

Aplicación práctica de tecnologías Big Data.

Caso práctico

La empresa Telco Max Spain es una operadora virtual de telecomunicaciones con más de 5 años de presencia en el mercado español. Al ser una operadora virtual, no tiene infraestructura de telecomunicaciones propia, sino que usa las redes de los operadores principales, pero cuenta con una cartera de más de 1 millón de clientes, tanto particulares como empresas.

Su negocio lo tiene dividido en 3 áreas: particulares, soho (pequeña y mediana empresa) y grandes cuentas.

En cuanto a sus sistemas, destacan:



[Stux](#) (Dominio público)

- ✓ Un CRM de particulares, donde almacena todos los contactos con clientes particulares, así como los contratos, servicios contratados, etc.
- ✓ Un CRM de clientes soho, donde almacena el mismo tipo de información que en el caso del CRM de particulares.
- ✓ Un CRM de grandes cuentas, que contiene información similar, aunque con un modelo de datos más complejo, ya que las grandes cuentas tienen jerarquías de cuentas y responsables de cada cuenta.
- ✓ Un sistema de facturación y contabilidad que cada mes toma los datos de los distintos CRM y genera las facturas y cargos asociados.
- ✓ Una plataforma de marketing en la que se guardan las campañas, los contactos con leads (no clientes con los que se tienen contactos comerciales) y todo el camino de los clientes desde que entran por

cualquier canal de contacto hasta que son clientes. A partir de ese punto, es el sistema CRM el que pasa a tener el control de las interacciones y los datos del cliente.

- ✓ Una plataforma de contact-center, que da soporte a las operaciones del equipo de televenta, y que contiene la información de los contactos que se tiene con los no clientes por este canal, y al igual que en el caso de la plataforma de marketing, cede la gestión del contacto al CRM una vez ese convierte en cliente.
- ✓ Un sistema ERP con las operaciones internas: nóminas, vacaciones, contrataciones, presupuestos, etc.
- ✓ Cada canal (la web, la aplicación móvil o las oficinas) tiene sus propias aplicaciones, que contienen información específica de ese canal (accesos a la web, operaciones realizadas, etc.) así como algo de información clientes solapada con el resto de sistemas.
- ✓ Además, recibe información de las operadoras y otros agentes externos a diferentes servidores SFTP, por ejemplo, con el tráfico generado en las redes para abonar el alquiler de las infraestructuras, y otros datos externos como informes de mercado, etc.
- ✓ Tiene un sistema informacional con diferentes procesos de ETL que toman la información de los sistemas CRM y de facturación, y que vuelcan los datos ya transformados y agregados a un Datawarehouse, con el que se construyen los informes o reportes de negocio, etc.

Pese a que el nivel de madurez de los diferentes sistemas es correcta y que tiene un buen sistema de reporting, saben que hay mucha información que no están utilizando para hacer los análisis, como puede ser la información de todos los sistemas que no son el CRM o el sistema de facturación, y saben que si tuvieran toda la información de un cliente, desde que se tiene el primer contacto con él, pasando por todas las iteraciones que se tiene con él como llamadas al contact center, el uso que hace de los servicios en la aplicación web, etc. podrían mejorar mucho la toma de decisiones, especialmente para mejorar una métrica que es fundamental en su negocio: la tasa de rotación o de baja de clientes.

Por ello, el Comité de Tecnología ha decidido invertir en una plataforma Big Data que les permita almacenar toda la información de los diferentes sistemas, y además, crear nuevos casos de uso como modelos predictivos de abandono de clientes, etc.

El equipo de Tecnología ya conoce Hadoop y la ha elegido como la tecnología que dará soporte a este nuevo sistema, pero deben en primer lugar definir la arquitectura que montarán, si harán un despliegue en cloud o en su propia infraestructura (también llamado on-premise), y si montarán un modelo de arquitectura como Data Lake, u otros modelos que están apareciendo últimamente, como Data Mesh.

El caso práctico que se plantea es el más habitual por el que las empresas deciden implantar una plataforma Big Data para poder mejorar la analítica y utilizar, en lugar de un subconjunto de los datos, todos los datos disponibles. Por ejemplo, el uso de modelos predictivos no es algo nuevo que haya surgido con las tecnologías Big Data. Durante

años, las empresas han estado haciendo modelos utilizando tecnologías como R o con herramientas de Data Mining sobre los datos que tenían en los sistemas informacionales (Data Warehouse o Datamarts). La principal diferencia con las tecnologías Big Data es que tradicionalmente, los sistemas informacionales no tenían toda la información porque o bien era demasiado voluminosa para poder almacenarla en estos sistemas (piensa, por ejemplo, en la transcripción de todas las llamadas al contact-center de un cliente, o el log de la navegación del cliente en todos los accesos que haya hecho en la web), o porque almacenarla para poder analizarla era demasiado caro. Con las tecnologías Big Data, las empresas pueden crear una plataforma que contenga toda la información de la empresa, y de esta manera, poder mejorar sus modelos.

Citas Para Pensar

¿Qué modelo predictivo de predicción de abandono funcionará mejor, uno que coja la información del CRM, que contiene los servicios contratados, la información del cliente como su nombre o dirección, y el estado de pago de las facturas, o un modelo que coja, además de toda esa información, las llamadas que ha hecho el cliente al contact-center (por ejemplo para poner una queja), la navegación que ha hecho en la web (por ejemplo, ha estado consultando la página de preguntas frecuentes, y concretamente, ha leído la pregunta de "¿cómo darse de baja?"?

Sí, la pregunta tiene fácil respuesta ;-)

Cuando una empresa decide mejorar su práctica analítica u otros casos de uso implantando una tecnología Big Data, debe analizar, además de las tecnologías existentes para poder seleccionar las más adecuadas, qué modelo de arquitectura utilizar, ya que este modelo definirá aspectos tan importantes como:

- ✔ Cómo será la integración de los datos.
- ✔ Cómo será la integración de los sistemas.
- ✔ Cómo se gestionarán cuestiones como la seguridad, el cumplimiento normativo, etc.
- ✔ Qué cambios en la organización a nivel de equipos o de procedimientos habrá que realizar.

En esta unidad vamos a esta parte menos tecnológica de la implantación de una plataforma como Hadoop en una organización:

- ✔ En primer lugar, conoceremos qué modelos de arquitectura son los más habituales y cómo está siendo la adopción de Hadoop o las tecnologías Big Data en general en las empresas.
- ✔ A continuación, estudiaremos las soluciones cloud como alternativa a las soluciones Hadoop on-premise (en la propia infraestructura).



[Ministerio de Educación y Formación Profesional](#) (Dominio público)

Materiales formativos de FP Online propiedad del Ministerio de Educación y Formación Profesional.

[Aviso Legal](#)

1.- Arquitecturas y modelos de despliegue.

Caso práctico

En su proceso de adopción de Hadoop en la empresa Telco Max Spain, el primer trabajo que debe hacer el equipo de arquitectura es definir cómo se va a implantar en cuanto a los datos que se almacenarán, quién será el responsable de los datos, cómo y quién los transformará, cómo se integrará con los sistemas que ya existen, etc.



[Gerd Altmann](#) (Dominio público)

Implantar una tecnología no sólo consiste en instalar, configurar y utilizar. Requiere un paso previo de definición.

Vamos a ver cómo será la adopción de Hadoop en Telco Max Spain.

El concepto de Big Data irrumpió con fuerza en el mercado en el año 2013 a nivel mundial, y un poco más tarde en España, en torno a un año más tarde. Hasta entonces, las empresas habían confiado la analítica a los sistemas Datawarehouse y Datamart, cargados mediante procesos de ETL.

Datawarehouse, Datamart y procesos ETL: el método tradicional

Como sabes, un sistema Datawarehouse es un repositorio de datos (más unas herramientas) que contiene información de la empresa, recogida de sistemas operacionales, y preparada para la toma de decisiones. El almacenamiento de los datos suele diseñarse para resolver consultas analíticas, como puede ser "dame, para cada código postal y mes del año, el número de clientes activos, el número de clientes morosos, el total de facturación así como el incremento o decremento frente al mismo mes del año anterior". Esta consulta, que puede ser habitual en un proceso de toma de decisiones, es muy compleja utilizando sistemas que están preparados para operar con datos atómicos, por ejemplo, para modificar los datos de un cliente concreto, añadir una

factura, etc. Además, en la mayoría de las ocasiones, requiere haber hecho unos cálculos previos para no tener que ejecutar cada consulta a demanda.

Un sistema Datawarehouse, por ejemplo, almacena los datos en un formato columnar en lugar de tabular, para poder resolver mejor las agregaciones o consultas por un campo, o dispone de un modelo físico de almacenamiento que no está basado en múltiples tablas relacionadas, sino con un modelo de estrella habitualmente, para facilitar las consultas sobre diferentes dimensiones de los datos.

Para que los datos de los sistemas operacionales, que como se ha dicho, están preparados para operaciones atómicas de alta, baja, modificación o consulta de registros aislados, se puedan adaptar a las estructuras de los sistemas Datawarehouse, existen unos procesos denominados ETL que, normalmente en ventana nocturna, cogen los datos del día de los sistemas operacionales (CRM, ERP, etc.), los transforman y los cargan en los sistemas Datawarehouse. Además, no sólo transforman los datos, sino que les aplican mecanismos de validación y limpieza, ya que es habitual que los datos de los sistemas operacionales no tengan un nivel de calidad adecuado para hacer analítica, y normalmente tienen muchos campos en blanco, valores incorrectos, números con mal formato, direcciones incorrectas, etc.

Los Datawarehouses tienen dos principales limitaciones:

- ✔ Son tecnologías caras, y el coste crece con el volumen de datos.
- ✔ Sólo permiten gestionar datos estructurados, es decir, provenientes de las bases de datos de los sistemas operacionales.

Además, dado que el objetivo de los Datawarehouse es almacenar y analizar todos los datos, lo habitual es que en un Datawarehouse se disponga de información de diferentes ámbitos: comercial, marketing, recursos humanos, etc. A menudo, para que cada departamento acceda sólo a la información específica que es de su interés, se realizan subconjuntos del Datawarehouse, denominados Datamarts, que son sistemas con su propio repositorios y datos.

Por último, sobre los repositorios de los Datawarehouses y Datamarts, la explotación de los datos se suele hacer de dos formas:

- ✔ La creación de informes, cuadros de mando o herramientas de análisis ad-hoc se realiza mediante herramientas de **Business Intelligence**, que son herramientas que requieren un primer paso de diseño del informe o del análisis, para que posteriormente un equipo de desarrollo haga la preparación de datos y la creación del informe.
- ✔ La creación de modelos predictivos o analítica más compleja se realiza mediante herramientas de **Data Mining**, que suelen disponer de capacidades para programar modelos a medida utilizando técnicas como machine learning.

La migración hacia arquitecturas Big Data

El modelo tradicional basado en herramientas de ETL, Datawarehouses y herramientas de Business Intelligence y Data Mining era o mejor dicho, es, un buen modelo para realizar análisis de los datos para la toma de decisiones, pero tiene varios problemas asociados:

- ✔ Sólo permite analizar datos estructurados, y cada vez hay un mayor número de

fuentes de datos no estructuradas que se quieren analizar: logs de aplicaciones, transcripciones de conversaciones, imágenes, vídeos, etc.

- ✔ Requiere mucha intervención de los equipos de tecnología o desarrollo, desde la construcción de los procesos ETL hasta la creación de los cuadros de mando o informes. Esto hace que desde que el negocio tiene una necesidad hasta que dispone de la herramienta para cubrir esa necesidad, el proceso puede durar demasiado tiempo.
- ✔ El coste de estos sistemas suele ser elevado, lo que conlleva, entre otras cosas, que no se tomen todos los datos para hacer los análisis, y por ejemplo, se descarten datos con una antigüedad de más de 2 años, ya que analizar toda la profundidad histórica tiene un coste inasumible.

La irrupción de tecnologías Big Data abrió un nuevo panorama a las empresas:

- ✔ Podían poner en valor el creciente número de fuentes de datos no estructurados.
- ✔ Podían utilizar todos los datos para poder hacer análisis.
- ✔ Al no requerir un esquema antes de su almacenamiento, o disponer de repositorios con esquema flexible, y unido con la aparición de nuevas herramientas de visualización, se puede reducir la intervención de los equipos de tecnología, dando más autonomía al negocio para construir casos de uso, y por lo tanto ganar eficiencia y reducir el tiempo de implantación de una necesidad o caso de uso.
- ✔ Reducir el coste de los sistemas orientados a la analítica o a la toma de decisiones.
- ✔ Permitían utilizar nuevas técnicas de analítica que requerían una mayor potencia de cálculo, como es el caso de las redes neuronales, mejorando, por lo tanto, los casos de uso que se podrían abordar, por ejemplo, motores de recomendación, modelos de simulación más precisos, etc.

Dentro de todo el espectro de tecnologías Big Data, como sabes, Hadoop ha sido la plataforma más utilizada.

A continuación vamos a ver cómo se adopta Hadoop y qué modelos de arquitectura se suele utilizar.

Autoevaluación

Indica si las siguientes afirmaciones son verdaderas o falsas.

Los sistemas Datawarehouse se reemplazan por sistemas Big Data

Verdadero Falso

Falso

Falso: aunque los sistemas Big Data mejoran algunas características de los Datawarehouses, estos últimos son sistemas que aportan valor en las empresas, por lo que su sustitución, en caso de realizarse, no se llevará a cabo a corto plazo.

El modelo tradicional basado en herramientas de ETL, Datawarehouses y herramientas de Business Intelligence y Data Mining gestiona y analiza datos estructurados

Verdadero Falso

Verdadero

Verdadero: estos sistemas no están preparados para analizar o gestionar datos no estructurados.

Las tecnologías Big Data suelen ser más baratas que las tecnologías clásicas de análisis o gestión de datos

Verdadero Falso

Verdadero

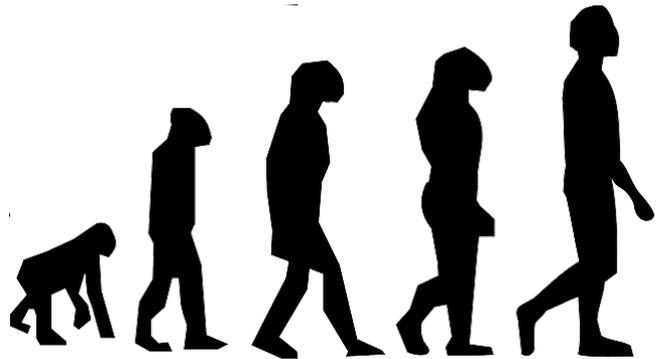
Verdadero: uno de los atractivos de las tecnologías Big Data es que su coste es inferior para las empresas.

1.1.- Adopción de tecnologías Big Data.

Como se ha indicado en el punto anterior, a partir del año 2014, las empresas españolas, especialmente las empresas de mayor tamaño, comenzaron a interesarse por las tecnologías Big Data, por sus beneficios y por sus dificultades de implantación.

En el caso de España, las primeras empresas que comenzaron a probar estas tecnologías fueron las del sector financiero, ya que son compañías con un nivel de madurez tecnológico muy

elevado, que tradicionalmente habían gestionado un gran volumen de datos (movimientos de cuentas, cotizaciones, etc.) pero a un coste muy elevado utilizando sistemas mainframe, y porque tienen un modelo de negocio muy apalancado en la tecnología y en la analítica, y para los que los casos de uso pueden tener un retorno de la inversión rápido.



[Ciker-Free-Vector-Images](#) (Dominio público)

En la **primera etapa, entre 2014 y 2017**, la mayor parte de las grandes empresas españolas, especialmente las del sector financiero, hicieron muchas pruebas de concepto sobre las tecnologías Big Data, principalmente utilizando plataformas Hadoop con distribución comercial de Cloudera. Estas pruebas de concepto permitieron a las empresas empezar a entender las tecnologías Big Data, tanto en sus beneficios como en sus dificultades o los cambios que requieren en las organizaciones. La realidad, en cualquier caso, es que pocas empresas monetizaron o tuvieron un retorno de la inversión inmediato, y vieron esta etapa más como una toma de contacto que como una necesidad de rentabilizar estas nuevas tecnologías.

El modelo arquitectónico que se utilizó en la primera fase, en la mayoría de los casos, fue el del **Data Lake**, que veremos más adelante, y que consiste en una plataforma que aglutina toda la información disponible de los sistemas, almacenada tal cual, sin transformaciones, y con procesos que permiten refinar la información y poner a analizarla con diferentes técnicas, e incluso los resultados finales servían para alimentar a los Datawarehouses existentes. Hadoop, en la mayor parte de los casos, sería como complemento al Datawarehouse.

Esta primera fase, y con la aproximación de Data Lake, permitió descubrir a las empresas la necesidad de poner mayor énfasis en el **Gobierno de Datos**, que si bien tradicionalmente se había implementado en mayor o menor medida en todas las compañías, el nuevo modelo, que suponía un movimiento y consumo de datos mucho más elevado, hacía más necesario gestionar o gobernar los datos para no caer en una situación descontrol y desuso de los datos existentes en las plataformas Big Data.

Asimismo, se puso énfasis en la **industrialización** de las tareas de los proyectos Big Data, es decir, en definir un modelo común de actividades en la construcción de proyectos de Big Data, de manera que las actividades comunes pudieran automatizarse u optimizarse para acelerar la creación de casos de uso y reducir la entropía entre proyectos.

Desde 2017 aproximadamente, surge una **segunda etapa** en las organizaciones en la que, una vez que la mayoría ya tenían una plataforma de datos y habían aprendido qué cambios requerían y qué podían ofrecer, se pone más interés en la creación de casos de uso y en generar valor para el negocio. A diferencia de la primera fase, donde el impulsor habían sido los equipos de tecnología, en esta segunda fase son los equipos de negocio los que impulsan la evolución de las arquitecturas de datos para poder mejorar los casos de uso de analítica que quieren abordar.

Esta segunda fase, de puesta en valor de las plataformas Big Data, está marcada por varios procesos dentro de las empresas:

- ✓ El uso de las áreas de negocio de los datos de las plataformas y su liderazgo en la construcción de proyectos o casos de uso.
- ✓ La creación de áreas de Gobierno de datos y la figura del Chief Data Officer (CDO), como garante de la calidad de los datos de la plataforma, su control para evitar redundancias o para democratizar el uso de los datos, o para el cumplimiento normativo en materia de datos, como el Reglamento General de Protección de datos (GDPR).
- ✓ A nivel de tecnologías, se pone foco en incorporar nuevas herramientas que faciliten el acceso a los datos para el negocio, así como aquellas orientadas a eficientar o industrializar la construcción de proyectos. Asimismo, se crean equipos de ingeniería de datos para satisfacer las demandas de datos por parte de los diferentes proyectos. Estos equipos de ingeniería tienen como principal cometido recoger la información de los diferentes sistemas de la empresa, procesarlo y prepararlos para su explotación por parte de los proyectos.

Por último, **desde el año 2019**, se está viviendo una **tercera etapa** en la que, si bien a nivel de negocio ya se tiene un nivel de madurez suficiente, se está poniendo más intensidad en la parte tecnológica en poder adaptar las plataformas a las necesidades cambiantes del día a día, principalmente, explorando el uso de soluciones cloud e incorporándolo a sus arquitecturas. La situación actual a nivel de plataformas de datos es, para las grandes empresas, que existen plataformas Big Data en infraestructura propia que conviven con soluciones Big Data en cloud, en lo que se denominan plataformas híbridas.

Cabe decir que cada empresa ha seguido unos pasos y tiempos propios, pero en general, las principales empresas han seguido lo indicado anteriormente.

Autoevaluación

Indica si las siguientes afirmaciones son verdaderas o falsas.

Al principio, el equipo que lideró la adopción de las tecnologías Big Data fue la dirección de la empresa

Verdadero Falso

Falso

Falso: en la primera fase fueron los equipos de tecnología los que principalmente impulsaron la adopción de las tecnologías Big Data.

En la segunda etapa, una vez que ya se había adoptado las tecnologías Big Data, se puso más foco en la creación de valor mediante casos de uso.

Verdadero Falso

Verdadero

Verdadero: la segunda etapa fue liderada por los equipos de negocio, que buscaban implementar nuevos casos de uso que ayudaran a conseguir mayor rentabilidad.

En la última etapa, en la que estamos a día de hoy, hay una migración muy fuerte a plataformas cloud.

Verdadero Falso

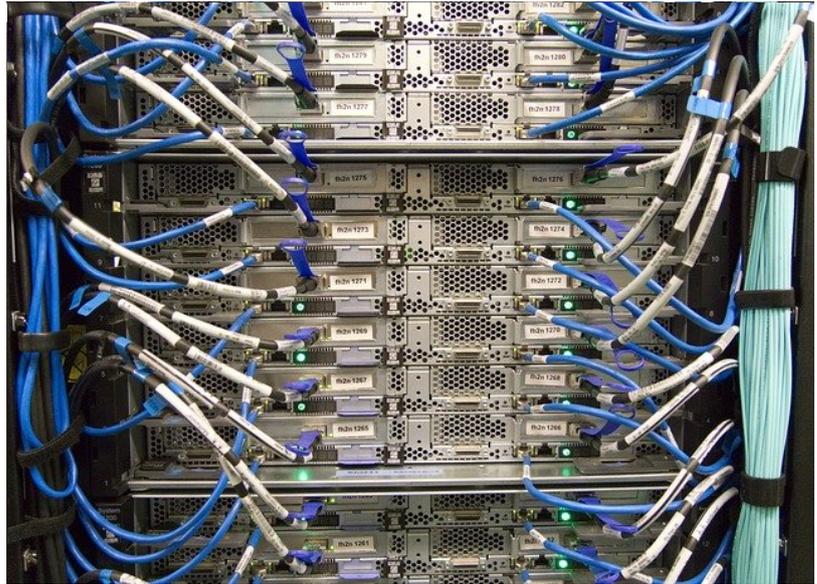
Verdadero

Verdadero: la migración a cloud es uno de los principales vectores de movimiento a nivel de tecnologías de datos hoy en día.

1.2.- Hadoop en la práctica.

Pasos en la adopción o incorporación

Una vez se ha decidido que Hadoop es la tecnología a utilizar como tecnología Big Data y para dar cobertura a los casos de uso que los sistemas tradicionales no pueden cubrir, se requiere una serie de pasos hasta su uso:



[dlohner](#) (Dominio público)

1.- **Identificación de las necesidades** de almacenamiento y procesamiento de datos para poder dimensionar correctamente la plataforma. Este punto es especialmente relevante en el caso de los despliegues de Hadoop en infraestructura propia. Para abordarlo, el método más habitual es hacer un catálogo de los casos de uso a abordar junto con sus necesidades en materia de datos, y hacer un inventario de los datos existentes en la organización, identificando cuáles son los candidatos a incorporar en la plataforma. A continuación, se determinará cuál es el volumen de datos que la plataforma debe gestionar.

2.- **Plan de capacidad y selección de hardware:** una vez se tiene la necesidad de almacenamiento, se realiza la selección del hardware. Para ello, tomando un estándar de 12 discos de 2 terabytes por nodo worker, se ajustará en función de si el clúster se utilizará mayoritariamente para almacenamiento o si habrá mayoritariamente procesamiento. Es decir, se trata de hacer la relación entre el almacenamiento y la capacidad de procesamiento para determinar qué densidad de almacenamiento tendrá cada nodo. Por ejemplo, en sistemas donde va a prevalecer el procesamiento de datos en tiempo real, y que almacenará los datos una vez hayan sido procesados, la densidad será baja (por ejemplo, se usarán discos de 1 terabyte por nodo), y en sistemas donde el objetivo es almacenar un gran volumen de datos, y posteriormente se ofrecerán algunas herramientas de análisis sobre ellos, la densidad será alta (se utilizarán, por ejemplo, discos de 4 terabytes por nodo). Una vez determinada la capacidad de cada nodo, se hará la división entre la necesidad o capacidad de datos estimada entre la capacidad real de cada nodo, teniendo en cuenta que la capacidad real de un nodo será la suma de la capacidad de sus discos dividido por el número de réplicas de cada bloque y quitando entre un 30 y 40% para dejar un espacio para trabajo o datos intermedios. Por ejemplo, si la necesidad es de 800 terabytes, y se ha decidido tener una densidad alta, con 36 terabytes por nodo brutas y un factor de replicación de 3, la capacidad real por nodo será de $12 \text{ terabytes} - 35\% = 7,8 \text{ terabytes}$ por nodo, así que se necesitarán 100 nodos aproximadamente.

3.- **Compra del hardware:** el proceso de compra suele requerir unas fases de

oferta, negociación y contratación. El tiempo de disponibilidad de los servidores puede ser elevado, de varias semanas e incluso meses.

4.- **Instalación física:** una vez se dispone del hardware, es necesario realizar la instalación física, que incluye la disponibilización de los racks o del espacio en el centro de datos, la instalación de las máquinas en los racks, la instalación de la arquitectura de red, así como el cableado de alimentación y red, y por último la refrigeración. Este proceso puede llevar varias semanas.

5.- **Instalación del software base:** una vez el hardware ha sido instalado, se realiza la instalación y configuración de los sistemas operativos en las máquinas, normalmente, maquetando una de ellas y copiando esta maqueta en el resto. La instalación de Hadoop requiere algún ajuste en el sistema operativo, aunque es una operación sencilla. Después de la instalación del sistema operativo, se instalarán diferentes programas para monitorización o gestión remota de las máquinas.

6.- **Instalación de Hadoop:** después de terminar la instalación del software base, se instala la distribución comercial utilizando el software de administración (Cloudera Manager o Ambari), que guían en la instalación de todos los servicios. Este proceso puede durar bastantes horas.

7.- **Pruebas de la plataforma:** una vez Hadoop se encuentra instalado en el clúster, es necesario validar todos los servicios, por ejemplo, realizando pequeños programas o pruebas sobre todos ellos, y comprobando que se ejecutan sin problemas.

8.- **Tuning:** antes de liberar la plataforma para su uso en toda la organización, se suele implementar una fase de tuning en el que utilizando ejemplos más complejos, se optimizan diferentes variables de configuración, tanto de sistema operativo como de Hadoop, para aumentar en la medida de lo posible el rendimiento de la plataforma en la ejecución de tareas. Estos tests y sus resultados se custodian para poder repetirlos con cierta periodicidad, a fin de detectar deterioros en el rendimiento con el paso del tiempo.

9.- **Preparación del entorno de trabajo:** en esta fase se incluyen tareas de creación de los usuarios, establecimiento de permisos, creación de colas de ejecución de YARN, creación de la estructura de directorios de trabajo, etc.

Entorno de trabajo

Hadoop suele desplegarse en las organizaciones como plataforma, no sólo para resolver un caso de uso particular, sino como un sistema en el que poder volcar información de múltiples fuentes para habilitar diferentes herramientas que permiten resolver un conjunto de casos de uso variados.

Los usuarios de Hadoop, entendiéndolo como los actores que pueden hacer uso del clúster para cualquier propósito son:

- ✔ Usuarios, habitualmente analistas de negocio, que acceden para consultar datos de HDFS o realizar consultas mediante Hive u otra herramientas de alto nivel.
- ✔ Usuarios con perfil más técnico, habitualmente data scientists, que acceden a Hadoop para consultar datos de HDFS, transformarlos o procesarlos internamente, desarrollar modelos de inteligencia artificial con los datos, aplicar los modelos y publicar los datos como tabla o mediante una herramienta de visualización que accede a la tabla.
- ✔ Usuarios de ingeniería de datos que acceden a Hadoop para lanzar procesos de

ingesta o procesos de transformación de datos de forma manual.

- ✓ Procesos automáticos desarrollados por los equipos de ingeniería que realizan ingestas o procesamiento de datos de forma desasistida, programada y coordinada.
- ✓ Procesos automáticos lanzados desde herramientas externas, por ejemplo, para la construcción de cuadros de mando o informes mediante la explotación de datos vía Hive o Impala.
- ✓ Aplicaciones que utilizan servicios del clúster para fines operacionales, por ejemplo, aplicaciones web accediendo a HBase para recuperar la información de un cliente.

Este entorno, con múltiples usuarios diferentes se conoce como **entorno multitenancy**, y requiere un mayor control del uso y los límites de uso de cada usuario para evitar que uno pueda interferir en la operativa habitual de otro, por ejemplo, que si un data scientist está entrenando un modelo que consume muchos recursos del clúster, no bloquee la ejecución de aplicaciones que requieren un tiempo de respuesta bajo o que son críticas.

La forma en la que se consigue un entorno multitenancy en Hadoop es mediante la aplicación de políticas como las siguientes:

- ✓ **Espacio de nombres:** se suele dar una estructura de directorios a HDFS de tal forma que cada usuario pueda modificar sólo los directorios que pertenezcan a su grupo, pero que pueda leer los directorios que sean comunes, con el objetivo de que un usuario no pueda modificar o alterar datos que puedan afectar a otros procesos. Un espacio de nombres habitual tiene una estructura de directorios en la que por un lado se tienen los datos ingestados, por otro los directorios de trabajo de las aplicaciones y los procesos, otros directorios con los datos ya procesados y públicos para el resto de la organización, y por último, directorios para que cada usuario pueda tener su espacio de trabajo.
- ✓ **Seguridad:** se suelen definir sistemas de autenticación (identificar qué usuarios pueden acceder) y de autorización (detección de qué puede realizar cada usuario). Lo habitual es configurar Hadoop para integrarse con Kerberos y un directorio activo, a fin de tener un sistema de autenticación común de todos los usuarios.
- ✓ **Colas de ejecución:** como recordarás, YARN permite configurar diferentes colas de ejecución de tal forma que cada aplicación o grupo de usuarios ejecute sus tareas en su cola, y que cada cola pueda tener un límite de recursos disponible, así como un mínimo de disponibilidad garantizada para ejecutar sus tareas. Por ejemplo, las aplicaciones que tienen que dar respuesta a los procesos de negocio pueden tener definido un mínimo de capacidad del 30% de los recursos totales, para garantizar que nunca habrá una situación en la que no disponga de recursos para ejecutar sus tareas, o que los científicos de datos utilicen una cola que tenga configurado una asignación máxima del 25% de los recursos disponibles, para que en caso de lanzar procesos demasiado pesados, éstos no interfieran con el resto de aplicaciones.
- ✓ **Gobierno de datos:** para poder organizar el uso de los datos, facilitar su acceso a cualquier agente interesado (y con privilegios) y evitar que haya descontrol con los datos, se crean equipos de gobierno de datos que tienen un papel fundamental en las plataformas multitenancy. Estos equipos suelen trabajar en publicar un catálogo de datos para que cualquier usuario pueda conocer qué datos hay en el clúster y qué características y significado tienen, así como en la definición de políticas y procedimientos para el acceso y el uso de los datos, controlar la evolución de los datos, etc.
- ✓ **Equipos horizontales de arquitectura y soporte:** habitualmente se diseñan equipos que dan soporte técnico a la plataforma, desde un punto de vista de arquitectura (cómo abordar casos de uso, cómo evolucionar la arquitectura, etc.), de

Autoevaluación

¿Qué quiere decir la afirmación de que Hadoop es multitenancy?

- Que un clúster Hadoop tiene muchos servidores que trabajan en paralelo.

- Que Hadoop permite trabajar con muchos ficheros de cualquier tipo.

- Que Hadoop permite que múltiples usuarios de diferente tipo utilicen la plataforma.

Mostrar retroalimentación

Solución

1. Incorrecto
2. Incorrecto
3. Correcto

1.3.- Modelos de despliegue.

Data Lake

Las tecnologías Big Data, concretamente Hadoop, en su origen, no estaban enfocadas a resolver casos de uso generales o a servir como plataforma, sino para resolver casos de uso concretos, como la indexación de páginas web o el análisis de grandes volúmenes de información para temas muy específicos.

Sin embargo, motivado por la concepción de Hadoop como plataforma y su crecimiento en cuanto a las herramientas disponibles sobre los datos almacenados, y la capacidad de atender diferentes aplicaciones y usuarios en un concepto multitenancy, que vino con la aparición de YARN, Hadoop comenzó a visualizarse como una pieza clave para mejorar los diferentes procesos analíticos que hasta ahora se estaban confiando al Datawarehouse, que presentaba problemas de escalado y de elevado coste. De esta manera, se empezó a visualizar Hadoop como una plataforma en la que poder volcar todos los datos disponibles en la empresa para posteriormente ser analizados por diferentes tecnologías y equipos. Además, la característica diferencial de Hadoop frente al Datawarehouse, por la que no es necesario especificar el esquema o la estructura de los datos en el momento de la escritura, así como la capacidad de almacenar y procesar cualquier tipo de dato, hizo que se empezara a ver a Hadoop como una plataforma centralizada para toda la organización en la que volcar todos los datos, y que ofreciera herramientas para que diferentes equipos pudieran hacer análisis con todos los datos disponibles.

Nace entonces el concepto de Data Lake.

Un Data Lake es un repositorio centralizado que permite almacenar todos los datos de una empresa a cualquier escala, sin modificarlos y sin tener que estructurarlos primero, y también permite ejecutar diferentes tipos de análisis: desde cuadros de mando y reporting hasta análisis en tiempo real y machine learning.

El concepto de Data Lake como modelo arquitectónico ha tenido mucho auge y la gran mayoría de las empresas de cierto volumen han utilizado Hadoop para implementar este tipo de arquitectura, ya que Hadoop ofrece la capacidad de almacenar un gran volumen de datos de cualquier tipo, así como analizarlos utilizando diferentes herramientas y por parte de múltiples usuarios.

El Data Lake utiliza parte de la filosofía del Datawarehouse en cuanto a aglutinar información de múltiples fuentes y ofrecer herramientas de análisis, pero tiene las siguientes mejoras:

- ✔ **Capacidad de escalado a un coste razonable:** el uso de tecnologías Big Data como Hadoop hace que la implementación de un Data Lake sea más económica que la de un Datawarehouse (un orden de magnitud más económica), y además, tiene una capacidad de almacenamiento de datos muy elevada.
- ✔ **Tipo de datos:** un Datawarehouse sólo almacena datos estructurados, mientras que un Data Lake puede almacenar cualquier tipo de dato.
- ✔ **Flexibilidad en los datos y en los procesos:** en un Data Lake, el esquema de los

datos se fija en la lectura, no en la escritura, lo que permite almacenar datos sin importar qué estructura tienen, pero sabiendo que pueden contener valor, para posteriormente analizarlos, frente a un Datawarehouse, que requiere conocer bien la estructura de los datos antes de su escritura. Además, en un Data Lake, los cambios en el formato o esquema de los datos es más fácil de abordarse que en un Datawarehouse.

- ✓ **Herramientas:** un Datawarehouse ofrece en general herramientas de análisis para analistas de negocio, mientras que un Data Lake tiene un conjunto de herramientas más amplio, con capacidad para análisis tradicional con informes o cuadros de mando, así como análisis basado en machine learning, en tiempo real, etc.

Asimismo, el concepto de Data Lake, al igual que el concepto de Datawarehouse, pretende tener un repositorio centralizado con todos los datos de la compañía, evitando de esta manera la situación más habitual que consiste en que cada herramienta o área tiene sus propios datos y la compartición entre áreas o aplicaciones es difícil, además de tener varias versiones diferentes del mismo dato.

El objetivo de un Data Lake es construir una plataforma de datos que permita recopilar todos los datos disponibles, tanto externos como internos, estructurados y no estructurados, para gestionarlos de forma centralizada y mejorar los procesos de toma de decisiones.

Íñigo Sanz (Dominio público)

Modelo de capas de un Data Lake

Para entender el funcionamiento o las capas de un Data Lake, es necesario entender las distintas fases por las que pasan los datos en un Data Lake, desde su incorporación en la plataforma hasta su explotación:



Íñigo Sanz (Dominio público)

En primer lugar, los datos son **ingestados** desde múltiples fuentes sin aplicar ninguna transformación en el proceso. Las fuentes pueden ser sistemas en streaming o ser datos en reposo, y los datos pueden ser estructurados o no estructurados. Una característica

importante de los Data Lakes es que los datos se almacenan tal cual se reciben, es decir, tal cual están en el origen. Estos datos se denominan **Raw Data**. Este es un punto de diferencia importante frente a los Datawarehouse, ya que estos últimos no disponen del dato en origen, sino que sólo disponen del dato ya preparado.

Las ingestas son organizadas e implementadas por un equipo de ingeniería de datos que debe conocer cómo recoger la información de las fuentes, y cómo almacenar los datos en crudo para poder ser recuperados en caso necesario. Para este propósito, los datos se suelen particionar, por ejemplo, agrupando los datos de origen por directorios por mes o año.

En cuanto a las **transformaciones** sobre los datos, una vez que los datos han sido almacenados en el Data Lake, se lanzan procesos para validar y procesar los datos para que puedan ser explotables por diferentes aplicaciones con fines analíticos. De esta manera, se suelen realizar los siguientes procesos:

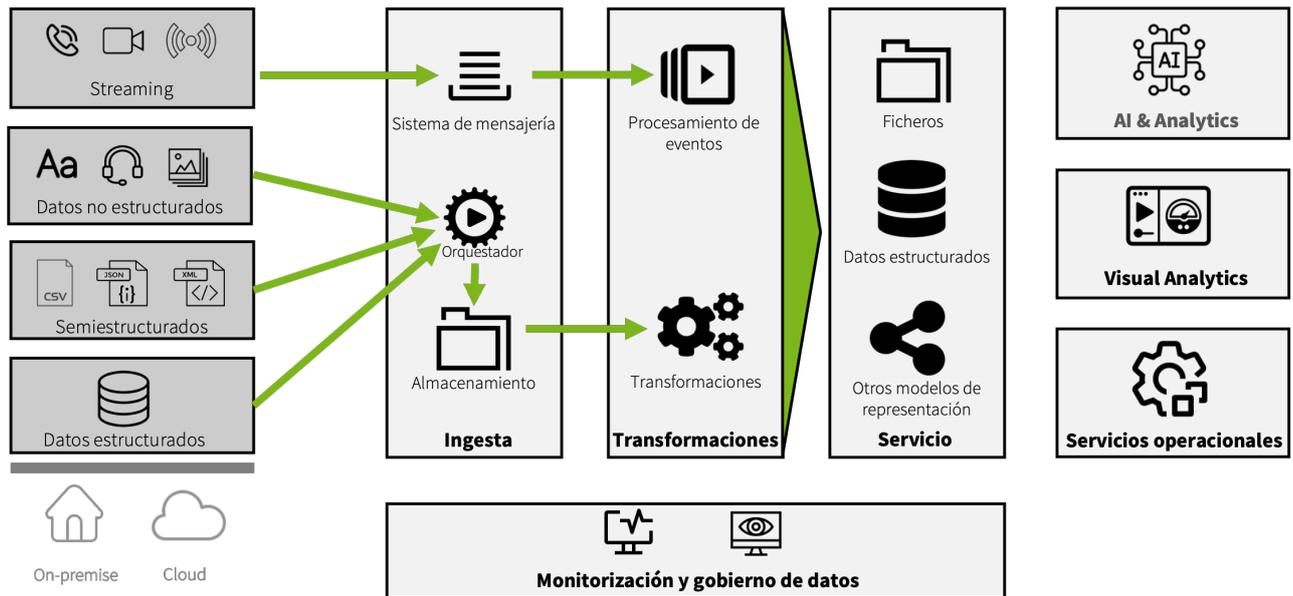
- ✔ **Validación:** se verifican los datos ingestados, tanto desde el punto de vista funcional (valores máximos, adecuación a las reglas de negocio, etc.) como desde el punto de vista técnico (nulos, vacíos, etc.).
- ✔ **Normalización:** se unifican los formatos y valores de los datos (es decir, números de teléfono, direcciones, nombres, etc.).
- ✔ **Limpieza:** los datos se transforman para eliminar espacios en blanco, caracteres no válidos, etc.
- ✔ **Enriquecimiento:** los datos se pueden utilizar para generar nuevos datos. Por ejemplo, una conversación transcrita para generar una métrica de sentimiento o uno más simple, como generar la edad a partir de la fecha de nacimiento.
- ✔ **Transformación:** los datos se pueden agregar para generar resúmenes o estadísticas.
- ✔ **Unificación y combinación:** los datos de varias fuentes se combinan por entidad, por ejemplo, los datos de un solo cliente se pueden combinar utilizando todos los datos de diferentes campañas o clientes.
- ✔ **Gestión de metadatos:** todas las transformaciones se etiquetan y registran para almacenar el linaje y el resto de metadatos y se almacenan y clasifican para facilitar el gobierno de datos.

Todos los procesos de transformación de los datos suelen ser liderados por un equipo de ingeniería de datos que, previo entendimiento del significado de los datos de origen y de las necesidades de análisis sobre los mismos, son capaces de ejecutar las transformaciones con herramientas que ofrece el Data Lake, como Apache Spark o Apache Hive.

Una vez los datos Raw han sido procesados, validados, limpiados, etc. son publicados en un **repositorio único** que pretende ser una fuente de la verdad de los datos y el lugar en el que todos los equipos consultan los datos, asegurándose de que todos ellos utilizan la misma información. Este repositorio suele tener un interfaz SQL pero también suele permitir el acceso a los datos en claro, directamente a los ficheros, para los casos en los que los data scientists desean hacer modelos sin una estructura de datos inicial. Asimismo, en un DataLake, los datos Raw también son accesibles para cualquier análisis, asumiendo que su nivel de calidad podría no ser óptimo.

Arquitectura de un Data Lake

La arquitectura física de un Data Lake suele ser la siguiente:



Íñigo Sanz (Dominio público)

La ingesta puede ser de dos tipos:

- ✓ Para la ingesta de datos en streaming, se utiliza un sistema de mensajería como Apache Kafka, para independizar la generación de los datos con su consumo, o un sistema como Apache Flume, para su volcado directo a HDFS.
- ✓ Los datos en reposo suelen ingestarse utilizando un orquestador que coordina y ejecuta planificadamente procesos de ingesta. El orquestador puede ser Apache Oozie, mientras que los procesos de ingesta pueden estar implementados con Sqoop (para ingesta de datos de bases de datos relacionales), o simplemente tareas ejecutadas como script que recogen ficheros de sistemas SFTP.

El procesamiento de los datos raw suele realizarse mediante herramientas de procesamiento masivo de datos como Apache Spark o Apache Hive. En el caso del procesamiento en tiempo real, se suele utilizar Apache Spark o Apache Flink para su implementación.

En cuanto al área de servicio, los datos ya refinados se suelen ofrecer tanto como fichero, utilizando HDFS como repositorio o un repositorio de objetos como Amazon S3, así como repositorios de acceso SQL como Hive o Impala, e incluso, en algunas ocasiones que los datos pueden representarse con otros modelos, alguna base de datos de grafos.

Autoevaluación

¿Cuál de las siguientes afirmaciones **no es correcta** en relación con los Data Lakes?

Hadoop es una buena plataforma para implementar un Data Lake.

- Los Data Lakes intentan ser un repositorio de datos único para toda la empresa.

- Frente a los Datawarehouses tradicionales, un Data Lake ofrece más funcionalidad.

- Un Data Lake es más fácil de gestionar que un Datawarehouse.

Mostrar retroalimentación

Solución

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Correcto

Data Mesh

Muchas organizaciones han invertido en Data Lake y un equipo de ingeniería de datos con la expectativa de impulsar su negocio apalancados en los datos.

Este equipo de ingeniería, que es el que se ocupa de las ingestas, las transformaciones de los datos y posteriormente en el ofrecimiento de los datos para que distintos grupo puedan utilizarlos, necesita conocer en detalle la naturaleza de los datos en origen, su significado, cómo se originan o cómo se deben transformar. La realidad es que sólo los equipos que trabajan en cada dominio u origen de los datos son los que conocen y entienden bien lo que son los datos, por lo que ese trabajo de transferencia de

conocimiento suele ser en ocasiones laborioso.

Además, el equipo de ingeniería de datos representa, en ocasiones, un cuello de botella para la implementación de los diferentes casos de uso ya que son ellos los que deben atender a la demanda del resto de áreas, alinear prioridades, gestionar dependencias o informar sobre cómo son y están los datos, y esta labor es difícilmente escalable.

El modelo de arquitectura del Data Lake es, realmente, un modelo centralizado, tanto en cuanto al equipo que gestiona los datos como a la plataforma o infraestructura que da soporte a toda la operativa.

Este modelo de arquitectura centralizada tiene ventajas:

- ✔ La especialización de los perfiles de ingeniería de datos, que permite tener una mayor productividad en las tareas de ingesta, transformación y ofrecimiento de los datos.
- ✔ La industrialización de los procesos de ingeniería de datos, que ofrece una mayor homogeneidad en las todas labores sobre los datos.
- ✔ La economía de escala, tanto en infraestructura como en equipos.

Sin embargo, tiene como principales problemas:

- ✔ La generación de dependencias sobre un único equipo de ingeniería de datos.
- ✔ La dificultad para escalar el modelo de operaciones sobre los datos.

Por estos dos principales motivos, surgió a principios de esta década el concepto de Data Mesh.

Una arquitectura Data Mesh es un enfoque descentralizado en el que cada dominio (gestión de clientes, ventas, logística, etc.) es responsable de preparar los datos y ofrecer al resto de equipos estos datos como producto, es decir, ofrecer al resto de equipos los datos para que puedan ser utilizados de la forma más sencilla posible.

Data Mesh es principalmente un enfoque organizacional. Sin embargo, la tecnología sigue siendo importante, ya que actúa como habilitador este tipo de arquitectura.

Data Mesh sigue cuatro principios básicos:

- ✔ **Propiedad y arquitectura descentralizada de datos orientada al dominio:** el principio de propiedad del dominio exige que los equipos de dominio asuman la responsabilidad de sus datos. Siguiendo la arquitectura distribuida impulsada por el dominio, la propiedad de los datos analíticos y operativos se traslada a los equipos de dominio, lejos del equipo de datos central. Esta aproximación permite escalar a medida que aumenta la cantidad de fuentes de datos, la cantidad de casos de uso y la diversidad de modelos de acceso a los datos, ya que simplemente es necesario aumentar los nodos autónomos en la malla.
- ✔ **Datos como producto:** el principio de datos como producto proyecta una filosofía de pensamiento de producto sobre datos analíticos. Este principio significa que hay consumidores para los datos más allá del dominio. El equipo de dominio es responsable de satisfacer las necesidades de otros dominios proporcionando datos de alta calidad y su documentación asociada para facilitar el uso de los datos.
- ✔ **Infraestructura de datos de autoservicio como plataforma:** la plataforma debe simplificarse para que los dominios puedan publicar sus datos o consumir los datos

de otros dominios de una forma sencilla, sin tener que resolver detalles técnicos de las transformaciones, los formatos, etc.

- ✔ **Gobierno computacional federado:** el principio de gobierno federado consiste en que en lugar de tener un único equipo de Gobierno de datos, esta disciplina es compartida por todos los dominios. El objetivo principal del gobierno federado es crear un ecosistema de datos que cumpla con las reglas de la organización y las regulaciones de la industria.

La realidad es que este modelo todavía no está muy implantado en las organizaciones, pero a día de hoy, muchas se están planteando si adoptar esta filosofía descentralizada para la gestión de los datos y sustituir los equipos de arquitectura o de ingeniería centralizados por equipos autónomos embebidos dentro de los diferentes dominios y coordinados para no generar una entropía en la plataforma global de datos. Asimismo, a nivel de plataforma tecnológica, el movimiento hacia una filosofía Data Mesh implica sustituir grandes infraestructuras on-premise por soluciones de plataforma de datos en la nube más flexibles y elásticas (que se adaptan a la demanda), y con los datos encapsulados y publicados mediante APIs de microservicios o mediante repositorios de objetos como Amazon S3.

Para saber más

Si quieres saber más sobre Data Mesh, que es un concepto muy novedoso en el mercado, puedes leer un artículo muy interesante en [este enlace](#).

1.4.- Gobierno de datos.

Los Data Lakes han proliferado mucho durante todos estos años. Un Data Lake puede ser una herramienta muy potente, en la que almacenar muchos datos de cualquier tipo, que incluso no conocemos, para poder analizarlos más tarde. Además, permite que haya una variedad de usuarios y de tecnologías de procesamiento y análisis.

Estos principios seguidos sin control pueden generar un problema que se conoce como el Data Swamp, o la ciénaga de datos. Las empresas que han almacenado datos en el Data Lake sin control (con la suposición de que probablemente el día de mañana tendrán utilidad), que han permitido la ejecución de procesos que transforman o consumen datos sin un procedimiento o forma de actuar, o que permiten a los usuarios o aplicaciones consumir los datos indiscriminadamente se enfrentan a un problema de descontrol del Data Lake, en el que no se sabe muy bien qué datos hay, qué datos son válidos, qué procesos están transformando los datos, cómo se obtienen los datos que se consumen o qué significan, y quién está accediendo o consumiendo los datos.

Esta situación ha puesto de manifiesto la necesidad de contar con equipos de Gobierno de Datos o Data Governance. Si bien esta necesidad no es exclusiva por la existencia de Data Lakes, es verdad que la revolución que los Data Lakes o el uso de tecnologías Big Data han llevado a cabo a acrecentado esta necesidad.

Data Governance es, según Gartner, el conjunto de procesos, roles, políticas, estándares y métricas que garantizan el uso eficiente y efectivo de los datos, alineado con los objetivos de las empresas.

Un equipo de Data Governance, que está liderado por la figura del CDO (Chief Data Officer), es un equipo que tiene, entre otros, los siguientes objetivos:

- ✔ **Seguridad** de los datos: garantizar la privacidad, la confidencialidad y el acceso apropiado.
- ✔ **Accesibilidad** de los datos: almacenar, proteger, indexar y permitir el acceso a los datos.
- ✔ Gestión de los **datos maestros y de referencia**: gestión de los datos compartidos para reducir la redundancia y garantizar una mejor calidad de los datos a través de la definición y el uso estandarizados de valores de datos.
- ✔ **Calidad** de los datos: asegurar que los datos son correctos, verídicos, completos y libres de errores.
- ✔ **Usabilidad** de los datos: facilitar el uso de los datos mediante la publicación de diccionarios y catálogos de datos que permiten conocer qué datos existen y qué significan.
- ✔ **Monitorizar** el uso de los datos por parte de los diferentes equipos.
- ✔ **Cumplimiento** regulatorio en materia de datos.



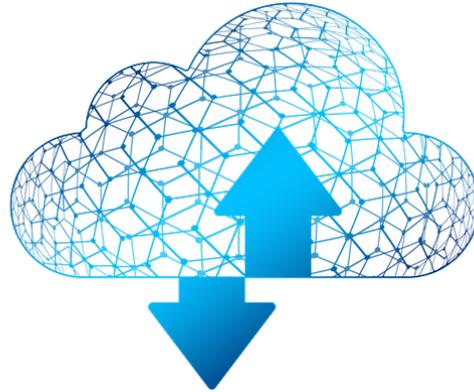
Íñigo Sanz (Dominio público)

Los equipos de Data Governance definen los procedimientos y las políticas sobre los datos, como pueden ser las políticas de purgado (borrado de datos), políticas de seguridad (quién puede ver qué dato), procedimiento o política de documentación, etc. así como proporcionan las herramientas para facilitar la gestión de las políticas y su cumplimiento. La suma de las políticas o procedimientos junto con las herramientas es lo que se conoce como el framework de gobierno de datos.

2.- Soluciones Hadoop-as-a-Service.

Caso práctico

Telco Max Spain ha desplegado una infraestructura Hadoop en sus propio centro de datos. Ha implementado muchos casos de uso y le está funcionando sin muchos problemas, pero la previsión es que en 2 meses, la plataforma no tendrá capacidad para almacenar todos los datos, y a nivel de procesamiento, con la puesta en producción las próximas semanas de varios proyectos de analítica predictiva, la previsión es que tampoco tendrá la potencia suficiente.



[Gerd Altmann](#) (Dominio público)

Por estos motivos, se están planteando migrar toda la infraestructura a un entorno cloud, ya que han leído que dispone de ciertas ventajas en materia de escalabilidad y elasticidad, aunque tienen dudas al haber escuchado que su coste puede ser más elevado.

Vamos a ver con ellos dos soluciones de Hadoop en cloud para valorar si merece la pena hacer un despliegue en la nube o, por el contrario, es mejor hacer un despliegue en la propia infraestructura.

Hadoop es una plataforma que inicialmente fue creada para su despliegue en infraestructura propia, también denominada on-premise. Sin embargo, los proveedores de soluciones cloud comenzaron a ofrecer, dentro de su portfolio de servicios, soluciones de Hadoop como servicio, que consiste en servicios por lo que se pueden levantar clústers Hadoop a demanda, que se arrancan en un corto periodo de tiempo (10-15 minutos) y que ofrecen las siguientes características:

- ✓ Son servicios en modalidad pago por uso, por lo que su coste es sólo las horas que está levantado el clúster por el precio de las máquinas levantadas. Como ejemplo, un clúster en Amazon Web Services de 20 nodos, con 122 gigabytes de memoria por nodo y 16 cores, tiene un coste de 32 euros por hora.
- ✓ Son servicios elásticos que se pueden escalar en cuanto a número de nodos a demanda, y por ejemplo, tener un clúster de 10 nodos durante el día, y por la noche, si se van a lanzar procesos complejos de ingesta y cálculos masivos, incrementarlo

a 30 nodos sólo las horas en las que las tareas de ingesta y procesamiento van a llevarse a cabo, y pagando por lo tanto sólo por ese incremento.

- ✔ Suelen tener distribuciones Hadoop que incluyen la mayoría de componentes del ecosistema, y que es muy fácil de configurar o utilizar porque ya vienen preconfigurados.
- ✔ Se integran fácilmente con el resto de los servicios de los proveedores de cloud, como puede ser los servicios de identidad, seguridad, control de costes, etc.

Por estas características, este tipo de soluciones está teniendo buena acogida en el mercado, siendo las principales distribuciones por utilización la de Amazon Web Services, denominada EMR, y la de Microsoft Azure, denominada HDInsight.

Para saber más

Un dato importante sobre los costes de las plataformas de Hadoop como servicio es que su modelo de coste por hora y nodo es muy ventajoso para reducir el tiempo de ejecución de tareas sin impactar en el coste.

Veamos un ejemplo: supongamos que tenemos un clúster de 20 nodos que tiene un coste de 32 euros por hora, y que tarda en ejecutar una tarea de ingesta de datos de una base de datos y la generación de un cuadro de mando posteriormente un total de 10 horas. El coste de ejecución de la tarea sería, por lo tanto, de 320 euros (32 euros x 10 horas).

Si la tarea se puede paralelizar, es decir, si son cálculos que añadiendo más nodos en paralelo se puede reducir linealmente su tiempo, podríamos incrementar el tamaño del clúster para tener 200 nodos en lugar de 20, y reduciendo el tiempo a 1 hora en lugar de 10 horas. El coste sería de un clúster 10 veces mayor sería de 320 euros, pero al pagar sólo una hora, el coste de ejecución de la tarea sería el mismo, 320 euros, ¡así que se puede conseguir reducir el tiempo de ejecución de la tarea sin que impacte en el coste!

Autoevaluación

Indica si la siguiente afirmación es verdadera o falsa:

Amazon EMR o Azure HDInsight permiten crear un clúster en mi propia infraestructura en cuestión de minutos

- Verdadero Falso

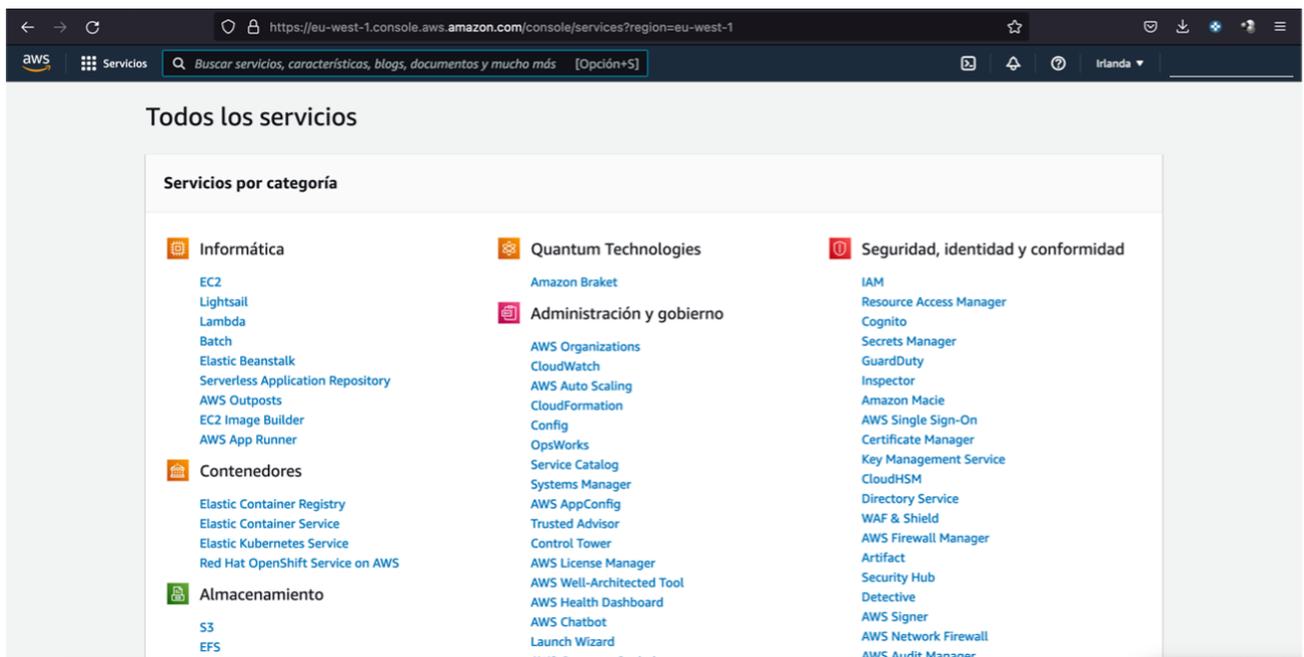
Falso

Falso: las soluciones de Hadoop como servicio ofrecen la posibilidad de crear clústers pero en infraestructura cloud, no en la propia infraestructura.

2.1.- Amazon EMR.

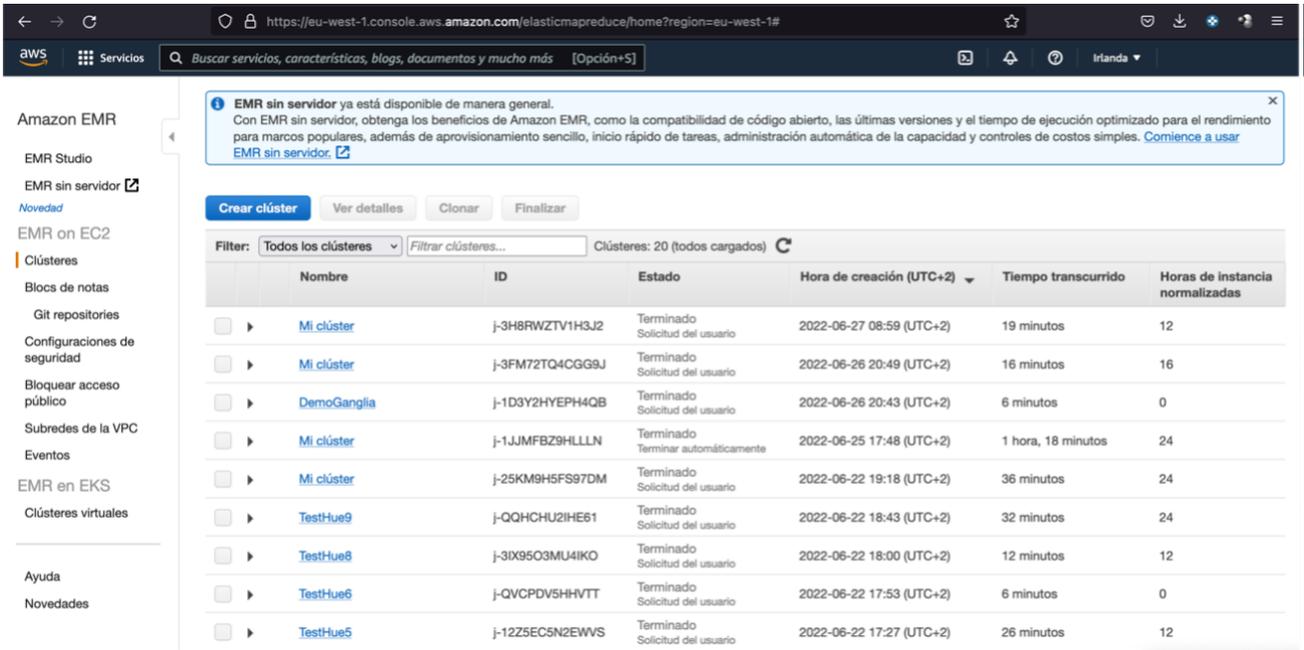
Amazon EMR, que son las siglas de Amazon Elastic MapReduce, es un servicio de Amazon Web Services que permite crear clusters Hadoop a demanda. Utiliza una distribución propia de Amazon que permite seleccionar los componentes que van a lanzarse en el cluster (Hive, Spark, etc.). Se ejecuta sobre máquinas EC2 (IaaS) y su coste asociado es el alquiler de las máquinas por horas más un sobrecoste de aproximadamente el 25%.

Para arrancar un clúster, en primer lugar hay que acceder a la consola de Amazon Web Services y seleccionar, dentro de los servicios de Análisis, EMR:



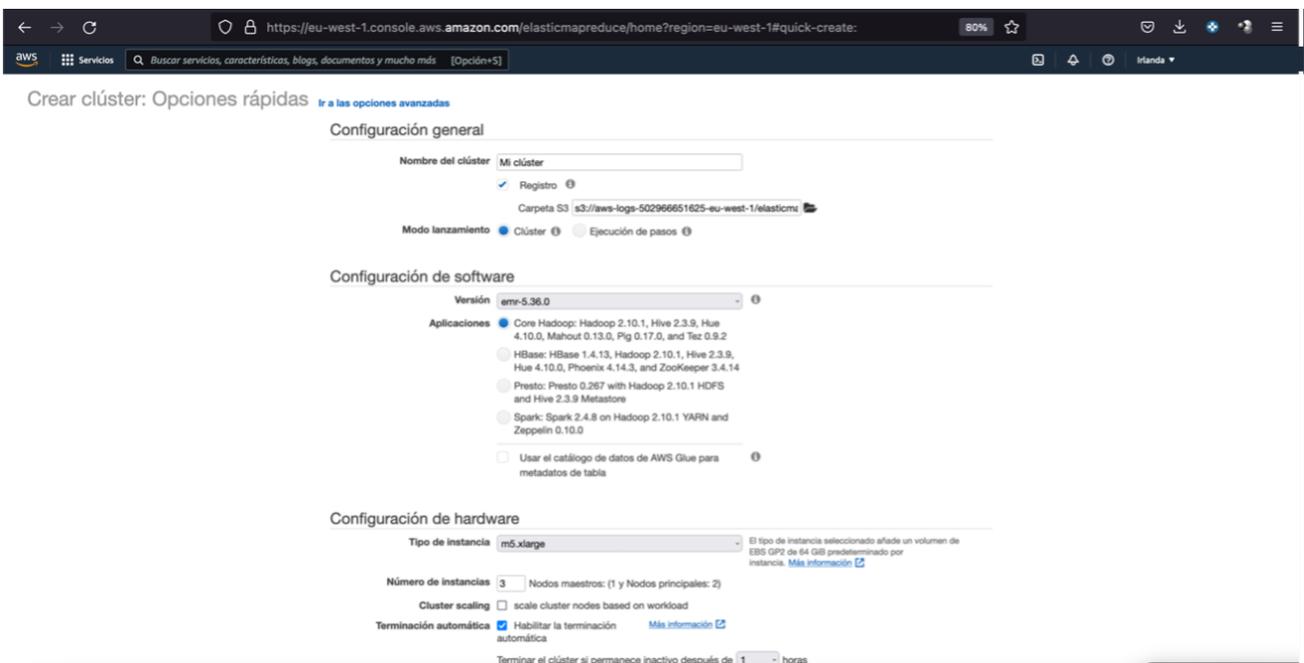
Íñigo Sanz (Dominio público)

A continuación se muestra la pantalla principal de EMR, donde se pueden ver los clústers activos o los lanzados recientemente:



Íñigo Sanz (Dominio público)

Para arrancar un nuevo clúster, hay que hacer clic en el botón “Crear clúster”



Íñigo Sanz (Dominio público)

En esta pantalla, se debe dar un nombre al clúster, elegir la versión de EMR (la versión influye en la versión de componentes de Hadoop que se instalarán), el tipo de clúster o el tipo de instancia sobre la que se ejecutará el clúster que estamos arrancando.

El tipo de clúster permite arrancar un clúster Hadoop especializado en un tipo de tareas o herramientas, permitiendo los siguientes tipos:

- ✔ **Core Hadoop:** es un tipo de clúster estándar, con los componentes principales del ecosistema.
- ✔ **HBase:** es un tipo de clúster orientado al procesamiento de datos basado en HBase

para su consumo operacional.

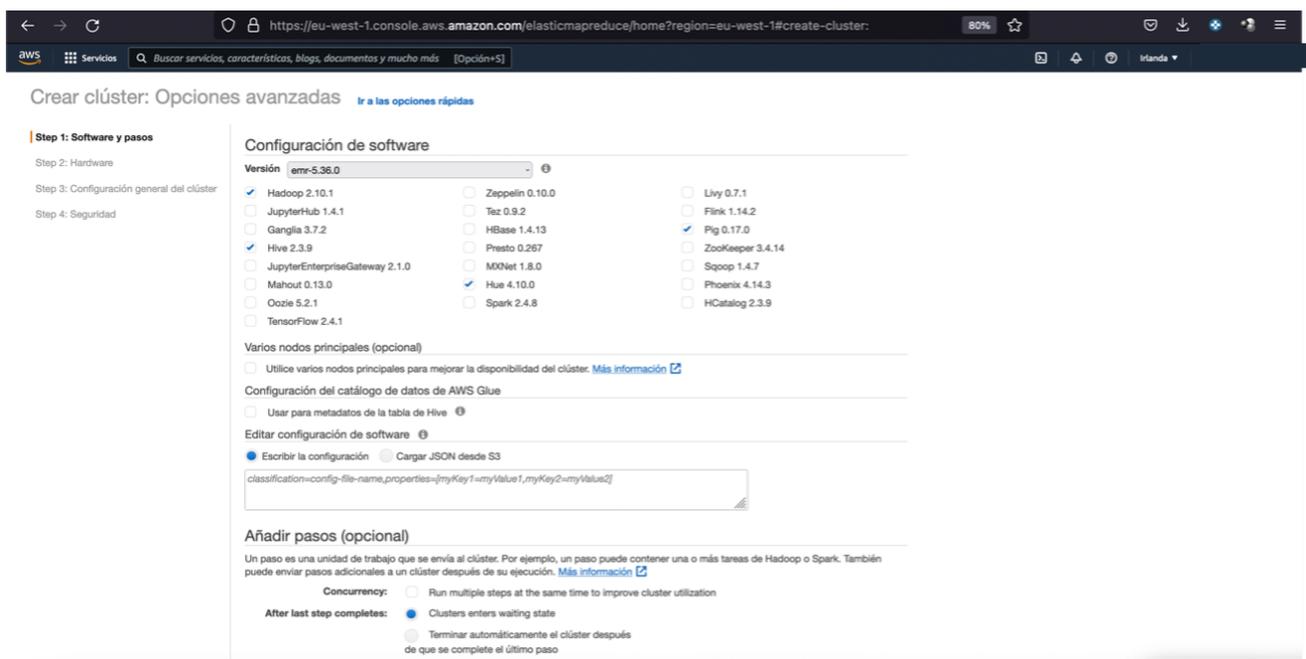
- ✓ **Presto:** Presto es un proyecto de Apache que permite acceso SQL sobre diferentes sistemas de almacenamiento. Este tipo de clúster arranca Presto así como otros componentes como HDFS o Hive.
- ✓ **Spark:** este tipo de clúster está especializado en ejecutar procesos de Spark sobre YARN, habitualmente para procesamiento de datos.

En cuanto a los tipos de instancia, se corresponden con los tipos de máquinas que Amazon ofrece como servidores en modalidad pago por uso. Cada instancia tiene unas capacidades en términos de memoria y procesador, siendo más caro cuanto mayor capacidad tiene.

En la pantalla anterior se permite, además, el número de nodos que compondrán el clúster así como da la opción de definir una política de autoescalado, por la que podremos añadir automáticamente más nodos al clúster en caso de que la carga de los mismos supere un umbral. Esta funcionalidad es especialmente útil para poder hacer crecer el clúster para adaptarlo a la carga real de forma automática. Esta política permite, además, reducir también el tamaño del clúster en los periodos en los que la carga es baja.

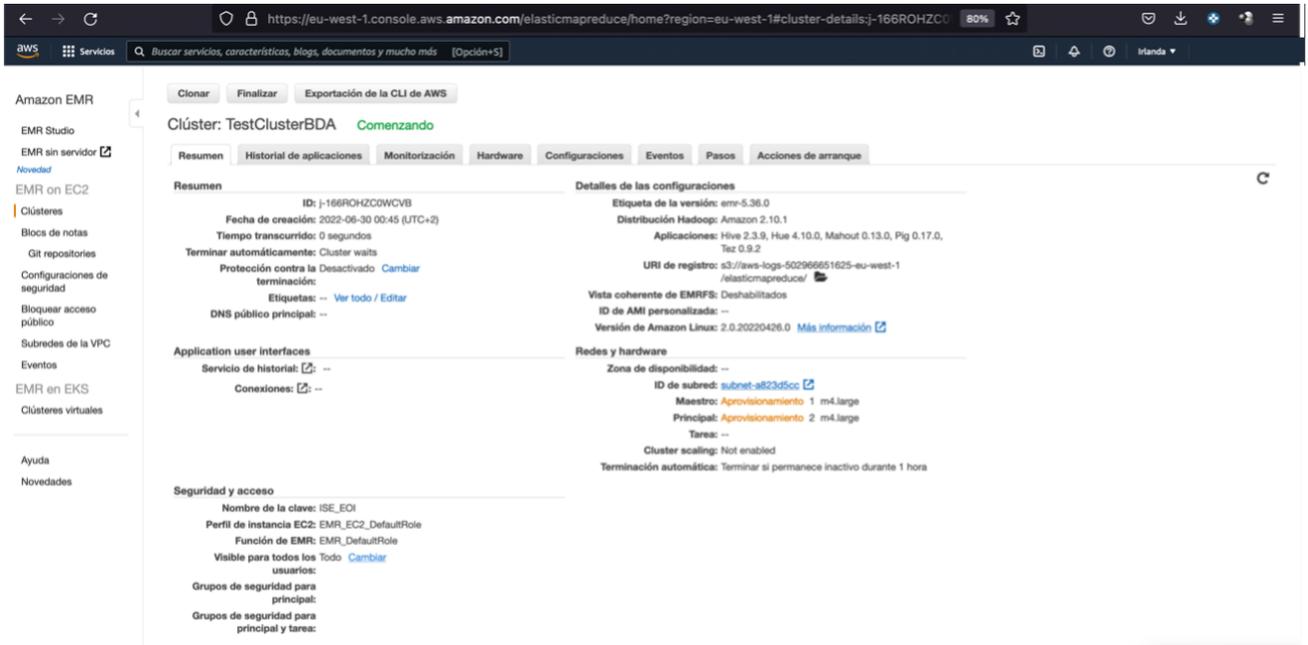
Por último, se puede configurar algún aspecto de seguridad como la clave privada que se utilizará para entrar por consola.

Es preciso indicar, además, que existe un modo de configuración avanzada con la que, entre otros, se puede seleccionar qué componentes del ecosistema se quiere instalar y arrancar en el clúster que vamos a crear.



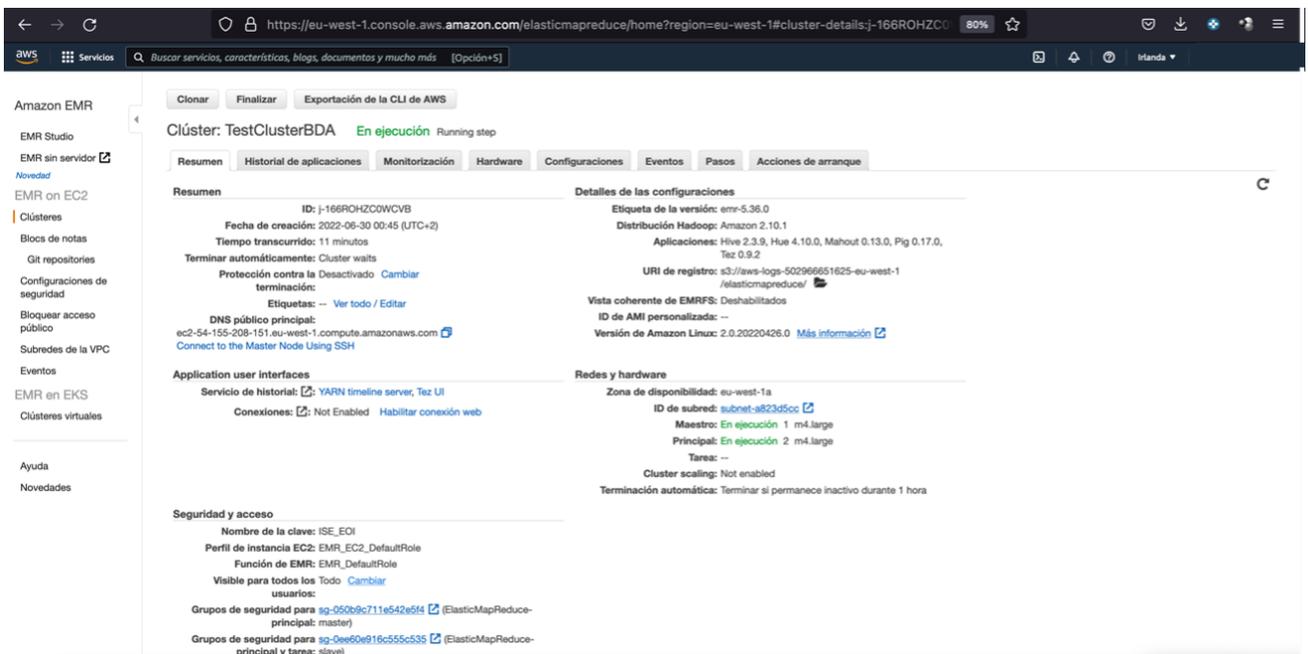
Íñigo Sanz (Dominio público)

Un vez hacemos clic en el botón “Crear clúster” de la parte inferior de la página, se inicia el proceso de aprovisionamiento y arranque del clúster.



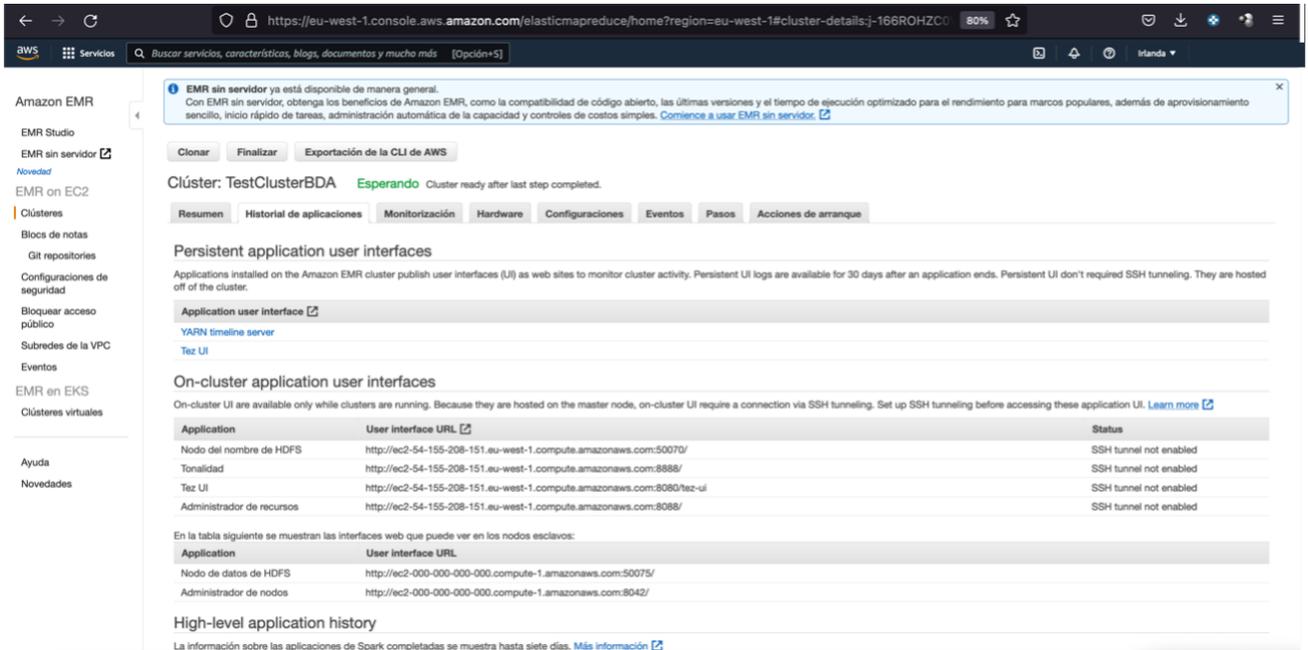
Íñigo Sanz (Dominio público)

Tras 10 minutos aproximadamente, el clúster ya se encuentra activo:



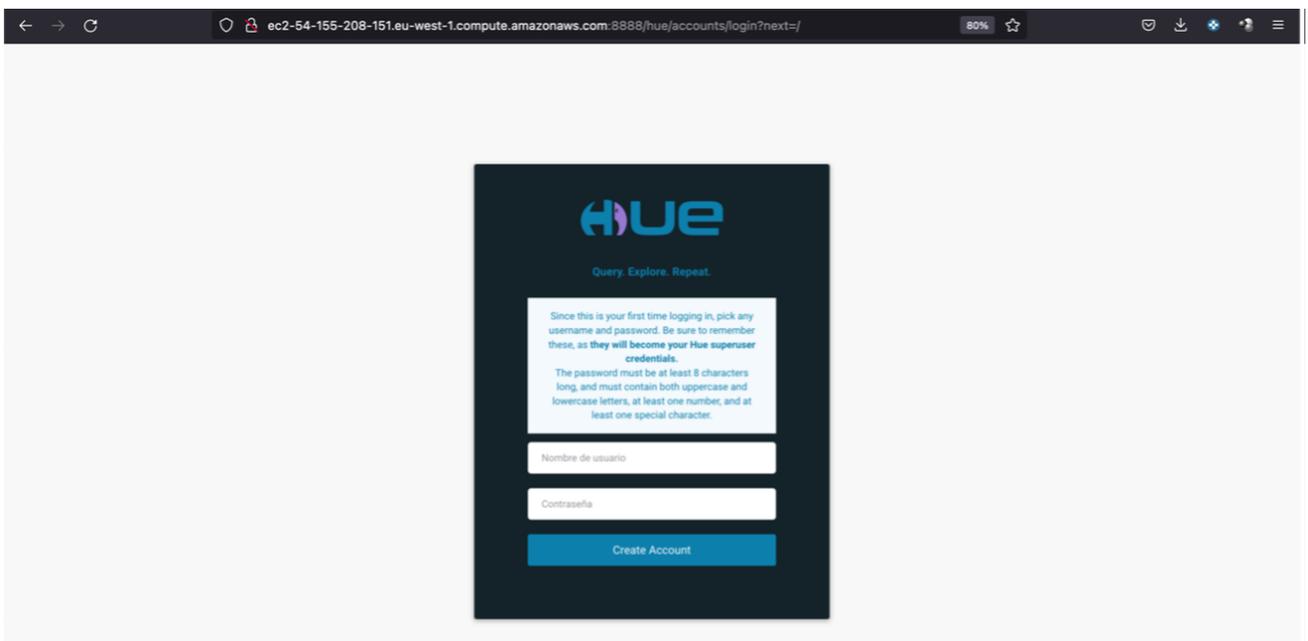
Íñigo Sanz (Dominio público)

En la segunda pestaña, denominada “Historial de aplicaciones”, podemos ver las direcciones de los diferentes interfaces web a los que podemos acceder:



Íñigo Sanz (Dominio público)

Por ejemplo, el segundo enlace que aparece es el de Hue (traducido al castellano como “Tonalidad”):



Íñigo Sanz (Dominio público)

Asimismo, podemos acceder al nodo master por SSH utilizando la clave privada que se dio de alta en el proceso de arranque:

```
Descargas — hadoop@ip-172-31-1-184:~ — ssh -i ISE_EOI.pem hadoop@ec2-54-155-208-151.eu-west-1.compute.amazonaws.com — 166x36
i.sanz@it506 Downloads % ssh -i ISE_EOI.pem hadoop@ec2-54-155-208-151.eu-west-1.compute.amazonaws.com
The authenticity of host 'ec2-54-155-208-151.eu-west-1.compute.amazonaws.com (54.155.208.151)' can't be established.
ED25519 key fingerprint is SHA256:m+GcIQ+pk0EAiXfRwbGp0Rb8uDi5D3tXK9eTuh57d8.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-54-155-208-151.eu-west-1.compute.amazonaws.com' (ED25519) to the list of known hosts.
Last login: Wed Jun 29 23:06:12 2022

  _| _|_ )
  _| ( /
  --\_____|
                Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEE MMMMMM      MMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::M:::::M      M:::::M R:::::R:::::R
EE::::::::::::::::::E M:::::M      M:::::M R:::::R:::::R
E:::::E      EEEEE M:::::M      M:::::M RR::::R      R::::R
E:::::E      M:::::M:M:M      M:::M R::::R      R::::R
E:::::EEEEEEEEEE M:::::M M:::M M:::M M:::::M R::::RRRRR::::R
E::::::::::::::::::E M:::::M M:::M:M      M:::::M R:::::R:::::RR
E:::::EEEEEEEEEE M:::::M M:::::M M:::::M R::::RRRRR::::R
E:::::E      M:::::M M:::M      M:::::M R::::R      R::::R
E:::::E      EEEEE M:::::M      MMM      M:::::M R::::R      R::::R
EE::::::::::::::::::E M:::::M      M:::::M R::::R      R::::R
E::::::::::::::::::E M:::::M      M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEE MMMMMM      MMMMMMM RRRRRR      RRRRRR

[hadoop@ip-172-31-1-184 ~]$ hadoop fs -ls /tmp
Found 3 items
drwxrwxrwt - yarn hdfsadmin group 0 2022-06-29 22:52 /tmp/entity-file-history
drwxrwxrwx - mapred mapred 0 2022-06-29 22:52 /tmp/hadoop-yarn
drwx-wx-wx - hive hdfsadmin group 0 2022-06-29 22:54 /tmp/hive
[hadoop@ip-172-31-1-184 ~]$
```

Íñigo Sanz (Dominio público)

Puedes ver en el ejemplo cómo se ha accedido a EMR y se ha ejecutado un comando “`hadoop fs -ls`”.

Por último, para parar un clúster, sólo hay que hacer clic en el botón “Finalizar”

Para saber más

Si quieres obtener más información sobre Amazon EMR, puedes consultar la [documentación oficial](#), o incluso seguir [el tutorial](#) donde se describe paso a paso un ejemplo de uso de un clúster EMR.

Autoevaluación

¿Cuál de las siguientes afirmaciones sobre EMR no es correcta?

- EMR permite arrancar clústers Hadoop rápidamente, por lo que es muy útil para hacer pruebas con Hadoop.
- EMR permite configurar qué componentes del ecosistema Hadoop arrancar.
- EMR puede adaptar el número de servidores a la carga real que esté soportando, por lo que sólo pagas por el uso real.

- EMR, que son las siglas de Elastic MapReduce, sólo permite MapReduce como framework para procesar datos.

La afirmación es correcta, por lo que no debe ser marcada (se pide las afirmaciones incorrectas).

La afirmación es correcta, por lo que no debe ser marcada (se pide las afirmaciones incorrectas).

La afirmación es correcta, por lo que no debe ser marcada (se pide las afirmaciones incorrectas).

El nombre del servicio, Elastic MapReduce puede hacer pensar que sólo permite MapReduce, pero no es correcto. Hay que reconocer que con el nombre, Amazon no ha estado muy acertado ;-)

Solución

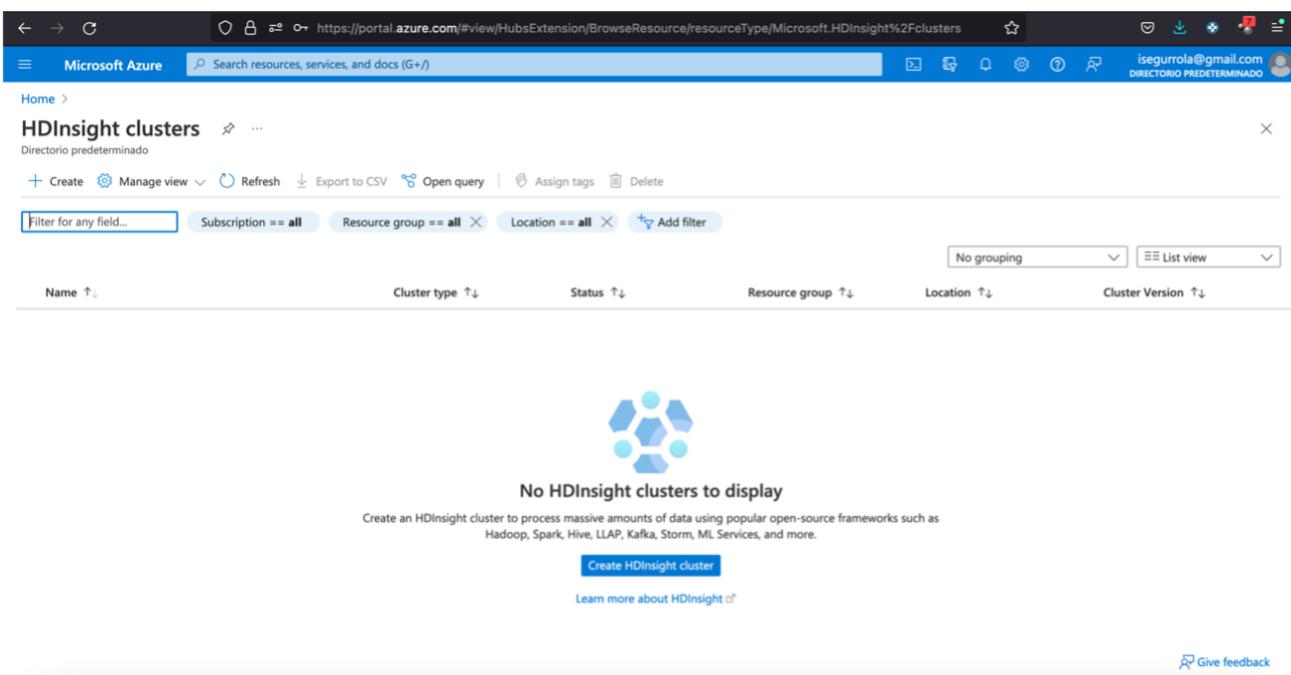
1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta

2.2.- Microsoft Azure HDInsight.

HDInsight es la solución de Microsoft Azure de Hadoop como servicio, que permite arrancar clústers Hadoop a demanda en una modalidad de pago por uso. Se podría decir que es un servicio muy parecido a Amazon EMR, aunque tiene algunas diferencias, entre las que destacan:

- ✓ HDInsight utiliza la distribución HDP de Hortonworks en lugar de una propia en el caso de EMR.
- ✓ HDInsight proporciona Ambari como herramienta de administración y monitorización, mientras que EMR proporciona una herramienta propia, mucho menos potente, y Ganglia para monitorización.

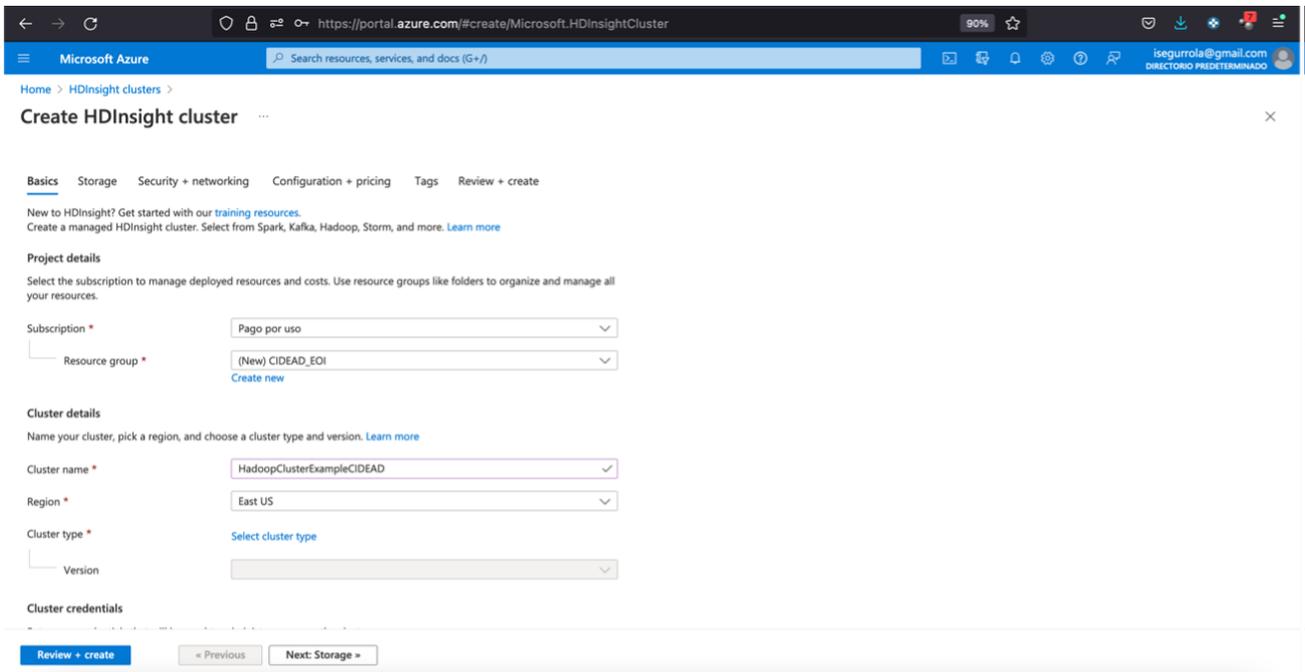
Para poder utilizar el servicio HDInsight, en primer lugar hay que tener una cuenta en Microsoft Azure, y posteriormente acceder al servicio HDInsight:



The screenshot shows the Microsoft Azure portal interface for HDInsight clusters. The page title is "HDInsight clusters" and it indicates a "Directorio predeterminado". There are navigation options like "Create", "Manage view", "Refresh", "Export to CSV", "Open query", "Assign tags", and "Delete". A search bar is present with the text "Filter for any field...". Below the search bar, there are filters for "Subscription == all", "Resource group == all", and "Location == all". A table header is visible with columns: "Name", "Cluster type", "Status", "Resource group", "Location", and "Cluster Version". The main content area displays a message: "No HDInsight clusters to display". Below this message, there is a button labeled "Create HDInsight cluster" and a link "Learn more about HDInsight". The page also includes a "Give feedback" link in the bottom right corner.

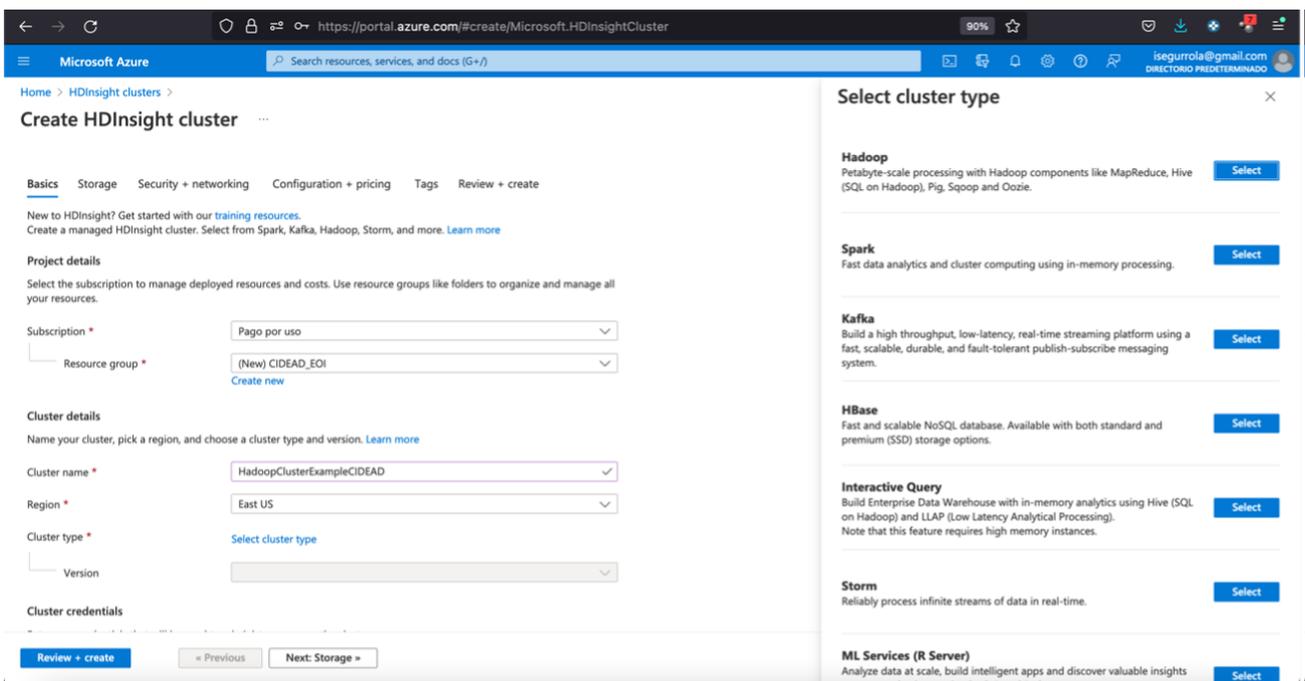
Íñigo Sanz (Dominio público)

Haciendo clic en el enlace "Create", se inicia el asistente para arrancar un clúster, teniendo una primera pantalla en la que se muestra un formulario para introducir el nombre del clúster, la región donde queremos que se despliegue o el tipo de clúster:



Íñigo Sanz (Dominio público)

HDInsight dispone de diferentes tipos de clúster Hadoop, que se diferencian en el tipo de componentes del ecosistema Hadoop que utilizan, y de algunas optimizaciones adaptadas al tipo de uso:



Íñigo Sanz (Dominio público)

La opción más general es la primera, la de tipo Hadoop.

A continuación pide crear una contraseña de administración:

Home > HDInsight clusters >

Create HDInsight cluster

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name *

Region *

Availability zone

Cluster type * **Hadoop**
[Change](#)

Version *

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Cluster login username *

Cluster login password *

Confirm cluster login password *

Secure Shell (SSH) username *

Use cluster login password for SSH

[Review + create](#) [< Previous](#) [Next: Storage >](#)

Íñigo Sanz (Dominio público)

Como siguiente paso, se puede definir qué sistema de almacenamiento utilizar:

Home > HDInsight clusters >

Create HDInsight cluster

Basics **Storage** Security + networking Configuration + pricing Tags Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type *

Selection method * Select from list Use access key

Primary storage account *
[Create new](#)

Container *

Data Lake Storage Gen1

Provide details for the cluster to access Data Lake Storage Gen1. The cluster will be able to access any Data Lake Storage Gen1 accounts that the chosen service principal has access to.

Data Lake Storage Gen1 access [Configure access settings](#)

Additional Azure Storage

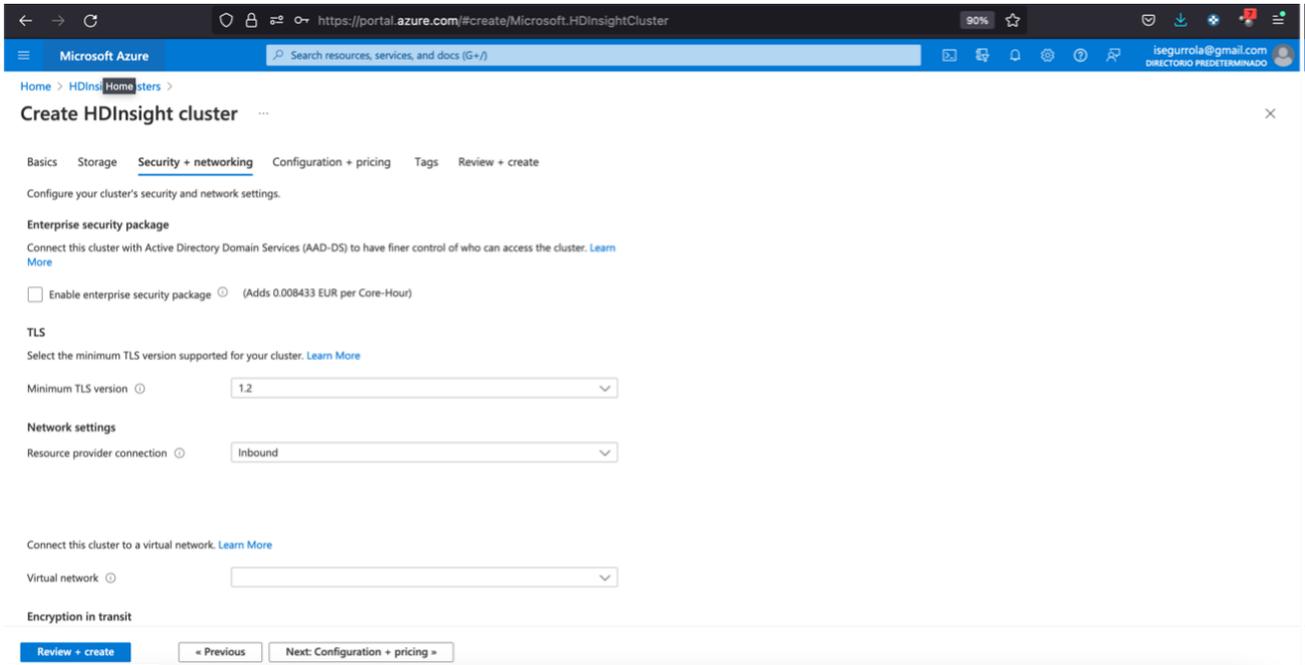
Link additional Azure Storage accounts to the cluster.

[Add Azure Storage](#)

[Review + create](#) [< Previous](#) [Next: Security + networking >](#)

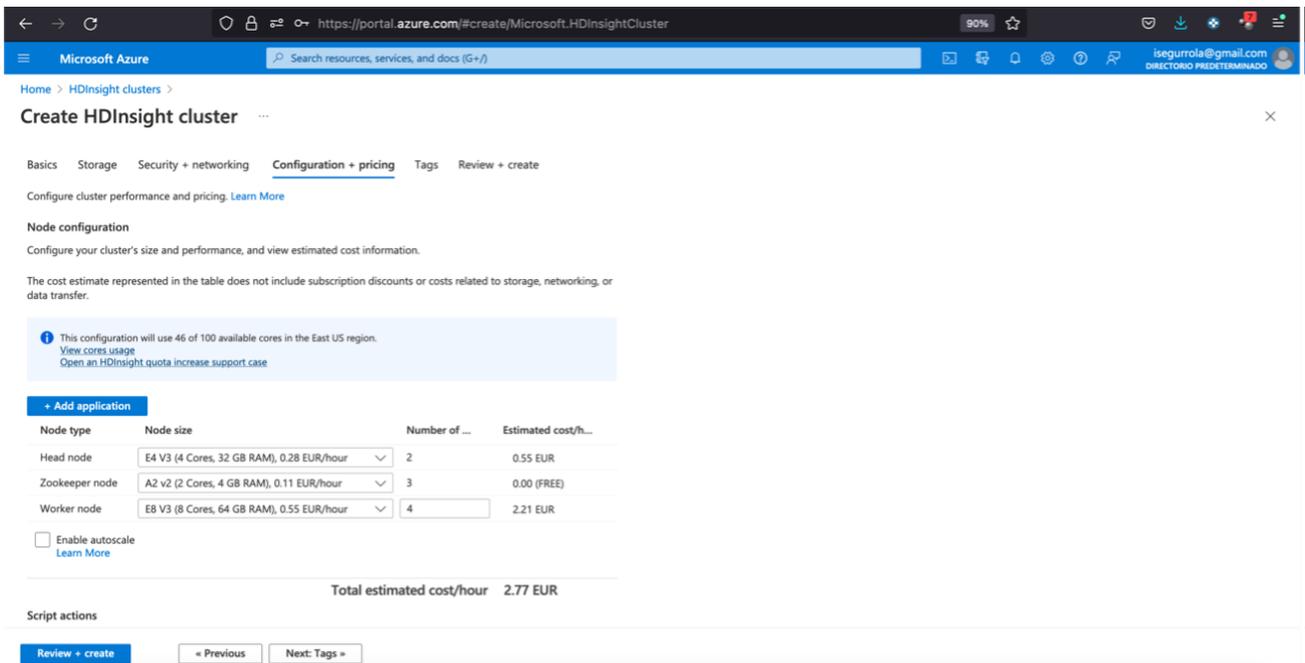
Íñigo Sanz (Dominio público)

A continuación se puede configurar la red, por si queremos añadir algún tipo de conectividad con nuestra infraestructura o crear grupos de redes, por ejemplo:



Íñigo Sanz (Dominio público)

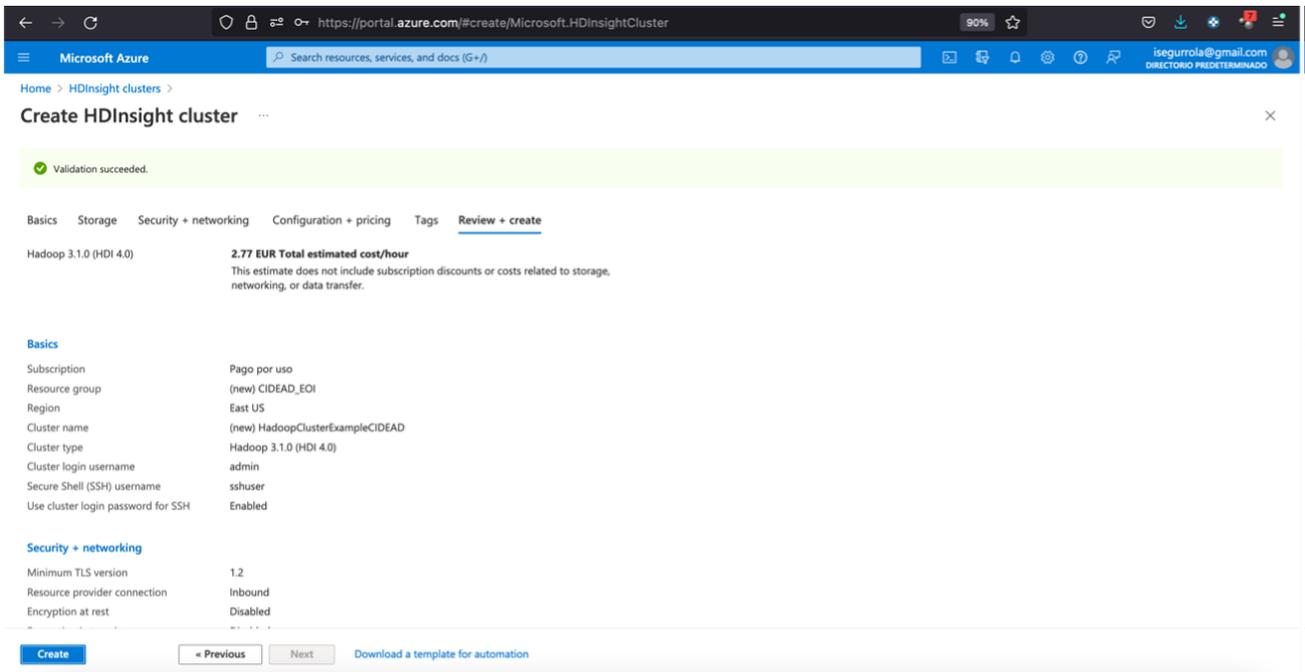
Como paso siguiente, se muestra una pantalla con el resumen y el precio del clúster que vamos a arrancar:



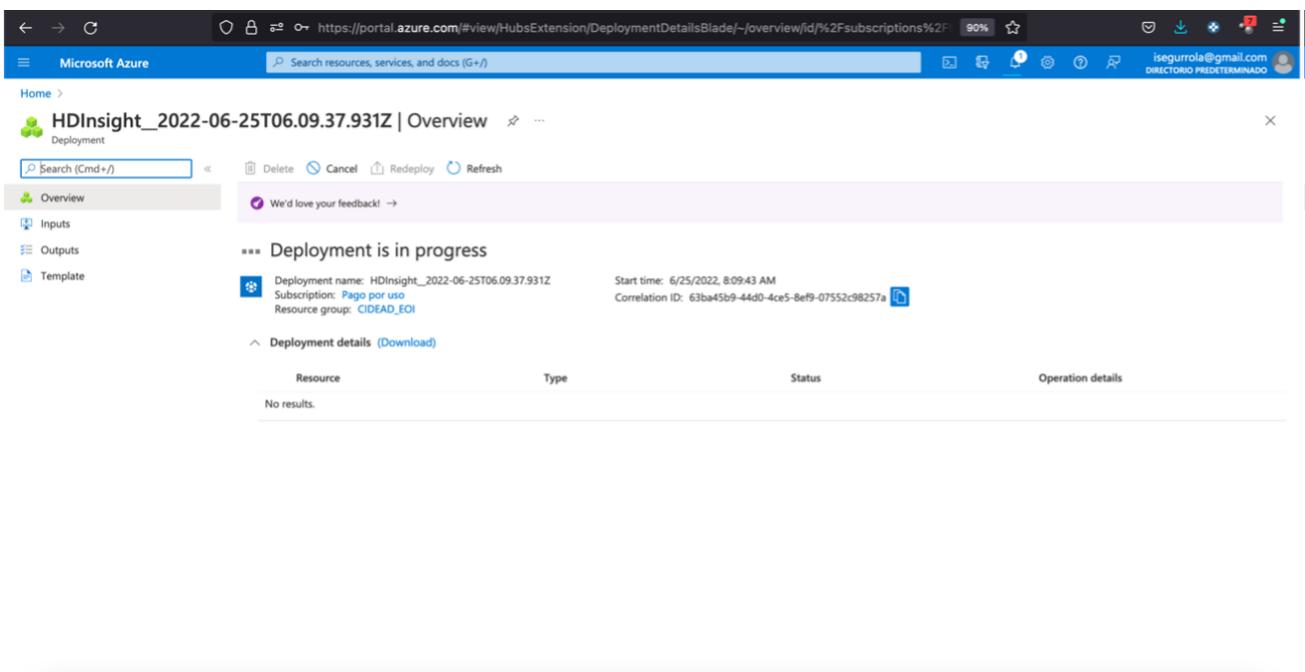
Íñigo Sanz (Dominio público)

Como puedes ver, en este caso se está arrancando un clúster con 2 nodos master y 4 nodos worker de 64 gigabytes de RAM que cuestan 2,77 euros la hora.

Como último paso, Azure valida que los datos introducidos son correctos y se lanza el arranque del clúster:



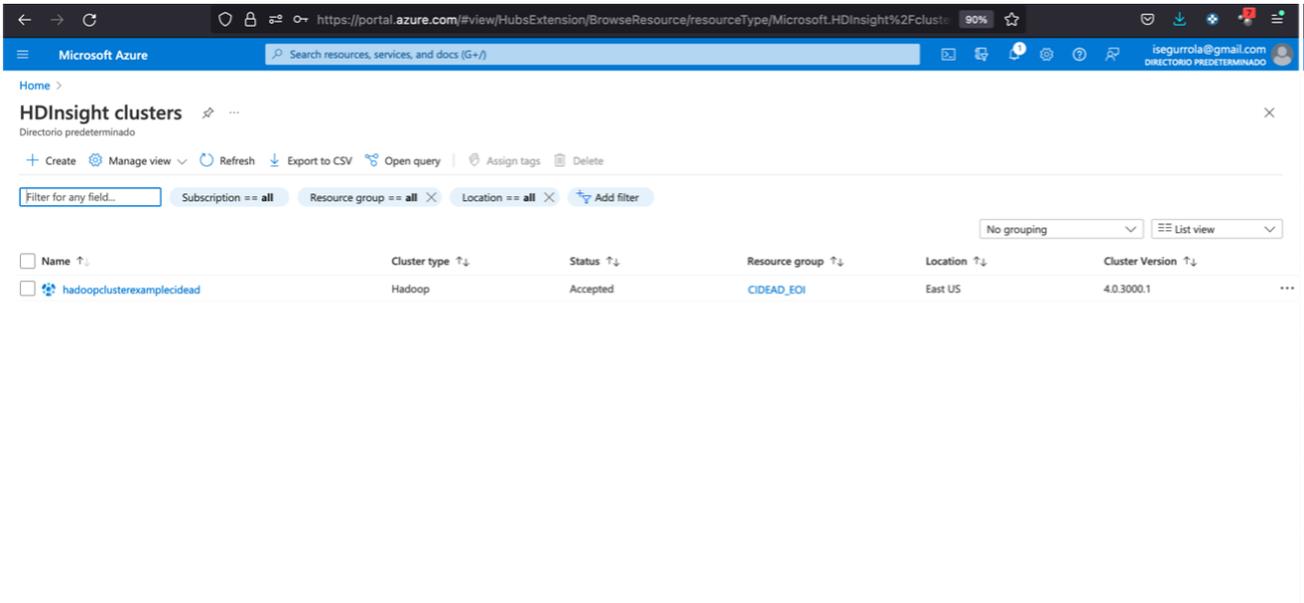
Íñigo Sanz (Dominio público)



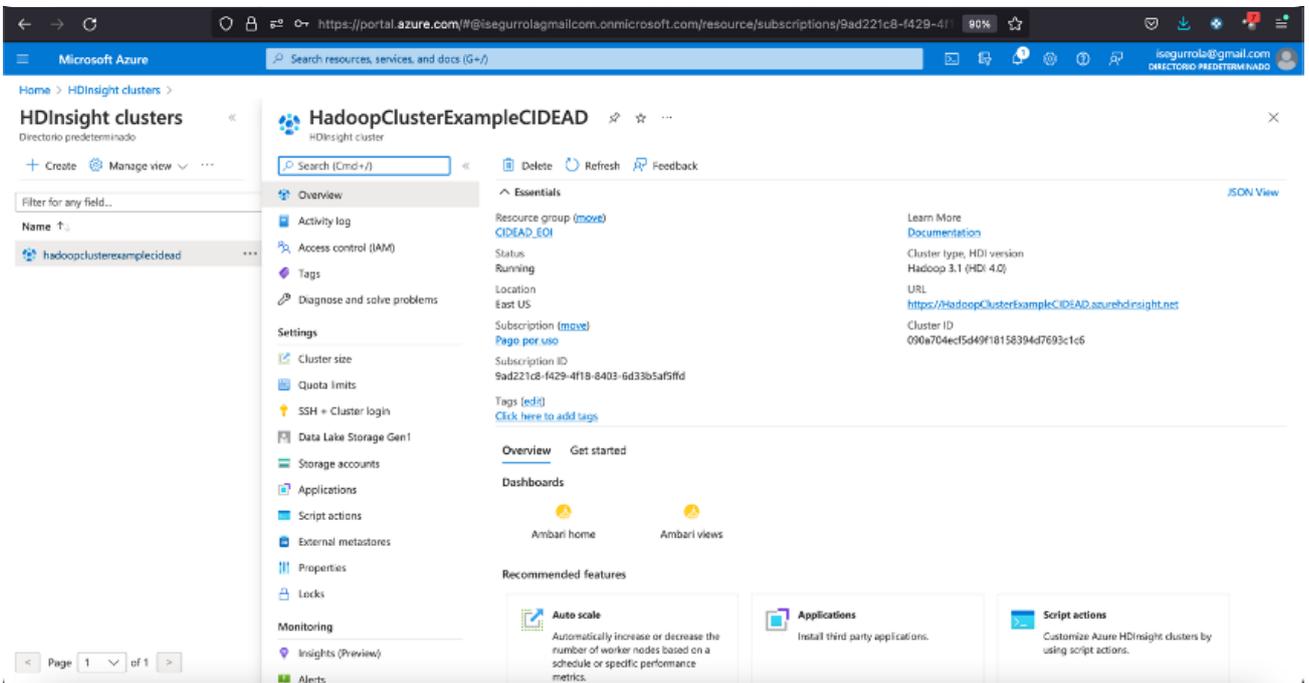
Íñigo Sanz (Dominio público)

El proceso de arranque dura unos 15 minutos, y a partir de entonces ya tendremos acceso a los nodos master, donde están desplegados los servicios como Hive o Hue, y podremos lanzar las tareas o utilizar el clúster de forma 100% funcional.

Volviendo a acceder al servicio de HDInsight en la consola, se puede ver la lista de clústers, pudiendo entrar en cada uno de ellos para ver su configuración o hacer cambios de dimensionamiento, etc.



Íñigo Sanz (Dominio público)



Íñigo Sanz (Dominio público)

The screenshot shows the 'Cluster size' configuration page for an HDInsight cluster. The table below summarizes the node types and their costs:

Node type	Node size	Number of nodes	Estimated cost/hour
Head node	E4 V3 (4 Cores, 32 GB RAM), 0.28 EUR/hour	2	0.55 EUR
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.11 EUR/hour	3	0.00 EUR (FREE)
Worker node	E8 V3 (8 Cores, 64 GB RAM), 0.55 EUR/hour	4	2.21 EUR
Total estimated cost/hour			2.77 EUR

Íñigo Sanz (Dominio público)

Para saber más

Si quieres conocer más cómo funciona HDInsight, en el siguiente enlace encontrarás la documentación oficial de Microsoft sobre HDInsight: [documentación oficial de HDInsight](#).

