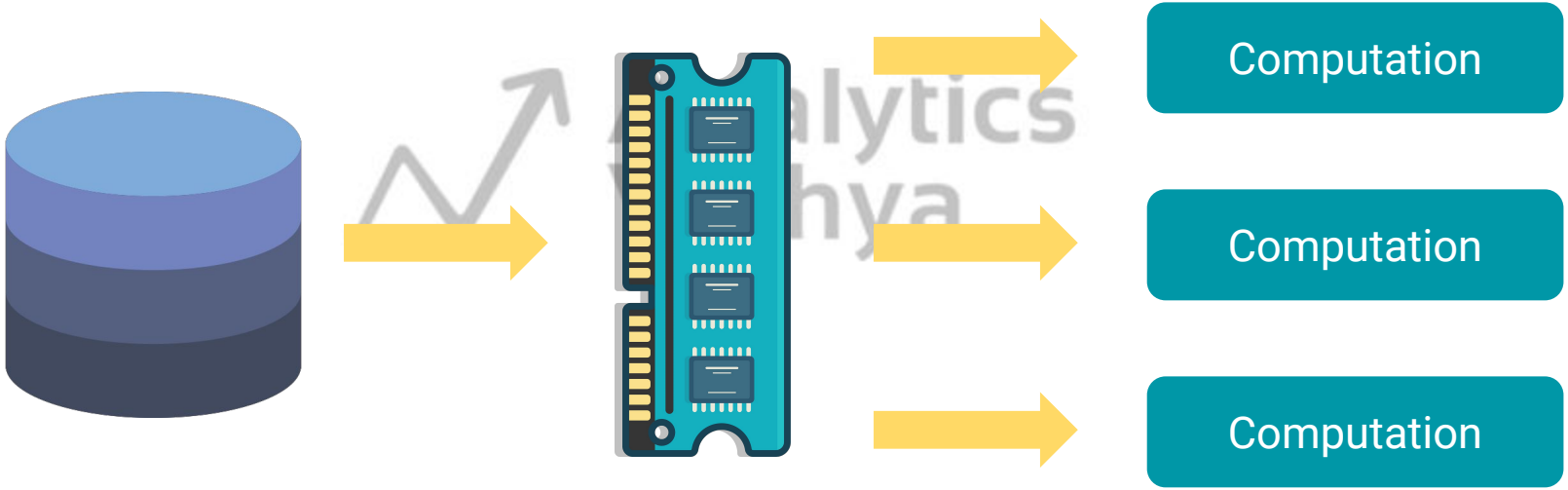
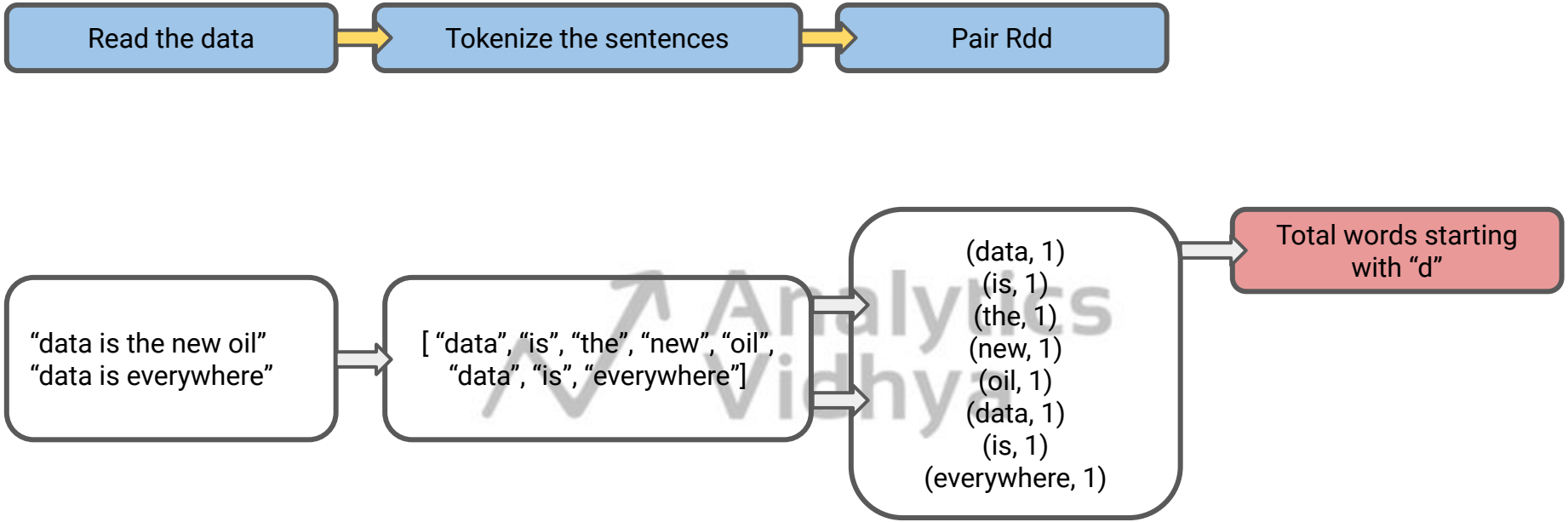
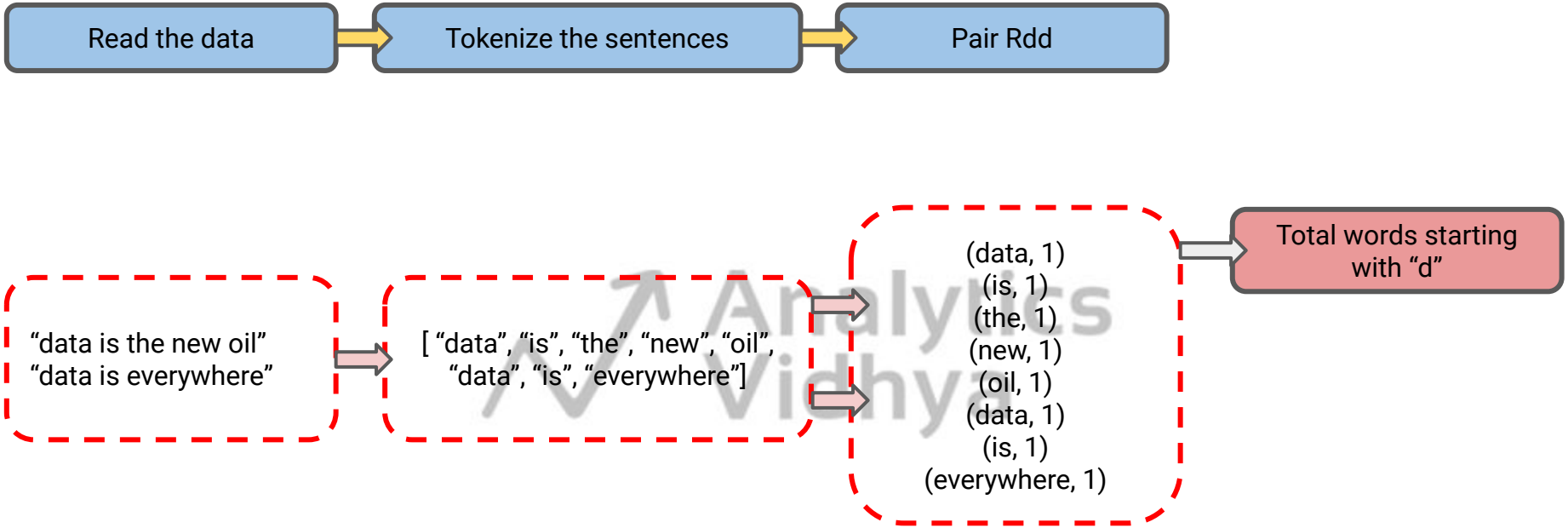
Caching and Persistence

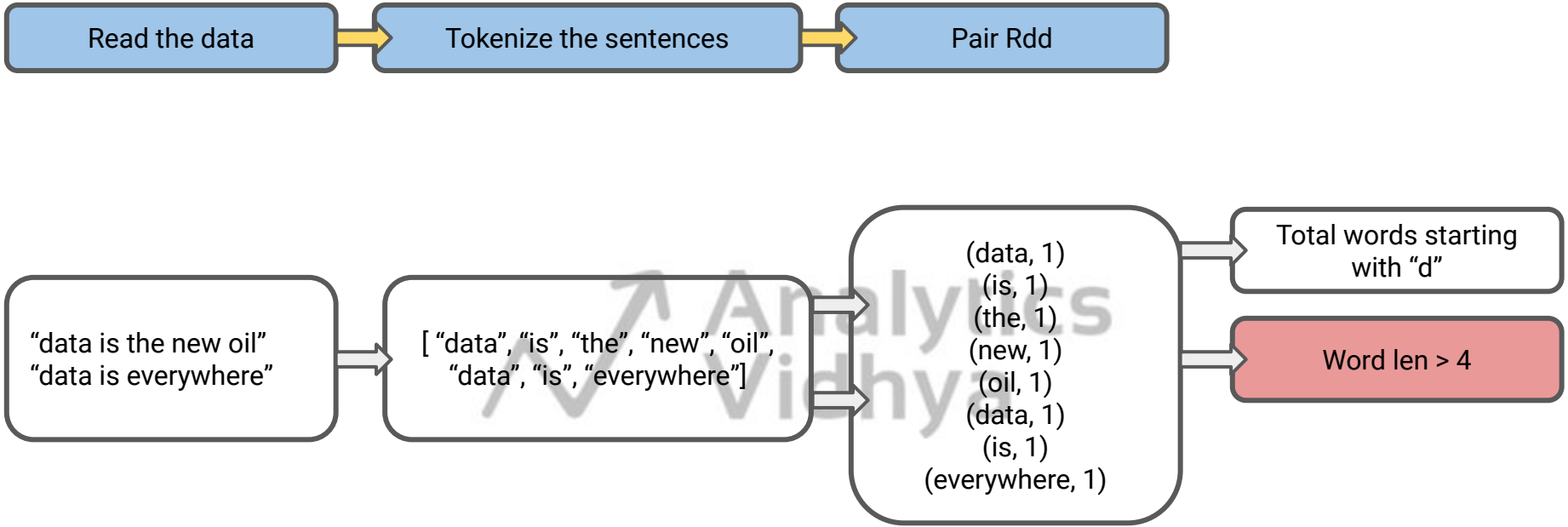
Cache

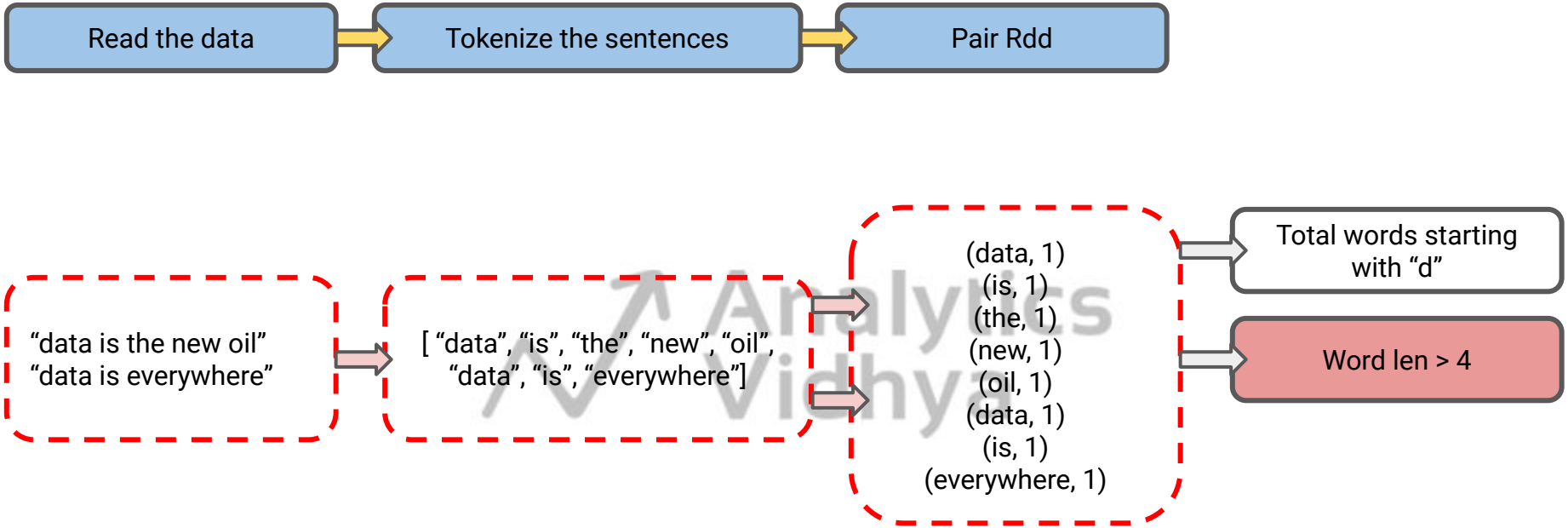
Persist

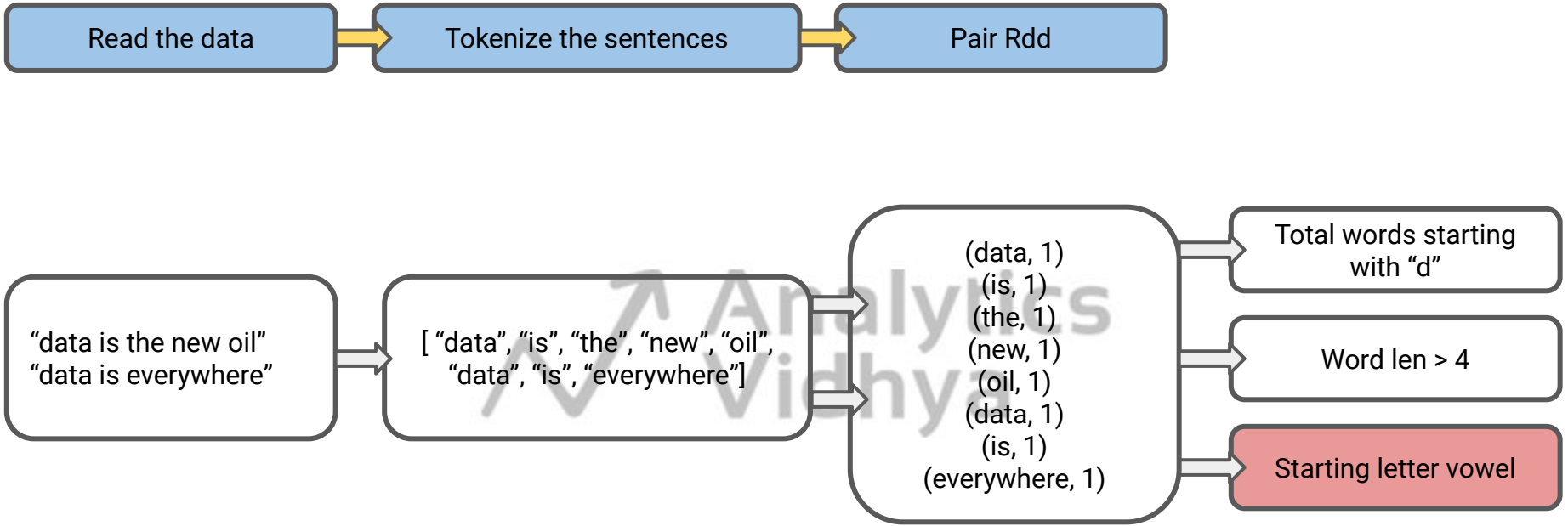


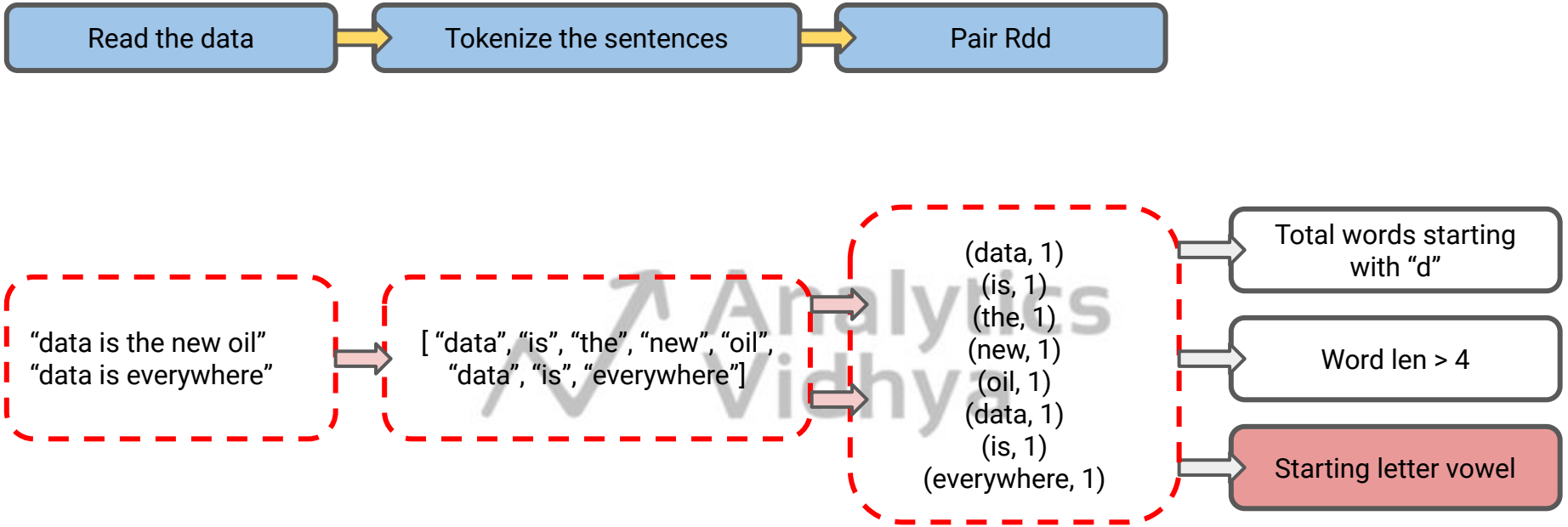

Read the data
Tokenize the sentences
Pair Rdd
“data is the new oil”
“data is everywhere”
[“data”, “is”, “the”, “new”, “oil”,
“data”, “is”, “everywhere”]
(data, 1)
(is, 1)
(the, 1)
(new, 1)
(oil, 1)
(data, 1)
(is, 1)
(everywhere, 1)
Total words starting with “d”

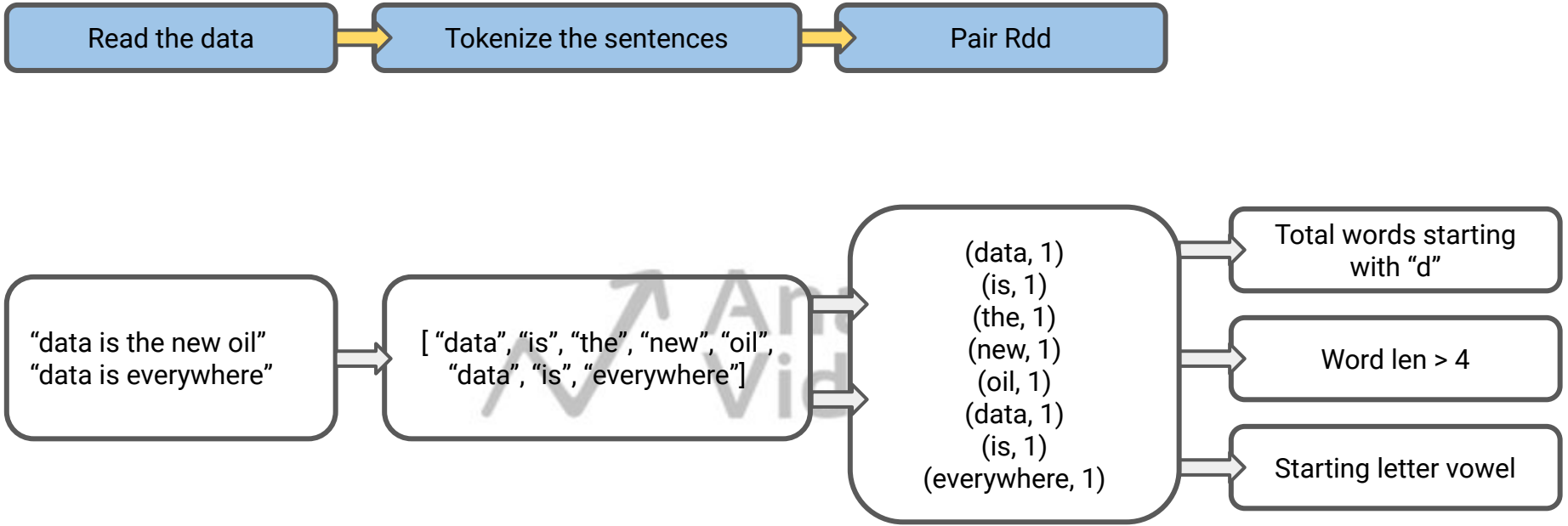



Read the data
Tokenize the sentences
Pair Rdd
“data is the new oil”
“data is everywhere”
[“data”, “is”, “the”, “new”, “oil”,
“data”, “is”, “everywhere”]
(data, 1)
(is, 1)
(the, 1)
(new, 1)
(oil, 1)
(data, 1)
(is, 1)
(everywhere, 1)
Total words starting with “d”
Word len > 4



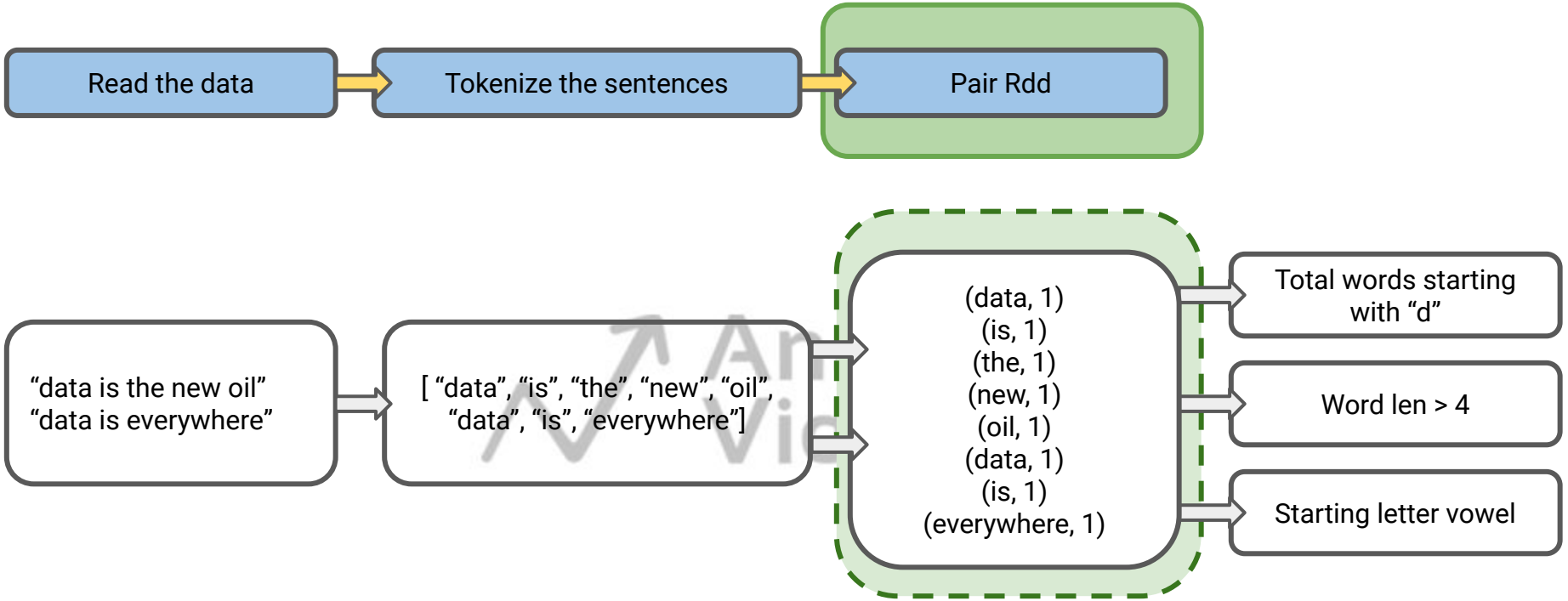

Read the data
Tokenize the sentences
Pair Rdd
“data is the new oil”
“data is everywhere”
[“data”, “is”, “the”, “new”, “oil”,
“data”, “is”, “everywhere”]
(data, 1)
(is, 1)
(the, 1)
(new, 1)
(oil, 1)
(data, 1)
(is, 1)
(everywhere, 1)
Total words starting with “d”
Word len > 4
Starting letter vowel

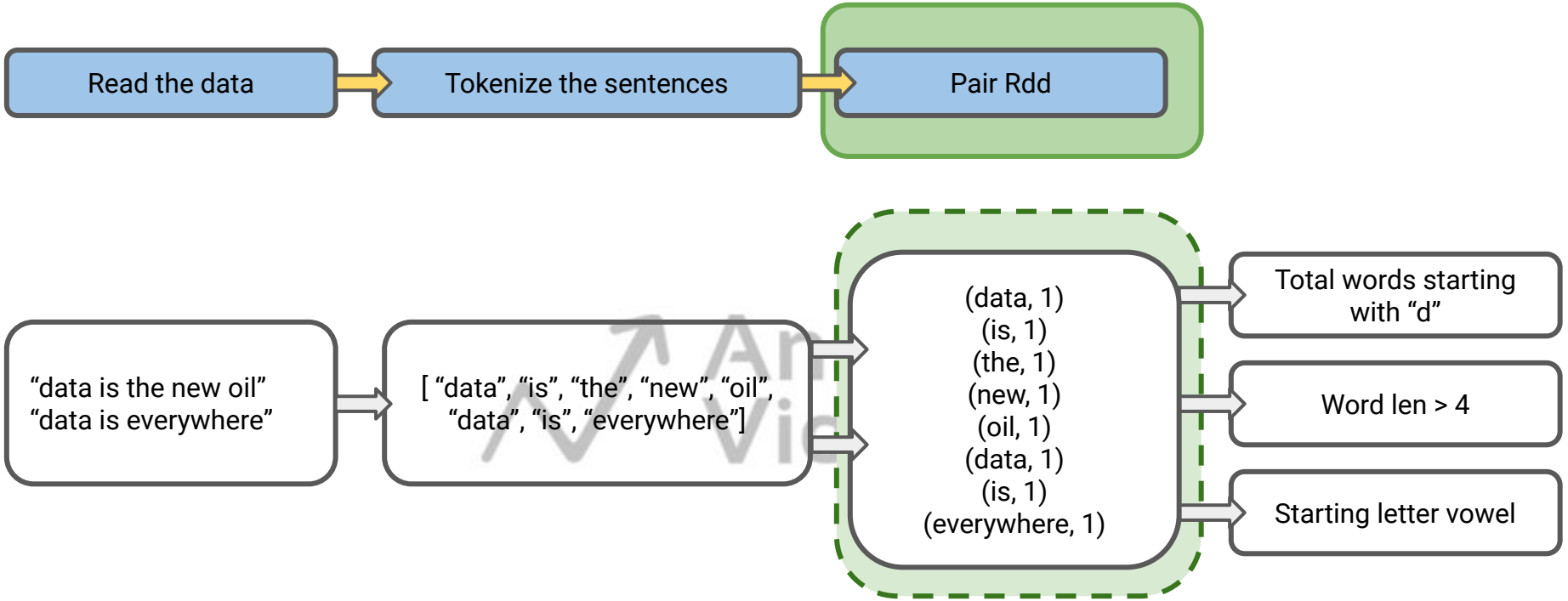

Read the data
Tokenize the sentences
Pair Rdd
“data is the new oil”
“data is everywhere”
[“data”, “is”, “the”, “new”, “oil”,
“data”, “is”, “everywhere”]
(data, 1)
(is, 1)
(the, 1)
(new, 1)
(oil, 1)
(data, 1)
(is, 1)
(everywhere, 1)
Total words starting with “d”
Word len > 4
Starting letter vowel

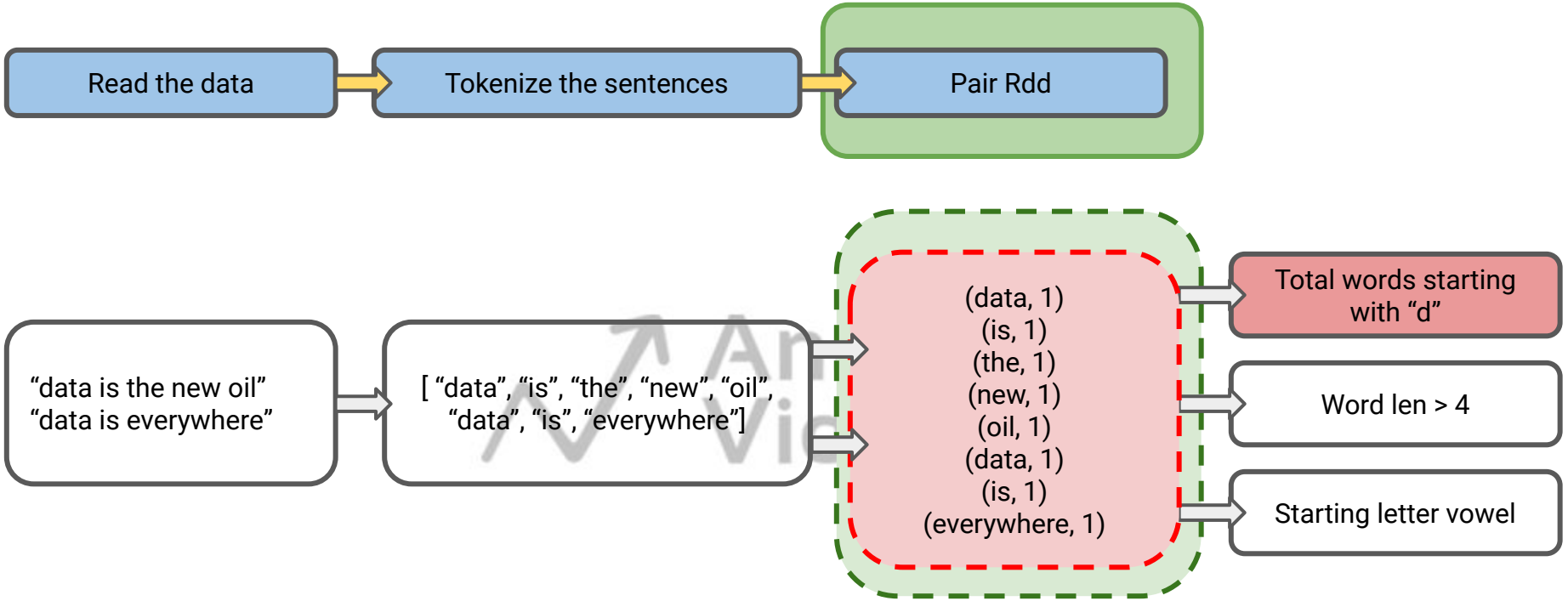


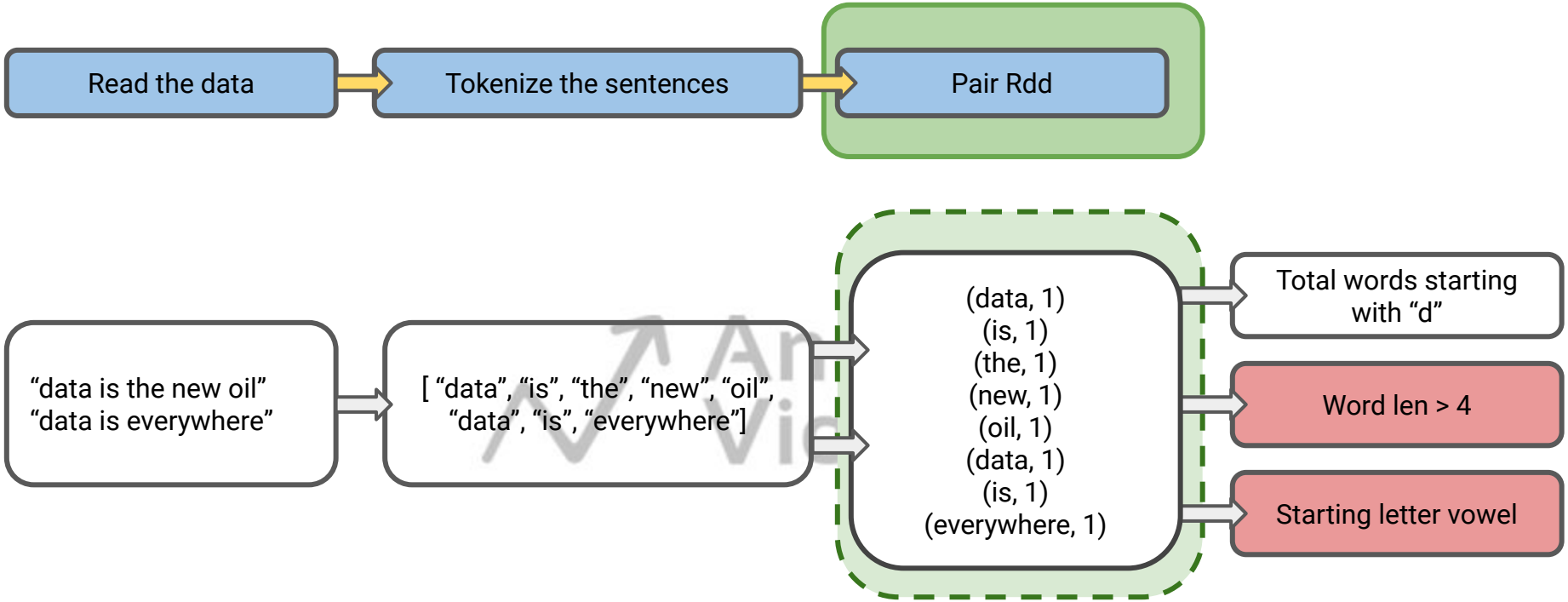
cache/persistence

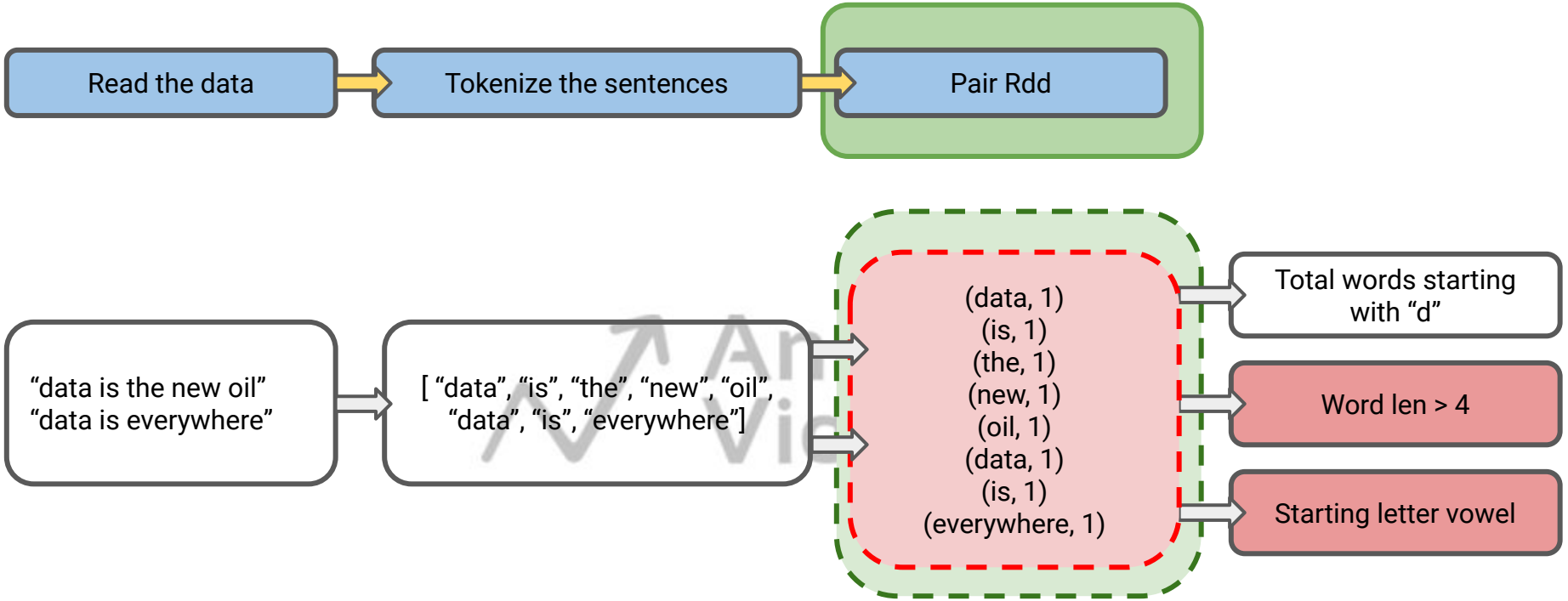
Analytics Vidhya





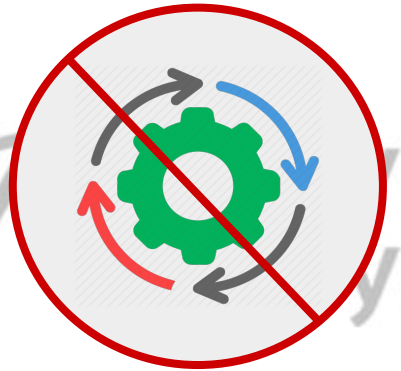






Benefits of Caching/Persistence

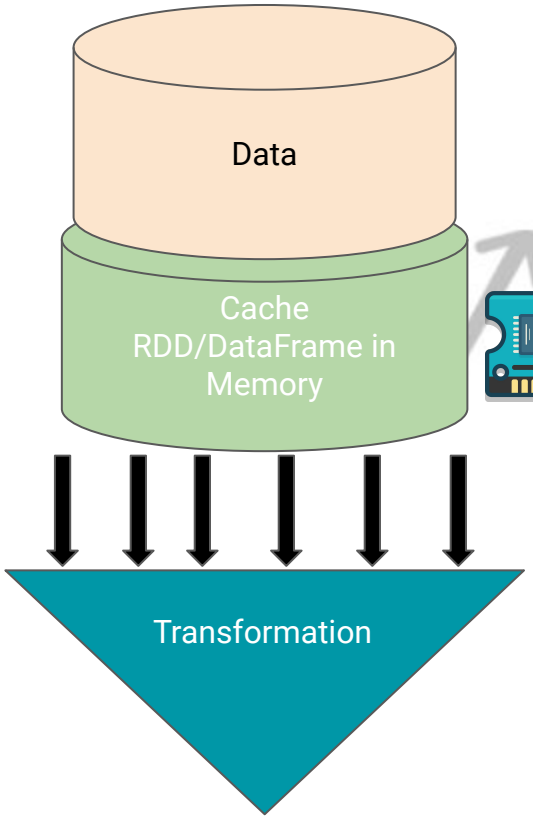
Saves the data by default in memory



Avoids re-computation of the whole lineage

Improves Performance

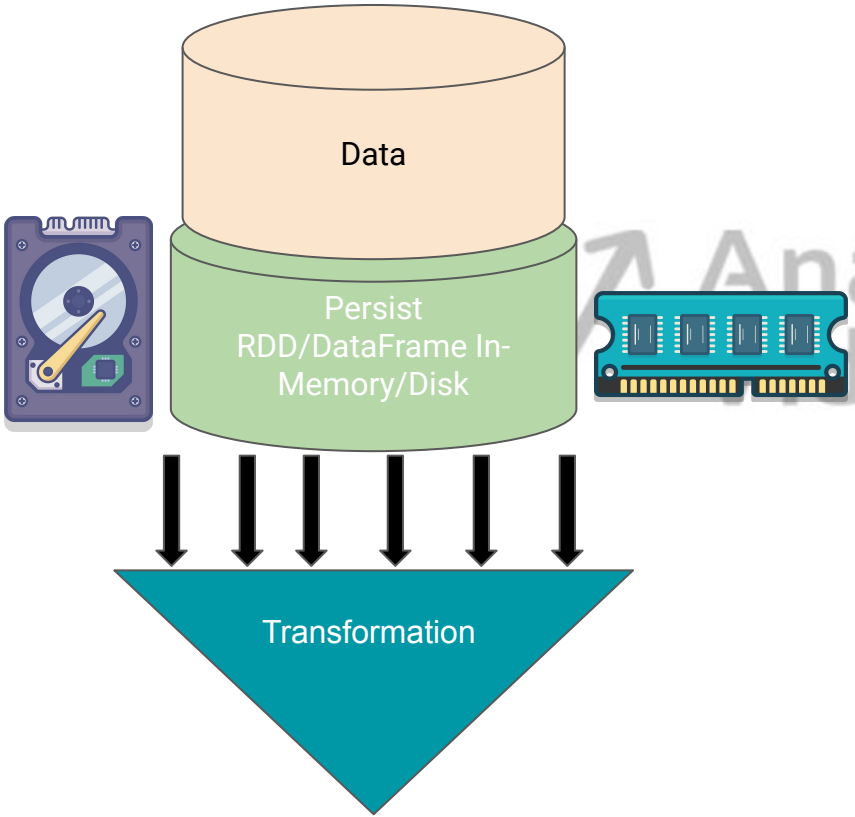
Cache



Syntax:

```
#cache RDD
rdd.cache()
```

Persistence



Syntax:

#persist RDD with default storage-level
rdd.persist()

#persist RDD with **MEMORY_AND_DISK_2**
rdd.persist(pyspark.StorageLevel.MEMORY_AND_DISK_2)

Thank You!!

Analytics
Vidhya