# Shared variables

Igor Yakushin
ivy2@uchicago.edu

December 27, 2019

# Shared variables

- Spark's second abstraction - shared variables
- By default variables are not shared between tasks
- Two kinds of shared variables supported:
  - Broadcast variables - can be used to cache a value in memory on all nodes
  - Accumulators - such as counters, sums; one can only "increment" those variables; can be used to store intermediate results of reduce operation

# Shared variables

```
lines = sc.textFile("README.md")
l1 = lines.map(lambda line: len(line.split()))
l1.reduce(lambda a,b: a if (a>b) else b)

pairs = lines.flatMap(lambda s: s.split()).map(lambda w: (w,1))
result = pairs.reduceByKey(lambda a,b: a+b)
print("\n".join(map(lambda x: "{} -> {}".format(*x),
                    result.collect())))

distData = sc.parallelize(list(range(1000))

b = sc.broadcast(list(range(10)))
b.value

a = sc.accumulator(0)
sc.parallelize(list(range(5))).foreach(lambda x: a.add(x))
a.value
```