upGrad

Hadoop & MapReduce Programming

Mayukh Chakraborty

Agenda

- Evolution of Big Data
- Hadoop Architecture
- HDFS Read Path
- HDFS Write Path
- YARN
- HDFS Commands
- Hadoop Tools & Use Cases
- MapReduce Programming
- Questions & Answers

Data Types

1. Structured

- Data stored in DB tables
- Constitutes 5% of all data being processed.

1. Semi-Structured

- XML/JSON/Log File data
- Constitutes 5-10% of all data being processed.

1. Unstructured

- Text, image, video data
- Constitutes more than 80% of all data being processed.

Quiz

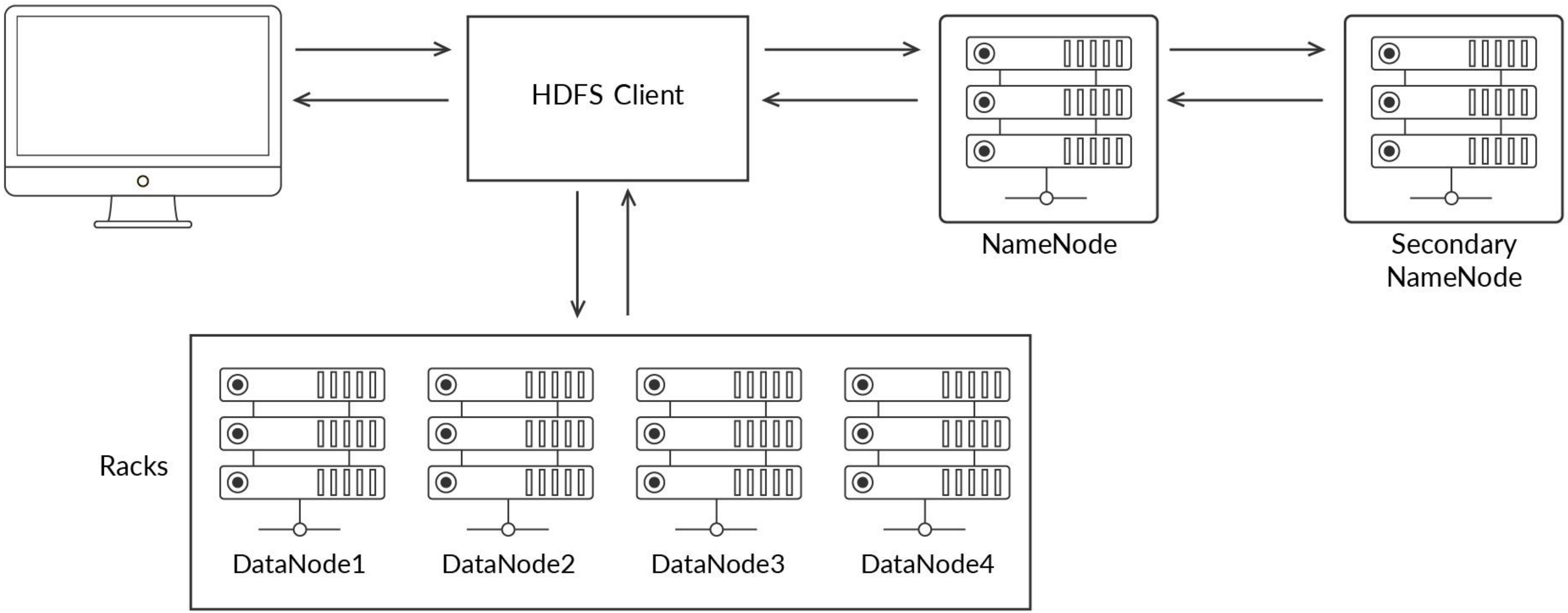
What kind of data, traditional DBMSs are most suitable for?

The Evolution of Distributed Systems

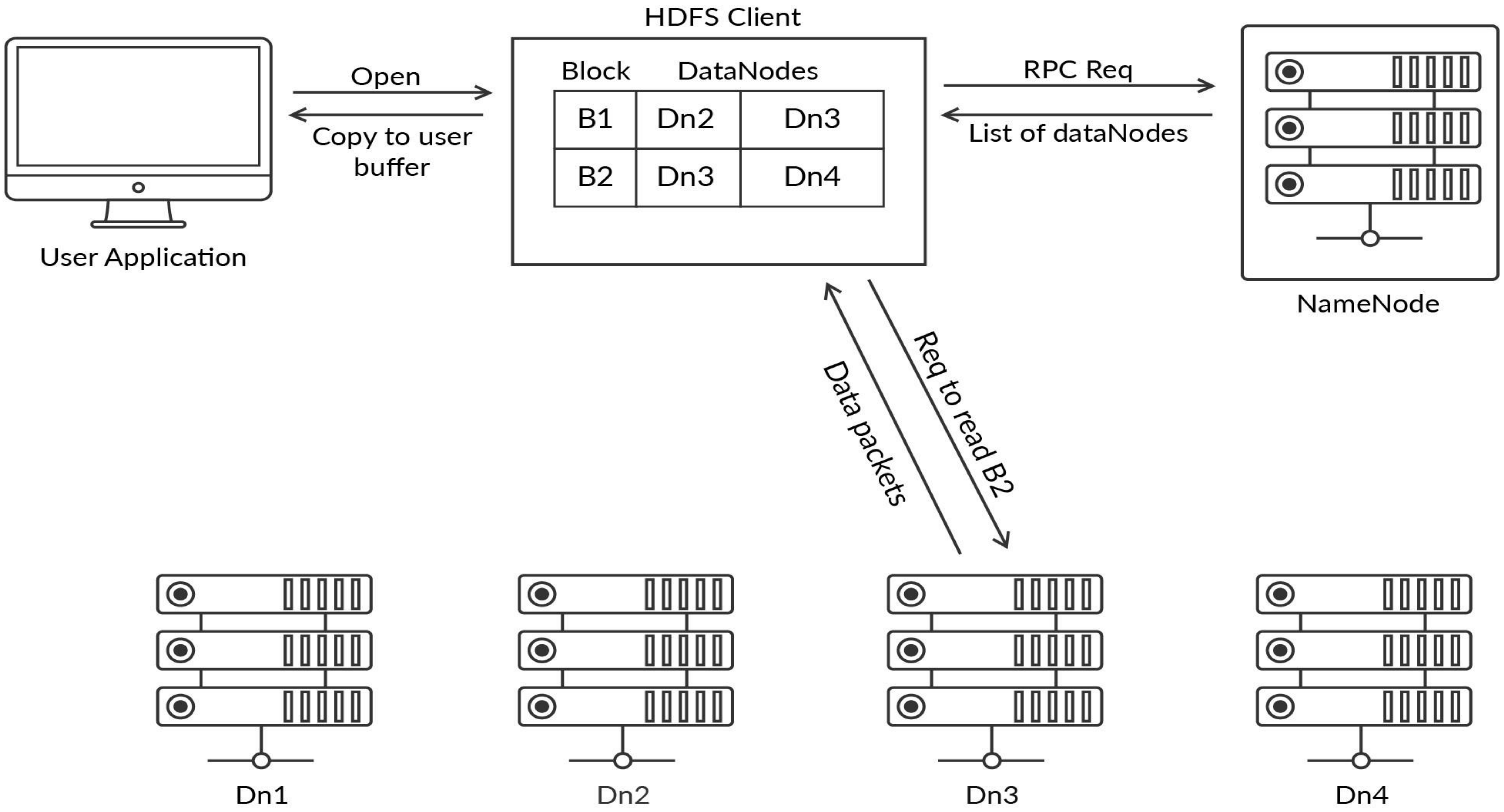
The 4Vs of Big Data

- **Volume**
  - The data cannot be stored and processed by a single machine.
- **Velocity**
  - The system should be capable of storing and processing data at high-speed.
- **Variety**
  - The System should be capable of handling structured, semi-structured and unstructured data.
- **Veracity**
  - Data can be of Questionable quality because of noise and bias.

Hadoop Architecture

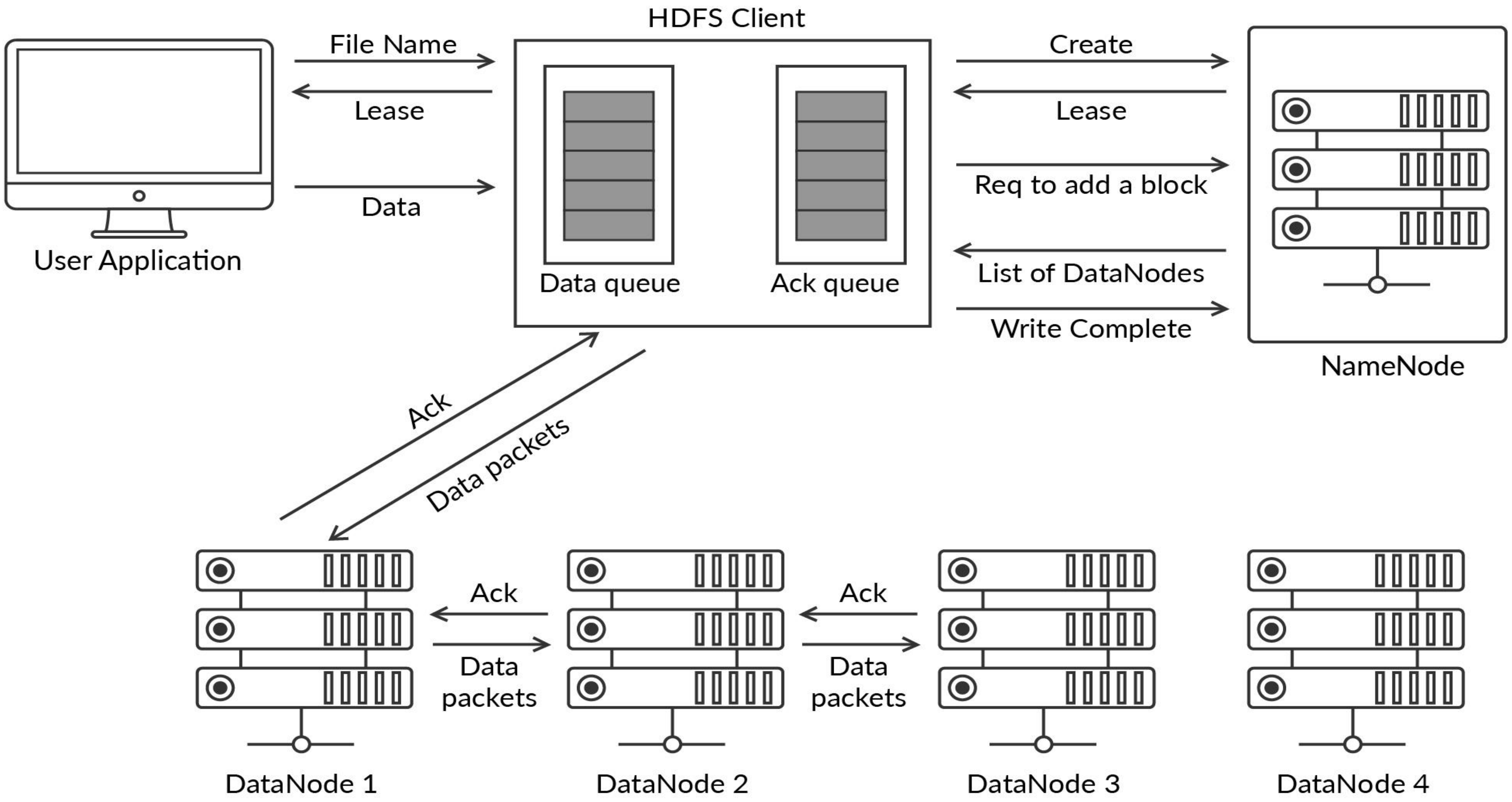


Hadoop Read Path

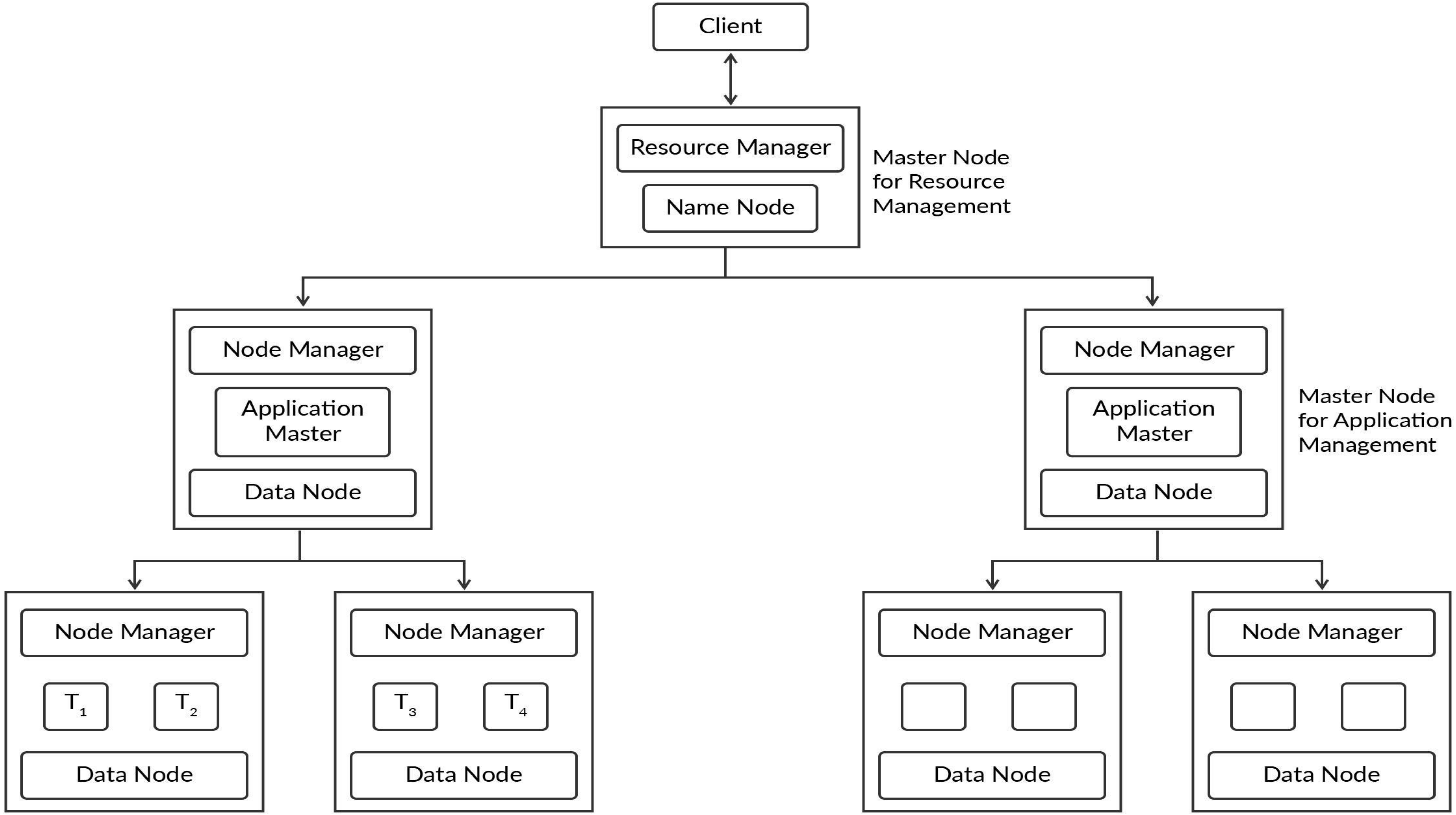


Block	DataNodes	
B1	Dn2	Dn3
B2	Dn3	Dn4

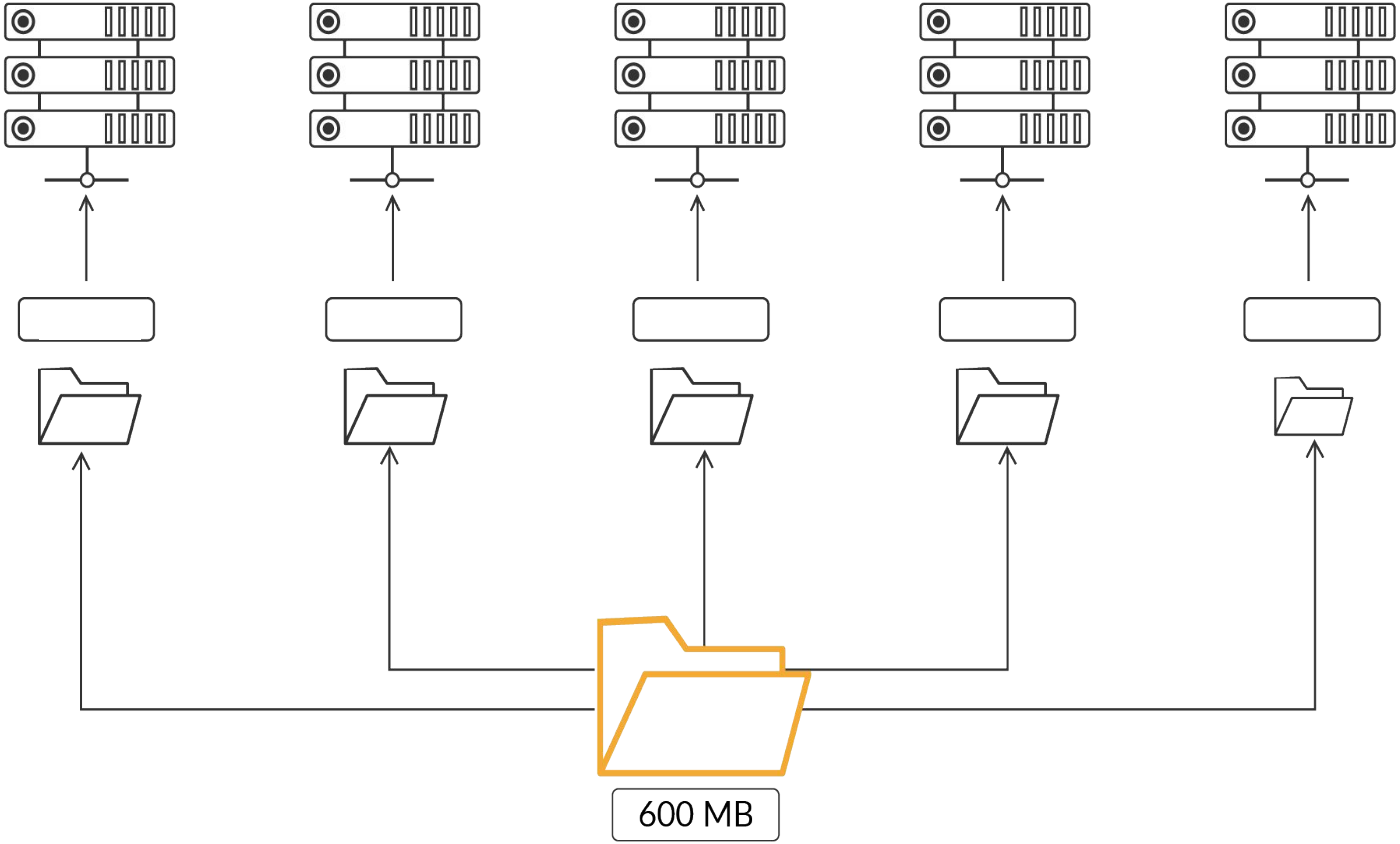
Hadoop Write Path



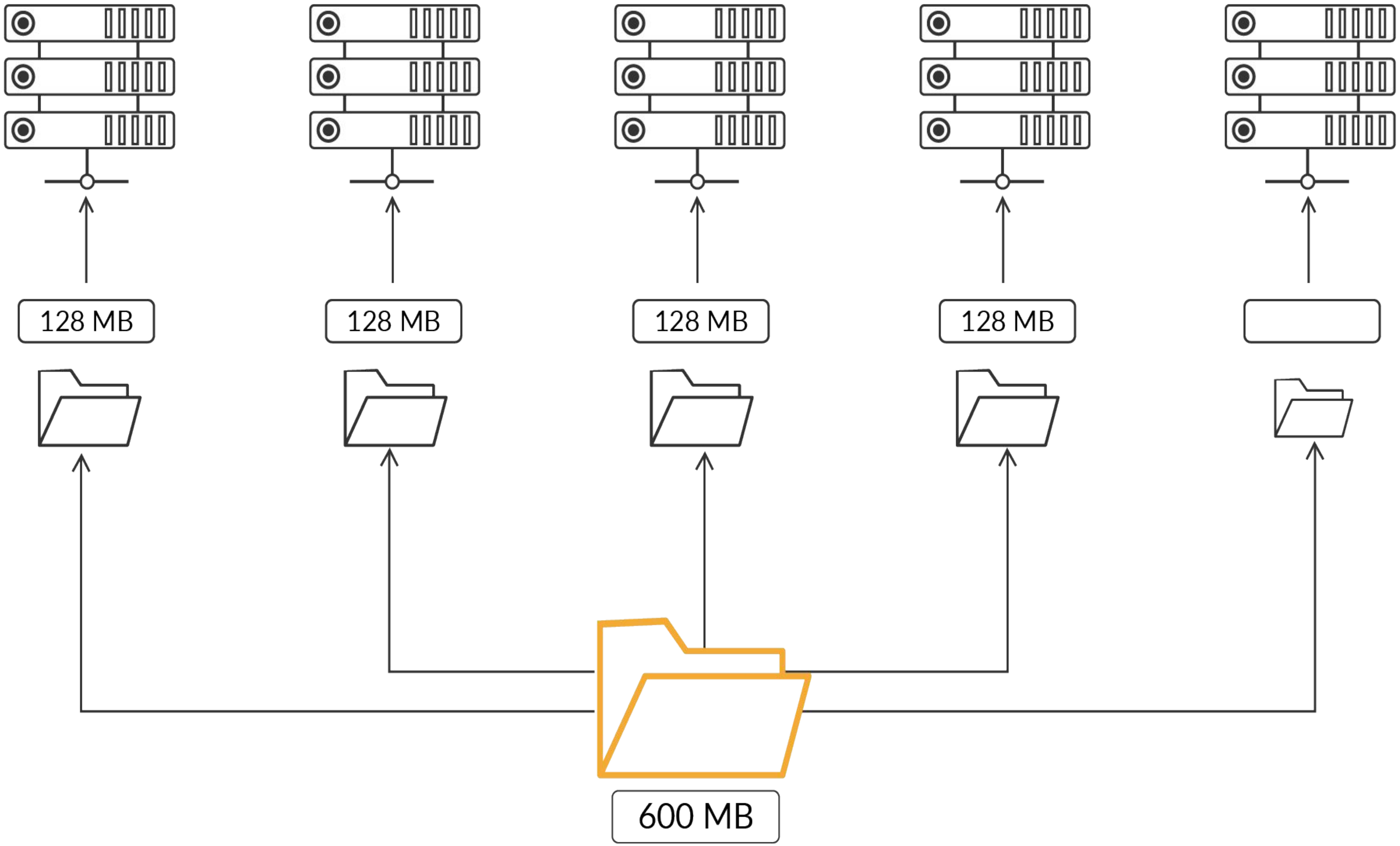
YARN Components



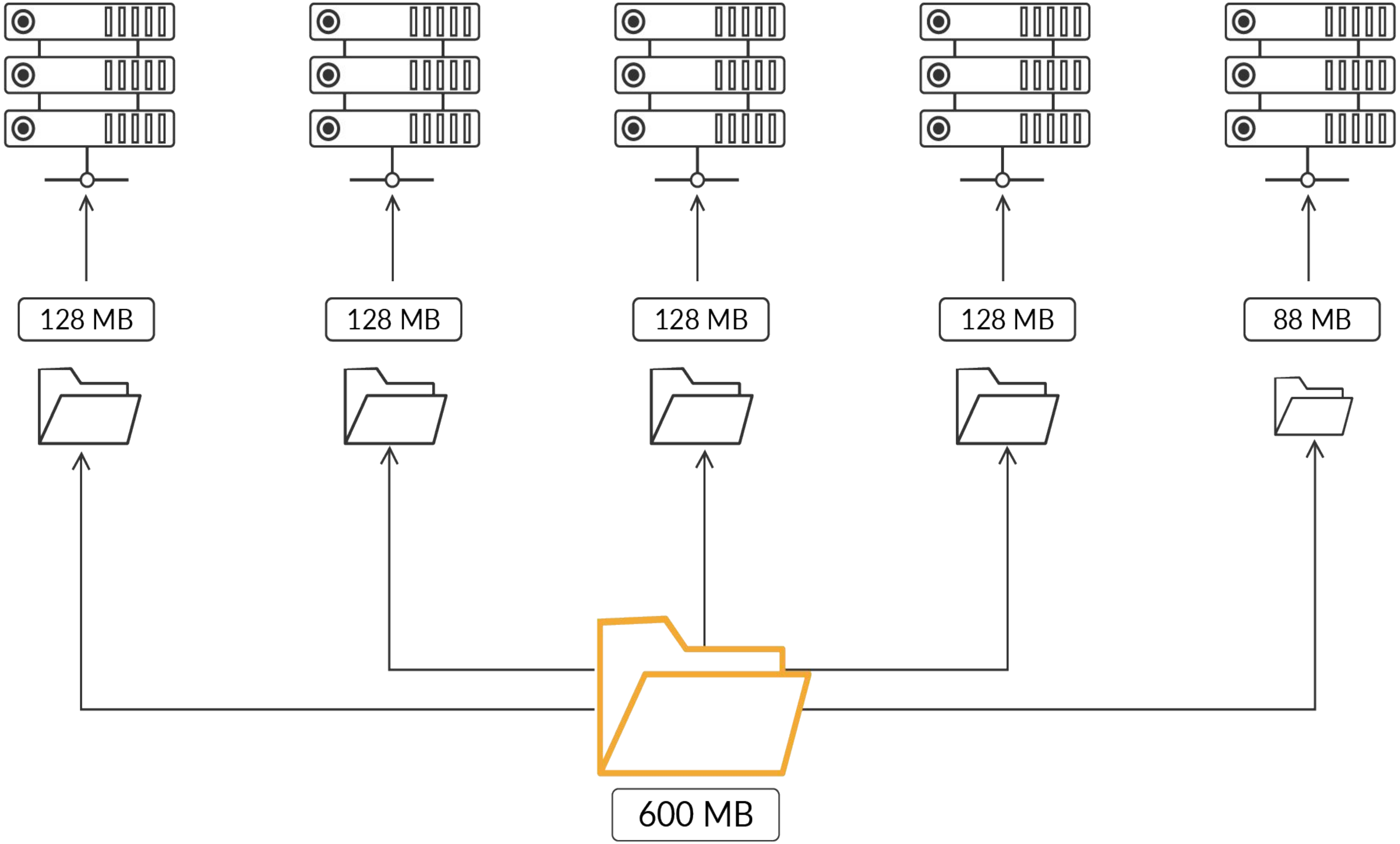
File Storage



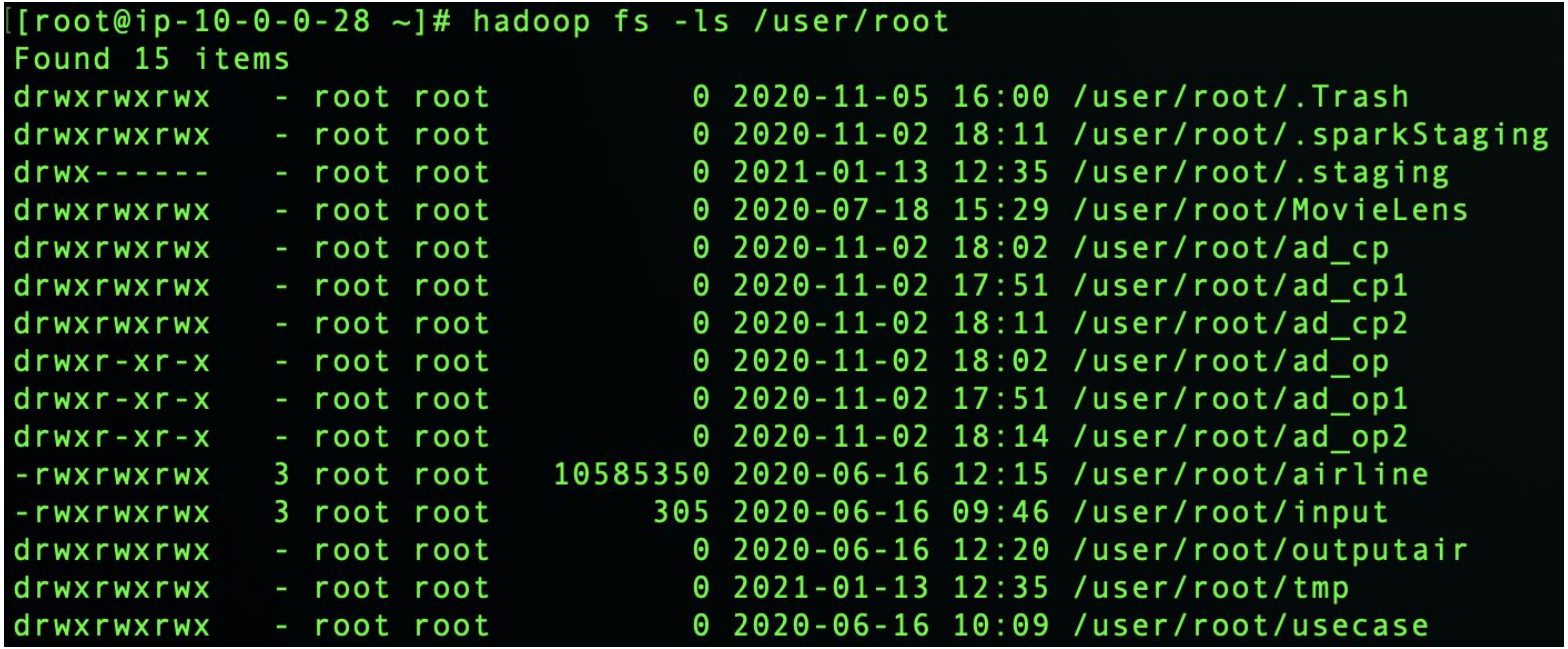
File Storage



File Storage



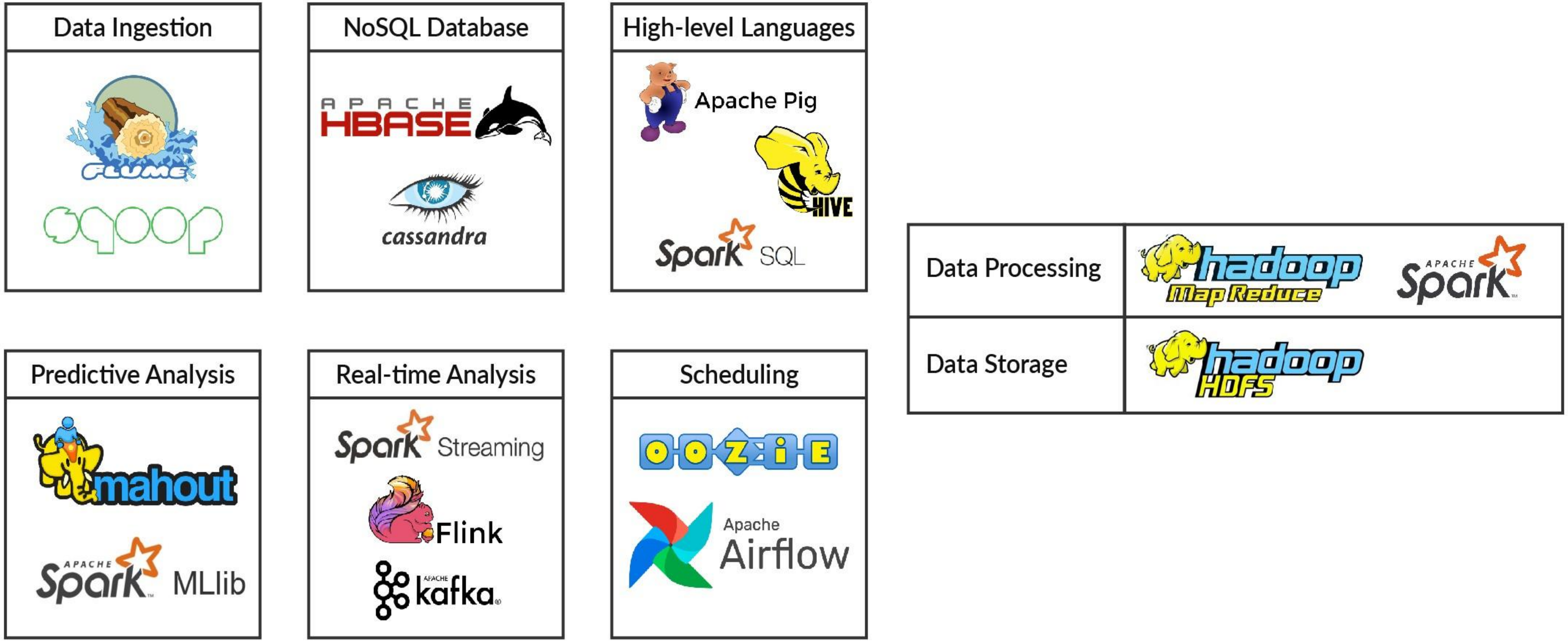
HDFS Commands



```
[[root@ip-10-0-0-28 ~]# hadoop fs -ls /user/root
Found 15 items
drwxrwxrwx   - root root          0 2020-11-05 16:00 /user/root/.Trash
drwxrwxrwx   - root root          0 2020-11-02 18:11 /user/root/.sparkStaging
drwx------   - root root          0 2021-01-13 12:35 /user/root/.staging
drwxrwxrwx   - root root          0 2020-07-18 15:29 /user/root/MovieLens
drwxrwxrwx   - root root          0 2020-11-02 18:02 /user/root/ad_cp
drwxrwxrwx   - root root          0 2020-11-02 17:51 /user/root/ad_cp1
drwxrwxrwx   - root root          0 2020-11-02 18:11 /user/root/ad_cp2
drwxr-xr-x   - root root          0 2020-11-02 18:02 /user/root/ad_op
drwxr-xr-x   - root root          0 2020-11-02 17:51 /user/root/ad_op1
drwxr-xr-x   - root root          0 2020-11-02 18:14 /user/root/ad_op2
-rwxrwxrwx   3 root root   10585350 2020-06-16 12:15 /user/root/airline
-rwxrwxrwx   3 root root        305 2020-06-16 09:46 /user/root/input
drwxrwxrwx   - root root          0 2020-06-16 12:20 /user/root/outputair
drwxrwxrwx   - root root          0 2021-01-13 12:35 /user/root/tmp
drwxrwxrwx   - root root          0 2020-06-16 10:09 /user/root/usecase
```

HDFS Commands Hand On Execution

Hadoop Tools & Use cases



Data Ingestion

NoSQL Database

High-level Languages

Predictive Analysis

Real-time Analysis

Scheduling

Data Processing

Data Storage

MapReduce Programming

Quiz

Give an example, where a combiner can’t be used.

Command For MR Program

```
hadoop jar
/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streami
ng-2.6.0-cdh5.15.1.jar \
-file mapper.py -mapper 'python mapper.py' \
-file combiner.py -combiner 'python combiner.py' \
-file reducer.py -reducer 'python reducer.py' \
-input /user/root/tmp/ages.txt \
-output /user/root/tmp/output_age_1
```

Command For MR Program with Partitioning

```
hadoop jar
/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5
.15.1.jar \
-file mapper_avg.py -mapper 'python mapper_avg.py' \
-file combiner.py -combiner 'python combiner.py' \
-file reducer_avg.py -reducer 'python reducer_avg.py' \
-input /user/root/tmp/ages.txt \
-output /user/root/tmp/output_age_1
-D mapreduce.map.output.key.field.separator=, \
-D num.key.fields.for.partition=3 \
-partitioner org.apache.hadoop.mapred.lib.KeyFieldBasedPartitioner
```

Questions & Answers