# Create MTG Trading Card Database By Crawler

Practical Exam

# Goal

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

- https://docs.magicthegathering.io/



http://gatherer.wizards.com/Pages/Card/Details.aspx?multiverseid=4711

```
<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">
      <img src="/Handlers/Image.ashx?size=medium&amp;name=U&amp;type=symbol" alt="Blue" align="absbottom" /></div>
</div>
  <div id="ctl00_ctl00_ctl00_MainContent_SubContent_SubContent_cmcRow" 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;"><img src="/Handlers/Image.ashx?size=small&amp;name=U&amp;type=symbol" alt="Blue" align="absbottom" />: Manta Riders gains flying until end of turn.</div></div>
</div>
</div>
```
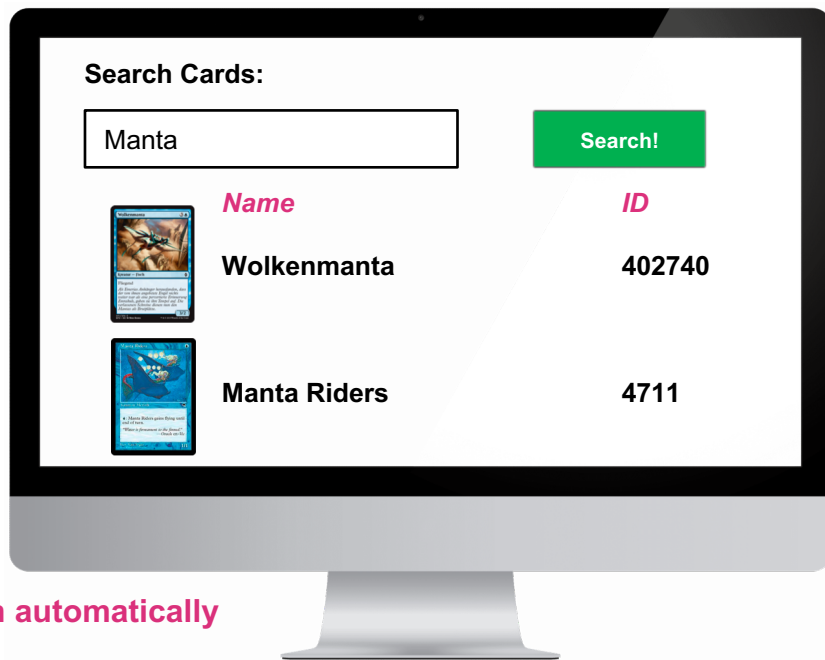
www.marcel-mittelstaedt.com

# Goal

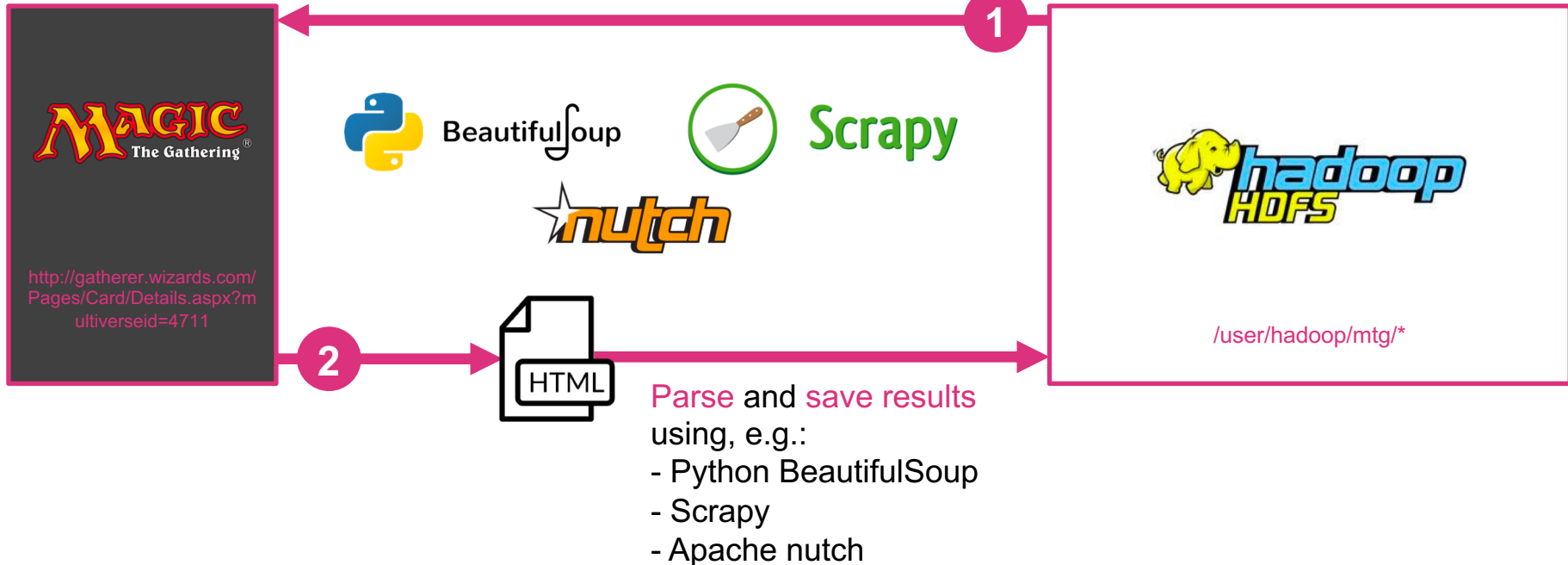We want to make use of this data to build a searchable database of all MTG trading cards.

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**

**Search Cards:**

| | Manta | | Search! |

| Name | ID |
| --- | --- |
| Wolkenmanta | 402740 |
| Manta Riders | 4711 |

Crawl HTML pages

1

2

http://gatherer.wizards.com/Pages/Card/Details.aspx?multiverseid=4711

/user/hadoop/mtg/*

Parse and save results using, e.g.:
- Python BeautifulSoup
- Scrapy
- Apache nutch

# Dataflow: 3. Enhance Data And Save Results



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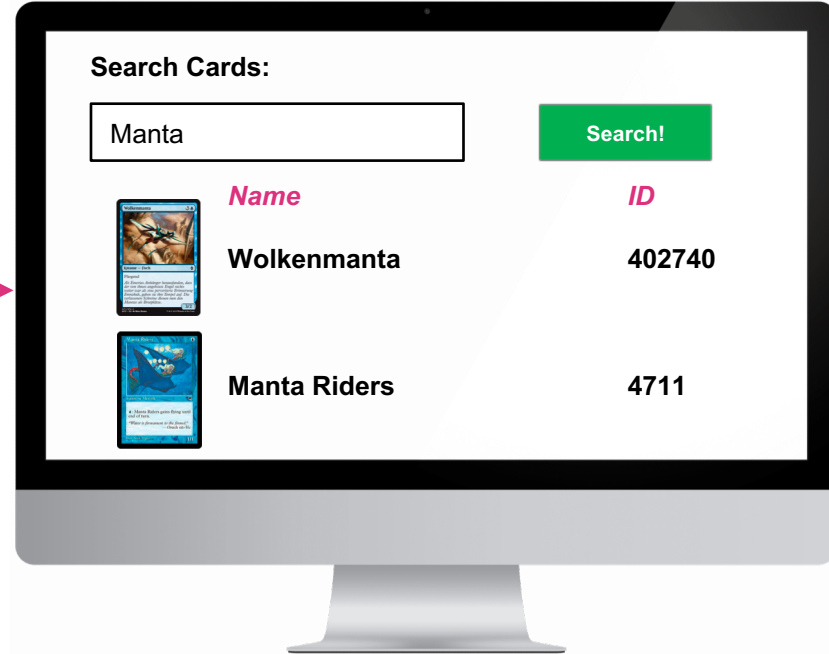- save everything to a end-user database (e.g. *MySQL*, *MongoDB*)

/user/hadoop/mtg/*

**www.marcel-mittelstaedt.com**

# Dataflow: **4. Provide Simple Web Interface**



- 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**