data integration
and processing

data goes
through
operations

- apply function
- work from one format to another
- join data with other datasets
- filter values out of a data set

Transformations
*some are
aggregations

Data Transformations

After this video you will be able to..

- List common data transformations within big data pipelines

- Design a conceptual data processing pipeline using the basic data transformations

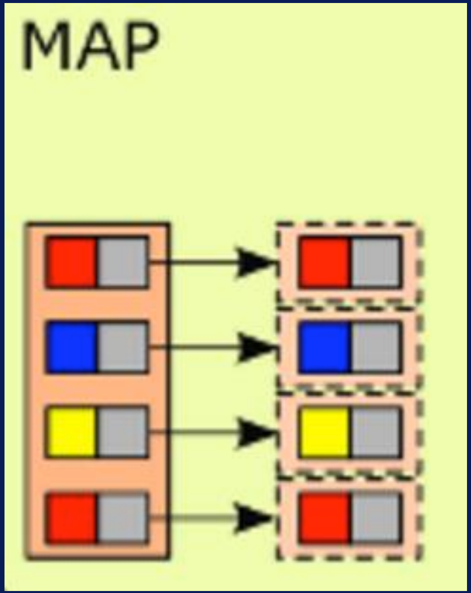
Transformation are Tools to shape your data

functions

convert from one form to another

⊕

Map — basic



Apply same operation to each member of a collection

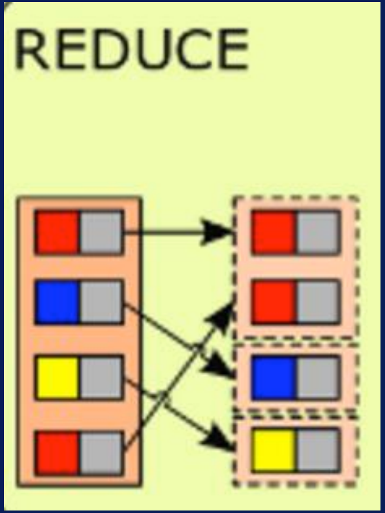
- Color each member of a set
- Discount each product's price by 5%
- Apply formatting to each document in a folder

- each data set is executed separately

same opperation to each element

Image from: http://www.stratosphere.eu

Reduce



‘Collecting’ things that have same ‘key’

Key: Colors

- Collect blocks as per their colors

- collectively apply the same process to objects of similar nature

Word Count Example

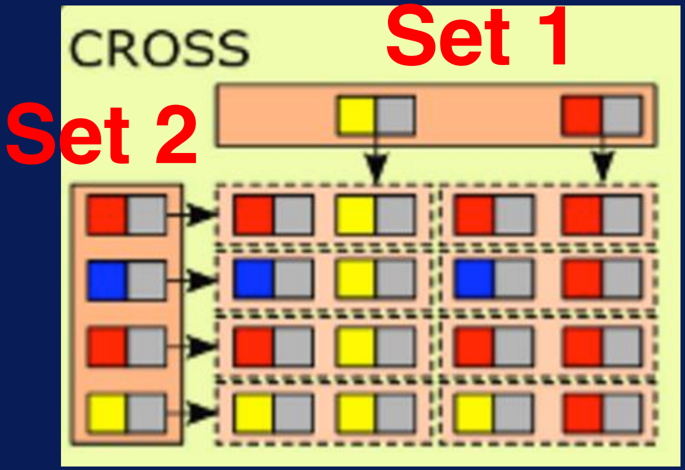
Key: Words

- Sum frequency counts of words

* map and reduce are types of transformations that work on a single list of key and data pairings

Image from: http://www.stratosphere.eu

Cross/ Cartesian

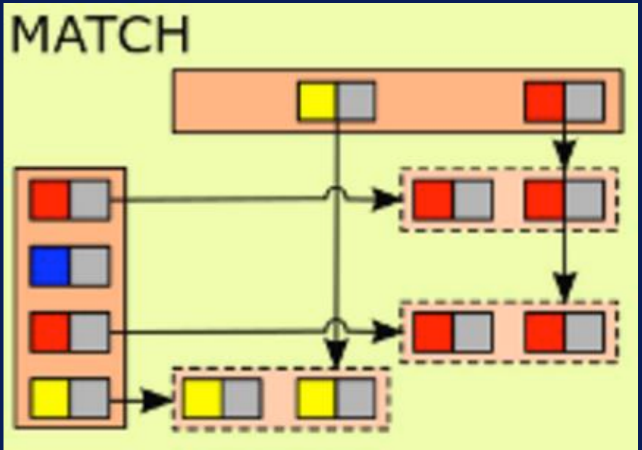


→ Multiplication

Do some process to each pair from two sets

Image from: http://www.stratosphere.eu

Match/Join



→ Selective Multiplication

Do some process to each <u>pair</u> from two sets – which have same ‘key’

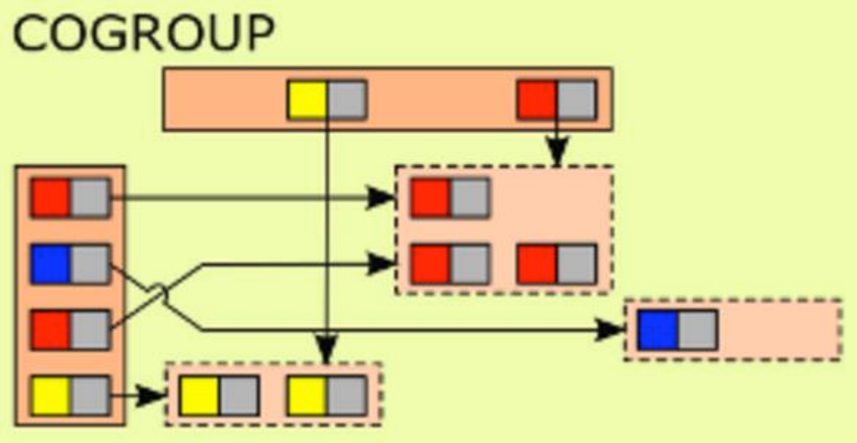
just grouping together
the data partitions with the same key

* selective in forming pairs – every pair
must have something in common

Key

Image from: http://www.stratosphere.eu

Co-Group



- Collect similar things first
- Apply a process to each collection

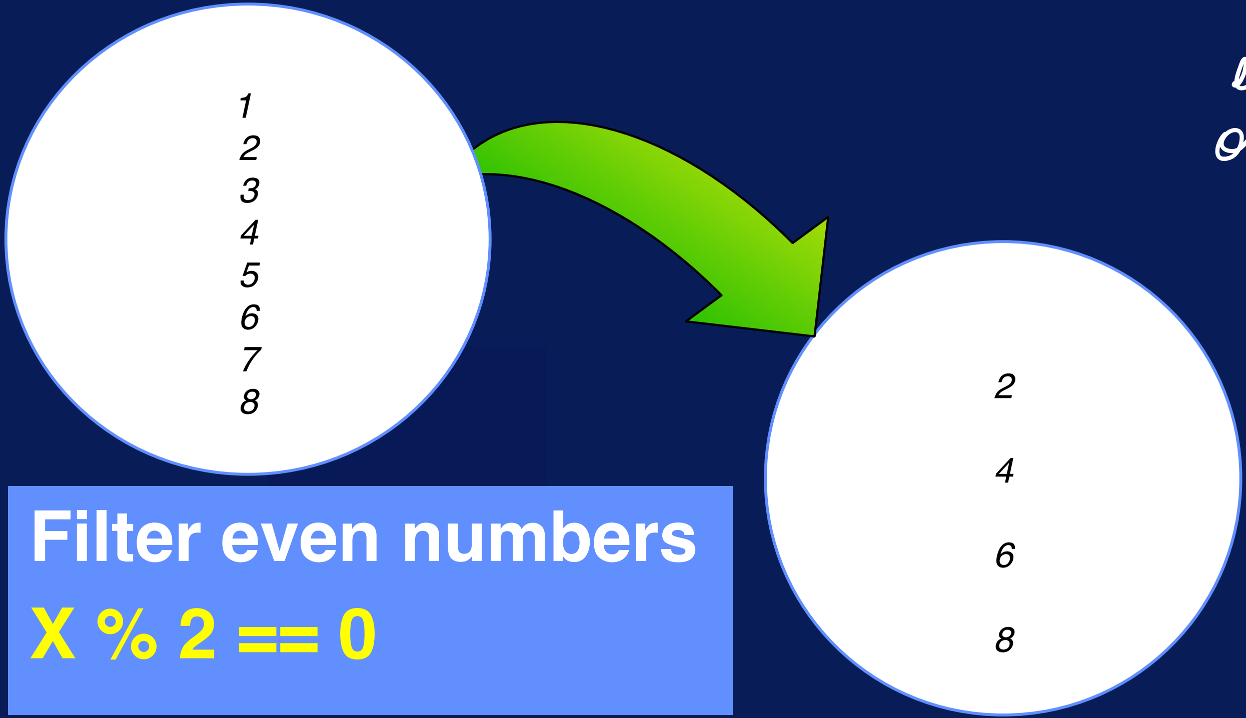
→ Group common items

listing even if they
dont exist in both
datasets

Image from: http://www.stratosphere.eu

Filter

Select elements that match a criteria

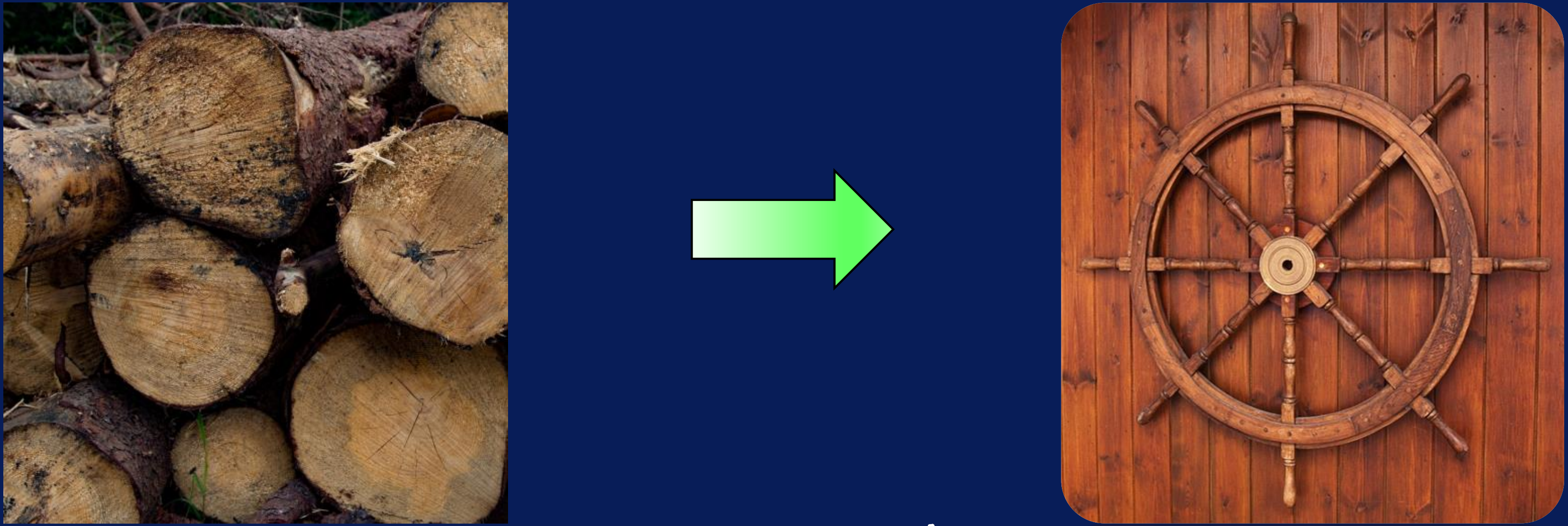


Filter even numbers

X % 2 == 0

like a test,
only elements that
pass the test
are shown

Basic Transformations → Get Results



effectiveness of transformation is in pipelining them in a way that helps solve problem

↳ as you would perform a series of tasks on a real block