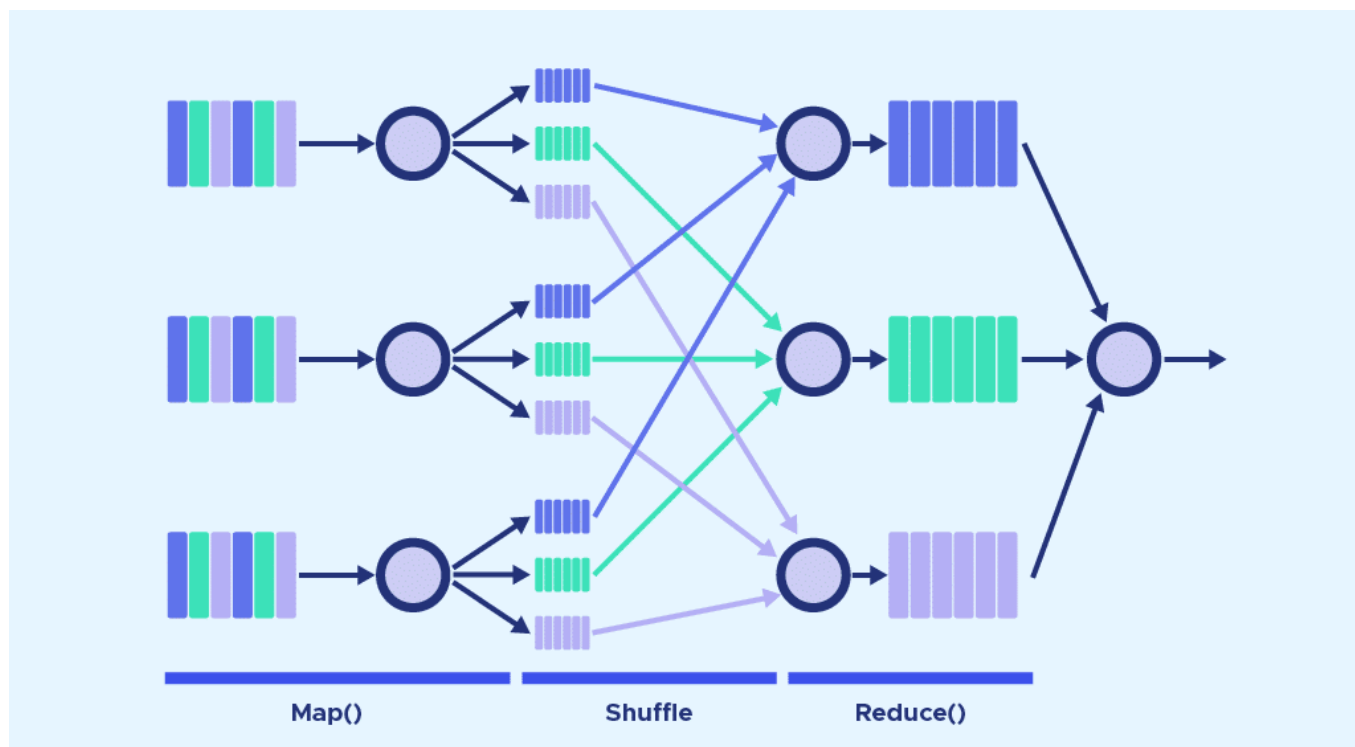


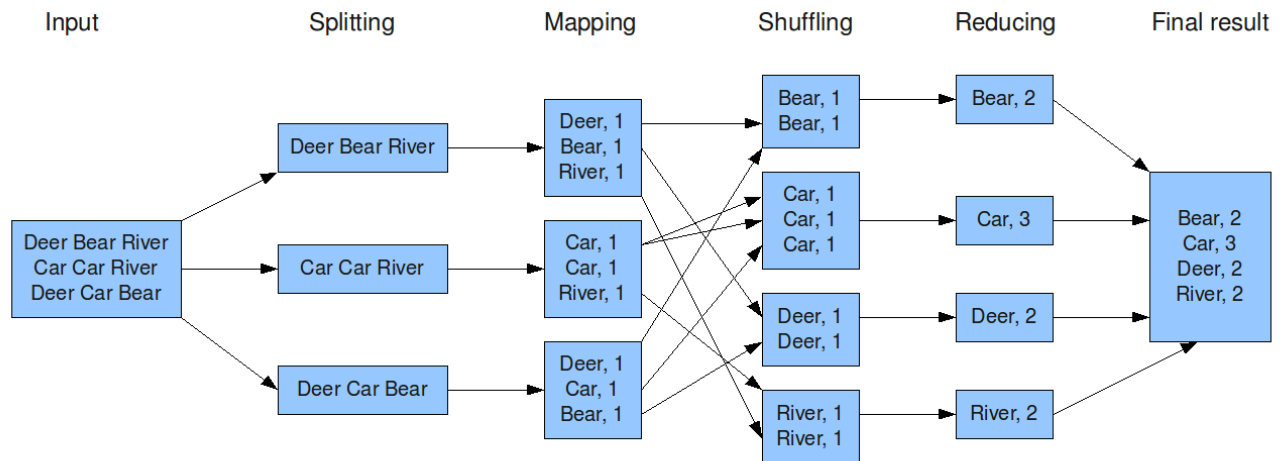
## Rappel : MapReduce, c'est quoi ?

Étape	Action
<b>Map</b>	Transformer chaque élément en paire clé-valeur (k, v)
<b>Shuffle</b>	Répartition/tri des données par clé
<b>Reduce</b>	Agréger les valeurs pour une même clé



Exemple

The overall MapReduce word count process



## ✓ En Spark DataFrame : équivalent avec `groupBy` + `agg`

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode, split, col
import time

# Initialisation de la session Spark
spark = SparkSession.builder.appName("WordCountSparkDF").getOrCreate()
# spark.sparkContext.setLogLevel("ERROR")
print("--- DÉBUT DU JOB SPARK ---")
start_time_spark = time.time()

# Lire le fichier texte (chaque ligne devient une ligne de DataFrame dans la colonne "value")
df = spark.read.text("input.txt")

# Word Count avec l'API DataFrame
word_counts_df = df.select(
    explode(split(col("value"), r"\s+")).alias("word")
).filter(col("word") != "") \
.groupBy("word") \
.count() \
.orderBy(col("count").desc())

# Collecter les 10 premiers résultats
results_spark = word_counts_df.limit(10).collect()

end_time_spark = time.time()
print(f"--- FIN DU JOB SPARK (terminé en {end_time_spark - start_time_spark:.4f} secondes) ---")

# Afficher les 10 premiers résultats
print("\n--- Résultat Spark (10 premiers éléments) ---")
for row in results_spark:
```


```
print(f"{row['word']}\t{row['count']}")
```

```
# Arrêter la session Spark  
spark.stop()
```

## DataFrame vs RDD pour MapReduce

Aspect	RDD	DataFrame / SQL
Verbosité	map / flatMap / reduceByKey	groupBy + agg
Optimisation	Peu d'optimisations	Catalyst optimiser (SQL engine)
Performance	Moins bonne sur gros volume	Meilleure (optimisé, lazy)
Lisibilité	Moins lisible	Très lisible
Fonctionnalité	Plus bas niveau (plus flexible)	API haut niveau (SQL-like)

### En résumé

Question	Réponse
Peut-on faire du MapReduce avec DataFrame ?	 Oui, et c'est même recommandé
Fonctions équivalentes à map/reduce ?	select , explode , groupBy , agg
Performance ?	Meilleure que RDD dans la plupart des cas