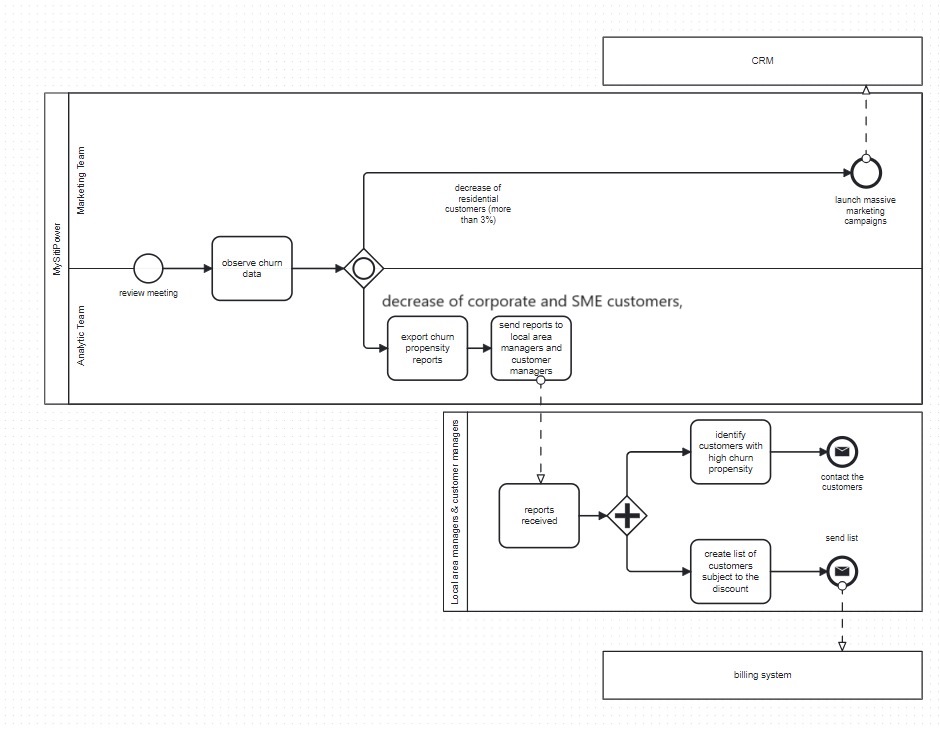
**PART 2 PROJECT: SOLUTION**

**Solution to Request 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.*



**Solution to Request 2: Google Cloud Architecture**

As a public cloud service, I chose Google Cloud, which offers several distinctive advantages over other public cloud service providers, such as:

* **Advanced services ecosystem: Google Cloud**  offers a comprehensive ecosystem of advanced services, including artificial intelligence and machine learning with Google Cloud AI Platform, data analytics with BigQuery, and scalable compute and storage services such as Compute Engine and Google Cloud Storage. This wide range of services allows developers to implement complex solutions without having to depend on multiple vendors.
* **Security and compliance**: Google Cloud provides a robust security framework with advanced encryption, role-based access, and event monitoring capabilities. Additionally, Google Cloud complies with numerous security and privacy standards, such as GDPR, HIPAA, and ISO 27001, ensuring regulatory compliance for a wide range of industries and markets.
* **Performance and scalability**: Google Cloud offers a high-speed global network with reduced latency thanks to its distributed network infrastructure. This results in high performance and fast response times for applications hosted on Google Cloud. In addition, the elastic scalability allows applications to handle rapidly growing workloads without compromising performance.
* **Continuous innovation**: Google is at the forefront of technological innovation and invests heavily in research and development to constantly improve its cloud services. This results in a continuous stream of new features and services, allowing users to take advantage of the latest technologies to maintain their competitiveness in the market.
* **Competitive** pricing: Google Cloud offers a competitive pricing structure with flexible pay-as-you-go billing options. In addition, it offers discounts and special offers for long-term workloads and customers committed to using its services on a regular basis.

Some of the possible drawbacks of Google Cloud include the learning curve to make the best use of all of its features, the lack of some specialized services found in other cloud providers, and the need to integrate with other external services and platforms. However, considering its strengths and growing adoption in the industry, Google Cloud remains a highly competitive and attractive choice for many businesses and organizations.

In addition, Google Cloud comes with many security, privacy, and compliance features. Here are some specific examples:

* **Encryption:** Google Cloud automatically encrypts data in transit and at rest.
* **Key management:** Google Cloud Key Management Service (KMS) and Cloud HSM make it easy to manage encryption keys.
* **Access control:** Google Cloud Identity & Access Management (IAM) allows you to control who has access to your cloud resources and what type of access they have.
* **Two-factor authentication (2FA):** Google offers two-factor authentication to add an extra layer of security.
* **Logging and monitoring:** Google Cloud Logging and Monitoring allow you to keep track of activities in your cloud environment.
* **Compliance:** Google Cloud complies with global standards such as GDPR, HIPAA, ISO, among others.
* **Backup and restore:** Google Cloud offers services like Persistent Disk Snapshots for backing up your data.



The proposed Architecture is based on Google Cloud components:

Image Containing Text, Diagram, Character, Line

Auto-generated description

1. From the external manufacturer, the raw data actually passes to the Digital Data Platform via the Data Landing block. The Batch Data Transfer (Cloud Storage) component is used to transfer daily data from the CRM to the Data Lakehouse, but also for real-time API data.
2. The data transferred to the Data Landing then passes to **the Data Manipulation** via the ETL Data Pipelines (Google Dataflow). This data manipulation process is managed and coordinated by the Workflow Orchestration (Cloud Composer).
3. Subsequently, the manipulated data is stored in the Data Lake block, traversing the various layers:

* **L0 (**Raw Storage): Raw data is retained without any transformation using the Raw Data Persistence (Cloud Storage) component.

1. **L1 (Curated):** Processed and structured data is stored in a relational database using Google Cloud SQL.
2. **L2:** Aggregated and pre-processed data is stored in a high-performance analytical environment using Google BigQuery.
3. The manipulated data is then available for Data **Consumption** via Data Manipulation.
4. **Data Consumption** can access data from the L2 Data Lake for analysis via the AI Platform Engine.
5. The External consumer makes a request to view the final data to the **Data Exposure/Visualization**, which can take over it through the Custom API Development component (Google Endpoints).
6. The  **Data Exposure** handles this request via a link to the **Data Consumption**, which will forward the data in the next step.
7. The results of the analysis are now made available from **Data Consumption** to **Data Exposure/Visualization**. The data is first visualized and analyzed via the Data Visualization Tool (Google Data Studio)
8. They are then further processed in the Enterprise BI Tool (Google BigQuery) for a more detailed and in-depth analysis.
9. Finally, the data is provided to the **External Consumer** for external use and final display.

Description of the proposed Cloud Architecture, based on Google Cloud, and its components:

**1. Data Landing**:

* **Batch** data transfer: This component is used to transfer daily data from the CRM system, pricing system, and also real-time data from API to the data lakehouse.

**2. Data Manipulation**:

* **ETL Data Pipelines**: This component is critical for processing and preparing raw data from the CRM system and pricing system for later storage and analysis.
* **Workflow Orchestration**: Used to coordinate workflow between different components and services of the architecture.

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

* **Raw data persistence (L0):** Used to preserve raw data from the CRM system and pricing system without any transformation.
* **Relational database and transactional engine (L1):** This layer is used to store processed and structured data to enable data analysis and manipulation.
* **Analytic engine (L2):** Used to store aggregated and pre-processed data, ready for advanced analysis and processing by business intelligence systems.

**4 Data Consumption:**

* **Custom ML Model**: Used to analyze data and provide customized predictions or models based on the customer's specific needs.
* **AI Platform Engine**: Used to implement artificial intelligence and machine learning algorithms to analyze and derive insights from data.

**5. Data Exposure/Visualization:**

* **Data Visualization Tool**: This component allows you to create interactive visualizations of your data for analysis and presentation. It is useful for communicating the results of the analysis in a clear and intuitive way.
* **Enterprise BI Tool**: Provides tools for business analysis and data-driven reporting. It's important for gaining in-depth insights and supporting business decisions.
* **Custom API development**: You may need to develop custom APIs to enable real-time data transfer from your CRM system.

**Justification for the choices, with corresponding components in Google Cloud:**

* **Batch** data transfer: Chosen to efficiently transfer large volumes of data from the CRM system and pricing system to the data lakehouse, ensuring optimized management of daily workloads. 🡪 **Google Cloud Storage**
* **ETL Data Pipelines**: Essential for processing and preparing raw data for later analysis, ensuring data quality and consistency within the data lakehouse. 🡪 **Google Cloud Dataflow**
* **Workflow Orchestration**: Used to coordinate workflow between different components and services of the architecture, ensuring efficient management of data manipulation tasks. 🡪 **Google Cloud Composer**
* **Raw data persistence (L0):** Chosen to preserve raw data from the CRM system and pricing system without any transformation, preserving the integrity of the original data. 🡪 **Google Cloud Storage**
* **Relational database and transactional engine (L1):** This layer is used to store processed and structured data to enable data analysis and manipulation, providing a solid foundation for querying data. 🡪 **Google Cloud SQL**
* **Analytic engine (L2):** Chosen to store aggregated and pre-processed data, ready for advanced analysis and processing by business intelligence systems, facilitating the generation of meaningful insights. 🡪 **Google BigQuery**
* **Custom ML Model**: Used to analyze data and provide predictions or models tailored to the customer's specific needs, allowing for greater customization of data analytics solutions. 🡪 **Google Cloud AI Platform**
* **AI Platform Engine**: Chosen to implement artificial intelligence and machine learning algorithms to analyze and derive insights from data, enabling a deeper understanding of data patterns and patterns. 🡪 **Google Cloud AI Platform**
* **Data Visualization** Tool: This component allows you to create interactive visualizations of your data for analysis and presentation, making it easier for end users to understand your data. 🡪 **Google Data Studio**
* **Enterprise BI Tool**: Provides tools for business analysis and data-driven reporting, allowing for fast and accurate generation of business reports and in-depth analytics. 🡪 **Google BigQuery**
* **Custom API development**: Necessary to integrate the CRM system and allow the real-time transfer of customer events, ensuring smooth communication between the different systems. 🡪 **Google Cloud Endpoints**

**Calculation of Associated Costs for Google Cloud**:

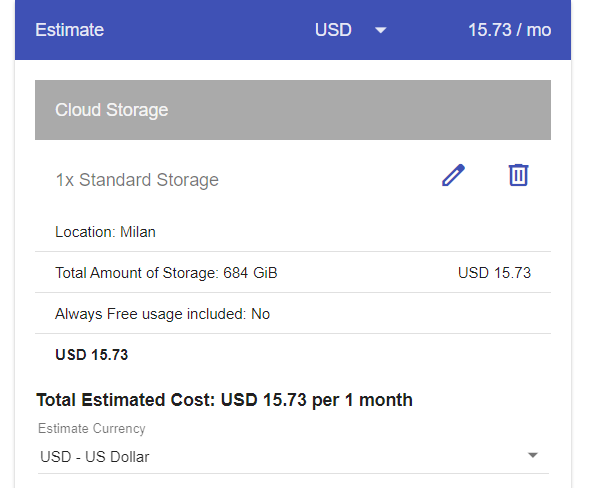
**Data Landing**:

* Batch data transfer (**Google Cloud Storage**): To calculate the total amount of storage needed, you need to add up the volume of data generated each day by customers and events, plus the size of the CSV file of pricing data.

**Customer registration data:** 450,000 customers \* 50 KB /customer = 22,500,000 KB = 22.5 GB

**Real-time customer events:** 5000 events/day \* 40 KB/event = 200,000 KB = 0.2 GB

**Pricing data: 100 MB = 0.1 GB**  Adding it all up, we have a total of about 22.8 GB of data per day, so per month it would be: 22.8GB \* 30days = 684 GB 🡪 the Total Estimated Cost: USD 15.73 for 1 month🡪



**Data Manipulation**:

* ETL Data Pipelines (**Google Cloud Dataflow**):
  + **Region**: "Milan (europe-west3)"
  + **Job type**: "Batch" - Suitable for processing CSV files on a daily basis.
  + **Shuffle enabled**: Selected – to perform join or group by operations on the data.
    - **Data Processed**:
      * Pricing data CSV file size: 100 MB per day.
      * Total size of customer registration data: 450,000 customers \* 50 KB/customer.
      * Total real-time customer event size: 5000 events/day \* 40 KB/event.
      * First of all, let's convert everything to KB:
        + 100 MB per day is 100 \* 1024 = 102,400 KB.
        + 450,000 customers \* 50 KB/customer is 22,500,000 KB.
        + 5000 events/day \* 40 KB/event is 200,000 KB.
      * Now let's add it all up: 102,400 KB (CSV file) + 22,500,000 KB (customer registration data) + 200,000 KB (real-time customer events) = 22,802,400 KB.
      * Finally, we convert to GB: 22,802,400 KB / 1024 (to convert to MB) / 1024 (to convert to GB) = about 21.74 GB.
      * So, the total amount of "Data Processed" should be around 21.74 GB per day.
  + **Hours the job runs per month**: 30 - (1 hour per day for 30 days).
  + **Number of worker nodes used by job**: 2 - Starting with 2 is a good place to start.
  + **Worker node instance type**: "n1-standard-1 (vCPUs: 1, RAM: 3.75 GB)"
  + **System... PD:** 200 - This value represents the amount of persistent storage available to the worker nodes.

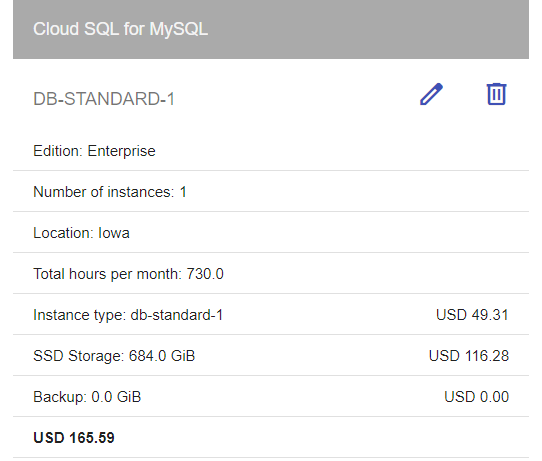
Image Containing Text, Screenshot, Font, Number

Auto-generated description

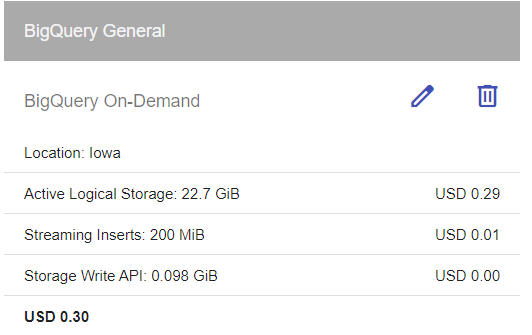
* Workflow Orchestration (**Google Cloud Composer**): The number of jobs and the complexity of the workflow will depend on the data processing logic and the frequency of updates.

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

* Raw data persistence (L0) (**Google Cloud Storage**): Storage is already calculated for the Data Landing component.
* Relational database and transactional engine (L1) (**Google Cloud SQL):** The cost will depend on the configuration of the virtual machine and the amount of data stored.



* Analytic engine (L2) (**Google BigQuery**): The cost will depend on the amount of data stored and the complexity of the queries.



**Data Consumption**:

* Custom ML Model (**Google Cloud AI Platform**): The cost will depend on the complexity of the machine learning models and the training time.
* AI Platform Engine (**Google Cloud AI Platform**): The cost will depend on the number of inferences or predictions made by machine learning models.

**Data Exposure/Visualization**:

* Data Visualization Tool (**Google Data Studio**): The cost will depend on how often you create reports and dashboards.
* Enterprise BI Tool (**Google BigQuery): The cost will depend on the configuration of the virtual machine and the amount of data stored on BigQuery.**
* Custom API development (**Google Cloud Endpoints**): considering the following information:

1. Average number of customers: 450,000
2. Average size of customer registration data per customer: 50 KB
3. Average number of real-time customer events: 5000 per day
4. Average size of each event: 40 KB
5. Average price data CSV file size: 100 MB

To calculate the number of requests per month, we need to consider two types of requests:

1. Customer registration requests: Each customer generates a daily registration operation, so the number of requests per month will be the average number of customers multiplied by the number of days in a month.

* + Number of requests per month for customer registrations: 450,000 customers \* 30 days (average of days in a month) = 13,500,000 requests per month.

2. Real-time customer event requests: Each customer event generates a request. Since we have 5000 events per day, we can multiply this by the average number of days in a month.

* + Number of requests per month for real-time customer events: 5000 events/day \* 30 days = 150,000 requests per month.

So, the total requests per month for both types of operations will be the sum of the two:

Total requests per month = 13,500,000 (customer registrations) + 150,000 (live customer events) = 13,650,000 requests per month.

However, since the field requires the number of requests per month to count in millions, we need to divide the total by 1 million:

Number of requests per month (in millions) = 13,650,000 / 1,000,000 = 13.65

Image Containing Text, Screenshot, Font, Number

Auto-generated description

Image That Contains Text, Font, Number, Line

Auto-generated description