CSC 369: Distributed Computing

Alex Dekhtyar

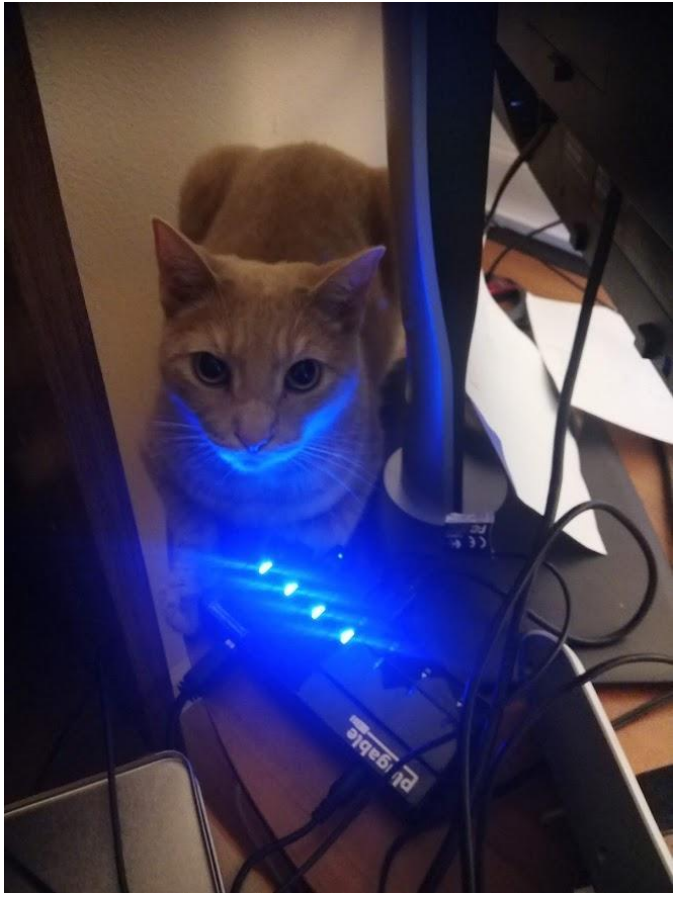
May The Fourth

Day 13: MapReduce→ Hadoop

May The Fourth Be With You


Petrified
Wood!

Housekeeping: Labs and Grading



MapReduce...What is it Good for?

MapReduce

$Map: K \times V \rightarrow \{K' \times V'\}$

$Reduce\ K \times (V)^* \rightarrow K \times (V)^*$

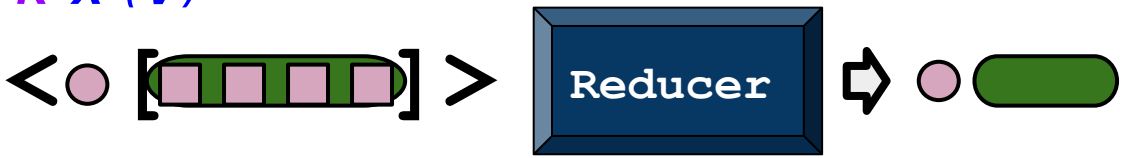
MapReduce

Map: K x V → {K' x V'}

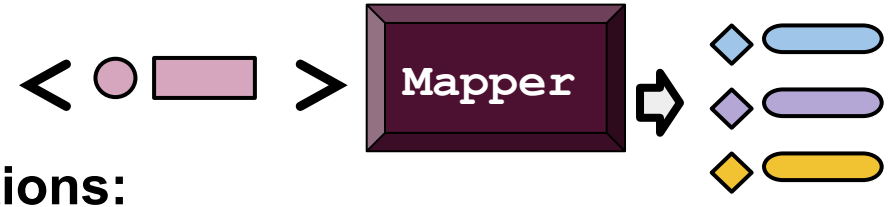
Reduce K x (V) → K x (V)**

Map: K x V → {K' x V'}

Reduce K x (V) → K x (V)**



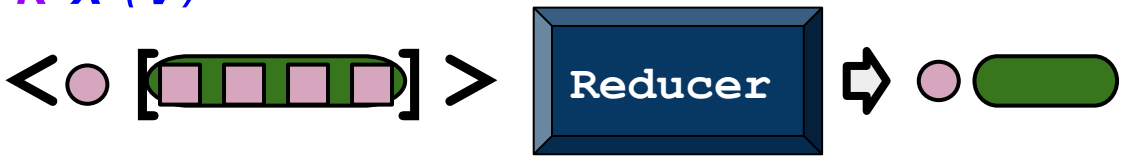
Map: K x V → {K' x V'}



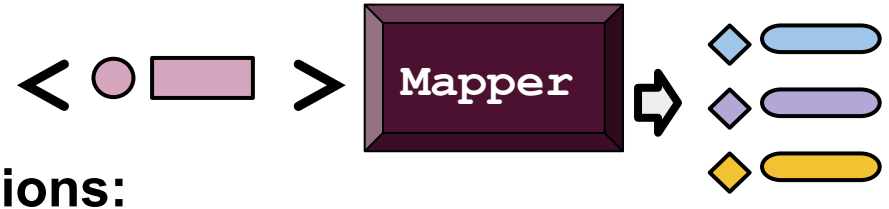
Record-by-record operations:

Filter/Selection
Projection

Reduce K x (V) → K x (V)**



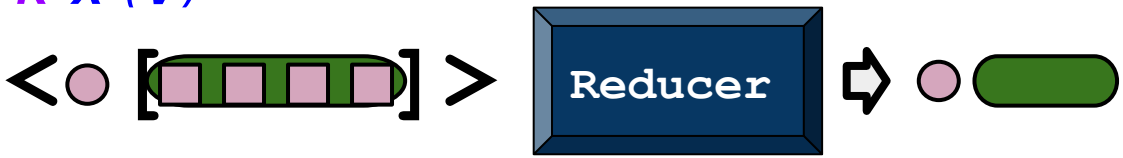
Map: K x V → {K' x V'}



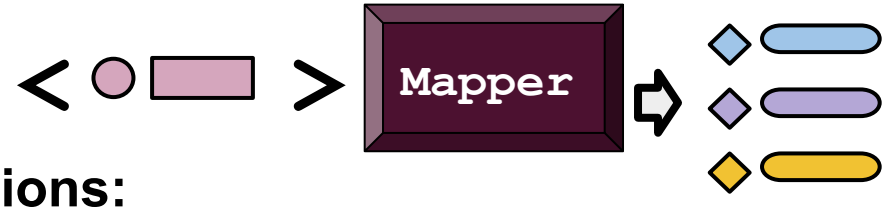
Record-by-record operations:

Filter/Selection
Projection

Reduce K x (V) → K x (V)**



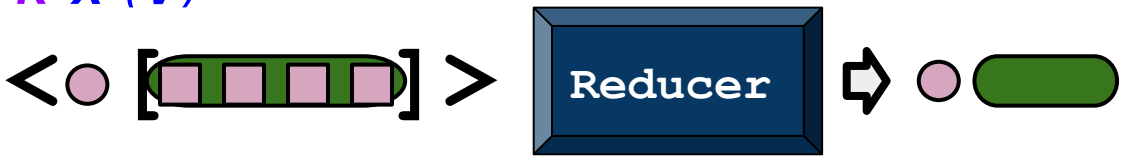
Map: K x V → {K' x V'}



Record-by-record operations:

Filter/Selection
Projection

Reduce K x (V) → K x (V)**

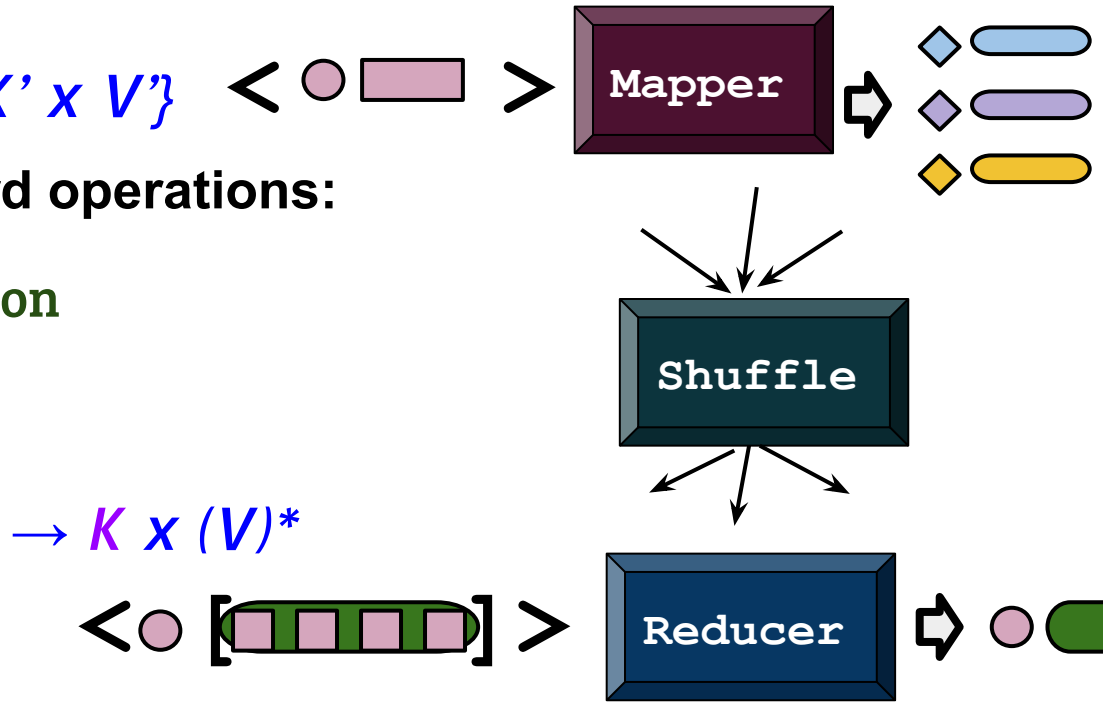


Aggregation

Map: K x V → {K' x V'}

Record-by-record operations:

Filter/Selection
Projection



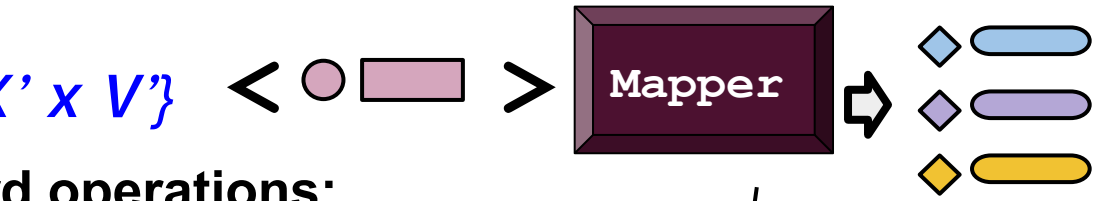
Reduce K x (V) → K x (V)**

Aggregation

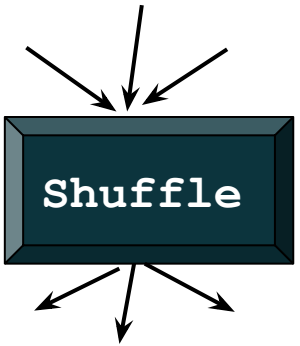
Map: K x V → {K' x V'}

Record-by-record operations:

Filter/Selection
Projection

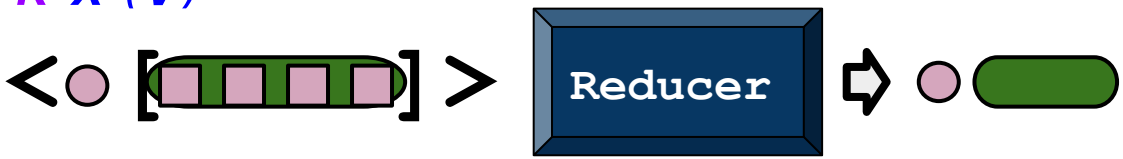


Grouping

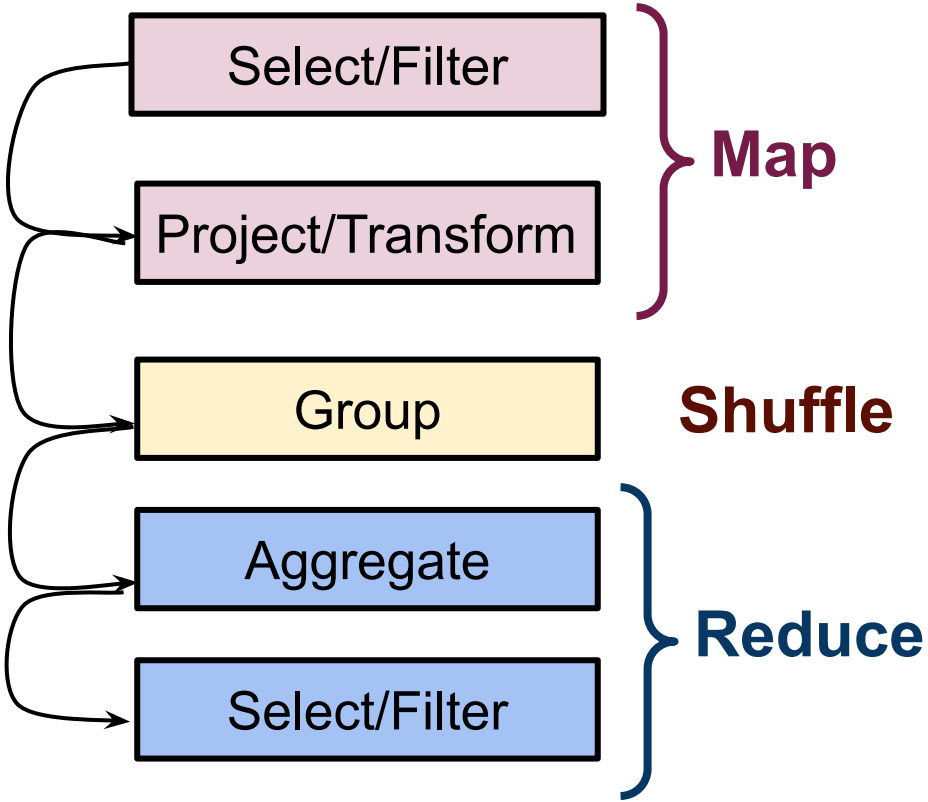


Reduce K x (V) → K x (V)**

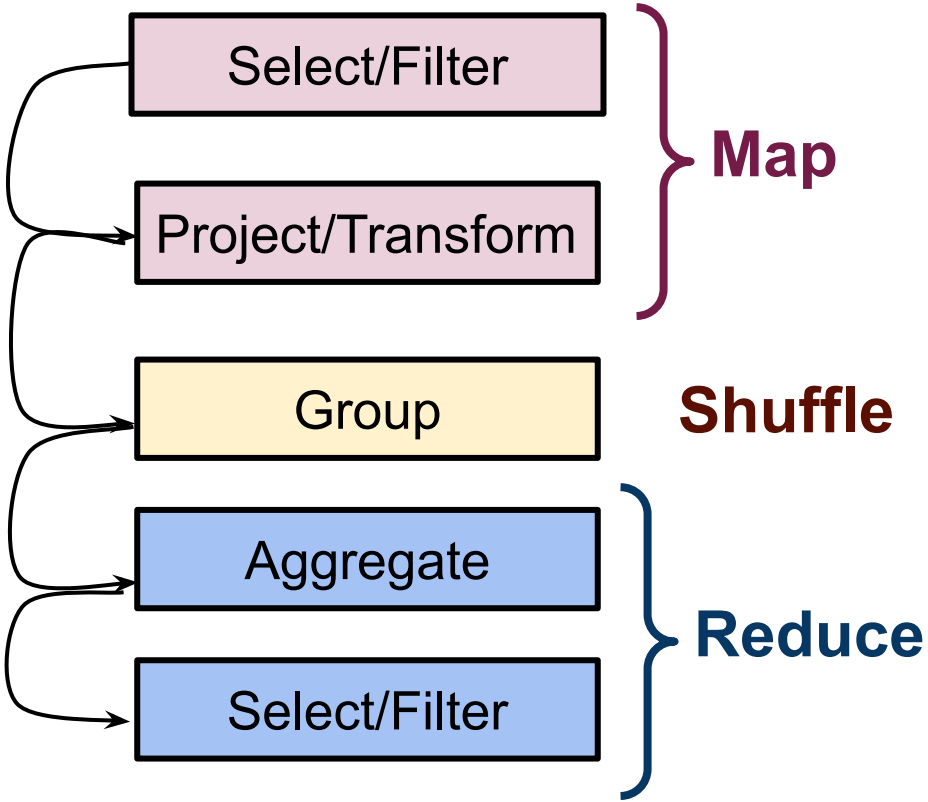
Aggregation



Data Processing Pipeline



Data Processing Pipeline



For each day from April 15 to April 30 find the ratio between the number of daily new deaths and the number of daily new cases.

Find all weeks during which the total number of new positive COVID-19 cases in California and Florida exceeded 10,000, and report the largest daily increase in cases for each such week

Our First Hadoop Steps Will be on such pipelines

Select/Filter

Project/Transform

Group

Aggregate

Select/Filter

Resource Manager (YARN)

Distributed Data Store (HDFS)

MapReduce

Other

ed

Resource Manager (YARN)

Distributed Data Store (HDFS)

(This is a simplified view)

Hadoop v1.0

MapReduce
Data Processing
& Resource Management

HDFS
Distributed File Storage

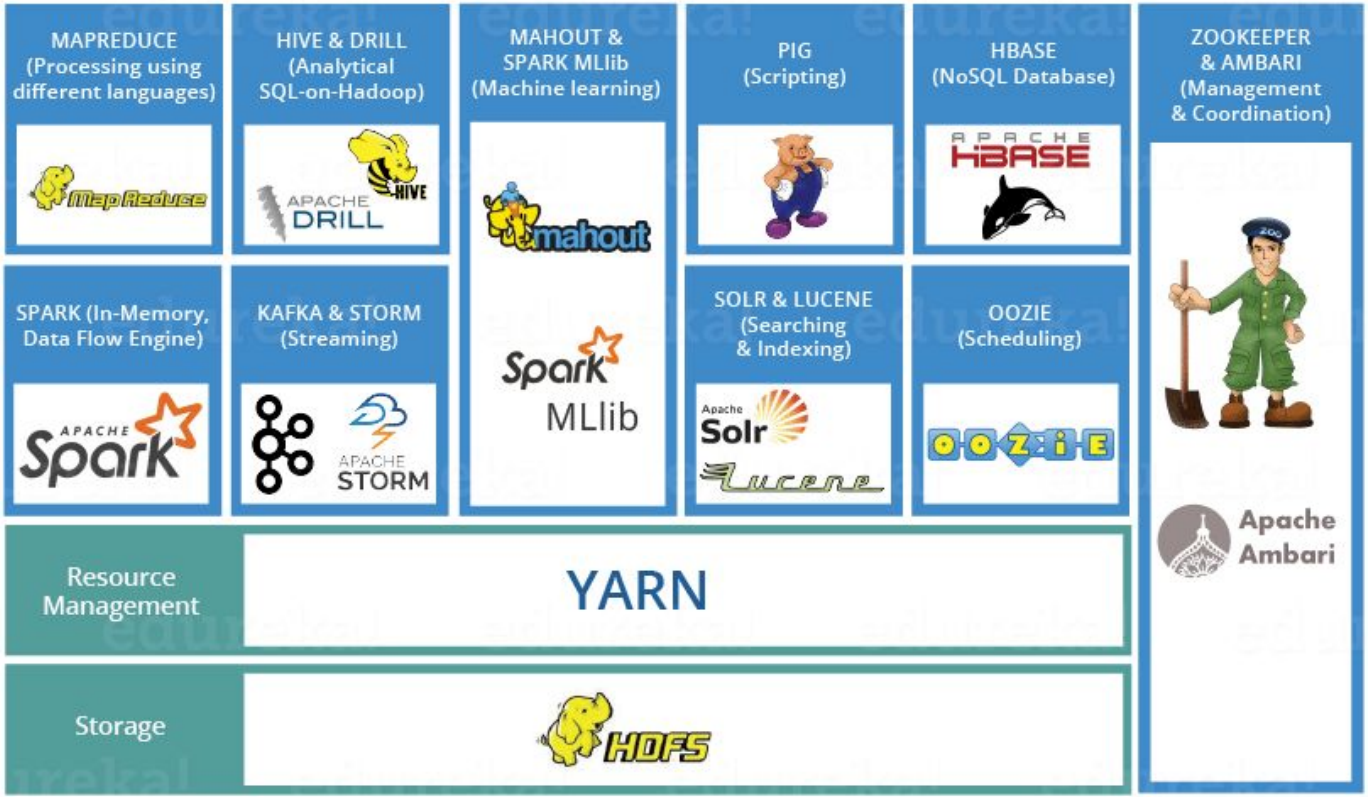
Hadoop v2.0

MapReduce

Other Data Processing Frameworks

YARN
Resource Management

HDFS
Distributed File Storage


MAPREDUCE (Processing using different languages)
HIVE & DRILL (Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib (Machine learning)
PIG (Scripting)
HBASE (NoSQL Database)
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM (Streaming)
SOLR & LUCENE (Searching & Indexing)
OOZIE (Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Sqoop
Structured Data
Apache Ambari

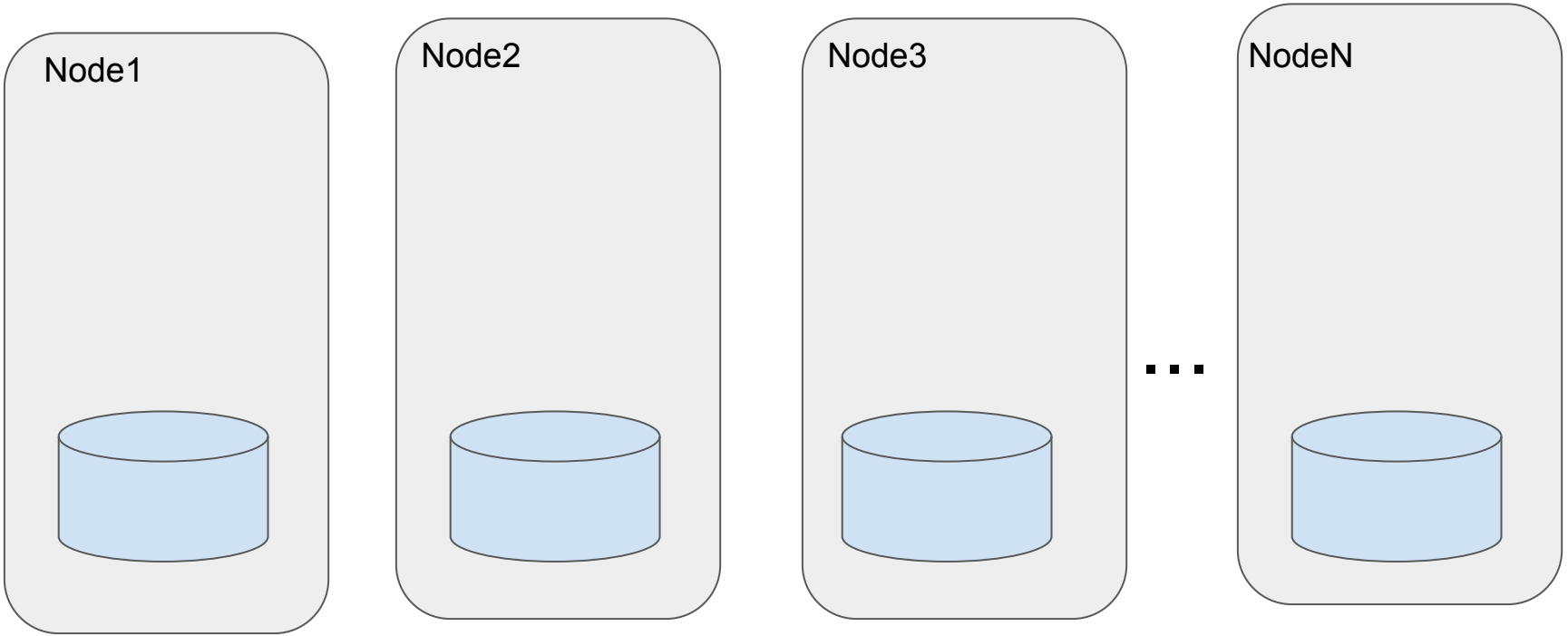
At first

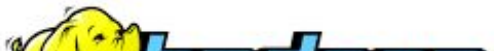
MapReduce

Resource Manager (YARN)

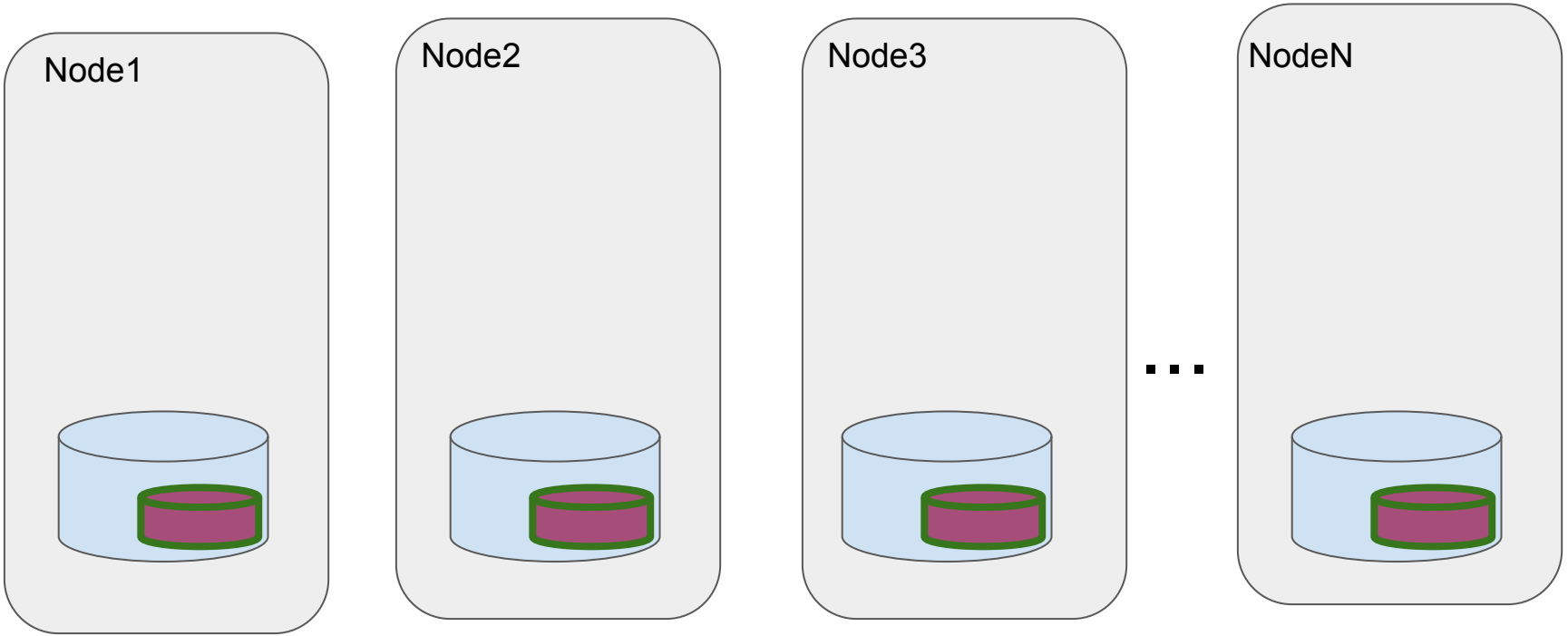
Distributed Data Store (HDFS)

Distributed Data Store (HDFS)

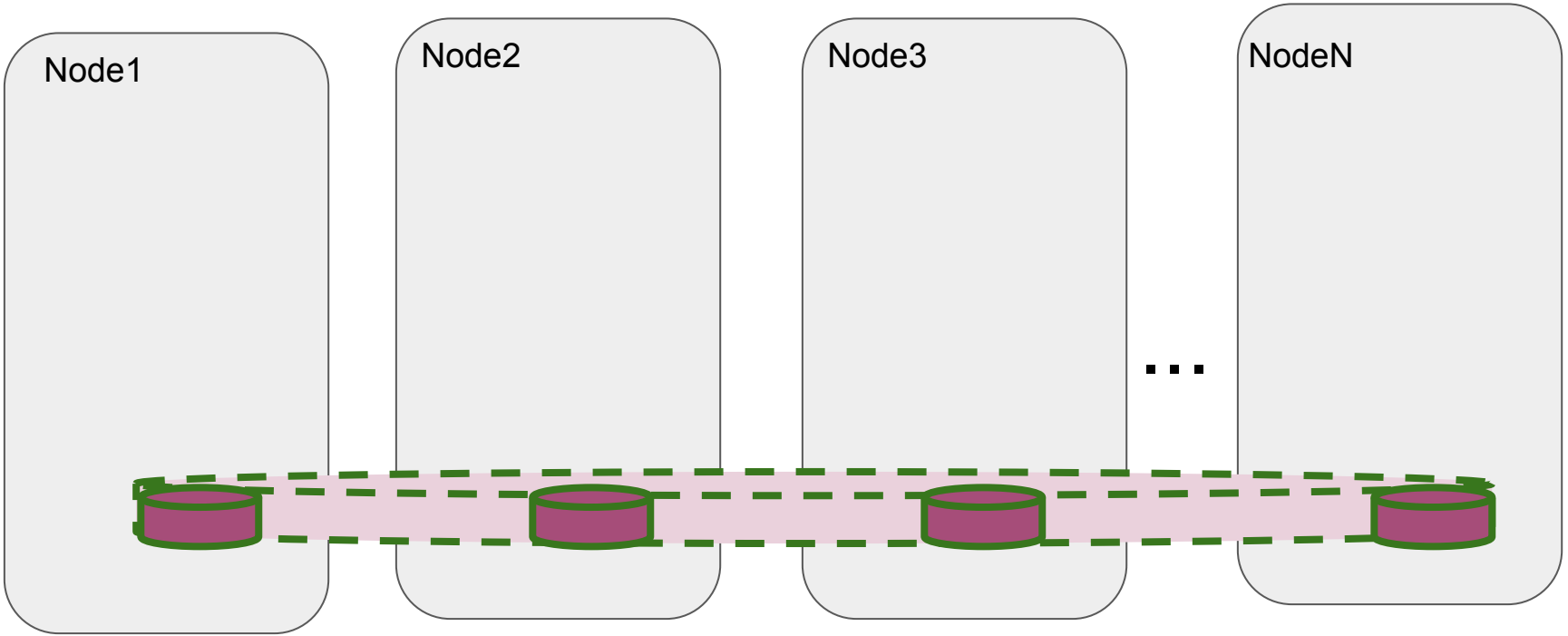




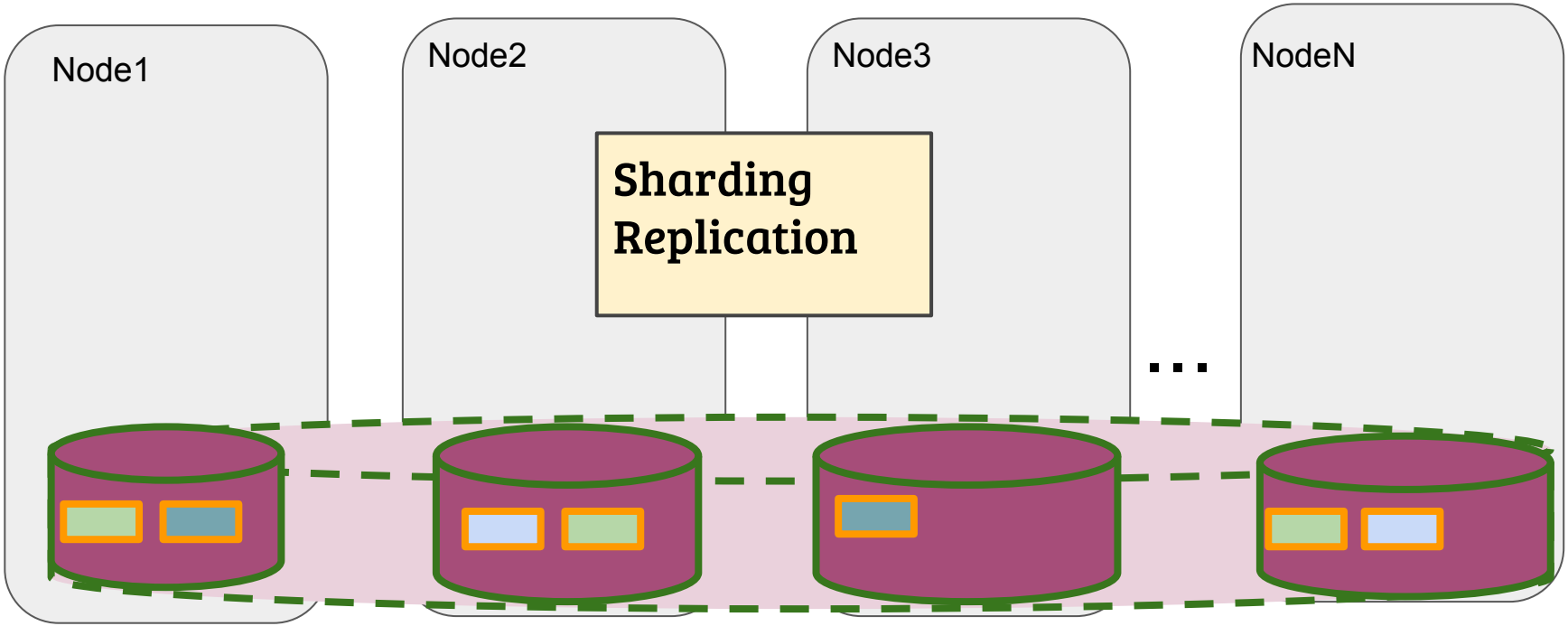
Distributed Data Store (HDFS)



Distributed Data Store (HDFS)



Distributed Data Store (HDFS)



Distributed Data Store (HDFS)

Command line interface

```
$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - dekhtyar hdfs          0 2020-04-02 12:00 .sparkStaging
drwx------   - dekhtyar hdfs          0 2020-05-04 04:05 .staging
drwxr-xr-x   - dekhtyar hdfs          0 2020-05-04 04:05 test
```

```
hdfs dfs -<command> <parameters>
```

Distributed Data Store (HDFS)

Command line interface

```
$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - dekhtyar hdfs          0 2020-04-02 12:00 .sparkStaging
drwx------   - dekhtyar hdfs          0 2020-05-04 04:05 .staging
drwxr-xr-x   - dekhtyar hdfs          0 2020-05-04 04:05 test
```

```
hdfs dfs -<command> <parameters>
```

See handout for hdfs commands

At first

MapReduce

Resource Manager (YARN)

Distributed Data Store (HDFS)

MapReduce

Today: Java Program DEMO

```
Job runner (main)
Mapper
Reducer
```

Wednesday: Details

```
Input
```