SPARK

- Apache Spark is a **cluster computing platform** designed to be fast and general purpose.
- Spark extends the MapReduce model to support more types of computations, including **interactive queries** and **stream processing**.
- One of the main features Spark offers for speed is the ability to run computations **in memory**,
- The system is also more **efficient** than MapReduce for complex applications running on disk.

Source: Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning spark: lightning-fast big data analysis*. " O'Reilly Media, Inc.".

Apache Spark

Lightning-fast unified analytics engine

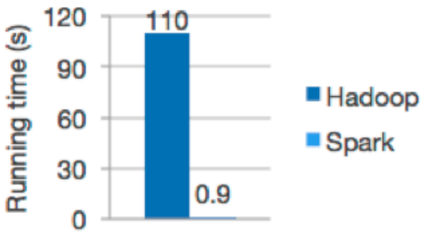
Download | Libraries | Documentation | Examples | Community | Developers | Apache Software Foundation

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Latest News

Spark 2.4.3 released (May 08, 2019)
Spark 2.4.2 released (Apr 23, 2019)
Spark 2.4.1 released (Mar 31, 2019)
Spark 2.3.3 released (Feb 15, 2019)

Archive

APACHECON
LAS VEGAS: Sept. 9-12, 2019
BERLIN: Oct. 22-24, 2019

Download Spark

Built-in Libraries:

SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Third-Party Projects