Valledupar mayo 12 2025

Universidad Popular del cesar

Materia: Base De Datos Avanzada

Proyecto Final: Análisis exploratorio con herramientas de procesado masivo de datos (actividad grupal de máximo 4 estudiantes)

Caso práctico: Análisis en tiempo real del comportamiento de compra en supermercados

**Objetivo:** Monitorizar y analizar en tiempo real el comportamiento de compra de los clientes en las distintas sucursales de una cadena de supermercados para optimizar promociones, logística y gestión de inventario.

### Tecnologías involucradas e integración

1. **Kafka (Ingesta de datos en tiempo real): Visitas a la pagina**
   * Cada vez que un cliente escanea un producto en caja, se genera un evento (producto, precio, cantidad, timestamp, sucursal).
   * Estos eventos se envían como mensajes a **Kafka topics**, uno por tipo de evento (ventas, devoluciones, promociones aplicadas).
2. **Spark Streaming (Procesamiento en tiempo real):**
   * Una aplicación de **Spark Structured Streaming** consume los datos en tiempo real desde los topics de Kafka.
   * Realiza transformaciones como agregaciones por categoría de producto, detección de compras sospechosas, y conteo por hora/sucursal.
3. **Hive (Almacenamiento batch y consulta histórica):**
   * Spark también escribe datos procesados (por día o por hora) en tablas de Hive en formato Parquet.
   * Hive se usa para hacer análisis históricos con grandes volúmenes de datos de compra.
4. **Impala (Consultas interactivas rápidas sobre Hive):**
   * Los analistas usan **Impala** para hacer consultas ad hoc sobre los datos en Hive (por ejemplo: “¿Qué categoría de productos vende más los lunes por la mañana?”).
   * Impala permite respuestas más rápidas que Hive tradicional para exploración de datos.
5. **BigQuery (Da**ons**shboard global y análisis avanzado):**
   1. Una vez al día, se exportan datasets colidados a **BigQuery** (desde Hive o directamente desde Spark) para alimentar dashboards en **Looker Studio**, **Power Bi** o **Tableau**.
   * BigQuery también se usa para entrenar modelos de ML sobre compras, estacionalidad o fidelización: **OPCIONAL**

A través de esta actividad, los estudiantes podrán poner en práctica los conceptos estudiados en las clases de teoría acerca de las herramientas ofrecidas para el manejo de grandes volúmenes de datos por los proveedores de cloud computing (GCS), aplicándolos a un análisis exploratorio de un conjunto de datos. El objetivo es iniciarse en el manejo de un sistema de Data Warehouse de tipo SaaS, en concreto el software BigQuery de Google Cloud, muy propicio para realizar análisis sin necesidad de conocimientos de programación. Además, se introducirá el uso de tecnologías ampliamente utilizadas en entornos de big data, como **Hive** e **Impala** para la consulta eficiente sobre grandes volúmenes de datos almacenados, **Kafka** como plataforma de ingesta de datos en tiempo real, y **Spark** para el procesamiento distribuido tanto en tiempo real como en modo batch, permitiendo así una comprensión integral de un pipeline de análisis de datos moderno.

Por tanto, BigQuery se erige como una tecnología accesible a un público más amplio, que permite consultar datos almacenados en sistemas de almacenamiento distribuidos como Google Cloud Storage —caso que ocupará esta práctica—, Google Drive, así como importar datos para ser almacenados y manejados por el propio software de BigQuery. Además de esto, se pretende que los estudiantes desarrollen su creatividad y sus capacidades como científicos de datos. Para ello, se enfrentarán a un nuevo conjunto de datos que tendrán que procesar para, posteriormente, utilizar su propia iniciativa sin una guía exacta de qué se pide sobre este análisis. Pueden Usar herramientas para trabajar datos distibuidos como Hive, o Impala, herramientas de Inteligencia de negocios como dremio entre otras que sean necesarias para el desarrollo de la actividad.

Conjunto de dataset sales formado por los ficheros *Visitas\_lote\_01.csv Visitas\_lote\_02.csv y Lote\_Predicciones\_Ingresos.csv*. Trabajarás en la creación y la sustentación de un **Cuadro de Mando Integral (CMI)** que te ayudará a visualizar y analizar KPIs clave, lo que te permitirá proponer decisiones estratégicas para mejorar los ingresos y la experiencia del cliente.

Descripción de los Datos: El dataset que utilizaremos contiene información detallada sobre las visitas de los usuarios al sitio web, incluyendo datos sobre dispositivos, canales de marketing, y transacciones realizadas.

Objetivos de la Actividad

* Aplicar los conceptos análisis de datos en un contexto práctico.
* Desarrollar un Cuadro de Mando Integral (CMI) basado en datos reales para guiar decisiones estratégicas.
* Realizar un análisis de datos de comportamiento de usuarios y ventas para proponer decisiones estratégicas fundamentadas.
* Utilizar herramientas de Business Intelligence para visualizar y analizar KPIs clave.

**Ficheros de datos**

* *Visitas\_lote\_01.csv*: Contiene datos de visitas desde el 1 de agosto de 2016 hasta el 30 de abril de 2018.
* *Visitas\_lote\_02.csv*: Contiene datos de visitas desde el 1 de mayo de 2018 hasta el 15 de octubre de 2018.
* *Lote\_Predicciones\_Ingresos.csv*: Un archivo de ejemplo en el formato correcto para las predicciones de ingresos.

**Atributos principales de los dataset**

* ***fullVisitorId*** - Un identificador único para cada usuario de la Google Merchandise Store.
* ***channelGrouping*** - El canal a través del cual el usuario llegó a la tienda. date - La fecha en la que el usuario visitó la tienda.
* ***device*** - Las especificaciones del dispositivo utilizado para acceder a la tienda.
* ***geoNetwork*** - Esta sección contiene información sobre la geografía del usuario.
* ***socialEngagementType*** - Tipo de participación, ya sea "Socialmente comprometido" o "No socialmente comprometido".
* ***totals*** - Esta sección contiene valores agregados a lo largo de la sesión.
* ***trafficSource*** - Esta sección contiene información sobre la fuente de tráfico desde la cual se originó la sesión.
* ***visitId*** - Un identificador para esta sesión. Esto es parte del valor que normalmente se almacena como la cookie
* ***\_utmb***. Este identificador es único solo para el usuario. Para un ID completamente único, deberías usar una combinación de ***fullVisitorId*** y ***visitId***.
* **visitNumber** - El número de sesión para este usuario. Si esta es la primera sesión, entonces se establece en 1.
* ***visitStartTime*** - La marca de tiempo (expresada en tiempo POSIX).
* ***hits*** - Esta fila y los campos anidados se llenan para todos los tipos de hits. Proporciona un registro de todas las visitas a la página.
* ***customDimensions*** - Esta sección contiene cualquier dimensión personalizada a nivel de usuario o de sesión que se haya configurado para una sesión. Este es un campo repetido y tiene una entrada para cada dimensión que se configura.
* ***totals*** - Este conjunto de columnas incluye principalmente datos agregados de alto nivel.

***Fichero de predicción****:* logaritmo natural de la suma de todas las transacciones por usuario para el periodo de diciembre 1, 2018 a enero 31, 2019.

**Actividad por desarrollar**

**Paso 1: Preparación y Entendimiento del Dataset**

* ***Tarea***: Cargar los archivos en una herramienta análisis de datos (Excel, Google Sheets, Power BI, Tableau, Talend, Python o la de su preferencia).
* ***Descripción*: Desglosar columnas JSON (como *totals* y *hits*) para extraer información relevante, como ingresos y número de interacciones.**
* ***Instrucción***: Identifica y organiza los datos de *fullVisitorId*, *channelGrouping*, y *totals.transactionRevenue*.
* **Resultado Esperado 1**: Un resumen en formato tabular que muestre ingresos totales por canal de marketing (*channelGrouping*) y por dispositivo (*device*), identificando tendencias clave.

**Paso 2: Desarrollo de un Cuadro de Mando Integral (CMI)**

* ***Tarea***: **Crear un Cuadro de Mando Integral (CMI) en la herramienta de BI que incluya cuatro perspectivas: Financiera, Clientes, Procesos Internos, y Aprendizaje y Crecimiento.**
* ***Instrucción***: Define y calcula los siguientes KPIs:

*Financiera:*

* KPI: Ingresos totales por usuario.
* Fórmula:



*Clientes:*

* KPI: Tasa de conversión por canal.
* Fórmula:



*Procesos Internos:*

* KPI: Tiempo medio de visita.
* Fórmula:



*Aprendizaje y Crecimiento:*

* KPI: Tasa de retención de usuarios.
* Fórmula:



* **Complementa: investiga sobre métricas o KPI y en función de los datos analizados, incluye al menos tres (3) KPI más en cada perspectiva del informe.**
* **Resultado Esperado 2**: Un Dashboard interactivo en una herramienta de BI que visualice estos KPIs, permitiendo una evaluación rápida del rendimiento empresarial.

| **Categoría** | **KPIs para procesos internos** |
| --- | --- |
| **Atención al cliente** | * [Costo de adquisición de clientes](https://tudashboard.com/costo-de-adquisicion-de-clientes/) * Costo de [retención de clientes](https://tudashboard.com/retencion-de-clientes-a-traves-del-analisis-de-datos/) * Tasa de generación de clientes potenciales * [Tasa de conversión](https://tudashboard.com/tasa-de-conversion/) de clientes * Cuota de mercado (frente a la competencia) * # Número de clientes * [Valor de vida del cliente](https://tudashboard.com/valor-del-tiempo-de-vida-del-cliente/) |
| **Financieros** | * Ingresos * Beneficios (netos, brutos, operativos) * ROI |
| **Procesos internos** | * Nuevas innovaciones * Gestión del valor ganado (EVM) * Tiempo de ciclo para realizar un pedido * Tiempos de inactividad del proceso (cuánto tiempo se pierde debido a tiempos de inactividad, problemas de personal/técnicos, etc.) * Calidad (problemas, defectos, tickets de asistencia al cliente) |
| **Empleados** (Se centra en los activos intangibles de la organización, es decir, las habilidades/capacidades internas necesarias para apoyar los procesos internos) | * Feedback 360 * Opiniones de los empleados (sobre la gestión de proyectos, la calidad, el equipo, el entorno de trabajo, la satisfacción de los empleados, etc.) |

Esta proyecto final permitirá a los estudiantes de la UPC integrar el análisis de datos en la toma de decisiones estratégicas, utilizando herramientas de BI para visualizar y evaluar KPIs clave en tiempo real.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Criterios | Descripción | Puntuación máxima (puntos) | Peso  % |
| Criterio 1  Paso 1: Preparación y Entendimiento del Dataset | Se evalúa la capacidad para cargar, desglosar y organizar correctamente los datos en la herramienta seleccionada.  - 2 puntos: Datos organizados correctamente, identificación clara de las variables clave, y resumen preciso de tendencias clave en formato tabular.  - 1.5 punto: Datos cargados y desglosados adecuadamente, con algunas inconsistencias menores en la organización o el resumen de tendencias.  - 1 punto: Datos parcialmente desglosados o mal organizados, con tendencias clave no claramente identificadas.  - 0-0,5 puntos: No se logró cargar o desglosar adecuadamente los datos, o el resumen de tendencias es confuso o incorrecto. | 2 | 20 % |
| Criterio 2  Paso 2: Desarrollo de Cuadro de Mando Integral (CMI) | Se valora la creación de un CMI que integre las cuatro perspectivas, con KPIs definidos y visualizados correctamente.  - 3 puntos: CMI completo, con KPIs bien definidos y calculados correctamente (del enunciados + nuevos); visualizaciones claras y relevantes en el dashboard.  - 2 puntos: CMI desarrollado con la mayoría de los KPIs definidos y calculados correctamente; visualizaciones comprensibles con algunas mejoras posibles.  - 1 puntos: CMI incompleto o con KPIs mal definidos; visualizaciones poco claras o que no reflejan adecuadamente los datos.  - 0-0,5 puntos: CMI mal estructurado, con KPIs incorrectos o sin visualizaciones adecuadas. | 3 | 30 % |
| Criterio 3  Paso 3: Simulación de Decisiones Estratégicas | Se evalúa la justificación de las decisiones estratégicas basadas en los datos y KPIs del CMI.  - 3 puntos: Decisiones estratégicas claramente justificadas con análisis detallado basado en los KPIs; propuesta bien fundamentada y coherente.  - 2 puntos: Decisiones estratégicas justificadas con referencia a los KPIs; análisis razonable, pero con algunas áreas de mejora en la justificación.  - 1 puntos: Decisiones estratégicas poco claras o mal justificadas; análisis insuficiente de los KPIs.  - 0-0,5 puntos: Decisiones estratégicas no justificadas o incoherentes con el análisis de los KPIs. | 3 | 30 % |
| Criterio 4  Calidad Visual, Estilo y Formato | Se evalúa la presentación visual del informe y del Dashboard, incluyendo la coherencia en los estilos, claridad visual, **y el uso adecuado de colores y gráficos**.  - 2 puntos: Informe y dashboard con un diseño profesional, coherente, y visualmente atractivo; uso adecuado de colores, estilos y formatos que facilitan la comprensión.  - 1 puntos: Diseño adecuado con algunos problemas menores en la coherencia o claridad visual; colores y formatos generalmente apropiados.  - 0.5 puntos: Diseño básico con problemas evidentes en la coherencia, estilo, o claridad visual; uso de colores o formatos que dificultan la comprensión.  - 0 puntos: Diseño pobre, incoherente, o confuso; uso inadecuado de colores y formatos que comprometen la claridad del informe. | 2 | 20 % |
|  |  | **10** | **100 %** |