Big Data!!!

Processing Big Data

Two basic tasks a program should accomplish:

- Handle storage
- Perform computations

Single-resource model

Usually we use:

- One processor
- One memory “bank”
- One disk/hard-drive.

Even if we have more than 1 processor or 1 disk we usually write programs that run on a single synchronous thread.

Processing Big Data

Google:

- Ten Billion Web pages
- Average size of Web page - 20KB
- Ten Billion * 20KB = 200TB
- Disk read speed (bandwidth) = 50 MB/Sec
- Time to read = 4 Million seconds =

46 days + 7 hours

Distributed Computing

The art of designing and building a system that will take advantage of multiple “nodes”

Distributed Computing

Goals:

- Improved performance
- High Availability
- Easy Scalability

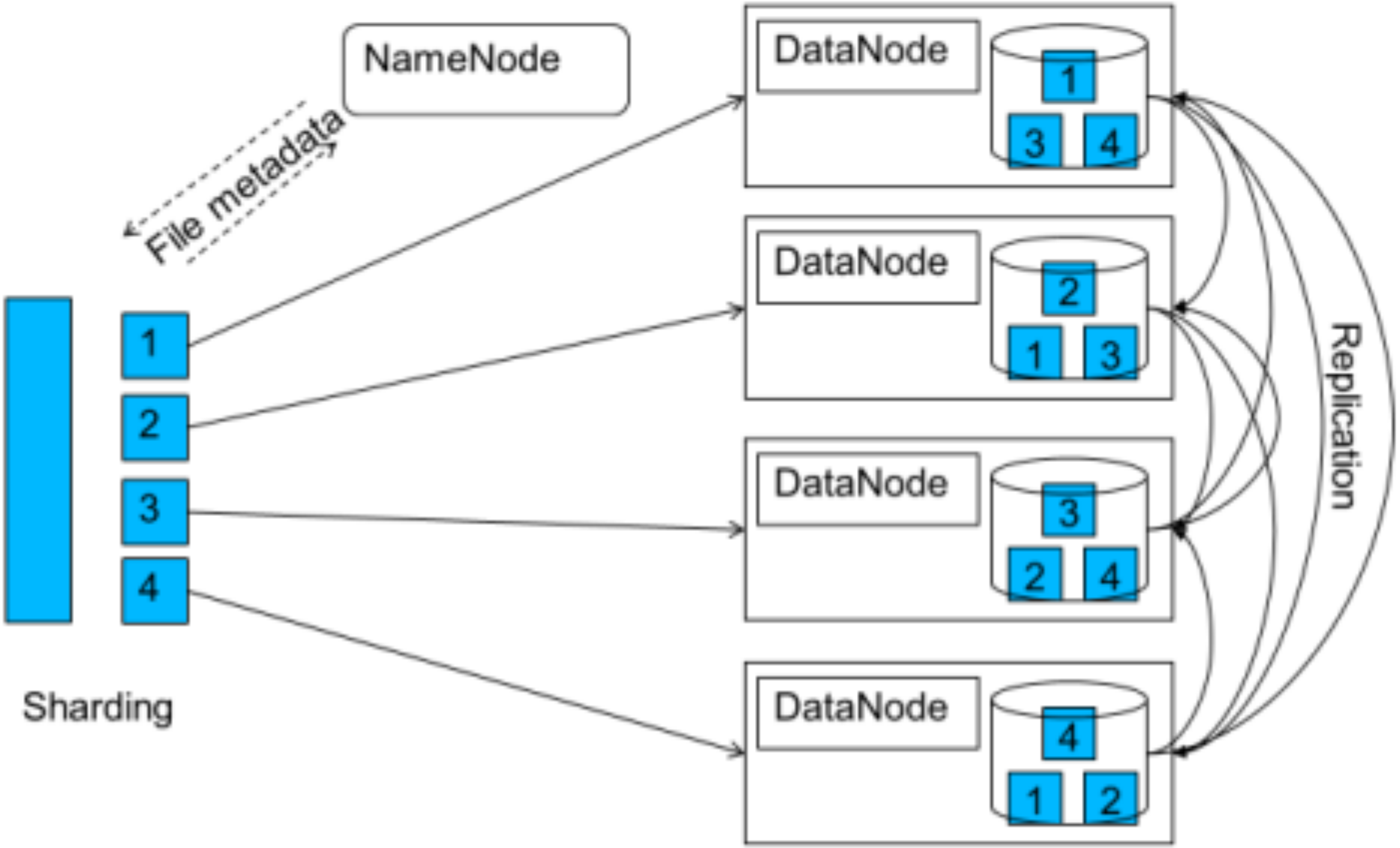
Hadoop: DFS + Map / Reduce

Framework for distributed computing

Distributed File System:

- Data kept in parts (sharding)
- Data is replicated to other nodes
- One “global” management system

Hadoop: DFS + Map / Reduce



Hadoop: DFS + Map / Reduce

Map / Reduce :- A-sync processes are:

- Map - scanning input data (Key->Value) and extracting info by Key
- Group and Sort by Key
- Reduce - aggregate / summarize data

Hadoop: DFS + Map / Reduce

In addition, Hadoop is:

- Scheduling execution of programs:
  - Local (“close” to data) execution and results storage
  - Managing queue of tasks
- Handling node failures:
  - Failure detection
  - Reseting (and re-queuing) tasks
- Managing inter-node communication

Hadoop Setup demo

Hadoop Setup demo

- Procuring a VM (Ubuntu)
- Connecting with SSH
- Updating Ubuntu
- Installing Hadoop - Prerequisites:
  - Java:
    - **apt-cache search openjdk**
    - **sudo apt-get install openjdk-8-jdk**
  - SSH/PDSH: **sudo apt-get install ssh**, **sudo apt-get install pdsh**
  - Rsync: **sudo apt-get install rsync**

Hadoop Setup demo

- Downloading Hadoop:
  - **sudo wget http://apache.osuosl.org/hadoop/common/stable/hadoop-2.7.3.tar.gz**
  - **sudo tar xzf hadoop-2.7.3.tar.gz**

- Pointing Hadoop to JAVA home:
  - **Edit ./etc/hadoop/hadoop-env.sh**
  - **export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64**

Hadoop Setup demo

- Get data:
  - `scp *.txt ubuntu@hadoop:~`
  - `sudo mkdir input`
  - `sudo mv ~/*.txt ./input`
- Running M/R job:

  -
    ```
    sudo bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.3.jar grep input output '[a-z.]+'
    ```
- Check output:
  - `cat output/*`