FACULTAD DE CIENCIAS NATURALES E INGENIERÍA

INGENIERÍA DE SISTEMAS

Informe final Práctica **Laboral: Dot Profiling**

Empresa: GrupoDot

Presentado por:
Kevin Steven Perez Pineda

Tutor:
Juan Leonardo Padilla Gomez

Profesor:
Edgar Jose Ruiz Dorantes

Fecha Realización prácticas:
01/Julio/2020 - 31/Diciembre/2020

10/Diciembre/2020

Versión 3

1. ENFOQUE DE LA COMPAÑÍA

Grupodot es una compañía que ofrece servicios tecnológicos mediante la inteligencia artificial, en específico en el machine learning por medio de la estructuración de modelos analiticos que de acuerdo a un gran volumen de datos puede predecir comportamientos a futuro. En la compañía se utilizan diversos lenguajes de programación y frameworks al igual que herramientas, pero en esencia se maneja Google Cloud Storage (GCP) el cual es una plataforma que ofrece diversos servicios de tecnología, desde creación de máquinas virtuales hasta alojamiento de Big Data.

2. MIS PRÁCTICAS LABORALES

Durante el desarrollo de mis prácticas laborales en la empresa Grupodot, me desempeñé como desarrollador Front-end y back-end en un proyecto interno de dicha empresa, el cual consistía en implementar un nuevo servicio o framework mediante una interfaz intuitiva, en donde mostraba análisis estadísticos de grandes volúmenes de datos los cuales estaban alojados en GCP (Google Cloud Platform), esta herramienta fue bautizada como Dot profiling en la que su principal objetivo es mitigar el uso de frameworks de terceros como el caso de padas-profiling, ya que el constante uso de esta implicaría grandes costos a la compañía.

Mi aporte en este proyecto fue considerable ya que apoyé al equipo en el desarrollo y definición de diferentes métodos, consultas SQL y clases CSS para llevar a cabo la implementación del mismo. Esta experiencia fue enriquecedora ya que al poder participar en el desarrollo de este proyecto pude entender cómo se crea un proyecto en el ámbito empresarial y cómo se desarrollan buenas prácticas en implementación de desarrollo.

3. RESUMEN

De acuerdo con los requerimientos de GrupoDot, se necesita un framework o software propio que solvente la necesidades del equipo de data analytics, el cual consiste en un análisis profundo de datos; este gran volumen de datos (BigData) están alojados en la plataforma GCP (Google Cloud Platform) en donde se solicita que esta herramienta muestre estos datos estadísticos de una forma fácil e intuitiva, ya que las diferentes herramientas del mercado que solventan estas necesidades no son tan eficientes para cargar grandes volúmenes de datos, además de que su uso representa costos significativos para la compañía. De acuerdo a lo anteriormente mencionado, el equipo de tecnología implementa un plan de trabajo en donde 4 de sus integrantes desarrollan un software que cumpla con las necesidades mencionadas en donde se dividió en diferentes fases las cuales consisten: Identificación de IDEs, frameworks, lenguajes de programación, gestores de bases de datos, editores de código, creación de funciones, clases, métodos que permita la conexión de la base de datos al back-end y manipulación de datos mediante consultas SQL, creación de un repositorio en GitLab para el buen manejo de versionamiento y distribución del equipo de trabajo, identificación de datos, gráficas y métricas para la visualización estadística, creación de una interfaz gráfica fácil e intuitiva para el usuario y creación de clases y métodos que componen el front-end.

Palabras clave: BigData, framework, software, consultas SQL, programación front - end, programación back – end.

3.1. ABSTRACT:

According to the requirements of GrupoDot, needed an own framework or software that solves the needs of the data analytics team, which consists of a deep data analysis; this large volume

of data (BigData) is hosted on the GCP platform (Google Cloud Platform) where it is requested that this tool show these statistical data in an easy and intuitive way, since the different market tools that solve these needs are not so efficient to load large volumes of data, and their use represents significant costs for the company. According to this, the technology team implements a work plan where 4 of its members develop a software that meets the needs where it was divided into different phases which consist of: Identification of IDEs, frameworks, languages of programming, database managers, code editors, creation of functions, classes, methods that allow the connection of the database to the back-end and manipulation of data through SQL queries, creation of a repository in GitLab for good management of versioning and distribution of the work team, identification of data, graphs and metrics for statistical visualization, creation of an easy and intuitive graphical interface for the user and creation of classes and methods that make up the front-end.

Keywords: BigData, framework, software, SQL queries, front - end programming, back - end programming.

4. OBJETIVOS

4.1. Objetivo General

Desarrollar una herramienta (dot profiling) que facilite el estudio estadístico de grandes volúmenes de datos (Big Data), mediante la implementación de diferentes estrategias de desarrollo, desde el 1ero de julio al 31 de diciembre de 2020.

4.2. Objetivos Específicos

- Definir entornos de desarrollo, que permitan visualizar el desarrollo de la herramienta.
- Ejecutar métodos de desarrollo, que posibiliten la definición de la estructura de la herramienta.
- Implementar estrategias que contribuyan al control de versionamiento de la herramienta.
- Fijar un ambiente visual intuitivo para el usuario.

5. PROCESO Y DESARROLLO PRÁCTICA

De acuerdo a lo solicitado por la compañía, se ideó un plan de trabajo dividido en diferentes fases con el fin de llevar a cabo la implementación del servicio satisfactoriamente. De acuerdo a lo anterior, se implementó en este plan de trabajo reuniones diarias (daily), en donde cada integrante del equipo se le asignaron actividades específicas del proyecto y mostraba sus avances o dificultades para llevarlas al cabo. Semanalmente se realizaba una sesión con los líderes del área de desarrollo para mostrar avances del proyecto.
El proyecto se llevó a cabo en 6 fases, descritas a continuación:

Fase 1: Análisis y definición de componentes y herramientas utilizadas para la implementación del software.
Para llevar a cabo esta fase, se definieron y se analizaron las herramientas para dar continuidad con la implementación del software, dichas herramientas son las que se mencionan a continuación:

- Lenguaje de programación back-end: python 3
- Lenguaje de programación front-end: Dash y css
- Framework manejo de funciones estadísticas: plotly
- Framework manejo de datos: pandas
- Manejo de versionamiento: GitLab

- Base de datos : Bigquery
- Editor de código: Visual Studio Code
- Sistema operativo: Linux ubuntu

Fase 2: Estructuración del proyecto.
Una vez definido el enfoque del proyecto, se creó una instancia virtual de python y se instalaron todas las dependencias ya definidas anteriormente; seguido a esto se estructuró el repositorio de acuerdo al modelo MVC (Modelo Vista Controlador) y mediante el cliente de BigQuery se pudo obtener la información de la base de datos en donde se definió el nombre del proyecto en la cual está alojada nuestra base de datos y el dataset en donde se encuentra nuestra tabla y el nombre de la tabla.

Fase 3: Definición de gráficas y visualización de datos.
Se crearon diferentes clases y métodos para la manipulación de datos a través de pandas-profiling y consultas sql; retornamos estos datos ya sean como un lista o un conjunto de datos (Dataset) con el fin de poder estructurar ya sea una gráfica y/o métrica para su visualización en el Front.

Fase 4: Control de versionamiento.
Para el buen manejo y control de versionamiento se cargó el proyecto en gitlab mediante comandos de git en el sistema operativo Linux ubuntu, en donde en donde se especificó un archivo “requirements.txt” en el cual está definido todos los componentes que conforman el proyecto con sus respectivas versiones.
Además de manejar un buen versionamiento del repositorio se decidió documentar de manera específica cada componente, variable, método de nuestro proyecto para llevar un mejor control y contribuir a las buenas prácticas de desarrollo.

Fase 5: Aplicación de gráficas definidas en la fase 3.
Ya una vez obtenida la información de la tabla, se procedió a definir qué funciones, gráficas y datos estadísticos son de importancia para el equipo de data analytics los cuales se componen en:

- Overview de todos los campos de las tablas en los que cada una contiene estadísticas descriptivas como kurtosis, Meda, coeficiente de correlación, entre otros
- Histograma
- diagrama o cuadro de correlación
- Gráficas de muestreo de correlaciones
- Diagramas de muestreo multivariado
- un overview de la tabla de 20 filas

Fase 6: Implementación del componente visual.
Mediante estilos CSS se definió la tipología y diseño del aplicativo (Imagen 1) y mediante los componentes HTML del framework Dash se definió un esquema o estructura general de la página para posteriormente mostrar cada uno de los datos, variables y gráficas que anteriormente se definieron.

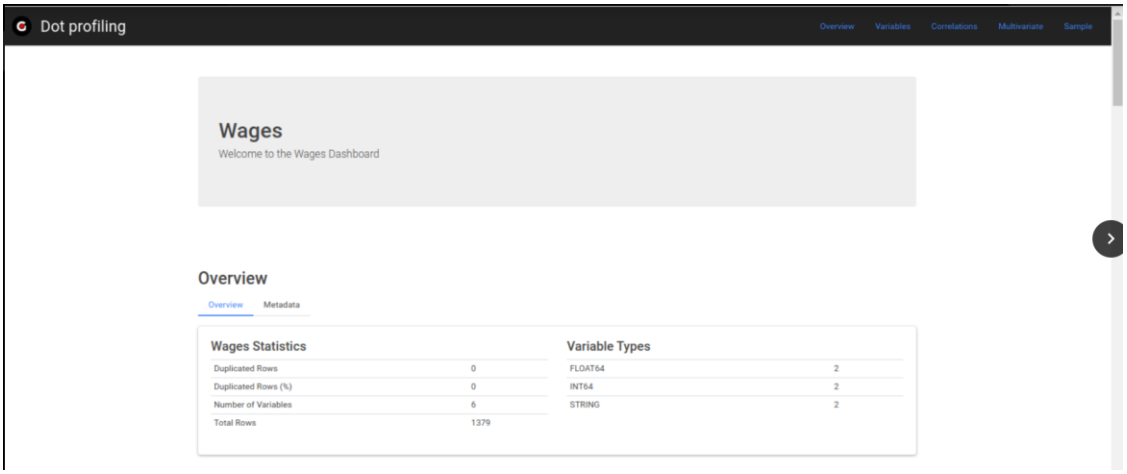

Imagen 1.

6. CONCLUSIONES

- Se logró desarrollar la herramienta de Dot Profiling, sin embargo, quedaron pendientes implementar mejoras de rendimiento.
- Los entornos de desarrollo definidos fueron phyton, dash, plotly, BigQuery, entre otros, los cuales permitieron visualizar de manera exacta, la estructura del software.
- Las gráficas y componentes estadísticos, jugaron un papel fundamental en el muestreo de los datos en la tabla consultada.
- La experiencia obtenida durante el desarrollo de la práctica laboral, fue muy enriquecedora, ya que pude ampliar los conocimientos obtenidos durante mi formación universitaria, y me mostró el énfasis que me gustaría seguir una vez termine mi pregrado.