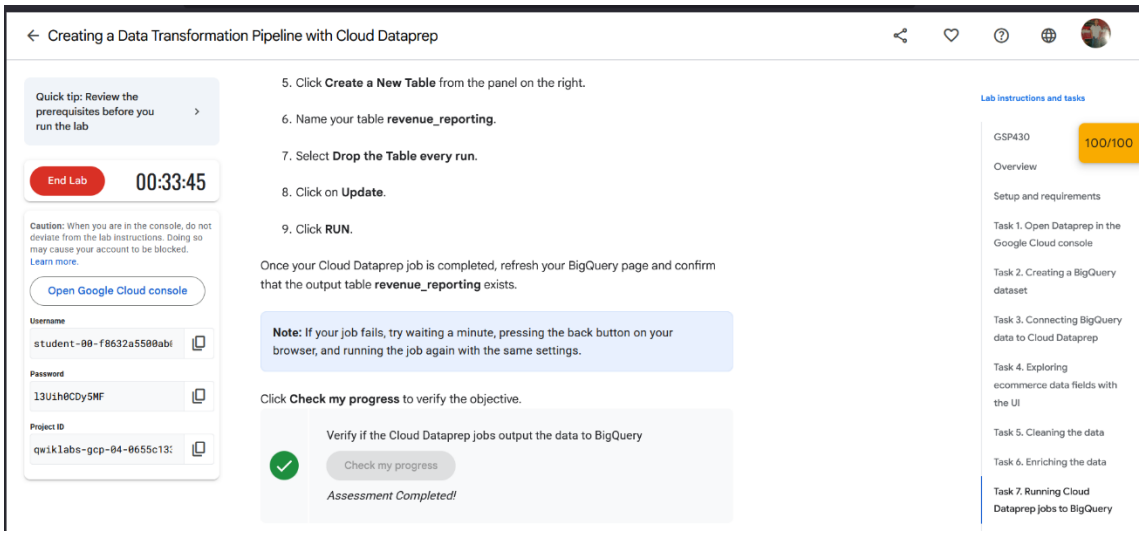
Google Skills Boost – Examen Final

Titulo: Creating a Data Transformation Pipeline with Cloud Dataprep



1. ¿Para qué se utiliza DataPrep?
   DataPrep es un servicio de Google Cloud Platform (GCP) que permite a los usuarios explorar, limpiar y preparar visualmente datos para análisis, informes y aprendizaje automáticos. Es una herramienta de autoservicio que simplifica el proceso de preparación de datos, ofreciendo una interfaz gráfica para transformar y limpiar datos sin necesidad de escribir código complejo.

2. ¿Qué cosas se pueden hacer con DataPrep?
   - DataPrep permite a los usuarios explorar los datos de manera visual, identificando patrones, anomalías y valores atípicos.
   - La herramienta facilita la limpieza de datos, permitiendo a los usuarios eliminar datos duplicados, corregir errores, manejar valores faltantes y estandarizar formatos.
   - DataPrep ofrece una amplia gama de transformaciones para modelar y enriquecer los datos, incluyendo la creación de nuevas columnas, la combinación de datos de diferentes fuentes y la aplicación de funciones de agregación.
   - DataPrep se integra con otros servicios de GCP como bigquery, cloud storage y Dataflow, facilitando el flujo de trabajo de datos desde la preparación hasta el análisis.

3. ¿Por qué otras herramientas lo podrías reemplazar? ¿Por qué?
   Las alternativas a DataPrep son las siguientes:
   - Gathr: Es una herramienta de integración de datos que simplifica la creación y gestión de pipeline de datos. Su interfaz de arrastrar y solar acelera la ingesta, transformación y carga de datos. Con

Gen AI Fabric de Gathr, también puedes aprovechar el aprendizaje automático y los servicios de IA generativa para realizar análisis de datos avanzados.

Características Principales

1. Rendimiento rápido: Gathr procesa datos con gran rapidez gracias a su base de Apache Spark. Puede procesar más de un millón de eventos por segundo tanto localmente como en nube.
2. Optimización de DevOps: Permite la monitorización y optimización continuas del rendimiento de DevOps. Esto le ayuda a detectar y resolver rápidamente problemas en su flujo de trabajo de datos.
3. Plataforma multifuncional: Gathr admite la ingesta de datos por lotes y streaming, la captura de datos modificados y el análisis basado en aprendizaje automático.

- **Airbyte:** Es una herramienta diseñada para optimizar flujos de trabajos de integración de datos. Ofrece una amplia biblioteca de más de 550 conectores prediseñados que permiten recopilar datos de múltiples fuentes y transferirlo al destino deseado. Para mejorar la calidad y consistencia de los datos, Airbyte permite realizar transformaciones mediante la integración con herramientas como dbt.

  Características principales:

  1. Flujos de trabajos con IA: Airbyte facilita la gestión de tus flujos de trabajo de IA. Te permite cargar datos no estructurados directamente en destino de almacenamiento vectorial como Pinecone o Milvus, así como en framework.
  2. **Cumplimiento normativo:** Airbyte cumple con las disposiciones de las regulaciones CCPA, RGPD e HIPAA para la seguridad de las operaciones de datos en todo el mundo. Esto garantiza protección de la privacidad y minimiza el riesgo de vulneraciones de la seguridad de los datos.

4. ¿Cuáles son los casos de usos comunes de DataPrep en GCP?

   4.1 Limpieza y transformación de datos:
   - 4.1.1 Eliminar duplicado, valores nulos o inconsistentes
   - 4.1.2 Normalización de formatos (fechas, números, textos)
   - 4.1.3 Estandarización de columnas (nombres, tipo de datos)
   - 4.1.4 Conversión de tipo de datos

   4.2 Preparación de datos para análisis
   - 4.2.1 Unir múltiples fuentes de datos (CSV, Big query, Cloud Storage)

4.2.2 Filtrado y segmentación de datos relevantes.

4.3 Análisis exploratorio y profiling de datos

4.3.1 Visualización de distribuciones, outliers, valores faltantes.

4.3.2 Estadísticas descriptivas automáticas sobre datasets cargados.

4.3.3 Detección de anomalías antes de usar los datos.

5. ¿Cómo se cargan los datos en DataPrep en GCP?

   Pasos para cargar datos:

   1. Acceder a Google cloud platform
   2. Menú de navegación→Ver todos los productos→Análisis→Alteryx desing cloud
   3. Crear un nuevo flujo (Create new flow)
   4. Colocarle un nombre descriptivo al flujo de trabajo, y una descripción (opcional).
   5. Add datasets→Import datasets, luego puedes elegir la fuente de datos.
   6. Selecciona el origen de datos (Archivo local, GCS, Big query).
   7. Define formato y opciones de lectura
   8. Una vez listo, se hace click en Import and Add to Flow.

6. ¿Qué tipo de datos se pueden preparar en DataPrep GCP?

   Tipos de datos principales en DataPrep

   - Enteros
   - Flotantes
   - Decimales
   - Booleanos
   - Cadena de Texto (string)
   - Fecha y Hora
   - Estructuras de Arreglos
   - Estructuras de objetos.

7. ¿Qué pasos se pueden seguir para limpiar y transformar datos en DataPrep en GCP?

   7.1 Importar datos a DataPrep

   7.2 Perfilado automático de datos, dataprep genera un data profile automático que muestra: Distribución de valores, Valores faltantes o nulos, Tipos de datos, Outliers, Duplicados, etc.

   7.3 Aplicar transformaciones desde la interfaz grafica haciendo click en las columnas que desea aplicar algunas transformaciones, ejemplo:

7.3.1 Eliminar datos inválidos o nulos.
7.3.2 Eliminar duplicados, usando la función Remove duplicate rows.
7.3.3 Corregir tipos de datos
7.3.4 Reemplazar valores, usando la función Find and replace.

7.2 Transformar datos

7.2.1 Crear columnas nuevas a partir de expresiones, cálculos o condiciones.

8. ¿Cómo se pueden automatizar tareas de preparación de datos en GCP?

8.1 Automatización Interna (Jobs Schedule)
Una vez realizado la limpieza y transformación de datos, se puede definir una programación de ejecución personalizada, agregando la frecuencia de ejecución, hora de ejecución zona horaria.

8.2 Automatización con cloud composer
Existe la posibilidad de disparar flujos de trabajos de DataPrep por medio de API, y se pueden orquestar por un DAG.

8.3 Utilizando combinación de cloud function o cloud scheduler y API

9. ¿Qué tipos de visualizaciones se pueden crear en DataPrep en GCP?

Los tipos de visualizaciones que se pueden crear en dataprep son los siguientes:

9.1 Visualizaciones de distribución por columnas
9.1.1 Histogramas (para valores numéricos)
9.1.2 Barras (para valores categóricos)
9.1.3 Distribución de frecuencia
9.1.4 Conteo de valores únicos
9.1.5 Porcentajes de valores nulos o faltantes
9.1.6 Outliers detectados visualmente.
9.2 Perfil de calidad de datos
9.2.1 Verdes: Valores válidos.
9.2.2 Azul: Valores únicos
9.2.3 Rojo: Valores inválidos
9.2.4 Gris: Valores faltantes o vacíos
9.3 Estadística resumida por columna
9.3.1 Mínimo, Máximo, Promedio y Mediana
9.3.2 Longitud promedio de Texto
9.3.3 Cantidad de valores nulos
9.3.4 Moda

9.4 Gráficos automáticos en la vista de perfil

9.4.1 Histogramas de frecuencias

9.4.2 Gráficos de líneas de tiempo (para fechas)

9.4.3 Diagramas de barras (para categorías)

9.4.4 Boxplots para detectar outliers

10. ¿Cómo se puede garantizar la calidad de datos en DataPrep en GCP?

10.1 Perfilado automático de datos

Cuando se carga un datasets, se genera un perfilado de datos de forma automática que incluye:

- Distribución de valores
- Conteo de nulos
- Tipos de datos detectados
- Outliers
- Duplicados

10.2 Validación visual con Data Quality Bar

10.3 Aplicar regla de limpieza de datos.

10.4 Validación con expresiones personalizadas.

Arquitectura

El gerente de Analítica te pide realizar una arquitectura hecha en GCP que contemple el uso de esta herramienta ya que le parece muy fácil de usar y una interfaz visual que ayuda a sus desarrolladores ya que no necesitan conocer ningún lenguaje de desarrollo. Esta arquitectura debería contemplar las siguiente etapas:

Ingesta: datos parquet almacenados en un bucket de S3 y datos de una aplicación que guarda sus datos en Cloud SQL.

Procesamiento: filtrar, limpiar y procesar datos provenientes de estas fuentes

Almacenar: almacenar los datos procesados en BigQuery

BI: herramientas para visualizar la información almacenada en el Data Warehouse

ML: Herramienta para construir un modelo de regresión lineal con la información almacenada en el Data Warehouse.