CNR-IIA: PROGETTO ARMONIA

Flusso dati in Cloud

Autore	Francesco D'Amore
Data di creazione	1-Ottobre-2022
Ultima revisione	30-Ottobre-2022
Titolo	D5.1 - Flusso dati in Cloud
Soggetto	WP5 - Infrastruttura ICT
Stato	Completato
Editore	CNR-IIA
Tipo	Deliverable
Identificazione	D5.1
Descrizione	
Contributi	Mariantonia Bencardino, Delia Evelina Bruno, Valentino Mannarino

INDICE

INTRODUZIONE E FINALITÀ

I dati prodotti dai sensori individuati nel deliverable D2.1 e assemblati come descritto nel deliverable D2.2 devo essere gestiti tramite risorse computazionali capaci di immagazzinare il dato e condividerlo con l'utente finale.

I processi e i dati nelle moderne architetture ICT (Information and Communication Technologies) vengono ospitati in infrastrutture Cloud dove le risorse computazionali vengono virtualizzate al fine di renderle flessibili e utilizzabili su richiesta dell'utente in base alle reali necessità.

In ARMONIA si vogliono sperimentare metodi di gestione dei dati ambientali orientati al cloud dove, successivamente all'inserimento, i dati verrebbero processati e analizzati. Il processo di acquisizione del dato tramite ponte IoT verrà descritto nel deliverable D5.2, mentre nel presente documento si descrive la configurazione della piattaforma Cloud per la gestione del dato acquisito.

Nella scelta di approcci orientati al Cloud Computing è importante considerare anche tematiche ambientali: acquistare macchine server per il processamento dei dati relativi ad un solo progetto specifico non rientra nelle buone pratiche atte a minimizzare l'impatto ambientale, tema di strettissima attualità che va considerato soprattutto per un progetto con le finalità di quello in oggetto: le risorse cloud, al contrario di quelle tradizionali possono essere rilasciate quando non sono più necessarie.

Dal punto di vista funzionale, l'obiettivo dell'architettura cloud descritta nel prossimo paragrafo è quello di acquisire dati grezzi in arrivo da canali IoT, immagazzinare i dati in data warehouse per il loro processamento e, infine, utilizzare sistemi di Business Intelligence (BI) per visualizzare i dati e condividerli con l'utente finale.

Il documento presente si focalizza sulla sola progettazione e configurazione della piattaforma cloud di ARMONIA: la descrizione dell'implementazione esula dagli scopi di questo deliverable e sarà oggetto uno successivo.

In alcuni casi verranno descritte soluzioni alternativi: in base alle specifiche esigenze, in fase realizzativa, queste scelte collesseranno in una specifica direzione implementativa.

ARCHITETTURA CLOUD BASED PER IL PROCESSAMENTO DEI DATI

Il fornitore Cloud che si è scelto per ARMONIA è Google Cloud Platform (GCP). Sono state considerate a tal proposito le problematiche relative al blocco tecnologico (Technological Lock In), cioè quel fenomeno che rende una soluzione tecnica strettamente dipendente dalle piattaforme di supporto individuate. Per mitigare il blocco tecnologico dovuto alla scelta di uno specifico rivenditore Cloud (Google nel caso di ARMONIA), sono stati individuati protocolli standard, come MQTT nel caso dei dati trasferiti tramite canali IoT, e componenti Open Source.

Si è inoltre optando per servizi cloud che hanno una chiara controparte anche su altre piattaforme concorrenti di quella prescelta.

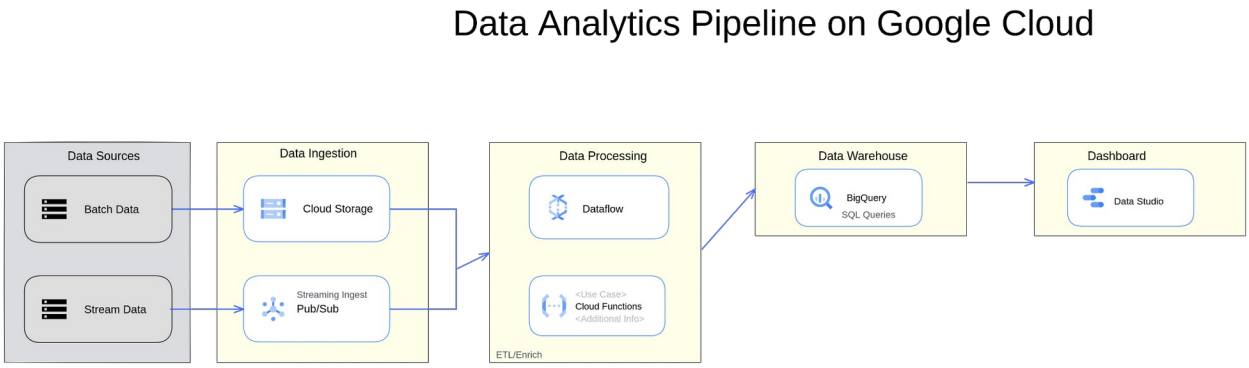
Nel caso di ARMONIA, i servizi principali scelti per l'architettura dati sono i seguenti:

- IoT Core: gestore dati IoT e Broker MQTT.
- Dataflow e/o Cloud Functions: processamento e pipeline di dati
- BigQuery: data warehouse
- Cloud Storage: Immagazzinamento standard dei dati

Nell'immagine che segue si descrive per blocchi il sistema progettato per ARMONIA su Google Cloud. Il flusso dei dati va da sinistra verso destra: inizialmente il dato viene raccolto da dispositivi

Single-Board Computer (Computer a Scheda Singola) e successivamente inviati verso il cloud utilizzando flussi in (quasi) tempo reale (*Stream Data*) tramite ponti IoT, come illustrato del deliverable D5.2.

Nel caso di indisponibilità della rete dati, le informazioni vengono raccolte nel device locale e successivamente trasferiti (*Batch Data*) in sistemi capaci di immagazzinare dati grezzi (Cloud



Storage).

I dati provenienti da canali IoT in (quasi) tempo reale vengono invece raccolti da da buffer che rendono più flessibile l'architettura, disaccoppiando le fasi di inserimento del dato da quelle di immagazzinamento e analisi.

In funzione della struttura del dato raccolto, ci si può trovare nella necessità di processare il dato con approcci diversi:

- ETL (Extraction, Transformation and Load) nel caso di dati scarsamente strutturati
- EL (Extraction and Load) nel caso di dati strutturati.

Nel primo caso la struttura dell'informazione raccolta necessità un processamento importante appena dopo l'inserimento nella piattaforma cloud e si dovrà usare, a tal fine, un componente come Dataflow che permette l'impiego di processi astratti di aggregazione e processamento.

Nel secondo caso il dato raccolto dai devices locali ha una struttura già adeguata all'immagazzinamento in Data Warehouse (Big Query) e saranno quindi sufficienti sistemi meno onerosi dal punto di vista computazionale per il passaggio verso sistemi di immagazzinamento, come ad esempio Cloud Function di GCP.

L'ultimo componente della piattaforma descritta sopra raccoglie il dato strutturato nel Data Warehouse (Big Query) e lo utilizza per creare strumenti per la condivisione del dato in Web Dashboard. A tal fine viene usato Data Studio, componente di GCP direttamente connesso al Data Warehouse.