Based on your reading of chapter 1 of Learning Spark 2nd edition, why is Apache Spark popular and what advantage does it have over Hadoop?

- Apache Spark is a distributed processing framework for big data workloads that is open-source. For quick analytic queries against any size of data, it uses in-memory caching and optimized query execution. Spark is popular because it is faster than other big data technologies, with the ability to execute several jobs parallelly to better fit Spark's in-memory architecture. In-memory processing in Sparks saves a lot of time and makes things easier and more efficient. Spark is developed concentrating on these four major characteristics mainly: Speed, Ease of use, Modularity, Extensibility.These four primary characteristics make Apache Spark so popular and widely used application for data processing in the today;s Big Data era. Many IT companies have jumped on board to support Spark, seeing an opportunity to expand their existing big data solutions into areas where Spark adds actual value, such as interactive querying and machine learning. Well-known organizations like IBM and Huawei have made large investments in the technology, and a rising number of startups are constructing businesses that rely entirely or partially on Spark.

  The framework of Spark makes it efficient or processing big data and being an open source application it is easily accessible for its user. Its distributed framework allows Apache Spark to function at higher capabilities than any other engine. Moreover, Apache Spark Facilitates in the coherence and simplicity of complex data pipelines.
- Spark may be used to implement machine learning tasks such as Nave Bayes and K-means computations, saving time and money. To work with huge datasets, Spark includes APIs and packages for graph processing, streaming, and machine learning. The programming languages Scala, Java, Python, SQL, and R can all be used with Spark.
- Explores and visualizes data sets via ad hoc or interactive queries. Additionally, utilizing MLlib, building, training, and assessing machine learning models, as well as implementing end-to-end data pipelines from a variety of data streams. Apache Sparks additionally improves and optimizes graph data sets and social network analysis. GraphX is a library for manipulating and performing graph-parallel computations on graphs (e.g., social network work graphs, routes and connection points, or network topology graphs).

Advantages of Apache Spark over Hadoop:

- Hadoop processes data using MapReduce, whereas Spark makes use of robust distributed datasets (RDDs).Hadoop uses a distributed file system (HDFS), which allows data to be stored on several machines. Because servers and machines may be added to accommodate increasing data quantities, the file system is scalable. Because Spark lacks a distributed file storage system, it is mostly utilized for computation on top of Hadoop. Spark does not require Hadoop to run, although it can be used alongside it because it can construct distributed datasets from HDFS files.
- Hadoop is wonderful for batch processing, but iterative processing is wasteful, hence Spark was built to address this. Spark programs run 100 times quicker in memory and 10 times faster on disk than Hadoop programs. Spark's performance is due to its in-memory processing. Instead, Hadoop MapReduce writes data to a disk, which is then read in the next cycle. It is substantially slower than Spark because data is reloaded from disk after each iteration.
- The Spark Streaming framework is designed to handle the velocity of large data by stream processing fault-tolerant, live data streams. Discretized Streams are a type of Spark Streaming data. DStreams are fault-tolerant and consistent. Hadoop's Map and Reduce and MapReduce tasks, on the other hand, take a long time to complete, raising latency. In MapReduce, data is disseminated and processed over a cluster, which increases processing time and slows it down.
- With fewer languages available, Hadoop is more difficult to utilize. MapReduce programs are written in Java or Python. When compared to Apache Spark, Hadoop is less flexible. Apache Spark is easier to utilize. Allows you to use the interactive shell mode. APIs can be created in Java, Scala, R, Python, and Spark SQL, among other languages. Hadoop is also slower than Spark. Bottlenecks can occur when data fragments are too big. The main library is Mahout, although Apache Spark is substantially quicker with in-memory processing and computations utilizing MLlib.

References:

Damji, J. S., Wenig, B., Das, T., & Lee, D. (2020). *Learning Spark: Lightning-Fast Data Analytics* (2nd ed.). O'Reilly Media.