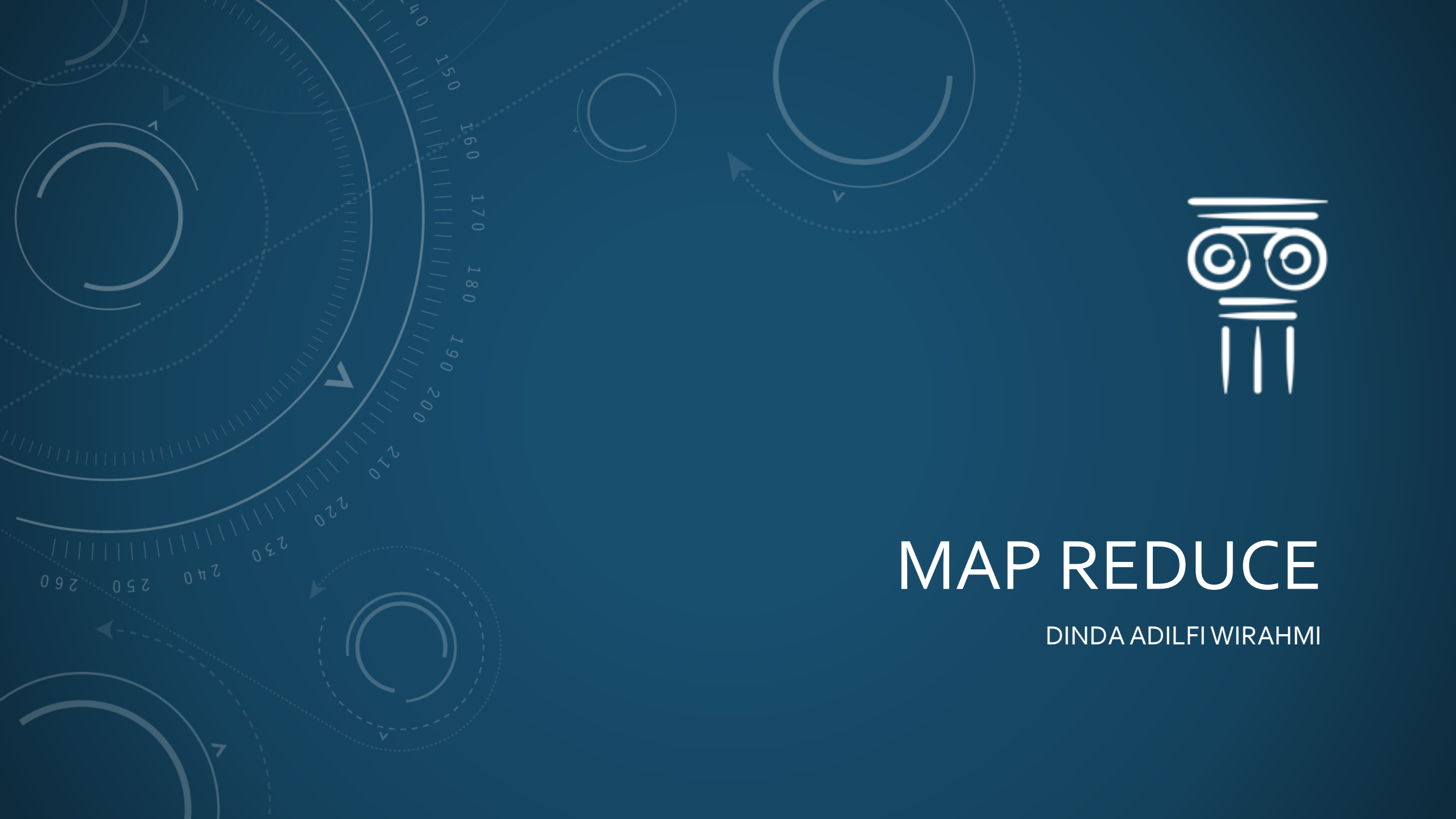

MAP REDUCE

DINDA ADILFI WIRAHMI

MAP REDUCE

MapReduce is a processing technique and a program model for distributed computing based on java. It is consist of a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. Map reduce consist of mapping, shuffle and reduce stage.

MAPPING STAGE

Process the input data that stored on Hadoop as HDFS. The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

SHUFFLE STAGE

Transfers the map output from Mapper to a Reducer in MapReduce. Data from the mapper are grouped by the key, split among reducers and sorted by the key. Every reducer obtains all values associated with the same key.

REDUCE STAGE

Process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

CODE

- Clone from https://github.com/adilfidinda/docker-hadoop

data > ProcessUnits.java

```
1   package hadoop;
2
3   import java.util.*;
4
5   import java.io.IOException;
6   import java.io.IOException;
7
8   import org.apache.hadoop.fs.Path;
9   import org.apache.hadoop.conf.*;
10  import org.apache.hadoop.io.*;
11  import org.apache.hadoop.mapred.*;
12  import org.apache.hadoop.util.*;
```

data/ProcessUnit.java

CODE

- On datanode volumes, add ./data:/hadoop/data/
- This step is for mounting data folder on docker volumes so it can running through docker

```
datanode:
  image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
  container_name: datanode
  restart: always
  volumes:
    - hadoop_datanode:/hadoop/dfs/data
    - ./data:/hadoop/data/
```

CODE

- Run docker-compose up –d

```
D:\docker-hadoop>docker-compose up -d
Docker Compose is now in the Docker CLI, try `docker compose up`

Creating network "docker-hadoop_default" with the default driver
Creating historyserver   ... done
                                                                    Creating datanode       ... done
                                                  Creating resourcemanager ... done
              Creating nodemanager     ... done                                          Creating namenode       ... done
```

- For get into root enter docker exec –it datanode bash and cd to folder “data”

```
D:\docker-hadoop>docker exec -it datanode bash
root@782cdfa5dac5:/# cd hadoop/data
```

- Create folder units to compile java code

```
root@782cdfa5dac5:/hadoop/data# mkdir units
```

CODE

- javac -classpath hadoop-core-1.2.1.jar -d units ProcessUnits.java
- jar -cvf units.jar -C units/ .

```
root@782cdfa5dac5:/hadoop/data# javac -classpath hadoop-core-1.2.1.jar -d units ProcessUnits.java
root@782cdfa5dac5:/hadoop/data# jar -cvf units.jar -C units/ .
added manifest
adding: hadoop/(in = 0) (out= 0)(stored 0%)
adding: hadoop/ProcessUnits$E_EMapper.class(in = 1980) (out= 814)(deflated 58%)
adding: hadoop/ProcessUnits$E_EReduce.class(in = 1661) (out= 678)(deflated 59%)
adding: hadoop/ProcessUnits.class(in = 1565) (out= 767)(deflated 50%)
```

- Create directory for input and copy sample.txt to directory input

```
root@782cdfa5dac5:/hadoop/data# hadoop fs -mkdir /input_dir
root@782cdfa5dac5:/hadoop/data# hadoop fs -put sample.txt /input_dir/sample.txt
2021-06-13 14:15:46,838 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

CODE

- javac -classpath hadoop-core-1.2.1.jar -d units ProcessUnits.java
- jar -cvf units.jar -C units/ .

```
root@782cdfa5dac5:/hadoop/data# javac -classpath hadoop-core-1.2.1.jar -d units ProcessUnits.java
root@782cdfa5dac5:/hadoop/data# jar -cvf units.jar -C units/ .
added manifest
adding: hadoop/(in = 0) (out= 0)(stored 0%)
adding: hadoop/ProcessUnits$E_EMapper.class(in = 1980) (out= 814)(deflated 58%)
adding: hadoop/ProcessUnits$E_EReduce.class(in = 1661) (out= 678)(deflated 59%)
adding: hadoop/ProcessUnits.class(in = 1565) (out= 767)(deflated 50%)
```

- Create directory for input and copy sample.txt to directory input

```
root@782cdfa5dac5:/hadoop/data# hadoop fs -mkdir /input_dir
root@782cdfa5dac5:/hadoop/data# hadoop fs -put sample.txt /input_dir/sample.txt
2021-06-13 14:15:46,838 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

CODE

- Sample.txt contain of data set

```
root@782cdfa5dac5:/hadoop/data# hadoop fs -cat /input_dir/sample.txt
2021-06-13 15:03:52,432 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1979    23      23      2       43      24      25      26      26      26      26      25      26      25
1980    26      27      28      28      28      30      31      31      31      30      30      30      29
1981    31      32      32      32      33      34      35      36      36      34      34      34      34
1984    39      38      39      39      39      41      42      43      40      39      38      38      40
1985    38      39      39      39      39      41      41      41      00      40      39      39      45root@782cdfa5dac5:/hadoop/data#
```

- Run hadoop jar units.jar hadoop.ProcessUnits input

```
root@782cdfa5dac5:/hadoop/data# hadoop jar units.jar hadoop.ProcessUnits /input_dir /output_dir
2021-06-13 14:16:33,427 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.25.0.3:8032
```

- If job run successfully

```
2021-06-13 14:16:41,663 INFO mapreduce.Job:  map 0% reduce 0%
2021-06-13 14:16:47,736 INFO mapreduce.Job:  map 100% reduce 0%
2021-06-13 14:16:53,776 INFO mapreduce.Job:  map 100% reduce 100%
2021-06-13 14:16:53,792 INFO mapreduce.Job: Job job_1623593386825_0002 completed successfully
```

CODE

- To view result enter hadoop fs –cat /output_dir/part-ooooo

```
root@782cdfa5dac5:/hadoop/data# hadoop fs -ls /output_dir
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-06-13 14:16 /output_dir/_SUCCESS
-rw-r--r--   3 root supergroup         24 2021-06-13 14:16 /output_dir/part-00000
root@782cdfa5dac5:/hadoop/data# hadoop fs -cat /output_dir/part-00000
2021-06-13 14:18:04,551 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1981    34
1984    40
1985    45
```

FAILURE IN MAP REDUCE – TASK FAILURE

The child JVM reports the error back to its parent task tracker before it exits. The error ultimately makes it into the user logs. The task tracker marks the task attempt as failed, freeing up a slot to run another task.

FAILURE IN MAP REDUCE – TASK TRACKER FAILURE

Failure of a task tracker is another failure mode. If a task tracker fails by crashing or running very slowly, it will stop sending heartbeats to the job tracker or send them very infrequently

FAILURE IN MAP REDUCE – JOB TRACKER FAILURE

Failure of the job tracker is the most serious failure mode. Hadoop has no mechanism for dealing with job tracker failure it is a single point of failure so in this case all running jobs fail.