Introduction to Apache Spark

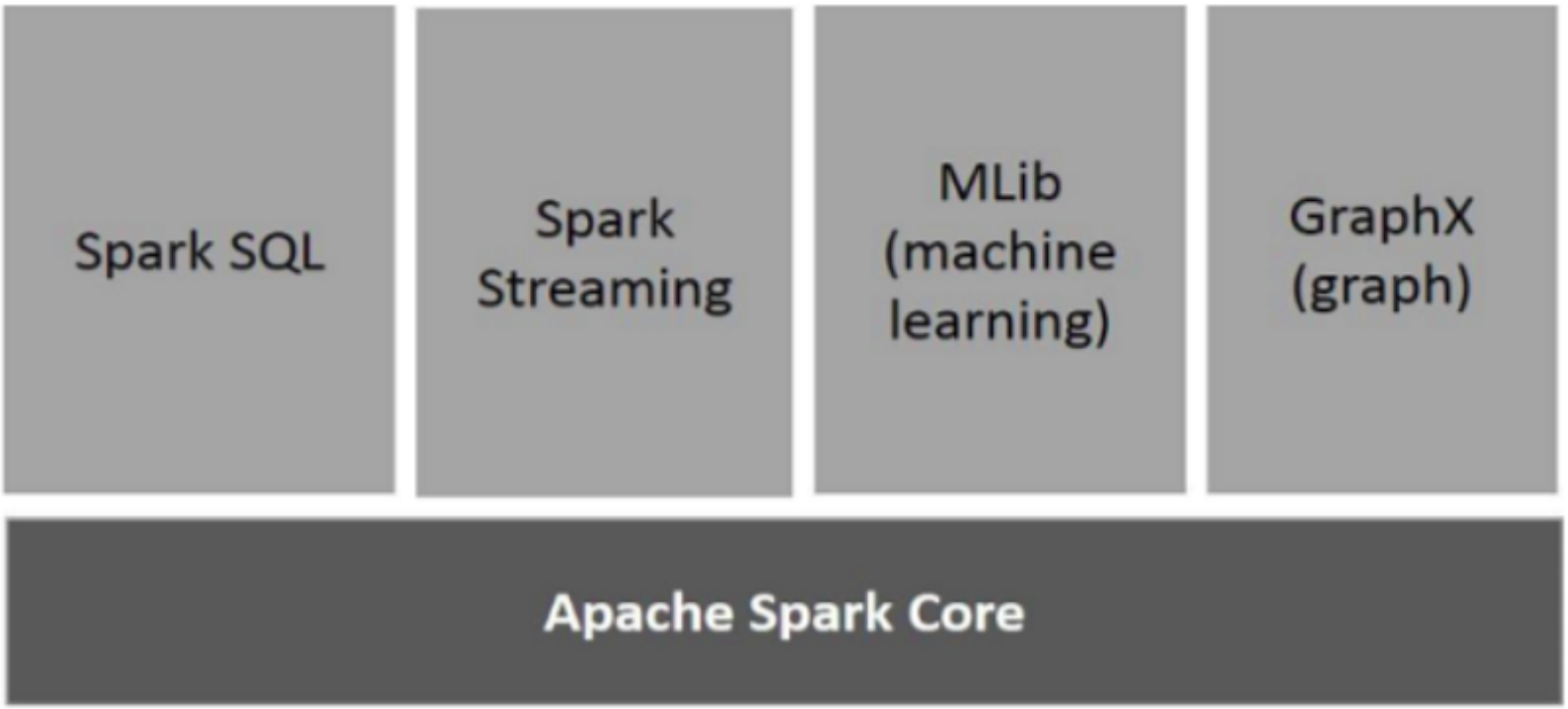
Ott Toomet

Computational Demography
2017-02-16

What is Apache Spark

- In-memory cluster computing
  - Shuffles around code, not data
- Can use various data sources (including distributed file systems, databases)
- Includes *MapReduce* framework (like Hadoop)
- Also includes advanced processing, ML
- Popular and rapidly developing

Components


Spark SQL
Spark Streaming
MLib (machine learning)
GraphX (graph)
Apache Spark Core

Spark Core

- 1 Driver + many executors
  - Executors may be on different nodes
  - Related memory needs
- Written in scala
  - API for scala, java, python, R
- RDD: Resilient Distributed Dataset
  - Distributed: partitioned across the spark cluster
  - Resilient: if a partition fails, only that partition is recovered
- Lazy evaluation
  - Postpone hard computations as long as possible

Components

- Spark SQL
  - DataFrames (R API)
  - SQL interface
    - Use standard SQL queries
- Spark Streaming
  - Real-time input, real-time output
  - Made of RDD-s
- Spark ML
  - Collection of ML algorithms
- graphX
  - Graph data structure

Conclusion

- Allows to access big datasets
  - Sometimes works rather well
  - Sometimes extremely hard to get to work
- Documentations is *bad*
  - Even Stackoverflow not too helpful
- Feels to be in the beta stage
  - Bugs, functionality missing, etc
  - Is this what bleeding edge means?