Petabyte-Scale Processing with Hadoop

An Open Source Perspective

Jimmy Lin
September 2

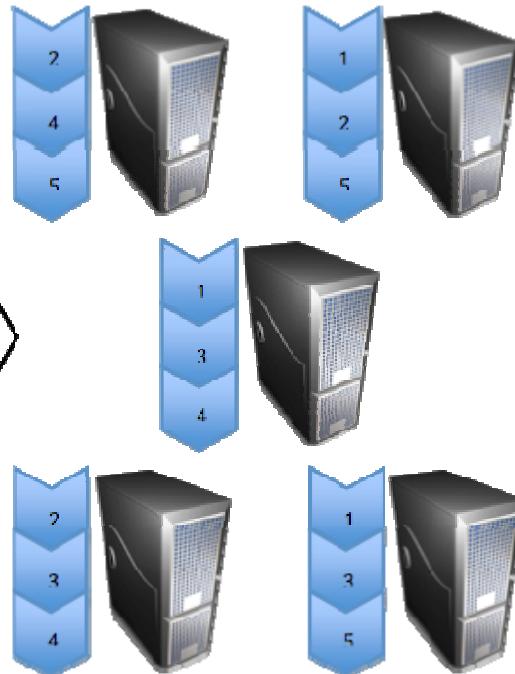# tabyte-Scale Processing: Big Ideas

ommodity hardware

hare-nothing architecture

ove data to processors

doop: HDFS and MapReduce
(n source embodiment of those ideas)

DFS

HDFS →

MapReduce

Input key*value pairs

Input key*value pairs

Data store 1 … Data store n

map

map

(key 1, values...) (key 2, values...) (key 3, values...)

(key 1, values...) (key 2, values...)

== Barrier == : Aggregates intermediate values by output key

key 1, intermediate values

key 2, intermediate values

ke interm va

reduce

reduce

reduce

final key 1 values

final key 2 values

final key 3 values
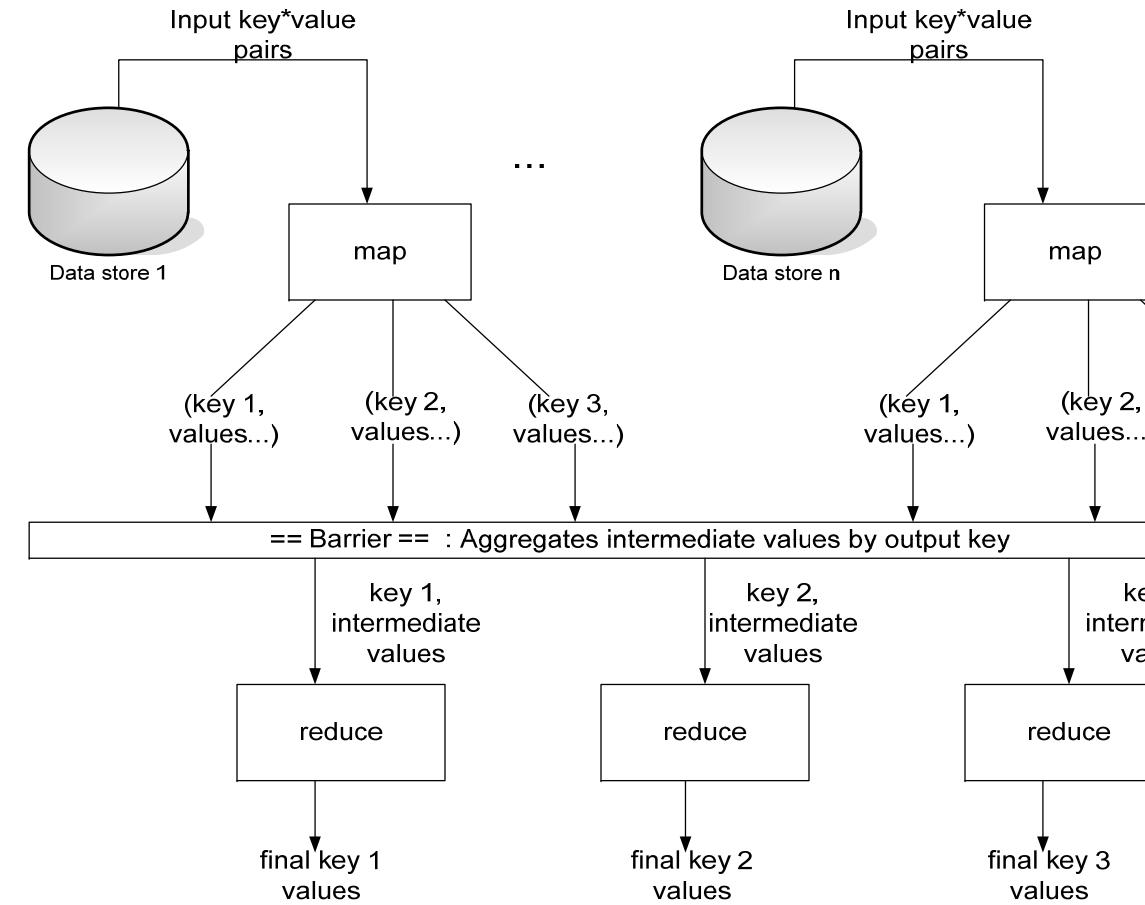
• stribute data blocks among servers
• plication for reliability

• High-level programming abstraction
• Execution framework handles "detai

brant Open Source Ecosystem

right tool for the job)

ve

g

Base

okeeper

se Study 1: Yahoo

adoop runs production "Webmap"

ata Points

er 25,000 nodes running Hadoop

ndreds of thousands of jobs per day

pical HDFS cluster: 1,400 nodes, 2 PB capacity

enchmarks (May, 2009)

B sort: 1,460 nodes in 62 seconds

B sort: 3,658 nodes in 16.25 hours

se Study 2: Facebook

ata Points

- 0+ servers running Hadoop
- PB under Hadoop/Hive management, +15 TB new data per day

hat for?

- ta collection: server logs, web crawls, etc.
- ocessing pipeline: ad optimization, summaries, etc.
- hoc analyses

w much does one petabyte cost?

sidering hardware only…)

aw disks: ~$100k

B SATA drive, ~$100

un X4540: ~$1m

unit: 2x6-cores, 32 GB RAM, 48 TB disk, ~$48k (21 units)

2 cores, 672 GB RAM

adoop cluster: ~$336k

mmodity server: 2x4-cores, 16 GB RAM, 12 TB disk, ~$4k (84 u

2 cores, 1.3 TB RAM

w much does one petabyte cost?

 consider software now…)

arallel database vender: ~$50m

pical pricing: $50k per TB

adoop: $0

wever, Cloudera would love to support you…