Getting started with Apache Spark

- Neeraj Bhadani

Hotels.com™

Neeraj Bhadani - Bio

- Working as BIG Data Engineer in Expedia Group (Hotels.com Brand)
- More than 10 years of Industry experience.
- LinkedIn : https://www.linkedin.com/in/neerajbhadani/

Index

- Limitations of map-reduce
- What is apache spark
- Spark components
- Architecture
- RDD : Resilient Distributed Dataset
- RDD operations
- Anatomy of spark job

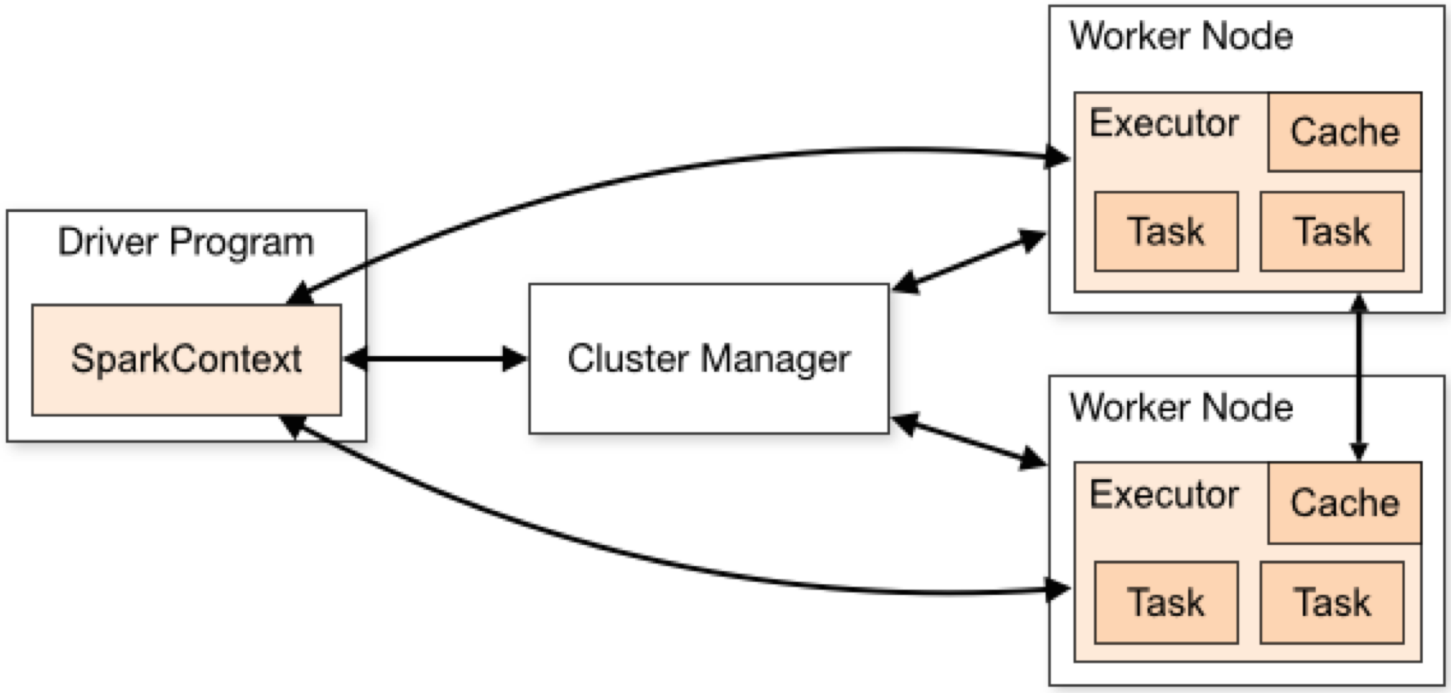
Limitations of MapReduce

- Slow Processing Speed
- Support for Batch Processing only
- Not efficient for iterative processing
- Not Easy to Use
- Unsuitable for trivial operations

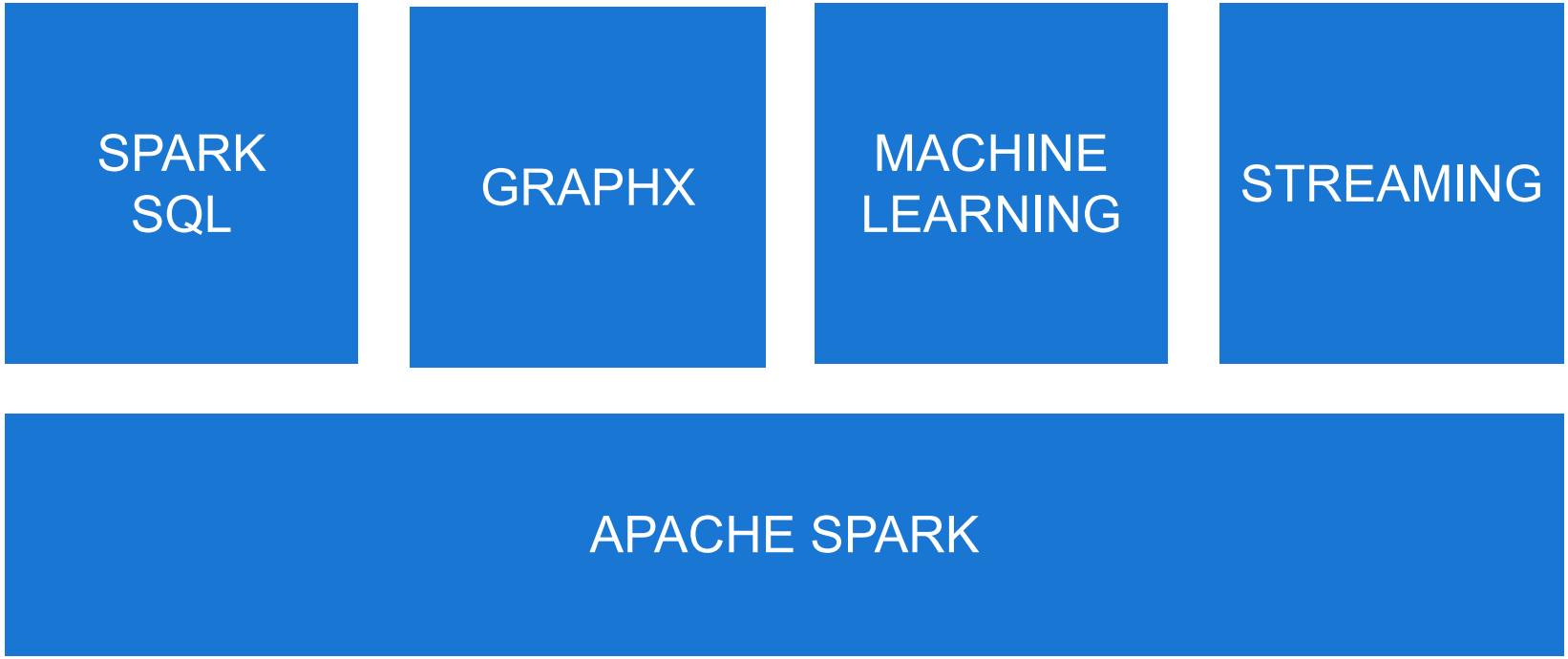
What is Apache Spark

- Initially started at UC Berkeley in 2009
- Fast and general purpose cluster computing system
- Most popular for running Iterative Machine Learning Algorithms
- Provides high level APIs in
  - Java
  - Scala
  - Python
  - R
- Integration with Hadoop and its eco-system and can read existing data

Spark components

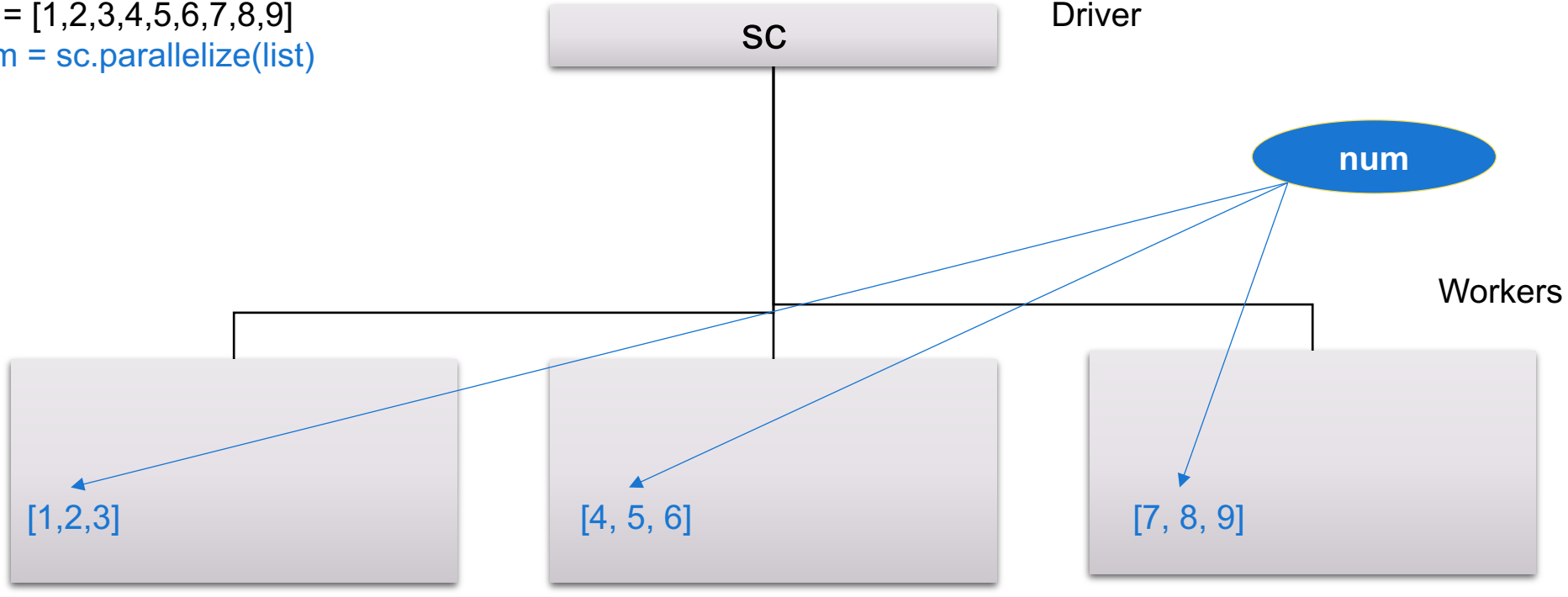


Architecture



RDD : Resilient distributed dataset

list = [1,2,3,4,5,6,7,8,9]
num = sc.parallelize(list)



RDD : Operations

Transformation : lazy

- Map
- Filter
- Join
- Flatmap
- Etc….

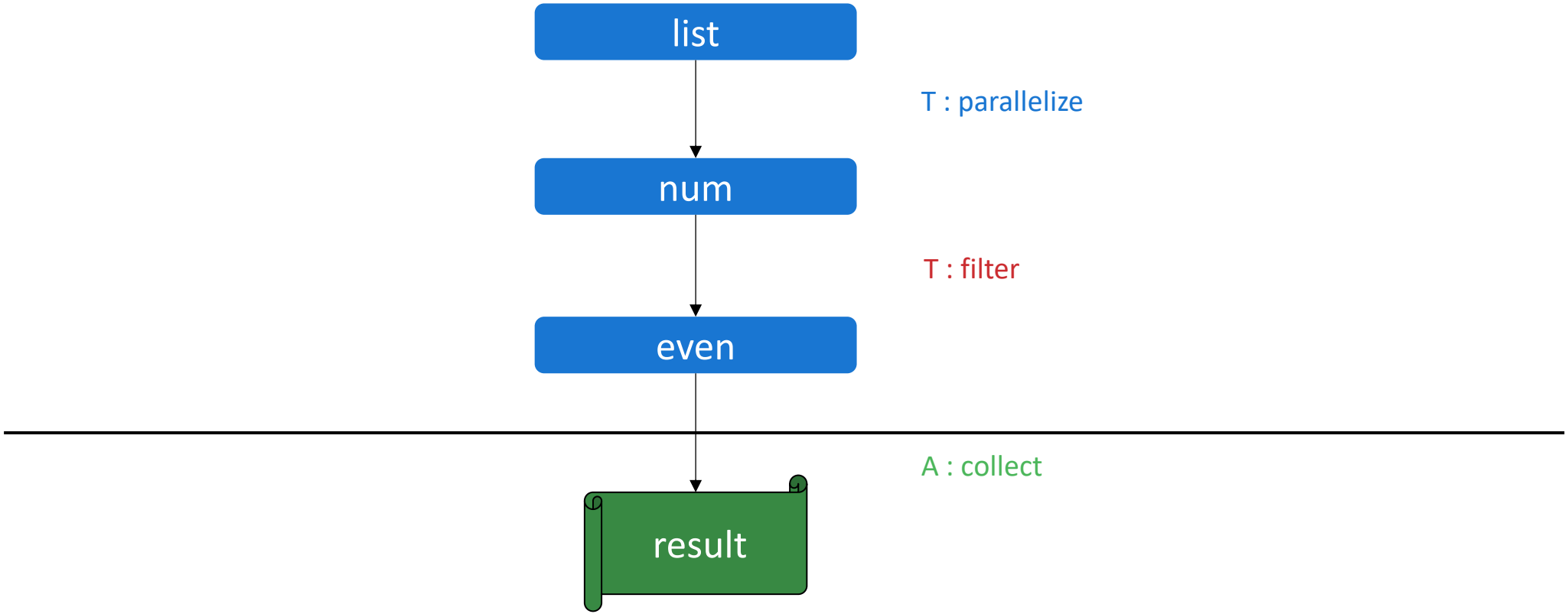
Actions : immediately

- Collect
- Count
- First
- Take
- Etc….

Sample code

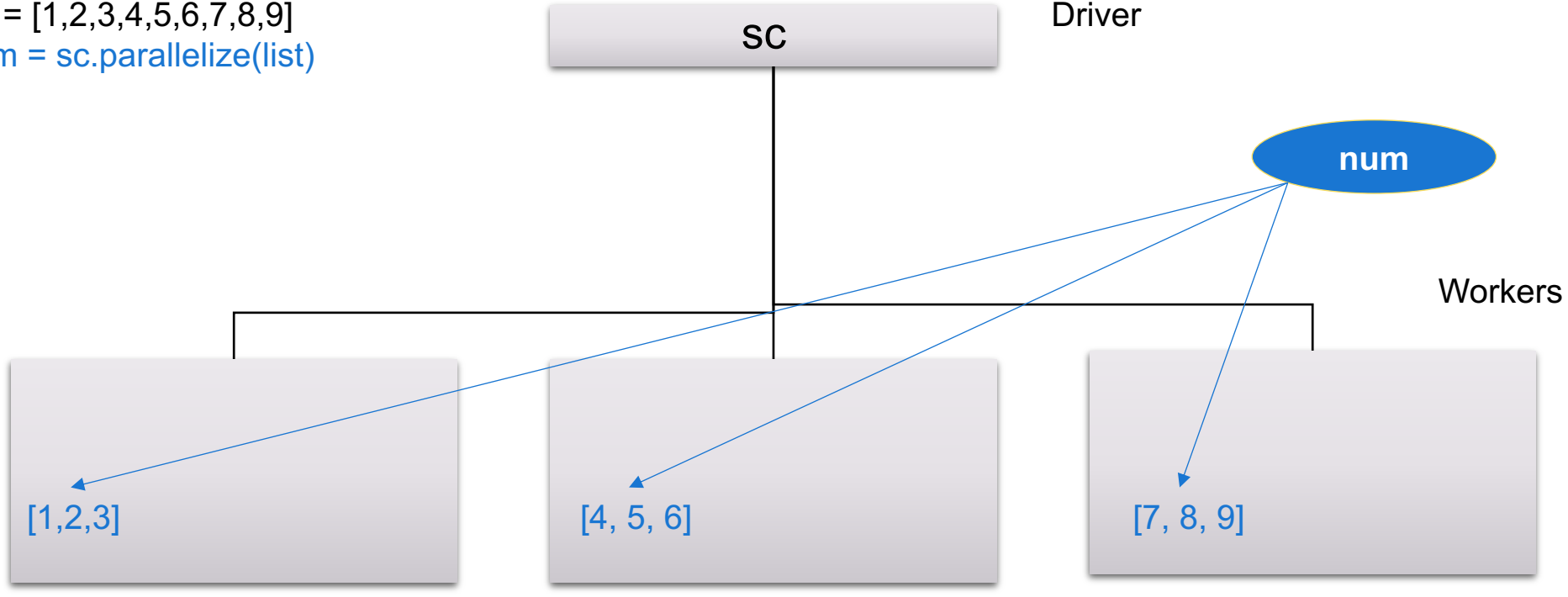
```
# python List
list = [1,2,3,4,5,6,7,8,9]

# Create RDD
num = sc.parallelize(list)

# Filter RDD
even = num.filter(lambda x : x%2 == 0)

# Collect result
even.collect()
```

DAG : Directed Acyclic Graph



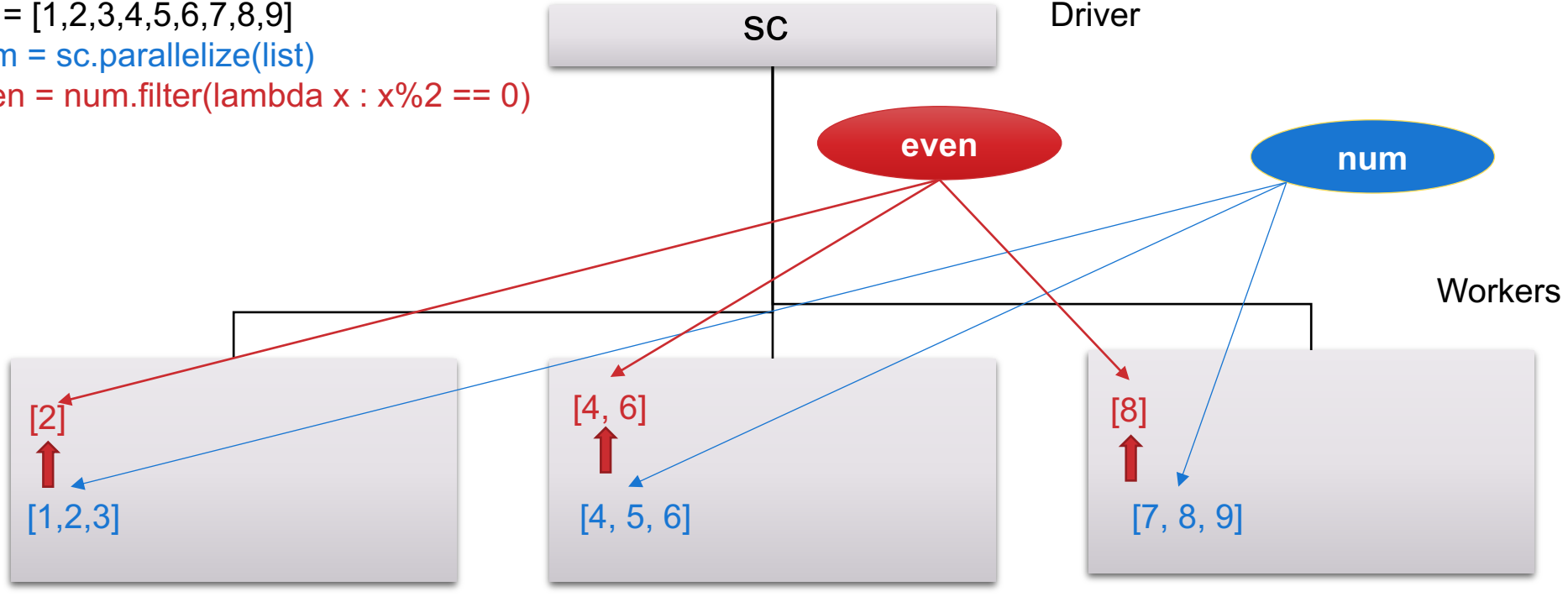
RDD : Resilient distributed dataset

list = [1,2,3,4,5,6,7,8,9]
num = sc.parallelize(list)



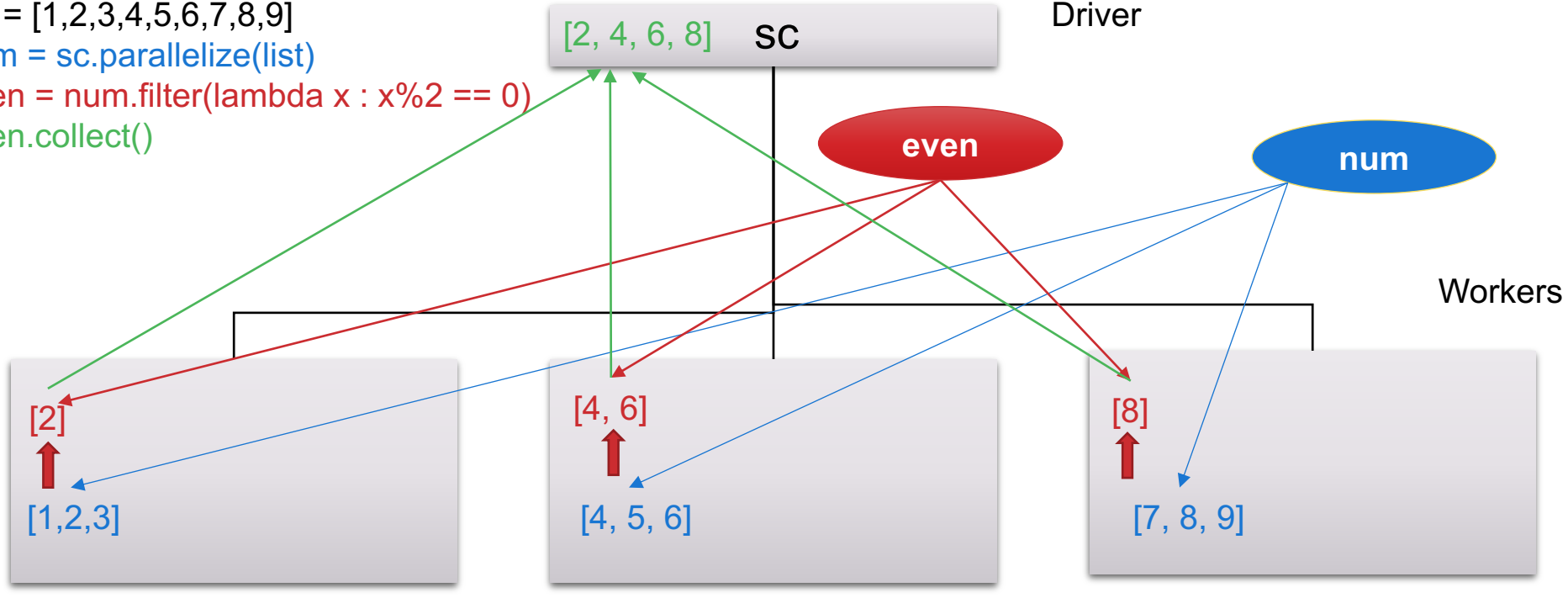
RDD : Resilient distributed dataset

```
list = [1,2,3,4,5,6,7,8,9]
num = sc.parallelize(list)
even = num.filter(lambda x : x%2 == 0)
```

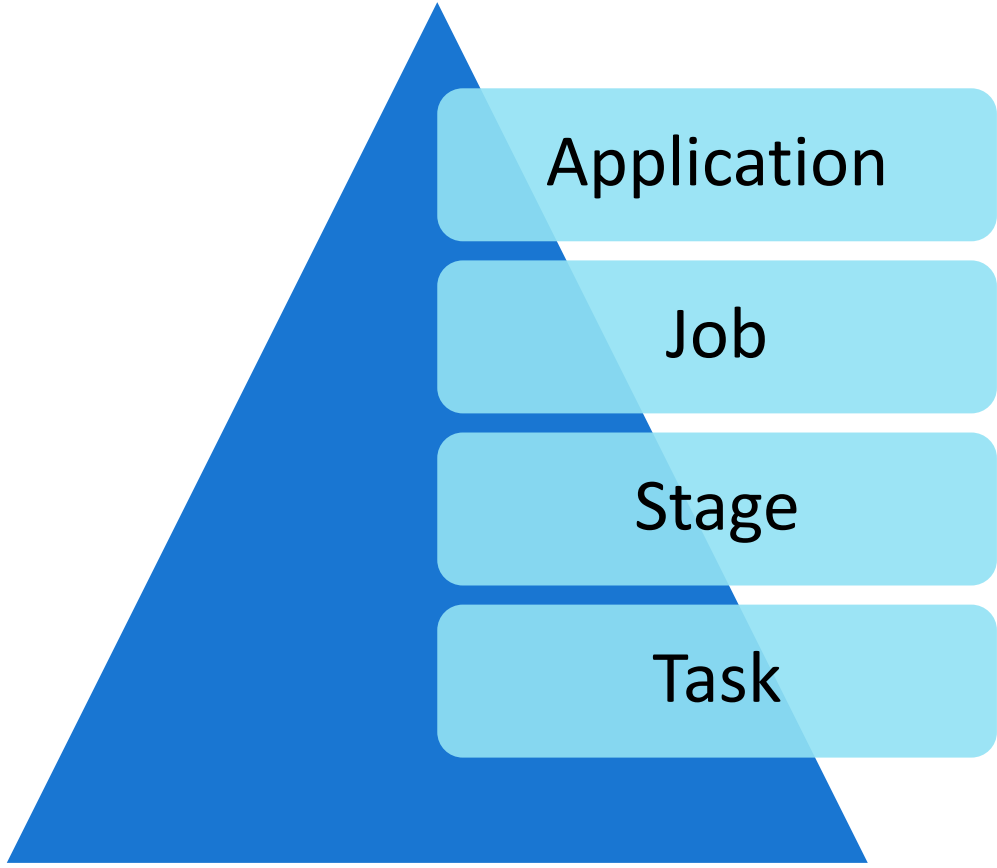


RDD : Resilient distributed dataset

```
list = [1,2,3,4,5,6,7,8,9]
num = sc.parallelize(list)
even = num.filter(lambda x : x%2 == 0)
even.collect()
```



Anatomy of spark job




DEMO

Feedback : QR Code

Thank You !