Modelamiento de datos, adentrándose en la predicción:

Enunciado:

Una vez conocidas las herramientas de procesamiento de grandes volúmenes de datos, los analistas ya quieren empezarlas a utilizar. Para esto, quieren explorar diferentes metodologías de predicción con grandes volúmenes de datos e incorporar algunos análisis avanzados que permitan tomar mejores decisiones. Metodologías hay muchas, así como herramientas hay muchas, por lo que es importante conocer aquellas que son quizás más populares y utilizadas según la industria.

Objetivos:

- Seleccionar una metodología para análisis de grandes volúmenes de datos y utilizarla para responder alguna pregunta de negocio basados en los datos disponibles.
- Realizar la configuración de un ambiente de desarrollo con el uso de databricks.

Las principales metodologías disponibles para el análisis de grandes conjuntos de datos y Big Data son: CRISP – DM, DELTA y MAMBO.

CRISP – DM: Es una metodología altamente orientada a la elaboración de modelos de machine learning, se centra en un entendimiento claro y conciso tanto del problema como del negocio.

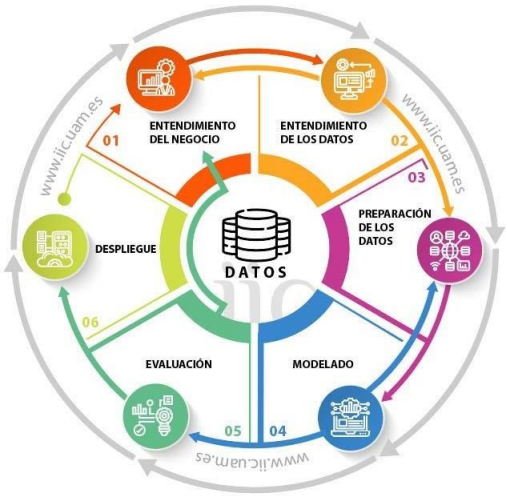

Fig 1. Metodologia CRISP – DM

Por su parte DELTA es un marco de trabajo o Framework el cual describe tres caras diferentes las cuales deberiamos de tener en cuenta a la hora de realizar el despliegue de un modelo: el cliente, el empleado y el proveedor.

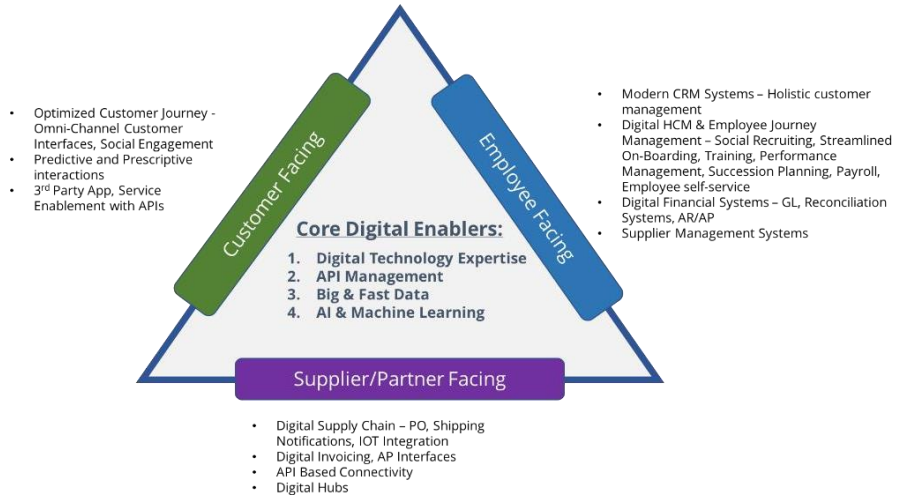

Fig 2. Delta Framework

Finalmente esta la metodologia MAMBO, la cual se centra en obtener informacion relevante para el negocio por medio de los datos, mas que en el uso de modelos de machine learning.

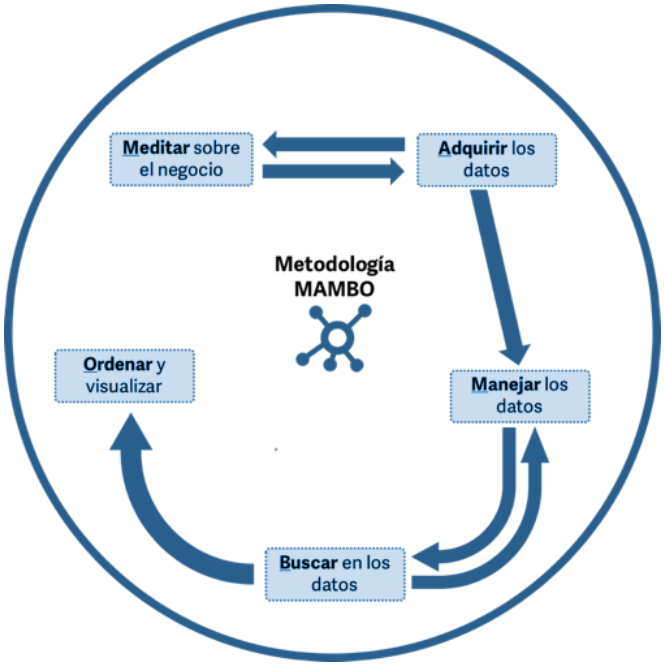

Fig 3. Metodologia MAMBO

Para el desarrollo de este ejericio en particular me apoyare en la metodologia MAMBO.

Fase 1: Meditar sobre el negocio

Dados los datos que poseemos podemos inferir que se tratan de datos pertenecientes a una empresa de retail o ventas al por menor, la cual posee sedes fisicas en diferentes ciudades de Colombia.

Fase 2: Adquirir los datos

Ya poseemos un set de datos.

Fase 3: Manejar los datos

Se procede a analizar si es posible o no la imputacion de los datos nulos, de lo contrario estos se elminan.

Fase 4: Buscar en los datos

Es en esta fase donde por medio del data wrangilng hacemos que los datos nos cuentan la informacion que poseen, informacion valiosa para el negocio.

En este caso se hallo cual fue la ciudad en la cual mas se ha vendido; la ciudad en la cual se registran mas ventas es Medellin con 27.699 ventas, respectivamente.

Fase 5: Ordenar y visualizar

En esta fase es donde nos ayudamos de herramientas como Power BI, Tableu o Click para la visualizacion de la informacion descubierta.

Configurando un ambiente en DataBricks:

El siguiente video explica paso a paso la configuracion de un ambiente de desarrollo en DataBricks: https://www.youtube.com/watch?v=DAV3p8rhWhM