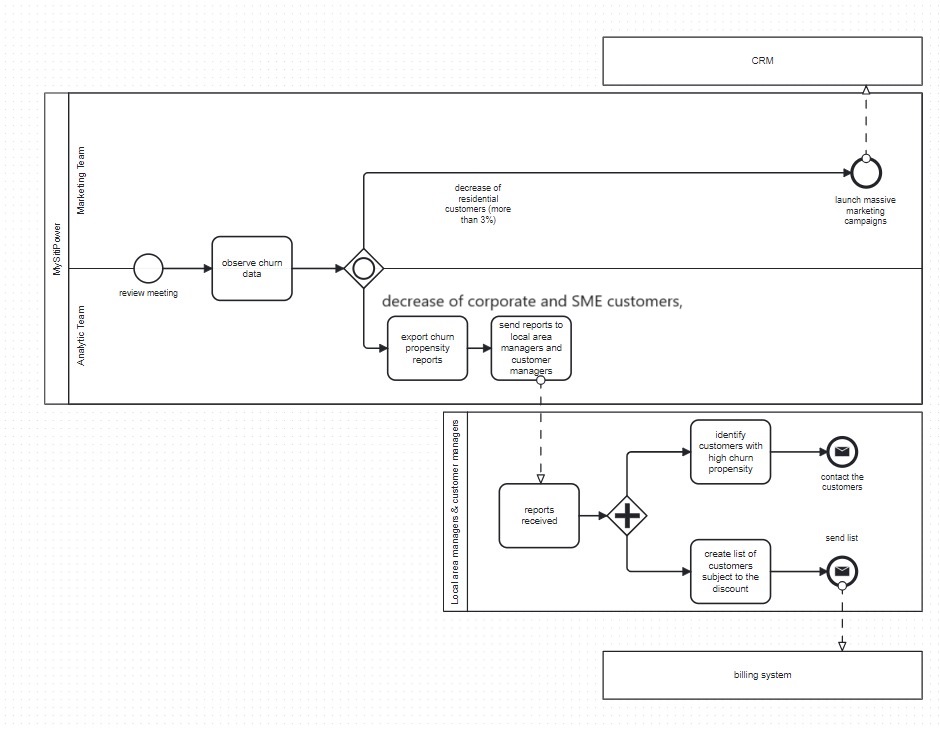
**PARTE 2 PROGETTO: SOLUZIONE**

**Soluzione della Richiesta 1: BPMN**

*Every 2 weeks, MystiPower Marketing Team organizes a recurring review meeting together with the Analytic Team to observe churn data across different customer segments: corporate, SME and residential customers. If churn data shows a significant decrease of residential customers compared to previous Information Systems Design and Big Data Exam date: 2024-02-28*

*period (more than 3%), the Marketing team launches massive marketing campaigns on the CRM system, if data shows a significant decrease (more than 3%) in corporate and SME customers, Analytic Team is required to export churn propensity reports to be sent to local area managers and customer managers to proactively contact the customers with high churn propensity, simultaneously, a temporary 20% discount for this customers is automatically applied by means of a direct integration between the data platform and the billing system which receives the list of customers subject to the discount.*



**Soluzione della Richiesta 2: Architettura Google Cloud**

Come servizio di cloud pubblico ho scelto Google Cloud che offre diversi vantaggi distintivi rispetto ad altri fornitori di servizi cloud pubblici, come ad esempio:

* **Ecosistema di servizi avanzati:** Google Cloud offre un ecosistema completo di servizi avanzati, tra cui intelligenza artificiale e machine learning con Google Cloud AI Platform, analisi dati con BigQuery, e servizi di elaborazione e storage scalabili come Compute Engine e Google Cloud Storage. Questa vasta gamma di servizi consente agli sviluppatori di implementare soluzioni complesse senza dover dipendere da più fornitori.
* **Sicurezza e conformità**: Google Cloud fornisce un solido framework di sicurezza con funzionalità avanzate di crittografia, accesso basato sui ruoli e monitoraggio degli eventi. Inoltre, Google Cloud è conforme a numerosi standard di sicurezza e privacy, come GDPR, HIPAA e ISO 27001, garantendo la conformità normativa per una vasta gamma di settori e mercati.
* **Performance e scalabilità**: Google Cloud offre una rete globale ad alta velocità con una latenza ridotta grazie alla sua infrastruttura di rete distribuita. Questo si traduce in prestazioni elevate e tempi di risposta veloci per le applicazioni ospitate su Google Cloud. Inoltre, la capacità di scalabilità elastica consente alle applicazioni di gestire carichi di lavoro in rapida crescita senza compromettere le prestazioni.
* **Innovazione continua**: Google è all'avanguardia dell'innovazione tecnologica e investe pesantemente in ricerca e sviluppo per migliorare costantemente i suoi servizi cloud. Questo si traduce in un flusso continuo di nuove funzionalità e servizi, consentendo agli utenti di sfruttare le ultime tecnologie per mantenere la loro competitività sul mercato.
* **Prezzi competitivi**: Google Cloud offre una struttura di prezzi competitiva con opzioni flessibili di fatturazione a consumo. Inoltre, offre sconti e offerte speciali per carichi di lavoro a lungo termine e per i clienti impegnati ad utilizzare i suoi servizi su base regolare.

Alcuni dei possibili svantaggi di Google Cloud includono la curva di apprendimento per utilizzare al meglio tutte le sue funzionalità, la mancanza di alcuni servizi specializzati presenti in altri cloud provider e la necessità di integrarsi con altri servizi e piattaforme esterne. Tuttavia, considerando i suoi punti di forza e la sua crescente adozione nel settore, Google Cloud rimane una scelta altamente competitiva e attraente per molte aziende e organizzazioni.

Inoltre, Google Cloud include molte funzionalità di sicurezza, privacy e conformità. Ecco alcuni esempi specifici:

* **Crittografia:** Google Cloud crittografa automaticamente i dati in transito e a riposo.
* **Gestione delle chiavi:** Google Cloud Key Management Service (KMS) e Cloud HSM permettono di gestire facilmente le chiavi di crittografia.
* **Controllo degli accessi:** Google Cloud Identity & Access Management (IAM) permette di controllare chi ha accesso alle tue risorse cloud e che tipo di accesso hanno.
* **Autenticazione a due fattori (2FA):** Google offre l’autenticazione a due fattori per aggiungere un ulteriore livello di sicurezza.
* **Logging e monitoraggio:** Google Cloud Logging e Monitoring permettono di tenere traccia delle attività nel tuo ambiente cloud.
* **Conformità:** Google Cloud è conforme a standard globali come GDPR, HIPAA, ISO, tra gli altri.
* **Backup e ripristino:** Google Cloud offre servizi come Persistent Disk Snapshots per il backup dei dati.



L’Architettura proposta è basata sulle componenti Google Cloud:

Immagine che contiene testo, diagramma, Carattere, linea

Descrizione generata automaticamente

1. Dal produttore esterno, i dati grezzi passano effettivamente al Digital Data Platform tramite il blocco **Data Landing**. La componente Batch Data Transfer (Cloud Storage) è utilizzata per trasferire i dati giornalieri dal CRM al Data Lakehouse, ma anche per i dati in tempo reale del’API.
2. I dati trasferiti al Data Landing passano poi al **Data Manipulation** tramite le ETL Data Pipelines (Google Dataflow). Questo processo di manipolazione dei dati è gestito e coordinato dal Workflow Orchestration (Cloud Composer).
3. Successivamente, i dati manipolati vengono memorizzati nel blocco **Data Lake**, attraversando i vari livelli:

* **L0 (Raw Storage)**: I dati grezzi vengono conservati senza alcuna trasformazione utilizzando la componente Raw Data Persistence (Cloud Storage).

1. **L1 (Curated)**: I dati elaborati e strutturati vengono memorizzati in un database relazionale utilizzando Google Cloud SQL.
2. **L2:** I dati aggregati e pre-elaborati vengono memorizzati in un ambiente analitico ad alte prestazioni utilizzando Google BigQuery.
3. I dati manipolati sono quindi disponibili per il **Data Consumption** tramite il Data Manipulation.
4. Il **Data Consumption** può accedere ai dati del Data Lake L2 per l'analisi tramite l'AI Platform Engine.
5. Il consumer Esterno fa una richiesta di visualizzazione dei dati finali al **Data Exposure/Visualization**, che può prenderla in carico tramite il componente Custom API Development (Google Endpoints).
6. Il **Data Exposure** gestisce questa richiesta tramite un collegamento con il **Data Consumption**, che inoltrerà i dati nel passaggio successivo.
7. Adesso i risultati dell'analisi sono resi disponibili dal **Data Consumption** al **Data Exposure/Visualization**. I dati vengono prima visualizzati e analizzati tramite il Data Visualization Tool (Google Data Studio)
8. Poi essi vengono elaborati ulteriormente nell'Enterprise BI Tool (Google BigQuery) per un'analisi più dettagliata e approfondita.
9. Infine, i dati vengono forniti all'**External Consumer** per l'utilizzo esterno e la visualizzazione finale.

Descrizione della Architettura di Cloud proposta, basata su Google Cloud, e delle sue componenti:

**1. Data Landing**:

* **Batch data transfer**: Questo componente è utilizzato per trasferire i dati giornalieri dal sistema CRM, dal sistema dei prezzi e anche dati in tempo reale da API al data lakehouse.

**2. Data Manipulation**:

* **ETL Data Pipelines**: Questo componente è fondamentale per elaborare e preparare i dati grezzi provenienti dal sistema CRM e dal sistema dei prezzi per l'archiviazione e l'analisi successiva.
* **Workflow Orchestration**: Utilizzato per coordinare il flusso di lavoro tra diversi componenti e servizi dell'architettura.

**3. Data Lakehouse (con L0, L1, L2)**:

* **Raw data persistence (L0)**: Utilizzato per conservare i dati grezzi provenienti dal sistema CRM e dal sistema dei prezzi senza alcuna trasformazione.
* **Relational database and transactional engine (L1)**: Questo livello è utilizzato per memorizzare i dati elaborati e strutturati per consentire l'analisi e la manipolazione dei dati.
* **Analytic engine (L2)**: Utilizzato per memorizzare i dati aggregati e pre-elaborati, pronti per l'analisi avanzata e l'elaborazione da parte dei sistemi di business intelligence.

**4 Data Consumption:**

* **Custom ML Model**: Utilizzato per analizzare i dati e fornire previsioni o modelli personalizzati in base alle specifiche esigenze del cliente.
* **AI Platform Engine**: Utilizzato per implementare algoritmi di intelligenza artificiale e machine learning per analizzare e trarre insight dai dati.

**5. Data Exposure/Visualization:**

* **Data Visualization Tool**: Questo componente consente di creare visualizzazioni interattive dei dati per l'analisi e la presentazione. È utile per comunicare i risultati dell'analisi in modo chiaro e intuitivo.
* **Enterprise BI Tool**: Fornisce strumenti per l'analisi aziendale e la creazione di report basati sui dati. È importante per ottenere insight approfonditi e supportare le decisioni aziendali.
* **Custom API development**: Potrebbe essere necessario sviluppare API personalizzate per consentire il trasferimento di dati in tempo reale dal sistema CRM.

**Motivazione delle scelte, con corrispettivi componenti in Google Cloud:**

* **Batch data transfer**: Scelto per trasferire grandi volumi di dati in modo efficiente dal sistema CRM e dal sistema dei prezzi al data lakehouse, garantendo una gestione ottimizzata dei carichi di lavoro giornalieri. 🡪 **Google Cloud Storage**
* **ETL Data Pipelines**: Essenziale per elaborare e preparare i dati grezzi per l'analisi successiva, garantendo la qualità e la coerenza dei dati all'interno del data lakehouse. 🡪 **Google Cloud Dataflow**
* **Workflow Orchestration**: Utilizzato per coordinare il flusso di lavoro tra diversi componenti e servizi dell'architettura, garantendo una gestione efficiente delle attività di manipolazione dei dati. 🡪 **Google Cloud Composer**
* **Raw data persistence (L0):** Scelto per conservare i dati grezzi provenienti dal sistema CRM e dal sistema dei prezzi senza alcuna trasformazione, preservando l'integrità dei dati originali. 🡪 **Google Cloud Storage**
* **Relational database and transactional engine (L1)**: Questo livello è utilizzato per memorizzare i dati elaborati e strutturati per consentire l'analisi e la manipolazione dei dati, fornendo una base solida per l'interrogazione dei dati. 🡪 **Google Cloud SQL**
* **Analytic engine (L2):** Scelto per memorizzare i dati aggregati e pre-elaborati, pronti per l'analisi avanzata e l'elaborazione da parte dei sistemi di business intelligence, facilitando la generazione di insight significativi. 🡪 **Google BigQuery**
* **Custom ML Model**: Utilizzato per analizzare i dati e fornire previsioni o modelli personalizzati in base alle specifiche esigenze del cliente, consentendo una maggiore personalizzazione delle soluzioni di analisi dei dati. 🡪 **Google Cloud AI Platform**
* **AI Platform Engine**: Scelto per implementare algoritmi di intelligenza artificiale e machine learning per analizzare e trarre insight dai dati, consentendo una comprensione più approfondita dei modelli dei dati e dei pattern. 🡪 **Google Cloud AI Platform**
* **Data Visualization** Tool: Questo componente consente di creare visualizzazioni interattive dei dati per l'analisi e la presentazione, facilitando la comprensione dei dati da parte degli utenti finali. 🡪 **Google Data Studio**
* **Enterprise BI Tool**: Fornisce strumenti per l'analisi aziendale e la creazione di report basati sui dati, consentendo una generazione rapida e accurata di report aziendali e analisi approfondite. 🡪 **Google BigQuery**
* **Custom API development**: Necessario per integrare il sistema CRM e consentire il trasferimento in tempo reale degli eventi dei clienti, garantendo una comunicazione fluida tra i diversi sistemi. 🡪 **Google Cloud Endpoints**

**Calcolo dei Costi associati per Google Cloud**:

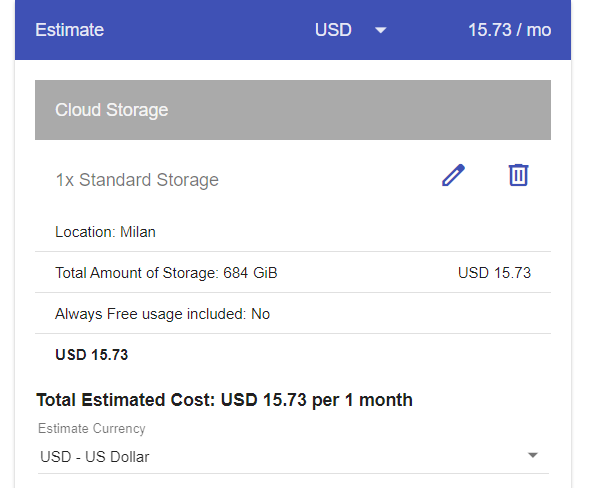
**Data Landing**:

* Batch data transfer (**Google Cloud Storage**): Per calcolare la quantità totale di storage necessaria, bisogna sommare il volume di dati generato ogni giorno dai clienti e dagli eventi, più la dimensione del file CSV dei dati sui prezzi.

**Dati di registrazione del cliente:** 450.000 clienti \* 50 KB/cliente = 22.500.000 KB = 22,5 GB

**Eventi dei clienti in tempo reale:** 5000 eventi/giorno \* 40 KB/evento = 200.000 KB = 0,2 GB

**Dati sui prezzi:** 100 MB = 0,1 GB 🡪 Sommando tutto, abbiamo un totale di circa 22,8 GB di dati al giorno, quindi al mese sarebbero: 22.8GB \* 30giorni = 684 GB 🡪 il Total Estimated Cost: USD 15.73 per 1 month



**Data Manipulation**:

* ETL Data Pipelines (**Google Cloud Dataflow**):
  + **Region**: “Milan (europe-west3)”
  + **Job type**: “Batch” - Adatto per l’elaborazione di file CSV su base giornaliera.
  + **Shuffle enabled**: selezionato – per eseguire operazioni di join o group by sui dati.
    - **Data Processed**:
      * Dimensione del file CSV dei dati sui prezzi: 100 MB al giorno.
      * Dimensione totale dei dati di registrazione del cliente: 450.000 clienti \* 50 KB/cliente.
      * Dimensione totale degli eventi dei clienti in tempo reale: 5000 eventi/giorno \* 40 KB/evento.
      * Prima di tutto, convertiamo tutto in KB:
        + 100 MB al giorno sono 100 \* 1024 = 102,400 KB.
        + 450.000 clienti \* 50 KB/cliente sono 22,500,000 KB.
        + 5000 eventi/giorno \* 40 KB/evento sono 200,000 KB.
      * Ora sommiamo tutto: 102,400 KB (file CSV) + 22,500,000 KB (dati di registrazione del cliente) + 200,000 KB (eventi dei clienti in tempo reale) = 22,802,400 KB.
      * Infine, convertiamo in GB: 22,802,400 KB / 1024 (per convertire in MB) / 1024 (per convertire in GB) = circa 21.74 GB.
      * Quindi, la quantità totale di “Data Processed” dovrebbe essere di circa 21.74 GB al giorno.
  + **Hours the job runs per month**: 30 - (1 ora al giorno per 30 giorni).
  + **Number of worker nodes used by job**: 2 - Iniziare con 2 è un buon punto di partenza.
  + **Worker node instance type**: “n1-standard-1 (vCPUs: 1, RAM: 3.75 GB)”
  + **System… PD**: 200 - Questo valore rappresenta la quantità di storage persistente disponibile per i nodi del worker.

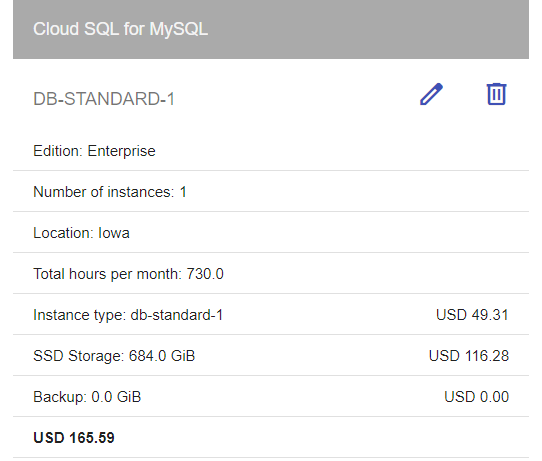
Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

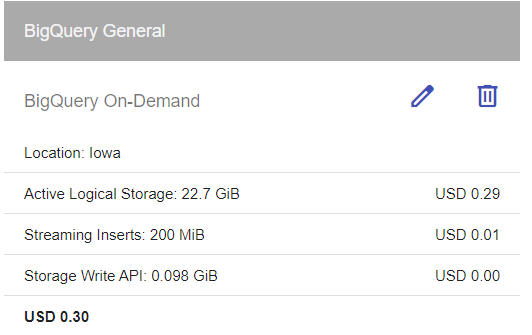
* Workflow Orchestration (**Google Cloud Composer**): Il numero di job e la complessità del flusso di lavoro dipenderanno dalla logica di elaborazione dei dati e dalla frequenza degli aggiornamenti.

**Data Lakehouse (con L0, L1, L2):**

* Raw data persistence (L0) (**Google Cloud Storage**): Lo spazio di archiviazione è già calcolato per la componente di Data Landing.
* Relational database and transactional engine (L1) (**Google Cloud SQL**): Il costo dipenderà dalla configurazione della macchina virtuale e dalla quantità di dati archiviati.



* Analytic engine (L2) (**Google BigQuery**): Il costo dipenderà dalla quantità di dati memorizzati e dalla complessità delle query.



**Data Consumption**:

* Custom ML Model (**Google Cloud AI Platform**): Il costo dipenderà dalla complessità dei modelli di machine learning e dal tempo di addestramento.
* AI Platform Engine (**Google Cloud AI Platform**): Il costo dipenderà dal numero di inferenze o predizioni effettuate dai modelli di machine learning.

**Data Exposure/Visualization**:

* Data Visualization Tool (**Google Data Studio**): Il costo dipenderà dalla frequenza di creazione di report e dashboard.
* Enterprise BI Tool (**Google BigQuery**): Il costo dipenderà dalla configurazione della macchina virtuale e dalla quantità di dati archiviati su BigQuery.
* Custom API development (**Google Cloud Endpoints**): considerando le seguenti informazioni:

1. Numero medio di clienti: 450.000
2. Dimensione media dei dati di registrazione del cliente per ogni cliente: 50 KB
3. Numero medio di eventi dei clienti in tempo reale: 5000 al giorno
4. Dimensione media di ogni evento: 40 KB
5. Dimensione media del file CSV dei dati sui prezzi: 100 MB

Per calcolare il numero di richieste al mese, dobbiamo considerare due tipi di richieste:

1. Richieste di registrazione dei clienti: ogni cliente genera un'operazione di registrazione giornaliera, quindi il numero di richieste al mese sarà il numero medio di clienti moltiplicato per il numero di giorni in un mese.

* + Numero di richieste al mese per le registrazioni dei clienti: 450.000 clienti \* 30 giorni (media dei giorni in un mese) = 13.500.000 richieste al mese.

2. Richieste di eventi dei clienti in tempo reale: Ogni evento dei clienti genera una richiesta. Dato che abbiamo 5000 eventi al giorno, possiamo moltiplicare questo per il numero medio di giorni in un mese.

* + Numero di richieste al mese per gli eventi dei clienti in tempo reale: 5000 eventi/giorno \* 30 giorni = 150.000 richieste al mese.

Quindi, il totale delle richieste al mese per entrambi i tipi di operazioni sarà la somma dei due:

Totale delle richieste al mese = 13.500.000 (registrazioni dei clienti) + 150.000 (eventi dei clienti in tempo reale) = 13.650.000 richieste al mese.

Tuttavia, poiché il campo richiede il numero di richieste per mese in conteggio di milioni, dobbiamo dividere il totale per 1 milione:

Numero di richieste per mese (in milioni) = 13.650.000 / 1.000.000 = 13.65

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, Carattere, numero, linea

Descrizione generata automaticamente