16.1.2

Big Data Technologies

You and Jennifer know that this project has outgrown Excel, SQL, and other databases you have used before. This means that you get to explore brand-new tools and think about data in a whole new way. To begin, you decide to explore the technologies that support the big data ecosystem. This way you can decide what technologies you will need to answer SellBy's questions.

Apache Hadoop (Hadoop) is one of the most popular open source frameworks, with numerous technologies for big data. Google developed Hadoop to process large amounts of data by splitting data across a distributed file system.

We'll start with the three main components of Hadoop:

- **Hadoop Distributed File System (HDFS)** is a file system used to **store data** across server clusters (groups of computers). It is scalable (which means it handles influxes of data), fault-tolerant (handles hardware failure), and distributed (spread across multiple servers connected by a common core).
- **MapReduce** is a programming model and processing technique for big data. MapReduce enables processing the large amount of data spread across the cluster in the HDFS by performing the same task for each file system.
- **Yet Another Resource Negotiator (YARN)** manages and allocates resources across the clusters and assigns tasks.

Hadoop distributes for the storage and processing of data through a cluster, which is a group of connected computers that work together to store and perform tasks on a dataset.

Hadoop is quite difficult to set up. You need to set up all three main components across multiple machines, as well as make sure each one has sufficient resources and is configured for optimal performance. Because of this, it may not be the right technology for your startup. However, you know your client will ask about it—"Hadoop" is a popular buzzword, after all—so it's important to have a baseline knowledge of it.

NOTE

Visit the **Hadoop official website** (https://hadoop.apache.org/) to see other projects offered by Hadoop.