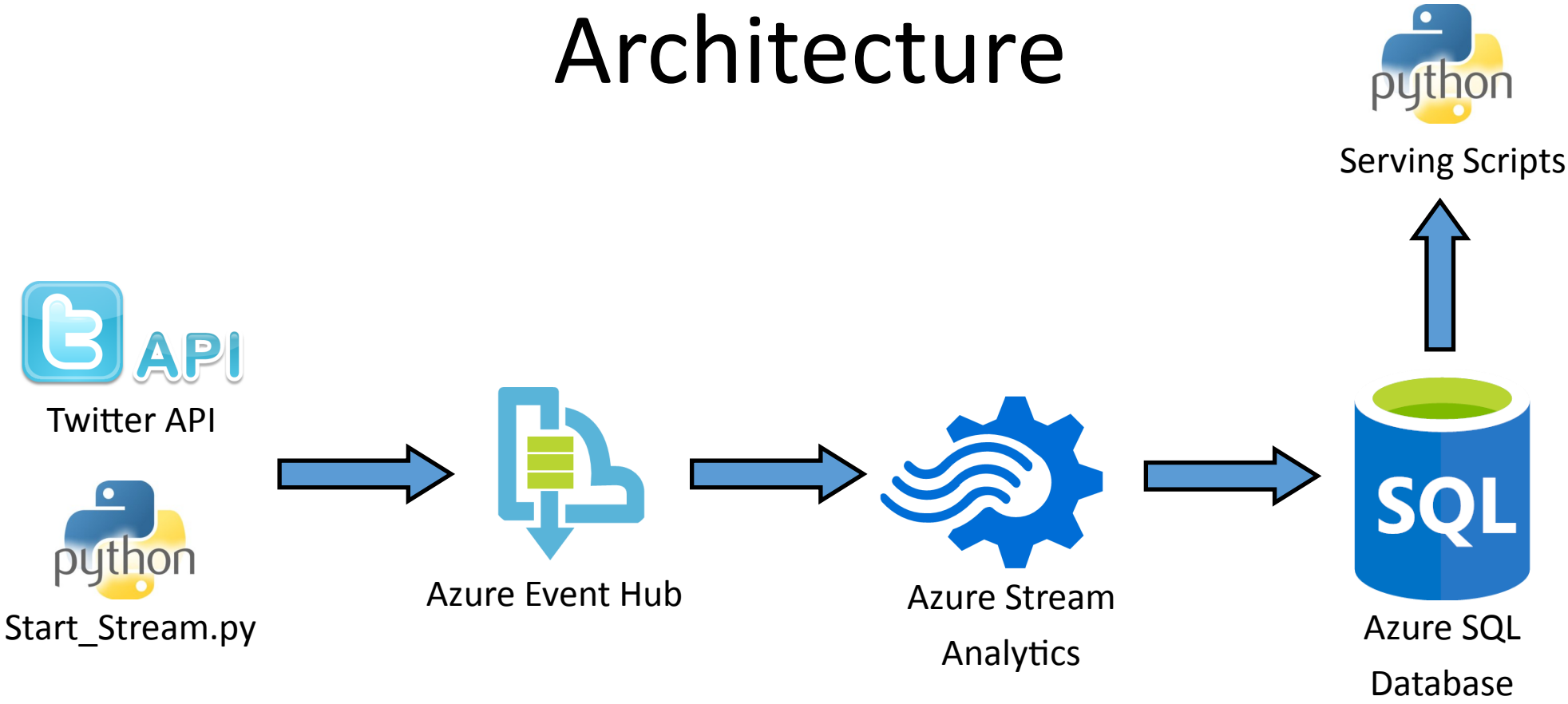
Architecture



As we talked about, I used Microsoft Azure for this exercise. This ended up being a great learning experience.

Start_Stream.py - Starts the stream of tweets (at least it does on my computer. It should work on others too). The script also processes the tweet to parse, clean and capture the words of the tweet itself and the count of those words. The script then sends that map/reduced tweet to the Azure Event Hub for processing. I will leave the Azure event hub, stream analytics job and SQL database running until I receive a grade in order to capture, process and store the data from the Start_Stream.py script.

Azure Event Hub—Captures and queues data for processing.

Azure Stream Analytics—This product from Microsoft captures streaming data, processes it and stores it. The processing has a convenient SQL interface (see attached screenshot). By the way, I did find out that Azure offers Apache Storm through HDInsight, but only after I was about 90% done.

finalresults.py - Functions like the one required for assignment. This uses an ODBC connection to connect the Azure SQL database that captures the tweet words and their counts. This will run without Start_Stream.py and should work on any computer (needs pandas library).

histogram.py - Functions like the one required for assignment. Connects to the Azure SQL database the same way as finalresults.py.