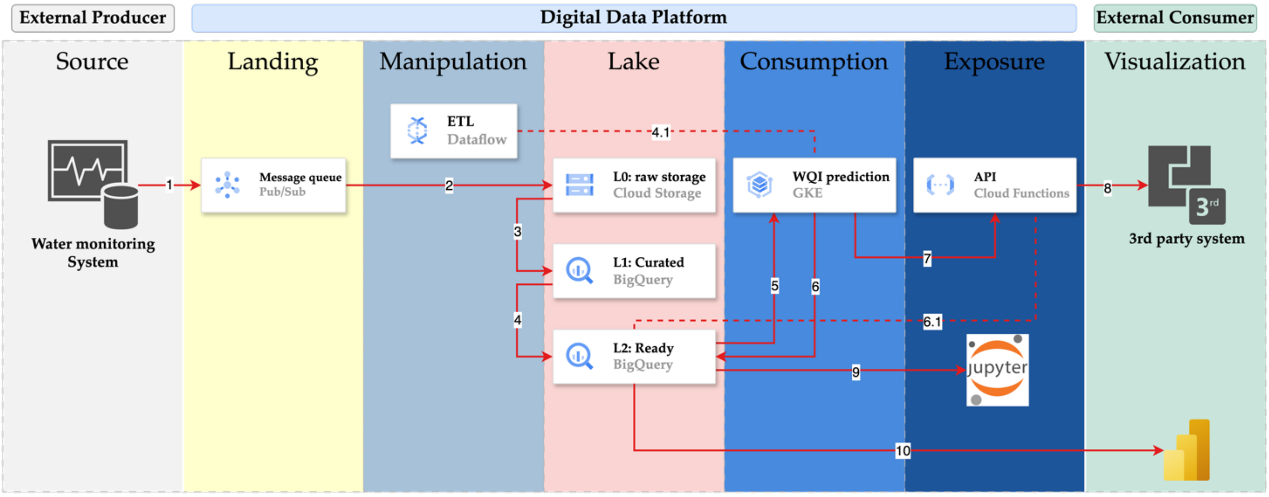
Tools and Technologies

PREZIOSA ALESSIA (590012)

Politecnico di Bari – xTech bip

Data Platform Architecture



1. An external water monitoring system collects data from sensors all over the country and pushes data in real-time: the data stream is integrated into the data platform via a ***Pub/Sub*** component that permits asynchronous communication. The publisher is, in this case, the external system which sends events regardless of how and when these will be processed.
2. Every 15 minutes, ***Dataflow*** (subscriber of the topic) reads newly arrived records and saves them as a text file (.csv) in lake L0 (Raw Storage) applying an initial elaboration regarding technical data quality. Data is stored in standardized raw format and masked with respect to regulations. A ***Cloud Storage*** bucket is used: it is not high-performance, but it's economical and can allow the storage of unstructured data in compressed format.
3. ***Dataflow*** performs refinements on data and stores them, in appended mode, in the L1 layer (Curated) of ***BigQuery***, a distributed, managed and high-performance Data Warehouse service capable of hosting structured data.
4. ***Dataflow*** aggregates and enriches data based on similar characteristics storing them into the L2 (Ready) ***BigQuery*** dataset. Here data is consumable and ready for Machine Learning algorithms, it can be analyzed by Data Engineers, published and exported on business tools.

4.1 The elaboration pipeline triggers the algorithm deployed (and exposed) on ***Google Kubernetes Engine***.

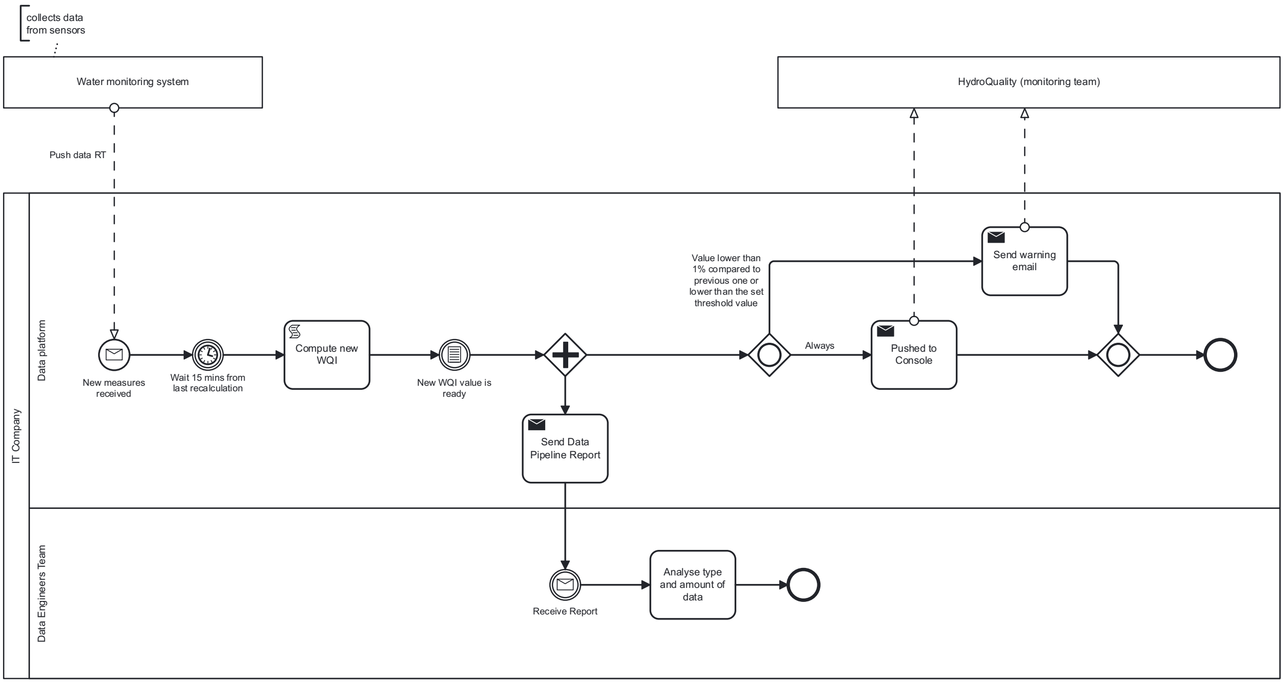
5. A custom microservice (***GKE***) predicts the WQI value using L2 dataset.
6. New and updated WQI values are stored in L2 dataset to be used and analyzed, along with the corresponding parameters, by the exposure and visualization layers (respectively in ***Jupyter Notebook*** (9) and ***PowerBI*** dashboard (10))

6.1 ***Cloud Functions*** is triggered by the operation of UPDATE on L2 dataset after ***GKE*** writes the newly computed WQI values; the data exposure layer updates data on 3rd party system through custom API for real time serving (8)

7. ***GKE*** can also invoke ***Cloud Functions*** if the new value is lower than 1% compared to the previous value or is lower than the set threshold value and send an email to the External Monitoring Team to warn them (8)

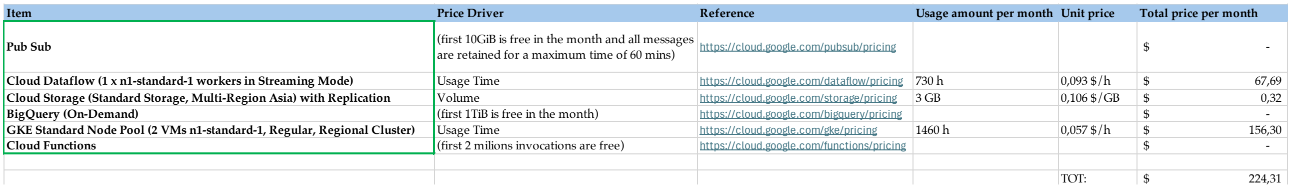
BPMN process

In the following picture, the BPMN process is represented.


collects data from sensors
Water monitoring system
HydroQuality (monitoring team)
Push data RT
IT Company
Data platform
Data Engineers Team
New measures received
Wait 15 mins from last recalculation
Compute new WQI
New WQI value is ready
Send Data Pipeline Report
Value lower than 1% compared to previous one or lower than the set threshold value
Always
Pushed to Console
Send warning email
Receive Report
Analyse type and amount of data

The Water monitoring system and the HydroQuality Agency have been designed as black boxes.

Running Costs



Item	Price Driver	Reference	Usage amount per month	Unit price	Total price per month
Pub Sub	(first 10GiB is free in the month and all messages are retained for a maximum time of 60 mins)	https://cloud.google.com/pubsub/pricing			$ -
Cloud Dataflow (1 x n1-standard-1 workers in Streaming Mode)	Usage Time	https://cloud.google.com/dataflow/pricing	730 h	0,093 $/h	$ 67,69
Cloud Storage (Standard Storage, Multi-Region Asia) with Replication	Volume	https://cloud.google.com/storage/pricing	3 GB	0,106 $/GB	$ 0,32
BigQuery (On-Demand)	(first 1TiB is free in the month)	https://cloud.google.com/bigquery/pricing			$ -
GKE Standard Node Pool (2 VMs n1-standard-1, Regular, Regional Cluster)	Usage Time	https://cloud.google.com/gke/pricing	1460 h	0,057 $/h	$ 156,30
Cloud Functions	(first 2 milions invocations are free)	https://cloud.google.com/functions/pricing			$ -
				TOT:	$ 224,31

Pub/Sub: daily, 1.25MB of data transit on average and messages (both acknowledged and unacknowledged) are retained for no more than 60 minutes. It doesn't imply any type of costs.
Dataflow: Since messages are of few KBs, 1 x n1-standard-1 workers in Streaming Mode is chosen for the job. It must work 24/7 during all month.
Cloud Storage: It's a Standard Storage with Multi-Region Replication, with an occupied space of 3GB per month
BigQuery: 1st TB of query is free; no costs are due.
Google Kubernetes Engine: 2VMs per node pools in a Regional Cluster is established.
Cloud Functions: no costs are due.