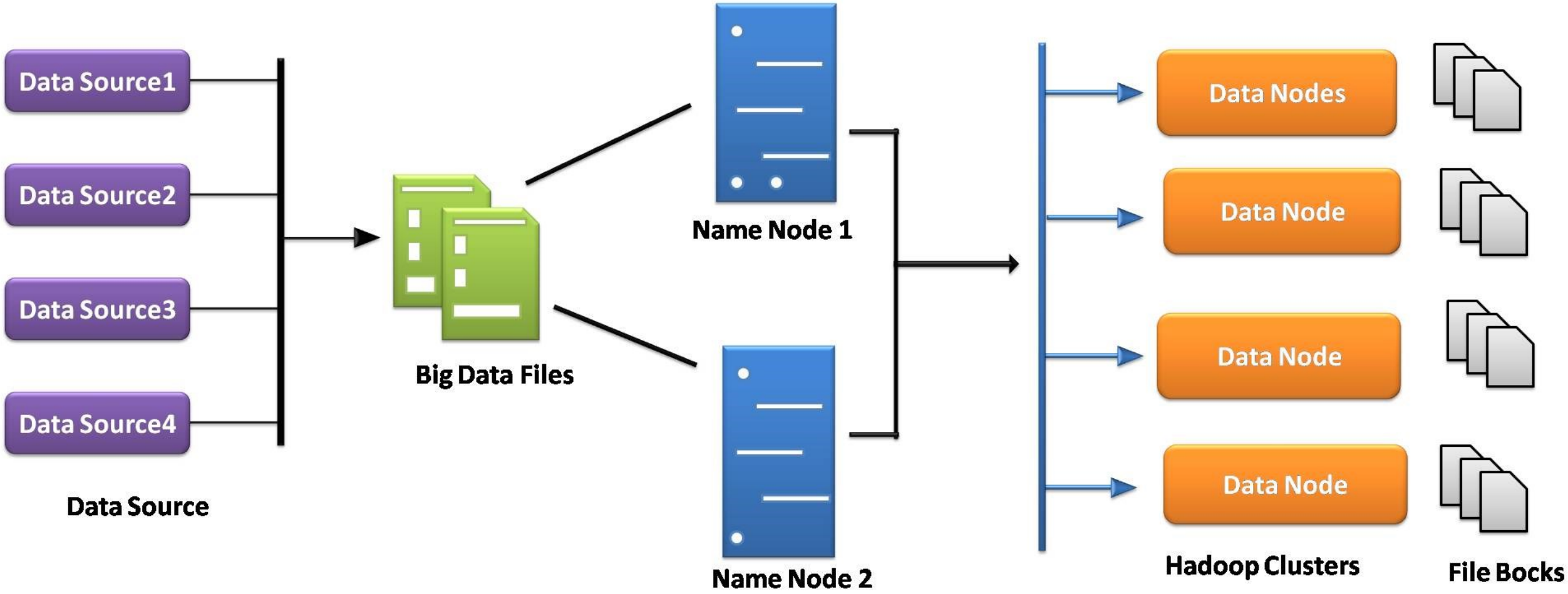
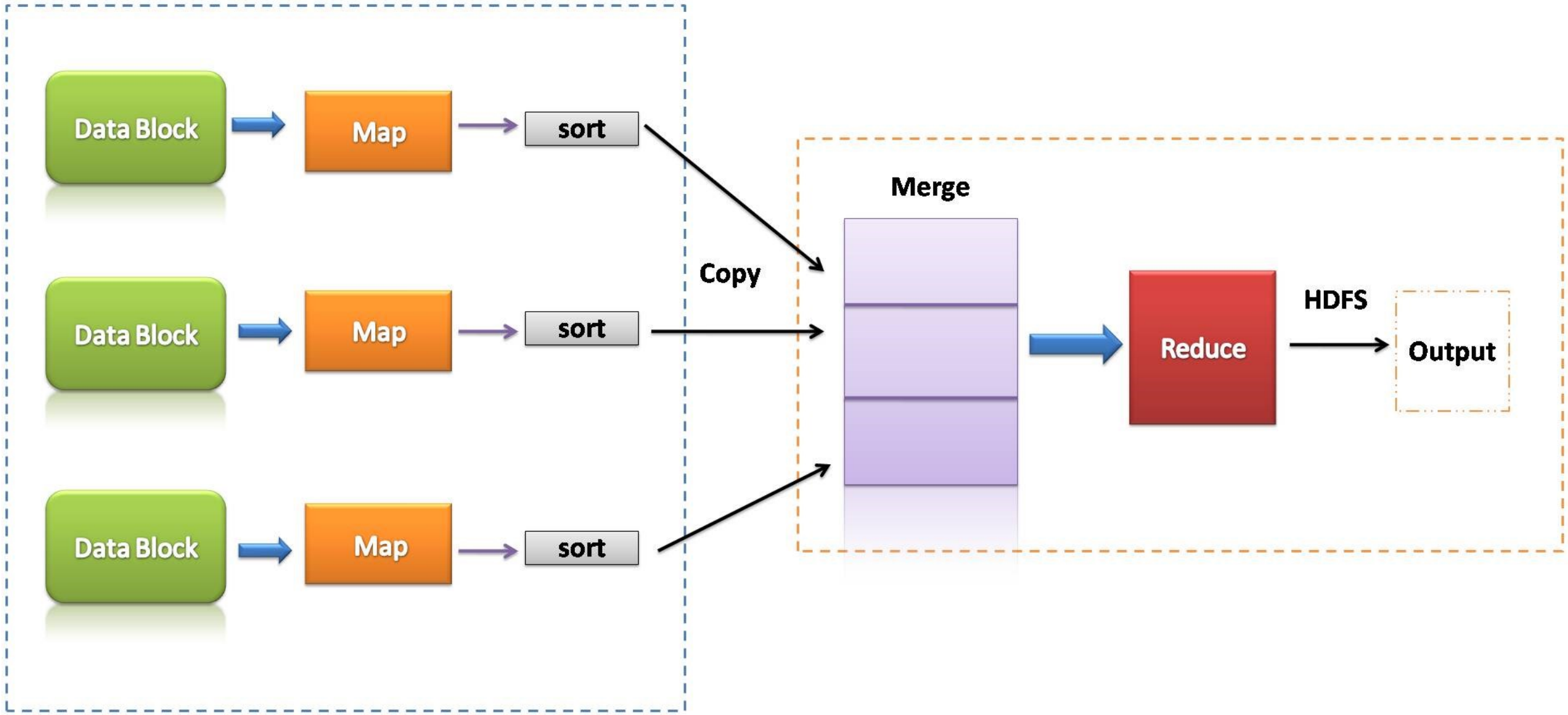
Hadoop System Architecture




Data Block
Map
sort
Data Block
Map
sort
Data Block
Map
sort
Copy
Merge
Reduce
HDFS
Output

Recap

- Hadoop data types
- Anatomy of a Hadoop job
- Hadoop jobs, end to end
- Software development workflow

MapReduce: Recap

- Programmers must specify:

  map (k, v) → <k', v'>*

  reduce (k', v') → <k', v'>*

  - All values with the same key are reduced together

- Optionally, also:

  partition (k', number of partitions) → partition for k'

  - Often a simple hash of the key, e.g., hash(k') mod n
  - Divides up key space for parallel reduce operations

  combine (k', v') → <k', v'>*

  - Mini-reducers that run in memory after the map phase
  - Used as an optimization to reduce network traffic

- The execution framework handles everything else…

```
docker run -it -d --name hadoop-local -p 9864:9864 -p 9870:9870 -p 8021:8021 -P --hostname localhost -v D:\:/mnt/d nyubigdata/hadoop-single-node:0.1.0
```

Java Hadoop

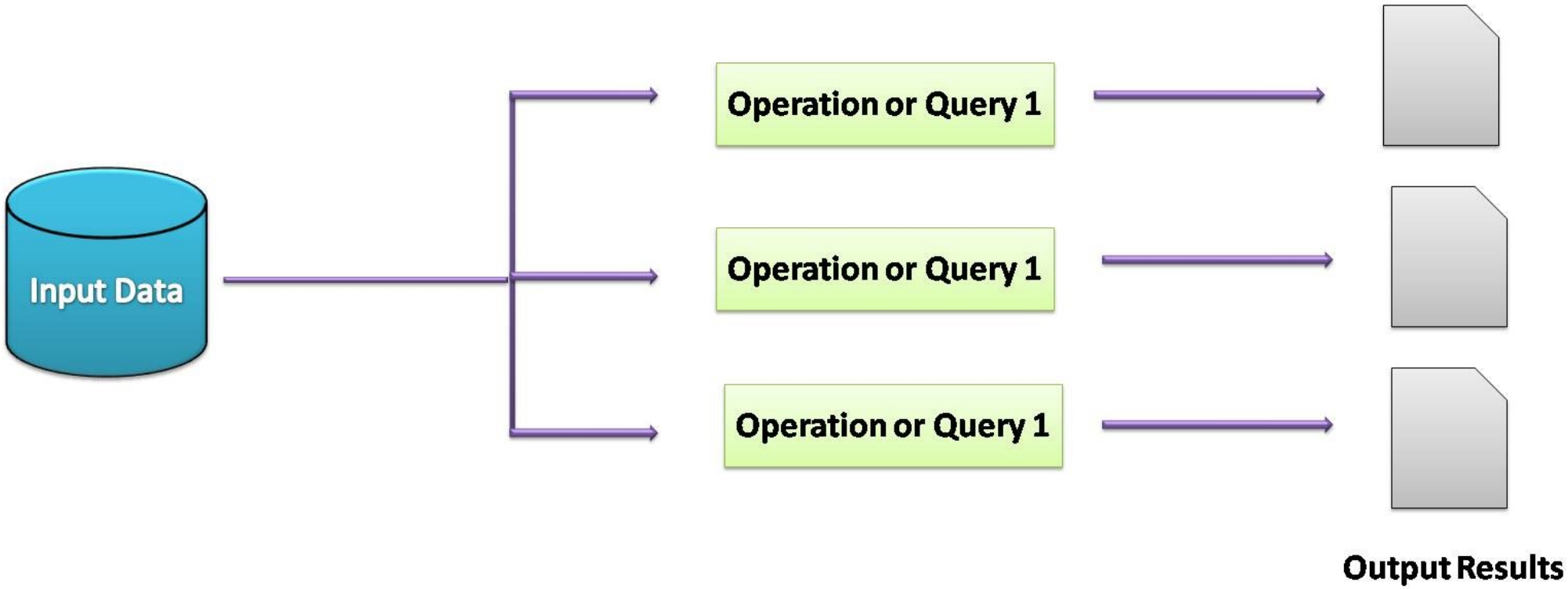
Hadoop Streaming

Python Hadoop

PyDoop

Spark

MapReduce Uses Disk I/O Operations



Apache Spark Uses In-Memory