Twitter Analytics

--team Voyager

Project introduction

Social media like facebook and twitter have become an important way to distribute news. Also people nowadays tend to post their feeling in social networks. Consequently, data collected from social networks implies the most up-to-date topics and public feelings about them. In this project, we try to approach popular topics or a specific topic analytics with real-time twits.

Motivation

Twitter provides public API for developers to obtain data. With its stream API, we can access real-time data from twitter to achieve live data analysis. With this idea, we find a proper framework used to complete streaming analytic—Spark.

Around 6000 twits are sent every second. To process these streaming data, the framework we choose should be able to deal with very large data set. Spark is scalable to large clusters. So it can deal with huge dataset both in computing and storage. Second-scale latencies are needed in live analytics. In this aspect, spark is famous for its process speed. Instead of disk-accessing, the computation spark adapted is memory-data accessing. Simply by reducing IO operation, spark is almost 100 times faster than Hadoop. As a scalable programming framework which is able to deal with huge dataset, spark has excellent strategy to ensure fault tolerance in an efficient way. When it comes to data storage, unlike Hadoop mapreduce that replicate data to achieve fault-tolerance, spark records data generation to recreate error data. If no data fault happens, the cost is zero. Moreover, with Spark, not only streaming data analytics can be achieved, batch data analysis is also integrated.

About Spark

Figure1 shows the eco-system of Spark. At the bottom is Spark Core which is the general execution engine for the whole platform. It provides Java, Scala, and Python APIs for ease of development.

Above this core, spark also provides four libraries to support a variety of applications. Spark DataFrames makes SQL-like query available in distributed file system. Spark Streaming library enables powerful interactive and analytical applications across both streaming and historical data, while inheriting Spark’s ease of use and fault tolerance characteristics. It readily integrates with a wide variety of popular data sources, including Twitter. MLlib and GraphX provide machine leaning and graph computation library respectively, which facilitate abundant well-performed popular Algorithms. With various APIs and readily to use libraries, fancy twitter analytics development is potentially to achieve.

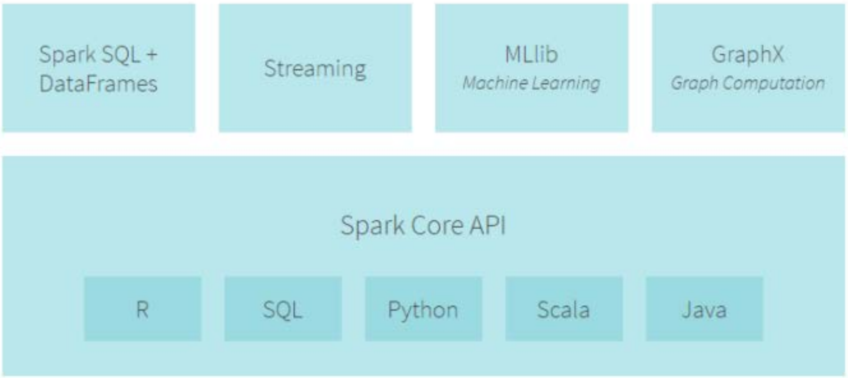

Figure1. Spark eco-system

Approach

As mentioned above, the data we use in this project is real-time stream data from twitter. In fetched twits, we can get more than text postings. Twitter APIs provide almost every feature appeared on its website, for instance, hash tagged topics and geo locations. Taking advantage of these features, we can explore how much a subject been discussed, people's feeling about a certain subject and more interestingly, how these subjects were distributed. Figure 2 shows the general workflow.

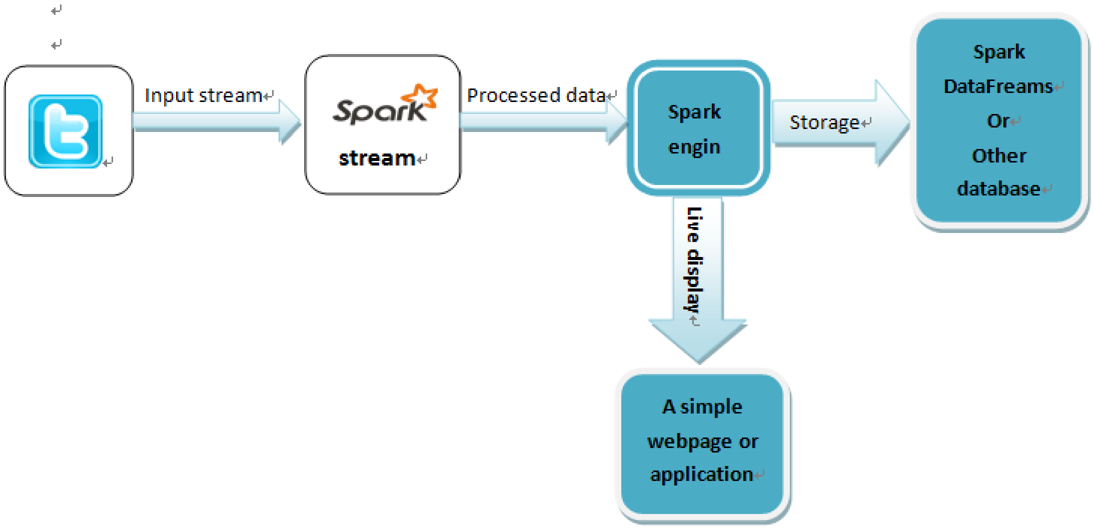

Figure2. General workflow

Step1.

Get familiar with twitter API and spark stream—real-time word count.

In this stage, the biggest challenge are to set up spark platform and to know how to use twitter API. With these resources, we can easily implement a real-time word count application.

Step2.

As above application been implemented, we should find a method to show the outcome of word-counting. We can store these result in a database and display them later. But it is by no means the best way. Showing the real-time data analysis with a live illustration is the ideal. So this could be the challenge of this step.

Step3.

Sentiment analysis can be done under the support of spark machine learning. And also, we find a training dataset from Kaggle that can be used to contribute our own sentiment analysis model.

However, since none of us has the experience with spark MLlib, this approach could be a serious issue.

Fortunately, we have a backup approach. Under the direction of Dr. Sharma, we find IBM Bluemix works really well with sentiment analytics. Since it is readily get access with twitter stream API, Bluemix can be used in a very convenient way—we just need to drag the components we need and connect them. We also noticed that it can connect with Spark directly, but for this part, we need to explore further.

Step4.

Geo location of twits is provided by API. If we can take advantage of this information, we can analyze topics' region feature. The problem here is that geo location provided by twitter API is just coordinates. It could be difficult to implement it with a specific city or even state. But even though location analysis cannot be achieved in an accurate way, we can still do some rough analytics.