Enterprise data stores for data ingestion

Large data stores are the norm in large enterprises. The concept of a data lake reflects this reality. Data lakes are very large collections of data stored in their natural formats, usually as object blobs or files. Today's data scientist must be proficient in building data pipelines that tap directly into such large collections of raw data, then process the data to gain insights.

Along with data lakes, technologies such as Apache Hadoop enable large enterprises to store very large amounts of data, and to access the data quickly for analysis. Hadoop has two advantages that make it useful in large enterprises. First, it is designed from the ground up to be fault tolerant. A Hadoop cluster runs on an array of individual commodity servers designed to cleanly fail over without loss of data or processing power. Second, Hadoop clusters allow for parallel execution of data analysis code against the blocks of data stored in the cluster. This enables the rapid execution of complex analyses against huge amounts of data.

While many data ingestion pipelines draw data directly from sources such as data lakes and Hadoop clusters, data scientists in large enterprises will sometimes work with data engineers to build a *data warehouse*. A data warehouse keeps data gathered and integrated from different sources (e.g., a data lake) and stores the large number of records needed for long-term usage by machine learning systems. A data warehouse is typically built using data extractions, data transformations and data loads. After selecting data from the sources of origin, data ingestion procedures resolve problems in the data and ready it for research and modeling.

Modern large enterprises have adopted sophisticated data management processes and systems to handle very large amounts of data. With large datasets and complex use cases, data ingestion involves the ability to use data from a wide variety of sources, mixing and matching those sources to create data pipelines that feed machine learning models.