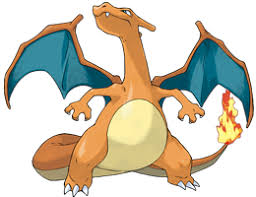
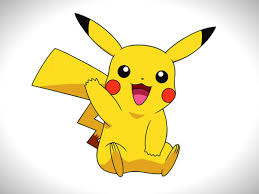
**Final Technical Project Report for ScrapeMon ETL Project**

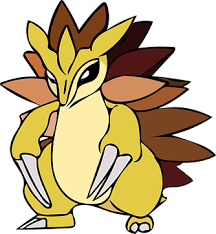
|  |
| --- |
| **Project Assignment says:** Due to the short timeline, teamwork will be crucial to the success of this project!  Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week. |



Steve



Sasha



Tad



Charlie

**Project Assignment says:**

#Project Proposal

Before you start writing any code, remember that you only have one week to complete this project. View this project as a typical assignment from work. Imagine a bunch of data came in and you and your team are tasked with migrating it to a production data base.

Take advantage of your Instructor and TA support during office hours and class project work time. They are a valuable resource and can help you stay on track.

**First Day: (Wednesday April 3rd, 2019)**

Using the following websites: <https://www.serebii.net/> and [https://www.pokeapi.co](https://www.pokeapi.co/) we will scrape data for 151 Pokemon.

Then, after scraping the data off of both websites, we will put the accumulated data into two dictionaries. After that, we will transform the dictionaries into DataFrames. Finally, we will turn the DataFrames into SQL or NoSQL tables.

During class, the group that had worked together during the SnakeBait project agreed to work together on the ETL project. Initial items identified as challenges by the team members from the ReadMe.md document included:

* *Short time frame*. Experiences from the first project took longer than estimated/expected.
* Determining the individual *tasks to be performed*.
* *Assigning responsibilities* to team members (not be able to gauge the confidence/strength of the team members for specific skills required).
* *Topic and selection of Data files* to be downloaded and merged. Considerations included access and availability of HTML & API data of the topic area.

The team considered the available assets to address the Project parameters and utilizing team skills

* Due to the short time frame, *Strong working relationships* formed during the SnakeBait project, would overcome time and functioning limitations.
* Tasks were voluntarily accepted (*self-assigned*) by members comfort level and perceived ability.
* Making sure that *Charlie* had NO input on topic selection.
* Sasha had proposed the *Pokemon© game characters* as a topic area and merging data that would be widely available from a diversity of sources.
* Initially proposed was the data to be screen-scraped from two different web sites. Input from TA’s and Instructor were sought out and it was suggested and accepted that the data could be obtained from a screen scrape and an API source demonstrating a diversity of data types to be merged in the ETL process.
* We met as a team on Thursday using Zoom, agreeing to meet on Friday.
  + At this meeting we agreed to the following assignments:
    - Sasha the API’s extraction.
    - Steve to scrape the HTML screens.
    - Tad to join the data into a common database table – Ben had suggested NoSQL Mongo. Ben proposed that we use MongoDB and that we consider using two different types of technology to demonstrate our understanding of ETL.
    - Charlie ?? Chhhhhaaaarrrrrllllliiieeee!!! Oh, Charlie would write this report and help where needed.
* We re-connected as a team using Zoom on Friday.

We reviewed team progress and made some minor changes.

At the 1st Zoom meeting (based on the type of the data we were gathering we would join Pokemon Characters and names with unique data from a second data site.

**Project Assignment says:** \* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

It was proposed by Tad that using Mongo as the database might be more difficult than using MySQL.

We settled on a Pokemon API (1st data set) and a widely used and detailed Pokemon user site for screen scraping (2nd data set). Then the plan called for TRANSFORMING(if some cleanup could be done at this point) the data into 2 data frames in python to be EXTRACTED to a database using MySQL Additional cleanup if necessary).

**Project Assignment says:**

Your project must use 2 or more sources of data.

\* The sources of data that you will extract from.

Based on the final database data table we anticipated using a join based on the unique Pokemon character number assigned to all characters. Relations might be used. The value to this ETL process is that we get unique data from one source and match it correctly to the other.

This document represents the Final Technical Project Report with the above (and additional) information and steps required to reproduce our ETL process

The two data sources chosen were:

1. <https://pokeapi.co/> for API’s producing pokeapi.csv data file.
2. <https://www.serebii.net> for screen scraping producing pokeweb.csv data file.
3. The Jupyter Notebook used was:

<https://github.com/sashabridges/ScrapeMon/blob/master/GottaScrapeEmAll%20Sasha%20Version.ipynb>

Please upload the report to Github and submit a link to Bootcampspot. (On Monday).