Introduction to HDFS

EECS 4415

Big Data Systems

Tilemachos Pechlivanoglou

tipech@eecs.yorku.ca

What is HDFS

- distributed file system
  - *fault tolerant*
  - *scalable*
  - *extremely easy to expand.*

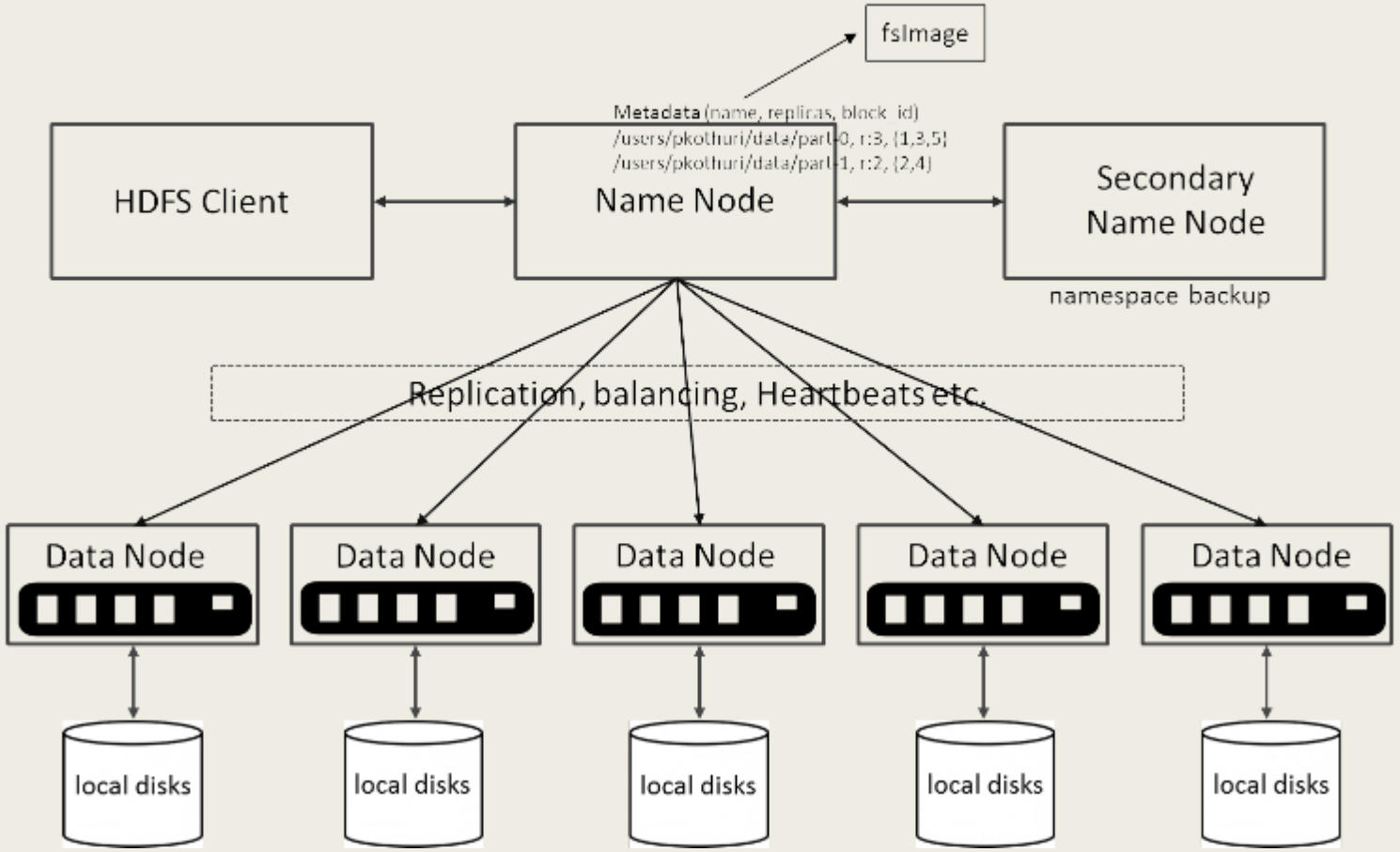
- designed to ‘just work’

HDFS components

- NameNode :– is the heart of an HDFS filesystem, it maintains and manages the file system metadata. E.g; what blocks make up a file, and on which datanodes those blocks are stored.

- DataNode :- where HDFS stores the actual data, there are usually quite a few of these.

HDFS Architecture



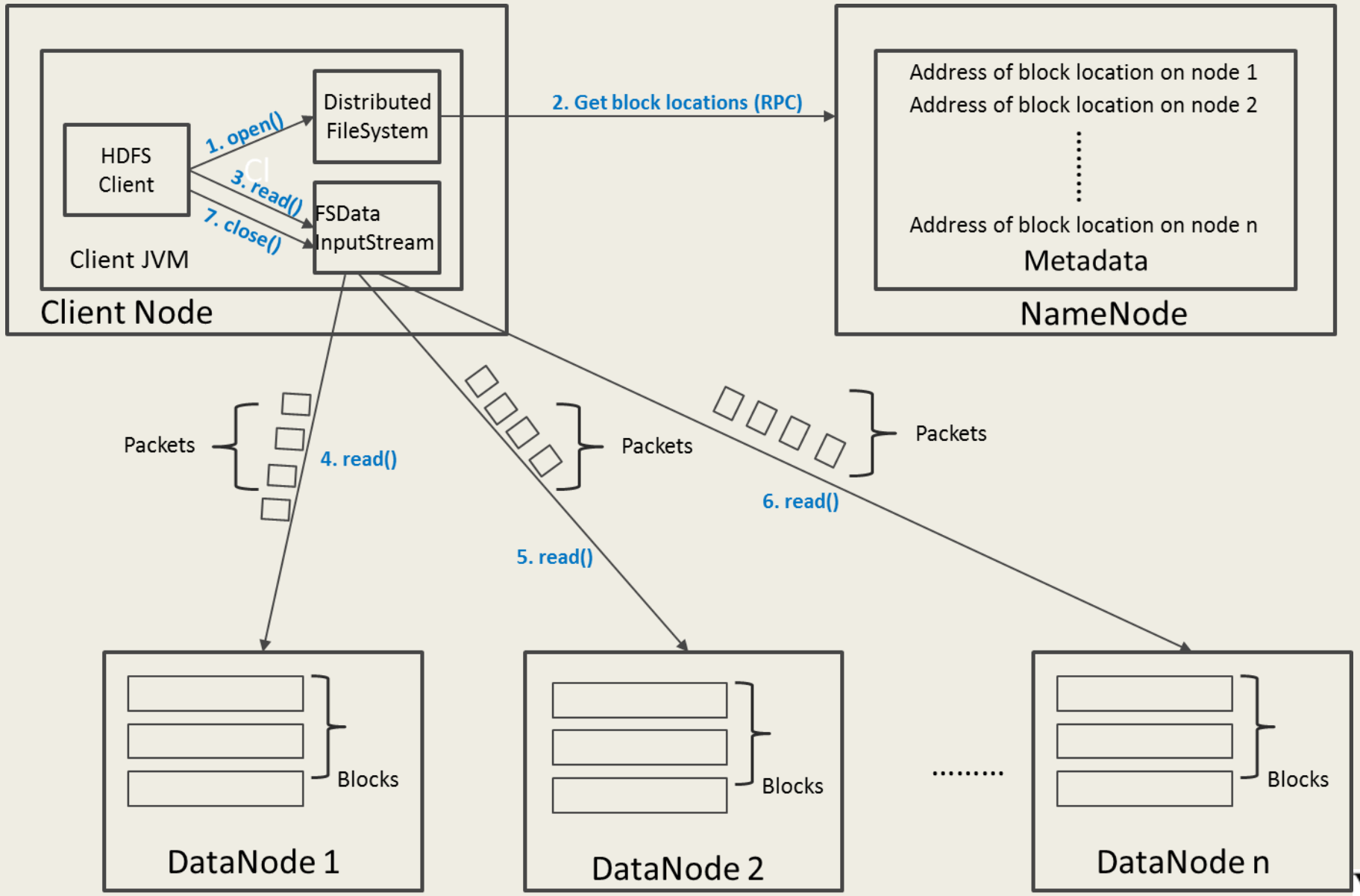
HDFS features

- Failure tolerance
- Scalability
- Space
- Industry standard

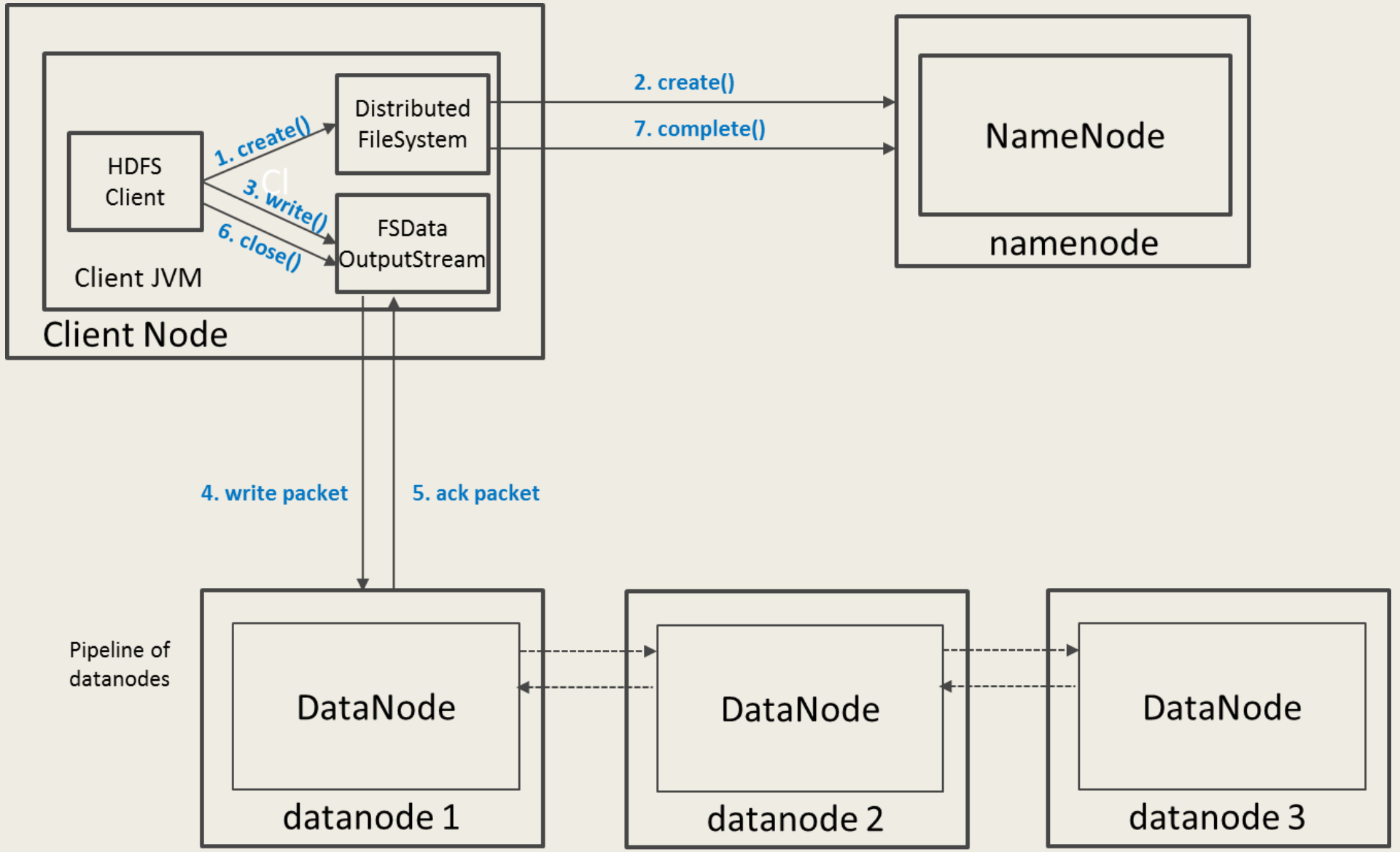
HDFS features

- Each file written into HDFS is split into data blocks
- Each block is stored on one or more nodes
- Copies can be replicated
  - *this way they are not lost if a server goes down*

Read Operation



Write Operation



Configuration

HDFS Defaults

- Block Size – 64 MB
- Replication Factor – 3
- Web UI Port – 50070

HDFS conf file - `/etc/hadoop/conf/hdfs-site.xml`

```
<property>
   <name>dfs.namenode.name.dir</name>
   <value>file:///data1/cloudera/dfs/nn,file:///data2/cloudera/dfs/nn</value>
</property>

<property>
   <name>dfs.blocksize</name>
   <value>268435456</value>
</property>

<property>
   <name>dfs.replication</name>
   <value>3</value>
</property>

<property>
   <name>dfs.namenode.http-address</name>
   <value>itracXXX.cern.ch:50070</value>
</property>
```

HDFS commands

- There are two types of shell commands
- User Commands
  - ***`hdfs dfs`*** *– runs filesystem commands on the HDFS*
  - ***`hdfs fsck`*** *– runs a HDFS filesystem checking command*
- Administration Commands
  - ***`hdfs dfsadmin`*** *– runs HDFS administration commands*

HDFS commands

List directory contents

```
hdfs dfs –ls
hdfs dfs -ls /
hdfs dfs -ls -R /var
```

Display disk space used

```
hdfs dfs -du -h /
hdfs dfs -du /hbase/data/hbase/namespace/
hdfs dfs -du -h /hbase/data/hbase/namespace/
hdfs dfs -du -s /hbase/data/hbase/namespace/
```

HDFS commands

Copy data to HDFS

```
hdfs dfs -mkdir tdata
hdfs dfs -ls
hdfs dfs -copyFromLocal tutorials/data/geneva.csv tdata
hdfs dfs -ls –R
```

Copy back to local filesystem

```
cd tutorials/data/
hdfs dfs –copyToLocal tdata/geneva.csv geneva.csv.hdfs
md5sum geneva.csv geneva.csv.hdfs
```

HDFS commands

Removing a file

```
hdfs dfs -rm tdataset/tfile.txt
hdfs dfs -ls –R
```

Write to hdfs from stdin

```
echo "blah blah blah" | hdfs dfs -put - tdataset/tfile.txt
hdfs dfs -ls –R
hdfs dfs -cat tdataset/tfile.txt
```

HDFS admin commands

Get report

```
hdfs dfsadmin –report
```

Get information of one node

```
hdfs dfsadmin -getDatanodeInfo
localhost:50020
```

Links and other material

- Python tutorials:
  - https://www.w3schools.com/python/
  - https://www.learnpython.org/
- Python documentation:
  - https://docs.python.org/3/

Thank you!

Questions?