MAGNE+IC™

Apache Spark Demo

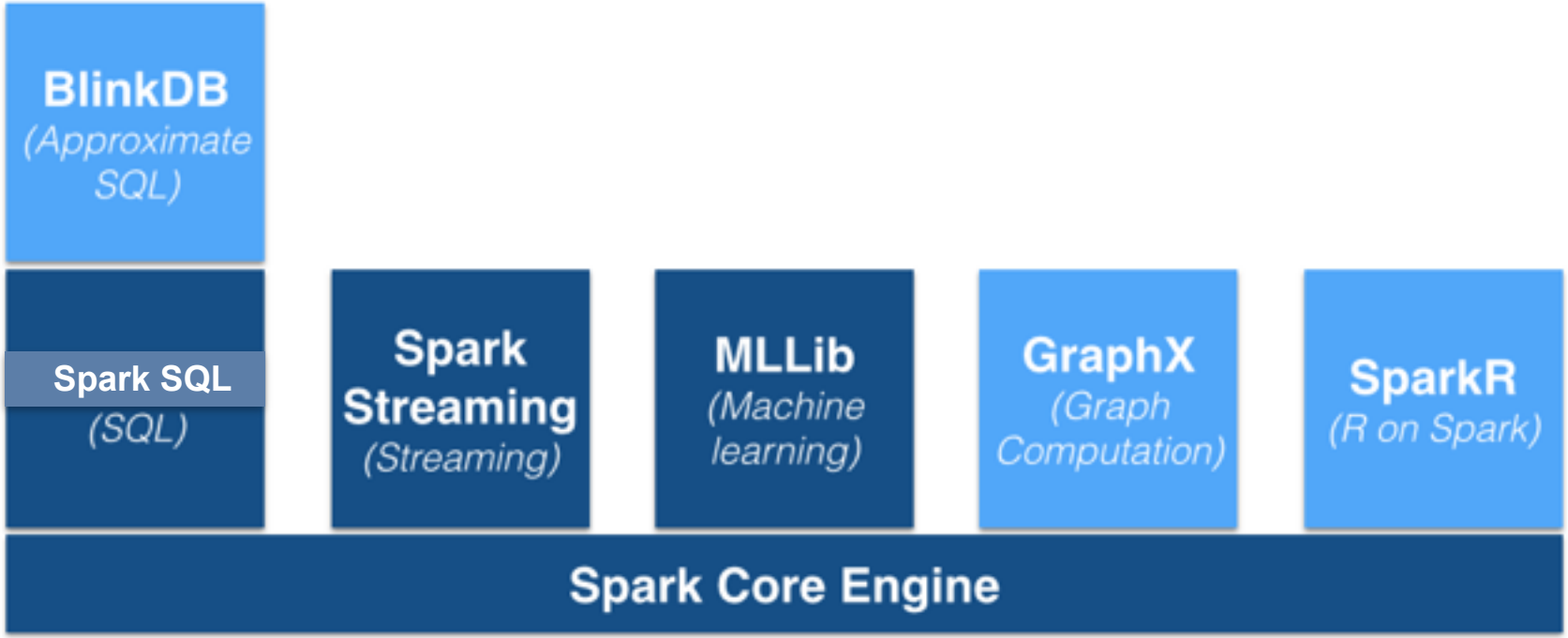
Apache Spark Demo

- About Apache Spark
- Write small program to illustrate API

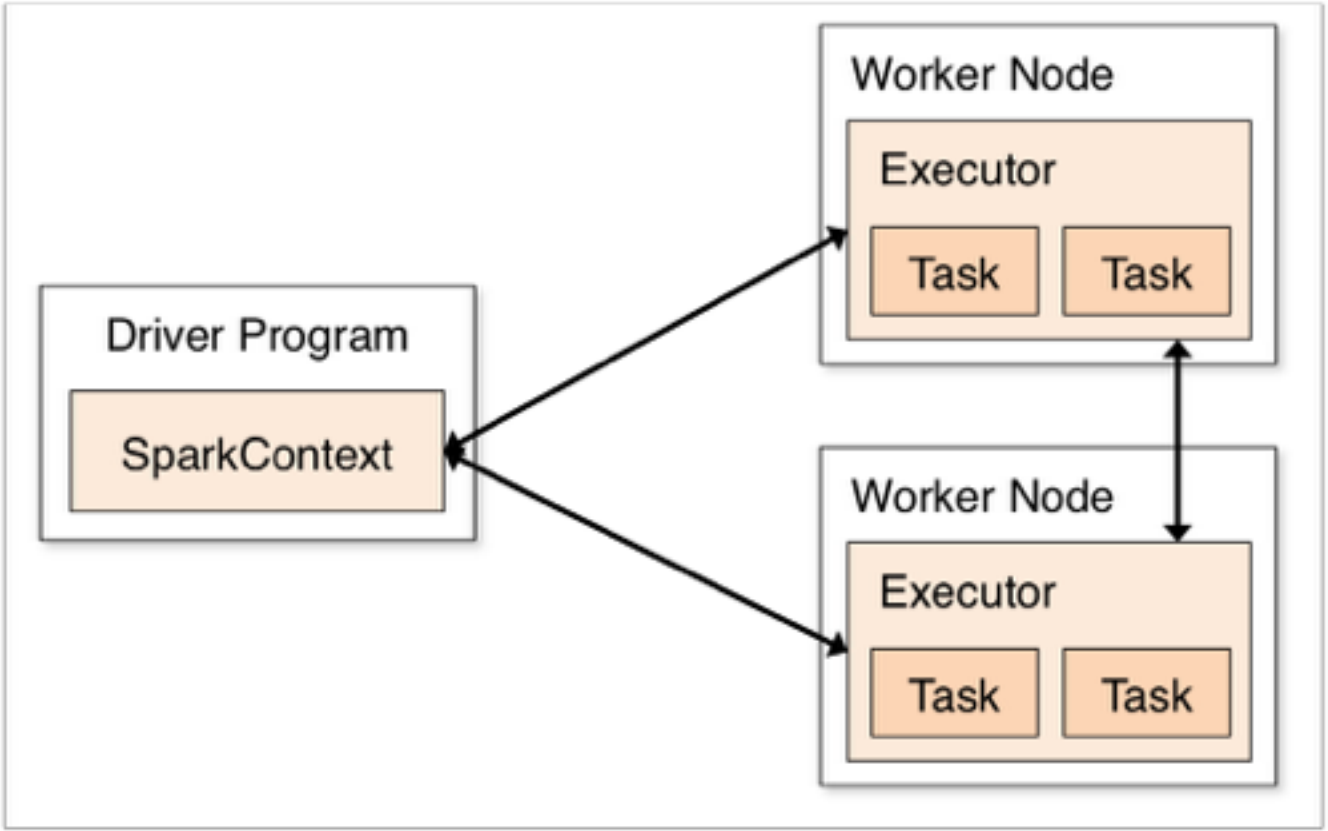
Apache Spark

- Open Source
- fast and general-purpose cluster computing system
- in memory cluster computing
  - May be up to 100 times faster then Map Reduce (for iterative algorithms)
- Supports large data sets
  - Works with Hadoop/HDFS
- same code for batch, interactive and online/streaming
- APIs: Scala, Java, Python

Apache Spark ecosystem



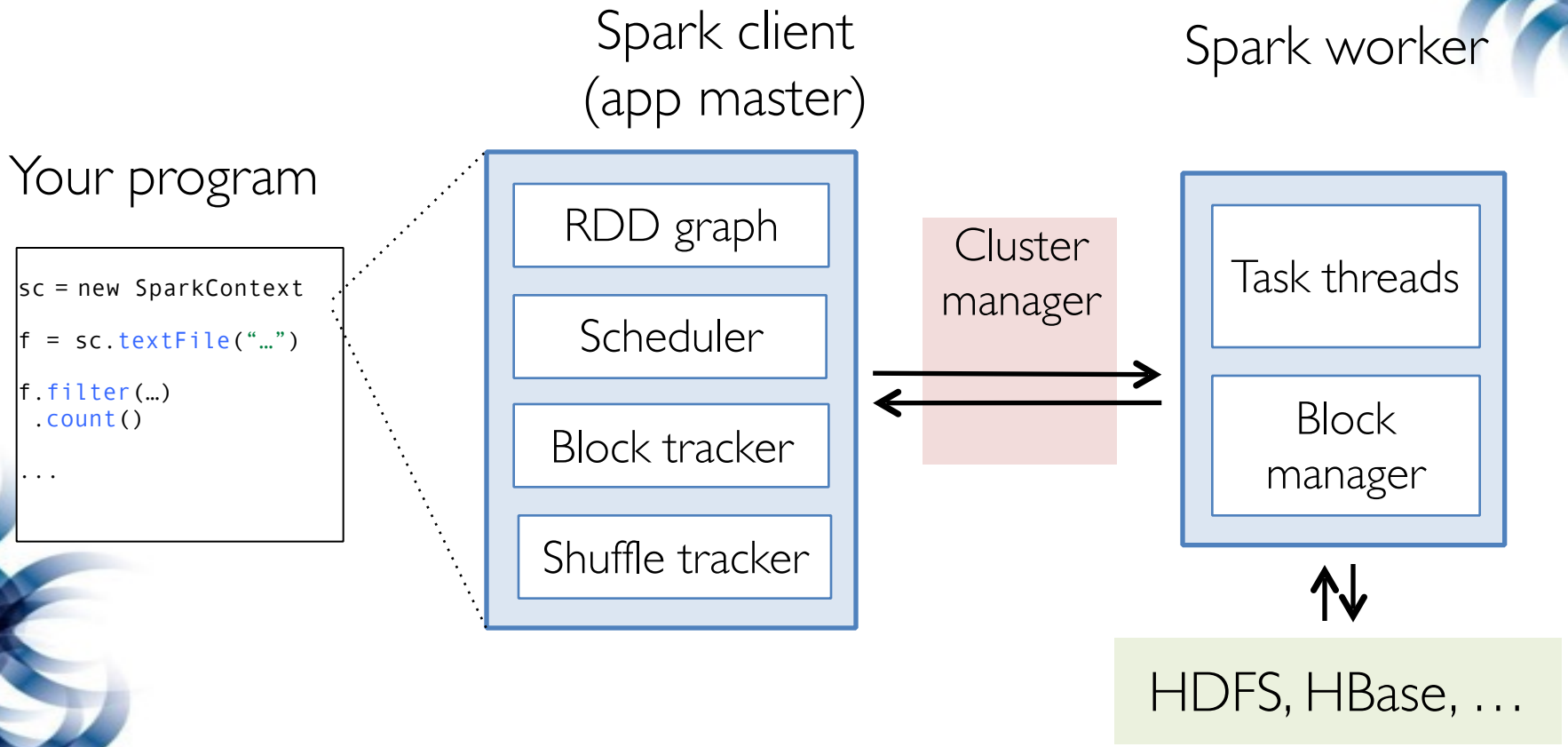
Distributed Execution Model



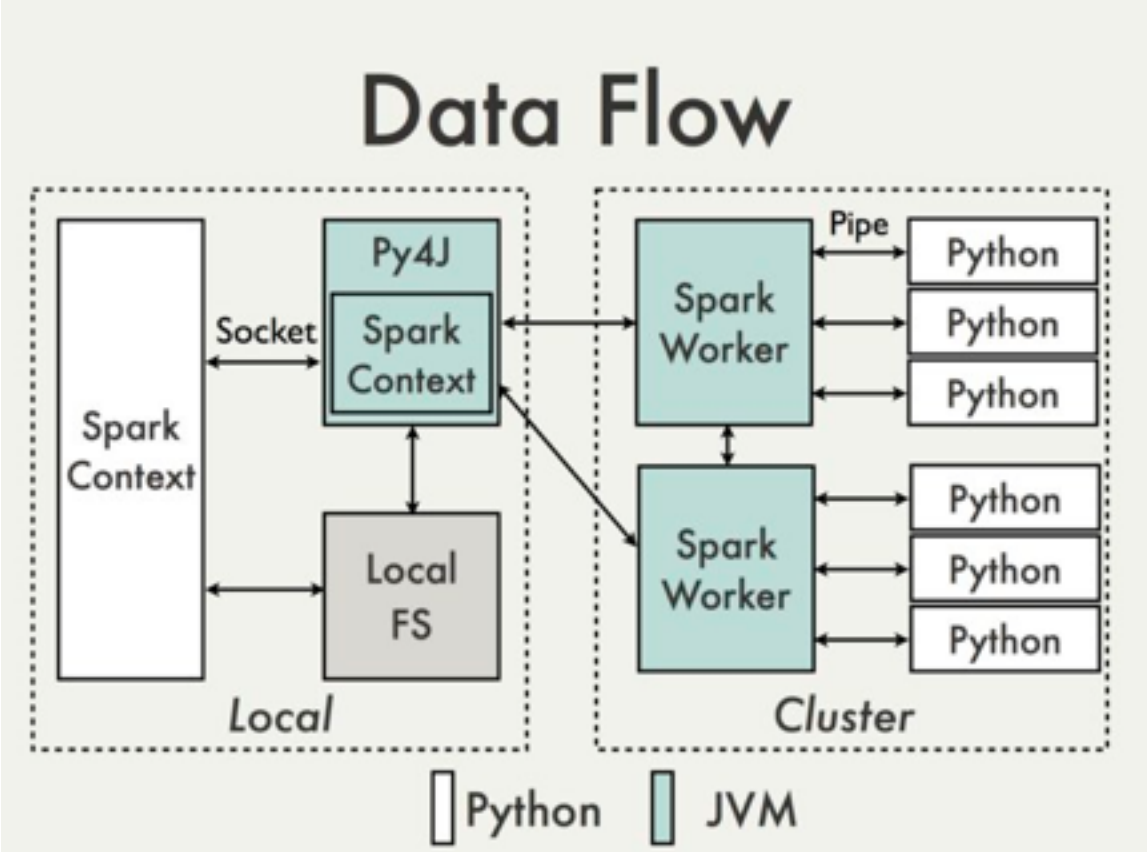
How to run Spark program

- Supported languages
  - Scala
  - Java
  - Python
- Interactive shells
  - Scala - sparkshell
  - Pyspark - Python shell
- Stand alone programs

Spark program execution



PySpark Architecture

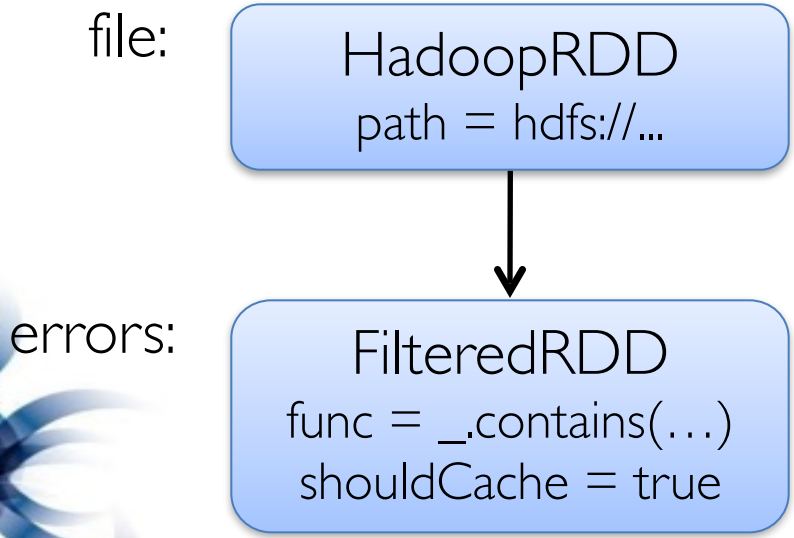


Resilient Distributed Dataset

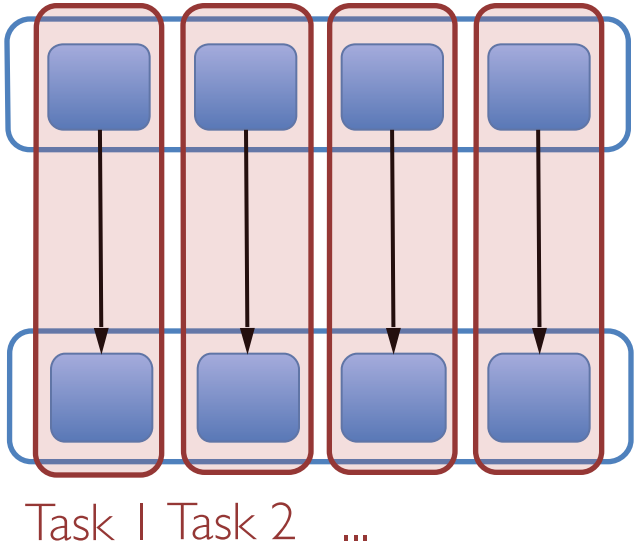
- collection of distributed elements
- split into multiple partitions
- can contain any type of Python, Java or Scala objects (serializable)
- Storage models
  - in memory
  - persisted to disk
  - recomputed if lost
- data locality aware

Resilient Distributed Dataset

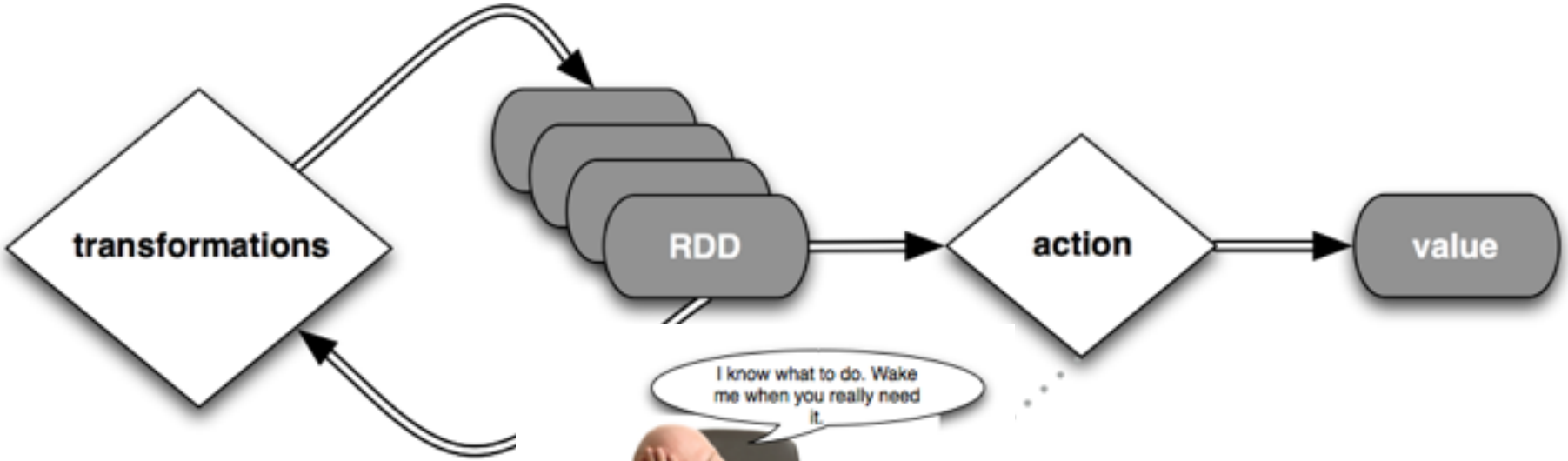
Dataset-level view:



Partition-level view:



Resilient Distributed Dataset



Demo program

- Calculate histogram of user requests frequencies
  - Count the number of unique users by the number of times their do requests
- ipython notebook demo
  - http://master:18888

Spark SQL demo

- ipython notebook demo
  - http://master:18888

urls

- Spark Cluster Console
  - http://master:8080
- Ganglia Monitoring
  - master:5080/ganglia

Questions

Thank you