

Introducción a la Ciencia de Datos

Dr. Leon Felipe Palafox Novack Ipalafox@up.edu.mx

0

Noticias del día

Qué ha pasado en el mundo de Data Science?







SCIENCE & HEALTH & ARTS & EDUCATION & ECONOMY

TECHNOLOGY WELLNESS HUMANITIES ECONOMY

UT News > Press Releases > Dell Medical School Launches Data Hub to Accelerate Biomedical Research, Advance Health

Dell Medical School Launches Data Hub to Accelerate Biomedical Research, Advance Health

Aug. 7, 2018



AUSTIN, Texas – Dell Medical School at The University of Texas at Austin is accelerating innovation and research by creating a Biomedical Data Science Hub to help solve complex research and clinical problems.

Impoins having a complicated scientific question: How do we predict who will be discoved with Type



1

Anuncios parroquiales

Proyecto Final



Proyecto Final



El objetivo es que trabajen en el proyecto a lo largo del curso.

Conforme vayamos aprendiendo las herramientas, se recomienda las practiquen con sus propios datos.

Vayan formando sus equipos oportunamente.



Recursos



- Los tutoriales de Python van a estar disponibles toda la clase
- El machote para el reporte final ya esta en la pagina web.



2

Datos

Tenemos que entender nuestra materia prima

The world's most important resource is no longer oil, is data

The Economist

Qué es un dato?



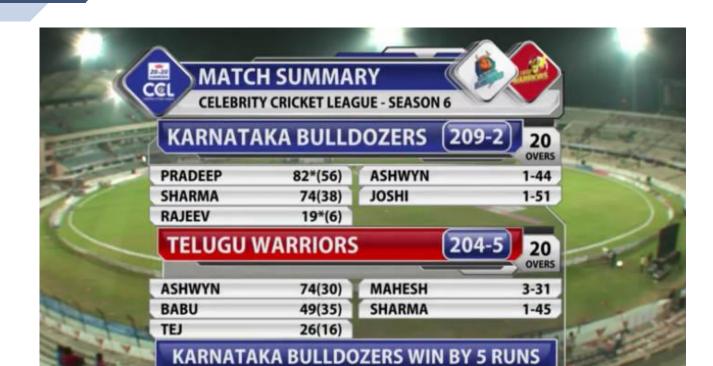
- Es la unidad mínima de información?
 - Es aquello, que dado un contexto, nos puede proveer de información.
 - El contexto por lo general se puede inferir.
 - Ejemplos:
 - Una foto
 - Un marcador















Cuantos tipos de datos hay?



- Hay muchas definiciones acerca de los tipos de datos que hay:
 - Continuos y discretos
 - Big and Small

Aqui vamos a hablar de estructurados, no estructurados y semi estructurados.



Datos Estructurados y No estructurados

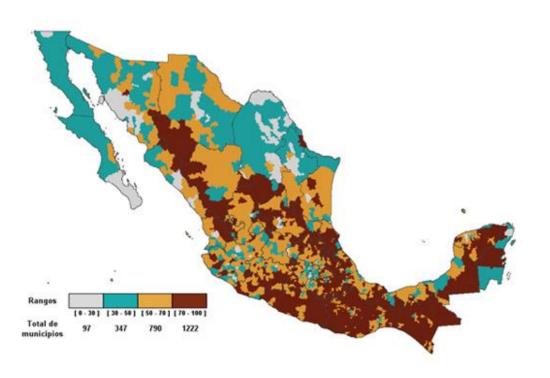






Datos Estructurados y No estructurados





Pobreza a nivel municipio (CONEVAL)





- Se estima que el grueso de los datos que existen son no estructurados
- Mucho valor del negocio esta escondido en datos no estructurados
- Imagenes, Audio, Video













- Son datos que cuentan con una estructura predefinida.
 - Estructuras de Datos (También llamados modelos de datos)
 - Colas, Listas, Arreglos, Pilas (stack)
 - Tablas como hojas de cálculo, un archivo de excel, etc
 - Mapas cartográficos tipo GIS (Google Earth)

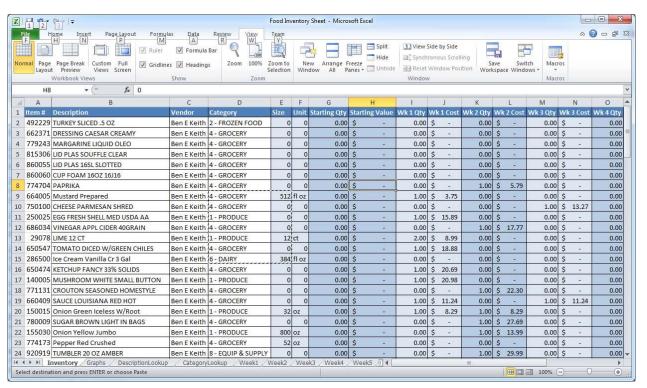




Son aquellos que organizan elementos de los datos y tienen patrones de como se relacionan uno con otro, y como se relacionan con el resto del mundo.

- También se dice que un dato estructurado es aquel que se asocia a un modelo de dato.
 - Un modelo de dato da estructura y capacidad de relación al dato



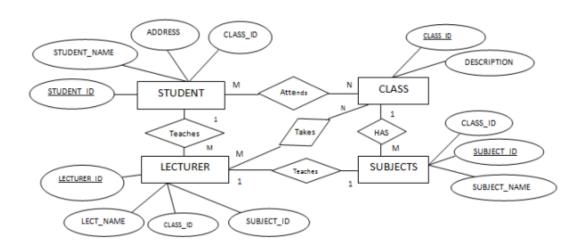








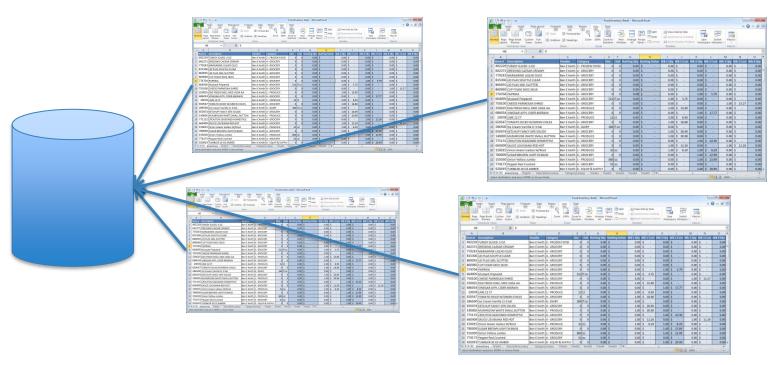
- Bases de Datos
 - Diagrama Entidad Relación



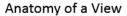


Base de Datos

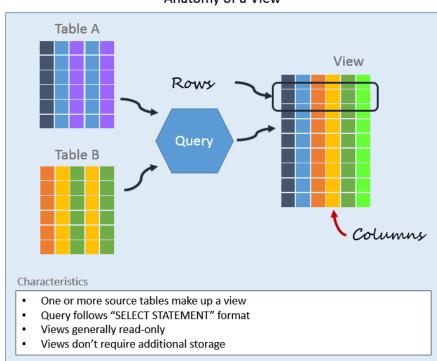
















Para que querriamos una vista?

Por que es importante conocer el diagrama entidad-relación?

Que otras cosas necesita una base de datos?





- Componentes de una base de datos:
 - Diagrama Entidad Relación
 - Catálogo de Datos
 - Bitácora de Datos



Datos semi-estructurados



Hay varios formatos, pero en la industría estan los dos mas usados: "Rail Booking": (

→ JSON

```
"Rail Booking": {
    "reservation": {
        "ref_no": 1234567,
        "time_stamp": "2016-06-24T14:26:59.125",
        "confirmed": true
},
    "train": {
        "date": "07/04/2016",
        "time": "09:30",
        "from": "New York",
        "to": "Chicago",
        "seat": "57B"
},
    "passenger": {
        "name": "John Smith"
},
    "price": 1234.25,
    "comments": ["Lunch & dinner incl.", "\"Have a nice day!\""]
}
```



Datos semi-estructurados





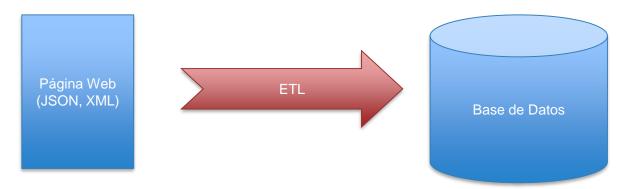
```
<?xml version="1.0"?>
<quiz>
<qanda seq="1">
 <question>
  Who was the forty-second
  president of the U.S.A.?
 </question>
 <answer>
  William Jefferson Clinton
 </answer>
</ganda>
<!-- Note: We need to add
 more questions later.-->
</quiz>
```



Usos de JSON y XML



Se utilizan primordialmente para transportar datos de un lado al otro







ETL



- **Extract, Transform and Load:**
 - Que creen que hace un ETL?

Por que necesitamos ETLs?

Creen que hay negocio haciendo ETLs?



ETL



- Un ETL (Extract, Transform, Load) se encarga de extraer los datos de una fuente, y transformarlos para que se coloquen en otra.
- Dichas transformaciones son por lo general simples y no involucran algoritmos complejos.
- Un uso muy común de un ETL es cuando se migran bases de datos o arquitecturas.



Data Lake vs Data Warehouse



- En la industria de Big Data hay dos conceptos:
 - Data Lake: Almacena TODOS los datos que se pueda, estructurados y no estructurados.
 - Data Warehouse: Almacena sólo datos estructurados. Tiene un esquema mucho mas estricto acerca de que tipo de datos se van a almacenar.
- En un Data Lake es muy dificil encontrar información, mientras que en un Data Warehouse es fácil. (No simple, sólo facil)



Como transformar datos?



- Que se necesita para hacer que un dato no estructurado tenga estructura?
 - Que se necesita para que su album de fotos tenga estructura?
 - Que se necesita para que un libro tenga estructura?
 - Que se necesita para que una canción tenga estructura?



Introducción a Python

La herramienta del Data Scientist