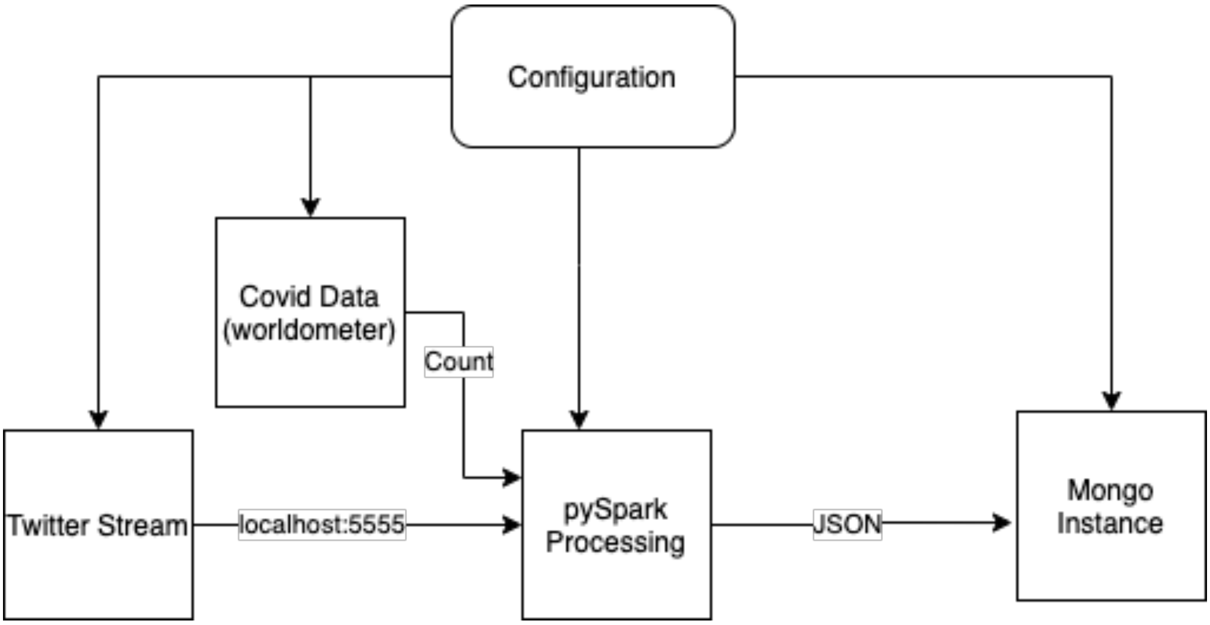
TWITTER STREAM DATA COLLECTION

Objective:

Collect & Process the twitter data along with the daily coronavirus data and Store it to the Mongo DB Instance

Architecture:

Config Driven Twitter Data Stream Collection

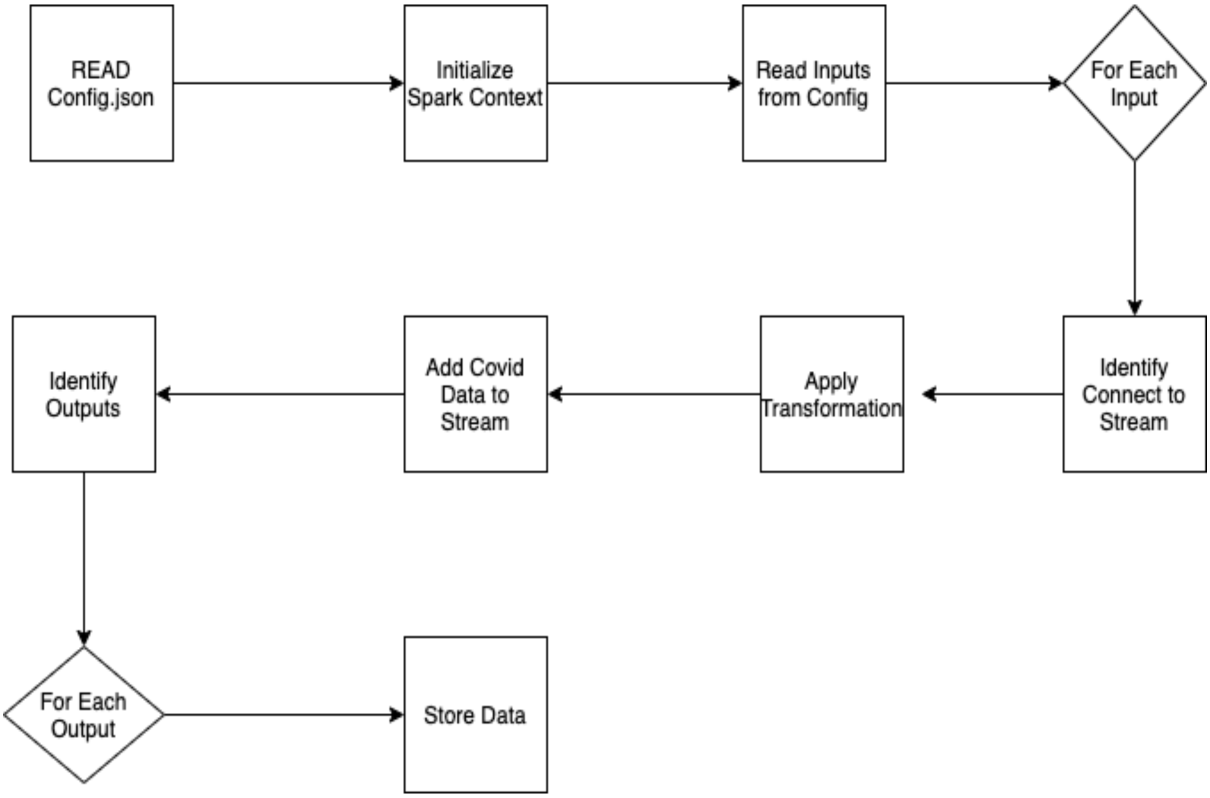


Setup Guidelines:

1. Setup Virutal Env : python3 -m venv /Users/darshan/pythonEnv
2. Activate it : source /Users/darshan/pythonEnv /bin/activate
3. Install dependencies : pip install -r requirements.txt
4. Setup Mongo : *docker run -p 27027:27017 mongo:4.0.0*
5. start the twitter stream python twitter_stream_simulator.py
6. start the tweet processing python twitter_data_processing.py

Implementation

1. Read the config.json file
2. Initialize the spark context by reading data from config
3. Take the input stream metadata from config
4. For Each Input stream apply transformation
5. Add covid data to each stream
6. Store to the output.



twitter_stream_simulator.py

Produces the twitter stream

twitter_data_processing.py

Config Driven Data Processing Engine

data_stream_provider.py
Provides the pyspark stream object by connecting to the stream

data_store_provider.py
Provides the Database Instance

Advantages of Implementation:

1. New Data Stream can be easily Added
2. New Outputs can be easily configured
3. Transformation/filtering criteria can be modified easily

Sample Config.json

```
{
  "app_name": "TwitterStreamProcessing",
  "batch_duration": 20,
  "inputs": [
    {
      "name": "twitter_data",
      "ip": "localhost",
      "port": 5555,
      "type": "stream",
      "sub_type": "socket",
      "filtering": true,
      "filter_strings": [
        "#",
        "RT:",
        "http://\\S+|https://\\S+"
      ],
      "output_field": "content"
    }
  ],
  "covid_input": {
    "url": "https://www.worldometers.info/coronavirus/",
    "output_field": "total_case_count"
  },
  "outputs": [
    {
      "name": "mongoDB",
      "ip": "127.0.0.1",
      "port": 27027,
      "type": "database",
      "sub_type": "mongoDB",
      "schema_name": "test",
      "table_name": "twitter_data"
```

```
  }
]
```

Mongo DB Sample Output

```
{ "_id" : ObjectId("60978e01511d30a7c8996f08"), "content" : [ "A collection of textile
samples lay spread out on the table - Samsa was a travelling salesman - and above it there
hung a picture that he had recently cut out of an illustrated magazine and housed in a nice,
gilded frame.", " His many legs, pitifully thin compared with the size of the rest of him, waved
about helplessly as he looked.", "The bedding was hardly able to cover it and seemed ready
to slide off any moment.", "It wasn't a dream.", "\"What's happened to me?\" he thought. ", "
One morning, when  Gregor Samsa woke from troubled dreams, he found himself
transformed in his bed into a horrible vermin.He lay on his armour-like back, and if he lifted
his head a little he could see his brown belly, slightly domed and divided by arches into stiff
sections.", "It showed a lady fitted out with a fur hat and fur boa who sat upright, raising a
heavy fur muff that covered the whole of her lower arm towards the viewer.", "The bedding
was hardly able to cover it and seemed ready to slide off any moment.", " One morning,
when  Gregor Samsa woke from troubled dreams, he found himself transformed in his bed
into a horrible vermin.He lay on his armour-like back, and if he lifted his head a little he could
see his brown belly, slightly domed and divided by arches into stiff sections.", "It wasn't a
dream.", "His room, a proper human room although a little too small, lay peacefully between
its four familiar walls." ], "timestamp" : ISODate("2021-05-09T12:53:40.214Z"),
"total_case_count" : 158337486 }
```