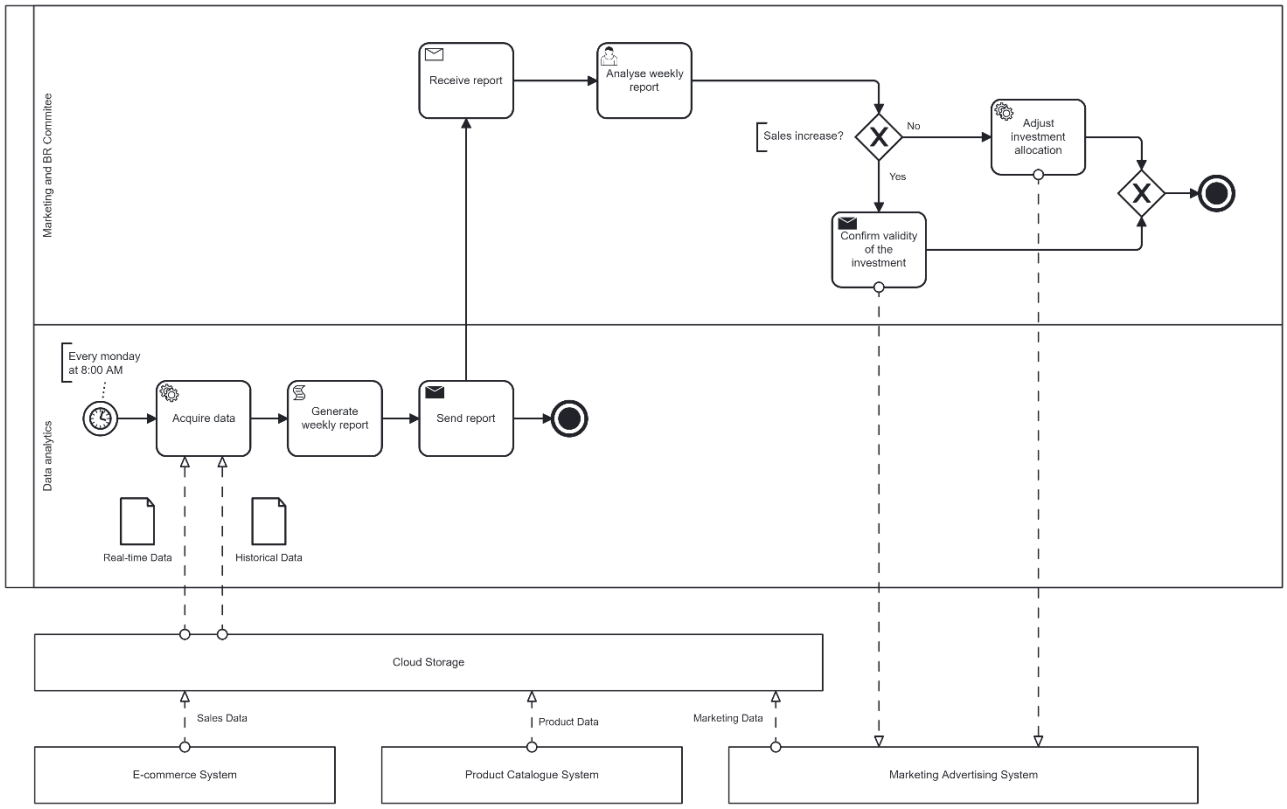
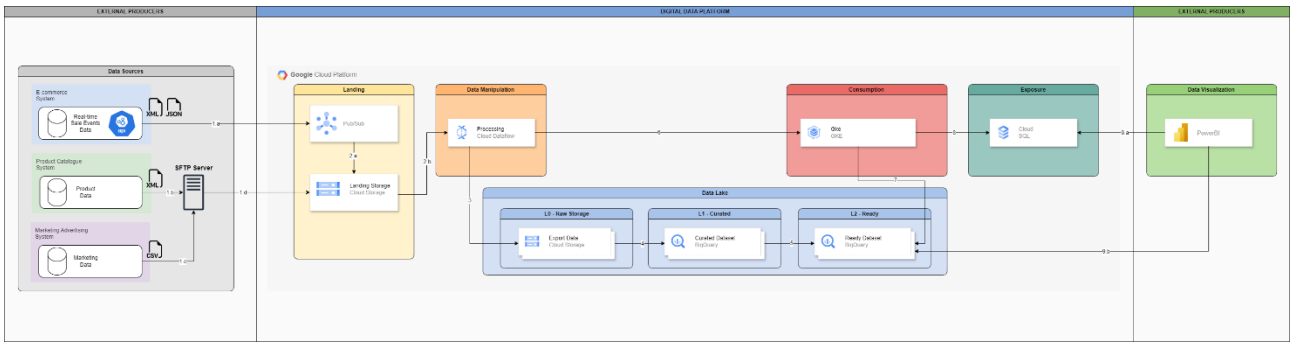
BPMN, Data Platform Architecture and Cost Estimation

Fabiano Mangini

BPMN



Data Platform Architecture



Functional Layers:

Data Landing Layer:

- Entry point for external data sources.
- Allows ingestion of batch and streaming data.

Data Manipulation:

- Manages data transformations and quality checks.
- Orchestrates workflow and custom data processing pipelines.
- Ensures data readiness for further processing within the Data Platform.

Data Lakehouse:

- Stores data persistently within the Data Platform in layers: L0, L1, and L2.
- Utilizes Google Cloud Storage for L0 and BigQuery for L1/L2.
- Transforms data from raw to curated formats for efficient storage and querying.
- Facilitates cost-effective storage and streamlined data access.

Data Consumption:

- Retrieves data from the Data Lakehouse for business use cases and analytics.
- Enables exploration and testing by data science teams and engineers.
- Implements custom application components and AI models for advanced analytics.

Data Exposure/Visualization:

- Securely exposes data to third parties via custom APIs or gateway components.
- Enables real-time serving and integration with external services.
- Enhances collaboration and innovation by extending data access to diverse ecosystems.

Components description:

Google Cloud Pub/Sub: This component enables asynchronous communication between services, facilitating event-driven systems at scale.

Google Cloud Storage: It offers durable and highly available storage for various data types, ensuring scalability and cost-effectiveness.

Google Cloud Dataflow: This service handles data transformations and enrichments in both stream and batch modes, implementing quality checks on incoming data streams.

Google BigQuery: Serving as a serverless data warehouse, BigQuery stores curated and ready data for efficient querying and analysis, fitting into the L1 and L2 layers of the Data Lakehouse architecture.

Google Kubernetes Engine (GKE): GKE manages containerized applications in a flexible, scalable environment, supporting automated operations and monitoring features.

Google Cloud SQL: As a fully managed relational database service, Cloud SQL handles storing and managing data, supporting reporting and data exposure functionalities within the ecosystem.

Data Flow:

1. Real-time data from the E-commerce system is sent to Google Cloud Pub/Sub, while data from Product Catalog and Marketing Advertising systems pass through the SFTP server and are sent to Google Cloud Storage.
2. Google Pub/Sub sends the received data to the Google Cloud Storage and Google Cloud Dataflow processes data after being triggered.
3. Raw data is stored in the L0 layer of the Data Lakehouse.
4. Processed data is appended into L1 of the Data Lakehouse in BigQuery, maintaining a historical record.
5. Data moves to L2 of the Data Lakehouse in BigQuery for further processing and utilization by machine learning models and visualization tools like Power BI.
6. GKE is triggered for data consumption after processing.
7. ML algorithm retrieves data from L2 to predict revenue trends based on advertising investments and competitor sales.
8. ML model features and output are saved in Google Cloud SQL.
9. Power BI reports are generated using data from Google Cloud SQL and the L2 layer of the Data Lakehouse and can be visualized by Marketing and Business Representative Committee to perform analysis and decision-making activities.

Costs Estimation

Item	Price Driver	Usage	Cost
Google Cloud Pub/Sub	Volume (MB)	750	0.44 €
Google Cloud Storage (Landing)	Volume (GB)	170	3.07€
Google Cloud Dataflow	Time (h)	30	2.29 €
Google Cloud Storage (L0)	Volume (GB)	170	3.07 €
Google BigQuery (L1)	Volume (TB)	5	25,00 €
Google BigQuery (L2)	Volume (TB)	5	25,00 €
GKE Standard Node Pool	Time (h)	730	23.54 €
Cloud SQL for PostgreSQL	Time (h)	730	10.23 €
Total			95.71 €

Costs Description:

- All costs are evaluated for a period of 1 month.
- Google Cloud Pub/Sub: Processes a maximum volume of 750 MB (50KB per event with 15.000 events per day) per day from the E-commerce System, this worst case is used for the cost evaluation.
- Google Cloud Storage – Landing/L0 layers: Stores transient data, requiring 170 GB per month to accommodate daily data influx of about 5.6GB (60.000 products with 80KB per product + 750 MB).
- Google Cloud Dataflow: Operates for 30 hours monthly with a single worker node and 6GB block storage per worker.
- Google BigQuery L1/L2 layer: Handles a query volume of 5TB monthly.
- GKE Standard Node Pool: Operates continuously, 24 hours a day, 7 days a week.
- Cloud SQL for PostgreSQL: Configured with "df-f1-micro" instance type and 20 GB SSD storage.