BIG DATA Y PYTHON

MÁSTER EN BIG DATA

201020

GABRIEL MARÍN DÍAZ

www.linkedin.com/in/XXX


hola

Presentación

Yo mismo

Nombre: Gabriel Marín Díaz

A qué me dedico…

- Channel Enablement Manager en Sage
- Profesor Asociado UCM

Perfil de LinkedIn: https://www.linkedin.com/in/gabrielmarindiaz/

CONTENIDO

Contenido

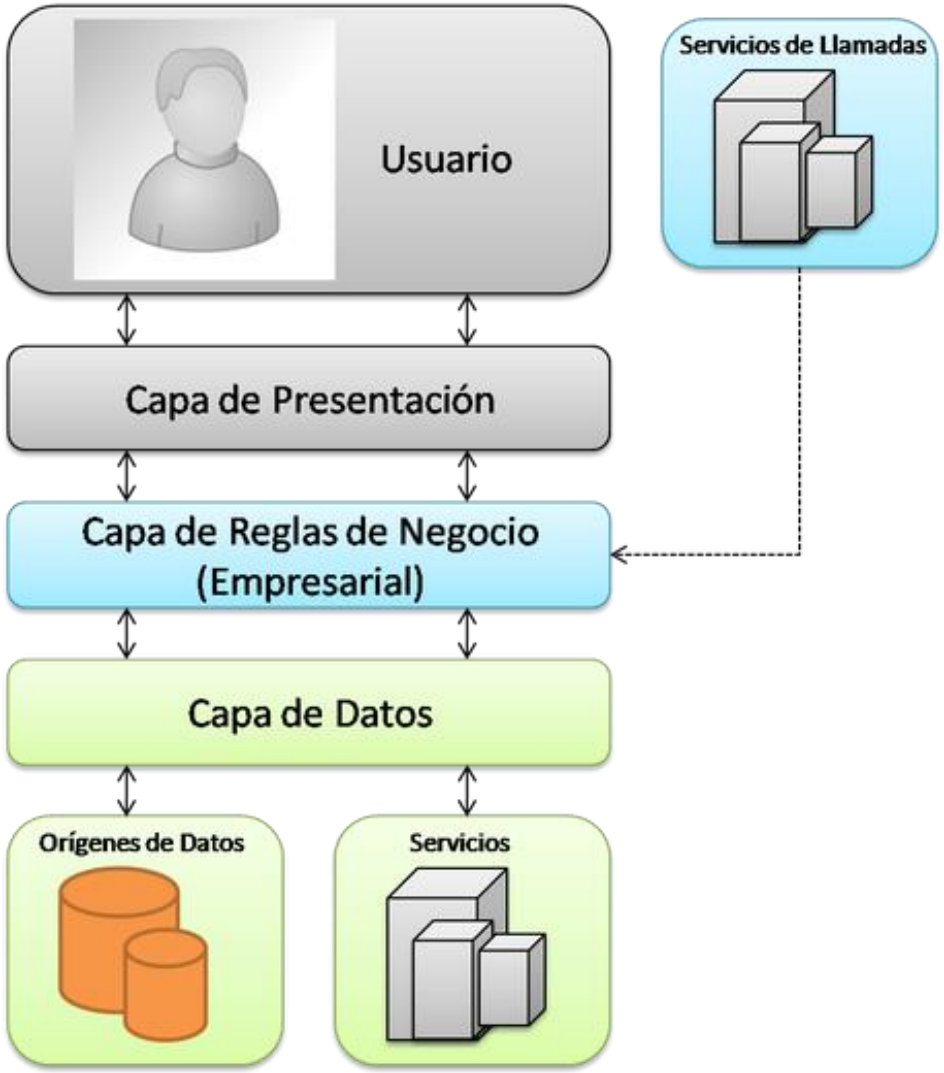
Resumen

Tema 1 – Visión General
Tema 2 – Introducción a SQL
Tema 3 – Introducción al Lenguaje Python
Tema 4 – HTML y Python
Tema 5 – Big Data y Python
Tema 6 – Procesamiento Distribuido (Spark)

Prácticas las realizaremos con Python, MySQL, MongoDB, Apache Spark

Arrancamos?

Arquitectura en capas



1. Capa de Presentación: Interacción entre el usuario y el software. Puede ser tan simple como un menú basado en líneas de comando o tan complejo como una aplicación basada en formas. Su principal función es mostrar información al usuario, interpretar los comandos de este y realizar algunas validaciones simples de los datos ingresados.

2. Capa de Reglas de Negocio (Empresarial): También denominada Lógica de Dominio, esta capa contiene la funcionalidad que implementa la aplicación. Involucra cálculos basados en la información dada por el usuario, datos almacenados y validaciones. Controla la ejecución de la capa de acceso a datos y servicios externos.

3. Capa de Datos: Esta capa contiene la lógica de comunicación con otros sistemas que llevan a cabo tareas por la aplicación. Para el caso de aplicaciones empresariales, está representado por una base de datos, que es responsable del almacenamiento persistente de información. Esta capa debe abstraer completamente a las capas superiores (negocio) del dialecto utilizado para comunicarse con los repositorios de datos (PL/SQL, Transact-SQL, etc.).

Contenido

Resumen

Tema 1 – Visión General
Tema 2 – Introducción a SQL
Tema 3 – Introducción al Lenguaje Python
Tema 4 – HTML y Python
Tema 5 – Big Data y Python
Tema 6 – Procesamiento Distribuido (Spark)

Prácticas las realizaremos con Python, MySQL, MongoDB, Apache Spark

ALGUNOS CONCEPTOS DE ESTADÍSTICA DESCRIPTIVA

Librería Pandas

ESTADÍSTICA

ALGUNOS CONCEPTOS DE ESTADÍSTICA DESCRIPTIVA COMO MEDIA, MEDIANA, MODA, VARIANZA, CUARTILES, DESVIACIÓN TÍPICA

GESTIÓN DE DATOS CON PANDAS

Librería Pandas

VAMOS A TRABAJAR CON EL FICHERO CSV base_datos_2008.csv CORRESPONDE CON LA INFORMACIÓN DE VUELOS DE EEUU Y CON LA LIBRERÍA PANDAS (pandas.ipynb)

Librería Matplotlib

VAMOS A TRABAJAR CON EL FICHERO CSV base_datos_2008.csv CORRESPONDE CON LA INFORMACIÓN DE VUELOS DE EEUU Y CON LA LIBRERÍA PANDAS (pandas.ipynb)

EJERCICIOS

Ejercicios

Utilizando el fichero de aerolíneas EEUU, realizar las siguientes actividades:

1. **Cargar el fichero en un dataframe.**
2. **Mostrar los primeros n-elementos y los últimos n-elementos del dataframe.**
3. **Entender y utilizar la función sample para reordenar un % de la base de datos, probad con distintos valores y analizar los resultados.**
4. **Describir los tipos de datos del dataframe.**
5. **Escoged una columna y detallar los valores en un máximo, mínimo, media en un rango del dataframe.**
6. **Mediante filtrado utilizar las funciones necesarias para obtener:**
   - **Aquellos vuelos que han sufrido un retraso en la salida mayor de 60 minutos .**
   - **Representar el número vuelos por ciudad (eje X) y su retraso (eje Y) utilizando gráficos más adecuados.**
   - **Hacer lo mismo para aquellos vuelos que han sufrido retraso en la salida y en la llegada de más de 60 minutos.**

Ejercicios

7. Averiguar el número de casos que no tienen ningún valor en la columna retraso de salida.
8. Obtener un nuevo dataframe que represente un subconjunto de datos del dataframe original en el que tengamos el número de vuelo, la fecha del vuelo, retraso en la salida, retraso en la llegada, la diferencia entre la llegada y la salida, ordenados por fecha y por aeropuerto de salida.
9. Crear una nueva columna en el dataframe que representa la diferencia entre el tiempo real del vuelo vs el tiempo programado.
10. Hacer lo mismo con las diferencias entre la hora programada de llegada y la hora real de llegada, y la hora progamada de salida y la hora real de salida.
11. Con las tres columnas nuevas creadas, más las columnas que representan la fecha y el número de vuelo crear un nuevo dataframe. Probar para ello la función adicción de dataframes.
12. Borrar las columnas recién creadas, comprobar que se hace un borrado real sobre el dataframe.
13. Obtener los estadísticos descriptivos de todos los vuelos cuyo origen está en IND, ATL, obtener los mismos para los distintos días de la semana.
14. Averiguar el número de datos repeti

Ejercicios

15. Probad la función dropna con el valor thresh, analizar cuál es su función para tratar datos perdidos.
16. Analizar la correlación de variables:
    - Retraso de salida, retraso de llegada.
    - Retraso de llegada vs distancia del vuelo.
    - Retraso de llegada vs tiempo de vuelo en el aire.
17. Analizar la correlación del conjunto de variables de vuelo.
18. Representar gráficamente las correlaciones.
19. Representar gráficamente en una nube de puntos la distancia recorrida en función del día del mes (eje x), retraso en la salida (eje y). Jugar con los parámetros de color, transparencia, añadir datos al gráfico.
20. Obtener el número de vuelos totales a la semana y representarlos en un diagrama de tarta. Obtener el porcentaje de vuelos de cada día de la semana en relación al total de días. Jugar con los parámetros de representación del diagrama de tarta.

Ejercicios

21. Hacer lo mismo con un diagrama de barras.
22. Obtener las curvas de desviación entre el tiempo programado de un vuelo y el tiempo real del mismo, el tiempo programado de salida y la hora de salida real, y la hora programada de llegada y la hora real. Seleccionar para ello el subconjunto de datos adecuado, agrupando por ejemplo por Origen.
23. Utilizar un histograma para representar los datos anteriores en función de la distancia, es decir, retraso de salida, retraso de llegada → Distancia; hora de salida programada, hora de salida real → Distancia; hora de llegada programada, hora de llegada real → Distancia.
24. Comparemos dos distribuciones utilizando kdeplot, elegir los parámetros que consideréis más adecuados, probad con distintos valores.
25. Representar en un diagrama de cajas en el eje X el retardo en la llegada versus Origen del vuelo para tres ciudades elegidas por vosotros. Interpretar los resultados.

Ejercicios

26. **Utilizar jointplot para representar en al menos cuatro ciudades de Origen el retraso en la llegada (eje y) vs retraso en la salida (eje x).**
27. **Hacer lo mismo para las variables hora estimada de llegada – hora real de llegada y hora estimada de salida – hora real de salida.**
28. **Probar las distintas configuraciones de gráfico ajustando el valor # kind { “scatter” | “reg” | “resid” | “kde” | “hex” } y ajustando los rangos de las variables.**
29. **Hacer un mapa de calor con los cuatro orígenes elegidos por mes y por semana de retrasos en salida, llegada, tiempo estimado vs tiempo real.**

¡Muchas Gracias!

GABRIEL MARÍN DÍAZ
LCDO. CIENCIAS FÍSICAS UCM

www.linkedin.com/in/gabrielmarindiaz/