3.1 Hadoop

Data Store	Store Format	Identify Way	Value
Hadoop	CSV, Parquet, JSON	File path (e.g., /data/merged_reviews.csv)	Contains key columns such as 'Name', 'Review', 'SkuInfo', 'Date', and 'StarCount'. Parquet and JSON formats are used for translated reviews and train/test datasets.

```python
def write_csv(self, df, path):
    df.write.csv(path, header = True)

def read_csv(self, path, sep=',', header=True, multiline=True):
    """
    Reads a CSV file into a DataFrame.

    :param path: The path to the CSV file.
    :param sep: The separator used in the CSV file (default is '
    :param header: Boolean indicating if the CSV file has a head
    :param multiline: Boolean for multiline support (default is
    :return: A DataFrame containing the CSV data.
    """
    return self.spark.read.csv(
        path,
        sep=sep,
        header=header,
        multiLine=multiline,
        inferSchema=True
    )

def write_parquet(self, df, path):
    """
    Writes a DataFrame to a Parquet file.

    :param df: The DataFrame to write.
    :param path: The path where the Parquet file will be saved.
    """
    df.write.parquet(path, mode='overwrite')

def read_parquet(self, path):
    """
    Reads a Parquet file into a DataFrame.

    :param path: The path to the Parquet file.
    :return: A DataFrame containing the Parquet data.
    """
    return self.spark.read.parquet(path)

def read_json(self, path):
    """
    Reads a JSON file into a DataFrame.

    :param path: The path to the JSON file.
    :return: A DataFrame containing the JSON data.
    """
    return self.spark.read.json(path)

def write_json(self, df, path):
    """
    Writes a DataFrame to a JSON file.

    :param df: The DataFrame to write.
    :param path: The path where the JSON file will be saved.
    """
    df.write.json(path)
```

3.2 Redis

Data Store	Store Format	Identify Way	Value
Redis	Key-Value Pair (String)	Unique key (e.g., review:<id>)	Contains key-value pairs where each key is a unique review ID, and the value is a JSON

			object with fields like name, review, SKU info, date, and star count.

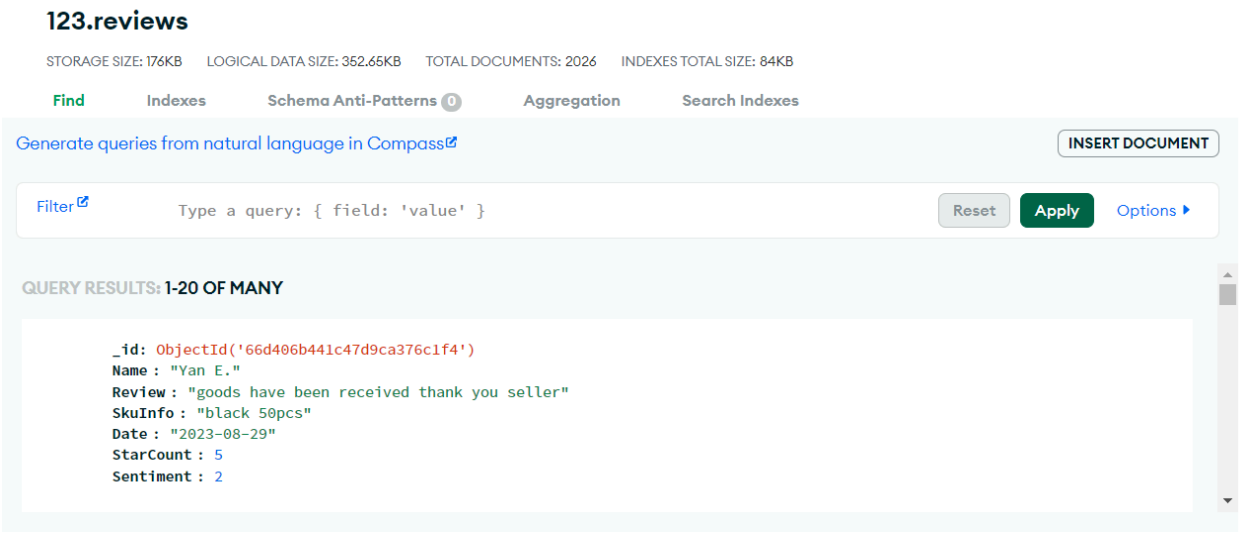
```
def store_dataframe(self, df, key_prefix="row"):
    """
    Stores a PySpark DataFrame in Redis as JSON strings.

    Parameters:
    - df: PySpark DataFrame, the DataFrame to store.
    - key_prefix: str, prefix for the Redis keys.
    """
    # Convert the Date column to a string format in the DataFrame
    df = df.withColumn("Date", date_format(col("Date"), "yyyy-MM-dd"))

    # Convert cleaned DataFrame to a list of dictionaries
    df_list = df.collect()
    data_to_store = [row.asDict() for row in df_list]

    # Store each row in Redis as a JSON string
    for i, data in enumerate(data_to_store):
        self.redis_client.set(f"{key_prefix}:{i}", json.dumps(data))

    print("Data stored in Redis successfully.")
    return len(data_to_store)


def load_data(self, num_rows, key_prefix="row"):
    """
    Loads data from Redis and converts it to a list of dictionaries.

    Parameters:
    - num_rows: int, the number of rows to load.
    - key_prefix: str, prefix for the Redis keys.

    Returns:
    - list of dictionaries representing the loaded data.
    """
    data = []
    for i in range(num_rows):
        json_data = self.redis_client.get(f"{key_prefix}:{i}")
        if json_data:
            data.append(json.loads(json_data))

    # Convert the date strings back to date objects (if needed)
    for item in data:
        if 'Date' in item and item['Date']:
            item['Date'] = datetime.strptime(item['Date'], "%Y-%m-%d").date()

    return data
```

```python
from redis_handler import RedisHandler

# Initialize Spark session
spark = SparkSession.builder.appName("Redis to PySpark DataFrame").getOrCreate()
redis_handler = RedisHandler(host='localhost', port=6379, db=0)

# Store the DataFrame in Redis
num_rows = redis_handler.store_dataframe(df_cleaned)

# Load the data back from Redis
loaded_data = redis_handler.load_data(num_rows)

# Convert the loaded data to a DataFrame
df_loaded = redis_handler.convert_to_dataframe(loaded_data, spark)
df_loaded.show(truncate=False)
```

```
24/09/07 18:13:36 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
Data stored in Redis successfully.
[Stage 49:>                                                         (0 + 1) / 1]
+----------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----------------------------+----------+---------+
|Name      |Review
|SkuInfo                      |Date      |StarCount|
+----------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----------------------------+----------+---------+
|Yan E.    |The goods have been received, thank you seller 😊😊
|Color Family:Black-50PCS     |2023-08-29|5        |
|Thevagi S.|all good
|Color Family:Grey            |2022-05-30|5        |
|s***.     |The quality is not very good, too thin, not genuine.  Considering the price, this is the quality!\n质量不是很好，太单薄了，不是正版的。 考虑到这个价格，就是这般质量吧!
|color_family:Black           |2022-10-30|3        |
|BC K.     |products packaging not see through, can't see the mask colour inside.
|Color Family:Black           |2022-04-03|4        |
|Jennie 8. |Good quality for the product,,,excellent services\nFrom the seller
|Color Family:Black           |2022-08-01|5        |
|S***.     |轻触并按住剪贴内容即可将其固定。未固定的剪贴内容将于 1 小时之后被删除。欢迎使用 Gboard 剪贴板，您复制的所有文本都会保存到这里。
|color_family:Black           |2024-01-29|5        |
|Eddie T.  |Good seller fast deliveryGood seller fast deliveryGood seller fast delivery
|color_family:Careion 3D Black|2023-09-18|5        |
|Thi N.    |I think it is good for me. i will use it and review it later for you guys
|color_family:White           |2023-08-31|5        |
|Narainis  |Suka suka suka. Maaf gambar tidak berkaitan tapi produk semua terbaik
|color_family:Hitam           |2023-04-05|5        |
|Khor S.   |thanks received
|color_family:Black           |2022-12-21|5        |
|Tham      |very thin
|Color Family:Black-50PCS     |2022-11-27|5        |
|*******898|It's unbelievable that seller sent  me short of 1 item out of only 20 I ordered.. I can only assume that it's done intentionally. The mask is only 3ply and not 4ply as adv
ertised and felt cheated by seller. I forgo my claim as the process is troublesome and its cost is minimal. The courier service really suckered and Lazada should ensure that buyers de
serve better delivery service.|Color Family:Purple          |2022-05-04|3        |
|Yap B.    |Fast shipping, will repurchase again.  Thank you
|color_family:Black           |2023-06-05|5        |
|W***.     |Selamat sampai thx seller\ncuma nipis sikit\nsesuai dg harga
|Color Family:Hitam           |2022-05-28|5        |
|N***.     |mask cukup seperti yang di pesan...harga murah boleh beli lagi
|Color:Grey                   |2022-04-01|5        |
|K***.     |pelik......kite order mask lain, tgok2 mask lain yg sampai....SGT2 kecewa la...walaupun murah...mask yg jenis lain yg dibagi tu, 50pcs 1.50 tp nipis giler...sy nk mask lai
n awak eh
|Color:Headloop Purple        |2022-04-10|5        |
|Theresa T.|The item just received today which packed neat.  The mask same as per advertised and comfortable to wear. Thank you Lazada for prompt delivery.
|Color Family:White-50PCS     |2023-11-26|5        |
|1***1     |I ordered on sept(50pcs) to check the size and design.. all were good. so I re ordered 100pcs. end up different design at nose bridge.dissappointed.
|Color Family:Black-50PCS     |2022-11-14|5        |
|Lun K.    |First time buy it, feel good
|Color:White                  |2022-04-25|5        |
|L***.     |This mask is big enough for man's face.\nIt is comfortable, good price too.\nThank you seller and courier guy.
|Color Family:Grey-50PCS      |2022-09-12|5        |
+----------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----------------------------+----------+---------+
only showing top 20 rows
```

3.3 MongoDB

Data Store	Store Format	Identify Way	Value
MongoDB	Document structure like JSON	_id as the unique key for each document	Contains several key-value pairs like the date of the review is the value of the “Date” key



```
[55]: # List some documents
mongo_handler.list_documents(limit=3)
# Retrieve data from MongoDB
mongo_data = mongo_handler.retrieve_data()
```

```
Listing 3 documents in the reviews collection:
[{'Date': '2023-08-29',
  'Name': 'Yan E.',
  'Review': 'goods have been received thank you seller',
  'Sentiment': 2,
  'SkuInfo': 'black 50pcs',
  'StarCount': 5,
  '_id': ObjectId('66d406b441c47d9ca376c1f4')},
 {'Date': '2022-05-30',
  'Name': 'Thevagi S.',
  'Review': 'all good',
  'Sentiment': 2,
  'SkuInfo': 'grey',
  'StarCount': 5,
  '_id': ObjectId('66d406b441c47d9ca376c1f5')},
 {'Date': '2022-04-03',
  'Name': 'BC K.',
  'Review': 'products packaging not see through can t see mask colour inside',
  'Sentiment': 2,
  'SkuInfo': 'black',
  'StarCount': 4,
  '_id': ObjectId('66d406b441c47d9ca376c1f6')}]
```

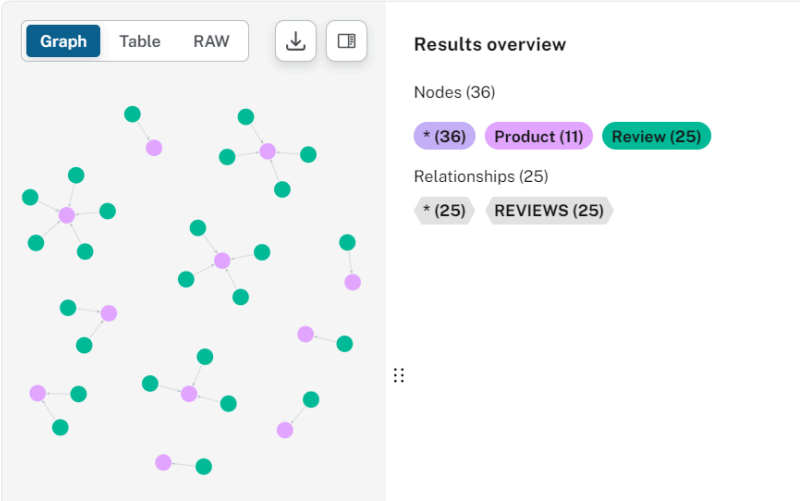
3.4 Neo4j

Data Store	Store Format	Identify Way	Value
Neo4j	Nodes and Relationships	Stores data as nodes and relationships. For instance, a :Product node and a :Review node are created, and a HAS_REVIEW relationship links the product to its associated review.	A :Product node might have properties such as skuInfo, while a :Review node includes properties like name, review, starcount, and date. The HAS_REVIEW relationship between them

			contains no additional properties.

```python
def create_product_nodes_and_relationships(self, data, batch_size=1000):
    cypher_query = """
    UNWIND $data as row
    MERGE (p:Product {skuInfo: row.SkuInfo})
    FOREACH(ignoreMe IN CASE WHEN row.Review IS NOT NULL THEN [1] ELSE [] END |
        CREATE (r:Review {
        name: row.Name,
        review: row.Review,
        starcount: row.StarCount,
        date: row.Date
        })
        MERGE (r)-[:REVIEWS]->(p)
    )
    """
    count_queries = {
        "Product": "MATCH (p:Product) RETURN COUNT(p) AS count",
        "Review": "MATCH (r:Review) RETURN COUNT(r) AS count"
    }
    try:
        with self.driver.session() as session:
            # Process data in batches
            for i in range(0, len(data), batch_size):
                batch = data[i:i + batch_size]
                session.run(cypher_query, data=batch)
                print(f"Batch {i//batch_size + 1} processed successfully.")

            # Count and print the nodes
            for label, query in count_queries.items():
                result = session.run(query)
                count = result.single()["count"]
                print(f"Total {label} nodes: {count}")
    except Exception as e:
        print(f"An error occurred: {e}")
```

```
neo4j $ MATCH p=()-[:REVIEWS]->() RETURN p LIMIT 25;
```



Neo4J Retrieving & Load

```
# Load reviews to Spark DataFrame
try:
    df = neo4j_handler.load_reviews_to_dataframe(spark)
    df.show()
except Exception as e:
    print(f"Error loading reviews: {e}")

# Close the connection
neo4j_handler.close()
```

```
+----------+--------------------+-----------+----------+---------+
|      Name|              Review|    SkuInfo|      Date|StarCount|
+----------+--------------------+-----------+----------+---------+
|     Nasih|             awesome|black 50pcs|2022-07-27|        5|
|Abinash M.|                  ok|black 50pcs|2022-09-28|        5|
|    Loh W.|   great design love|black 50pcs|2024-06-10|        5|
|     Md A.|        good product|black 50pcs|2023-07-15|        5|
| Garlic M.|doesnt match vide...|black 50pcs|2022-07-25|        1|
|  NorHa S.|black mask there ...|black 50pcs|2022-08-16|        5|
|     n***i|received good con...|black 50pcs|2022-06-28|        5|
|THERESA H.|fast delivery but...|black 50pcs|2022-10-03|        5|
|*******896|short three packs...|black 50pcs|2023-05-12|        4|
|  Chris C.|got packing no so...|black 50pcs|2023-09-26|        5|
|    Yan E.|goods have been r...|black 50pcs|2023-08-29|        5|
|    Greedy|           very thin|black 50pcs|2022-11-27|        5|
|     1***1|i ordered sept ch...|black 50pcs|2022-11-14|        5|
|     T***.|good service fast...|black 50pcs|2024-03-03|        5|
|    AgnesQ|inner layer mask ...|black 50pcs|2023-02-27|        5|
|    tan F.|no good quality l...|black 50pcs|2022-07-29|        1|
|   Arse A.|congratulations u...|black 50pcs|2022-07-06|        5|
|     L***.|nice mask fast de...|black 50pcs|2022-07-23|        5|
|     GO A.|not ply made chin...|black 50pcs|2023-02-25|        1|
|   Tang L.|everything good f...|black 50pcs|2023-06-09|        5|
+----------+--------------------+-----------+----------+---------+
only showing top 20 rows
```

3.5 Hbase

Data Store	Store Format	Identify Way	Value
Hbase	Key-Value Structure with Column Families	Each row in an HBase table is uniquely identified by a Row Key. In this project we are using uuid combine with key_hash which are the combination of 'Sentiment' and 'Review' as a row key	Consists of a Row Key and one or more Column Families with Columns.

```
def generate_row_key(self, row):
    """
    Generates a unique row key for storing in HBase.

    Parameters:
    - row: Row object containing the data.

    Returns:
    - str, the generated row key.
    """
    unique_id = uuid.uuid4()
    key_hash = hash((row['Sentiment'], row['Review']))
    return f"{unique_id}_{key_hash}"
```

[4]:
```
test_df  = hbase_handler.retrieve_from_hbase('test_data')

# Show the retrieved DataFrame
test_df.show(5)
test_df.count()
```

```
+--------------------+---------+-------------+--------------------+----------------+
|              Review|Sentiment|SkuInfo_index|              tokens|number_of_tokens|
+--------------------+---------+-------------+--------------------+----------------+
|okay okay okay qu...|        2|          4.0|[okay, okay, okay...|               9|
|delivery fast sat...|        2|          0.0|[delivery, fast, ...|              23|
|                  ok|        2|          7.0|                [ok]|               1|
|good seller fast ...|        2|          1.0|[good, seller, fa...|              10|
|                good|        2|          1.0|              [good]|               1|
+--------------------+---------+-------------+--------------------+----------------+
only showing top 5 rows
```