
Universidad
Tecmilenio®

Universidad Tecmilenio

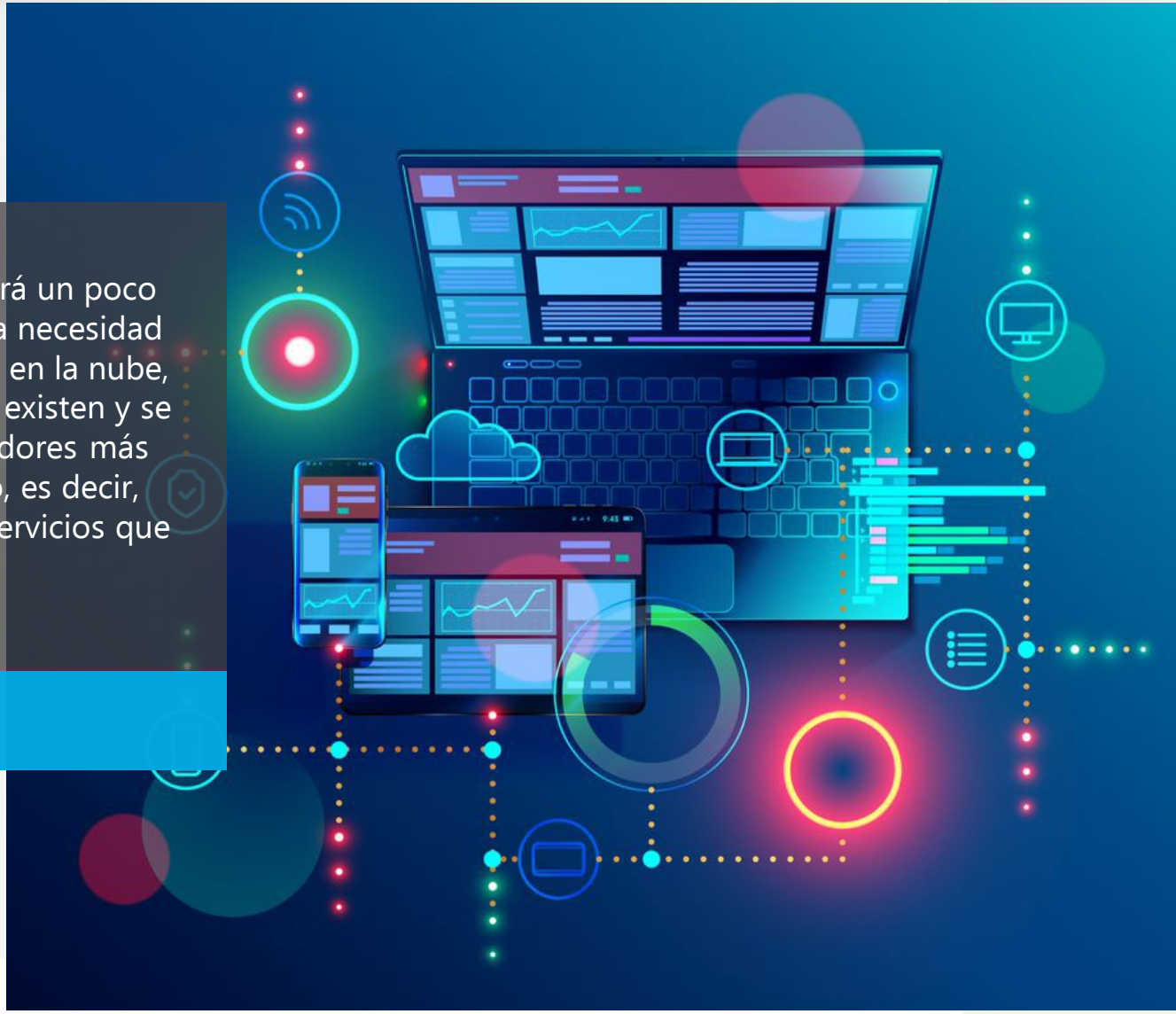
Infraestructura para Big Data

Introducción a Microsoft Azure

Semana 10

Introducción

En este tema se revisará un poco de historia acerca de la necesidad de contar con servicios en la nube, los tipos de nubes que existen y se abordarán los proveedores más grandes del mercado, es decir, Microsoft Azure y los servicios que ofrece.



Explicación

Cómputo en la nube, de acuerdo con el Instituto Nacional de Estándares y Tecnología de Estados Unidos (NIST, 2018), es un modelo para habilitar de manera conveniente y bajo demanda el acceso a través de la red a un grupo compartido de recursos de cómputo configurables, como servidores, almacenamiento, infraestructura de red, aplicaciones y servicios que puedan ser rápidamente provisionados y liberados con un esfuerzo mínimo de administración.

El modelo de cómputo en la nube se compone de cinco **características** esenciales (Newberry, 2017):

1. Autoservicio bajo demanda.
2. Acceso remoto.
3. Un catálogo de servicios.
4. Flexibilidad para crecer.
5. Servicios.

Tres tipos de servicios que se ofrecen (AWS, s.f.):

01. **PaaS (Platform as a Service):** el proveedor de cómputo en la nube ofrece una solución completa de infraestructura y aplicaciones a los clientes, un ejemplo es cuando las empresas contratan un servicio para contar con una aplicación como SAP, la cual se encuentra hospedada en los servidores del proveedor del servicio.

02. **IaaS (Infraestructure as a Service):** ofrece acceso a infraestructura de sistemas, como poder de cómputo, almacenamiento y equipos de comunicación.

03. **SaaS (Software as a Service):** se da acceso al uso de aplicaciones como un servicio, un ejemplo de esto puede ser la aplicación de correo que ofrece Google.

Explicación

Big Data y el cómputo en la nube

Para implementar una solución se requieren varios elementos de software y hardware, los cuales pueden llegar a necesitar mucha capacidad de cómputo, por lo cual, una de las estrategias que más se utilizan en las empresas que ya han implementado una solución de Big Data es hacer uso del cómputo en la nube, debido a la flexibilidad.

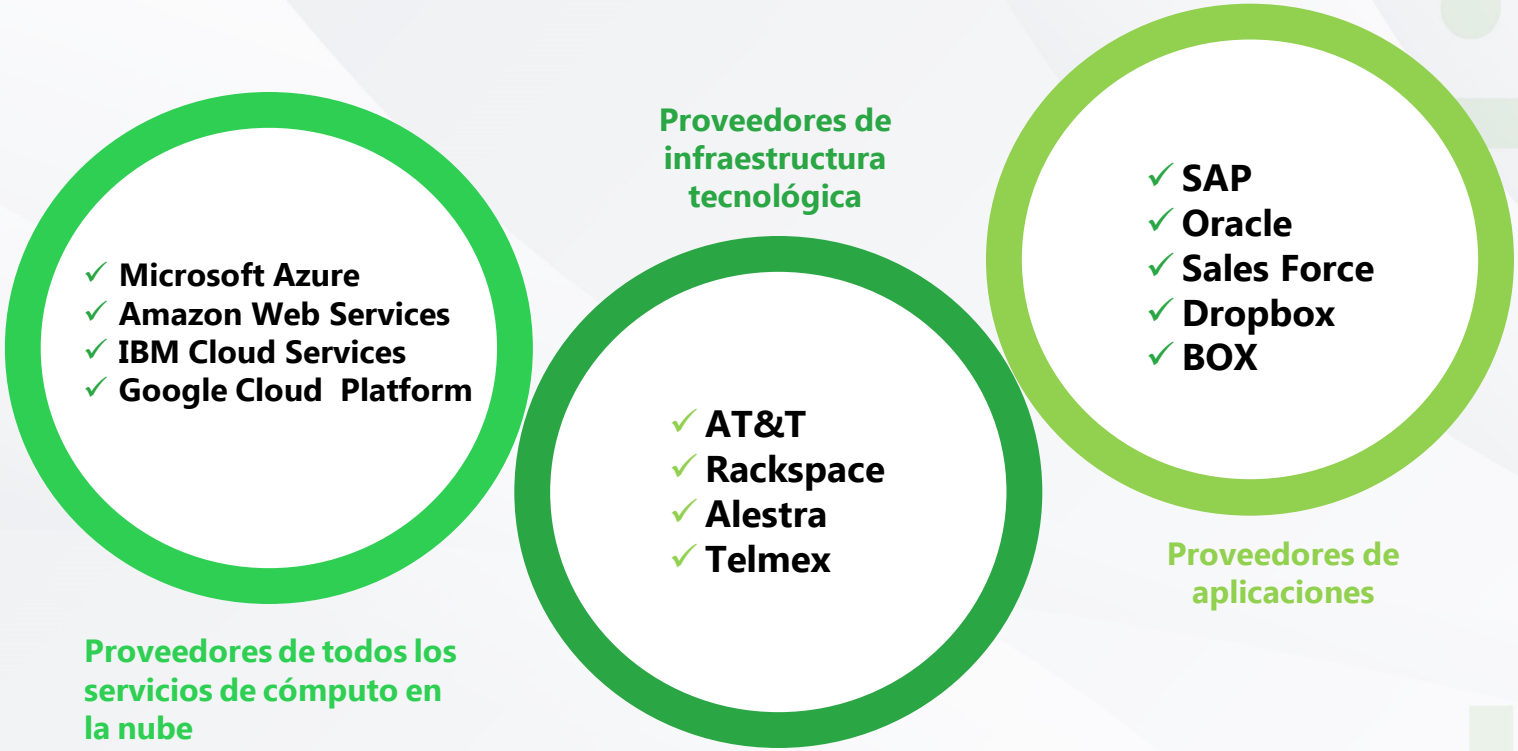
En la adopción de este modelo de cómputo en la nube podemos mencionar que existen tres tipos (Microsoft Azure, s.f.-a):

- La **nube pública** son aquellos servicios de cómputo en los que se contrata a un proveedor que puede ser de diversos tipos, y la empresa paga un cargo por un periodo de tiempo por su uso.
- La **nube privada** son aquellos servicios de cómputo con los que cuentan las empresas, ya sea en sus instalaciones o en las del proveedor del servicio.
- La **nube híbrida** es una combinación, es decir, una parte de los sistemas o infraestructura se encuentra en una nube pública y otra parte en la nube privada.

Explicación

Universidad Tecmilenio

Los **proveedores** de servicios de cómputo en la nube son variados y se separan en tres grandes grupos:



Explicación

Universidad Tecmilenio

Microsoft Azure vs. Amazon Web Services vs. Google

Microsoft Azure es el segundo proveedor de servicios en la nube en el mercado, está más abierto a ofrecer servicios para nubes públicas e híbridas.

AWS es el competidor más fuerte de servicios en la nube, su enfoque solo es en la nube pública.

Google Cloud tiene lento crecimiento, pero muy estable y con una experiencia en tecnología muy fuerte que lo hará muy atractivo en poco tiempo.

Explicación

Azure es un conjunto completo y en expansión constante de servicios de informática en la nube que ayuda a las organizaciones a afrontar sus desafíos empresariales.

Ofrece la flexibilidad de crear, administrar e implementar aplicaciones en una red mundial enorme con las herramientas y las plataformas que Microsoft tiene disponibles.

Con **Azure** se puede hacer lo siguiente (Microsoft Azure, s.f.-b):

- ✓ Prepararse para el futuro al tener soluciones innovadoras al alcance .
- ✓ Trabajar en un entorno híbrido (nube híbrida).
- ✓ Crear soluciones a la medida.
- ✓ Contar con una nube segura.

Soluciones que ofrece:

- Aplicaciones de nube híbrida.
- Bases de datos.
- Internet de las cosas.
- Aplicaciones de inteligencia artificial.
- SAP/Oracle en Azure.
- Blockchain.
- Herramientas de desarrollo de aplicaciones.
- Comercio electrónico.
- Análisis de datos.
- Aplicaciones para el negocio.
- Aplicaciones como servicio.
- Almacenamiento de datos.
- Plan de recuperación de desastres.
- Centro de datos.
- Infraestructura.
- Respaldos.

Cierre

El cómputo en la nube (*cloud computing*) es una tecnología que permite acceder de forma remota a archivos, software y procesamiento de datos mediante Internet, ya que no requiere de la instalación de computadoras locales adicionales.

Asimismo, este servicio puede utilizarse por empresas de todos tipos y tamaños.

Bibliografía

- AWS. (s.f.). *What is cloud computing?* Recuperado de https://aws.amazon.com/what-is-cloud-computing/
- Microsoft Azure. (s.f.-a). *What is cloud computing?.* Recuperado de https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/
- Microsoft Azure. (s.f.-b). *Información general.* Recuperado de https://azure.microsoft.com/es-mx/overview/#intro-to-cloud-computing
- NIST. (2018). *Final Version of NIST Cloud Computing Definition Published.* Recuperado de https://www.nist.gov/news-events/news/2011/10/final-version-nist-cloud-computing-definition-published
- Novkovic, G. (2017). *Five characteristics of cloud computing.* Recuperado de https://www.controleng.com/articles/five-characteristics-of-cloud-computing/

Universidad Tecmilenio

Infraestructura para Big Data

Configuración de Hadoop

Semana 10

Introducción

Distintas organizaciones que operan en diferentes giros, desde servicios bancarios hasta atención médica, se dirigen hacia una transformación completamente digital mediante el uso de Microsoft Azure.

En este tema se creará un clúster de Hadoop mediante la instalación del mismo en una máquina virtual.

Explicación

Crearás una cuenta para estudiantes en el servicio de la nube de **Microsoft**, que recibe el nombre de Microsoft Azure.

La cuenta para estudiantes de Microsoft Azure nos ofrece un crédito de 100 USD por 12 meses para utilizarlo en todos los recursos disponibles del servicio e incluye un gran número de aplicaciones y servicios gratuitos (Microsoft Azure, s.f.).

Para activar tu cuenta, entra al siguiente enlace:

Microsoft Azure. (s.f.).
Pruebe Azure for Students gratis hoy mismo. **Recuperado de** https://azure.microsoft.com/es-es/free/students/

Instalación de Hadoop

A continuación, se presentan los pasos para instalar Hadoop en una máquina virtual en la nube de tu elección, o bien, en tu equipo local.

1. Se generará una instancia o máquina virtual en la nube de tu elección.
2. Al crear las instancias en AWS, se deben asignar: Type: All traffic, source: anywhere.
3. En la línea de comandos entra a la primera de tus instancias EC2 e instala JAVA.
4. Asigna nombres a cada uno de los nodos.
5. Genera una llave pública y una privada para cada uno de los nodos.

Explicación

Asimismo, las siguientes acciones se realizarán por cada nodo:

1. Descargar Hadoop, revisando la versión a instalar en: https://hadoop.apache.org/releases.html
2. Para cada nodo, añade la siguiente línea: "HadoopInstallationFolder/etc/hadoop/core-site.xml" dentro del archivo de configuración. Modifica el archivo "core-site.xml" en el usuario *master* y en cada uno de los otros usuarios creados. Del mismo modo, reemplaza la IP con la dirección IP de tu nodo principal.
3. Limpia el contenido del archivo HadoopInstallation Folder/etc/hadoop/slaves y añade la IP privada de cada uno de los nodos *workers* y el master. Este paso especifica la lista de todas aquellas máquinas que pueden considerarse workers para el master.
4. Recuerda asignar la variable de entorno JAVA_HOME en el archivo: HadoopInstallationFolder/etc/hadoop/hadoop-env.sh
5. Por medio de línea de comandos en el master, entra al directorio HadoopFolder/bin/ y corre el siguiente comando para formatear al sistema de archivos HDFS:
   ./hadoop namenode -format
6. En el master, en el directorio HadoopFolder/sbin/ corre lo siguiente en línea de comando:
   . /start-dfs.sh
7. Verifica el estatus del clúster, escribiendo master_public_ip:50070 en un explorador web. Si una página con el mensaje "Namenode Information" aparece, quiere decir que has iniciado el clúster correctamente.

Acceder a los servidores implementados

Todas las máquinas virtuales (VM) son accesibles desde direcciones externas vía SSH por el puerto 22 desde su respectiva IP pública. Por ende, la lista de todos los puertos abiertos en los servicios de Hadoop puede encontrarse en Network Security Group, implementado en el grupo de recursos.

Explicación

Recuperar la clave privada SSH y los FQDN de Hadoop

Se debe generar una llave privada SSH y añadirla a todos los servidores para implementar un clúster de Hadoop desde **Ambari**, sin tener que instalar manualmente los agentes en todos los servidores que deseas en el clúster Hadoop.

Según los datos por defecto en la ARM Template, el nombre del Ambari Server (para este paso) será **rei-ambarisrv-iy.westeurope.cloudapp.azure.com**, mientras que el usuario de Linux será: **linuxadmin**.

1. Conéctate al Ambari server vía SSH, usando el DNS name asociado a la IP pública: <AMBARI_SERVER_NAME>.<LOCATION>.cloudapp.azure.com
2. Una vez que hayas entrado, cambia al usuario root.
3. Después de esto, podrás ver todos los FQDN Servers Hadoop instalados en la lista "/etc/hosts".
4. Posteriormente, corre el siguiente comando para obtener la llave privada que se generó durante la instalación de Ambari: cat /root/.ssh/id_rsa

Explicación

Por su parte, Apache Ambari habilita a los administradores del sistema para administrar y monitorear el clúster de Hadoop, además de integrarlo con la infraestructura existente en la empresa.

Por lo tanto, al usarlo podrás controlar el despliegue y la administración, así como la cancelación o suspensión de servicios relacionados con Hadoop Cluster, tales como:

- ✓ HDFS
- ✓ YARN + MapReduce2
- ✓ Tez
- ✓ Hive
- ✓ HBase
- ✓ Pig
- ✓ Sqoop
- ✓ Oozie
- ✓ ZooKeeper
- ✓ Falcon
- ✓ Storm

Se creará una URL de Ambari Web con la siguiente estructura:
<ambarie_server_name>.<location>.<server>:<port>

Una vez que la instalación de Ambari esté completa, usa el siguiente comando para acceder a la interfaz de usuario:
http://<AMBARI_SERVER_NAME>,<LOCATION>.cloudapp.azure.com:8080.

Recuerda que podrás acceder con las credenciales de usuario: admin, con contraseña: admin.

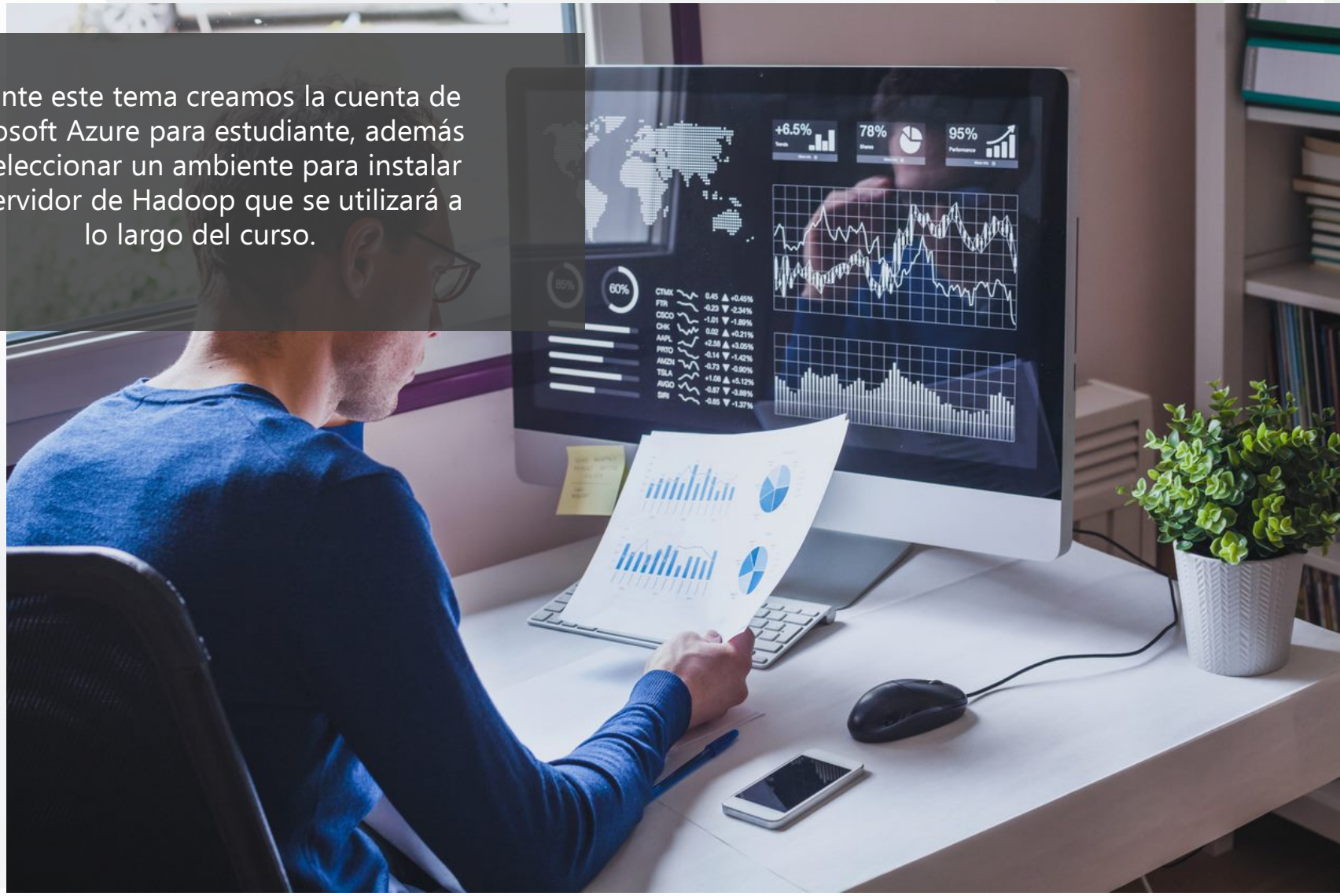
Explicación

Haz clic en el botón **Launch Install Wizard** para configurar un clúster, según White (2015):

1. Escribe un nombre con el que quieras reconocer a tu clúster.
2. Asegúrate de que la versión HDP 2.5.3.0 esté seleccionada y haz clic en el botón Siguiente para guardar y continuar.
3. Copia los FQDN de tus servidores de Hadoop en este espacio. De igual forma, copia la llave privada que obtuviste pasos arriba y haz clic en el botón de Registrar y Confirmar, respectivamente.
4. Después de completar ese proceso, aparecerán dos advertencias (*warnings*), las cuales se pueden ignorar. Ahora, haz clic en el botón Siguiente.
5. En la sección de selección de servicios tienes la opción de añadir o remover cualquiera de los servicios que desees instalar en el clúster. Por lo que, si deseas algún servicio que tenga una dependencia a otro servicio que no esté instalado, lo podrás hacer vía línea de comandos.
6. En la sección de asignar masters podrás asignar componentes de master a cualquier servidor que desees. En este caso, deja la configuración por defecto (*default*) y haz clic en el botón Siguiente.
7. En la siguiente sección podrás asignar los esclavos (*slaves*), por tanto, deja los valores por defecto y haz clic en Siguiente.
8. En la siguiente sección puede ser que veas un número en rojo al lado de los servicios, lo cual es un indicador de que dicho servicio requiere atención. Para este caso de instalación, es necesario que introduzcas el *password*. Posteriormente, haz clic en el botón Siguiente.
9. Revisa la configuración y haz clic en Deploy.
10. Una vez que la instalación esté completa (para el caso de este proceso), ignora las advertencias y haz clic en Siguiente.
11. Revisa el resumen de la instalación y haz clic en Completar.
12. **Al final, el estudiante será redireccionado al *dashboard* principal.**

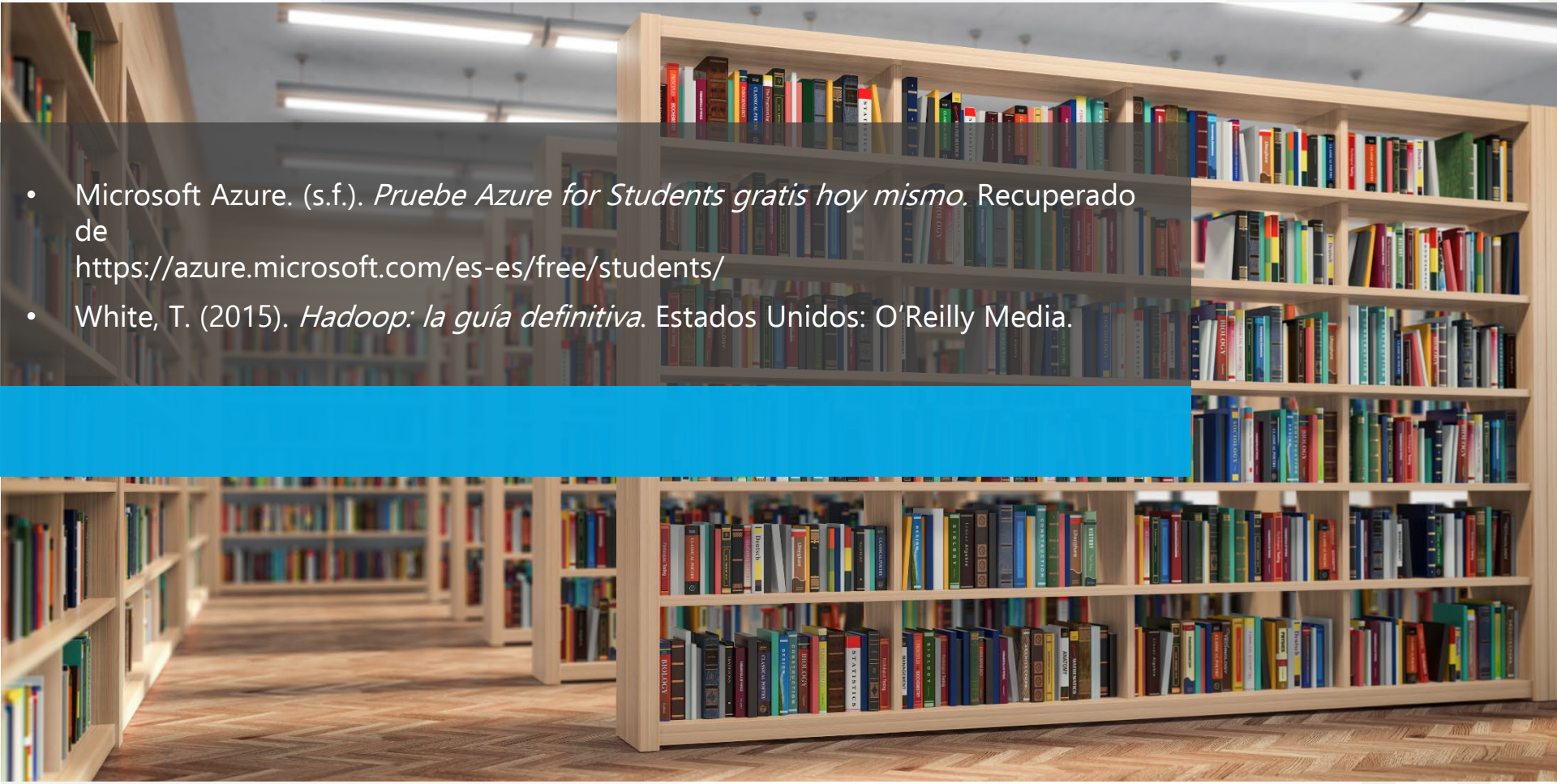
Cierre

Durante este tema creamos la cuenta de Microsoft Azure para estudiante, además de seleccionar un ambiente para instalar un servidor de Hadoop que se utilizará a lo largo del curso.



Bibliografía

- Microsoft Azure. (s.f.). *Pruebe Azure for Students gratis hoy mismo.* Recuperado de https://azure.microsoft.com/es-es/free/students/
- White, T. (2015). *Hadoop: la guía definitiva*. Estados Unidos: O'Reilly Media.



Universidad Tecmilenio

Infraestructura para Big Data

Comandos en HDFS (Hadoop Distributed File System)

Semana 10

Introducción

Aprenderás cuáles son los comandos más usados en HDFS y los comandos base proporcionados por Hadoop para el manejo de archivos y autenticación.

Explicación

Los **diez comandos más utilizados** en el sistema de archivos distribuidos Hadoop son los siguientes:

mkdir	copyFromLocal
ls	cat
get	mv
lcopyToLocals	cp
put	rm

Puedes encontrar estos comandos (y el resto de ellos), junto con la sintaxis para usarlos, en el siguiente enlace:

Apache Hadoop. (2021). *Overview.* Recuperado de https://hadoop.apache.org/docs/stable/api/overview-summary.html

Para la ejecución de todos los comandos se escribe primero **hadoop fs** y luego el <comando> (excepto si se quiere obtener la versión; en ese caso solo se escribirá hadoop)
(Apache Hadoop, 2021).

Para conocer más sobre Hadoop Distributed File System (HDSF), ve el siguiente video:

Oracle Learning. (2019, 25 de octubre). *Introduction to Hadoop Distributed File System (HDFS)* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=jFsYao4C3cw&ab_channel=OracleLearning

Explicación

`mkdir`	Es para crear uno o varios directorios en la ruta especificada.

mkdir - $ hadoop fs -mkdir [-p] <paths>

`ls`	Para obtener la lista de archivos y directorios que se encuentran en la ruta especificada.

ls - $ hadoop fs -ls <path>

`get`	Descarga el archivo desde el sistema de archivos de HDFS al sistema de archivos local.

get - $ hadoop fs -get <archivo_destino> <path_local>

`copyToLocal`	Es igual al comando get revisado anteriormente.

copyToLocal - $ hadoop fs -copyToLocal <archivo_destino> <path_local>

`put`	Para escribir o subir uno o varios archivos desde el sistema de archivos local al sistema de archivos de HDFS en el directorio especificado.

put - $ hadoop fs -put <archivo_local_1> <archivo_local_2> ... <archivo_local_n><path_destino>

`copyFromLocal`	Parecido al comando put, pero este es solo para un archivo.

copyFromLocal - $ hadoop fs -copyFromLocal <path_1> <path_2> ... <path_n> <path_destino>

`cat`	Muestra el contenido de un archivo.

cat - $ hadoop fs -cat <archivo>

`mv`	Mueve el archivo especificado.

mv - $ hadoop fs -mv [-f] <archivo_origen> <archivo_destino>

`cp`	Copia el archivo especificado.

cp - $ hadoop fs -cp [-f] <archivo_1> <archivo_2> ... <archive_n> <path_destino>

`rm`	Borra o remueve el archivo especificado.

Explicación

SSH (Secure Shell)

Es un protocolo de comunicación seguro para conectarse a escritorios remotos (SSH.COM, s.f.).

Para ejecutar los comandos se debe entrar a la máquina virtual donde se instaló Hadoop y abrir la consola.

Para utilizar SSH en Windows es necesario contar con una aplicación que se llama Cliente OpenSSH o una aplicación cliente que se llama Putty. Para hacerlo desde Linux o macOS solo es cuestión de ejecutar el comando ssh en cualquier terminal.

Pasos para instalar o revisar si está instalado el Cliente OpenSSH en Windows 10:

1. Entra a la sección de configuraciones y selecciona la opción de aplicaciones.
2. Haz clic en el texto que dice: Características opcionales.
3. Revisa si ya tienes instalada una característica opcional que se llama Cliente de OpenSSH. En caso de que no la tengas, búscala, haz clic en el botón de Instalar y después reinicia el equipo.
4. Una vez que puedas conectarte al clúster con el uso de SSH, vamos a obtener la liga para conectar. Para eso, ya que estás en la pantalla con el resumen de tu clúster, del lado izquierdo hay una opción que se llama SSH + Cluster login.
5. En la pantalla de SSH + Cluster login aparece una liga que hay que copiar para poder comunicarnos con el usuario y la contraseña que definiste en la configuración del clúster.

Explicación

Universidad Tecmilenio

Otra forma de conectarse para ver y administrar los directorios y los archivos del sistema de archivos distribuidos de Hadoop del clúster de HDInsight es a través de una aplicación que nos provee Microsoft Azure, que se llama **Storage Explorer**, la cual puedes descargar en el siguiente enlace:

Microsoft Azure. (s.f.). Azure Storage Explorer. Recuperado de https://azure.microsoft.com/en-us/features/storage-explorer/

- ✓ Una vez ingresado al sitio, selecciona el sistema operativo que usas, descarga e instala la aplicación en tu computadora.
- ✓ Ya instalada la aplicación, haz clic en la configuración de cuentas del lado izquierdo y agrega tu cuenta de estudiante de Tecmilenio.
- ✓ En la sección principal debajo de tu cuenta te aparecerán los recursos de almacenamiento que tengas activos en Microsoft Azure, ahí podrás ver la cuenta de almacenamiento (Storage Account) que creaste con tu clúster, después haz clic para ver su contenido.

Cierre

Ahora conoces los comandos más utilizados para la administración de archivos en el sistema de archivos distribuidos de Hadoop (HDSF), que es el componente principal del ecosistema Hadoop.

Asimismo, analizaste la explicación y la visualización de los comandos más socorridos para la administración de archivos en HDSF, mediante el uso del protocolo SSH para conectarte al clúster, el cual nos da la facilidad de usarlo en cualquier equipo Linux que cuente con la instalación de Hadoop.

Bibliografía

- Apache Hadoop. (2021). *Overview*. Recuperado de https://hadoop.apache.org/docs/stable/api/overview-summary.html
- SSH.COM. (s.f.). *SSH Protocol – Secure Remote Login and File Transfer*. Recuperado de https://www.ssh.com/ssh/protocol/