
# Introduction to PySpark

By Pratik Parmar

# Agenda

- Introduction to Spark
- Spark with Python
  - PySpark Shell
  - IPython/Jupyter notebooks
  - Python Scripts

# What is Apache Spark?

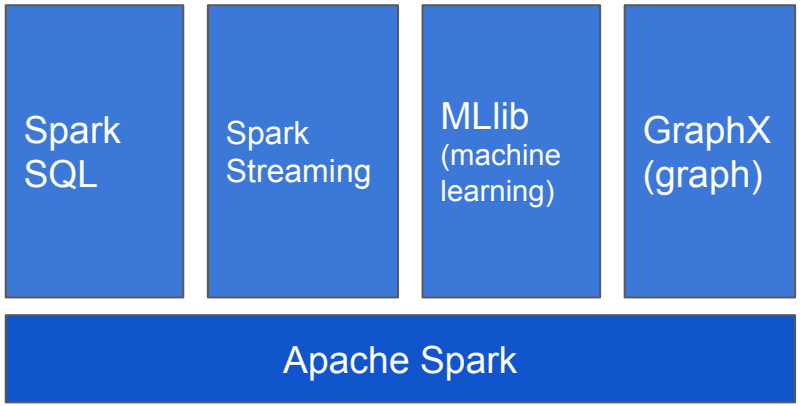
- Apache Spark™ is a fast and general engine for large scale data processing.
- Originally, developed in AMPLab at UC Berkely (2009), open sourced in 2010, transferred to Apache 2013.
- Claims to run programs upto 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

# Advantages over MapReduce

- Speed
- Ease of use - Scala, Java, Python, R
- Generality - Supports different use cases.
- Runs everywhere - Hadoop, Mesos, Standalone ...

# Spark Components

- Spark SQL and DataFrames
- Spark Machine Learning (MLlib)
- GraphX
- Spark Streaming



# Talk is Cheap.
# Show me the code.

Linus Torvalds

# Resources

- http://spark.apache.org/documentation.html
- https://databricks.com/spark/developer-resources
- http://spark-packages.org

# Pratik Parmar

Twitter: @hackyroot
GitHub: @hackyroot

# Thank You !!!