Apache Spark

Transformations and Actions

Spark Transformations

Driver **Spark Session**

spark.read.csv()

Timestamp	Age	Gender	Country	state
2014-08-27 11:29:31	37	Female	United States	IL
2014-08-27 11:29:37	44	M	United States	IN
2014-08-27 11:29:44	32	Male	Canada	
2014-08-27 11:29:46	31	Male	United Kingdom	
2014-08-27 11:30:22	31	Male	United States	TX

select(Age, Gender, Country, state)

Age	Gender	Country	state
37	Female	United States	IL
44	M	United States	IN
32	Male	Canada	
31	Male	United Kingdom	
31	Male	United States	TX

filter(Age<40)

Age	Gender	Country	state
37	Female	United States	IL
32	Male	Canada	
31	Male	United Kingdom	
31	Male	United States	TX

groupBy(country) .count()

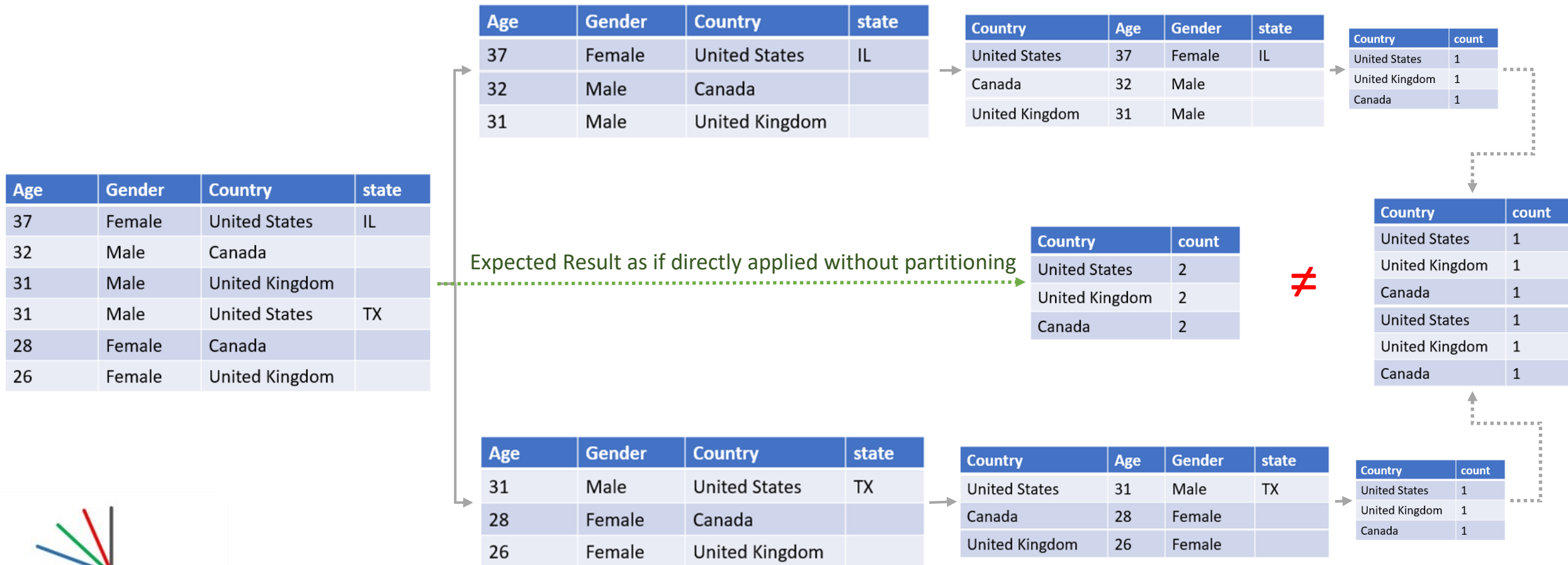| Country	Count
United States	2
Canada	1
United Kingdom	1
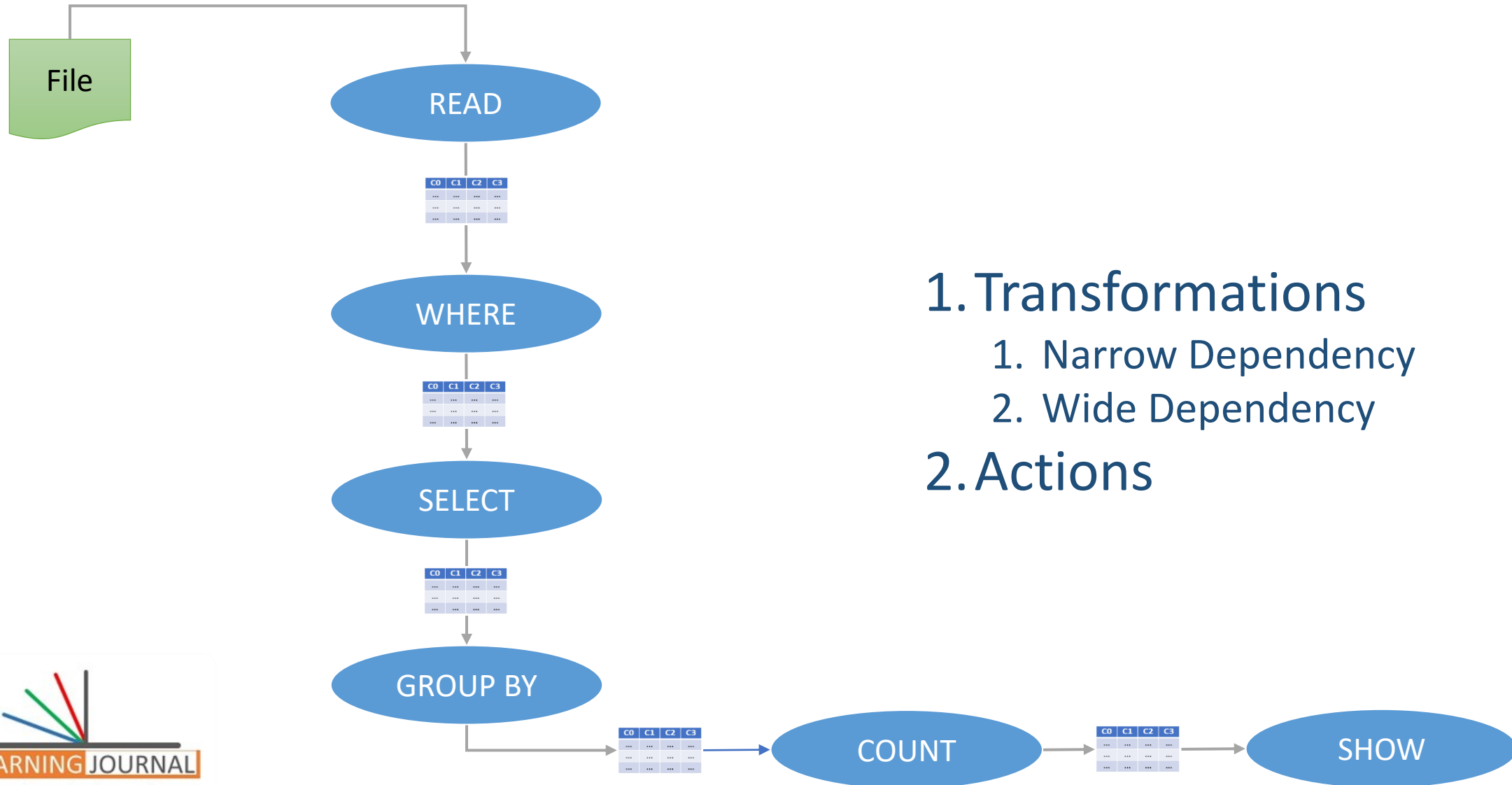
Narrow Dependency Transformation

A transformation performed independently on a single partition to produce valid results.

Age	Gender	Country	state
37	Female	United States	IL
44	Male	United States	IN
32	Male	Canada	
31	Male	United Kingdom	

where(Age<40)

Age	Gender	Country	state
37	Female	United States	IL
32	Male	Canada	
31	Male	United Kingdom	

Age	Gender	Country	state
37	Female	United States	IL
44	Male	United States	IN
32	Male	Canada	
31	Male	United Kingdom	
31	Male	United States	TX
45	Female	United Kingdom	
28	Female	Canada	
26	Female	United Kingdom	

Correct Result as if directly applied without partitioning

Age	Gender	Country	state
37	Female	United States	IL
32	Male	Canada	
31	Male	United Kingdom	
31	Male	United States	TX
28	Female	Canada	
26	Female	United Kingdom	

Age	Gender	Country	state
31	Male	United States	TX
45	Female	United Kingdom	
28	Female	Canada	
26	Female	United Kingdom	

where(Age<40)

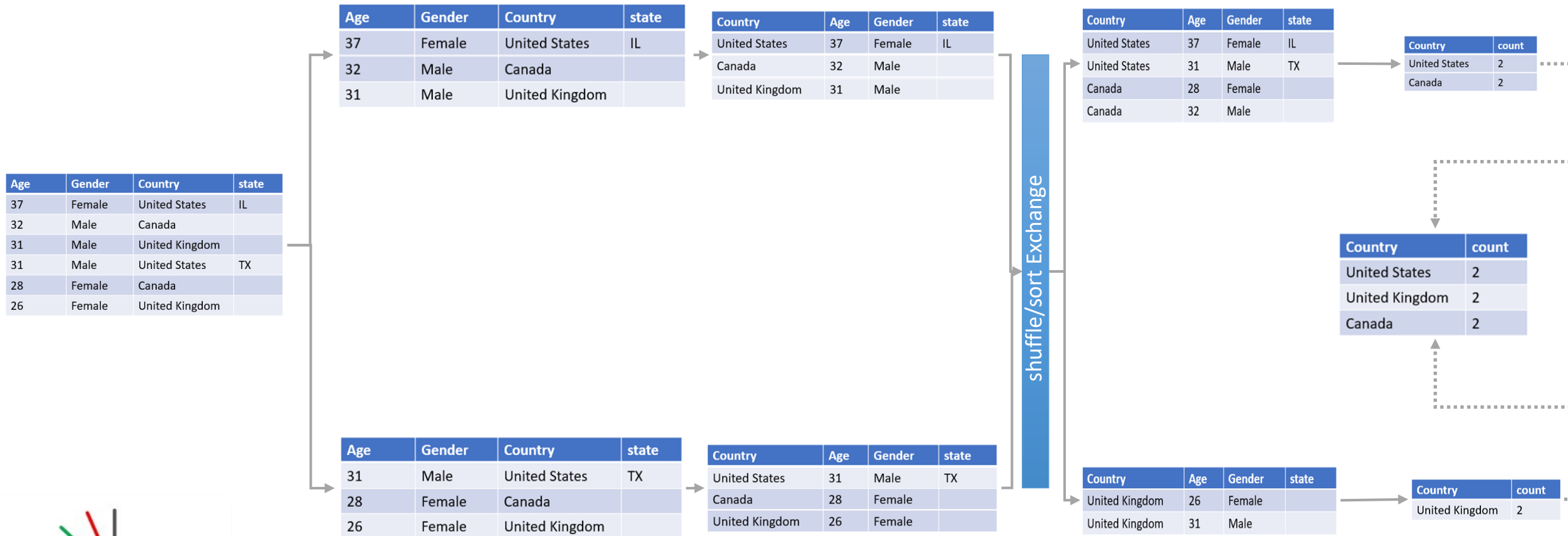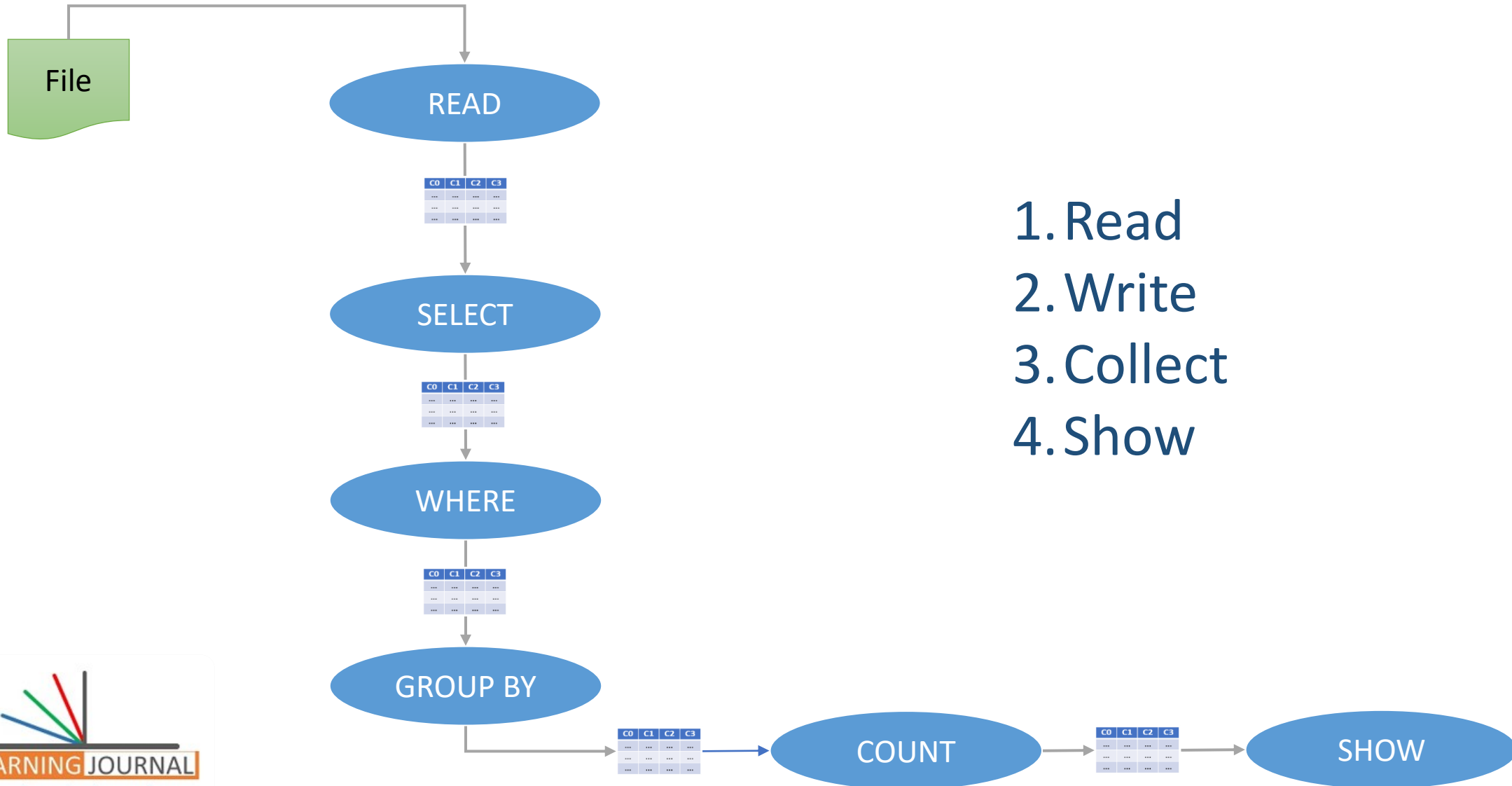| Age	Gender	Country	state
31	Male	United States	TX
28	Female	Canada	
26	Female	United Kingdom	

Wide Dependency Transformation

A transformation that requires data from other partitions to produce valid results.

Age	Gender	Country	state
37	Female	United States	IL
32	Male	Canada	
31	Male	United Kingdom	

Country	Age	Gender	state
United States	37	Female	IL
Canada	32	Male	
United Kingdom	31	Male	

Country	count
United States	1
United Kingdom	1
Canada	1

Age	Gender	Country	state
37	Female	United States	IL
32	Male	Canada	
31	Male	United Kingdom	
31	Male	United States	TX
28	Female	Canada	
26	Female	United Kingdom	

Expected Result as if directly applied without partitioning

Country	count
United States	2
United Kingdom	2
Canada	2

≠

Country	count
United States	1
United Kingdom	1
Canada	1
United States	1
United Kingdom	1
Canada	1

Age	Gender	Country	state
31	Male	United States	TX
28	Female	Canada	
26	Female	United Kingdom	

Country	Age	Gender	state
United States	31	Male	TX
Canada	28	Female	
United Kingdom	26	Female	

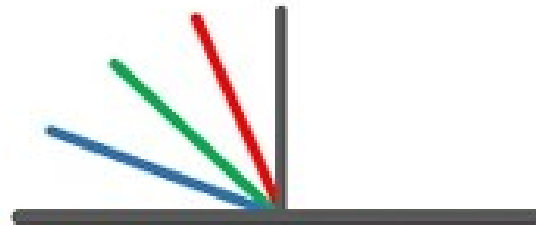| Country	count
United States	1
United Kingdom	1
Canada	1

Wide Dependency Transformation

A transformation that requires data from other partitions to produce valid results.

www.learningjournal.guru