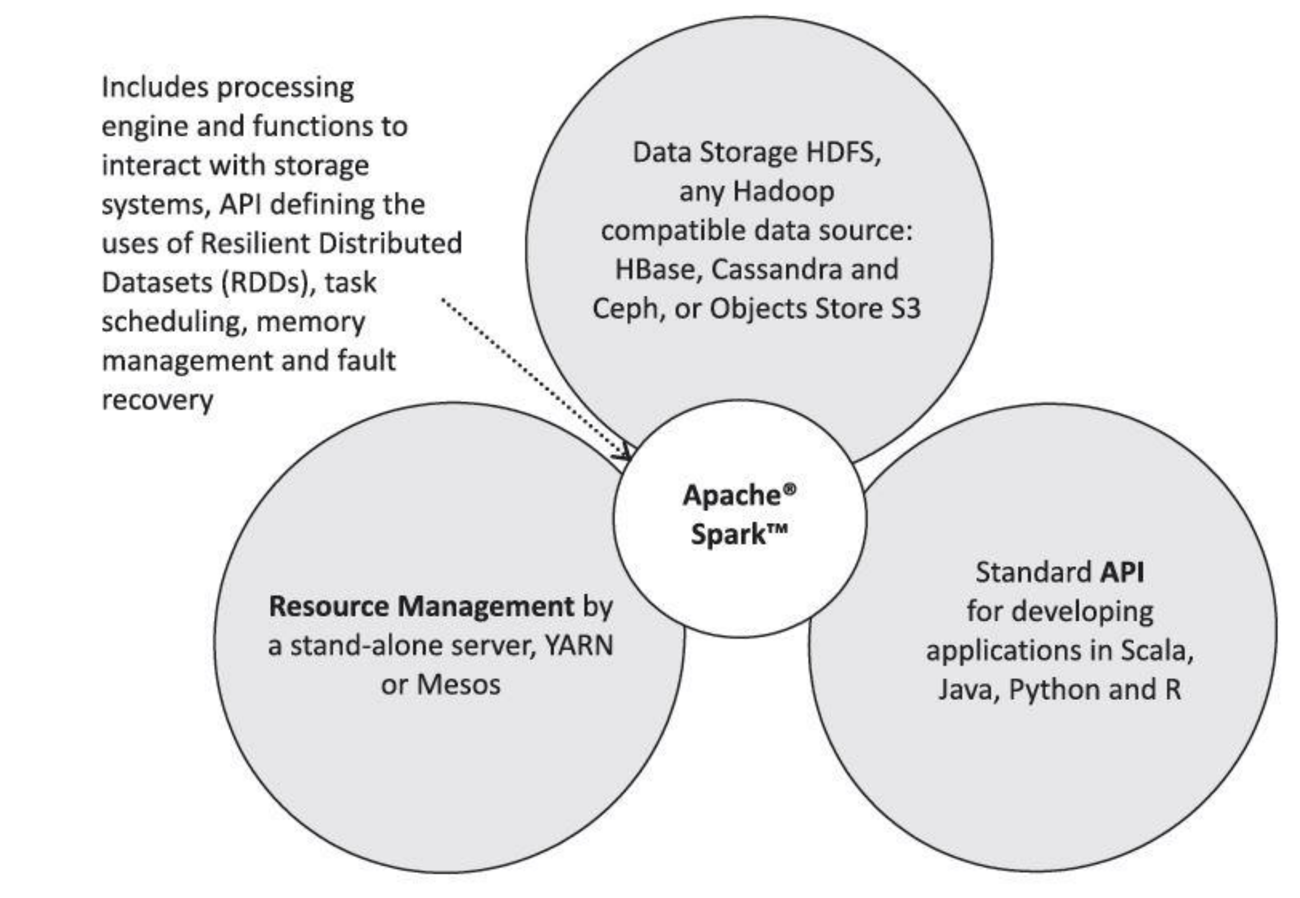
Lesson 2

Apache® Spark™ Main Components, Features, and Architecture Layers

Apache® Spark™

- A fast and general compute engine with a simple and expressive programming model.
- Powers the analytics applications up to 100 times faster
- Supports HDFS compatible data

Figure 5.1 Main components of the Spark architecture



Includes processing engine and functions to interact with storage systems, API defining the uses of Resilient Distributed Datasets (RDDs), task scheduling, memory management and fault recovery

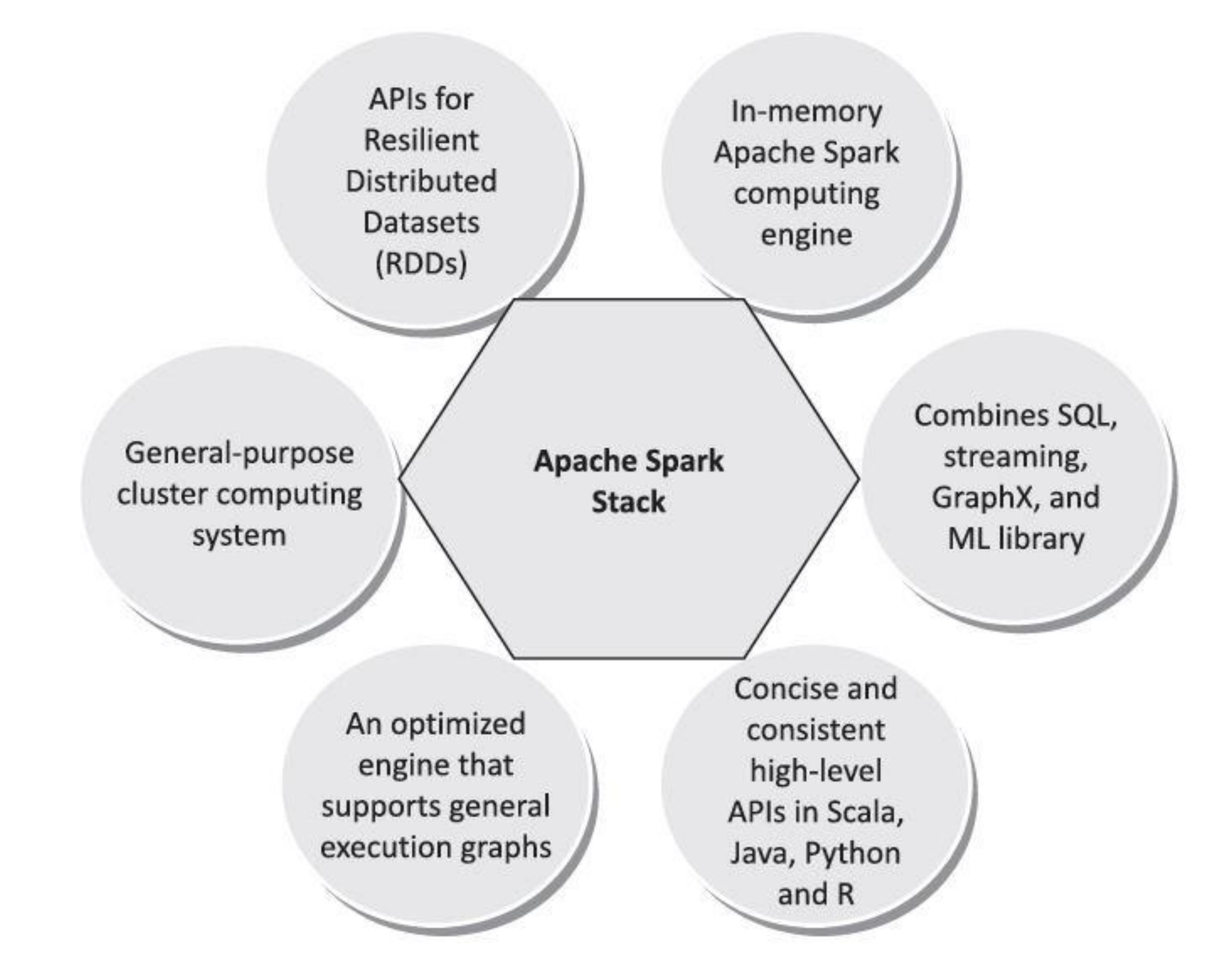
Data Storage HDFS, any Hadoop compatible data source: HBase, Cassandra and Ceph, or Objects Store S3

Apache® Spark™

Resource Management by a stand-alone server, YARN or Mesos

Standard **API** for developing applications in Scala, Java, Python and R

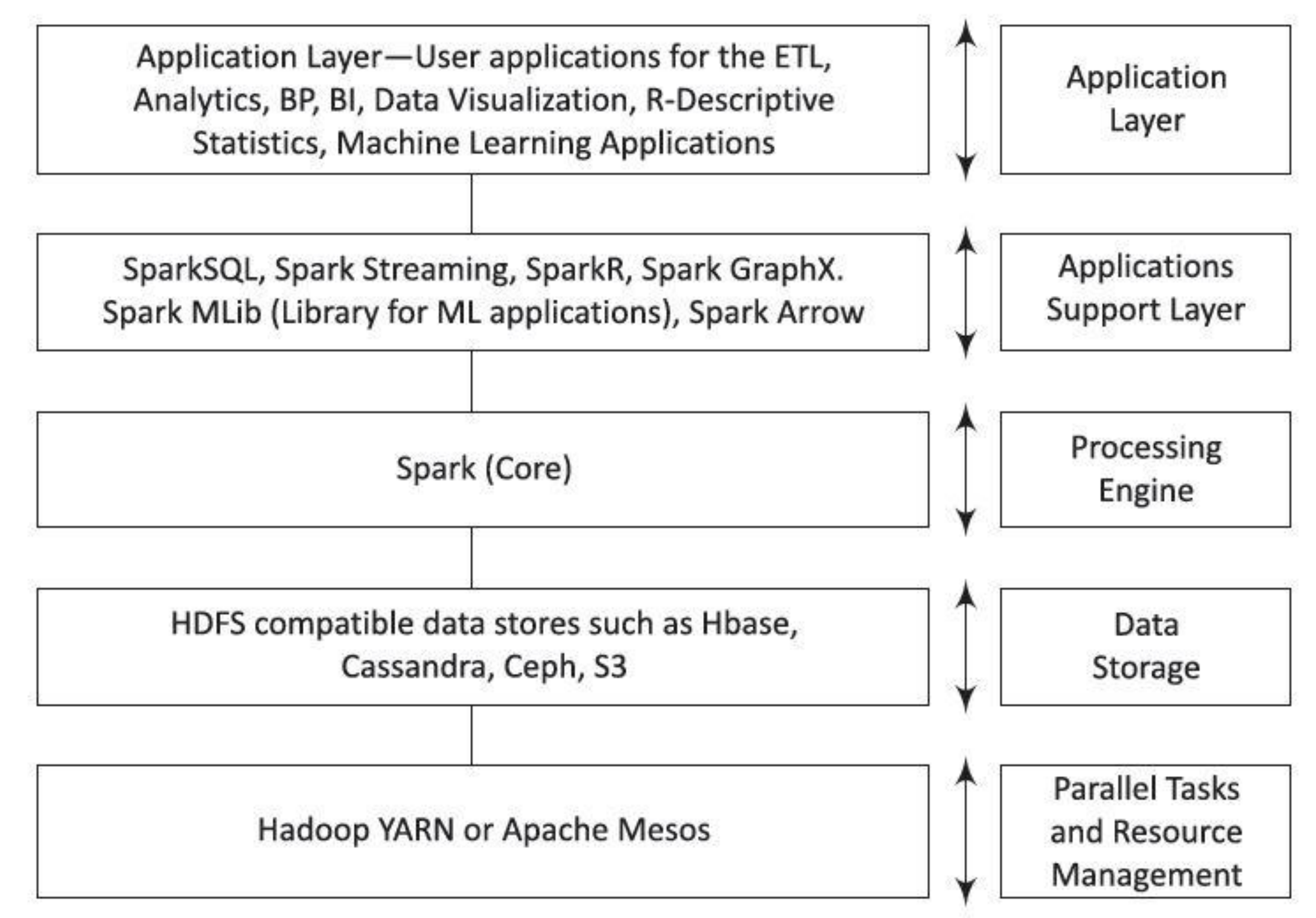
Figure 5.2 Main features of Spark


APIs for Resilient Distributed Datasets (RDDs)
In-memory Apache Spark computing engine
General-purpose cluster computing system
Apache Spark Stack
Combines SQL, streaming, GraphX, and ML library
An optimized engine that supports general execution graphs
Concise and consistent high-level APIs in Scala, Java, Python and R

Spark Software Stack

- The main components of Spark stack are SQL, Streaming, R, GraphX, MLib and Arrow at the applications support layer

Figure 5.3 Five-layer architecture for running applications using Spark stack



Application Layer—User applications for the ETL, Analytics, BP, BI, Data Visualization, R-Descriptive Statistics, Machine Learning Applications

Application Layer

SparkSQL, Spark Streaming, SparkR, Spark GraphX. Spark MLib (Library for ML applications), Spark Arrow

Applications Support Layer

Spark (Core)

Processing Engine

HDFS compatible data stores such as Hbase, Cassandra, Ceph, S3

Data Storage

Hadoop YARN or Apache Mesos

Parallel Tasks and Resource Management

Layer 1: management and scheduling of the resources

- Hadoop, YARN or Mesos facilitates the parallel running of the tasks and the management and scheduling of the resources

Layer 2: Data Store

- Such as HDFS, HBase, Cassandra, Ceph), or at the Objects Store Amazon S3

Layer 3: Spark core

- A processing engine using Data Store (layer 2) which provides the data to the processing engine using parallel running of tasks (layer 1)

Layer 4: Software Stack Components

1. **Spark SQL** for the structured data

The SQL runs the queries on Spark data in the traditional business analytics and visualization applications

2. **Spark Streaming** for processing real-time streaming data, micro-batches style of computing and processing

Uses the Dstream, a series of RDDs, to process the real-time data

Software Stack Components

3. **SparkR**, an R package used as light-weight front end for Apache Spark from R, APIs using through the RDD class

4. **Spark Mlib,** a scalable machine learning library, consisting of common learning algorithms and utilities, such as classification, regression, clustering, …..

Software Stack Components

5. **Spark GraphX,** a collection of graph and Graph analytics algorithms which extends to use of the Spark RDDs.

- 6. **Spark Arrow** for columnar in-memory analytics and enabling usages of vectorized UDFs (VUDFs), Arrow enables high performance Python UDFs for SerDe and data pipelines

Spark Supported File Formats

- Text file, Sequence File, CSV (Comma Separated Values) File, JSON file, Object file (for structured data, serializable and deserializable), TSV (Tab Separated Values) File

Summary

We learnt

- Spark main components
- Spark Features
- DataFrame
- RDDs
- Spark architecture

Summary

- In-memory processing for the analytics applications up to 100 times faster
- Spark stack of SQL, Streaming, R, GraphX, MLib and Arrow
- Supports HDFS compatible data: HDFS, HBase, Cassandra, Ceph), or at the Objects Store Amazon S3

End of Lesson 2 on **Apache® Spark™ Main Components, Features, and Architecture Layers**