**ETL Project (Magic the Gathering ETL)**

**Gibran Ahmad** **∙ Paul Dean ∙ Soobin Hwang ∙ Seunghwa Jun**

**EXTRACT** (Original Data Sources and Formats)

* **Magic The Gathering(MTG, collectable card games) from Kaggle**
* Source : <https://www.kaggle.com/mylesoneill/magic-the-gathering-cards>
* Format : JSON file ("./AllCards.json")
* **Card Price (War of the Spark)**
* Source : https://www.mtggoldfish.com/index/XLN\_GRN#paper
* Format : HTML file (pd.read\_html)

**TRANSFORM** (Data Cleaning or Transformation Process)

* **① Use ‘pd.read\_json’ to create a DataFrame of the cardinfo data**
* json\_file = "./AllCards.json"
* mtg\_df = pd.read\_json(json\_file)
* **② Tanspose the data, so that the card names go down vertically and into the index**
* mtg\_df = pd.DataFrame(mtg\_df.transpose())
* **③ Clean up the dataframe, select and rename columns**
* clean\_mtg\_df = mtg\_df[["index","convertedManaCost", "text", "type"]]
* clean\_mtg\_df = clean\_mtg\_df.rename(columns={"index":"cardname"})
* **④ Use pd.read\_html to create a Dataframe of the price information data**
* price\_url = 'https://www.mtggoldfish.com/index/XLN\_GRN#paper'
* price = pd.read\_html(price\_url)
* **⑤ Remove the first two rows of the dataframe as they are not relevant.**
* price\_table = price[2:].copy()
* df = pd.DataFrame(price\_table[0])
* **(6) Select columns that are relevant and set ID.**
* clean\_price\_df = df[['Card', 'Price']]
* clean\_price\_df.index.name = 'id'

**ETL Project** (2/2)

**LOAD** (Final Database and Tables)

* **Create SQLite database and Push Data inti it**
* Import squlalchemy and connecting to DBS
* Import methods needed to abstract classes into tables
* Declare column types
* **Set up a connection to the database**
* **Create the tables associated with the classes**
* **Push the pandas dataframe into the SQLite database**
* clean\_mtg\_df.to\_sql('cardinfo', con=engine, if\_exists='replace')
* clean\_price\_df.to\_sql('priceinfo', con=engine, if\_exists='replace')
* **DDD**

**Note : Please add the Load items**

**And you have to specify REASONS why this was chosen (Read HW Guidelines Readme)**

**Also, I think it would be great we can capture a db image here~**