los colonos franceses, declarando la independencia de este país y creando el primer estado moderno con gobernantes afroamericanos.

A partir de 1809,²³ los pueblos bajo dominio de España llevaron adelante una Guerra de Independencia Hispanoamericana, de alcance continental, que terminó con el surgimiento de varias naciones: Argentina, Bolivia, Colombia, Costa Rica, Panamá, Ecuador, Nicaragua, Paraguay, Perú, Uruguay y Venezuela. En 1844 y 1898 el proceso continuó con la independencia de Puerto Rico y Cuba, respectivamente.

En 1819, se creó el Gran Ecuador, que incluyó la mayor parte del actual Ecuador y Perú, denominado Gran Colombia, y que duró hasta 1830. En 1822, el Imperio del Brasil, al disolverse el Reino Unido de Portugal, Brasil y Algarve, se convirtió en la República Federal del Brasil. La monarquía fue abolida para establecer una república. Los países que formaron el Brasil se negocian en 1867 un proceso de independencia durante el siglo XX.

En la actualidad, España, también denominado Reino de España, es un país soberano, miembro de la Unión Europea, constituido en Estado social y democrático de derecho y cuya forma de gobierno es la monarquía parlamentaria. Su territorio, con capital en Madrid,[29] está

Figura 4-92: América con un enlace hacia España en Wikipedia

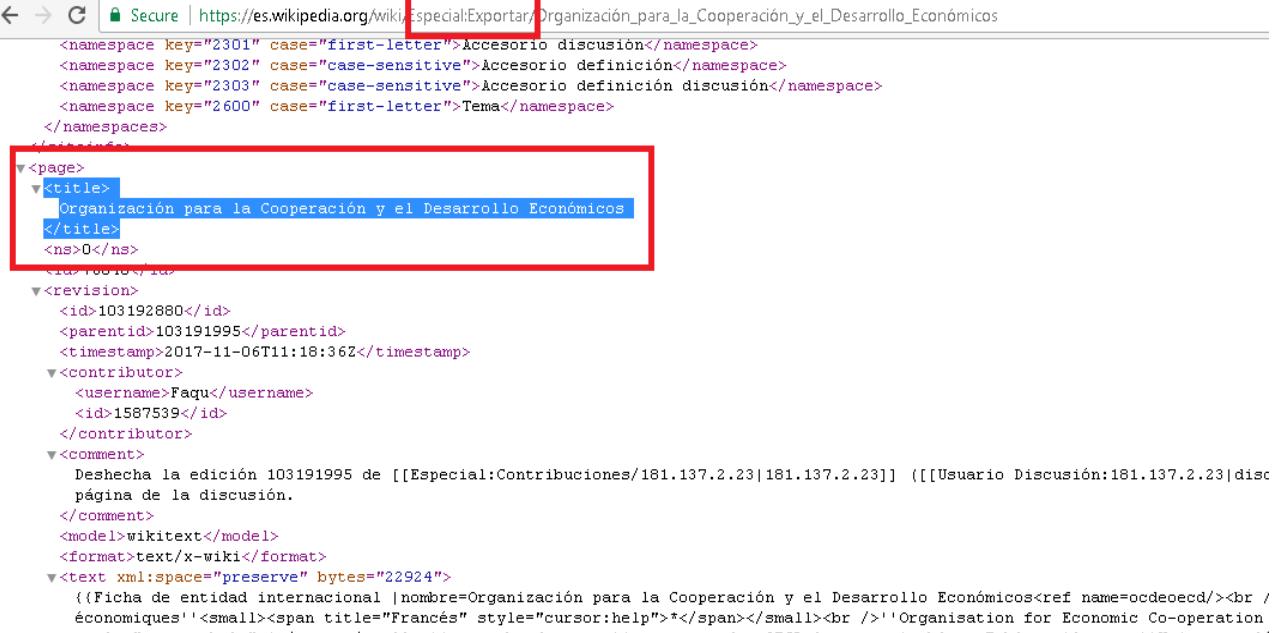
4.7.6. Análisis funcional de una iteración

Para realizar el análisis se escogió aleatoriamente la página correspondiente a la Organización para la Cooperación y el Desarrollo Económicos; en la imagen 4.93. se observa que al agregarle a la url /Special:Export después de la palabra wiki, vemos el xml correspondiente a dicha página.

The screenshot shows the Wikipedia article for the Organization for Economic Co-operation and Development (OECD). The page title is "Organización para la Cooperación y el Desarrollo Económicos". The content discusses the organization's history, its role in coordinating policies, and its headquarters in Paris. A sidebar on the left provides links to other Wikipedia pages related to the OECD. The URL in the browser's address bar is https://es.wikipedia.org/wiki/Organizaci%C3%B3n_para_la_Cooperaci%C3%B3n_y_el_Desarrollo_Econ%C3%B3mico. The word "Special:Export" is highlighted in orange, indicating the URL modification for extracting the XML representation of the page content.

Figura 4-93: Título y Special:Export

En detalle en la figura 4.94 se observa que el título de la página (llave) está en el tag title.



```

<?xml version="1.0" encoding="UTF-8"?>
<page>
  <title>Organización para la Cooperación y el Desarrollo Económicos</title>
  <ns>0</ns>
  <id>103191995</id>
  <revision id="103192880">
    <parentid>103191995</parentid>
    <timestamp>2017-11-06T11:18:36Z</timestamp>
    <contributor>
      <username>Fiqui</username>
      <id>1587539</id>
    </contributor>
    <comment>
      Deshecha la edición 103191995 de [[Especial:Contribuciones/181.137.2.23|181.137.2.23]] ([[Usuario Discusión:181.137.2.23|discusión]] de la página de la discusión.
    </comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text xml:space="preserve" bytes="22924">
      {{Ficha de entidad internacional |nombre=Organización para la Cooperación y el Desarrollo Económicos<ref name="ocdeoeecd"/><br />económicas|<small><span title="Francés" style="cursor:help">*</span></small><br />Organisation for Economic Co-operation and Development|<small><span style="cursor:help">*</span></small> |imagen-bandera=no |imagen-escudo=OECD logo.svg |emblema=Emblema |lema='''Mejores polí...
    </text>
  </revision>
</page>

```

Figura 4-94: Visualización del título en el tag title

En el resultado principal del programa podemos observar que la página previamente seleccionada para el análisis posee un PageRank de 113.59535 (figura 4.95).

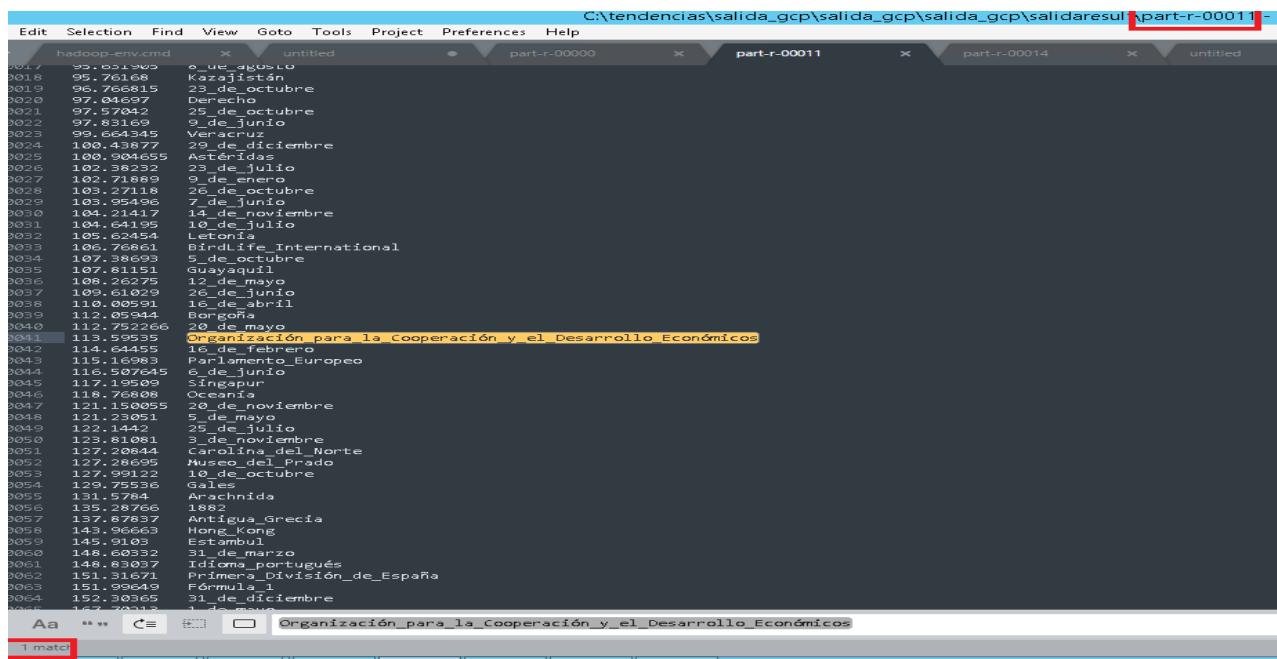


Figura 4-95: Pagerank de la página seleccionada para analizar

En la figura 4.96 se detalla una de las partes de la iteración 04 (Iteración escogida aleatoriamente); podemos observar que nuestra página es referenciada por ejemplo por la página cuyo título es "Biblioteca depositaria" (El programa cambió los espacios por underline).

C:\tendencias\salida_gcp\salida_gcp\salida_gcp\salidaalter04\part-r-00011 Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

path\00011

24515 Biblioteca bolivariana de merida 0.14999998 Biblioteca Bolivariana de Mérida
24516 Biblioteca_de_Autores_Cristianos 3.2351656 Nácar-Columba,regímen_franquista,Iglesia_católica,Editorial_Católica,Vulgata,Universidad_Pontificia_de_Salamanca,Ángel_Herrera_Oria,Jose_María_Sánchez_de_Munain,
24517 Biblioteca_de_El_Escorial 0.36600008 Real_Biblioteca_del_Monasterio_de_San_Lorenzo_de_El_Escorial
24518 Biblioteca_de_Olmedo 0.15100814 Bernar_Bíaz_del_Castillo,Historia_verdadera_de_la_conquista_de_la_Nueva_España,México,Hernán_Cortés,Valladolid,Alcalá_de_Henares,Valladolid,este,Bartolomé_de_Olmedo,Miguel_I
24519 Biblioteca_de_los_quijote 0.14999998 Biblioteca_de_Don_Quijote
24520 Biblioteca_de_la_Universidad_Adolfo_Ibañez 0.14999998 Biblioteca_de_la_Universidad_Adolfo_Ibañez
24521 Biblioteca_de_la_lengua_complementaria 0.14999998 Biblioteca_de_la_lengua_complementaria
24522 Biblioteca_de_la_música_gregoriana 0.14999998 Biblioteca_de_la_música_gregoriana
24523 Biblioteca del Casino de Manresa 0.18204801 provincia de Barcelona,Modernismo_catalán,Ignasi_Oms_i_Ponsa,Casino_de_Manresa,burguesia,Diputación_de_Barcelona,Manresa,1999,Biblioteca,_La_Gaixa
24524 Biblioteca del Condado de St_Leous 0.14999998 Biblioteca del Condado de San Luis
24525 Biblioteca del Estado_Ruso 1.5458322 Biblioteca Nacional_Puse,San_Petersburgo,libros,Dostoevsky,lengua_rusa,URSS,transliteración,biblioteca_nacional,Rusia,Federación_de_Rusia,Moscú
24526 Biblioteca del Instituto_Internacional_de_Madrid 0.18030358 Instituto_Internacional_de_Madrid,Instituto_Internacional_de_Madrid,Instituto_Internacional_de_Madrid,Residencia_de_Señoritas
24527 Biblioteca del Instituto_Nacional 0.76525897 Diego_Portales,Diego_Barros_Arenas,Eusebio_Lillo,Arturo_Frat,Patricio_Aylwin,Salvador_Allende,Eusebio_Lillo,Francisco_Bilbao,Gabriela_Mistral,Bernardo_O'Higgins,J
24528 Biblioteca del Monasterio_de_Admont 0.14999998 Admont,Austria,Abadia_de_Admont,Ilustración,Salzburgo,azobispo,Abadia_de_San_Pedro
24529 Biblioteca del Parque_Posadas 0.27749997 Organización_sín_Anexo_de_Lucreto,1975,25_de_agosto,Jorge_Batlle,Presidente_de_la_Republca,2000,nylon,Día_del_Libro,orientales,José_Artigas,1975,25_de_agosto,Uruguay,
24530 Biblioteca_depositaria 0.15166226 Organismo_internacional,Biblioteca_Nacional,Archivo_Nacional,Organización_de_Estados_Americanos,Organización_para_la_Cooperación_y_el_Desarrollo_Económicos,Unión_Europea,Bib
24531 Biblioteca_digital 11.621800 Real_Academia_Nacional_de_Medicina,Biblioteca_Virtual_de_la_Real_Academia_Nacional_de_Medicina,2002,Biblioteca_de_la_Universidad_de_Antioquia,2012,Ministerio_de_Defensas,Bibli
24532 Biblioteca_digital_de_humanidades 0.19409396 Humanidades,Universidad_Veracruzana,México,2010,Impuesto_Migración_humana,Estados_ Unidos_de_América,Poder_legislativo,Análisis_cuantitativo,Veracruz,Arizona,
24533 Biblioteca_digital_del_patrimonio_iberoamericano 0.14999998 Biblioteca_Digital_del_Patrimonio_Iberoamericano
24534 Biblioteca_digital_de_literatura 0.14999998 Biblioteca_digital_de_literatura
24535 Biblioteca_estándar_de_h... 0.13407252 WPS_Scholar,h.Tipo_de dato_lógico,h.Tipo_de dato_entero,stardg,h,EGLIBC,klibc,newlib,Biblioteca_estándar_de_C++,signah,Linux,uclibc,dietlibc,tipo_de datos,mac
24536 Biblioteca_jordi_rubio_i_balaguer 0.14999998 Biblioteca_Jordi_Rubio_i_Balaguer
24537 Biblioteca_mario_de_andrade 0.14999998 Biblioteca_Mario_de_Andrade
24538 Biblioteca_mitológica 16.12057 Griego_antiguo,eu.Wikisource.org/wik/CE%F8%24%CE%9B%CE%82%CE%BB%CE%9B%CE%BF%CE%80%CE%AE%CE%BA%CE%87_Texto_griego,Wikisource,PDF,Junta_de_Extramadura,Morfología_lingüística,
24539 Biblioteca_nacional_Francisco_Gavidia 0.14999998 Biblioteca_Nacional_Francisco_Gavidia
24540 Biblioteca_nacional_Francisco_Rey_Fahd 0.17244035 Riad,Biblioteca_nacional_Ivan_Vasovi,idioma_arabe,Biblioteca_nacional_de_Vietnam,Biblioteca_nacional_de_Vanuatu,1986,1983,Biblioteca_nacional_eسلو伐كية,Arabia_Saudita
24541 Biblioteca_nacional_central_de_florence 0.14999998 Biblioteca_Nacional_Central_de_Florence
24542 Biblioteca_nacional_de_Costa_de_Marfil 0.16813046 2008,2006,Costa_de_Marfil,idiomas_français,Cultura_de_Costa_de_Marfil,Abiyán,2009,Abiyán
24543 Biblioteca_nacional_de_Moldavia 0.16403274 Chisinaú,idioma_rumano,Cultura_de_Moldavia,Moldavia,Chisinaú
24544 Biblioteca_nacional_de_guatemala 0.14999998 Biblioteca_Nacional_de_Guatemala
24545 Biblioteca_nacional_de_polonia 0.14999998 Biblioteca_Nacional_de_Polonia
24546 Biblioteca_nacional_del_peru 0.14999998 Biblioteca_Nacional_del_Peru

Endfind 6 hits

Search "Organización para la Cooperación y el Desarrollo Económicos" (6 hits in 1 file)

C:\tendencias\salida_gcp\salida_gcp\salida_gcp\salidaalter04\part-r-00011 (6 hits)

Line 24530: Biblioteca_depositaria 0.15166226 Organismo_internacional,Biblioteca_Nacional,Archivo_Nacional,Organización_de_Estados_Americanos,Organización_para_la_Cooperación_y_el_Desarrollo_Económicos
Line 28087: Busan 23.931995 18_de_septiembre,GyeongsangdelSur,península_de_Cores,Japón,Taishimbo,(isla),Fukuoka,nar_de_China,Oriental,rio_Nakdong,Gyeongjeongsan,Clima_subtropical,Des
Line 53983: Dirección_General_de_Políticas_de_desarrollo_sostenible 0.16936485 desarrollo_sostenible,Ministerio_de_Asuntos_Exteriores_y_de_Cooperación,Administración_Pública_de_España,
Line 57628: Economía_de_Italia 1.2991914 - Producto_Interno_Bruto,USD,billón,Dolar,2017,USD,billón,Dolar,2017,italiana,2010,Fondo_Monetario_Internacional,Banco_Mundial,Wikipedia,Anos,1
Line 77277: Gerardo_Jiménez_Sánchez 0.14999998 Organización_para_la_Cooperación_y_el_Desarrollo_Económicos,Johns Hopkins University,Instituto_Nacional_de_Medicina_Genómica_en_América_L
Line 136006: Organización_for_Economic_Co-operation_and_Development 0.60728504 Organización_para_la_Cooperación_y_el_Desarrollo_Económicos

Figura 4-96: Parte de resultado de iteración 04

En otras partes de la iteración 04 (Iteración escogida aleatoriamente), podemos observar en la figura 4.97 que nuestra página es referenciada por ejemplo por la página cuyo título es "Gerardo Jiménez Sánchez".

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Windows ? C:\tendencias\salida_gcp\salida_gcp\salida_gcp\salida_dalter04\part-r-00011 Notepad++

part-r-00011

77261	Gerard_Sevillano	0.14999998	Força_Lleida_Club_Esportiu,Força_Lleida_Club_Esportiu,Club_Joventut_de_Badalona,Força_Lleida_Club_Esportiu,24_de_abril,1994,Barcelona,Força_Lleida_Club_Esportiu,LEB
77262	Gerard_Van_Swieten	0.17938329	Gerard_van_Swieten
77263	Gerard_van_Swieten	0.14999998	Gerard_van_Swieten
77264	Gerard_van_Spanedonck	0.14999998	Gerard_van_Spanedonck
77265	Gerardo_Amarilla	0.33240534	Partido_Nacional_(Uruguay),1969,Cámara_de_Representantes_(Uruguay),2016,1_de_marzo,departamento_de_Rivera,Universidad_Internacional_de_Andalucía,Universidad_de_la_Re
77266	Gerardo_Araña	0.15543213	Guadalope_Nettel,Juan_Villoro,Roberto_Bolívar,Javier_Sicilia,Radiador_Magazine,Cristina_García
77267	Gerardo_Arvalos	0.20001999	delantero,futbolista,1987,3_de_agosto,Paraguay,Cacique,Priera,División_de_Paraguay,Club_Sportivo_Luqueño,Delantero,Sportivo_Trinitense,2008,metro,Paraguay,Cacique,I
77268	Gerardo_Concejo	0.31276405	Ciudad_Oregon,California,Universidad_de_California_en_Los_Angeles,Nueva_York,Universidad_de_Columbia,1982,Sociedad_General_de_Escritores_de_Méjico,UNN,Universidad_Naci
77269	Gerardo_Cortés	0.45621586	2013,Club_de_Deportes_La_Serena,2014,Club_de_Deportes_Santiago_Morning,2014,2016,Coquimbo_Union,2017,10_de_abril,2016,Competición_(juego),Gol,Gol,Competición_(juego),Gol
77270	Gerardo_Gandini	1.5049497	Alberto_Ginastera,director_de_orquesta,compositor,pianista,2013,22_de_marzo,Buenos_Aires,1936,16_de_octubre,Argentina,Buenos_Aires,2000,1978,Piano,profesor,director_de_c
77271	Gerardo_García_(deportista)	0.14999998	Gerardo_García
77272	Gerardo_González	0.15660038	Gerardo_González_(boxeador)
77273	Gerardo_González	0.23453635	Gerardo_González_(actor),Gerardo_González_de_Vega,Gerardo_González_Aquino,Gerardo_González_(boxeador)
77274	Gerardo_Grote	0.14999998	Gerardo_Grote
77275	Gerardo_II_de_Rosellón	1.19121423	Gerardo_II
77276	Gerardo_Jáñez_Sánchez	0.14999998	Organización_para_la_Cooperación_y_el_Desarrollo_Económicos,John_Hopkins_University,Instituto_Nacional_de_Medicina_Genómica_en_América_Latina,Instituto_Nacional
77277	Gerardo_Jáñez_Sánchez	0.14999998	Organización_para_la_Cooperación_y_el_Desarrollo_Económicos,John_Hopkins_University,Instituto_Nacional_de_Medicina_Genómica_en_América_Latina,Instituto_Nacional_de_Futbol_de_Méjico,1955,Club_de_Fútbol_Atlante,1987,Club_de_Fútbol_Atlante,selección_de_fútbol_de_Méjico,1987,13_de_marzo,1987
77278	Gerardo_López_Solís_Gómez	0.14999998	Gerardo_Octavio_Solís_Gómez
77279	Gerardo_Octavio_Solís_Gómez	0.14999998	Puebla,Méjico,escritor
77280	Gerardo_Oribe	0.15873606	1928_Nueva_York,Tánger,La_Línea_de_la_Concepción,2008,Provincia_de_Cádiz,1948,filólogo,Odón_Betanzos_Palacios,Academia_Norteamericana_de_la_Lengua_Española,escritor,F
77281	Gerardo_Piñales	0.35215238	Nueva_York,Tánger,La_Línea_de_la_Concepción,2008,Provincia_de_Cádiz,1948,filólogo,Odón_Betanzos_Palacios,Academia_Norteamericana_de_la_Lengua_Española,escritor,F
77282	Gerardo_Pérez	0.3077106	Club_Atlético_Tiro_Federal,Club_Atlético_Rosario_Central,Gol,Competición_(juego),Gol,Competición_(juego),Gol,Competición_(juego),Central_Córdoba
77283	Gerardo_Rubén_Morales	0.15113898	Gerardo_Morales
77284	Gerardo_Sosa_Castañel	0.3521785	Juan_Manuel_Camacho_Berrán,Méjico,Hidalgo_(Méjico),Organización_Editorial_Mexicana,1991,1998,Juan_Manuel_Menes_Llaguno
77285	Gerardo_Torrado	0.7976404	Copa_Oro,Copa_FIFA_Confederaciones_Fútbol,1999,10_de_junio,Club_Universidad_Nacional,1997,6_de_septiembre,Mediocampista,North_American_Soccer_League_(II),2017,Agosto,25,
77286	Gerardo_Valecina_Cam	0.16951197	1956,1959,Mejican_Tolima,Golconde_(Colombia),Chocó,1972,21_de_enero,1972,26_de_agosto,Antioquia,Santo_Domingo_(Antioquia),Bogotá,1949,Vaupés,24_de_mayo,1953,Bogotá
77287	Gerardo_de_Jesús_Rojas_López	0.14999998	Gerardo_de_Jesús_Rojas_López
77288	Gerardo_dos_Santos	0.14999998	Gerardo_Franco,Cosme_dos_Santos
77289	Gerardo_Engel	0.14999998	1956,1959,Verardo_Bugaló
77290	Gerardo_F	0.58977175	ramírez,sinonimia,(biología),mstm,Costa Rica,Zootrophion,labelo,Pleurothallis_fulgens,inflorescencia,peciolada,hoja,vaina,epífito,Pleurothallis,Costa_Rica,Luer,Gerardo_a_m
77291	Geranthanthus	0.33545247	Cordia
77292	Geranthanthus_excellens	0.14999998	Cordia_trichotoma

Figura 4-97: Resultado de iteración 04; relación con la página de Gerardo Jiménez Sánchez

En la figura 4.98 se observa en Wikipedia la página de Gerardo Jiménez Sánchez, la cual contiene un enlace a nuestra página (Organización para la Cooperación y el Desarrollo Económicos); la referencia se realiza utilizando la sigla OCDE.

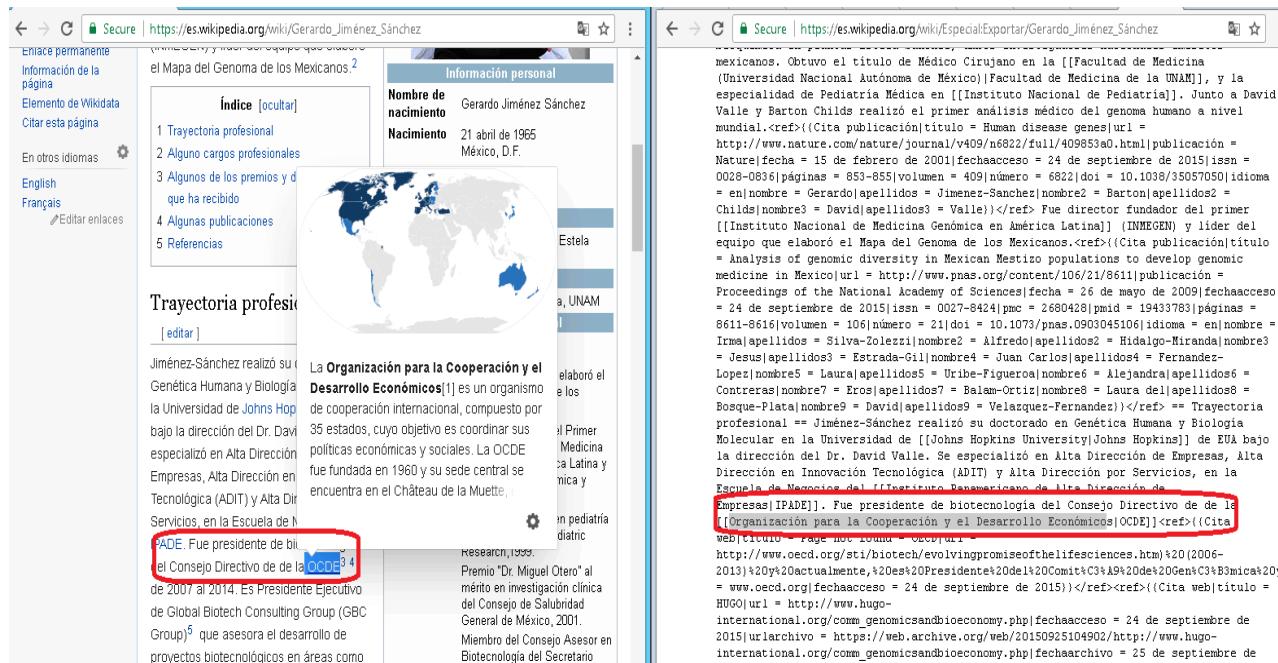


Figura 4-98: Referencia desde página de Gerardo Jiménez Sánchez

De igual manera podemos ver como nuestra página es referenciada desde la página denominada "Biblioteca depositaria" (Figura 4.99).

The figure consists of two side-by-side screenshots of a web browser. The left screenshot shows the Wikipedia article titled 'Biblioteca depositaria'. It features a large world map at the top, followed by several paragraphs of text. The right screenshot shows the same article's edit history page. It lists a single edit made by the bot 'BenjaBot' on October 29, 2016, at 12:27Z. The edit comment was '(Bot) Correcciones ortográficas'.

Figura 4-99: Referencia desde página Biblioteca depositaria

En la figura 4.100, se observa que en una parte de la Iteración 0 (Escogida aleatoriamente) nuestra página aparece referenciando muchas otras páginas como por ejemplo la de Brasil y la de Lituania.

The figure is a screenshot of a search results page. The search term used was 'Organización para la Cooperación y el Desarrollo Económicos'. The results are listed in a grid format, showing 149 hits across 13 files. The results include various organizations from around the world, such as 'Organización municipal de Catamarca', 'Organización política-administrativa de Alicante', and 'Organización territorial de Etiopía'. Each result entry includes a link to the full record.

Figura 4-100: Referencia a otras páginas

En la figura 4.101 podemos observar como en Wikipedia se reflejan las referencias que están en nuestra página hacia la página cuyo título es José Ángel Gurría.

The figure consists of two side-by-side screenshots of web browsers. The left screenshot shows the Wikipedia page for 'Organización para la Cooperación y el Desarrollo Económicos' (OECD). A red box highlights a link to 'José Ángel Gurría'. The right screenshot shows the Wikipedia page for 'José Ángel Gurría'. A red box highlights a link back to the OECD page. Both pages contain extensive text and lists of references.

Figura 4-101: Comparación de enlaces hacia la página de José Ángel Gurría

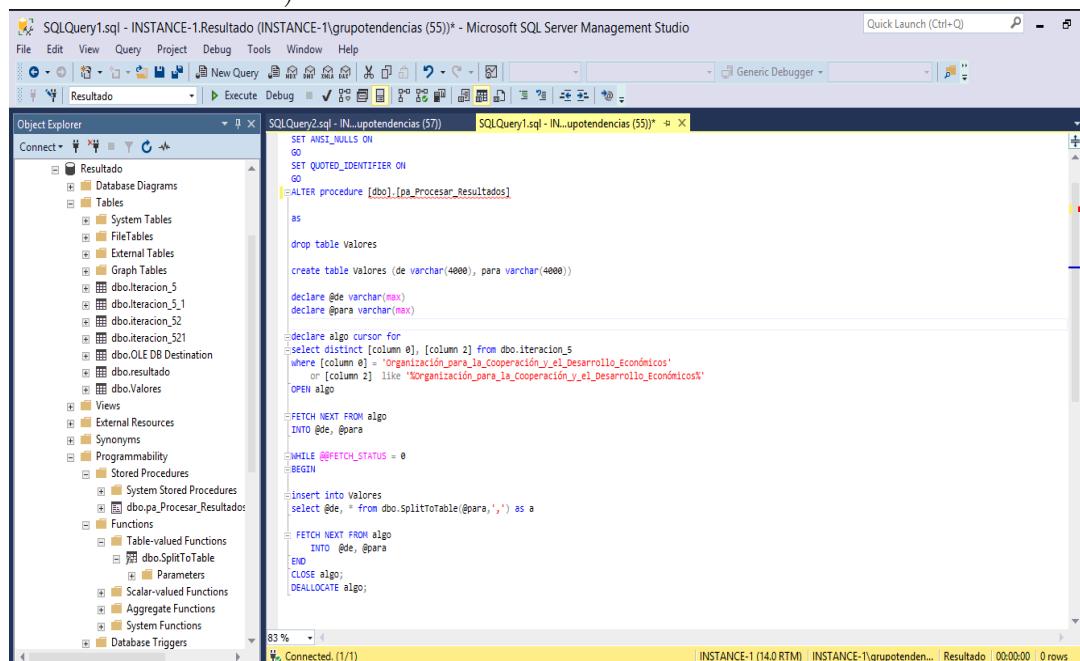
De igual manera se puede observar en Wikipedia como se reflejan las referencias que están en nuestra página hacia la página cuyo título es StatLinks (Figura 4.102)

The figure consists of two side-by-side screenshots of web browsers. The left screenshot shows the Wikipedia page for 'StatLinks'. A red box highlights a link to 'Organización para la Cooperación y el Desarrollo Económicos' (OECD). The right screenshot shows the Wikipedia page for 'Organización para la Cooperación y el Desarrollo Económicos' (OECD). A red box highlights a link to 'StatLinks'. Both pages contain extensive text and lists of references.

Figura 4-102: Comparación de enlaces hacia la página de StatLinks

4.7.7. Crear y Ejecutar Procedimiento almacenado y Función

El procedimiento almacenado (que invoca la función) permite tomar la información relacionada a la página escogida (Organización para la Cooperación y el Desarrollo Económicos) de la última iteración del proceso. Lo anterior es con el fin de poder realizar posteriormente una consulta que genera la entrada requerida en R (para el cálculo del PageRank y la visualización del resultado).



The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left lists database objects like Resultado, Tables, System Tables, etc. The central pane displays a T-SQL script for creating a stored procedure named `[dbo].[pa_Procesar_Resultados]`. The script includes declarations for variables `@de` and `@para`, a cursor named `algo` that selects distinct values from the `dbo.iteracion_5` table where column 0 contains 'Organización para la Cooperación y el Desarrollo Económicos' or column 2 contains 'Organización para la Cooperación y el Desarrollo Económicos', and a loop that inserts rows into the `Valores` table using the `dbo.SplitToTable` function. The status bar at the bottom indicates the connection is 'Connected. (1/1)'.

```

SQLQuery1.sql - INSTANCE-1.Resultado (INSTANCE-1\grupotendencias (55)) - Microsoft SQL Server Management Studio
File Edit View Query Project Debug Tools Window Help
Resultado | Execute Debug | ✓ |  |
SQLQuery2.sql - IN...upotendencias (57) | SQLQuery1.sql - IN...upotendencias (55)* | X
Object Explorer | SQLQuery2.sql - IN...upotendencias (57) | SQLQuery1.sql - IN...upotendencias (55)* | X
File Edit View Query Project Debug Tools Window Help
Resultado | Execute Debug | ✓ |  |
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER procedure [dbo].[pa_Procesar_Resultados]
as
drop table Valores
create table Valores (de varchar(4000), para varchar(4000))
declare @de varchar(max)
declare @para varchar(max)
declare algo cursor for
select distinct [column 0], [column 2] from dbo.iteracion_5
where [column 0] = 'Organización para la Cooperación y el Desarrollo Económicos'
or [column 2] like 'Organización para la Cooperación y el Desarrollo Económicos'
OPEN algo
FETCH NEXT FROM algo
INTO @de, @para
WHILE @@FETCH_STATUS = 0
BEGIN
insert into Valores
select @de, * from dbo.SplitToTable(@para,',') as a
FETCH NEXT FROM algo
INTO @de, @para
END
CLOSE algo;
DEALLOCATE algo;

```

Figura 4-103: Procedimiento almacenado en SQL Server

```

CREATE FUNCTION [dbo].[SplitToTable]
(
    @cadena as varchar(4000), @Delimitador varchar(1)
)

returns @ValueTable table ([Value] nvarchar(4000))

AS

begin
declare @extString nvarchar(4000)
declare @Pos int
declare @extPos int
declare @CommaCheck nvarchar(1)

--Inicializa
set @extString = ''
set @CommaCheck = right(@cadena,1)

set @cadena = @cadena + @Delimitador

--Busca la posición del primer delimitador
set @Pos = charindex(@Delimitador,@cadena)
set @extPos = 1

--Itera mientras exista un delimitador en el string
while (@Pos <> 0)
begin
    set @extString = substring(@cadena,1,@Pos - 1)

    insert into @ValueTable ([Value]) Values (@extString)

    set @cadena = substring(@cadena,@Pos + 1,len(@cadena))

    set @extPos = @Pos
    set @Pos = charindex(@Delimitador,@cadena)
end
return

```

Figura 4-104: La función en SQL Server

```

exec [pa_Procesar_Resultados]

```

The Messages pane shows the following output:

```

(144 rows affected)
(4 rows affected)
(22 rows affected)
(247 rows affected)
(274 rows affected)
(273 rows affected)
(272 rows affected)

```

Figura 4-105: Ejecución del procedimiento almacenado

4.7.8. Exportar muestra de iteración resultante de procedimiento almacenado

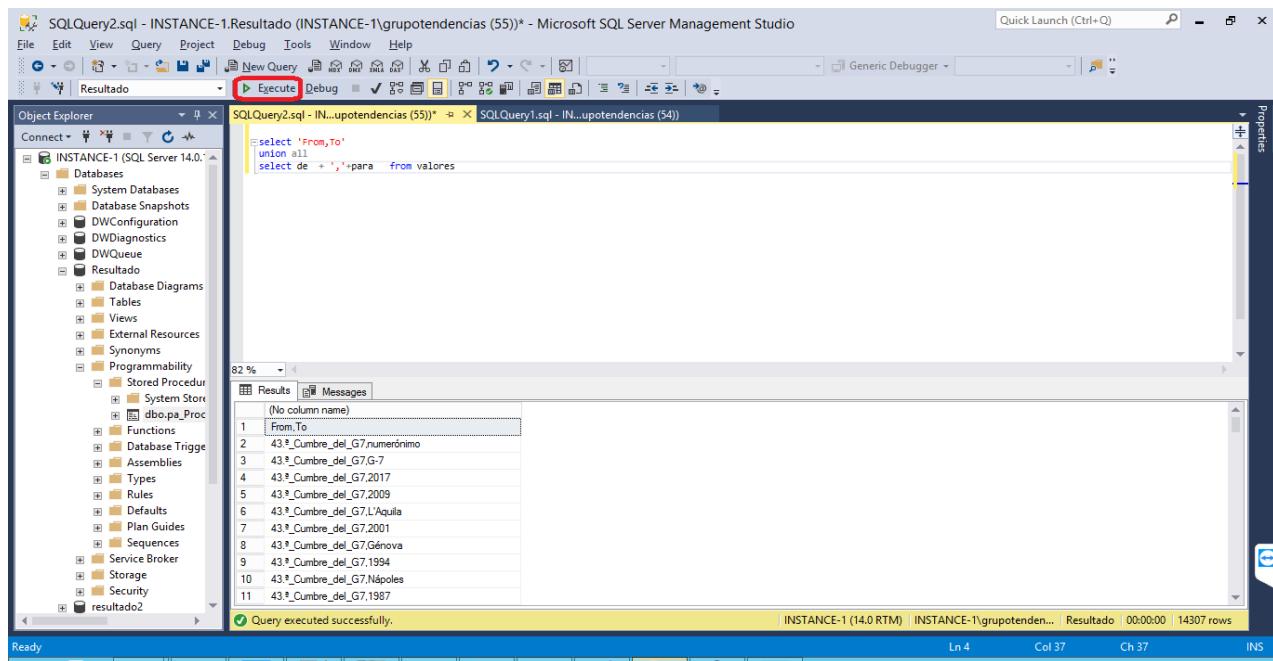


Figura 4-106: Ejecutar la consulta para generar la muestra

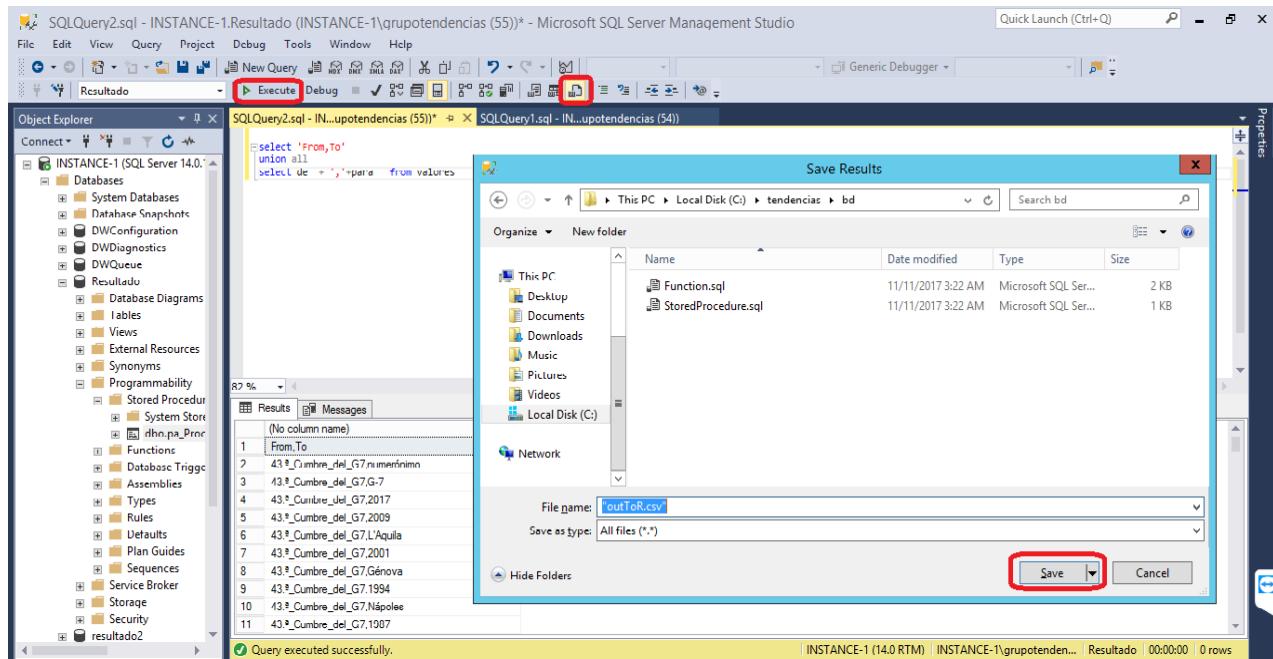
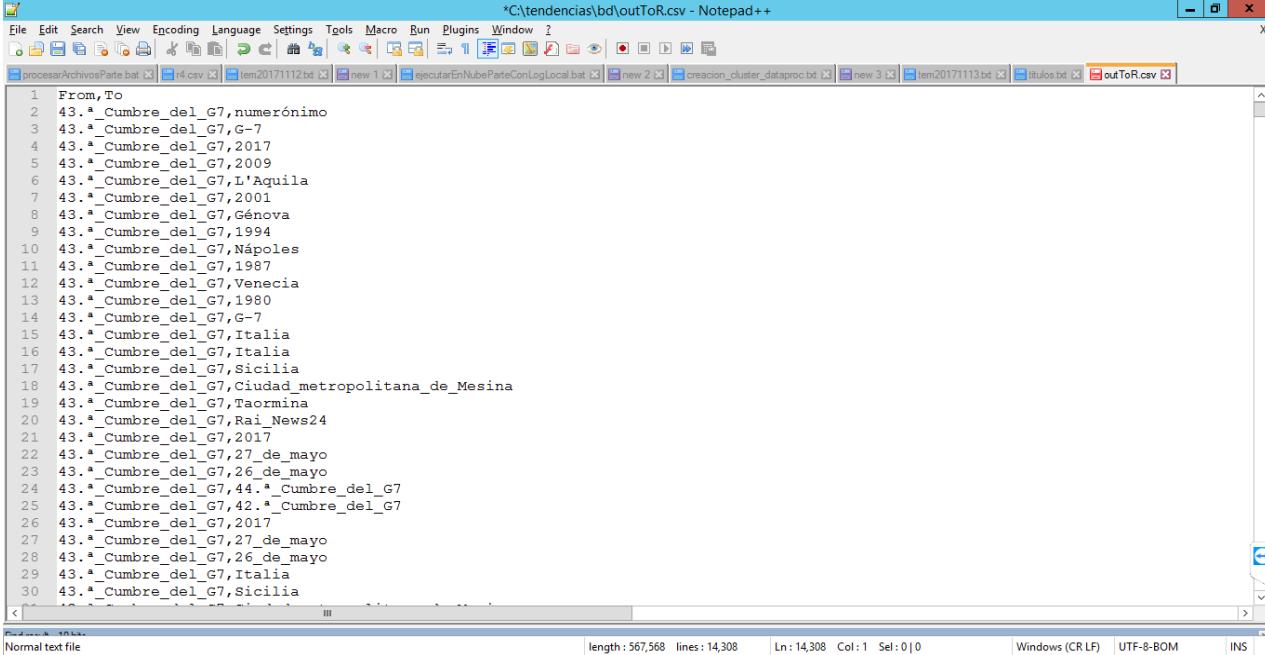


Figura 4-107: Guardar el resultado de la consulta en un archivo .csv



The screenshot shows a Notepad++ window with the title bar "C:\tendencias\bd\outToR.csv - Notepad++". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Tools, Macro, Run, Plugins, Window, and Help. The toolbar contains icons for file operations like Open, Save, Print, and Find. The status bar at the bottom displays "Normal text file", "length : 567,568 lines : 14,308 Ln : 14,308 Col : 1 Sel : 0 | 0", "Windows (CR LF)", "UTF-8-BOM", and "INS". The main text area contains approximately 30 numbered lines of text, mostly starting with "43.*_Cumbre_del_G7," followed by various locations and years. The lines are as follows:

```
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
```

Figura 4-108: Muestra final de iteración

4.8. Utilización de R

4.8.1. Importar csv de muestra de datos de iteración en R

Cargue de la última iteración en R para el caso escogido

```
EntradaR3 <-read_delim("C:/tendencias/r4.csv", ",",
escape_double = FALSE, trim_ws = TRUE)
```

4.8.2. Procesar datos en R

Procesamiento y análisis de resultados en R para el caso escogido

Gracias a R podemos observar de manera gráfica la relación entre las diferentes páginas de wikipedia, teniendo en cuenta el PageRank.

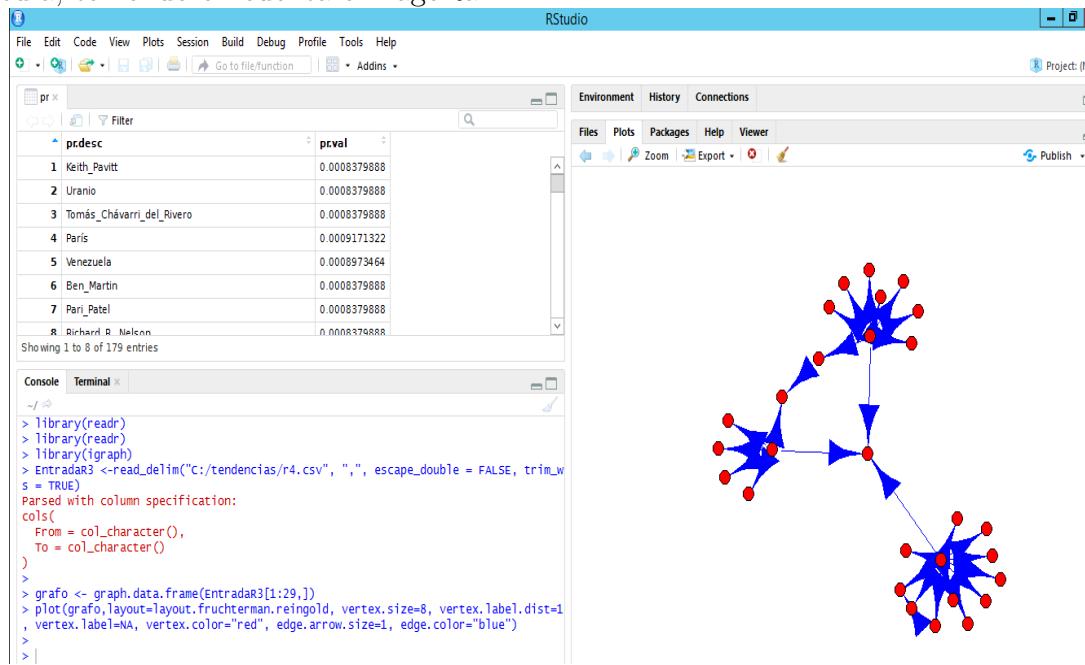


Figura 4-109: Vizualización de pequeña muestra de resultados

4.8.3. Graficar datos en R

Análisis funcional del resultado de R

Como se puede observar en la gráfica, cada nodo representa una página de Wikipedia (la llave es el título), y los enlaces entre los nodos corresponden a los links que hay desde cada página de Wikipedia a otras páginas.

Los nodos con mayor área representan las páginas de Wikipedia con mayor PageRank (calculado previamente teniendo en cuenta por ejemplo la cantidad de páginas que tienen enlace hacia la calificada y los enlaces que esta última tiene hacia otras).

Dada la gran cantidad de nodos es difícil visualizar el detalle de toda la gráfica, por lo cual es recomendable analizar de manera independiente diferentes muestras de la tabla del resultado aplicando filtros de interés (ejemplo: el top 10 de las páginas con mayor PageRank).

Posteriormente, se podrían analizar algunas páginas de Wikipedia de interés (dados los filtros aplicados) para determinar si existe algún patrón funcional o algunas características especiales que produzcan cierto efecto sobre el PageRank de dichas páginas. Ejemplo: podría llegar a concluirse que las páginas con mayor cantidad de referencias son aquellas cuyo contenido brinda mayor confianza.

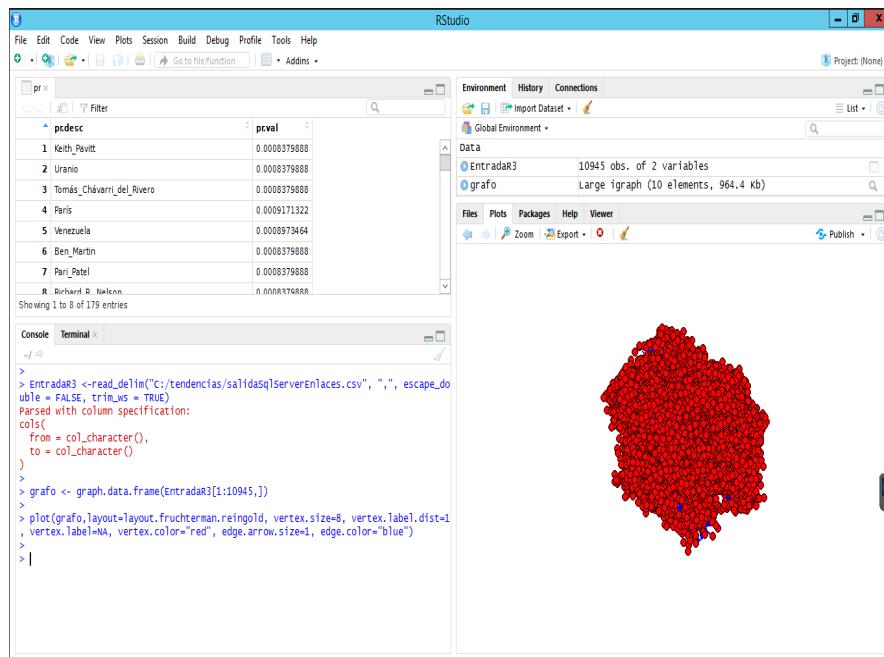


Figura 4-110: Vizualización del grafo

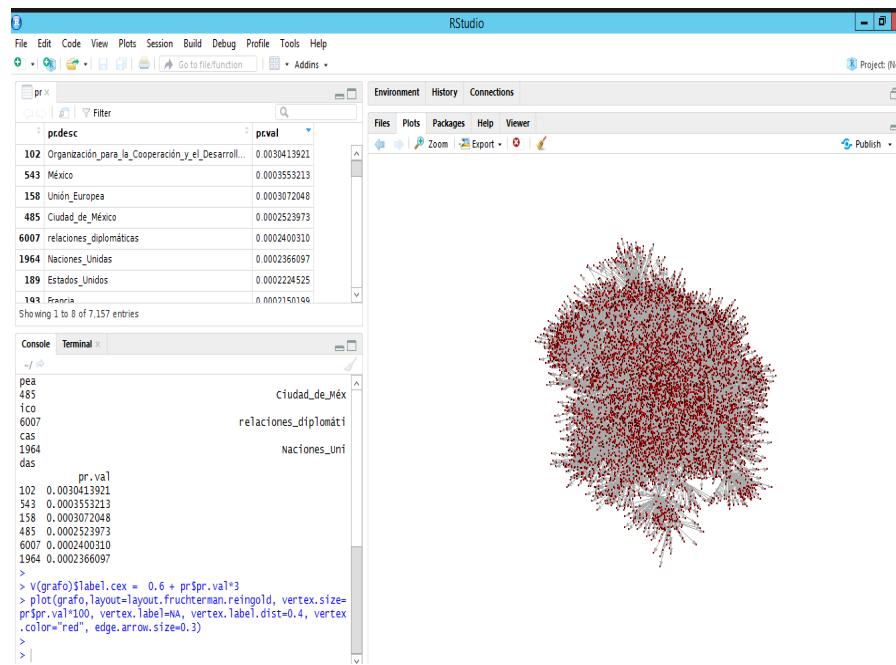


Figura 4-111: Visualización aplicando PageRank

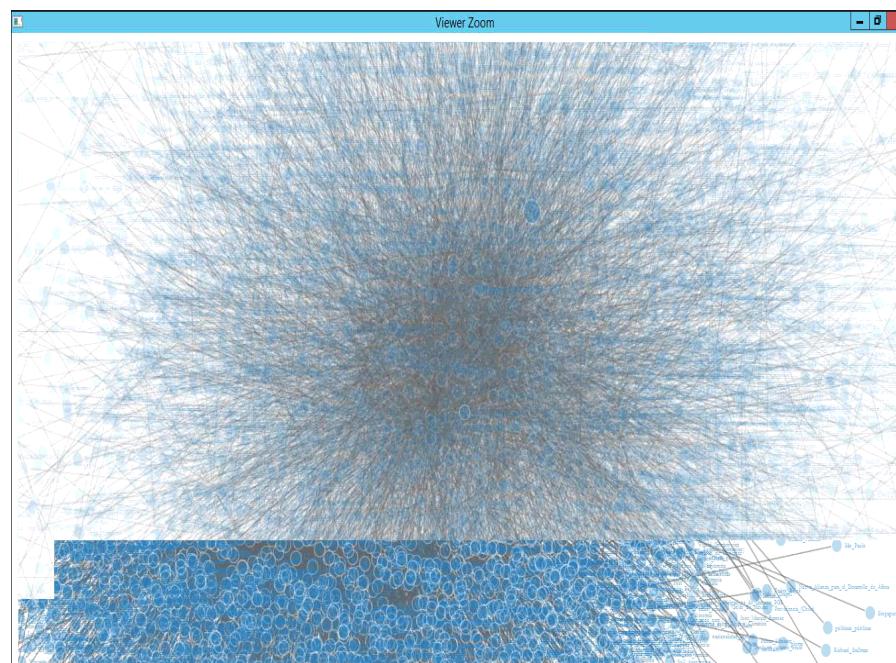


Figura 4-112: Visualización 3D aplicando PageRank

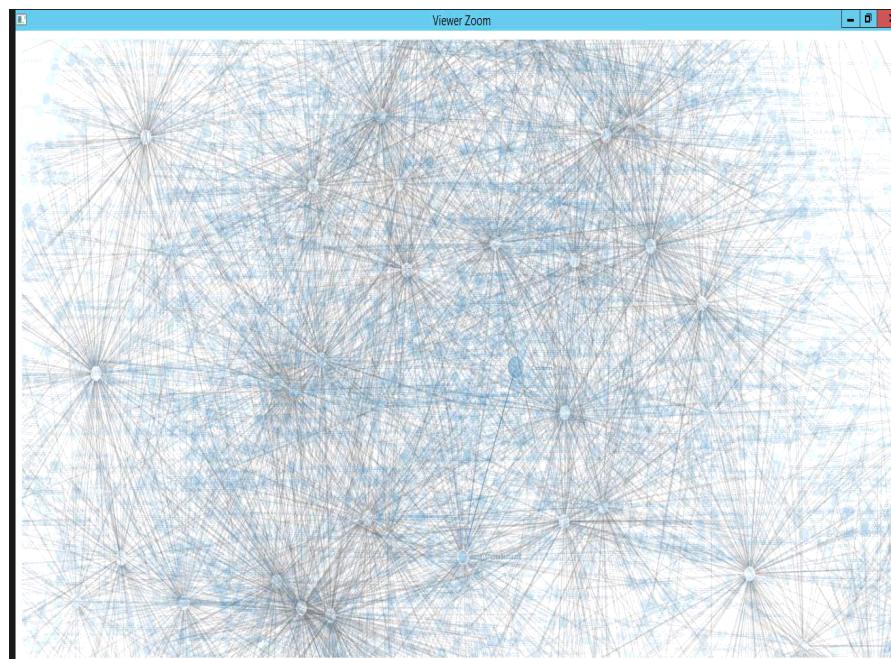


Figura 4-113: Visualización 3D aplicando PageRank con menor zoom

5 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 logró 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, entre otros.
- Aplicando el algoritmo de PageRank se determinaron cuáles páginas resultaban ser más 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.
- Como ejemplo, cabe mencionar que cuando Google está realizando diferentes búsquedas, está teniendo en cuenta el PageRank calculado en este proyecto.
- 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

- [1] APACHE. *welcome to apache hadoop*. 2017
- [2] GARCÍA LLORENTE, Daniel: *Procesamiento masivo de datos vía Hadoop*, La Universidad Politécnica de Cataluña (Universitat Politècnica de Catalunya),Facultat d'Informàtica de Barcelona (FIB), Tesis de Grado, 2017
- [3] LESKOVEC, J. ; RAJARAMAN, A. ; ULLMAN, J.D.: *Mining of Massive Datasets*. Cambridge University Press, 2014. – ISBN 9781107077232
- [4] LOCKWOOD, Glennk. *conceptual overview of map-reduce and hadoop*. 2015
- [5] CLOUD PLATFORM, Google. *por qué la infraestructura de google cloud es ideal para tu empresa*. 2017