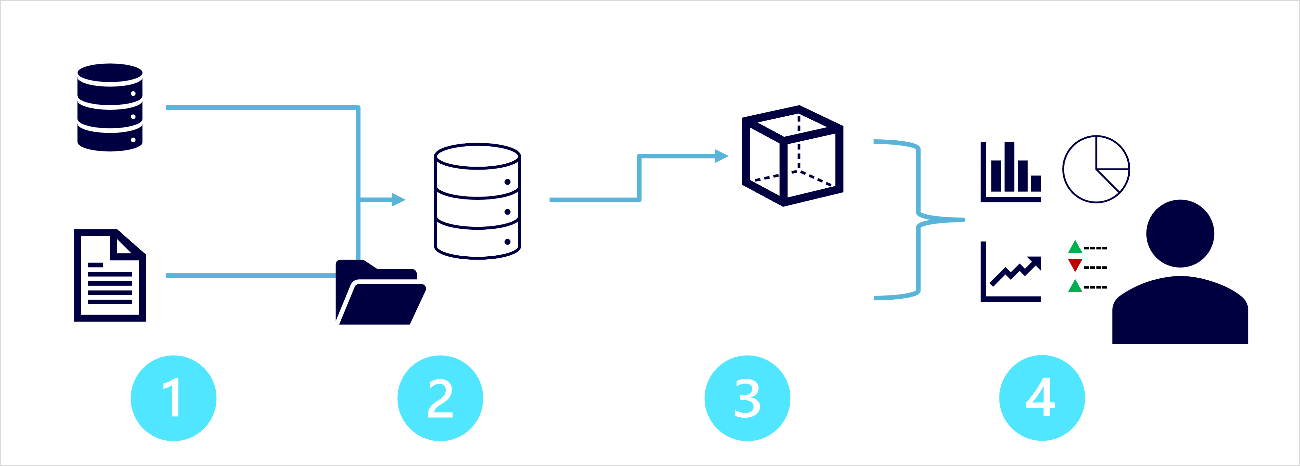
**Descripción del almacenamiento de datos moderno**

Completado100 XP

* 3 minutos

La arquitectura moderna de almacenamiento de datos puede variar, al igual que las tecnologías específicas que se usan para implementarla; de todos modos, en general, se incluyen los siguientes elementos:



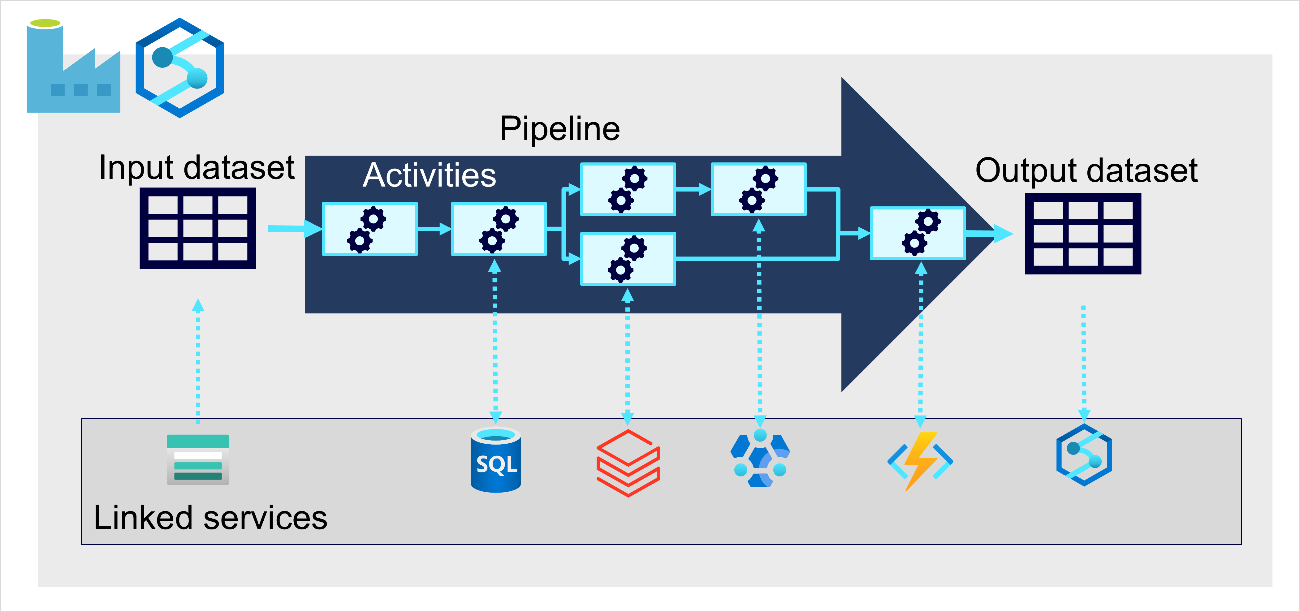
1. **Ingesta y procesamiento de datos**: los datos de uno o varios almacenes de datos transaccionales, archivos, flujos en tiempo real u otros orígenes se cargan en un lago de datos o en un almacenamiento de datos relacional. Normalmente, la operación de carga implica un proceso de *extracción, transformación y carga* (ETL) o de *extracción, carga y transformación* (ELT) en el que los datos se limpian, filtran y reestructuran para su análisis. En los procesos de ETL, los datos se transforman antes de cargarse en un almacén analítico, mientras que en un proceso de ELT los datos se copian en el almacén y, posteriormente, se transforman. En cualquier caso, la estructura de datos resultante está optimizada para las consultas analíticas. El procesamiento de datos suele realizarse mediante sistemas distribuidos que pueden procesar grandes volúmenes de datos en paralelo mediante clústeres de varios nodos. La ingesta de datos incluye el procesamiento por lotes de datos estáticos y el procesamiento en tiempo real de los datos de streaming.
2. **Almacén de datos analíticos**: los almacenes de datos para análisis a gran escala incluyen *almacenamientos de datos* relacionales, *lagos de datos* basados en sistema de archivos y arquitecturas híbridas que combinan características de almacenes de datos y lagos de datos (a veces bajo la denominación de *lagos de almacenamiento de datos* o *bases de datos de lago*). Los trataremos con más detalle más adelante.
3. **Modelo de datos analíticos**: aunque los analistas de datos y los científicos de datos pueden trabajar con los datos directamente en el almacén de datos analíticos, es habitual crear uno o varios modelos de datos que agreguen previamente los datos para facilitar la generación de informes, paneles y visualizaciones interactivas. A menudo, estos modelos de datos se describen como *cubos*, en los que los valores de datos numéricos se agregan en una o varias dimensiones (por ejemplo, para determinar las ventas totales por producto y región). El modelo encapsula las relaciones entre los valores de datos y las entidades dimensionales para admitir el análisis de tipo "rastrear agrupando datos/explorar en profundidad".
4. **Visualización de datos**: los analistas de datos consumen datos de modelos analíticos y directamente de almacenes analíticos para crear informes, paneles y otras visualizaciones. Además, los usuarios de una organización, que pueden no ser profesionales de la tecnología, pueden realizar informes y análisis de datos de autoservicio. Las visualizaciones de los datos muestran tendencias, comparaciones e indicadores clave de rendimiento (KPI) para una empresa u otra organización, y pueden tomar la forma de informes impresos, diagramas y gráficos en documentos o presentaciones de PowerPoint, paneles basados en web y entornos interactivos en los que los usuarios pueden explorar los datos visualmente.

**Exploración de canalizaciones de ingesta de datos**

Completado100 XP

* 5 minutos

Ahora que comprende un poco la arquitectura de una solución de almacenamiento de datos moderna y algunas de las tecnologías de procesamiento distribuido que se pueden usar para controlar grandes volúmenes de datos, es el momento de explorar cómo se ingieren los datos en un almacén de datos analíticos de uno o varios orígenes.



En Azure, la ingesta de datos a gran escala se implementa mejor mediante la creación de *canalizaciones* que organicen procesos de ETL. Puede crear y ejecutar canalizaciones mediante [Azure Data Factory](https://azure.microsoft.com/services/data-factory), o puede usar el mismo motor de canalización en [Azure Synapse Analytics](https://azure.microsoft.com/services/synapse-analytics) si quiere administrar todos los componentes de la solución de almacenamiento de datos en un área de trabajo unificada.

En cualquier caso, las canalizaciones constan de una o varias *actividades* que operan en los datos. Un conjunto de datos de entrada proporciona los datos de origen y las actividades se pueden definir como un flujo de datos que manipula incrementalmente los datos hasta que se genera un conjunto de datos de salida. Las canalizaciones utilizan *servicios vinculados* para cargar y procesar datos, y esto le permite usar la tecnología adecuada para cada paso del flujo de trabajo. Por ejemplo, puede usar un servicio vinculado de Azure Blob Store para ingerir el conjunto de datos de entrada y, posteriormente, usar servicios como Azure SQL Database para ejecutar un procedimiento almacenado que busque valores de datos relacionados, antes de ejecutar una tarea de procesamiento de datos en Azure Databricks o Azure HDInsight, o aplicar lógica personalizada mediante una función de Azure. Por último, puede guardar el conjunto de datos de salida en un servicio vinculado, como Azure Synapse Analytics. Las canalizaciones también pueden incluir algunas actividades integradas, que no requieren un servicio vinculado.

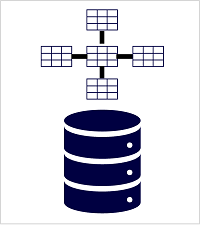
# Exploración de almacenes de datos analíticos

Completado100 XP

* 8 minutos

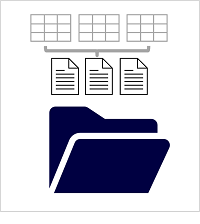
Hay dos tipos comunes de almacén de datos analíticos.

## Almacenamientos de datos



Un almacenamiento de datos es una base de datos relacional en la que los datos se almacenan en un esquema optimizado para el análisis de datos en lugar de en cargas de trabajo transaccionales. Normalmente, los datos de un almacén transaccional se organizan en un esquema en el que los valores numéricos se almacenan en tablas de hechos centrales, que están relacionadas con una o varias tablas de dimensiones que representan entidades por las que se pueden agregar los datos. Por ejemplo, una tabla de hechos podría contener datos de pedidos de ventas, que se pueden agregar por las dimensiones de cliente, producto, tienda y tiempo (lo que le permite, por ejemplo, encontrar fácilmente los ingresos totales mensuales de ventas por producto para cada tienda). Este tipo de esquema de tabla de hechos y dimensiones se denomina esquema de estrella; aunque a menudo se extiende a un esquema de copo de nieve mediante la adición de tablas adicionales relacionadas con las tablas de dimensiones para representar jerarquías dimensionales (por ejemplo, el producto puede estar relacionado con categorías de productos). Un almacenamiento de datos es una excelente opción si tiene datos transaccionales que se pueden organizar en un esquema estructurado de tablas y quiere usar SQL para consultarlos.

## Lagos de datos



Un lago de datos es un almacén de archivos, normalmente en un sistema de archivos distribuido para el acceso a datos de alto rendimiento. A menudo se usan tecnologías como Spark o Hadoop para procesar consultas en los archivos almacenados y devolver datos para informes y análisis. Estos sistemas suelen aplicar un enfoque de esquema en lectura para definir esquemas tabulares en archivos de datos semiestructurados en el punto donde se leen los datos para su análisis, sin aplicar restricciones cuando se almacenan. Los lagos de datos son excelentes para admitir una combinación de datos estructurados, semiestructurados e incluso no estructurados que quiere analizar sin necesidad de aplicar el esquema cuando los datos se escriben en el almacén.

### Enfoques híbridos

Puede usar un enfoque híbrido que combine características de lagos de datos y almacenamientos de datos en una base de datos de lago o un lago de almacenamiento de datos. Los datos sin procesar se almacenan como archivos en un lago de datos y una capa de almacenamiento relacional abstrae los archivos subyacentes y los expone como tablas, que se pueden consultar mediante SQL. Los grupos de SQL de Azure Synapse Analytics incluyen PolyBase, que permite definir tablas externas basadas en archivos de un lago de datos (y otros orígenes) y consultarlas mediante SQL. Synapse Analytics también admite un enfoque de base de datos de lago en el que puede usar plantillas de base de datos para definir el esquema relacional del almacenamiento de datos, al tiempo que almacena los datos subyacentes en un almacenamiento de lago de datos, separando el almacenamiento y el proceso de la solución de almacenamiento de datos. Los lagos de almacenamiento de datos son un enfoque relativamente nuevo en los sistemas basados en Spark y se habilitan mediante tecnologías como Delta Lake, que agrega funcionalidades de almacenamiento relacional a Spark, por lo que se pueden definir tablas que exijan esquemas y coherencia transaccional, admitan orígenes de datos de streaming y cargados por lotes y proporcionen una API de SQL para realizar consultas.

## Servicios de Azure para almacenes analíticos

En Azure, hay tres servicios principales que puede usar para implementar un almacén analítico a gran escala

[**Azure Synapse Analytics**](https://azure.microsoft.com/services/synapse-analytics) es una solución de un extremo a otro unificada para el análisis de datos a gran escala. Reúne varias tecnologías y funcionalidades, y esto permite combinar la integridad y la confiabilidad de los datos de un almacenamiento de datos relacional basado en SQL Server escalable y de alto rendimiento con la flexibilidad de una solución Apache Spark de código abierto y lago de datos. También incluye compatibilidad nativa para el análisis de registros y telemetría con grupos del Explorador de datos de Azure Synapse, así como canalizaciones de datos integradas para la ingesta y la transformación de datos. Todos los servicios de Azure Synapse Analytics se pueden administrar a través de una única interfaz de usuario interactiva denominada Azure Synapse Studio, que incluye la capacidad de crear cuadernos interactivos en los que se pueden combinar código de Spark y contenido de Markdown. Synapse Analytics es una excelente opción cuando se quiere crear una única solución de análisis unificada en Azure.

[**Azure Databricks**](https://azure.microsoft.com/services/databricks) es una implementación de Azure de la popular plataforma Databricks. Databricks es una completa solución de análisis de datos integrada en Apache Spark y ofrece funcionalidades nativas de SQL, así como clústeres de Spark optimizados para cargas de trabajo para el análisis de datos y la ciencia de datos. Databricks proporciona una interfaz de usuario interactiva a través de la cual se puede administrar el sistema y se pueden explorar los datos en cuadernos interactivos. Debido a su uso común en varias plataformas en la nube, es posible que quiera considerar el uso de Azure Databricks como almacén analítico si quiere usar la experiencia existente con la plataforma o si necesita operar en un entorno de varias nubes o admitir una solución portátil en la nube.

[**Azure HDInsight**](https://azure.microsoft.com/services/hdinsight) es un servicio de Azure que admite varios tipos de clústeres de análisis de datos de código abierto. Aunque no es tan fácil de usar como Azure Synapse Analytics y Azure Databricks, puede ser una opción adecuada si la solución de análisis se basa en varios marcos de código abierto o si necesita migrar una solución local existente basada en Hadoop a la nube.

**Nota**

Cada uno de estos servicios puede considerarse como un almacén de datos analíticos, en el sentido de que proporcionan un esquema y una interfaz a través de los cuales se pueden consultar los datos. Sin embargo, en muchos casos, los datos se almacenan realmente en un lago de datos y el servicio se usa para procesar los datos y ejecutar consultas. Algunas soluciones pueden incluso combinar el uso de estos servicios. Un proceso de ingesta de extracción, carga y transformación (ELT) puede copiar datos en el lago de datos y, posteriormente, usar uno de estos servicios para transformar los datos y otro para consultarlos. Por ejemplo, una canalización puede usar un trabajo de MapReduce que se ejecuta en HDInsight o un cuaderno que se ejecuta en Azure Databricks para procesar un gran volumen de datos en el lago de datos y, posteriormente, cargarlo en tablas de un grupo de SQL en Azure Synapse Analytics.