UNIT II

Hadoop Ecosystem

HADOOP ADVANTAGES

Unlimited data storage

1. Server Scaling Mode

a) Vertical Scale

b)Horizontal Scale

High speed processing system

All varities of data processing

1. Structural
2. Unstructural
3. semi-structural

DISADVANTAGE OF HADOOP

If volume is small then speed of hadoop is bad
Limitation of hadoop data storage
Well there is obviously a practical limit. But physically HDFS Block IDs are Java longs so they have a max of 2^63
and if your block size is 64 MB then the maximum size is 512 yottabytes.
Hadoop should be used for only batch processing
1. Batch process:-background process
Hadoop is not used for OLTP
OLTP process:-interactive with uses

HDFS FEATURES

- Highly fault-tolerant
- High throughput
- Suitable for applications with large data sets
- Streaming access to file system data
- Can be built out of commodity hardware

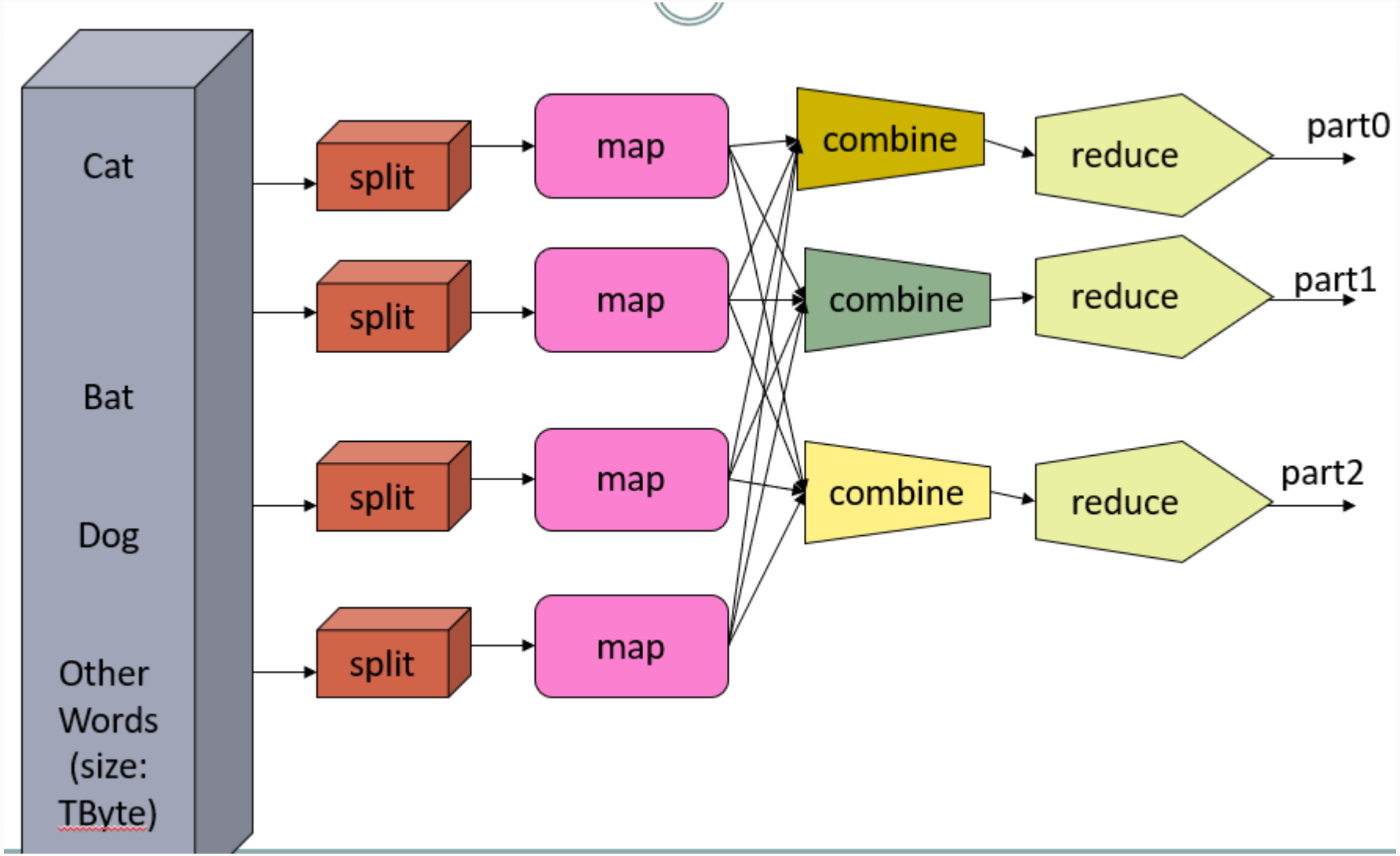
FAULT TOLERANCE

- Failure is the norm rather than exception
- A HDFS instance may consist of thousands of server machines, each storing part of the file system's data.
- Since we have huge number of components and that each component has non-trivial probability of failure means that there is always some component that is non-functional.
- Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

DATA CHARACTERISTICS

- Streaming data access
- Applications need streaming access to data
- Batch processing rather than interactive user access.
- Large data sets and files: gigabytes to terabytes size
- High aggregate data bandwidth
- Scale to hundreds of nodes in a cluster
- Tens of millions of files in a single instance
- Write-once-read-many: a file once created, written and closed need not be changed – this assumption simplifies coherency
- A map-reduce application or web-crawler application fits perfectly with this model.

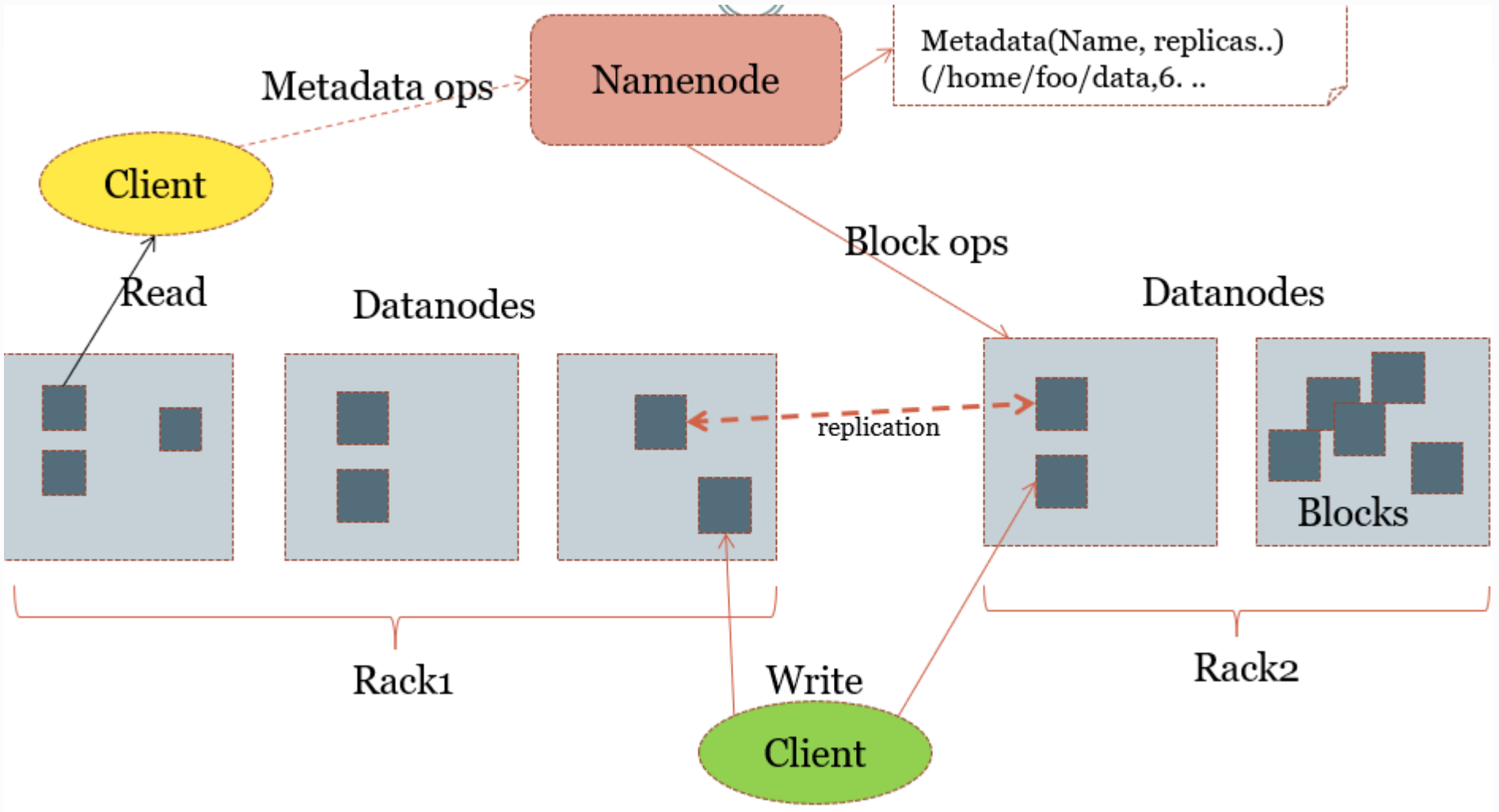
MAPREDUCE


Cat
Bat
Dog
Other Words (size: TByte)
split
split
split
split
map
map
map
map
combine
combine
combine
reduce
reduce
reduce
part0
part1
part2

HDFS ARCHITECTURE

- Master/slave architecture
- HDFS cluster consists of a single **Namenode**, a master server that manages the file system namespace and regulates access to files by clients.
- There are a number of **DataNodes** usually one per node in a cluster.
- The DataNodes manage storage attached to the nodes that they run on.
- HDFS exposes a file system namespace and allows user data to be stored in files.
- A file is split into one or more blocks and set of blocks are stored in DataNodes.
- DataNodes: serves read, write requests, performs block creation, deletion, and replication upon instruction from Namenode.

HDFS ARCHITECTURE



FILE SYSTEM NAMESPACE

Hierarchical file system with directories and files
Create, remove, move, rename etc.
Namenode maintains the file system
Any meta information changes to the file system recorded by the Namenode.
An application can specify the number of replicas of the file needed: replication factor of the file. This information is stored in the Namenode.

References

https://www.tandfonline.com/doi/abs/10.5437/08956308X5601005?journalCode=urtm20

https://www.igi-global.com/article/big-data-technologies-and-analytics/115517

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

https://www.geeksforgeeks.org/hadoop-ecosystem/

Thank You