Introduction to Apache Spark

Neeraj Bhadani

About Me

- Working as Data Scientist @ Expedia Group
- LinkedIn: https://www.linkedin.com/in/neerajbhadani/
- Medium: https://medium.com/@bhadani.neeraj.08

Agenda

- Hadoop Ecosystem
- Limitations of Map - Reduce
- What is Apache Spark
- Spark Components
- Architecture
- RDD: Resilient Distributed Dataset
- Anatomy of Spark Job

Hadoop Ecosystem

1. Spark SQL
2. Spark Streaming
3. GraphX
4. MLlib

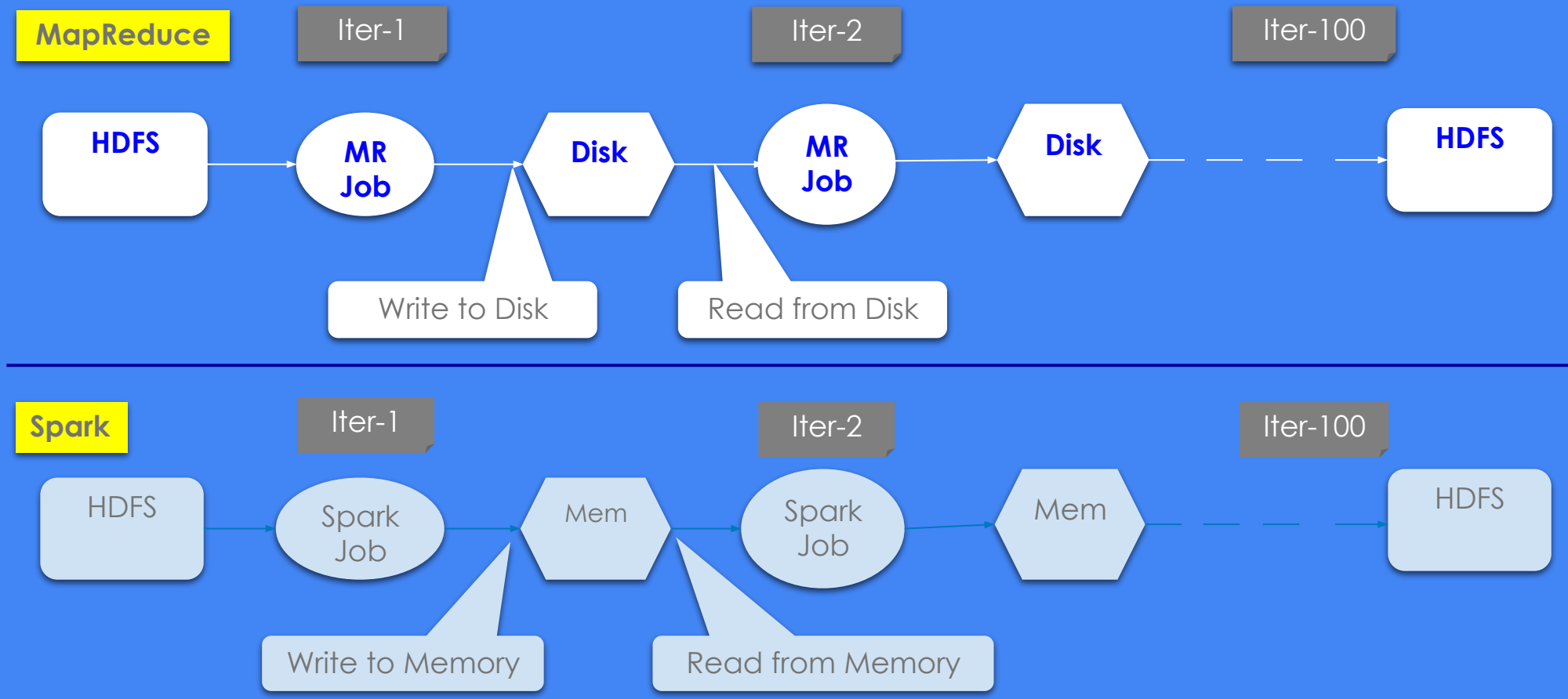
Limitations of Map-Reduce

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

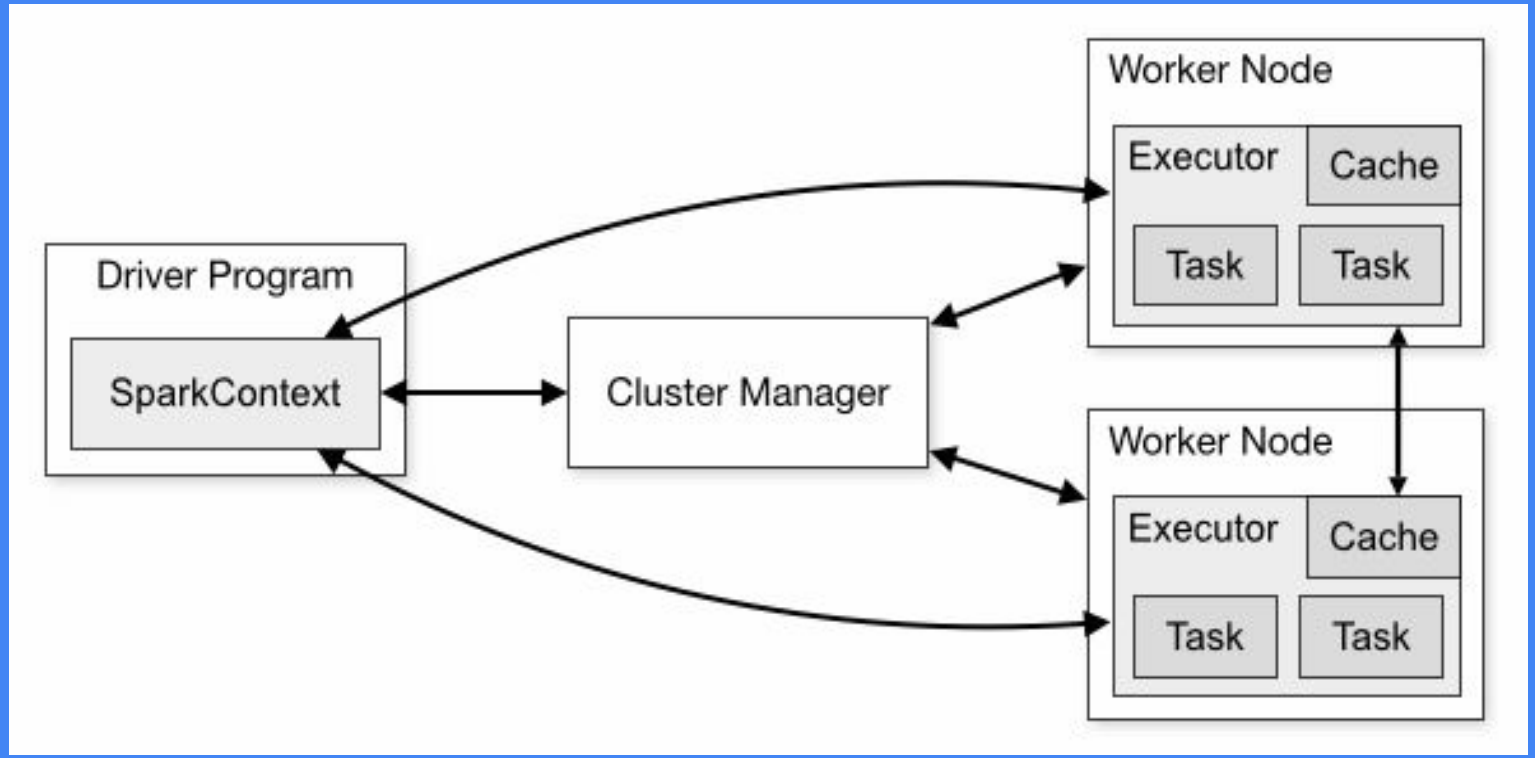
What is Apache Spark?

- 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
  - SQL
- Integration with Hadoop and its ecosystem

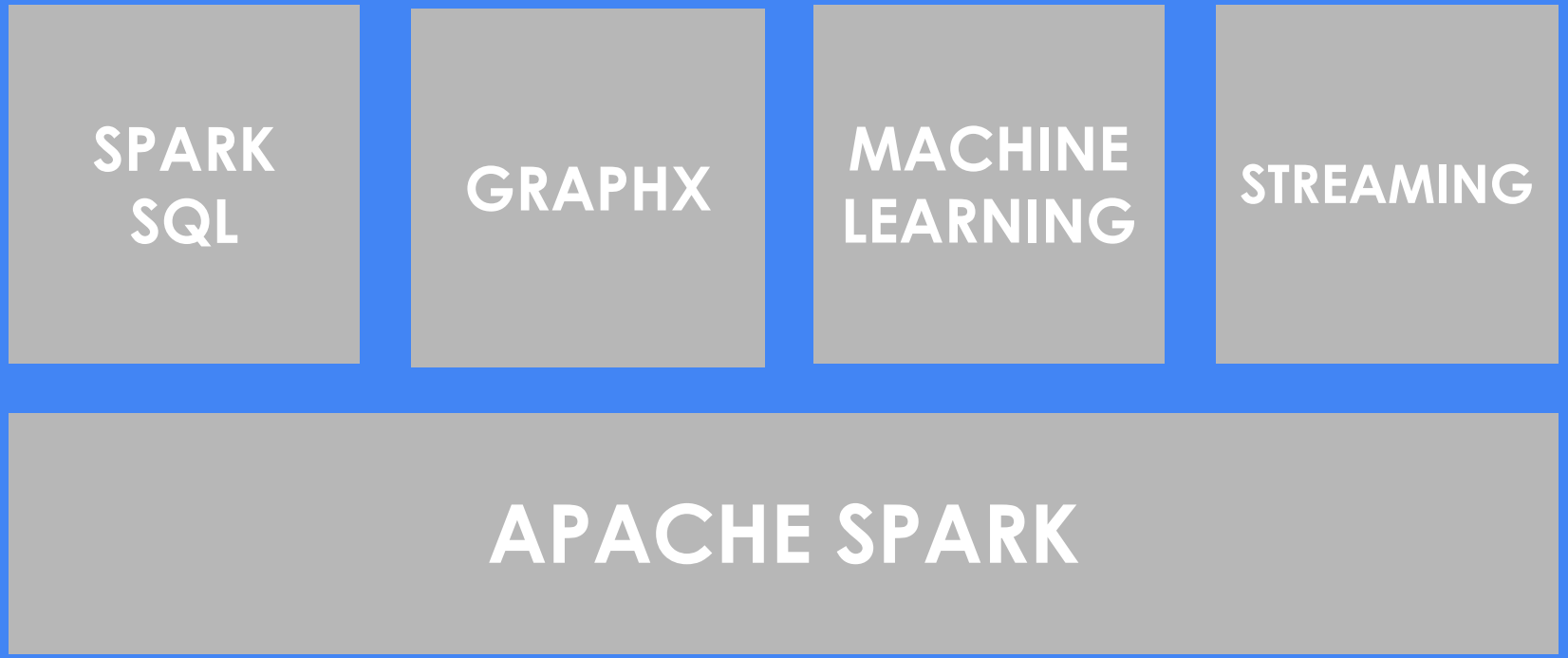
Spark – In Memory



Spark Components

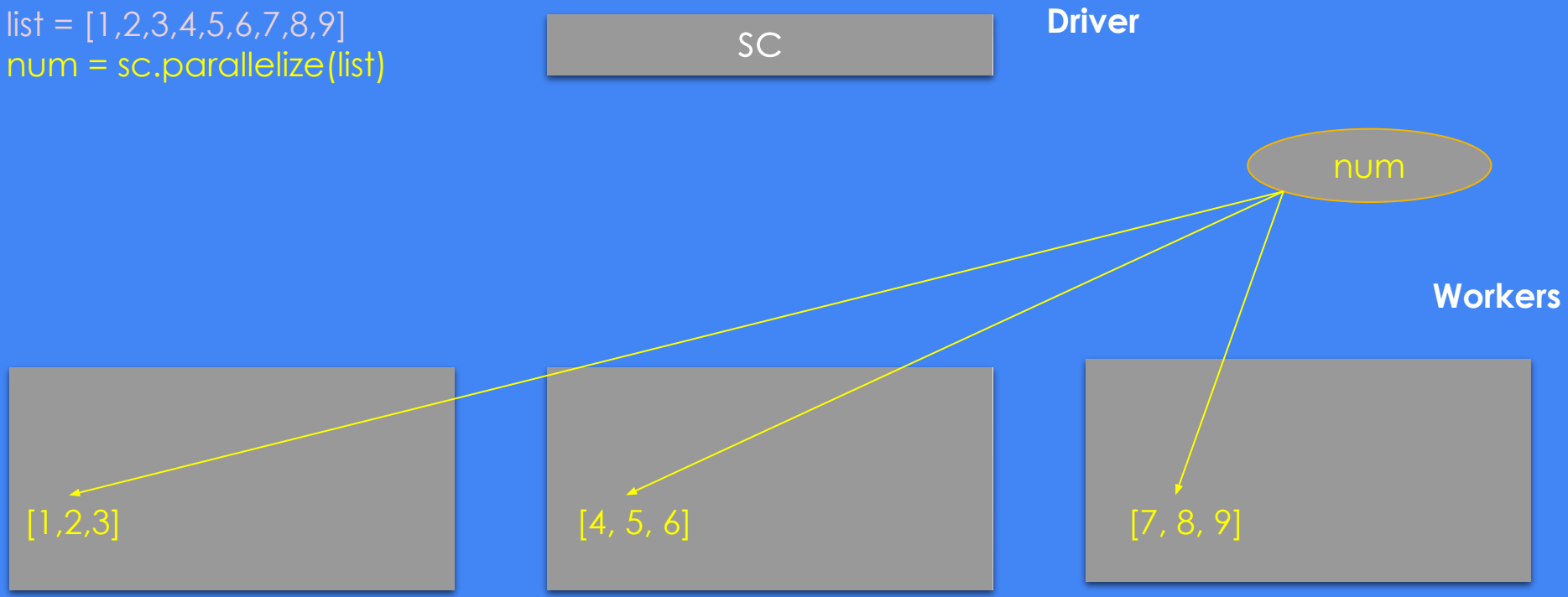


Architecture



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...

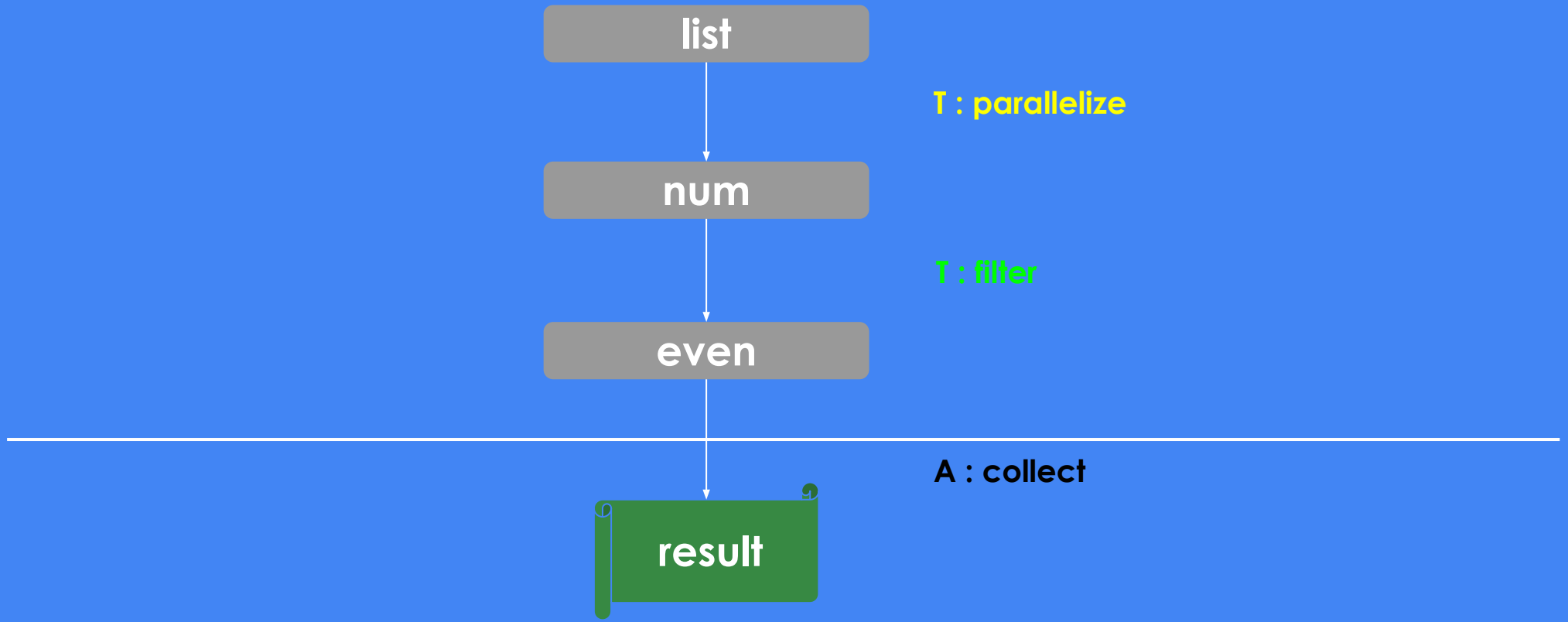
Actions: **Immediate**

- count
- collect
- first
- take
- etc...

Sample Code

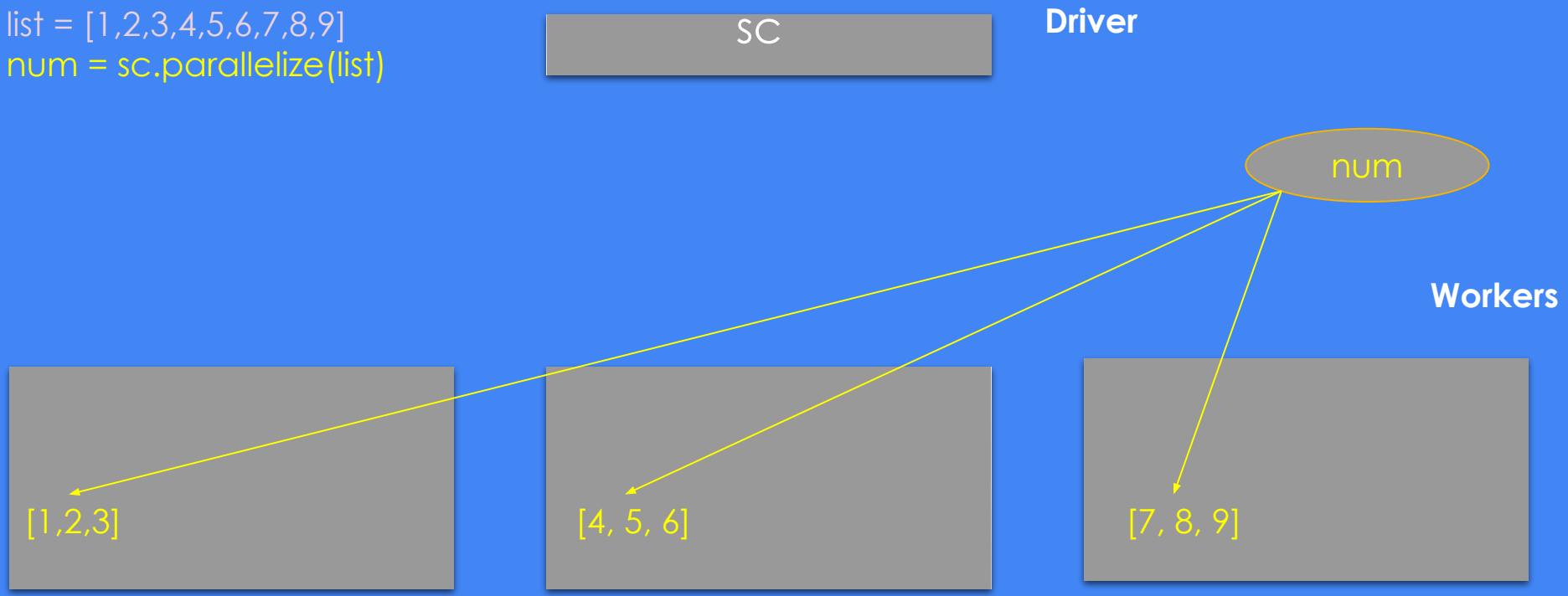
```
# Python List
list = [1,2,3,4,5,6,7,8,9]

# Create RDD
num = sc.parallelize(list)

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

# Collect result
even.collect()
```

DAG : Directed Acyclic Graph



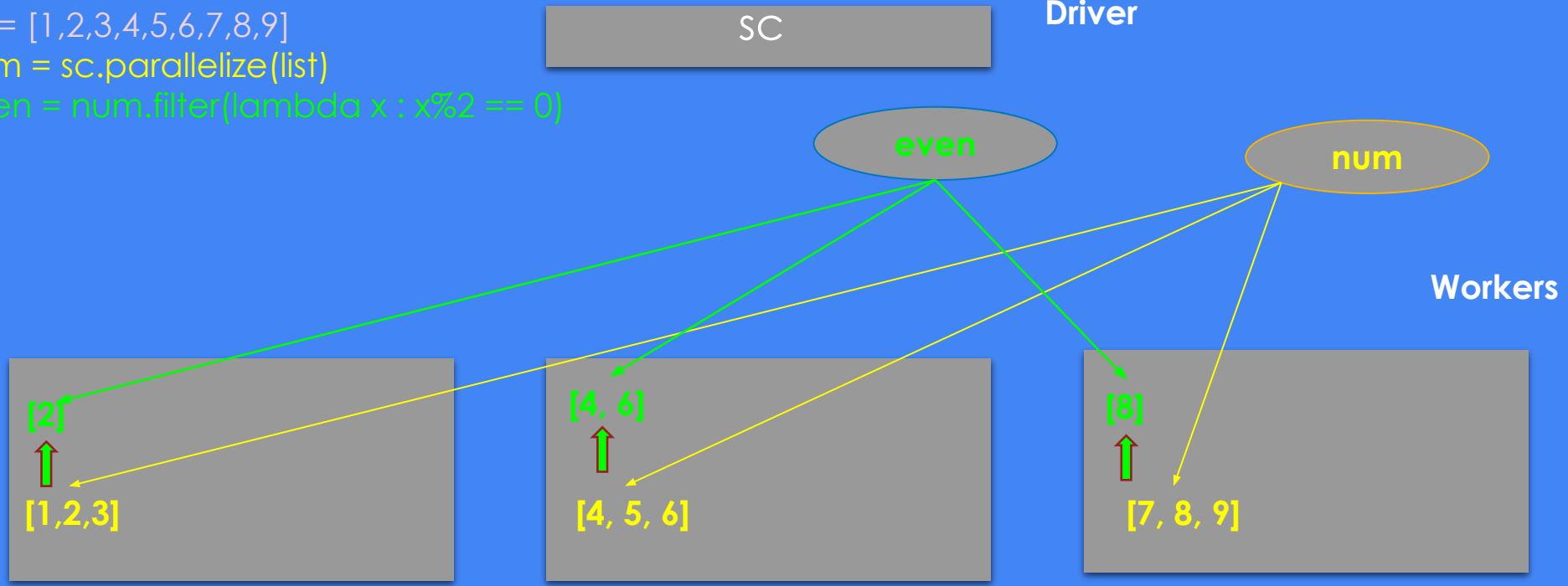
RDD : Resilient Distributed Dataset

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



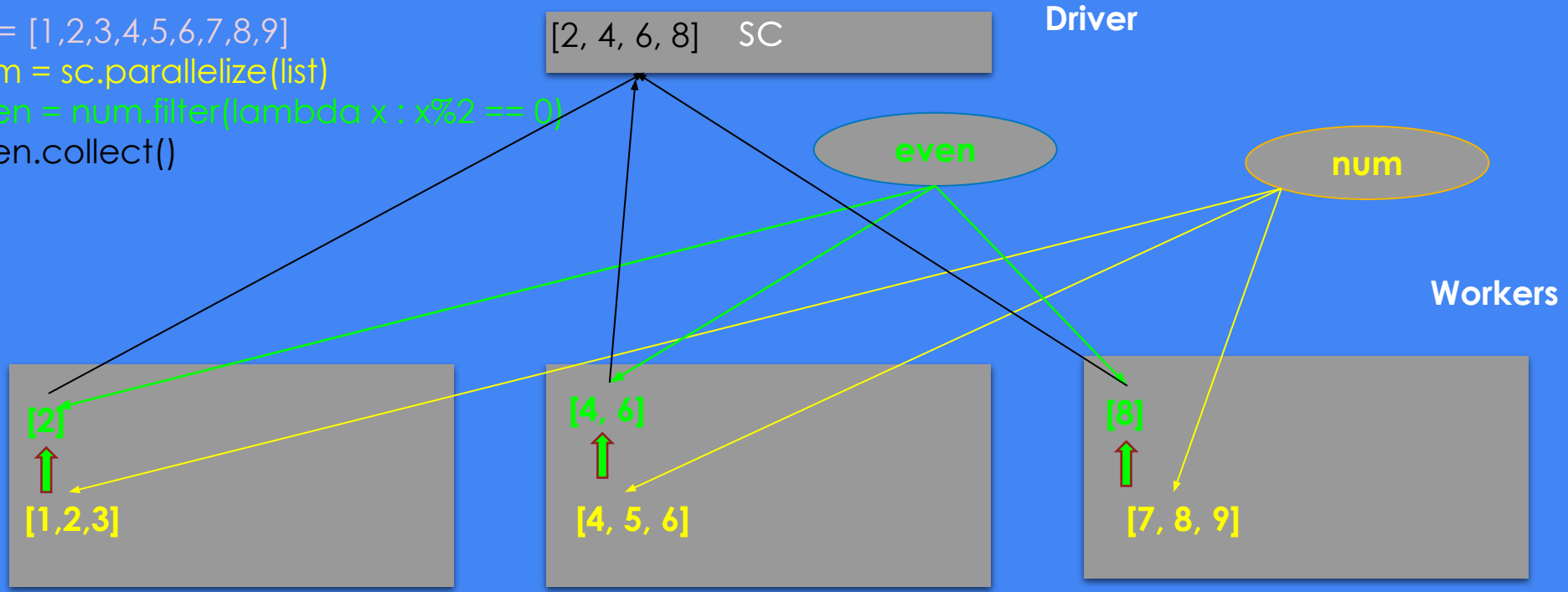
RDD : Resilient Distributed Dataset

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

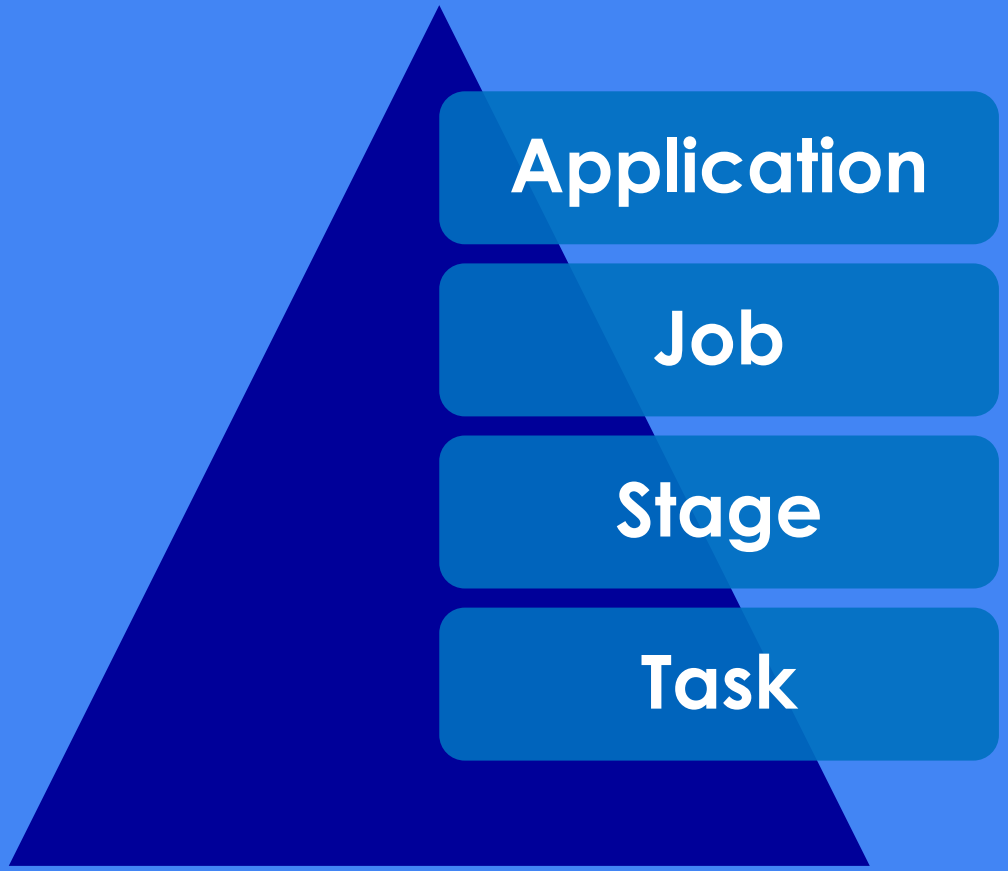


RDD : Resilient Distributed Dataset

```
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

Photo by Joshua Reddekopp on Unsplash

Questions?

- LinkedIn: https://www.linkedin.com/in/neerajbhadani/
- Medium: https://medium.com/@bhadani.neeraj.08

Thank You!

- LinkedIn: https://www.linkedin.com/in/neerajbhadani/
- Medium: https://medium.com/@bhadani.neeraj.08