Hadoop and

Erika and Nathan

History of Hadoop

- Started with Google
- Google utilized a DFS and MapReduce
- The Hadoop people basically stole it

Hadoop Distributed File System

- Hadoop is used for BIG DATA
- Data is stored across many nodes (cluster)
- Hadoop makes 2 back ups of your data
- Namenode calibrates the data across the cluster

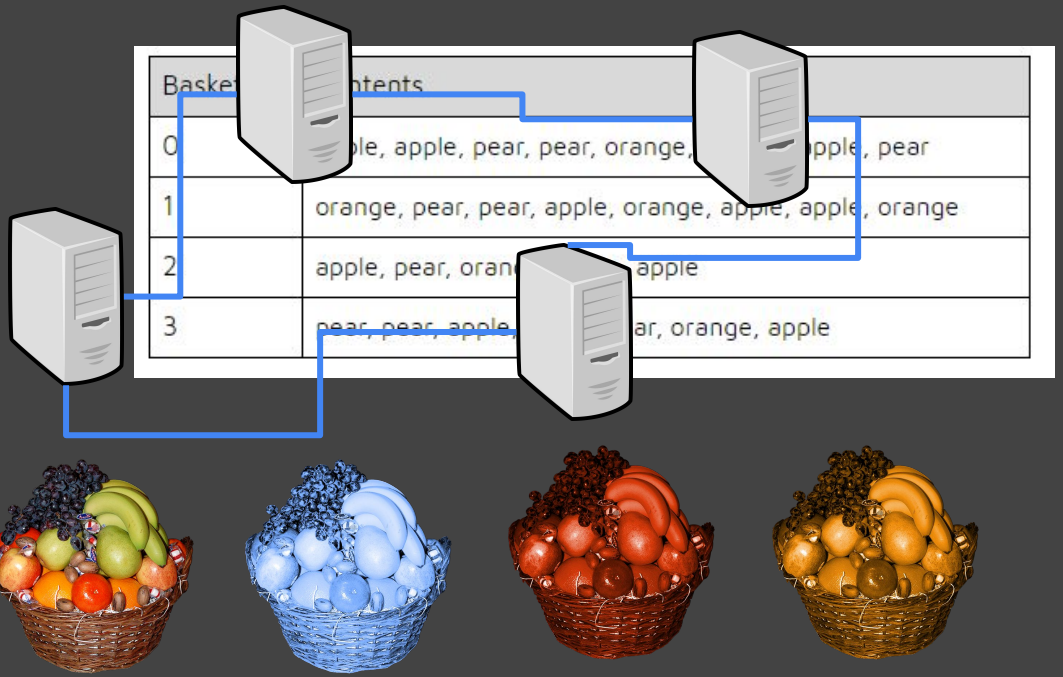
HDFS - A Farmer's Story

His "file"

Basket ID	Contents
0	apple, apple, pear, pear, orange, orange, apple, pear
1	orange, pear, pear, apple, orange, apple, apple, orange
2	apple, pear, orange, apple, apple
3	pear, pear, apple, apple, pear, orange, apple

64 MB 64 MB 64 MB 64 MB

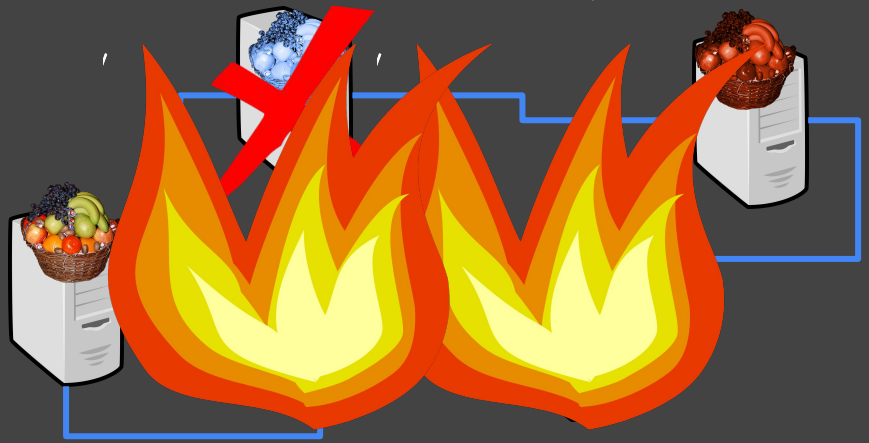
Hadoop Distributed File System



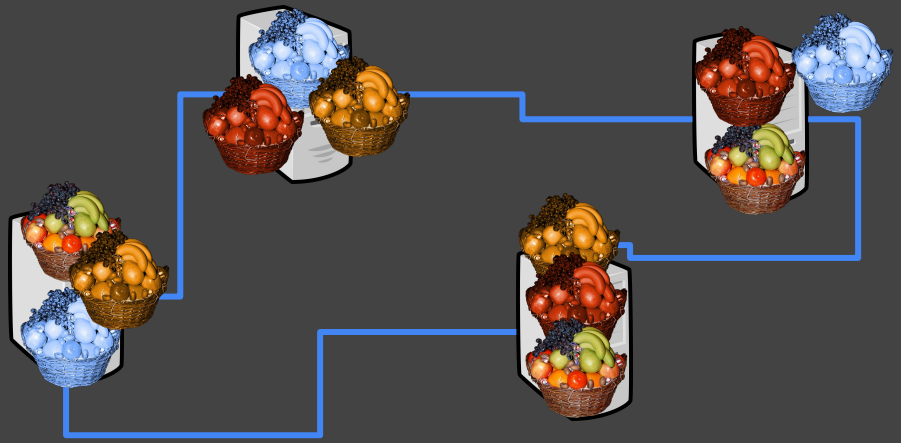
Basket	ntents
0	le, apple, pear, pear, orange, pple, pear
1	orange, pear, pear, apple, orange, apple, apple, orange
2	apple, pear, oran apple
3	ear, orange, apple

64 MB 64 MB 64 MB 64 MB

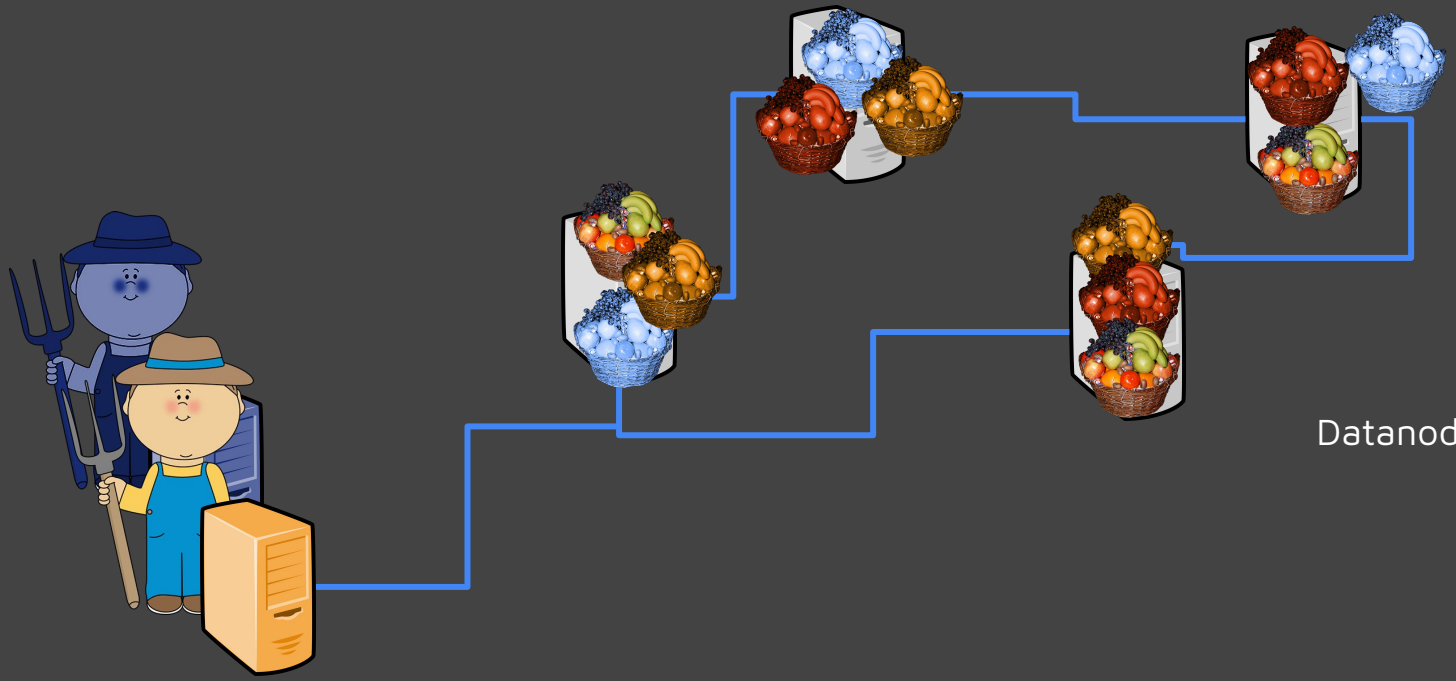
Hadoop Distributed File System



Hadoop Distributed File System



Hadoop Distributed File System



NameNode

Datanodes

MapReduce

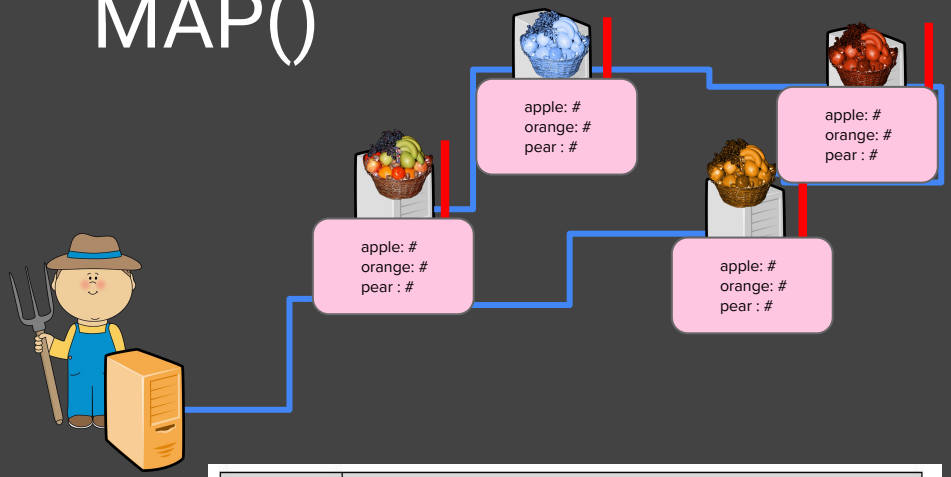
Map Sort/Shuffle Reduce

- Deals with big data more efficiently
- Utilizes multiple nodes for PARALLEL PROCESSES
- Divides/conquers (backs up)

MapReduce

Map Sort/Shuffle Reduce

MAP()



Basket ID	Contents
0	apple, apple, pear, pear, orange, orange, apple, pear
1	orange, pear, pear, apple, orange, apple, apple, orange
2	apple, pear, orange, apple, apple
3	pear, pear, apple, apple, pear, orange, apple

On each node:

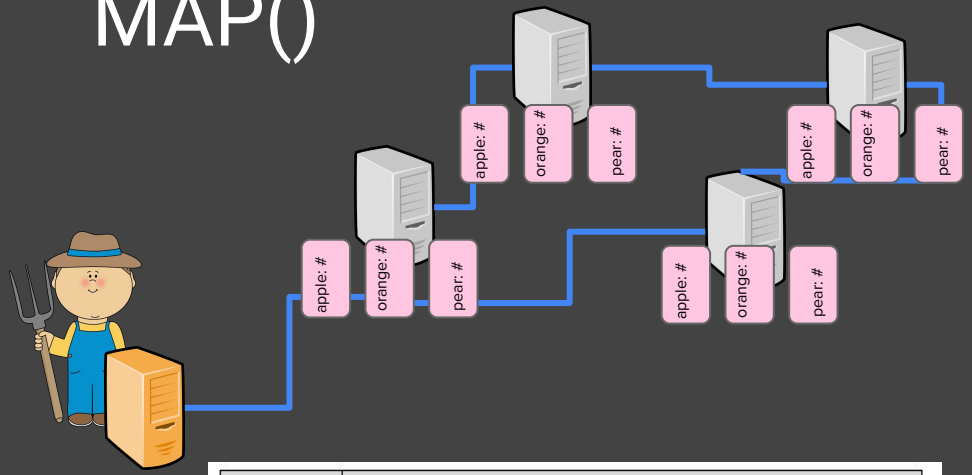
MAP( )

apple: #
orange: #
pear : #

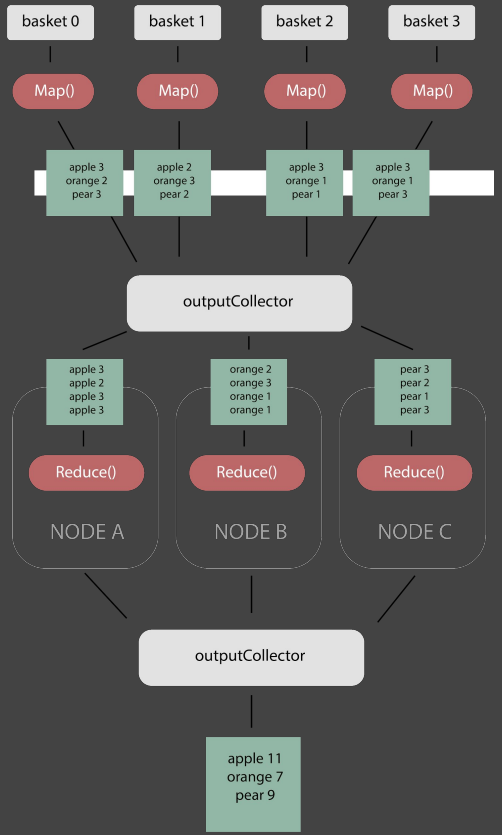
MapReduce

Map Sort/Shuffle Reduce

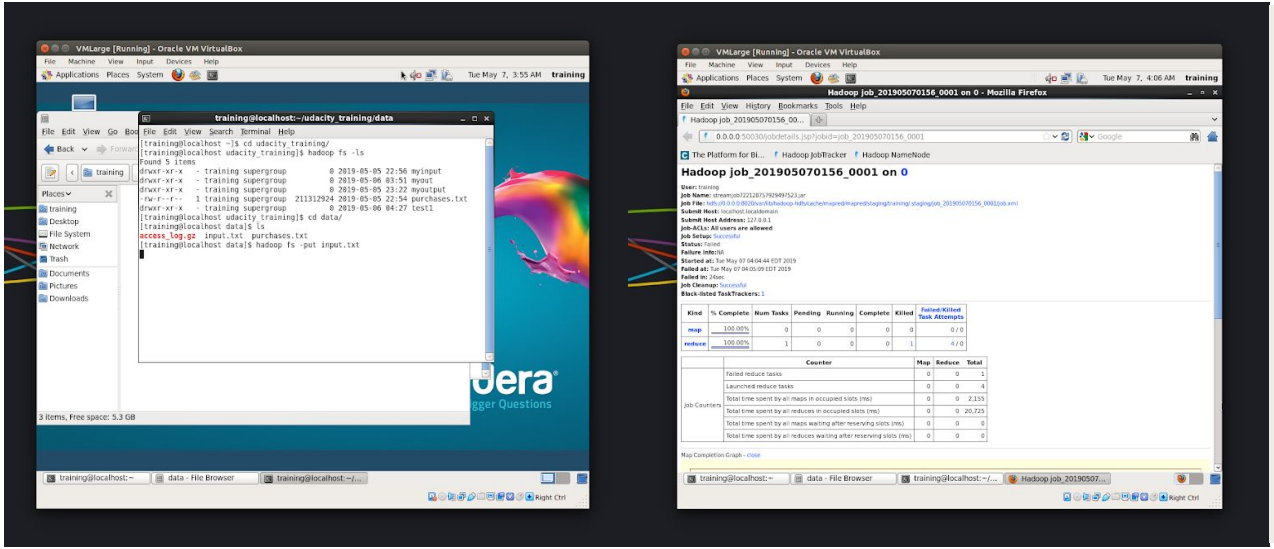
MAP()



Basket ID	Contents
0	apple, apple, pear, pear, orange, orange, apple, pear
1	orange, pear, pear, apple, orange, apple, apple, orange
2	apple, pear, orange, apple, apple
3	pear, pear, apple, apple, pear, orange, apple



Demo:


VMLarge [Running] - Oracle VM VirtualBox
training@localhost:~/udacity_training/data
[training@localhost ~]$ cd udacity_training/
[training@localhost udacity_training]$ hadoop fs -ls
Found 5 items
drwxr-xr-x - training supergroup 0 2019-05-05 22:56 myinput
drwxr-xr-x - training supergroup 0 2019-05-06 03:51 myout
drwxr-xr-x - training supergroup 0 2019-05-05 23:22 myoutput
-rw-r--r-- 1 training supergroup 211312924 2019-05-05 22:54 purchases.txt
drwxr-xr-x - training supergroup 0 2019-05-06 04:27 test1
[training@localhost udacity_training]$ cd data/
[training@localhost data]$ ls
access_log.gz input.txt purchases.txt
[training@localhost data]$ hadoop fs -put input.txt
3 items, Free space: 5.3 GB
Hadoop job_201905070156_0001 on 0 - Mozilla Firefox
Hadoop job_201905070156_0001 on 0
Hadoop job_201905070156_0025 on 0 - Mozilla Firefox

Class Interaction Yay

Guess: 1 Star 3 Stars 5 Stars

★ ★ ★ ★ ★

★ ★ ★ ★ ★

★ ★ ★ ★ ★

Class Interaction Yay

Guess: 1 Star 3 Stars 5 Stars

★★★☆☆

★★★★★

★☆☆☆☆

QUESTIONS?