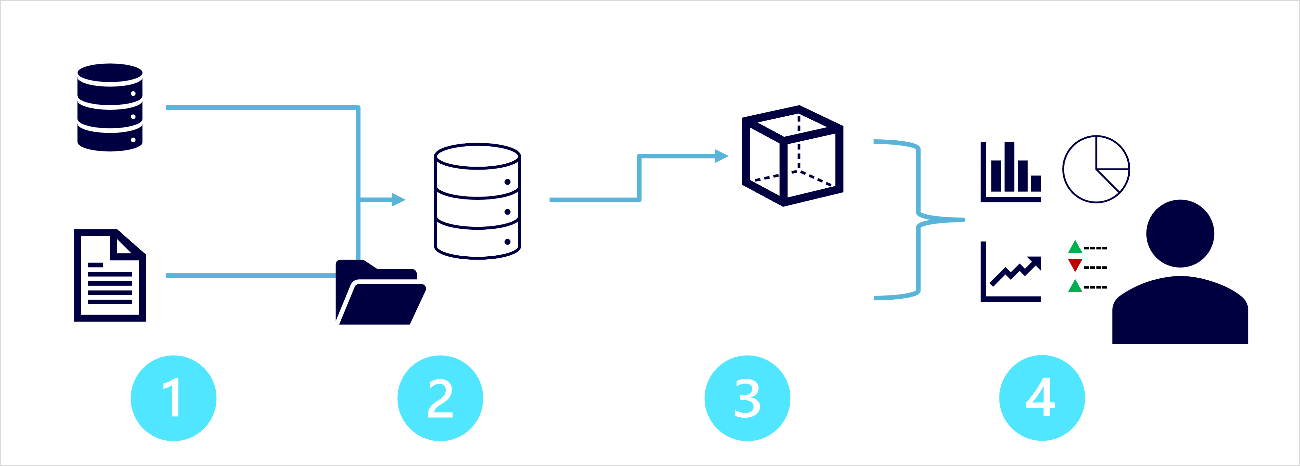
**Descripción del almacenamiento de datos moderno**

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:



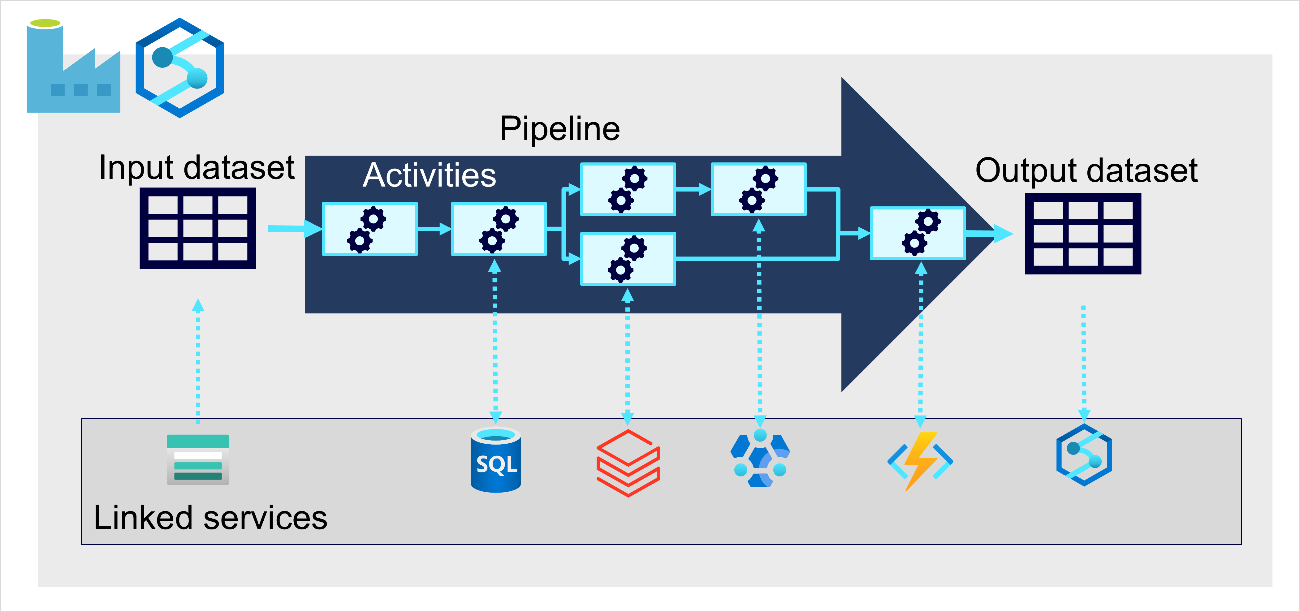
1. **Ingesta y procesamiento de datos**: los datos de uno o varios almacenes de datos transaccionales, archivos, flujos en tiempo real u otros orígenes se cargan en un lago de datos o en un almacenamiento de datos relacional. Normalmente, la operación de carga implica un proceso de *extracción, transformación 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.