Informe Técnico: 18 de Mayo 2024

Luis Fernando Suárez
Automatización para la Integración de Datos para IA
Universidad Central
Maestría en Analítica de Datos
Bogotá, Colombia
lsuarezp3@ucentral.edu.co

May 28, 2024

Contents

1 Noticias y Observatorio

1.1 Google lanza Project Gameface en Android

Google anunció durante su evento anual Google I/O el lanzamiento de Project Gameface para Android, una innovadora tecnología que permite a los usuarios controlar aplicaciones y juegos mediante movimientos faciales y expresiones. Este "mouse" avanzado representa un avance significativo en la accesibilidad y la interacción con dispositivos móviles.

1.2 Davivienda lanza 'El profe de finanzas', el asistente de IA para educación financiera

Davivienda presentó "El profe de finanzas", un asistente de inteligencia artificial diseñado para educación financiera. Este asistente, el primero de su tipo en Colombia y la región, utiliza tecnología de ChatGPT para ofrecer respuestas instantáneas y personalizadas que se adaptan a las necesidades específicas de cada usuario, mejorando así la educación financiera de sus clientes.

1.3 Lavado a cielo abierto (Openwashing) en la IA

El debate sobre si los modelos de inteligencia artificial deberían ser de código abierto sigue creciendo. Elon Musk demandó a OpenAI y a su director ejecutivo, Sam Altman, alegando que la empresa se desvió de su misión original de ser abierta. La administración Biden está evaluando los riesgos y beneficios asociados con los modelos de IA de código abierto, lo que resalta la importancia de la transparencia en el desarrollo de estas tecnologías.

2 Exposición sobre Apache Hive

En la clase se discutieron las siguientes soluciones relacionadas con Apache Hive:

2.1 Soluciones

Apache Hive es una herramienta de data warehousing que facilita el manejo de grandes volúmenes de datos almacenados en HDFS (Hadoop Distributed File System). Proporciona una interfaz similar a SQL para consultar datos almacenados, lo que permite a los usuarios realizar análisis complejos sin necesidad de escribir programas en MapReduce.

2.2 BigQuery

BigQuery es un almacén de datos completamente gestionado de Google que permite realizar análisis superrápidos en grandes conjuntos de datos. Ofrece una arquitectura sin servidores, donde los usuarios pueden ejecutar consultas

SQL estándar para analizar datos en terabytes y petabytes de escala de forma eficiente.

2.3 Dataprep

Google Dataprep es una herramienta basada en la nube para la preparación de datos. Permite a los usuarios limpiar, transformar y enriquecer datos visualmente, facilitando la creación de flujos de trabajo de datos antes de analizarlos con otras herramientas de Google Cloud.

2.4 Google Composer

Google Composer es un servicio de orquestación de flujos de trabajo gestionado basado en Apache Airflow. Facilita la creación, planificación y monitoreo de flujos de trabajo complejos en la nube, integrándose con otros servicios de Google Cloud.

2.5 Data Fusion

Google Cloud Data Fusion es un servicio de integración de datos completamente gestionado que permite crear y gestionar canalizaciones de datos visualmente. Ofrece conectores predefinidos y transformaciones para facilitar la integración de datos entre diferentes fuentes y destinos.

3 Metodologías Ágiles

Las metodologías ágiles son enfoques de gestión de proyectos que promueven la entrega incremental, la colaboración constante y la capacidad de respuesta al cambio. Estas metodologías son ampliamente utilizadas en el desarrollo de software y en proyectos de ciencia de datos debido a su enfoque flexible y adaptativo.

3.1 Scrum y Lean

Scrum y Lean son dos de las metodologías ágiles más comunes. Scrum se enfoca en la gestión de proyectos mediante sprints cortos y reuniones diarias para asegurar el progreso constante. Lean, por otro lado, busca maximizar el valor y minimizar el desperdicio en los procesos de desarrollo.

3.2 Sprint

Un sprint es un período de trabajo definido, generalmente de una a cuatro semanas, durante el cual se completa una cantidad específica de trabajo. Los sprints permiten a los equipos enfocarse en objetivos a corto plazo y realizar ajustes rápidos en función del feedback.

3.3 Canvas como Herramienta Visual

El Canvas es una herramienta visual que ayuda a los equipos a planificar y ejecutar proyectos. Proporciona una vista clara de los objetivos, tareas y progresos de cada sprint, facilitando la colaboración y la transparencia.

3.4 Metodologías en Ciencia de Datos

En ciencia de datos, las metodologías ágiles ayudan a manejar la naturaleza iterativa de los proyectos. El ciclo de vida de un proyecto de ciencia de datos incluye la recopilación de datos, el análisis exploratorio, el modelado, la validación y la implementación.

3.5 Metodologías de Gestión

- **KDD (Knowledge Discovery in Databases)**: Se enfoca en la extracción de conocimiento útil de grandes volúmenes de datos.
- **CRISP-DM (Cross-Industry Standard Process for Data Mining)**: Proporciona un proceso estandarizado para la minería de datos, desde la comprensión del negocio hasta la evaluación y el despliegue.
- **SEMMA (Sample, Explore, Modify, Model, Assess)**: Desarrollado por SAS, esta metodología se centra en el análisis de datos para crear modelos predictivos.

4 MLOps

4.1 Video de YouTube: ¿Qué es DevOps y CI/CD?

El video explica los conceptos de DevOps y CI/CD (Integración Continua y Entrega Continua), resaltando cómo estos enfoques mejoran la colaboración entre los equipos de desarrollo y operaciones. DevOps promueve una cultura de colaboración y automatización, mientras que CI/CD garantiza que los cambios en el código se integren y se desplieguen de manera continua y eficiente.

4.2 Roles en DevOps

- **Product Owner**: Responsable de definir la visión del producto y gestionar el backlog.
- **Scrum Master**: Facilita el proceso Scrum, asegura que el equipo siga las prácticas ágiles y elimina impedimentos.
- **Equipos**: Compuestos por desarrolladores, operadores y otros roles necesarios para entregar el producto.

5 Taller de DevOps

Para entregar en el aula virtual la próxima semana, se propuso crear un tablero con tres columnas: "Complete", "To Doing" y "To Do". Se diseñaron sprints semanales para organizar el trabajo del equipo. Se discutieron los retos y dificultades que podrían surgir, y se creó una ruta de cierre para asegurar el éxito del proyecto. La colaboración y la comunicación constante fueron clave para superar los desafíos y mantener el proyecto en curso.