Vicente Rubén Del Pino Ruiz, May 14, 2014

**TOPIC: Big Data Case Study in Social Media**

**Concentration: Twitter, AWS and Hadoop**

**Project Description:**

I will use AWS and Hadoop for realizing near real time analytics over data downloaded from Twitter.

|  |
| --- |
| **Problem:**  Download movie related data from Twitter, upload them to a Hadoop cluster and realize analytics over the data in near real time.  **My Goal:**  Show how integrate and analyze a high volume of data from social media using Hadoop. |

**Overview of steps:**

1. Setup a Hadoop cluster in AWS.
2. Create and setup the download process for Twitter.
3. Upload data to HDFS.
4. Train sentiment model with training set (out of scope for this project)..
5. Calculate sentiment for new tweets (out of scope for this project).
6. Transform Twitter data to Parquet format.
7. Analyze data using Impala.

**Data Set:**

Twitter data download in real time.

**Hardware:**

AWS Cluster using three m1.large nodes.

**Software:**

The software used will be the following:

1. Custom software using Python and Java.
2. Hadoop HDFS.
3. Hadoop Parquet format.
4. Hadoop Mahout.
5. Hadoop PIG
6. Hadoop Impala

**Benefits:**

Being able to analyze social media content in near real time helps to have a better understanding of the public engagement with the term searched.

**Cons:**

Social Media generates a big volume of data. Manipulate and analyze this data is a difficult task to be achieved in a local machine; this is why for this project I will use a cluster.