Redshift and AWS Lake Formation

Finally, though technically a cloud storage solution, I do find it necessary to at least mention data warehouses and data lakes in the storage section as they are often utilized to store and manage large volumes of data (usually data that is miscellaneous in nature). A purpose that keen readers may recall was mentioned as the defining characteristic of storage in the background section of the chapter.

Now, data warehouses are designed to assist with business intelligence and data analytics tasks and usually is responsible for storing data, performing simple quantitative analyses on the data and visualizing data. Amazon's data warehousing solution is called **Redshift** (discussed already in the OLAP section), and focuses on operating more so as a business and data analysis tool than as a service for the storage of data. Redshift is built around a cluster of computing nodes which sit on top of Redshift's data store allowing us to prepare, perform simple arithmetic on and visualize quantitative data at the petabyte-scale.

Data lakes on the other hand, allow an organization to store virtually any type of data at an almost unlimited scale. They are used when storing data pertaining to a wide range of services and belonging to a diverse set of formats (such as databases, BI tools and multimedia files) all in a single location. They are highly adaptable and can change with ease to meet new requirements. **AWS Lake Formation** is Amazon's managed service for the creation and deployment of data lakes, and is usually utilized when organizations using AWS require a repository to hold vast amounts of both processed and unprocessed data from a myriad of unrelated sources, which companies sometimes have to do, in order to comply with regulatory standards and policies.