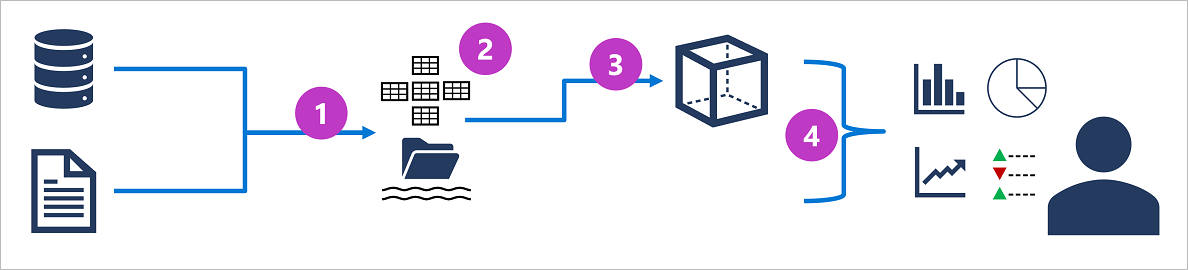
**Explore analytical data processing**

Analytical data processing typically uses read-only (or read-*mostly*) systems that store vast volumes of historical data or business metrics. Analytics can be based on a snapshot of the data at a given point in time, or a series of snapshots.

The specific details for an analytical processing system can vary between solutions, but a common architecture for enterprise-scale analytics looks like this:



1. Operational data is extracted, transformed, and loaded (ETL) into a data lake for analysis.
2. Data is loaded into a schema of tables- typically in a Spark-based *data lakehouse* with tabular abstractions over files in the data lake, or a *data warehouse* with a fully relational SQL engine.
3. Data in the data warehouse may be aggregated and loaded into an online analytical processing (OLAP) model, or *cube*. Aggregated numeric values (*measures*) from fact tables are calculated for intersections of *dimensions* from dimension tables. For example, sales revenue might be totaled by date, customer, and product.
4. The data in the data lake, data warehouse, and analytical model can be queried to produce reports, visualizations, and dashboards.

*Data lakes* are common in large-scale data analytical processing scenarios, where a large volume of file-based data must be collected and analyzed.

*Data warehouses* are an established way to store data in a relational schema that is optimized for read operations – primarily queries to support reporting and data visualization. *Data Lakehouses* are a more recent innovation that combine the flexible and scalable storage of a data lake with the relational querying semantics of a data warehouse. The table schema may require some denormalization of data in an OLTP data source (introducing some duplication to make queries perform faster).

An OLAP model is an aggregated type of data storage that is optimized for analytical workloads. Data aggregations are across dimensions at different levels, enabling you to *drill up/down* to view aggregations at multiple hierarchical levels; for example to find total sales by region, by city, or for an individual address. Because OLAP data is pre-aggregated, queries to return the summaries it contains can be run quickly.

Different types of user might perform data analytical work at different stages of the overall architecture. For example:

* Data scientists might work directly with data files in a data lake to explore and model data.
* Data Analysts might query tables directly in the data warehouse to produce complex reports and visualizations.
* Business users might consume pre-aggregated data in an analytical model in the form of reports or dashboards.

**Khám phá Xử lý Dữ liệu Phân tích (Analytical data processing)**

Xử lý dữ liệu phân tích thường sử dụng các hệ thống chỉ đọc (hoặc chủ yếu đọc) lưu trữ khối lượng lớn dữ liệu lịch sử hoặc số liệu kinh doanh. Phân tích có thể dựa trên ảnh chụp nhanh của dữ liệu tại một thời điểm nhất định hoặc một loạt ảnh chụp nhanh.

Các chi tiết cụ thể cho một hệ thống xử lý phân tích có thể thay đổi giữa các giải pháp, nhưng kiến trúc chung cho phân tích quy mô doanh nghiệp trông giống như thế này:

1.   Dữ liệu hoạt động được trích xuất, chuyển đổi và tải (ETL) vào hồ dữ liệu để phân tích. 2.   Dữ liệu được nạp vào lược đồ bảng - thường nằm trong kho dữ liệu dạng Spark với các khái trừ dạng bảng trên các file trong hồ dữ liệu hoặc kho dữ liệu với công cụ SQL quan hệ đầy đủ. 3.   Dữ liệu trong kho dữ liệu có thể được tổng hợp và nạp vào mô hình xử lý phân tích trực tuyến (OLAP) hoặc khối (cube). Các giá trị số được tổng hợp (đo lường) từ các bảng dữ liệu thực tế được tính toán cho các giao điểm của các chiều từ các bảng chiều. Ví dụ, doanh thu bán hàng có thể được tổng theo ngày, khách hàng và sản phẩm. 4.   Dữ liệu trong hồ dữ liệu, kho dữ liệu và mô hình phân tích có thể được truy vấn để tạo báo cáo, hình ảnh và bảng điều khiển.

**Hồ dữ liệu (Data lakes):** Thông dụng trong các tình huống xử lý phân tích dữ liệu quy mô lớn, nơi cần thu thập và phân tích khối lượng lớn dữ liệu dựa trên file.

**Kho dữ liệu (Data warehouses):** Là một cách thiết lập để lưu trữ dữ liệu trong lược đồ quan hệ được tối ưu hóa cho các hoạt động đọc - chủ yếu là các truy vấn để hỗ trợ báo cáo và trực quan hóa dữ liệu. **Kho dữ liệu dạng Lakehouse (Data Lakehouses):** Là một đổi mới gần đây hơn, kết hợp lưu trữ linh hoạt và có thể mở rộng của hồ dữ liệu với ngữ nghĩa truy vấn quan hệ của kho dữ liệu. Sơ đồ bảng có thể yêu cầu một số chuẩn hóa dữ liệu trong nguồn dữ liệu OLTP (giới thiệu một số bản sao để thực hiện truy vấn nhanh hơn).

**Mô hình OLAP:** Là một kiểu lưu trữ dữ liệu được tổng hợp, được tối ưu hóa cho khối lượng công việc phân tích. Dữ liệu được tổng hợp trên các chiều ở các cấp độ khác nhau, cho phép bạn khoan lên/xuống để xem các tổng hợp ở nhiều cấp bậc theo thứ bậc; ví dụ để tìm tổng doanh số theo vùng, theo thành phố hoặc theo một địa chỉ riêng lẻ. Vì dữ liệu OLAP được tổng hợp sẵn nên các truy vấn để trả về các bản tóm tắt có trong đó có thể chạy nhanh chóng.