Herramientas del Big Data

Introducción a la Ingeniería de Datos

Contenido de la clase

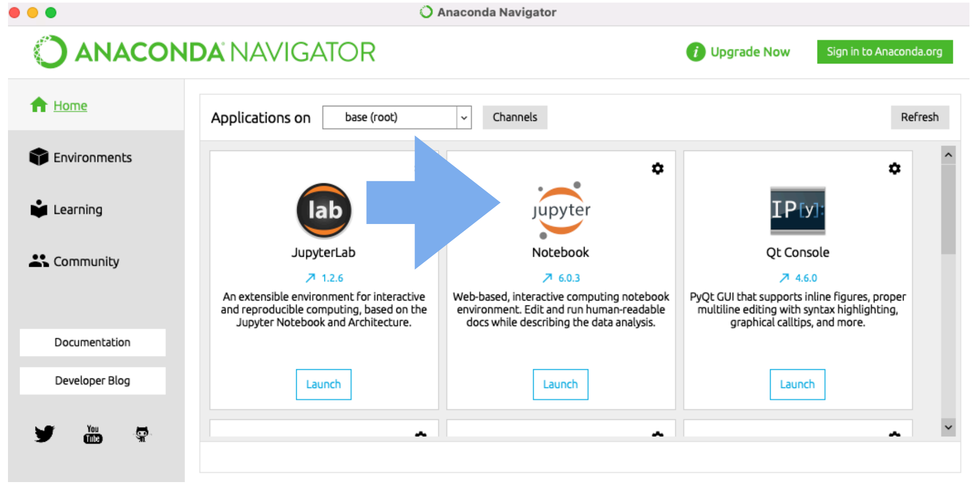
- Introducción a la Ingeniería de datos
- Tipos de datos
- Manejo de datos (Script)

Ingeniería de datos

Antesala

Requisitos

- Por favor descargue Python https://www.python.org/
- Por favor descargue Anaconda https://anaconda.org/anaconda/python
- Desde Anaconda abra Jupyter Notebook
- El Jupyter será nuestra libreta de notas y desarrollo para esta clase



Ingeniería de datos

- La ingeniería de datos tiene que ver con procesos de bases de datos e integración de soluciones.
- En el Big Data las soluciones se desarrollan en la nube, por lo cual acá veremos algunos conceptos con small data.
- Una de las funciones principales de la ingeniería de datos son :
  1. Obtener los datos
  2. Limpiar y estructurar los datos
  3. Crear los pipelines para automatizar los procesos
  4. Generar un mínimo descriptivo de los datos.

Ingeniería de datos

Importancia

- La raíz del trabajo en Big Data se define como ETL :
  * Extraer;
  * Transformar;
  * Cargar

- Los anteriores pasos definen el Pipeline del trabajo que siempre se debe hacer a la hora de entender un problema de Big Data.

FIREBLAZE AI SCHOOL

Data Engineers

Also known as Data Architects

• Tools that need to be mastered •

Python

hadoop

NoSQL

• Skills that need to be mastered •

Programming

Data Mining

Database architecture

Statistical modeling & regression analysis

Tipos de datos

Tipos de datos

Herramientas del Big Data

- Primitivos : Booleans, floats , str, int
- Estructurados : Bases de datos
- Semi Estructurados : Json , APIS
- No estructurados : HTML, TXT
- Cualitativos , Cuantitativos.

Fuentes de datos

Fuentes de datos

- Los datos se obtienen a través de API.
- Un ejemplo es el calendario de Google.
- Otra forma de descargar datos son los Logs.
- Otra fuente de datos es el User Analytics.

ETL

- Extraer: Lectura de Bases de datos
  + Bases de datos de aplicaciones
  + CRM

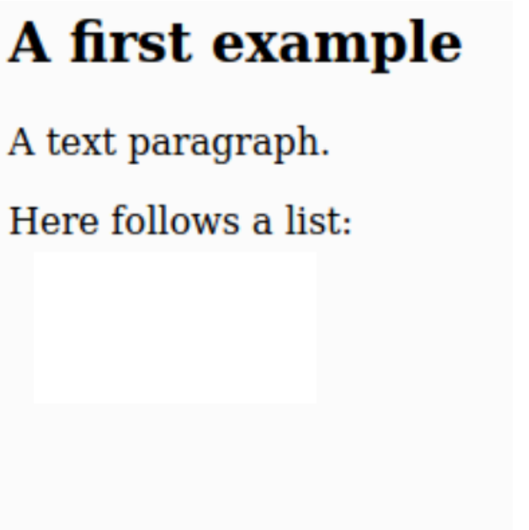
- Transformar : Estructura de los datos
  + Limpieza
  + Correctos formatos

- Cargar: Llevar la data a un Warehouse.

Web Scraping

- Es una forma de extraer información de la WEB
- Se basa en el entendimiento de :
  + HTML : Estructura de la página (Header, Tab)
  + CSS : Estética de la página
  + Javascript : Computo a la página
  + JSON : Datos.

```
<html>
  <body>
    <h2>A first example</h2>
    <p>A text paragraph.</p>
    <p>
      Here follows a list:
    </p>
  </body>
</html>
```



Ejemplo tomado de Datacamp

A first example

A text paragraph.

Here follows a link.

```
...
    <p>
      Here follows a
      <a href="https://google.com">link</a>.
    </p>
...
```

Se identifica el Tag

```
{html_document}
<html>
[1] <body> \n    <h2>A first example</h2>\n    <p>A text paragraph.</p>\n   ...
```

Se genera la lectura de los datos

Se construye la Base de datos

¿Cómo extraer info de la web?

- Request library -> Python
- Rvest -> R
- Se envía una solicitud para descargar los datos

```
[3]: requests.get("https://www.semana.com/")

[3]: <Response [200]>
```

Puedes descargar información de ahí

```
requests.get("https://www.algunapagina.com")

'<Response [400]>'
```

No hiciste bien la solicitud

```
import bs4
```

```
soup = bs4.BeautifulSoup(out.text,'html.parser')
```

```
print("Titulos")
print("-"*32)
print(soup.title.text)
print("-"*32)
```

```
Titulos
--------------------------------
Semana.com -  Últimas Noticias de Colombia y el Mundo
--------------------------------
```

```
Elements  Console  »  ⚠ 7
<div class="header-box">
  <div class="header-section">
    <div class="header-top">…</div>
    <div class="sticky-observer"></div>
    <div class="header-nav">
      <nav class="header-navbar">
        <a class="navbar-item active" href
          ::before
          <span>Últimas noticias</span> ==
        </a>
        <a class="navbar-item" href="/tv/"
        <a class="navbar-item" href="/econ
        ">…</a>
        <a class="navbar-item" href="/impr
        …</a>
        <div class="navbar-item hidden-md
        </div>
```

```
[46]: ['/vida-moderna/articulo/ya-las-vio-estas-son-las-series-favoritas-de-bill-gates-en-
      netflix/202141/',
       '/educacion/articulo/educacion-virtual-6-de-cada-10-docentes-cree-que-acompanamient
      o-de-padres-ha-sido-regular/202118/',
       '/mundo/articulo/rapero-podria-pagar-cuatro-anos-de-carcel-por-cortarle-el-pene-a-s
      u-amigo-a-cambio-de-dinero/202159/',
       '/gente/articulo/habla-el-cuidador-de-los-perros-de-lady-gaga-que-recibio-un-dispar
      o-durante-el-robo/202119/',
       '/nacion/',
       '/nacion/articulo/cortes-de-luz-hoy-martes-2-de-marzo-en-13-zonas-de-bogota/20215
```

```python
news2 = []
for new in news:
    news2.append("www.semana.com"+new)
```

```
'www.semana.com/vida-moderna/articulo/ya-las-vio-estas-son-las-series-favoritas-
ill-gates-en-netflix/202141/',
'www.semana.com/educacion/articulo/educacion-virtual-6-de-cada-10-docentes-cree-
acompanamiento-de-padres-ha-sido-regular/202118/',
'www.semana.com/mundo/articulo/rapero-podria-pagar-cuatro-anos-de-carcel-por-cor
e-el-pene-a-su-amigo-a-cambio-de-dinero/202159/',
'www.semana.com/gente/articulo/habla-el-cuidador-de-los-perros-de-lady-gaga-que-
bio-un-disparo-durante-el-robo/202119/',
'www.semana.com/nacion/',
'www.semana.com/nacion/articulo/cortes-de-luz-hoy-martes-2-de-marzo-en-13-zonas-
ogota/202155/',
```