Spark IN ACTION

Petar Zečević
Marko Bonaći

MEAP

MANNING

MEAP Edition
Manning Early Access Program
Spark in Action
Version 7

For more information on this and other Manning titles go to
www.manning.com

welcome

Dear reader,

Thank you for purchasing *Spark in Action* during MEAP. Writing this book has been a great experience for us and we are excited that the book has reached this stage. Although we are releasing only the first two chapters, we hope you will find the book useful right away. We also hope that you will take advantage of the Author Online forum to help us make it better.

We make minimum assumptions about your skill level. You should have some programming experience, if not with Scala, then preferably with Java. All the examples are in Scala, but Scala knowledge is not necessary because we will be explaining the main building blocks of the Scala language as we come across them. The corresponding Python and Java examples will be available on the book's GitHub repository. Some experience with distributed systems (namely Hadoop) is preferable (though not necessary) because that way you will be able to grasp Spark's architecture more quickly.

Apache Spark is a platform applicable to an increasing number of problems in the Big Data world. So before building real-life applications with Apache Spark, a lot of ground has to be covered first. We hope that the book will be a well-balanced mix of "theory" and practice. We are trying to make it an "in action" book as much as possible.

The book is divided into four parts.

- In Part 1, you can find a gentle introduction to programming with Apache Spark, and a more detailed overview of its core API.
- In Part 2, we will describe Spark's components: SQL, Streaming, GraphX, and MLlib.
- In Part 3, we will describe different Spark runtime options, as well as ways to manage and monitor its execution.
- In Part 4, you will build two real-world Spark applications.

We will be reading and responding to your comments on the Author Online forum. Feel free to give us your opinion about the released chapters and what you expect to see in the following ones. Help us write a better book!

Regards,
—Marko Bonaći and Petar Zečević

PS: Even though English is not our native language, have no fear! Before publication, all files will be edited and any language issues corrected.

brief contents

1

Introduction to Apache Spark

This chapter covers

- What Spark brings to the table
- Spark components
- Spark program flow
- Spark ecosystem
- Overview of this book

You probably already know that Apache Spark is usually defined as a fast, general purpose distributed computing platform. Sounds a bit like marketing speak on the first glance, but we could hardly come up with a more appropriate label to put on the Spark box.

Apache Spark really did bring a revolution to the Big Data space. You have probably heard how Spark makes efficient use of memory and that it is able to execute equivalent jobs 10 to 100 times faster than Hadoop's MapReduce. On top of that, Spark's creators managed to abstract away the fact that you are dealing with a cluster of machines, and instead presented you with a set of collections-based APIs. Working with Spark's collections feels like working with local Scala, Java, or Python collections, but Spark's collections actually reference data distributed on many nodes. And operations on these collections get translated to complicated parallel programs without the user being necessarily aware of the fact, which is a truly powerful concept.

In this chapter, we first shed some light on the main Spark features and compare Spark to its natural predecessor: Hadoop's execution engine, MapReduce. Then we give you a brief exposé of Spark's components and show you how a typical Spark program executes using a simple "hello world" example. Finally, we will briefly explore Hadoop's ecosystem—a collection

of tools and languages used together with Hadoop for Big Data operations—to see how Spark fits in.

1.1 What is Spark?

Apache Spark is an exciting new technology that is rapidly superseding Hadoop's MapReduce as the preferred Big Data processing platform. Hadoop is an open-source, distributed, Java computation framework consisting of HDFS, Hadoop's distributed file system, and MapReduce, its execution engine. Spark is a fast and distributed general purpose computing platform. Spark's unique design, which allows for keeping large amounts of data in memory, offers tremendous performance improvements. Spark programs can be 100 times faster than their MapReduce counterparts.

Using Spark's elegant API and runtime architecture, you can write distributed programs in a manner similar to writing the local ones. As we already said, Spark's collections abstract away the fact that they are potentially referencing data distributed on a great number of nodes. Spark also allows you to use functional programming methods, which are a great match for data processing tasks.

By supporting Python, Java, Scala, and, most recently, R, Spark is open to a wide range of users: to the science community that traditionally favors Python and R, to the still-widespread Java community, and to people using the increasingly popular Scala, which offers functional programming on the Java Virtual Machine (JVM).

Finally, Spark combines MapReduce-like capabilities for batch programming, real-time data-processing functions, SQL-like handling of structured data, graph algorithms, and machine learning, all in a single framework. This makes it a one-stop-shop for most of your Big Data-crunching needs. It’s no wonder, then, that Spark is one of the busiest and fastest-growing Apache Software Foundation projects today.

But some applications are not appropriate for Spark. Because of its distributed architecture, Spark necessarily brings some overhead to the processing time. This overhead is negligible when handling large amounts of data, but if you have a dataset that can be handled by a single machine (which is becoming ever more likely these days), it might be more efficient to use some other framework optimized for that kind of computation. Also, Spark was not made with online transaction processing (OLTP) applications in mind: fast and numerous atomic transactions. It is better suited for online analytical processing (OLAP): batch jobs and data mining.

But because you’re reading this book, you probably already know all that. You came to learn how to use the darn thing, and we hope this book will help you in that goal. We have done our best to write a comprehensive guide to Spark architecture, its components, its runtime environment, and its API, while providing concrete examples and real-life case studies. By reading it, and more importantly, by sifting the examples through your fingers, you’ll gain the knowledge and skills necessary for writing your own high-quality Spark programs and managing Spark applications.

1.2 Spark components

As we said previously, Spark simplifies the picture by providing many Hadoop ecosystem functions through several purpose-built components. These are Spark Core, Spark SQL, Spark Streaming, Spark GraphX, and Spark MLlib—depicted in figure 1.1.

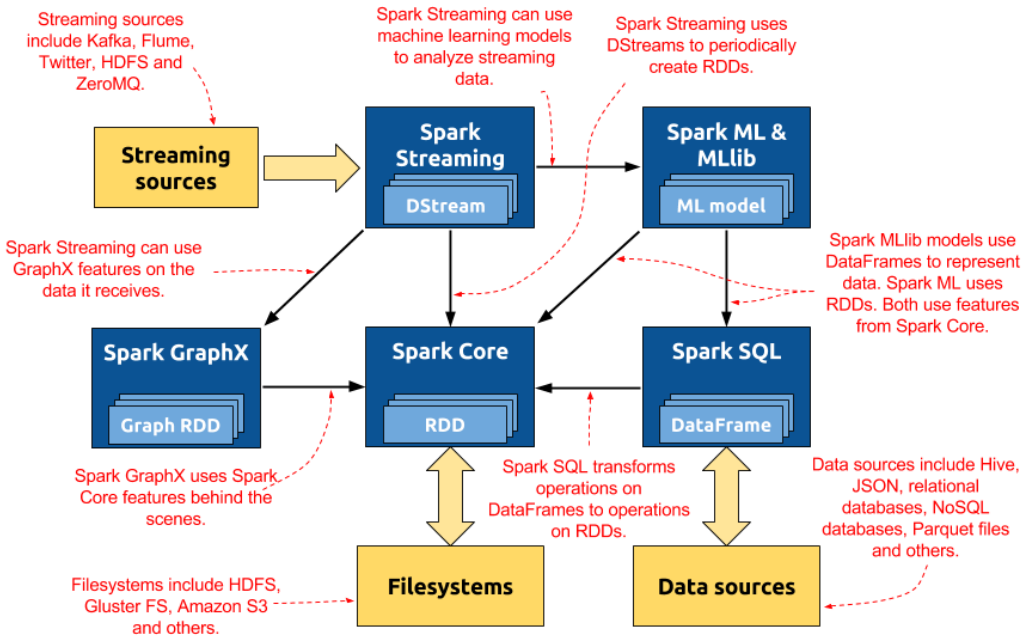

Figure 1.1: Main Spark components and various runtime environments and storage options

A brief description of each Spark component follows.

1.2.1 Spark Core

Spark Core contains basic Spark functionalities required for running jobs and needed by other components. The most important of these is the RDD, or *resilient distributed dataset*[1], which is the main element of the Spark API. It's an abstraction of a *distributed* collection of items with operations and transformations applicable to the dataset. It is *resilient* because it is capable of rebuilding datasets in case of node failures.

[1] RDDs are explained in chapter 2. As they're the fundamental abstraction of Spark, they're also covered in detail in chapter 4.

Spark Core contains logic for accessing various file systems, such as HDFS, Gluster FS, Amazon S3 and so on. It also provides a means of information sharing between computing nodes with broadcast variables and accumulators. Other fundamental functions, such as networking, security, scheduling, and data shuffling, are also part of the Spark Core.

1.2.2 Spark SQL

Spark SQL is the newest Spark component, but actively developed. It provides functions for manipulating large sets of distributed, structured data using an SQL subset supported by Spark and Hive SQL (HQL). With DataFrames introduced in Spark 1.3, which simplified handling of structured data and enabled radical performance optimizations, Spark SQL became one of the most important Spark components. Spark SQL can also be used for reading and writing data to and from various structured formats and data sources, such as JavaScript Object Notation (JSON) files, Parquet files (an increasingly popular file format that allows for storing schema along with the data), relational databases and Hive.

Operations on DataFrames at some point translate to operations on RDDs and execute as ordinary Spark jobs.

Spark SQL provides a query optimization framework called Catalyst that can be extended by custom optimization rules. Spark SQL also includes a Thrift server, which can be used by external systems, such as business intelligence tools, to query data through Spark SQL using classic JDBC and ODBC protocols.

1.2.3 Spark Streaming

Spark Streaming is a framework for ingesting real-time streaming data from various sources. The supported streaming sources include HDFS, Kafka, Flume, Twitter, ZeroMQ, and custom ones. Its operations recover from failure automatically, which is important for online data processing.

Spark Streaming represents streaming data using *discretized streams* (or DStreams), which are capable of periodically creating RDDs with the data that came in during the last time window.

Spark Streaming can be combined with other Spark components in a single program, unifying real-time processing with machine learning, SQL, and graph operations, which is something unique in the Hadoop ecosystem.

1.2.4 Spark MLlib

Spark MLlib is a library of machine-learning algorithms grown from the MLbase project at UC Berkeley. Supported algorithms include logistic regression, naive Bayes classification, support vector machines (SVM), decision trees, random forests, linear regression, k-means clustering, and others.

Apache Mahout is an existing project offering implementations of distributed machine-learning algorithms running on Hadoop. Although Apache Mahout is more mature, Spark MLlib

and Mahout both include a similar set of machine-learning algorithms. But with Mahout migrating from MapReduce to Spark, they are bound to be merged in the future.

Spark MLlib handles machine learning models used for transforming datasets, which are represented as RDDs or DataFrames.

1.2.5 Spark GraphX

Graphs are data structures comprising vertices and the edges connecting them. GraphX provides functions for building graphs, represented as "graph RDDs": EdgeRDD and VertexRDD. GraphX contains implementations of the most important algorithms of graph theory, such as page rank, connected components, shortest paths, SVD++, and others. It also provides the Pregel message-passing API, the same API for large-scale graph processing implemented by Apache Giraph, a project with implementations of graph algorithms and running on Hadoop.

1.3 Spark program flow

Let's see what a typical Spark program looks like. Imagine that a 300 MB log file is stored in a three-node HDFS cluster. HDFS automatically splits the file into 128 MB parts (*blocks*, in Hadoop terminology) and places each part on a separate node of the cluster [2] (figure 1.2). Let us assume Spark is running on YARN, inside the same Hadoop cluster.

Figure 1.2: Storing a 300 MB log file in a three-node Hadoop cluster

[2] Though it's not relevant to our example, we should probably mention that HDFS replicates each block to two additional nodes (if the default replication factor of 3 is in effect).

A Spark data engineer is given the task of analyzing how many errors of type `OutOfMemoryError` have happened during the last two weeks. She knows that the log file contains the last two weeks of logs of her company's application server cluster. She sits down at her laptop, puts on a bandana, cracks her fingers a bit, and starts to work.

She first starts her *Spark shell*, and through it, establishes a connection to the Spark cluster. Next, she loads the log file from HDFS (figure 1.3) by using this (Scala) line:

```
val lines = sc.textFile("hdfs://path/to/the/file")
```

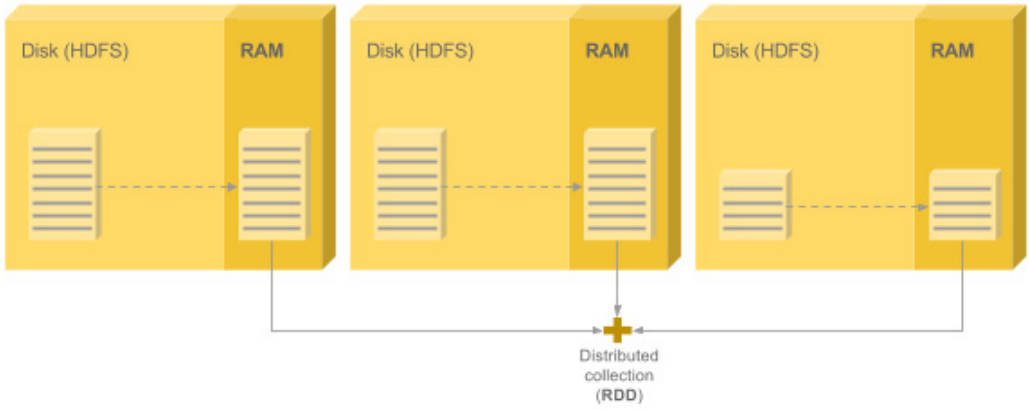

Figure 1.3: Loading a text file from HDFS

To achieve maximum *data locality,*[3] the loading operation asks Hadoop for locations of each block of the log file, and then transfers all the blocks into RAM of the cluster's nodes.

Now Spark has a reference to each of those blocks (*partitions*, in Spark terminology) in RAM. The sum of those partitions is a distributed collection of lines from the log file referenced by a resilient distributed dataset (RDD). Simplifying a bit, we can say that RDDs allow you to work with a distributed collection the same way you would work with any local, non-distributed one. You don't have to worry about the fact that the collection is distributed, nor do you have to handle node failures yourself.

Besides automatic fault-tolerance and distribution, the RDD provides an elaborate API, which allows you to work with a collection in a functional style. You can filter the collection, map over it with a function, reduce it to a cumulative value, subtract, intersect or create a union with another RDD, and so on.

[3] Data locality is honored if each block gets loaded in the RAM of the same node where it resides in HDFS. The whole point is to try to avoid having to transfer large amounts of data over the wire.

Back to our example. We said that our Spark data engineer now has a reference to the RDD, so in order to find the error count, she first wants to remove all the lines that don't have an `OutOfMemoryError` substring. This is a job for the `filter` function, which she calls like this:

```
val oomLines = lines.filter(l => l.contains("OutOfMemoryError")).cache()
```

After filtering the collection so it contains the subset of data that she needs to analyze (figure 1.4), she calls `cache` on it, which tells Spark to leave that RDD in memory across jobs. Caching is the basic component of Spark's performance improvements we mentioned before. The benefits of caching the RDD will become apparent later.

Figure 1.4: Filtering the collection to contain only lines containing the OutOfMemoryError string

Now, she is left with only those lines that contain the error substring. For this simple example, we'll ignore the possibility that the `OutOfMemoryError` string might occur in multiple lines of a single error. Our data engineer counts the remaining lines and reports the result as the number of out-of-memory errors that occurred in the last two weeks:

```
val result = oomLines.count()
```

Spark enabled her to perform distributed filtering and counting of the data with only three lines of code. Her little program was executed on all three nodes in parallel.

If she now wants to further analyze lines with `OutOfMemoryErrors`, and perhaps call `filter` again (but with other criteria) on an `oomLines` object that was previously cached in memory, Spark won't load the file from HDFS again, as it would normally do, but will just load it from the cache.

1.4 The Spark revolution

Although the last decade saw Hadoop's wide adoption, Hadoop is not without its shortcomings. Although powerful, as its mascot, the yellow elephant illustrates, it is slow. This has opened

the way for newer technologies, such as Spark, to solve the same challenges Hadoop solves, but more efficiently. In the next few pages, we'll discuss Hadoop's shortcomings and how Spark answers those issues.

The Hadoop framework, with its Hadoop Distributed File System (HDFS) and MapReduce data-processing engine, was the first framework that brought distributed computing to the masses. Hadoop solved the three main problems facing any distributed data-processing endeavor:

- *Parallelization*—How to perform subsets of the computation simultaneously
- *Distribution*—How to distribute the data
- *Fault-tolerance*—How to handle component failure

NOTE Appendix A describes MapReduce in more detail.

On top of that, Hadoop clusters are often made of commodity hardware, which makes Hadoop easy to set up. That's why the last decade saw its wide adoption.

1.4.1 MapReduce's shortcomings

Although Hadoop is the foundation of today's Big Data revolution and is actively used and maintained, it still has its shortcomings and they mostly pertain to its MapReduce component. MapReduce job results need to be stored in HDFS before they can be used by another job. For this reason, MapReduce is inherently bad with iterative algorithms.

Furthermore, many kinds of problems do not easily fit MapReduce's two-step paradigm, and decomposing every problem into a series of these two operations can be difficult. The API can be cumbersome at times.

Hadoop is a rather low-level framework, so a myriad of tools has sprung up around it: tools for importing and exporting data, higher-level languages and frameworks for manipulating data, tools for real-time processing, and so on. They all bring additional complexity and requirements with them, which complicates any environment.

Spark solves many of these issues.

1.4.2 What Spark brings to the table

Spark's core concept is an in-memory execution model that enables caching job data in memory, instead of fetching it from disk every time, as MapReduce does. This can speed the execution of jobs up to 100 times,[4] compared to the same jobs in MapReduce, and it has the biggest effect on iterative algorithms such as machine learning, graph algorithms, and other types of workloads that need to reuse data.

[4] See "Shark: SQL and Rich Analytics at Scale", by Reynold Xin et al., https://amplab.cs.berkeley.edu/wp-content/uploads/2013/02/shark_sigmod2013.pdf

For example, imagine you have city map data stored as a graph. The vertices of this graph represent points of interest on the map, and the edges represent possible routes between them, with associated distances. And imagine you need to find a spot for a new ambulance station that will be situated as close as possible to all the points on the map. That spot would be the center of your graph. It can be found by first calculating the shortest path between all the vertices, and then finding the *farthest point distance* (the maximum distance to any other vertex) for each vertex, and finally finding the vertex with the smallest farthest point distance. Completing the first phase of the algorithm, finding the shortest path between all vertices, in a parallel manner is the most challenging (and complicated) part, but not impossible.[5] In the case of MapReduce, you'd need to store the results of each of these three phases on disk (HDFS). Each subsequent phase would read the results of the previous one from disk. But with Spark, you can find the shortest path between all vertices and simply cache that data in memory. Then the next phase can use that data from memory, find the farthest point distance for each vertex, and cache its results. Finally, the last phase could go through this last cached data and find the vertex with the minimum farthest point distance.

You can imagine the performance gains compared to reading and writing to disk every time.

Spark performance is so good that it recently (October 2014) won the Daytona Gray Sort contest and set a new world record (jointly with TritonSort, to be fair) by sorting 100 TB in 1,406 seconds (see http://sortbenchmark.org/).

SPARK'S EASE OF USE

The Spark API is much easier to use than the classic MapReduce API. To implement the classic word-count example from Appendix A as a MapReduce job, you'd need three classes: the main class that sets up the job, a `Mapper`, and a `Reducer`, each ten lines long, give or take a few.

[5] See " A Scalable Parallelization of All-Pairs Shortest Path Algorithm for a High Performance Cluster Environment", by T. Srinivasan et al., http://www.greymind.com/Publications/A%20Scalable%20Parallelization%20of%20All-Pairs%20Shortest%20Path%20Algorithm%20for%20a%20High%20Performance%20Cluster%20Environment.pdf

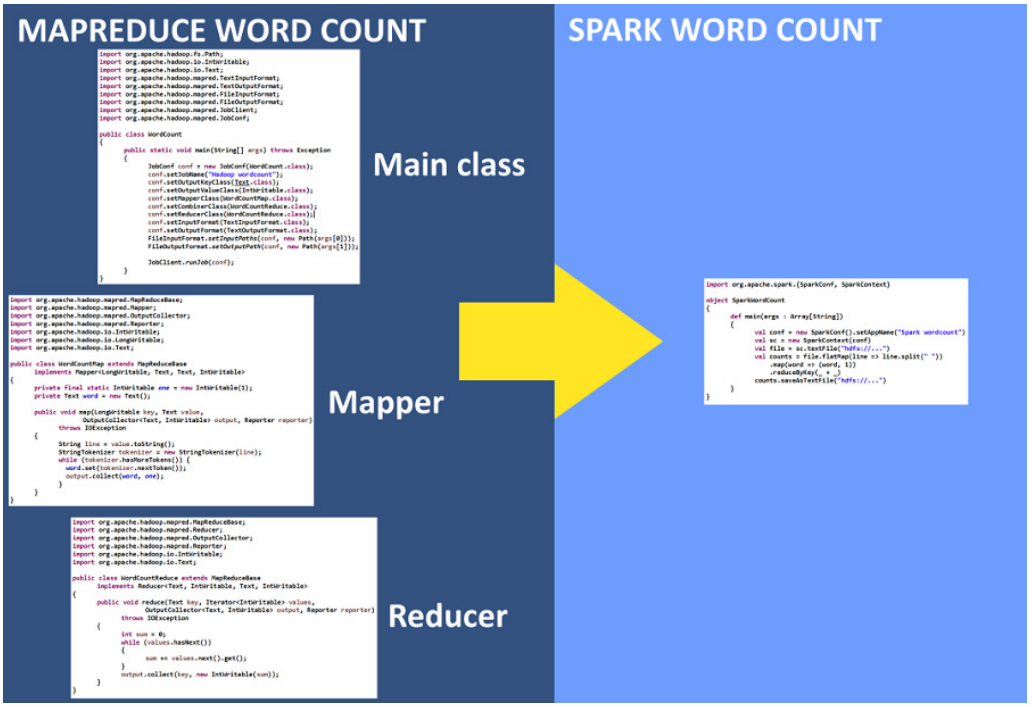

Figure 1.5: A word count program demonstrates Spark's conciseness and simplicity. The program implemented in Hadoop's MapReduce framework is on the left, and as a Spark Scala program on the right

By contrast, the following is all it takes for the same Spark program written in Scala:

```
val conf = new SparkConf().setAppName("Spark wordcount")
val sc = new SparkContext(conf)
val file = sc.textFile("hdfs://...")
val counts = file.flatMap(line => line.split(" "))
    .map(word => (word, 1)).countByKey()
counts.saveAsTextFile("hdfs://...")
```

Figure 1.1. shows this graphically.

Spark supports Scala, Java, Python and R programming languages, so it is accessible to a much wider audience. Although Java is supported, Spark can take advantage of Scala's versatility, flexibility, and functional programming concepts that are a much better fit for data analysis. Python and R are widespread among data scientists and in the scientific community, which brings those users on par with Java and Scala developers.

Furthermore, the Spark shell (REPL, or read-eval-print loop) offers an interactive console that can be used for experimentation and idea testing. There's no need for compilation and deployment just to find out something isn't working (again). REPL can even be used for launching jobs on the full set of data.

Finally, Spark can run on several types of clusters: Spark standalone cluster, Hadoop's YARN (which stands for "yet another resource negotiator") and Mesos. This gives it additional flexibility and makes it accessible to a larger community of users.

SPARK AS A UNIFYING PLATFORM

An important aspect of Spark is its combination of many functionalities of the tools in the Hadoop ecosystem into a single unifying platform. The execution model is general enough that the single framework can be used for stream data processing, machine learning, SQL-like operations, and graph and batch processing. Many roles can work together on the same platform, which helps in bridging the gap between programmers, data engineers, and data scientists.

And the list of functions that Spark provides is continuing to grow.

SPARK ANTI-PATTERNS

Spark isn't suitable, though, for asynchronous updates to shared data [6] (such as online transaction processing, for example) because it has been created with batch analytics in mind. (Spark streaming is just batch analytics applied on data in a time window.) Tools specialized for those use cases will still be needed.

Also, if you don't have a large amount of data, Spark might not be needed at all, because it needs to spend some time setting up jobs, tasks, and so on. Sometimes a simple relational database or a set of clever scripts can be used to process your data more quickly than a distributed system such as Spark. But data often has a tendency to grow, and it might outgrow your RDBMS (relational database management system) or your clever scripts rather quickly.

1.5 Spark ecosystem

We already mentioned the Hadoop ecosystem. Some of its most important tools are shown in figure 1.6. Infrastructure tools, providing basic data storage, synchronization and scheduling functions, are marked with dark blue. Interface tools, or those that can be used to transfer data between Hadoop and other systems, are marked with light yellow. Analytic tools, or those that provide data transformation and manipulation functions, are marked with light blue and brown backgrounds. Finally, Ambari, a management tool, is marked with blue.

[6] "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing", by Matei Zaharia et al., www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf

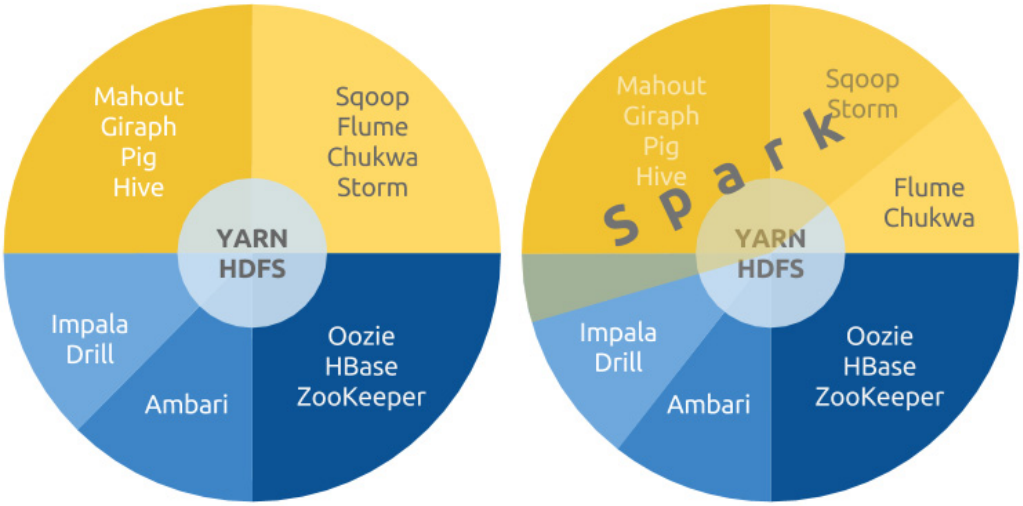

Figure 1.6: Basic infrastructure, interface, analytic and management tools in the Hadoop ecosystem with some of its functionalities that Spark incorporates or makes obsolete

Please note that this figure is by no means complete.[7] One could argue that we failed to add one tool or another, but a complete list of tools would be hard to fit in this section. We believe, though, that this list represents a good subset of the most prominent tools in the Hadoop ecosystem.

If you compare the functionalities of Spark components with the tools in the Hadoop ecosystem, you can see that some of the tools are suddenly superfluous. For example, Apache Giraph can be replaced by Spark GraphX and Spark MLlib can be used instead of Apache Mahout. Apache Storm's capabilities overlap greatly with those of Spark Streaming so in many cases Spark Streaming can be used instead.

Apache Pig and Apache Sqoop aren't needed anymore, as the same functionalities are covered by Spark Core and Spark SQL. But even if you have legacy Pig workflows and need to run Pig, the Spork project enables you to run Pig on Spark.

Spark has no means of replacing the infrastructure and management of the Hadoop ecosystem tools, though. Impala and Drill can coexist alongside Spark, especially with Drill's coming support for Spark as an execution engine.

We already said that Spark doesn't need to use HDFS storage. Besides HDFS, Spark can operate on data stored in Amazon S3 buckets and plain files. More exciting, it can also use

[7] If you're interested, you can find a hopefully complete list of Hadoop-related tools and frameworks at hadoopecosystemtable.github.io.

Tachyon, which is a memory-centric distributed file system, or other distributed file systems, such as GlusterFS.

Another interesting fact is that Spark doesn't have to run on YARN. Apache Mesos and the Spark standalone cluster are alternative cluster managers for Spark. Apache Mesos is an advanced distributed systems kernel bringing distributed resource abstractions. It can scale to tens of thousands of nodes with full fault-tolerance. Spark Standalone Cluster is a Spark-specific cluster manager that is used in production today on multiple sites.

So if we switch from MapReduce to Spark and get rid of YARN and all the tools that Spark makes obsolete, what's left of the Hadoop ecosystem? Or to put it this way: are we slowly moving toward a Spark ecosystem?

1.6 Summary

- Apache Spark is an exciting new technology that is rapidly superseding Hadoop's MapReduce as the preferred Big Data processing platform
- Spark programs can be 100 times faster than their MapReduce counterparts
- Spark supports Java, Spark, Python and R languages
- Writing distributed programs with Spark is similar to writing the local Java, Spark or Python programs
- Spark provides a unifying platform for batch programming, real-time data-processing functions, SQL-like handling of structured data, graph algorithms, and machine learning, all in a single framework
- Spark is not appropriate for small data sets nor should you use it for OLTP applications
- Spark components are Spark Core, Spark SQL, Spark Streaming, Spark MLlib and Spark GraphX
- Resilient Distributed Datasets (RDDs) are Spark's abstraction of distributed collections
- Spark supersedes some of the tools in Hadoop ecosystem

In the next chapter, we'll take you by the hand while you install Spark, give you an overview of the Spark interactive shell, and help you start writing your first Spark programs.