Spark Use Cases and Resources

Here are a few resources about different Spark use cases:

- **Data Analytics**
- **Machine Learning**
- **Streaming**
- **Graph Analytics**

You Don't Always Need Spark

Spark is meant for big data sets that cannot fit on one computer. But you don't need Spark if you are working on smaller data sets. In the cases of data sets that can fit on your local computer, there are many other options out there you can use to manipulate data such as:

- **AWK** - a command line tool for manipulating text files
- **R** - a programming language and software environment for statistical computing
- **Python PyData Stack**, which includes pandas, matplotlib, numpy, and scikit-learn among other libraries

Sometimes, you can still use pandas on a single, local machine even if your data set is only a little bit larger than memory. Pandas can read data in chunks. Depending on your use case, you can filter the data and write out the relevant parts to disk.

If the data is already stored in a relational database such as **MySQL** or **Postgres**, you can leverage SQL to extract, filter and aggregate the data. If you would like to leverage pandas and SQL simultaneously, you can use libraries such as **SQLAlchemy**, which provides an abstraction layer to manipulate SQL tables with generative Python expressions.

The most commonly used Python Machine Learning library is **scikit-learn**. It has a wide range of algorithms for classification, regression, and clustering, as well as utilities for preprocessing data, fine tuning model parameters and testing their results. However, if you want to use more complex algorithms - like deep learning - you'll need to look further. **TensorFlow** or **PyTorch**.

Spark's Limitations

Spark has some limitation.

Spark Streaming’s latency is at least 500 milliseconds since it operates on micro-batches of records, instead of processing one record at a time. Native streaming tools such as **Storm**, **Apex**, or **Flink** can push down this latency value and might be more suitable for low-latency applications. Flink and Apex can be used for batch computation as well, so if you're already using them for stream processing, there's no need to add Spark to your stack of technologies.

Another limitation of Spark is its selection of machine learning algorithms. Currently, Spark only supports algorithms that scale linearly with the input data size. In general, deep learning is not available either, though there are many projects integrate Spark with Tensorflow and other deep learning tools.

Hadoop versus Spark

The Hadoop ecosystem is a slightly older technology than the Spark ecosystem. In general, Hadoop MapReduce is slower than Spark because Hadoop writes data out to disk during intermediate steps. However, many big companies, such as Facebook and LinkedIn, started using Big Data early and built their infrastructure around the Hadoop ecosystem.

While Spark is great for iterative algorithms, there is not much of a performance boost over Hadoop MapReduce when doing simple counting. Migrating legacy code to Spark, especially on hundreds of nodes that are already in production, might not be worth the cost for the small performance boost.

Beyond Spark for Storing and Processing Big Data

Keep in mind that Spark is not a data storage system, and there are a number of tools besides Spark that can be used to process and analyze large datasets.

Sometimes it makes sense to use the power and simplicity of SQL on big data. For these cases, a new class of databases, know as NoSQL and NewSQL, have been developed.

For example, you might hear about newer database storage systems like **HBase** or **Cassandra**. There are also distributed SQL engines like **Impala** and **Presto**. Many of these technologies use query syntax that you are likely already familiar with based on your experiences with Python and SQL.

In the lessons ahead, you will learn about Spark specifically, but know that many of the skills you already have with SQL, Python, and soon enough, Spark, will also be useful if you end up needing to learn any of these additional Big Data tools.