## **2** **Architecture of the Proposed Solution**

**2.1** **Data Analytics Architecture**

**Stream Layer:**

* Real-time input of twitter streams using Twitter’s API (Tweepy)
* Use Spark Streaming, part of Apache’s language-integrated API, to provide scalable and fault-tolerant cleaning and processing of real-time tweet streams.
* Apply ML Algorithms using Spark’s MLib’s RDD-based and/or Dataframe-based APIs

**Batch Layer:**

* Store batches of raw data in Spark’s columnar storage layout.
* Store and process structured data in a relational database model using Spark SQL for efficient querying
* Clean and process stored data using MapReduce algorithms in Spark
* Create a predictive machine learning model with logistic regression and classification using Spark’s MLib

**Data Visualization:**

* Visualize data using D3, MatPlotLib, and other Python modules

**2.2** **Data Collection/Ingestion, Storage, Processing, Serving, and Visualization**

**Data Collection/Ingestion:**

* **Streaming Data:** Receive tweet streams using Twitter’s API (Tweepy) and Twitter Intelligence Tool (TWINT)
* **Batch Data:** Receive massive amounts of batch data (e.g. old tweets, economic data...etc.) from Kaggle and web-scraped tweets.

**Data Storage:**

* Store cleaned and processed data using Spark's columnar storage layout
* Store structured tweets in a relational model to be queried fast

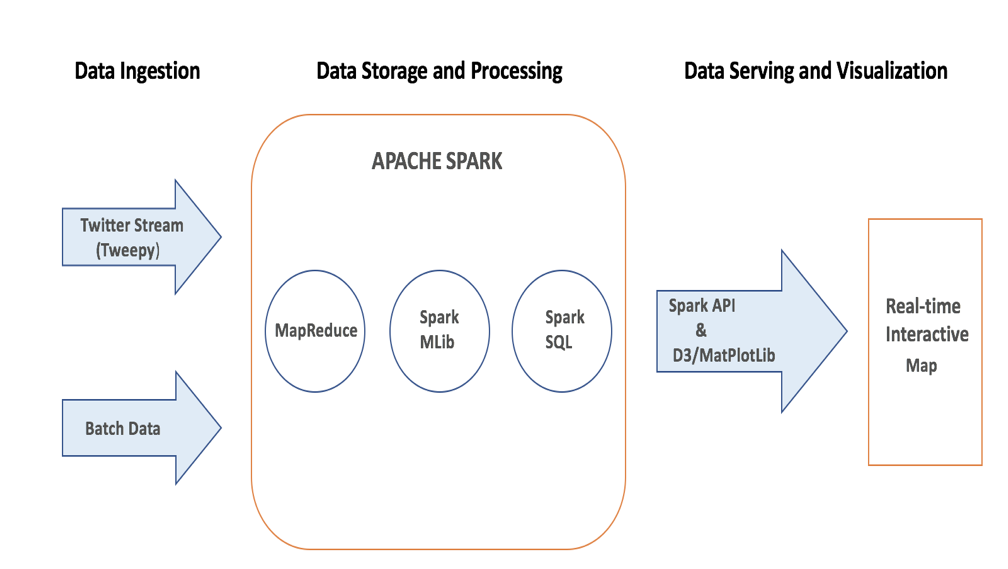
**Data Processing:**

* Clean and process data using MapReduce algorithms in Spark
* Run logistic regression on data to create a predictive ML model using Spark's MLib
* Use Pandas and other Python modules to process data efficiently

**Data Serving and Visualization:**

* Serve and visualize data in the browser as an interactive geographic map by integrating D3 and MatPlotLib modules with related Spark API

**2.3** **Architecture and Data Flow of the System**

****

**2.4** **Anticipated Difficulties in the Initial Approach**

* Streaming tweets with geo-locations can impose a limitation on the accuracy of the data because only a small number of users enable location permissions on Twitter.
* Since Canada is a bilingual country, streaming tweets in French may require additional translation/cleaning algorithms.
* Mapping Twitter's geo-locations with Canada's electoral districts (ridings) maybe a challenging task. However, given the latitude and longitude of all electoral districts, we will be able to map the ridings to the tweet location quite accurately.