

Create MTG Trading Card Database By Crawler

Practical Exam




Goal

magicthegathering.io provides up-to-date information regarding all MTG trading cards available:




- <https://gatherer.wizards.com/Pages/>

Manta Riders


Details | Sets & Legality | Language | Discussion



Oracle

Card Name: Manta Riders
Mana Cost: 
Converted Mana Cost: 1
Types: Creature — Merfolk
Card Text:  Manta Riders gains flying until end of turn.
Flavor Text: "Water is firmament to the finned."
—Oracle en-Vec
P/T: 1 / 1
Expansion:  Tempest
Rarity: Common
Artist: Kaja Foglio

Printed

Community Rating:

Community Rating: 2,630 / 5 (27 votes)
[Click here to view ratings and comments.](#)

```
<div class="smallGreyMono" style="margin-top: 5px;">
<b class="ft"><b></b></b>
<div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_nameRow" class="row">
  <div class="label">
    Card Name:</div>
    <div class="value">
      Manta Riders</div>
    </div>
    <div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_manarow" class="row manarow">
      <div class="label" style="line-height: 25px;">
        Mana Cost:</div>
      <div class="value">
        </div>
      </div>
      <div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_cmcrw" class="row">
        <div class="label" style="font-size: .7em;">
          Converted Mana Cost:</div>
        <div class="value">
          1</div>
        </div>
        <div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_typerow" class="row">
          <div class="label">
            Types:</div>
            <div class="value">
              Creature — Merfolk</div>
            </div>
            <div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_textrow" class="row">
              <div class="label">
                Card Text:</div>
                <div class="value">
                  <div class="cardtextbox" style="padding-left:10px;">: Manta Riders gains flying until end of turn.</div></div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```

<https://gatherer.wizards.com/Pages/Card/Details.aspx?multiverseid=4711>

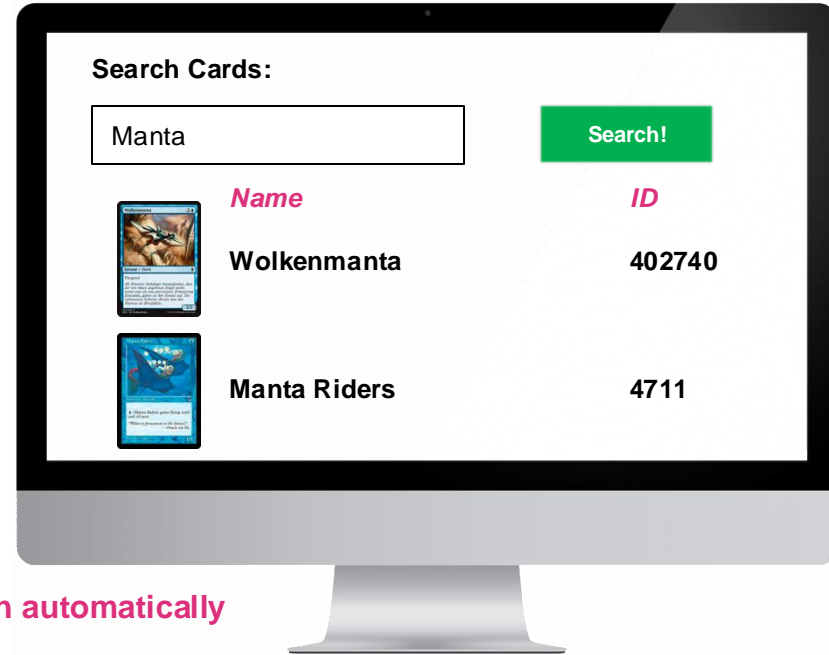


Goal

We want to make use of this data to build a searchable database of all MTG trading cards.

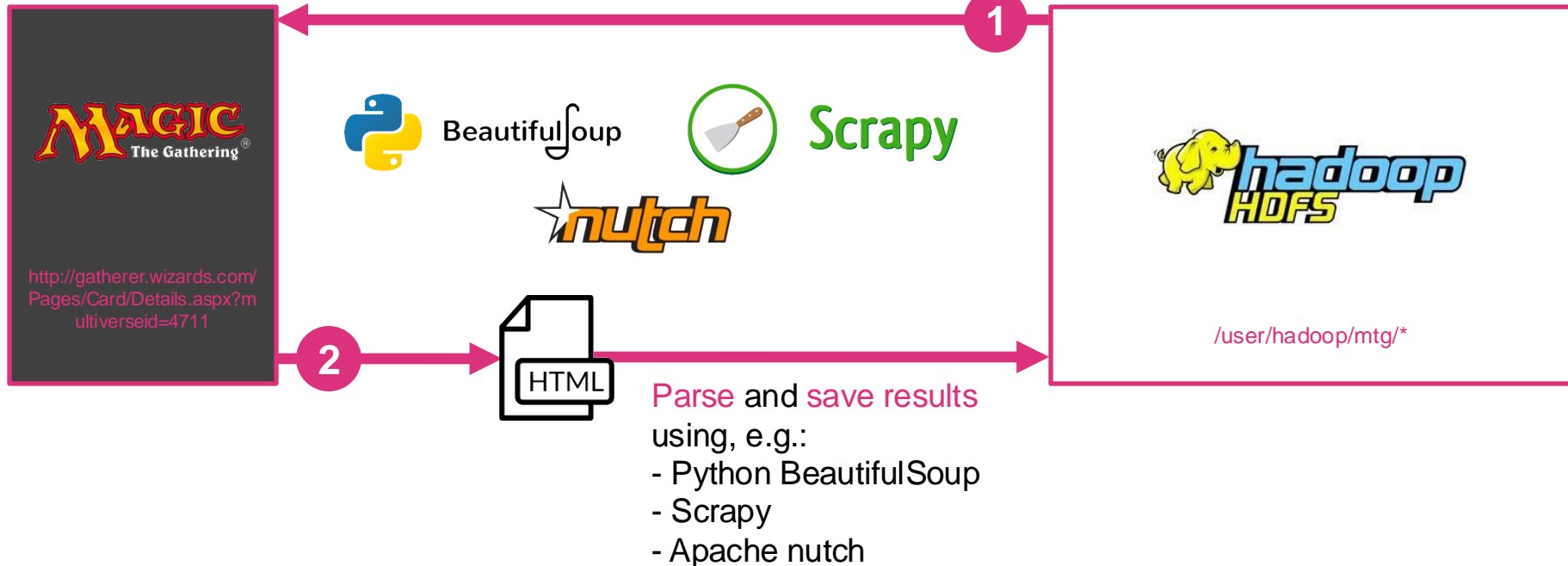
Workflow:

- **Crawl** data from gatherer.wizards.com
- **Parse** required **information** and save them to **HDFS** (queryable through Hive)
- **Export** MTG data to **end-user database** (e.g. MySQL, MongoDB...)
- Provide a simple **HTML Frontend** which is able to:
 - read from end-user database
 - process user input (card name, text or artist)
 - **display search results**
- The whole data workflow **must be implemented** within an ETL **workflow tool** (e.g. Pentaho Data Integration or Airflow) and **run automatically**



Dataflow: 1. Get MTG Data

Crawl HTML pages



Dataflow: 3. Enhance Data And Save Results



/user/hadoop/mtg/*



1

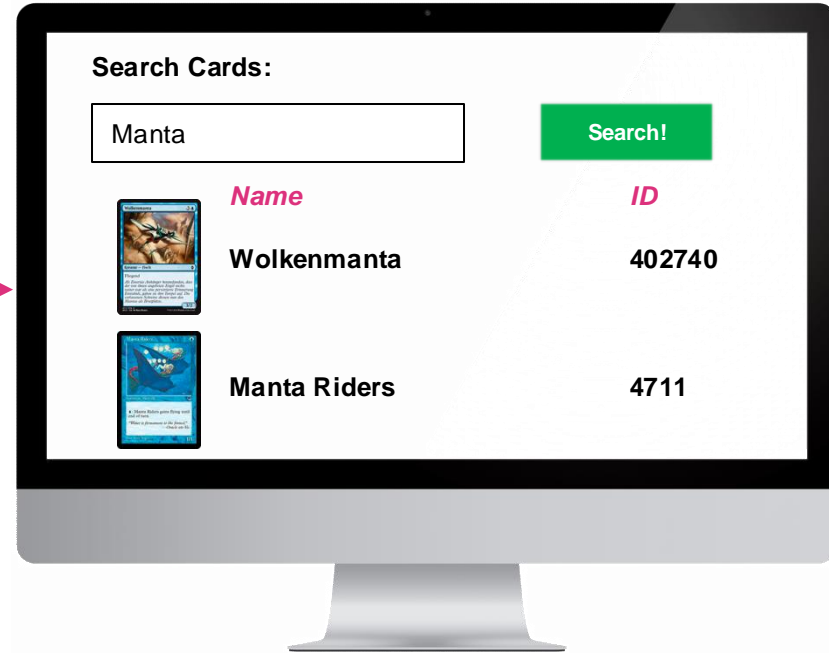
- enhance data (e.g. for later querying)
- use *Hive*, *Python*, *Spark* or *PySpark*
- save everything to a end-user database (e.g. *MySQL*, *MongoDB*)



Dataflow: 4. Provide Simple Web Interface



1



- Provide a simple **HTML Frontend** which is able to:
 - read from end-user database
 - process user input (card name, text or artist)
 - **display search results**