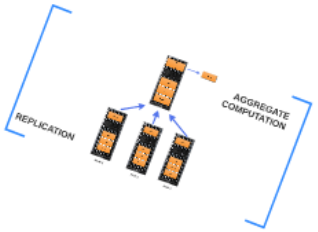
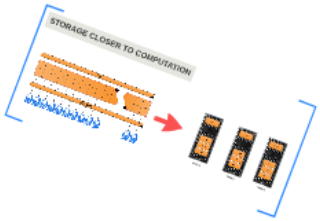







PROBLEM ANALYZE SOLUTION

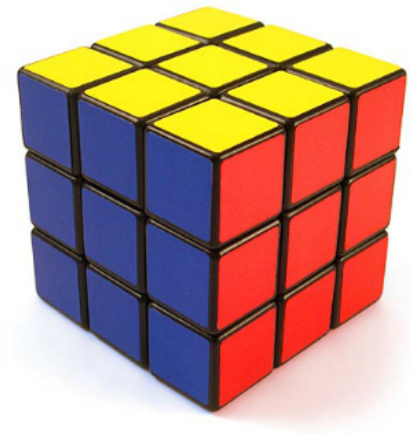



www.hadoopinrealworld.com

# UNDERSTANDING BIG DATA PROBLEM

PROBLEM

ANALYZE



SOLUTION

# SAMPLE BIG DATA PROBLEM

- Stocks Dataset - Day by day stock information for several symbols for several years
- Size - 1 TB
- Problem - Find out Maximum closing price for each stock symbol

```
ABCSE,B7J,2008-10-28,6.48,6.74,6.22,6.72,44300,5.79
ABCSE,B7J,2008-10-27,6.21,6.78,6.21,6.40,55200,5.51
ABCSE,B7J,2008-10-24,6.39,6.66,6.21,6.40,67400,5.51
ABCSE,B7J,2008-10-23,6.95,6.95,6.50,6.59,59400,5.68
ABCSE,B7J,2008-10-22,6.92,7.17,6.80,6.80,55300,5.86
ABCSE,B7J,2008-10-21,7.20,7.30,7.10,7.10,54400,6.11
ABCSE,B7J,2008-10-20,6.94,7.31,6.94,7.12,45700,6.13
ABCSE,B7J,2008-10-17,6.43,6.93,6.42,6.90,57700,5.94
ABCSE,B7J,2008-10-16,6.61,6.69,6.21,6.53,83200,5.62
ABCSE,B7J,2008-10-15,6.84,6.90,6.36,6.36,78900,5.48
ABCSE,B7J,2008-10-14,7.15,7.32,6.93,6.96,74700,5.99
ABCSE,B7J,2008-10-13,6.00,6.57,6.00,6.57,75700,5.66
ABCSE,B7J,2008-10-10,5.05,5.72,4.79,5.72,158400,4.93
ABCSE,B7J,2008-10-09,6.30,6.41,6.00,6.02,140500,5.18
ABCSE,B7J,2008-10-08,5.60,6.47,5.60,6.28,292000,5.41
ABCSE,B7J,2008-10-07,7.59,7.59,6.66,6.69,89900,5.76
ABCSE,B7J,2008-10-06,7.83,7.90,7.00,7.40,159600,6.37
```

# EXECUTION TIME

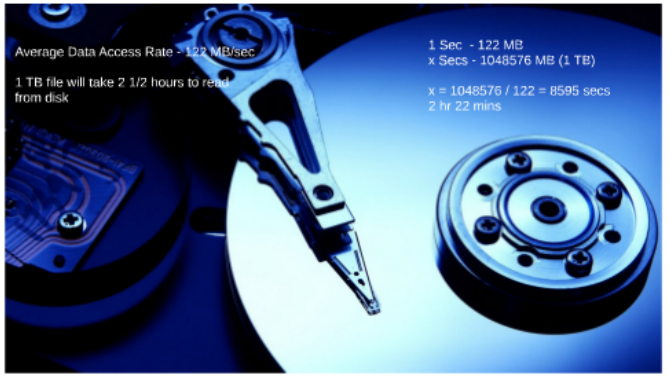
**Data access rate**

+

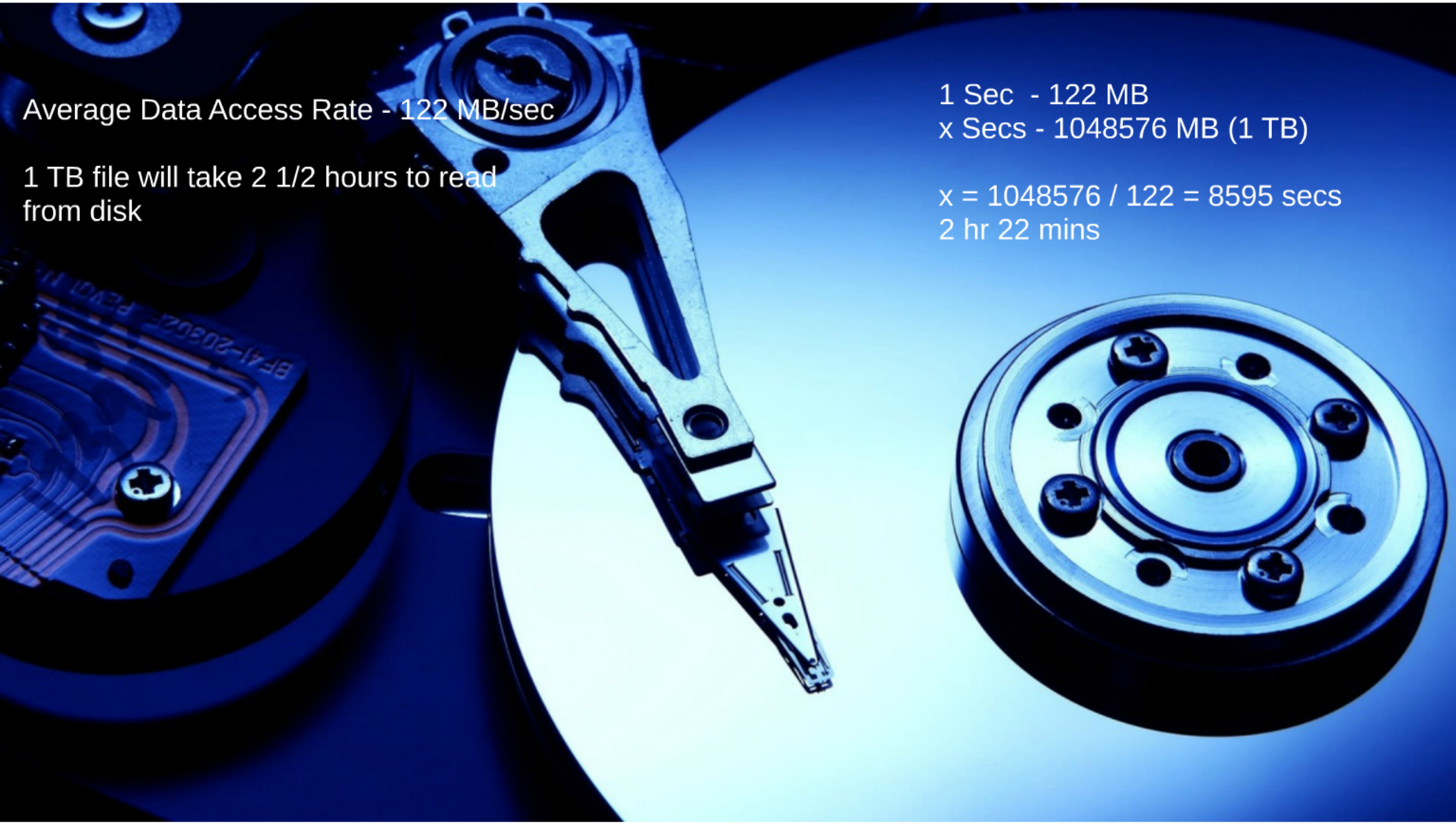
**Program computation time (~60 mins)**

+

Network Bandwidth.. etc..




> 3 hrs


Average Data Access Rate - 122 MB/sec
1 TB file will take 2 1/2 hours to read from disk
1 Sec - 122 MB
x Secs - 1048576 MB (1 TB)
x = 1048576 / 122 = 8595 secs
2 hr 22 mins

# EXECUTION TIME

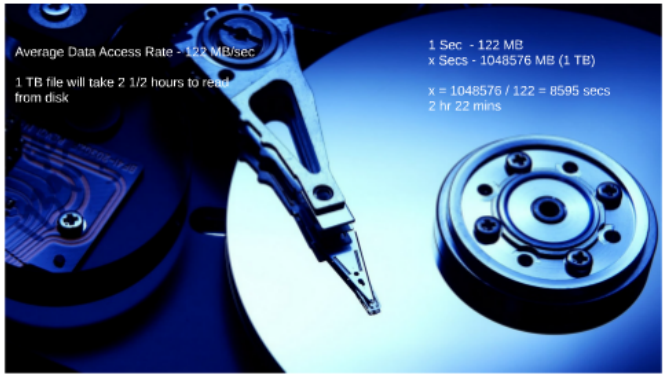
**Data access rate**

+

**Program computation time (~60 mins)**

+

Network Bandwidth.. etc..




> 3 hrs

# HOW ABOUT THIS?

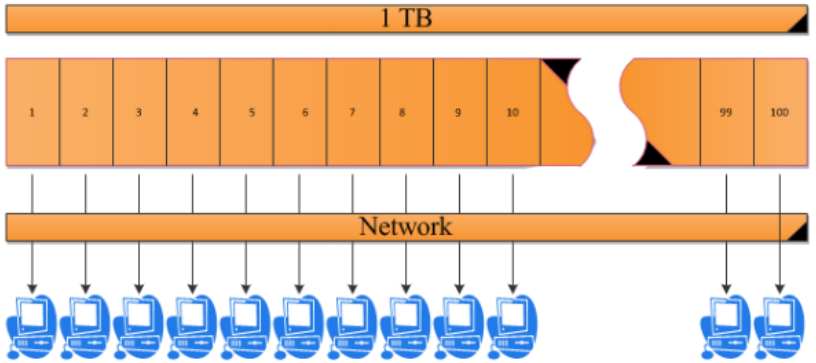
Split 1 TB file in to 100 equal sized blocks and read them parallely

Time to read = 150 mins /100

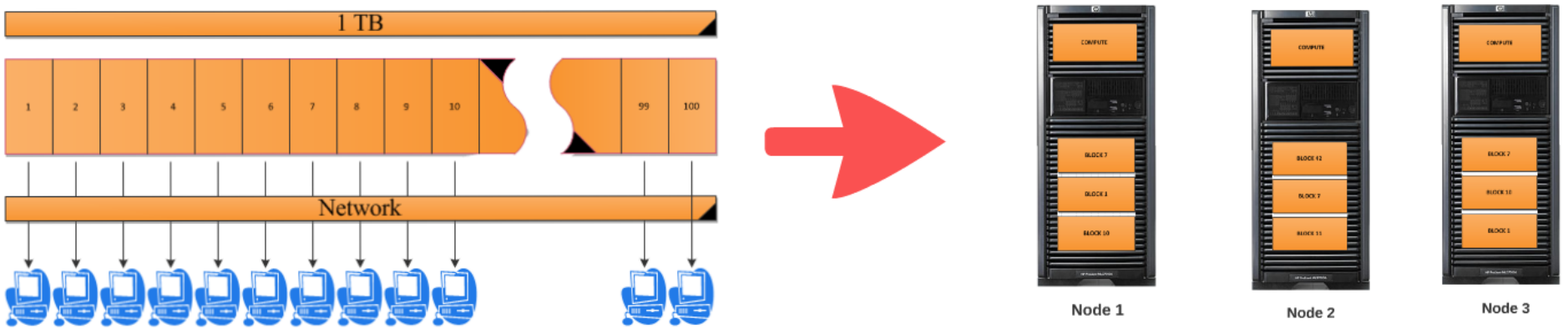
< 2 minutes

Computation Time = 60 mins /100

< 1 minute



# STORAGE CLOSER TO COMPUTATION



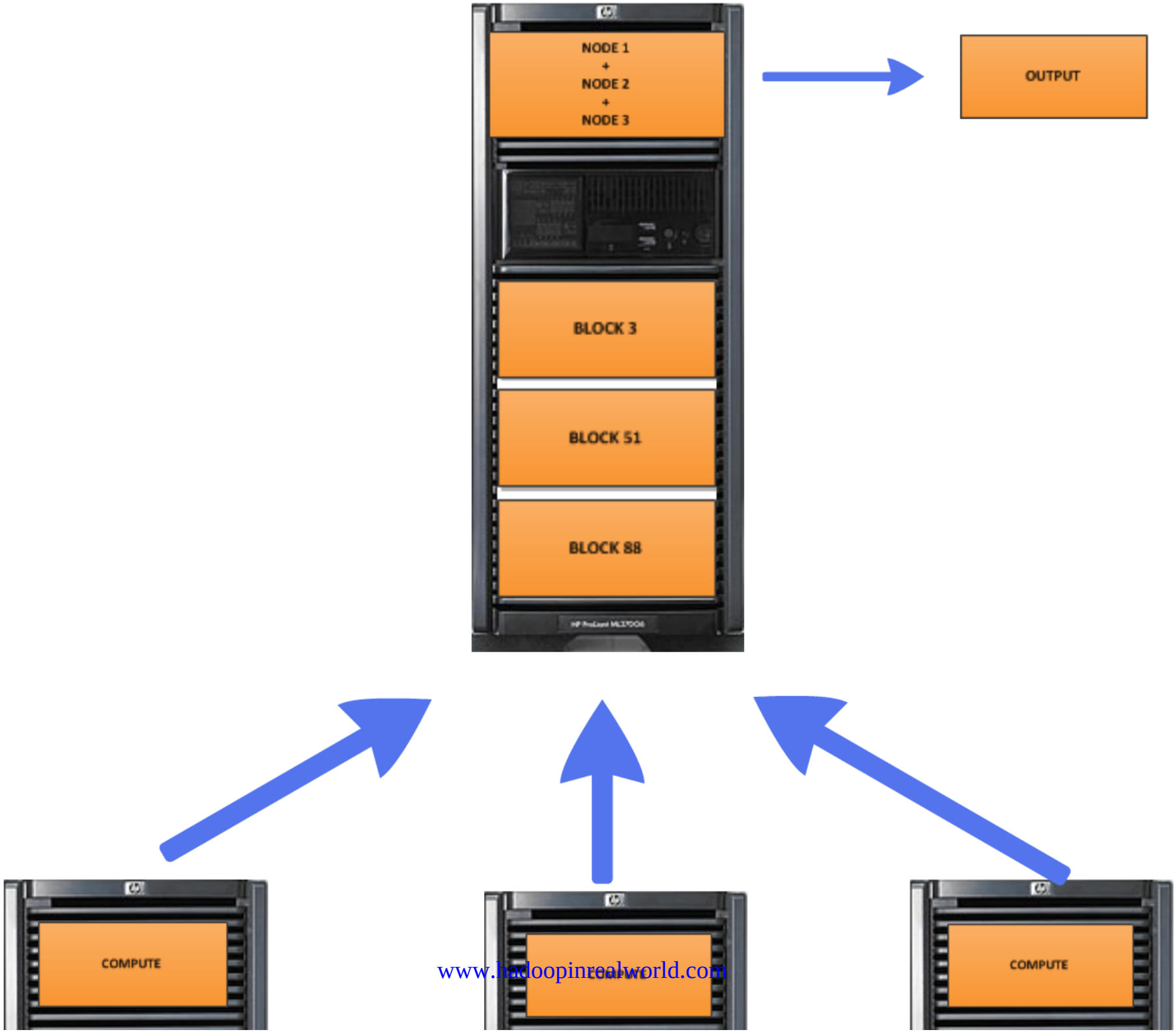
# AGGREGATE COMPUTATION

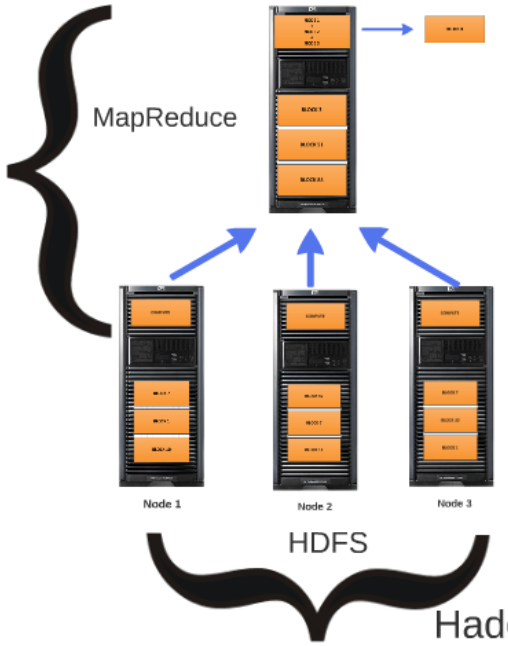

# REPLICATION


Node 1

Node 2

Node 3






**HDFS - Reliable Shared Storage**

**+**

**MapReduce - Distributed Computation**

**=**

Hadoop is a framework for distributed processing of large data sets across clusters of commodity computers