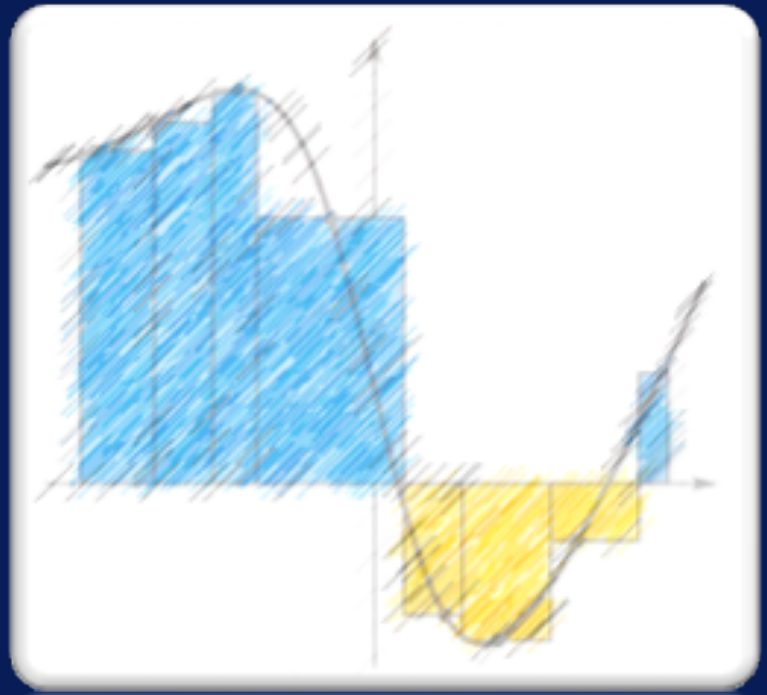
Aggregations in Big Data Pipelines

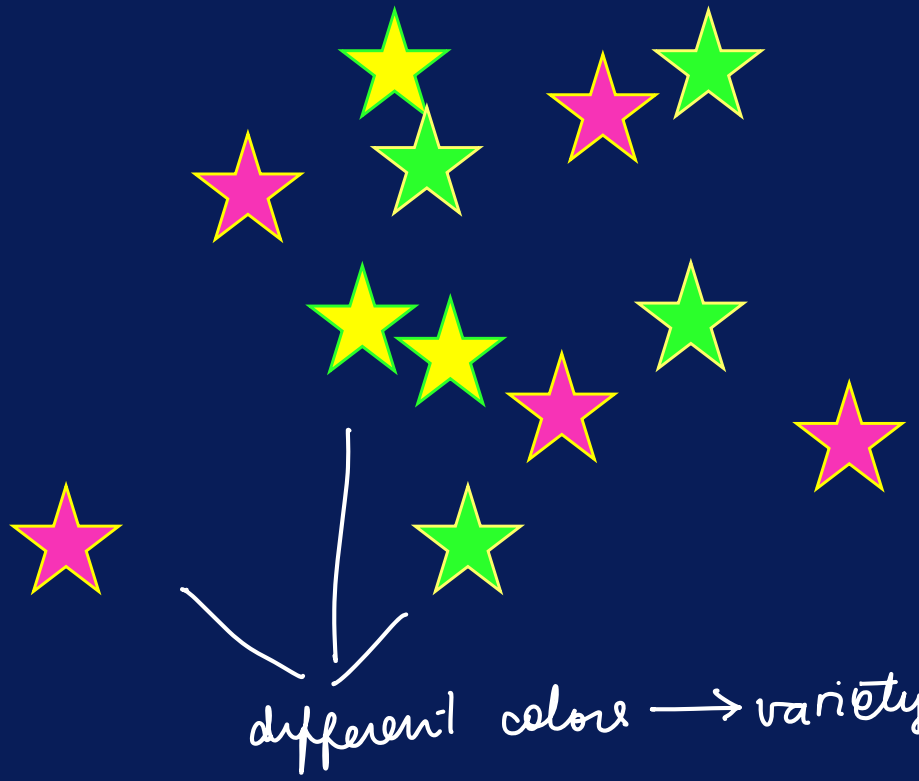


After this video you will be able to..

- Compare and select the Aggregation operation that you require to solve your problem
- Explain how you can use Aggregations to compact your dataset and reduce volume (in many cases)
- Design complex operations in your pipeline using a series of Aggregations

What is Aggregation ?

- any operation on a data set that performs a specific transformation taking all related data elements into consideration



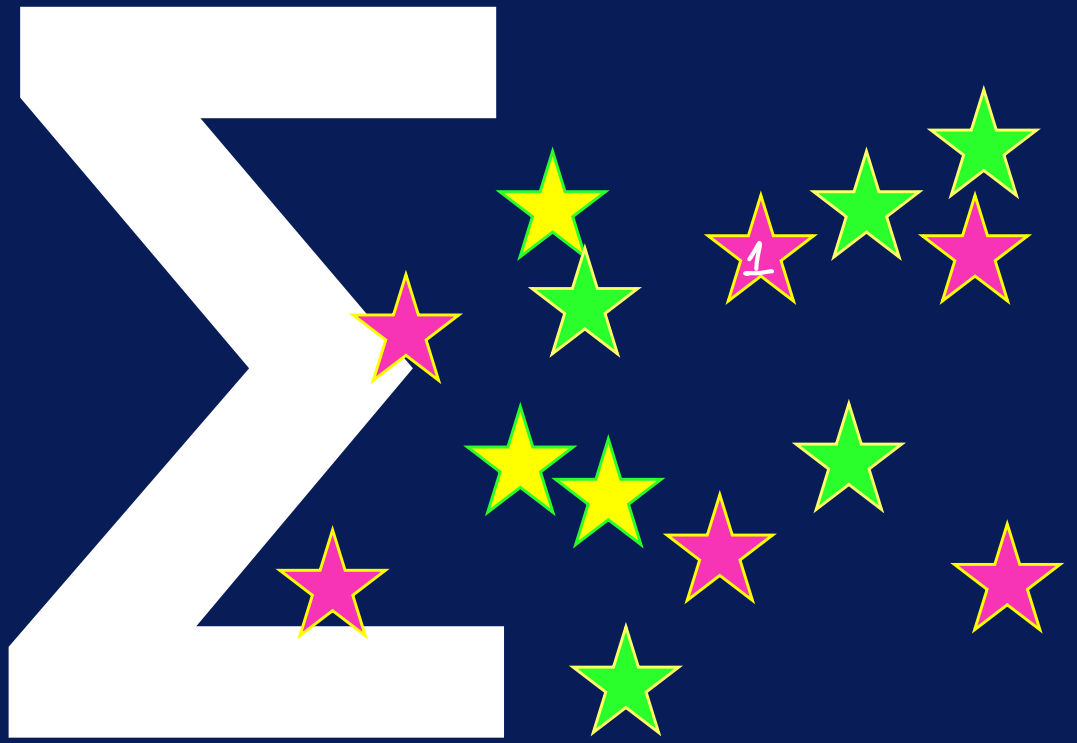
Symbol for any transformation

f can take the shape of different transformations

Aggregation → f (all elements)

applying a transformation f that takes all the elements of data as input

Aggregation → f (all elements)



f

14

Symbol for summation

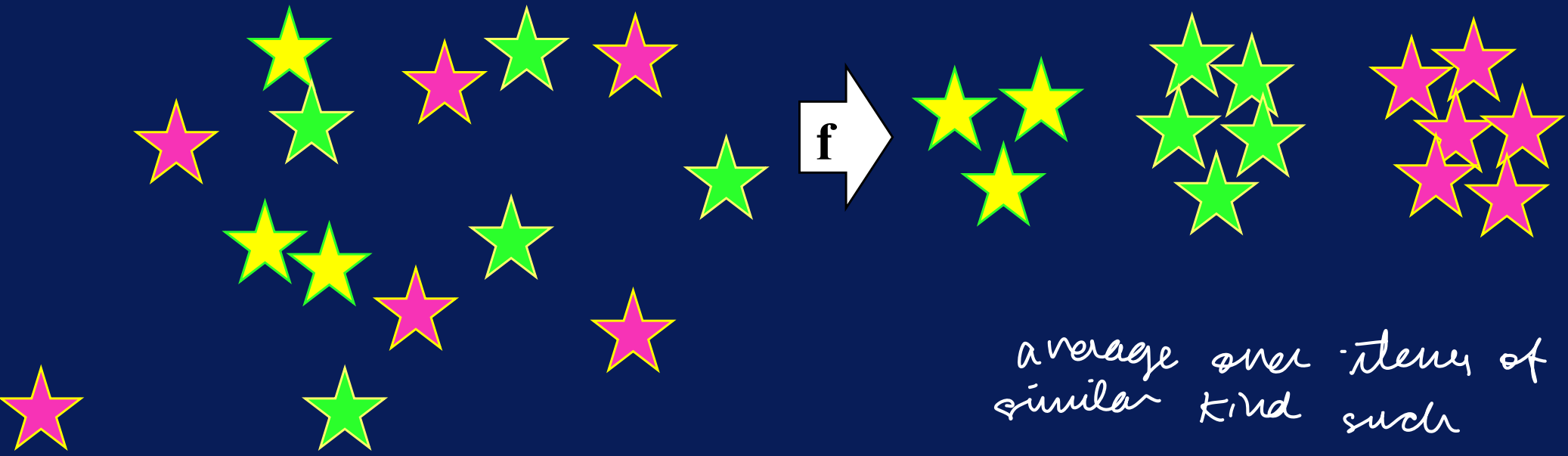
one of the simplest aggregation

GROUP BY

summation of individual star colors

→ R: 3 tuples, (star color, count) >

AVERAGE # PER COLOR



average over items of similar kind such

OTHER TRANSFORMATIONS

MAX

MIN

STANDARD DEVIATION

Connecting Aggregations

SUM → MAX → MAX(SUM)

maximum of the sum

SUM → MIN → MIN(SUM)

* can always perform aggregation as a series of operations

BOOLEAN AGGREGATION

AND 10110100110101101011011010111010  0

OR 10110100110101101011011010111010  1

SETS

* don't allow duplicate values

STRINGS

UNION

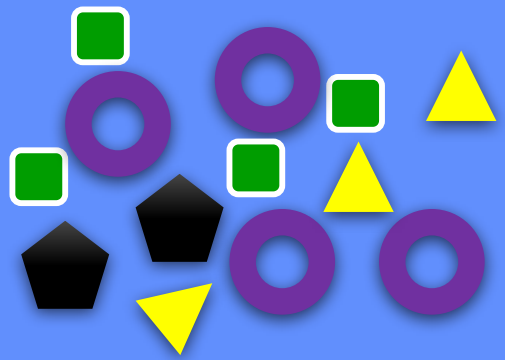
INTERSECTION

DIFFERENCE

CONCATENATION

Aggregations → Organized & Compact Data

↳ important tool when dealing with large data sets



f

AGGREGATED OUTPUT

Variety & Volume

Actionable Insights

by choosing the right aggregation, you can generate compact and meaningful insights that enable faster and effective decision making in business

* in most cases aggregation results in smaller output data sets