Yandex

Actions

Driver & executors

› Driver program runs your Spark application

› Driver delegates tasks to executors to use cluster resources

› In local mode, executors are collocated with the driver

› In cluster mode, executors are located on other machines

› More in the next lesson

Actions

› Triggers data to be materialized and processed on the executors and then passes the outcome to the driver

› Example: actions are used to collect, print and save data

Frequently used actions

- collect*()*
  - collects items and passes them to the driver
  - for small datasets! all data is loaded to the driver memory

- take(n: *Int)*
  - collects only n items and passes them to the driver
  - tries to decrease amount of computation by peeking on partitions

- top(n: *Int)*
  - collects n largest items and passes them to the driver

- reduce(f: *(T, T) ➞ T)*
  - reduces all elements of the dataset with the given associate, commutative binary function and passes the result back to the driver

Frequently used actions

- saveAsTextFile(path: *String*)
  - each executor saves its partition to a file under the given path with every item converted to a string and confirms to the driver
- saveAsHadoopFile(path: *String*, outputFormatClass: *String*)
  - each executor saves its partition to a file under the given path using the given Hadoop file format and confirms to the driver

Frequently used actions

- foreach(f: *T* ➞ *()*)
  - each executor invokes f over every item and confirms to the driver
- foreachPartition(f: *Iterator[T]* ➞ *()*)
  - each executor invokes f over its partition and confirms to the driver

Quiz

Summary

- Actions trigger computation and processing of the dataset
- Actions are executed on executors and they pass results back to the driver
- Actions are used to collect, save, print and fold data

BigDATAteam