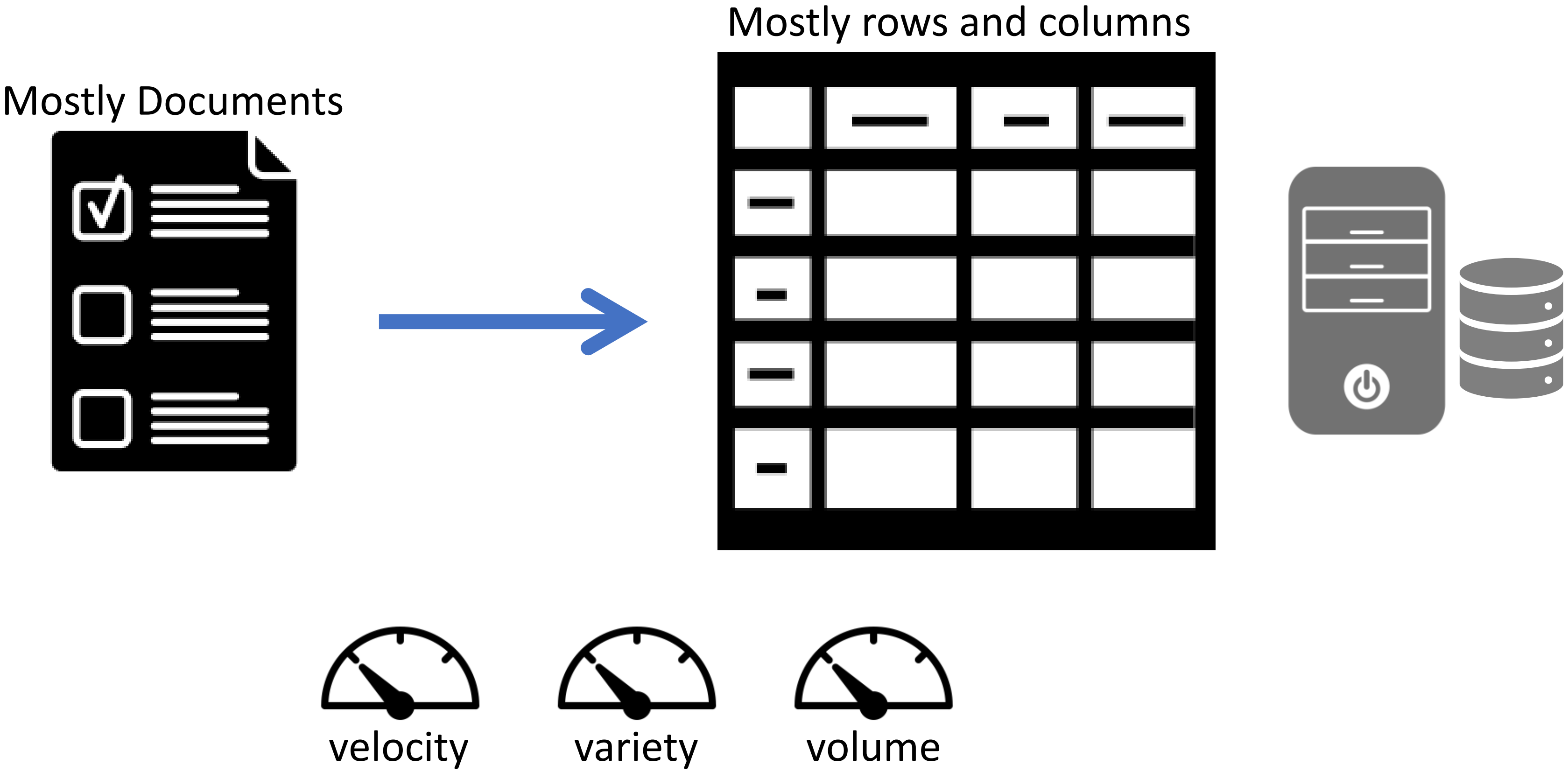
Hadoop

Old data – Pre web



Mostly Documents

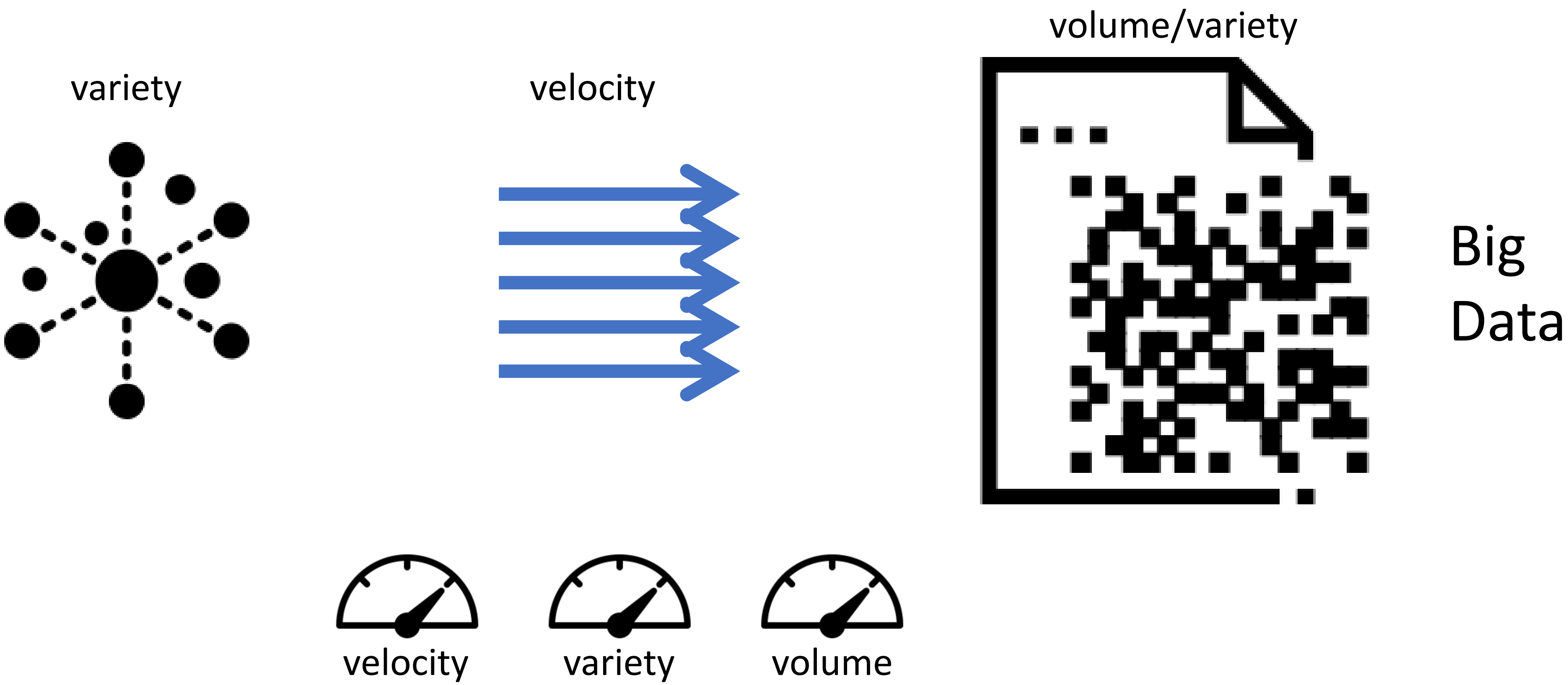
Mostly rows and columns

velocity

variety

volume

New data – Post web



variety

velocity

volume/variety

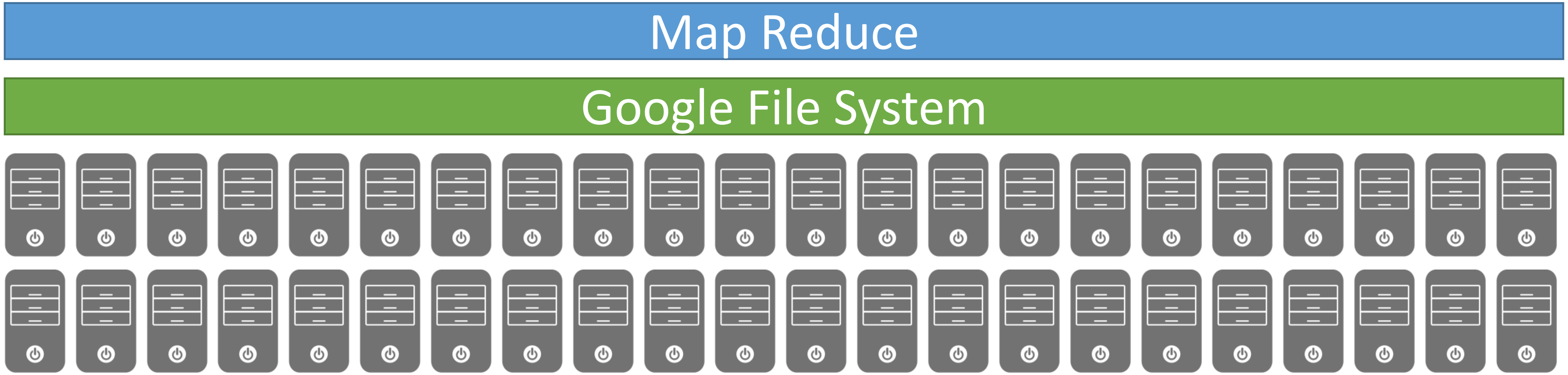
Big
Data

velocity

variety

volume

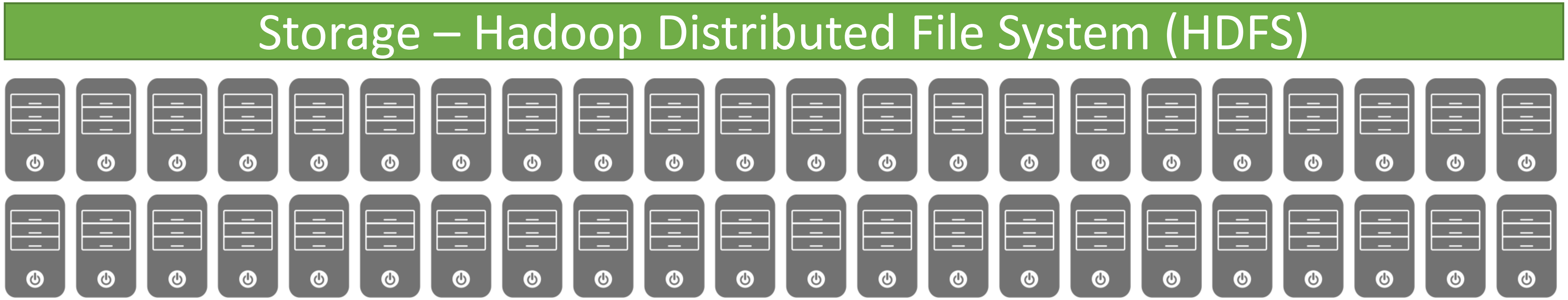
Google



Map Reduce

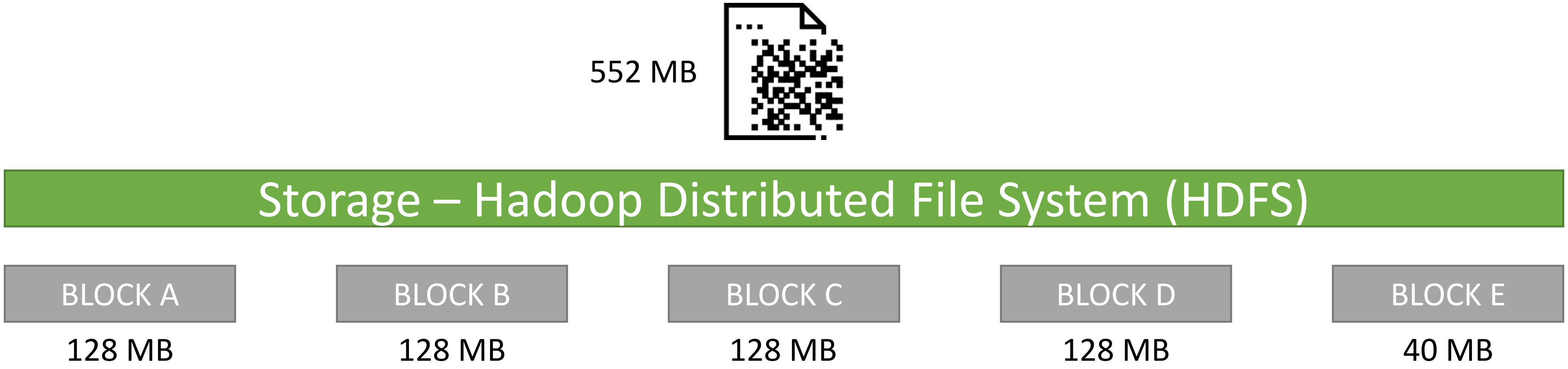
Google File System

Open version - Hadoop



Storage – Hadoop Distributed File System (HDFS)

Stored in blocks

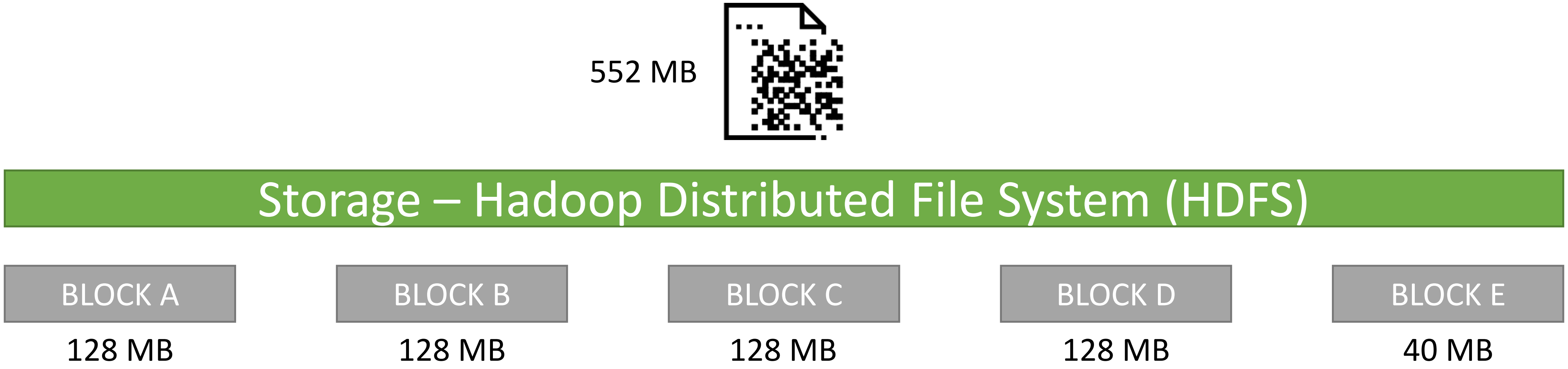


552 MB

Storage – Hadoop Distributed File System (HDFS)

BLOCK A	BLOCK B	BLOCK C	BLOCK D	BLOCK E
128 MB	128 MB	128 MB	128 MB	40 MB

Stored in blocks



552 MB

Storage – Hadoop Distributed File System (HDFS)

BLOCK A

128 MB

BLOCK B

128 MB

BLOCK C

128 MB

BLOCK D

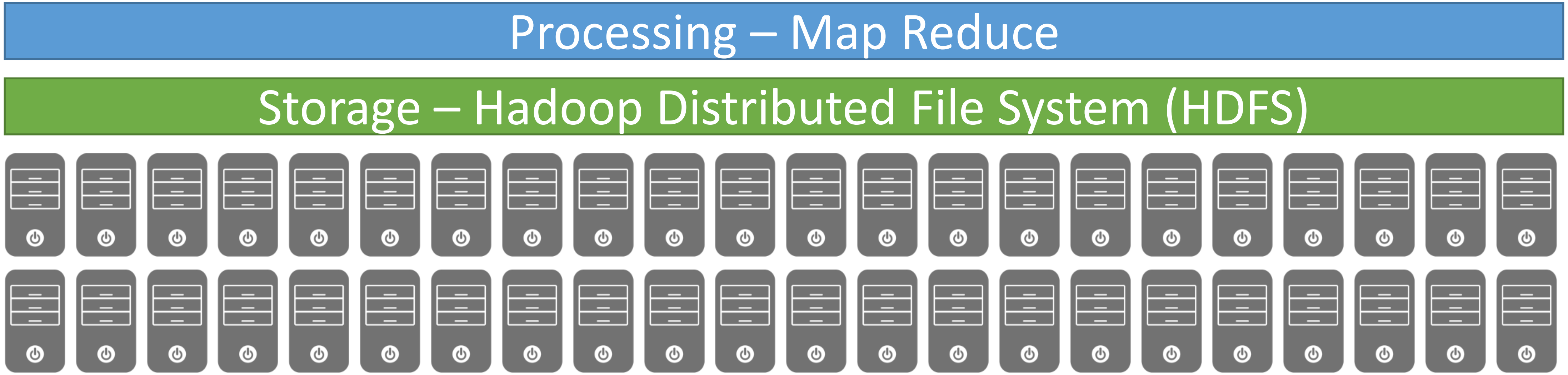
128 MB

BLOCK E

40 MB

Redundancy built in for blocks

Open version - Hadoop



Processing – Map Reduce

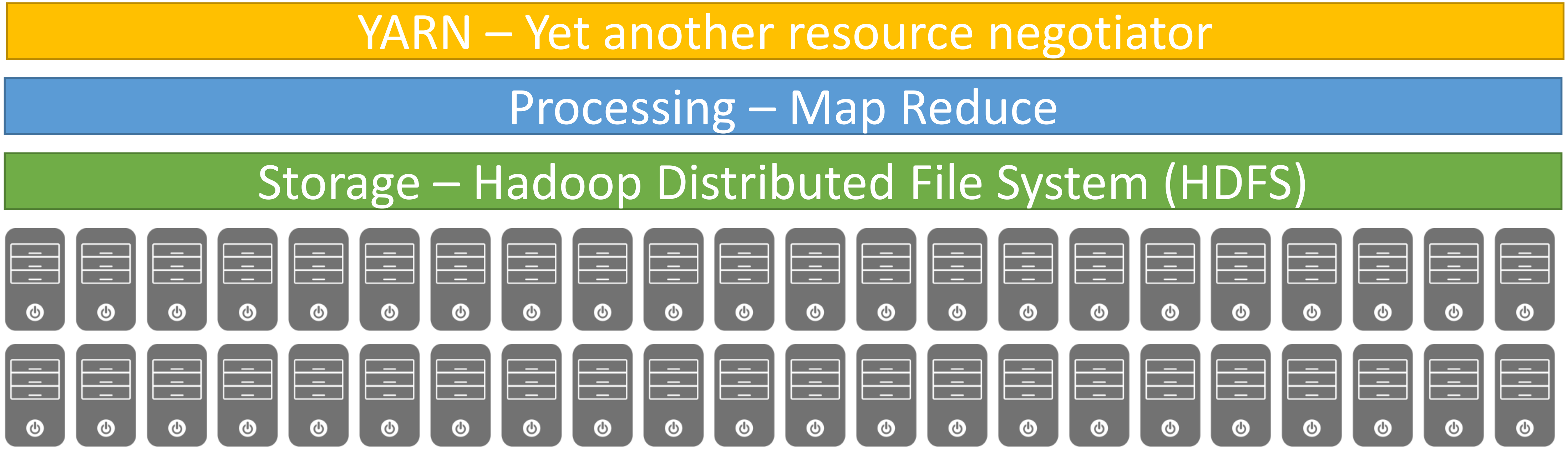
Storage – Hadoop Distributed File System (HDFS)

Open version - Hadoop

Processing – Map Reduce

INPUT	SPLIT	MAP	SORT	REDUCE
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness	It was the best of times	it, 1, was, 1, the, 1, best, 1, of, 1, times, 1	it it it it …	{'it': 4, 'was': 4, 'the': 4, 'of': 4, 'times,': 2, 'age': 2, 'best': 1, 'worst': 1, 'wisdom,': 1, 'foolishness': 1}
	it was the worst of times	it, 1, was, 1, the, 1, worst, 1, of, 1, times, 1		
	it was the age of wisdom	It, 1, was, 1, the, 1, age, 1, of, 1, wisdom, 1		
	it was the age of foolishness	it, 1, was, 1, the, 1, age, 1, of, 1, foolishness, 1		

Open version - Hadoop



YARN – Yet another resource negotiator

Processing – Map Reduce

Storage – Hadoop Distributed File System (HDFS)

What is Hadoop?

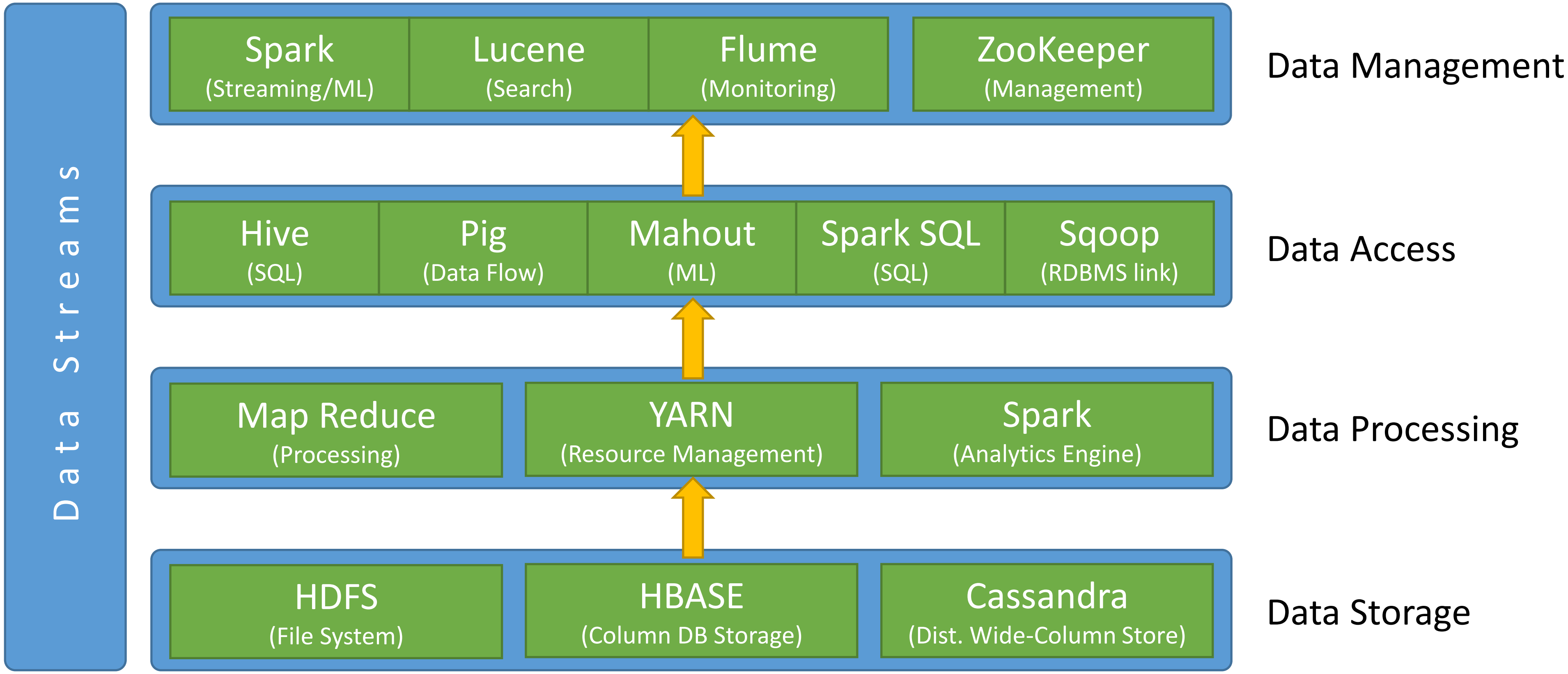
An open source software platform for distributed storage and distributed processing of very large data sets on computer clusters build from commodity hardware

Democratized big data

Hadoop democratized computing power and made it possible for companies to analyze and query big data open source software and inexpensive, off-the-shelf hardware.

A viable alternative to the proprietary data warehouse (DW). Organizations could now store and process huge amounts of data, increased computing power, fault tolerance, flexibility in data management, and lower costs compared to DW’s, and greater scalability

Hadoop Ecosystem



Data Streams

Data Management

Spark (Streaming/ML)

Lucene (Search)

Flume (Monitoring)

ZooKeeper (Management)

Data Access

Hive (SQL)

Pig (Data Flow)

Mahout (ML)

Spark SQL (SQL)

Sqoop (RDBMS link)

Data Processing

Map Reduce (Processing)

YARN (Resource Management)

Spark (Analytics Engine)

Data Storage

HDFS (File System)

HBASE (Column DB Storage)

Cassandra (Dist. Wide-Column Store)

HADOOP – DOCKER

Images from Big Data Europe

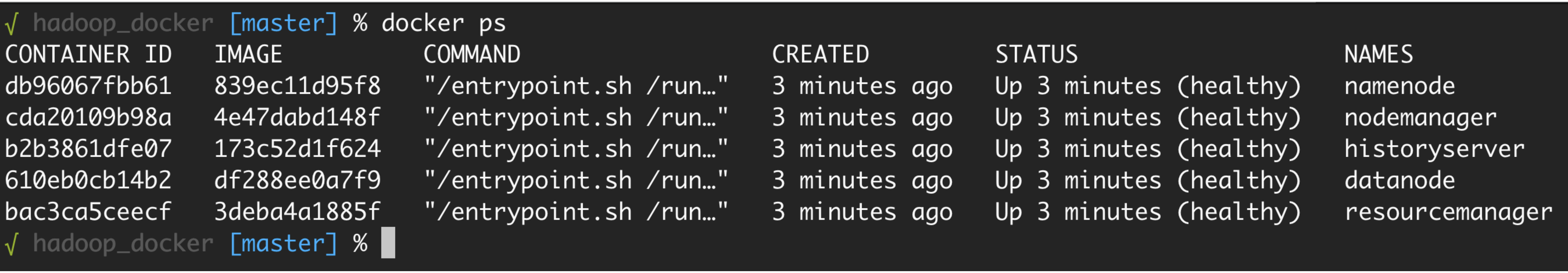
- GitHub project address
  - https://github.com/big-data-europe
- Repository
  - https://github.com/big-data-europe/docker-hadoop

CREATE CONTAINERS

Run docker compose file

```
$ docker-compose up
```

List containers

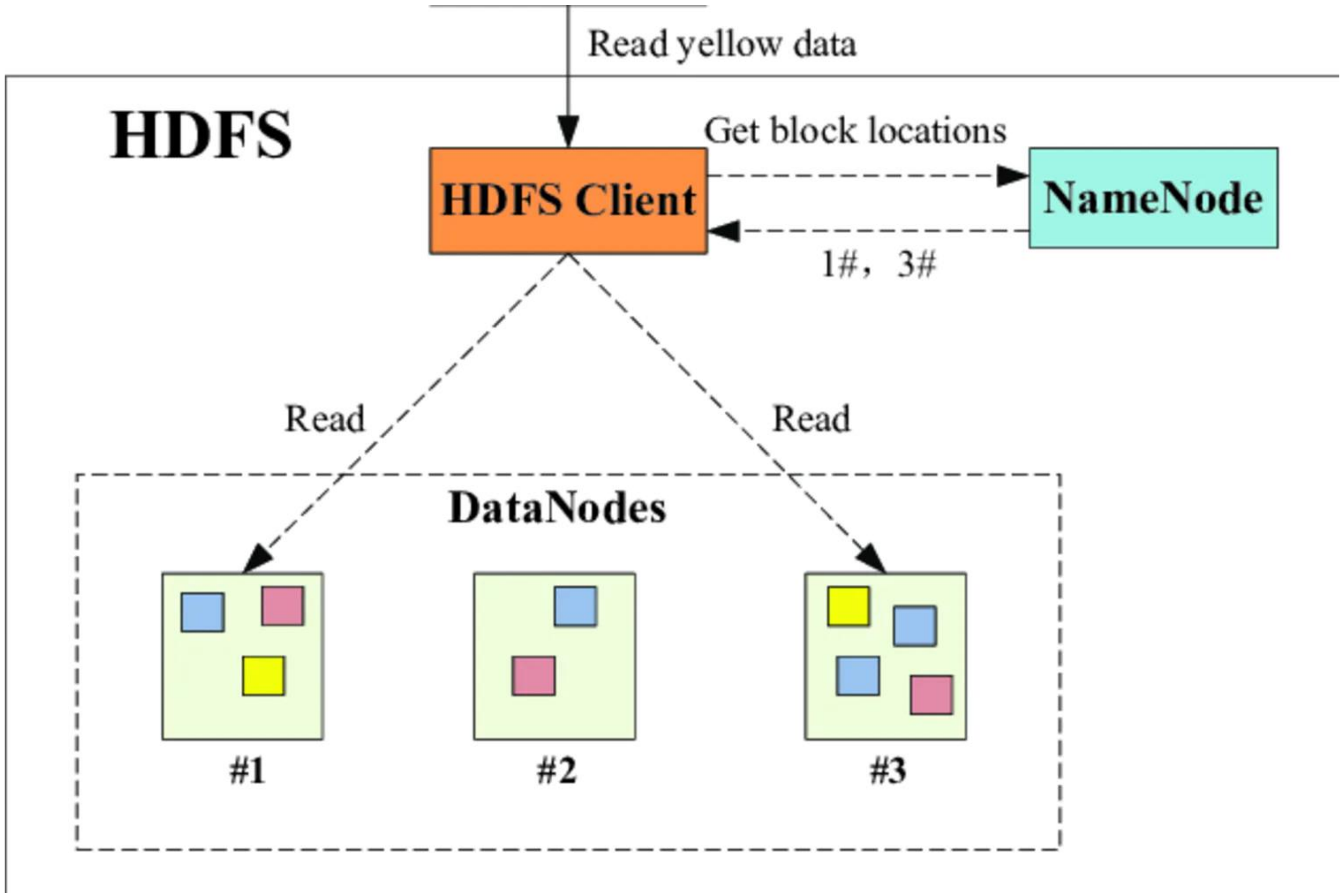


```
√ hadoop_docker [master] % docker ps
CONTAINER ID   IMAGE          COMMAND                  CREATED         STATUS                   NAMES
db96067fbb61   839ec11d95f8   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   namenode
cda20109b98a   4e47dabd148f   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   nodemanager
b2b3861dfe07   173c52d1f624   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   historyserver
610eb0cb14b2   df288ee0a7f9   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   datanode
bac3ca5ceecf   3deba4a1885f   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   resourcemanager
√ hadoop_docker [master] %
```

Containers

- Hadoop cluster with:
  - 1 HDFS namenode (or primary node to manage the secondary)
  - 3 secondary (datanodes)
  - 1 YARN resourcemanager
  - 1 historyserver
  - 1 nodemanager

Architecture



CURRENT STATUS

View current status

```
http://localhost:9870/
```

ENTER COINTAINER CREATE INPUT

Create word frequency input

Enter container namenode

```
$ docker exec -it namenode bash
```

Create two files

```
$ mkdir input
$ echo "it was the best of times it was the worst of times" >input/f1.txt
$ echo "it was the age of wisdom it was the age of foolishness" >input/f2.txt
```

Load to HDFS

```
# create the input directory on HDFS
$ hadoop fs -mkdir -p input

# to put the input files to all the datanodes on HDFS
$ hdfs dfs -put ./input/* input
```

WORD COUNTING PROGRAM

Counting word program

```
$ curl -L http://some.url --output some.file
```

```
$ curl -L
    https://repo1.maven.org/maven2/org/apache/
    hadoop/hadoop-mapreduce-examples/2.7.1/
    hadoop-mapreduce-examples-2.7.1-sources.jar
    --output
    hadoop-mapreduce-examples-2.7.1-sources.jar
```

Word counting program

```
$ curl -L http://some.url --output some.file
```

https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-mapreduce-examples/2.7.1/hadoop-mapreduce-examples-2.7.1-sources.jar

RUN PROGRAM

Run program

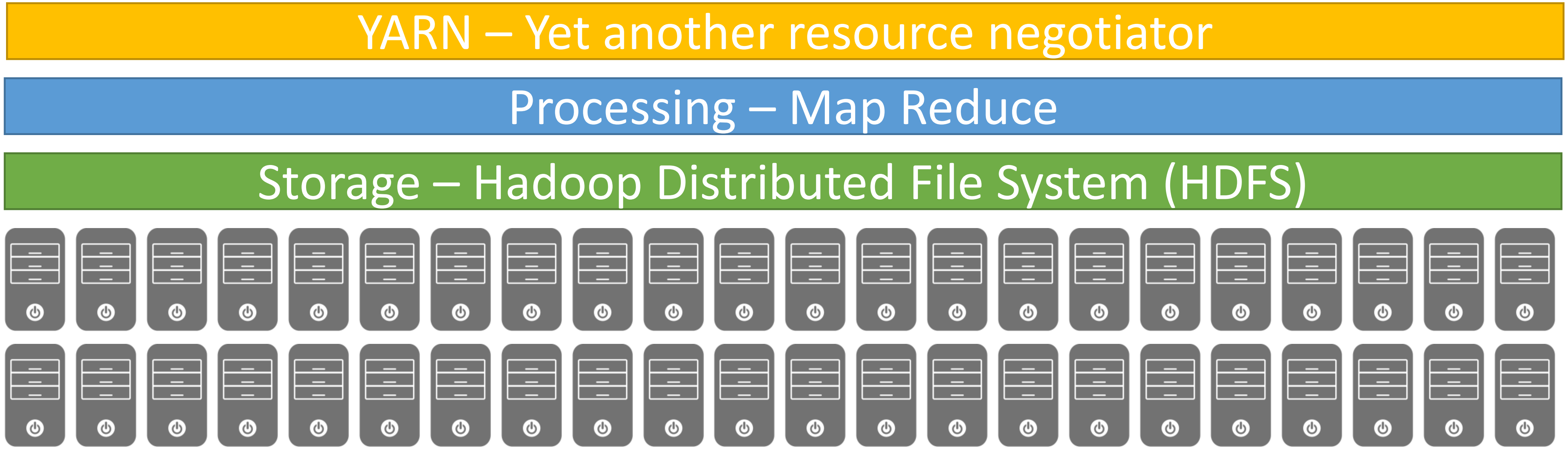
```
$ hadoop jar
    hadoop-mapreduce-examples-2.7.1-sources.jar
    org.apache.hadoop.examples.WordCount
    input
    output
```

View results

```
$ hdfs dfs -cat output/part-r-00000
```

OVERVIEW

Open version - Hadoop

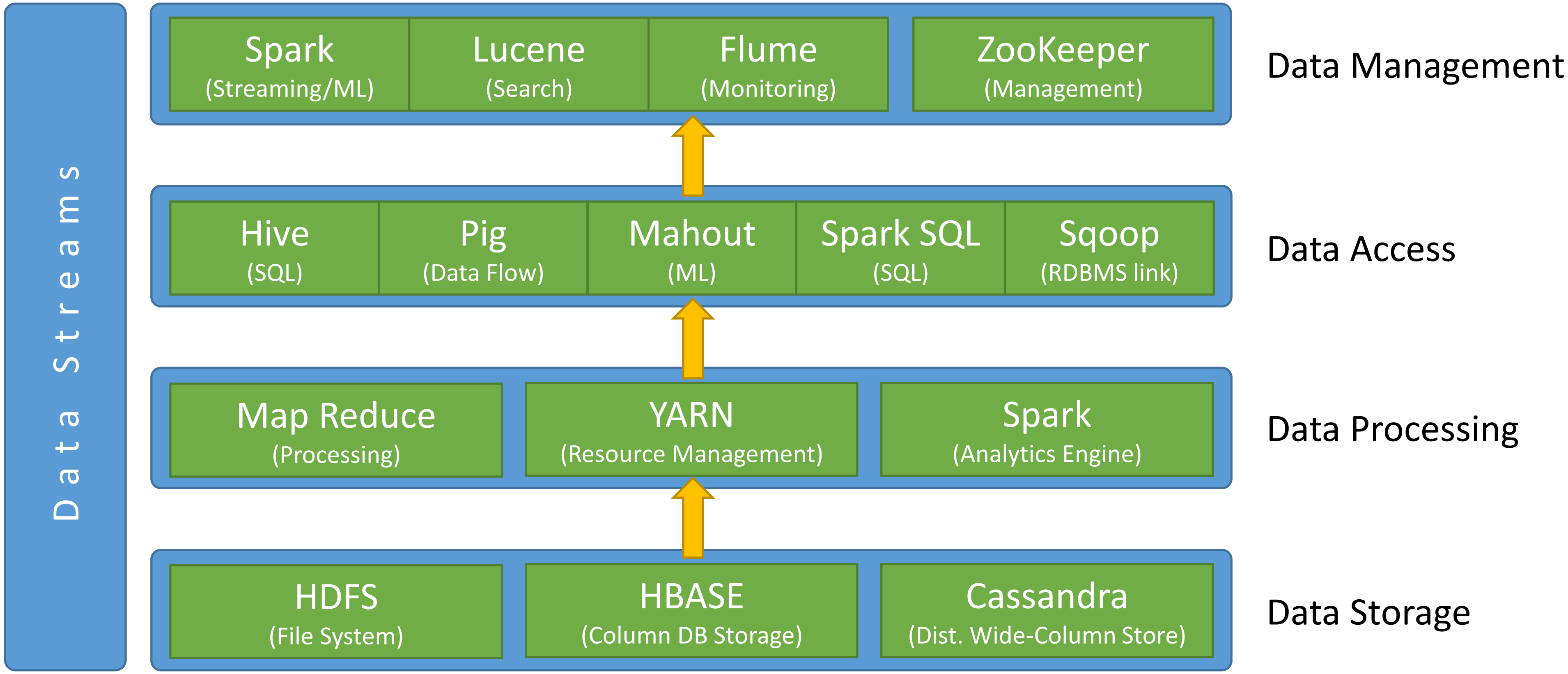


YARN – Yet another resource negotiator

Processing – Map Reduce

Storage – Hadoop Distributed File System (HDFS)

Hadoop Ecosystem



Your turn:

Download Moby Dick by Herman Melville

From Project Gutenberg

```
https://www.gutenberg.org/
```