Hadoop/MapReduce: Translating DB Operations to Hadoop Jobs

DB Operations

- Select
- Projection
- Grouping and Aggregation
- Duplicate Elimination
- Join

Selection: σ

- **Select: σ_c *(R)*:**
  - Select subset of tuples from R that satisfy **selection condition** *c*

R

A	B	C	D
α	α	1	7
α	β	5	7
β	β	12	3
β	β	23	10

$\sigma_{((A=B) \wedge (D>5))}(R)$

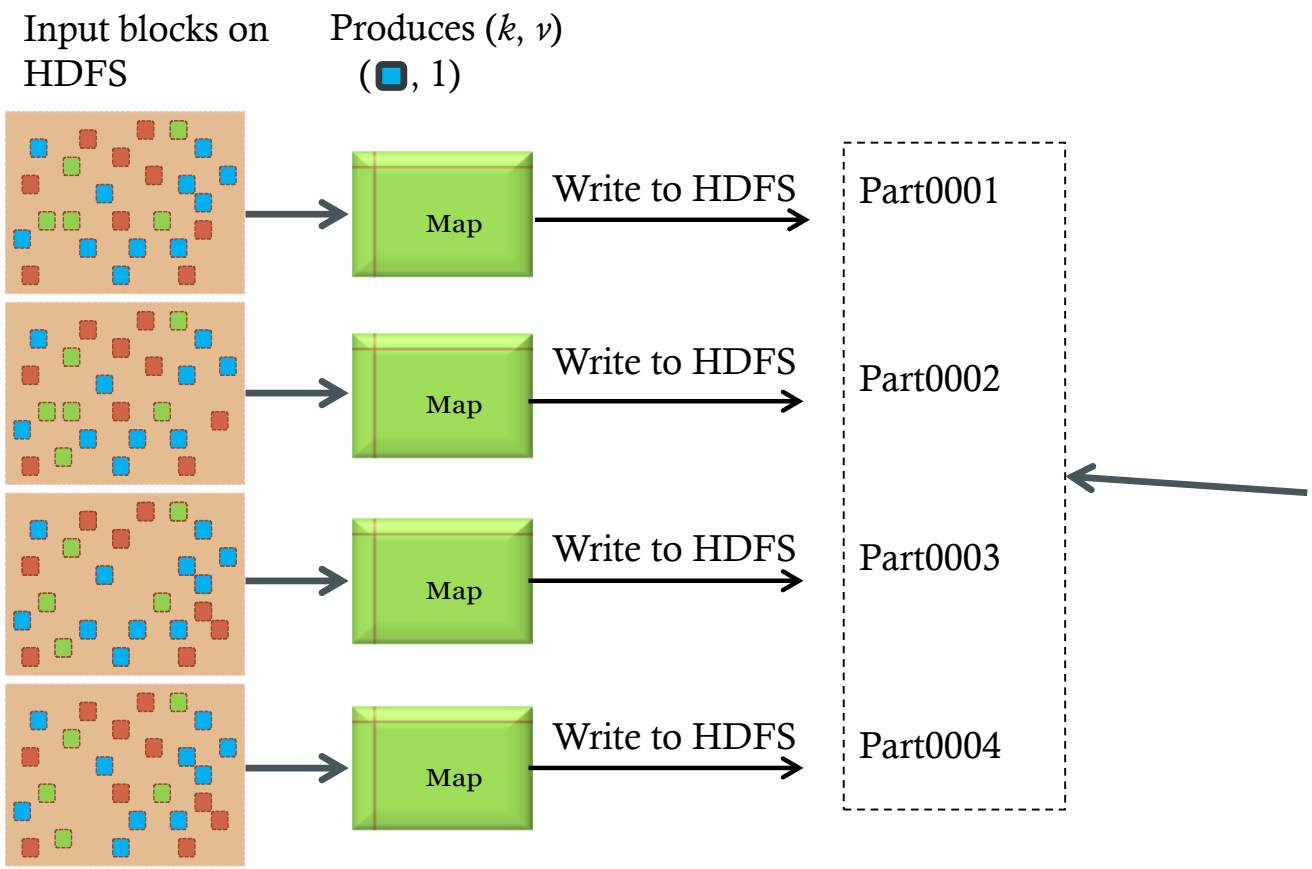
A	B	C	D
α	α	1	7
β	β	23	10

```
Select *
From R
Where R.A = R.B
And R.D > 5;
```

In Hadoop, Selection is implemented as a Map-Only Job

Remember: Color Filter

Job: Select only the blue and the green colors



Projection: π

- $\pi_{A1, A2, \ldots, An}(R)$
  - returns all tuples in R, but only columns A1, A2, …, An

Rename column A to V

Compute this expression and call it X

$\pi_{C,\ V \leftarrow A,\ X \leftarrow C*3+B}(R)$

Select C, A as V, C*3+B as X
From R;

R

A	B	C
1	2	5
3	4	6
1	2	7
1	2	8

C	V	X
5	1	17
6	3	22
7	1	23
8	1	26

In Hadoop, Projection is implemented as a Map-Only Job

Grouping & Aggregation

- **Aggregation function** takes a collection of values from a GROUP of records and returns a single value for that group:
  - **avg**: average value
  - **min**: minimum value
  - **max**: maximum value
  - **sum**: sum of values
  - **count**: number of values
- **Grouping & Aggregate operation** in relational algebra
  - $\gamma_{g1,g2,\ldots gm,\ F1(A1),\ F2(A2),\ \ldots Fn(An)}(R)$

Grouping & Aggregation Operator: Example

R

A	B	C
α	α	7
α	β	7
β	β	3
β	β	10

S

branch_name	account_number	balance
Perryridge	A-102	400
Perryridge	A-201	900
Brighton	A-217	750
Brighton	A-215	750
Redwood	A-222	700

$\gamma_{sum(c)}(R)$

sum(c)
27

$\gamma_{branch\_name,sum(balance)}(S)$

branch_name	sum(balance)
Perryridge	1300
Brighton	1500
Redwood	700

In Hadoop, Grouping & Aggregation is implemented as a Map-Reduce Job

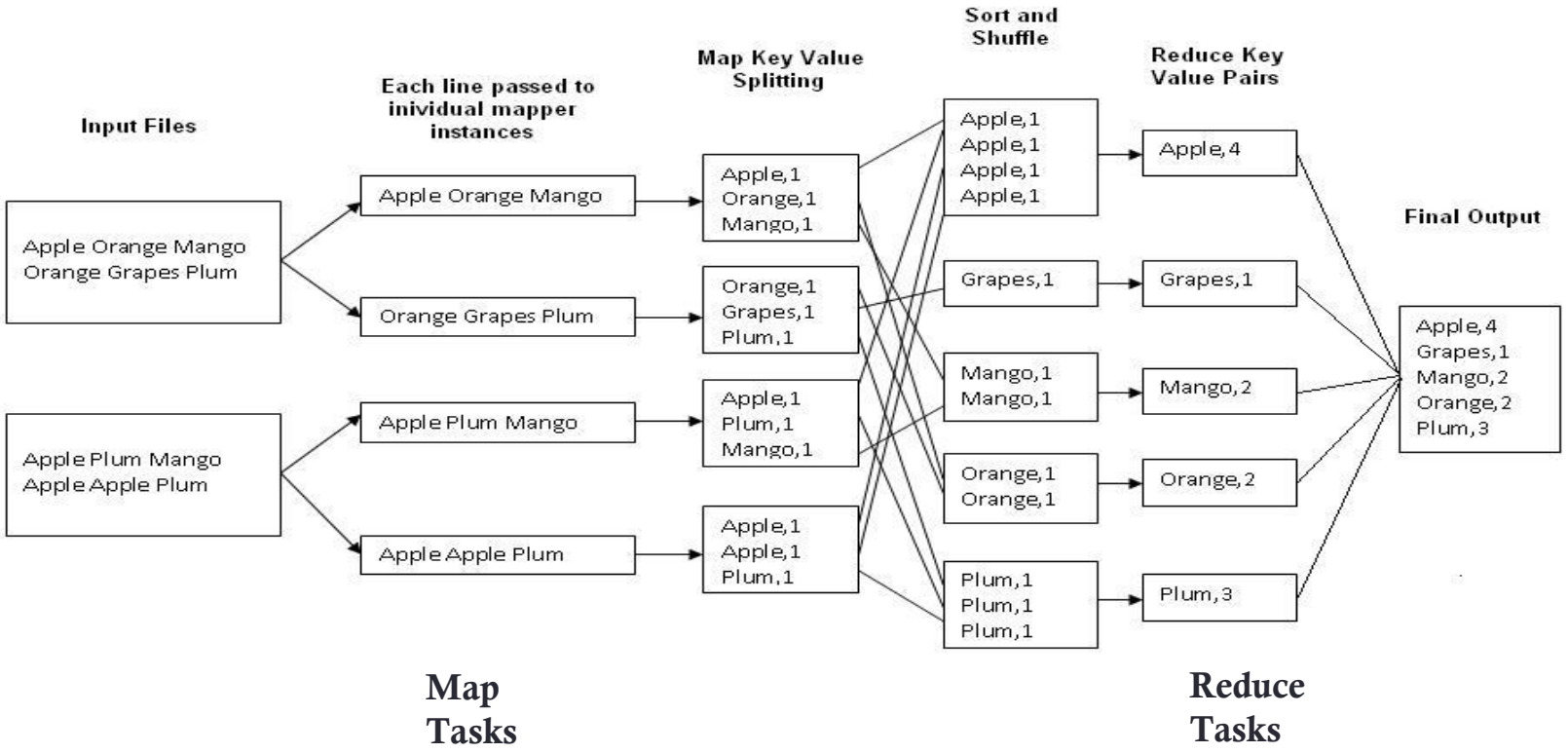
Select sum(C)
From R;

Select sum(balance)
From S
(

What is the key/value?

Back to Word Count

- **Job: Count the occurrences of each word in a data set**



Duplicate Elimination: δ (R)

- Delete all duplicate records

Select Distinct *
From R;

In Hadoop, duplicate elimination is implemented as a Map-Reduce Job

R

A	B
1	2
3	4
1	2
1	2

d (R)

A	B
1	2
3	4

What is the key/value?

Map (Key= hash code of the tuple, Value= tuple itself).

Join: $R \bowtie_C S$

R

A	B
1	2
3	2

S

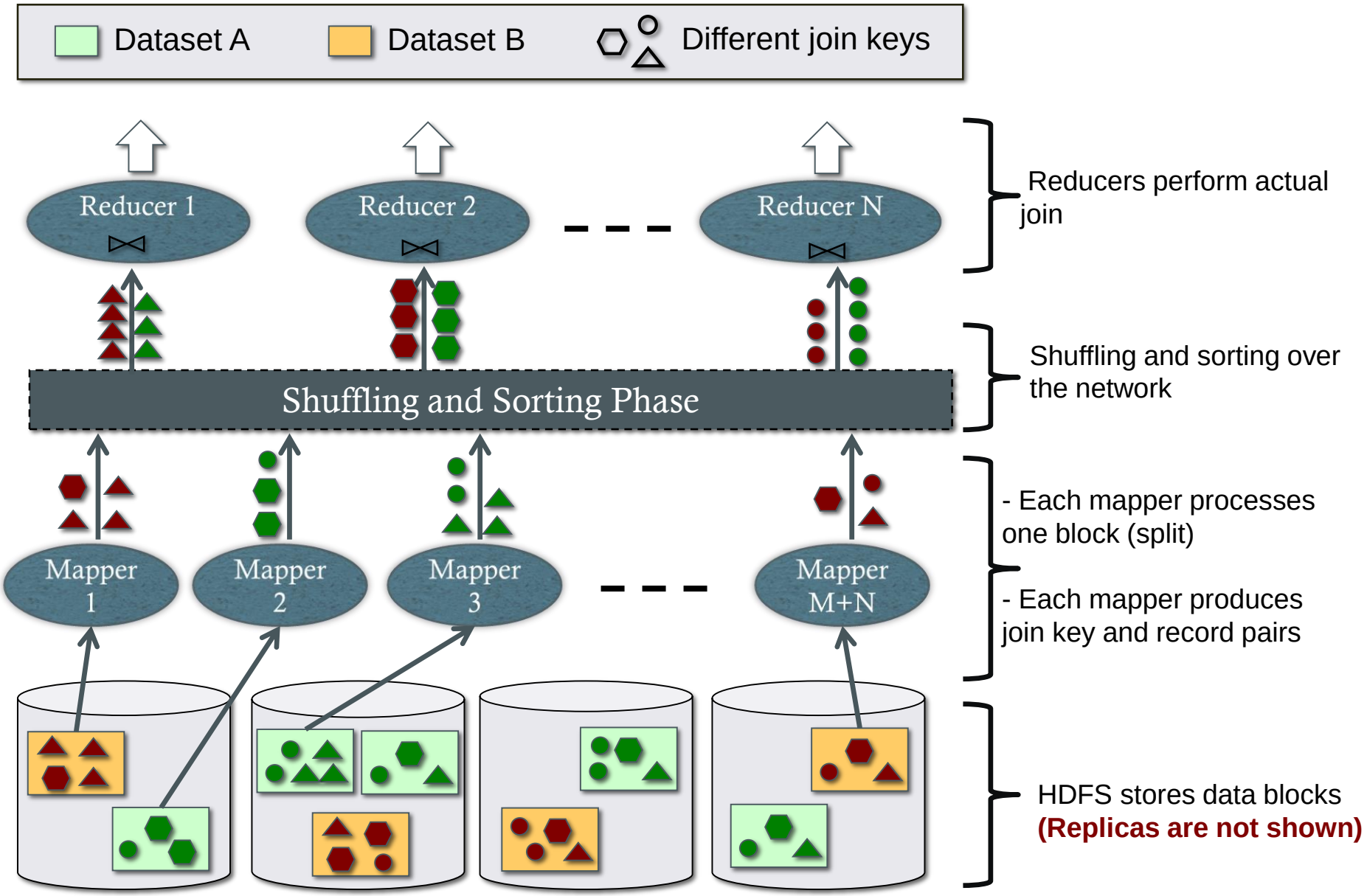
D	C
2	3
4	5
4	5

$R \bowtie_{R.A>=S.C} S$

A	B	D	C
3	2	2	3

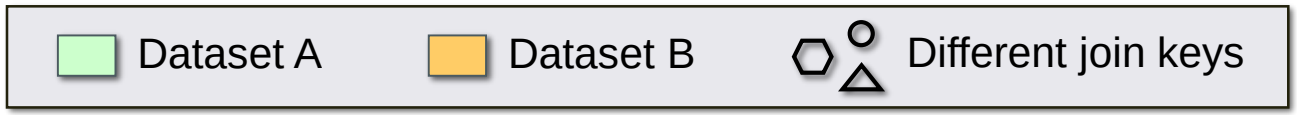
Several alternate join logic and algorithms exist.
Several variations of Hadoop Join implementations possible.

Joining Two Large Datasets: Re-Partition Join

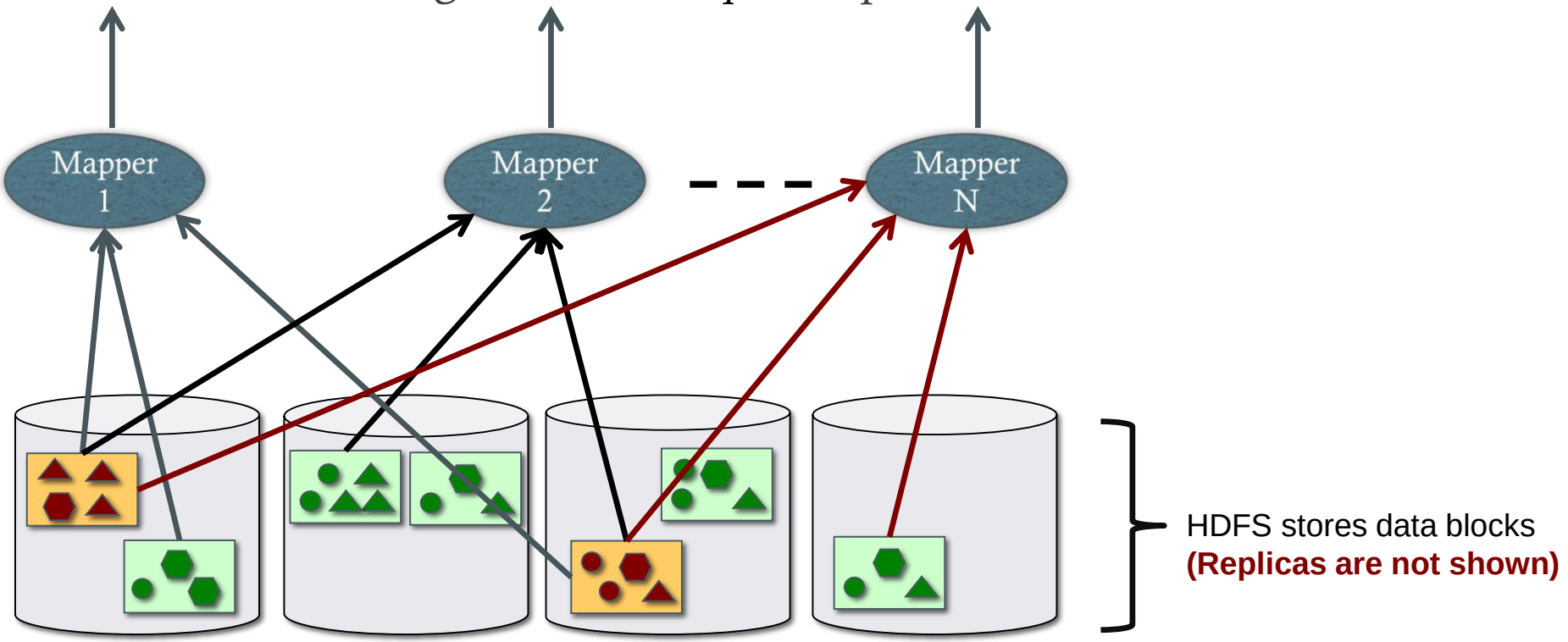


Joining Large Dataset (A) with Small Dataset (B)

Broadcast/Replication Join



- Every map task processes one block from A and the entire B
- Every map task performs the join ***(MapOnly job)***
- Avoid the shuffling and reduce *expensive* phases



Translating DB Operations to Hadoop Jobs (**Summary**)

- Select (Filter) → Map-only job
- Projection → Map-only job
- Grouping and aggregation → Map-Reduce job
- Duplicate Elimination → Map-Reduce job
- Join → Map-Reduce job