CREDEMHACK

Specifiche Tecniche per la Sottomissione delle Soluzioni

Per garantire un processo di valutazione equo e coerente, tutte le sottomissioni vengono testate tramite una pipeline automatizzata. È fondamentale aderire strettamente a queste specifiche per assicurare che la vostra soluzione possa essere valutata correttamente.

1. Obiettivo della Sfida

L'obiettivo è sviluppare un modello in grado di processare un insieme di documenti per:

- **Classificare** ogni documento.
- **Estrarre** da ogni documento il Nome, Cognome e la Data.
- **Arricchire** i dati usando i file di anagrafica forniti.
- **Produrre** un file di output finale (DocumentsOfRecord.dat) con un formato specifico.

2. Il File di Output: DocumentsOfRecord.dat

Questo è il *deliverable* principale che verrà analizzato per la valutazione. Deve rispettare le seguenti regole:

- **Encoding**: UTF-8.
- **Delimitatore di campo**: Pipe (|).
- **Struttura**: Il file **deve** contenere due sezioni, separate da due righe di intestazione METADATA esatte:

Sezione 1: DocumentsOfRecord

Intestazione:
METADATA|DocumentsOfRecord|PersonNumber|DocumentType|Country|DocumentCode|DocumentName|DateFrom|DateTo|SourceSystemOwner|SourceSystemId

Dettaglio Campi:

Campo	Valore	Note
METADATA	MERGE	Valore fisso.
DocumentsOfRecord	DocumentsOfRecord	Valore fisso.
PersonNumber	*Es*. 007681926	Person Number del dipendente, da recuperare dal file Elenco Personale
DocumentType	*Es*. Polizza sanitaria	Valore basato sulla classificazione, dalla colonna Cluster del file Cluster Docs.
Country	*Es*. Italy	Valore basato sulla classificazione, dalla colonna Country Document HCM del file Cluster Docs.
DocumentCode		ID univoco composto da: PersonNumber_DateFrom_DocumentType. DateFrom in formato YYYYMMDD.
DocumentName	*Es*. MARIO ROSSI	Nome e Cognome estratti dal documento, in maiuscolo.
DateFrom	Es. 2025/04/26	Data estratta dal documento con formato YYYY/MM/DD.
DateTo	*Nullo*	Lasciare vuoto.

SourceSystemOwner	PEOPLE	Valore fisso.
SourceSystemId		Stesso valore del DocumentCode.

Sezione 2: DocumentAttachment

Intestazione:
METADATA|DocumentAttachment|PersonNumber|DocumentType|Country|DocumentCode|DataTypeCode|URLorTextorFileName|Title|File|SourceSystemOwner|SourceSystemId 22

Dettaglio Campi:

- I campi da METADATA a DocumentCode e da SourceSystemOwner a SourceSystemId sono identici a quelli della sezione precedente.

Campo	**Valore**	**Note**
DataTypeCode	FILE	Valore fisso.
URLorTextorFileName	Es. NOME_FILE.pdf	Nome originale del file processato.
Title	Es. NOME_FILE.pdf	Nome originale del file processato.
File	Es. NOME_FILE.pdf	Nome originale del file processato.

Esempio

Nome file	**Cluster**	**Nominativo**	**Data**
MARIO_ROSSI_2023-05-10_Polizza sanitaria.pdf	Polizza sanitaria	Rossi Mario	2023/05/10

GINEVRA_BIANCHI_2022-11-20_ Polizza sanitaria.pdf	Polizza sanitaria	Bianchi Ginevra	2022/11/20

DocumentsOfRecord.dat

```
METADATA|DocumentsOfRecord|PersonNumber|DocumentType|Country|DocumentCode|DocumentName
|DateFrom|DateTo|SourceSystemOwner|SourceSystemId
MERGE|DocumentsOfRecord|999931077|Polizza sanitaria|Italy|999931077_20230510_Polizza
sanitaria|Mario Rossi|2023/05/10||PEOPLE|999931077_20230510_Polizza sanitaria

MERGE|DocumentsOfRecord|999930919|Polizza sanitaria|Italy|999930919_20221120_Polizza
sanitaria|Ginevra Bianchi|2022/11/20||PEOPLE|999930919_20221120_Polizza sanitaria

METADATA|DocumentAttachment|PersonNumber|DocumentType|Country|DocumentCode|DataTypeCod
e|URLorTextorFileName|Title|File|SourceSystemOwner|SourceSystemId

MERGE|DocumentAttachment|999931077|Polizza sanitaria|Italy|999931077_20230510_Polizza
sanitaria|FILE|MARIO_ROSSI_2023-05-10_Polizza
sanitaria.pdf|MARIO_ROSSI_2023-05-10_Polizza
sanitaria.pdf|MARIO_ROSSI_2023-05-10_Polizza
sanitaria.pdf|PEOPLE|999931077_20230510_Polizza sanitaria

MERGE|DocumentAttachment|999930919|Polizza sanitaria|Italy|999930919_20221120_Polizza
sanitaria|FILE|GINEVRA_BIANCHI_2022-11-20_Polizza
sanitaria.pdf|GINEVRA_BIANCHI_2022-11-20_Polizza
sanitaria.pdf|GINEVRA_BIANCHI_2022-11-20_Polizza
sanitaria.pdf|PEOPLE|999930919_20221120_Polizza sanitaria
```

---

3. Casi Particolari da Gestire

La vostra soluzione deve gestire correttamente i seguenti scenari:

- **Se un dipendente non viene trovato** nell'anagrafica Elenco Personale, i campi devono essere popolati come segue:
  - **PersonNumber:** Nessun dipendente
  - **DocumentType:** SCARTATO
  - **Country:** *campo vuoto*
  - **DocumentName:** Nessun dipendente
- **Se un campo tra Nome, Cognome o Data non è estraibile dai documenti**, usare i segnaposto rispettivi NONAME, NOLASTNAME, NODATE. In questo caso bisogna attenersi anche al punto precedente e quindi non estrarre (anche se presenti!) i dati dall'Elenco Personale.

Fate attenzione a questi casi particolari perché potreste ritrovarvi a dover gestire più di una casistica (e potrebbe fare la differenza!)

---

4. Requisiti Tecnici del Container

A. Immagine Docker

- **Autonoma**: L'immagine deve essere *self-contained*, includendo codice, requirements.txt, e tutti i file statici necessari (es. i CSV di anagrafica).
- **Eseguibile**: Deve essere un processo non interattivo che parte, esegue l'elaborazione e termina automaticamente.
- **Caricata su Artifact Registry**: L'immagine deve essere caricata correttamente nel repository condiviso.

B. Comportamento all'Esecuzione

Il nostro sistema avvierà il vostro container fornendo tre variabili d'ambiente che il vostro codice deve leggere e utilizzare:

- **INPUT_BUCKET**: Il bucket GCS da cui leggere i file di test.
- **OUTPUT_BUCKET**: Il bucket GCS su cui scrivere l'output finale.
- **RUN_ID**: L'ID univoco dell'esecuzione, da usare come nome della cartella di output.

```
# Importate subito le variabili nel vostro main.py che orchestra la soluzione
import os
run_id = os.environ.get("RUN_ID")
input_bucket = os.environ.get("INPUT_BUCKET")
output_bucket = os.environ.get("OUTPUT_BUCKET")

# Esempi di come dovete testare i documenti e salvare la soluzionne
input_bucket = storage_client.bucket(input_bucket_name)
...
print(f"Caricamento di 'solution.zip' su gs://{output_bucket_name}/{run_id}/")
```

C. Output Finale

Il vostro container deve produrre un solo file:

- Nome: solution.zip.
- Contenuto: Deve contenere il file DocumentsOfRecord.dat e una cartella BlobFiles con i documenti originali elaborati.
- Posizione di Upload: Il file deve essere caricato nel percorso gs://[OUTPUT_BUCKET]/[RUN_ID]/solution.zip.

---

5. Processo di Sottomissione

Build & Push dell'Immagine: Eseguite il build della vostra immagine e caricatela su Artifact Registry usando il comando fornito. Assicuratevi di sostituire **nometeam**.

Il progetto di riferimento è `credemhack-gcp`

```
gcloud builds submit --tag
europe-west1-docker.pkg.dev/credemhack-gcp/hackathon-solutions/credemh
ack-[nometeam]:v1
```

Avvio della Valutazione: Pubblicate un messaggio sul topic Pub/Sub new-submission-topic, usando l'URI esatto dell'immagine che volete far valutare.

```
gcloud pubsub topics publish new-submission-topic
--message='{"tag":"europe-west1-docker.pkg.dev/credemhack-gcp/hackatho
n-solutions/credemhack-[nometeam]:v1"}'
```

Se la sottomissione è andata a buon fine, troverete il vostro punteggio nella leaderboard.

In bocca al lupo!

Il team

CREDEMHACK