**Hadoop**

# Hadoop

Apache Hadoop es una plataforma de software de código abierto para el almacenamiento distribuido y procesamiento distribuido de grandes conjuntos de datos en clústeres informáticos construidos del hardware de productos básicos.  Los servicios de Hadoop prevén almacenamiento de datos, procesamiento de datos, acceso a los datos, gestión de datos, seguridad y operaciones.

## BENEFICIOS

Algunas de las razones por las que las organizaciones usan Hadoop es su capacidad para almacenar, administrar y analizar grandes cantidades de datos estructurados y no estructurados de manera rápida, confiable, flexible y de bajo costo.

## Caracteristicas de Hadoop

* **Escalabilidad y rendimiento**: el procesamiento distribuido de datos locales para cada nodo en un clúster permite a Hadoop almacenar, gestionar, procesar y analizar datos a escala de petabytes.
* **Confiabilidad**: los grandes clústeres informáticos son propensos a la falla de los nodos individuales en el clúster. Hadoop es fundamentalmente resistente: cuando un nodo falla, el procesamiento se redirige a los nodos restantes en el clúster y los datos se vuelven a replicar automáticamente en preparación para futuras fallas de nodo.
* **Flexibilidad**: a diferencia de los sistemas de administración de bases de datos relacionales tradicionales, no es necesario crear esquemas estructurados antes de almacenar datos. Puede almacenar datos en cualquier formato, incluidos los formatos semiestructurados o no estructurados, y luego analizar y aplicar el esquema a los datos cuando se leen.

**Bajo costo**: a diferencia del software propietario, Hadoop es de código abierto y funciona con hardware básico de bajo costo.



## Capacidades de Hadoop

### **ALMACENAMIENTO DE DATOS**

El Sistema de archivos distribuidos de Hadoop ( [HDFS](https://es.hortonworks.com/apache/hdfs/) ) proporciona almacenamiento escalable, tolerante a fallas y rentable para su gran lago de datos. Fue diseñado para abarcar grandes grupos de servidores de productos básicos que aumentan hasta cientos de petabytes y miles de servidores. Al distribuir el almacenamiento a través de muchos servidores, el recurso de almacenamiento combinado puede crecer linealmente con la demanda sin dejar de ser económico en cada cantidad de almacenamiento.

### **PROCESAMIENTO DE DATOS**

[MapReduce](https://es.hortonworks.com/apache/mapreduce/) es el marco original para la escritura masiva de aplicaciones paralelas que procesan grandes cantidades de datos estructurados y no estructurados almacenados en HDFS. MapReduce puede aprovechar la localidad de los datos, procesándola cerca del lugar donde está almacenada en cada nodo del clúster para reducir la distancia sobre la que debe transmitirse.

### **ACCESO Y ANÁLISIS DE DATOS**

Las aplicaciones pueden interactuar con los datos en Hadoop utilizando SQL por lotes o interactivo ( [Apache Hive](https://es.hortonworks.com/apache/hive/) ) o acceso de baja latencia con NoSQL ( [Apache HBase](https://es.hortonworks.com/apache/hbase/) ). Hive permite a los usuarios comerciales y analistas de datos utilizar sus herramientas preferidas de análisis, presentación de informes y visualización de negocios con Hadoop. Los datos almacenados en HDFS en Hadoop se pueden buscar utilizando [Apache Solr](https://es.hortonworks.com/apache/solr/) .

### **GOBERNANZA Y SEGURIDAD DE DATOS**

El ecosistema de Hadoop amplía el acceso y procesamiento de datos con potentes herramientas para la administración e integración de datos que incluyen administración de seguridad centralizada ( [Apache Ranger](https://es.hortonworks.com/apache/ranger/) ) y etiquetado de clasificación de datos ( [Apache Atlas](https://es.hortonworks.com/apache/atlas/) ), que combinan políticas dinámicas de acceso a datos que previenen proactivamente las violaciones de acceso a datos. La seguridad del perímetro de Hadoop también está disponible para integrarse con los sistemas de seguridad empresariales existentes y controlar el acceso de los usuarios a Hadoop ( [Apache Knox](https://es.hortonworks.com/apache/knox/) ).

# Casos de uso

### Datos de empresa y optimización del almacenamiento

El almacén de datos de la empresa (EDW) es el repositorio de datos central de la organización construida para apoyar las decisiones de negocios. EDW contiene datos relacionados con las áreas que la empresa quiere analizar. Para un fabricante, podría ser de clientes, productos o lista de datos materiales. EDW se construye al extraer los datos de un número de sistemas operativos. Conformen se alimentan los datos en el EDW, se convierten, formatean y resumen para presentar una visión corporativa única. Se agregan los datos en el almacén de datos conforme pasa el tiempo en forma de instantáneas y un almacén de datos de la empresa contiene normalmente datos de 5 a 10 años. Una arquitectura de almacén de datos Hadoop permite el análisis más profundo y reporte avanzado de estos diversos conjuntos de datos.

El almacén de datos de la empresa se ha convertido en un componente estándar de la arquitectura de datos corporativa. Sin embargo, la complejidad y el volumen de datos ha planteado algunos retos interesantes para la eficacia de las soluciones existentes de EDW.   
  
Darse cuenta del potencial transformador de los macrodatos depende de la capacidad de gestionar la complejidad de las corporaciones mientras movilizan las fuentes de datos para todos los tipos como los sociales, web, IoT y más. La integración de las nuevas fuentes de datos en el sistema existente EDW impulsará a las corporaciones dar unas analíticas más profundas. Más importante aún, la optimización de EDW que utilizan Hadoop proporciona un entorno rentable con un rendimiento óptimo, escalable y flexible.

ELEMENTOS DE SOLUCIONES

### **Plataforma de datos de Hortonworks**

La arquitectura del almacén de datos de Hadoop para la gobernanza e integración de los datos, la gestión de los datos, el acceso de los datos, la seguridad y las operaciones— creados para la integración profunda con su tecnología de centro de datos existente.

### **Syncsort**

### ETL de alto rendimiento para acceder y abordar a las empresas de datos tradicionales a HDP. **JethroData**

Motor de alto rendimiento analítico para BI interactivo sobre Hadoop Big Data.

### **SERVICIOS PROFESIONALES**

Orientación experta y soporte para probar rápidamente el valor de su nueva arquitectura y maximizar el valor de los datos comprobados y validadeos de Hortonworks.

**HORTONWORKS DATAFLOW**

# HORTONWORKS DATAFLOW

Hortonworks DataFlow (HDF) proporciona la única plataforma de extremo a extremo que recolecta, cura, analiza y actúa sobre datos en tiempo real, en las instalaciones o en la nube, con una interfaz visual de arrastrar y soltar. HDF es una solución integrada con Apache Nifi/MiNifi, Apache Kafka, Apache Storm y Druid.

El HDF, plataforma de análisis de datos en tiempo real de retransmisión incluye sistemas de gestión de flujo de datos, procesamiento de retransmisión y servicios empresariales.

## Características

* Recopila y manipula los flujos de macrodatos de Internet de las cosas con seguridad y eficientemente al mismo tiempo que proporciona visibilidad operacional en tiempo real, gestión y control.
* Permite construir aplicaciones analíticas de retransmisión en minutos para capturar perspectivas perecederas en tiempo real sin necesidad de escribir una sola línea de código.
* Gestiona los ecosistemas HDF y HDP con panel de gestión integral de aprovisionamiento, seguimiento y gestión.

### PLATAFORMA DE RECOPILACIÓN AGNÓSTICA INDEPENDIENTE INTEGRADA DE FUENTE DE DATOS

HDF tiene capacidades de recolección de datos con funciones completas que están retransmitiendo datos agnósticos e integrados con más de 220 procesadores. Los macrodatos de Internet de las cosas se puede recopilar de fuentes dinámicas y distribuidas de diferentes formatos, esquemas, protocolos, velocidades y tamaños y de tipos como máquinas, dispositivos de geolocalización, flujos de clic, archivos, fuentes sociales, archivos de registro y vídeos.

### RESPUESTA EN TIEMPO REAL A LOS DATOS EN MOVIMIENTO

Con HDF, la recopilación de datos no es un proceso tedioso. Se puede manejar datos en pleno vuelo con un panel de control visual para ajustar fuentes, unir y dividir retransmisiones y priorizar el flujo de datos. HDF también puede agregar datos contextuales a sus flujos para un análisis y perspectiva más completa. Las rutas de procedencia y auditoría de datos proporcionan cumplimiento y resolución de problemas de [seguridad y gobernanza](https://es.hortonworks.com/solutions/security-and-governance/), según sea necesario en tiempo real. Integrado con Apache NiFi, MiNiFi, Kafka y Storm, HDF está listo para el procesamiento de eventos de alto volumen para análisis y acciones inmediatas. Kafka permite diferentes tasas de creación de datos y entrega mientras que Storm proporciona [análisis de datos en tiempo real de retransmisión](https://es.hortonworks.com/solutions/data-ingestion/) y perspectivas inmediatas a gran escala.

### PROTEGER POR COMPLETO DESDE ORIGEN HASTA EL DESTINO

Enrutamiento y flujo de datos seguros de extremo a extremo desde la fuente hasta el destino, con autorización de usuario discreta y cadena de custodia visual detallada y en tiempo real. Utilice el interfaz de usuario visual de HDF para cifrar datos, enviarlos a Kafka, configurar búferes y gestionar la congestión para que los datos puedan priorizarse dinámicamente y enviarse de forma segura. HDF permite el acceso de datos basado en funciones que permite que las empresas compartan de forma dinámica y segura porciones selectas de datos pertinentes. HDF fácilmente puede implementar gestión de flujos y aplicaciones de retransmisión en un entorno protegido por Kerberos sin demasiado cargo operacional.

### CREAR APLICACIONES DE ANÁLISIS DE RETRANSMISIÓN SIN CÓDIGO

HDF incluye un módulo completo de análisis de retransmisión, Streaming Analytics Manager (SAM), para crear aplicaciones analíticas de retransmisión que realizan correlación de eventos, enriquecimiento de contexto, concordancia con patrones compleja, incorporaciones analíticas y crear alertas/notificaciones cuando se descubren perspectivas. SAM hace que la construcción de analítica de retransmisión sea fácil para los desarrolladores de aplicaciones, DevOps y analistas de negocio para construir, desarrollar, colaborar, analizar, implementar y administrar aplicaciones en minutos sin escribir una sola línea de código. Los analistas usan gráficos preconstruidos para construir rápidamente el análisis y crear cuadros de mando, mientras que DevOps puede administrar y supervisar el rendimiento de aplicaciones de forma inmediata.

**HORTONWORKS DATA PLATFORM**

# HORTONWORKS DATA PLATFORM (HDP)

HDP es la única distribución de Apache™ Hadoop de código abierto y lista para empresas basada en una arquitectura centralizada ([YARN](https://es.hortonworks.com/apache/yarn/)). HDP responde a las necesidades completas de los datos en espera, impulsa las aplicaciones en tiempo real de los clientes y ofrece un análisis robusto que acelera la toma de decisiones y la innovación.

Hortonworks le ofrece la flexibilidad de ejecutar la misma plataforma de código abierto líder en la industria para obtener información de datos en el centro de datos, así como en la nube pública de su elección. Tanto si se trata de Microsoft Azure, Amazon Web Services o Google Cloud Platform, su organización puede aprovechar la agilidad de la nube y, al mismo tiempo, ejecutar sus cargas de trabajo locales para convertirse en una empresa híbrida basada en datos. Con nuestra arquitectura de datos moderna y lista para empresas, los clientes tendrán una experiencia consistente tanto en el centro de datos como en la nube.

## Características

**Abierto**

* Hortonworks está comprometido con un enfoque abierto al desarrollo de software que impulsa la innovación
* HDP permite que las empresas implementen, integren y trabajen con volúmenes sin precedentes de datos estructurados y no estructurados
* HDP ofrece software de clase empresarial que fomenta la innovación y evita el bloqueo de proveedor

**CENTRAL**

* HDP se basa en una arquitectura centralizada apoyada por YARN que asigna recursos entre varias aplicaciones
* YARN maximiza la ingestión de datos al permitir que las empresas analicen los datos para apoyar casos de uso diverso
* YARN coordina los servicios en todo el clúster para [operaciones, gobernanza de datos y seguridad](https://es.hortonworks.com/solutions/security-and-governance/)

**INTEROPERABLE**

* HDP es interoperable con un amplio ecosistema de proveedores de nubes y centro de datos
* HDP minimiza el gasto y el esfuerzo necesario para conectar al cliente de infraestructura con capacidades de procesamiento y datos de HDP
* Con HDP, los clientes pueden preservar su inversión en arquitectura TI existente a medida que adoptan Hadoop

**LISTO PARA LA EMPRESA**

* HDP proporciona monitoreo y administración de clúster centralizados de Hadoop
* Con HDP, la gestión de seguridad y datos está integrada en la plataforma
* HDP asegura que la seguridad se administra sistemáticamente a través de los