Microsoft

Action Plan | Predisposizione Monitoraggio xBOT

Submitted to	Submitted by	Submission date
Poste Italiane	Microsoft Customer and Success Team	13/12/2021

Table of Contents

Contents

Microsoft

Scopo del documento

Il presente documento fornisce un action plan per riportare a norma il monitoraggio del sistema di ChatBot ad oggi in produzione; Poste Italiane richiede questo intervento in quanto l'attuale monitoring del sistema al momento non è in linea con gli attuali standard.

Aree di intervento

A valle del meeting del 30/11/2021 sono state individuati 4 ambiti di intervento, per i quali vengono di seguito forniti sia i dettagli del problema sia la proposta di procedura per la normalizzazione.

Timeout da Genesys verso BOT

Il punto di ingresso rispetto ad un utente che fa una richiesta via voce o via chat è **Genesys**, configurato per smistare le richieste al BOT; sistema di cui **Accenture** segue gli sviluppi mentre la gestione in esercizio è gestita da **PosteItaliane**.

Se la comunicazione non viene completata entro un timeout configurato (20 secondi per Genesys, 10 secondi per il server Tomcat), l'utente viene trasferito ad un operatore umano. Le persone del "business" di Poste per le varie iniziative hanno dei sistemi di reportistica/alert lato Genesys per intercettare questi timeout; di conseguenza sono state ricevute segnalazioni discordanti in quanto vengono ricevuti alert dal prodotto che non hanno riscontri con il monitoraggio attuale lato risorse BOT su Azure.

Action plan

L'action plan proposto per questo ambito si articola in differenti step che possono essere svolti parallelamente:

- Raccolta delle informazioni e verifica dell'attuale ambiente di hosting, relative configurazioni e log applicativi attualmente disponibili per la piattaforma Genesys; a valle della verifica dove opportuno si può intervenire rafforzando la raccolta delle informazioni e convogliarle verso un unico **Log Analytics Workspace** configurato con una opportuna retention policy.
- Rafforzare il monitoraggio sugli endpoint di ingresso lato Azure, ovvero diversi AppService corrispondenti ai Bridge o al BOT monitorando le metriche canoniche:
    - Richieste HTTP
        - Duration
        - Status Code
        - Exceptions
    - Compute
        - Memory
        - CPU Percentage
        - HTTP Queue
        - Etc.

Altro punto da monitorare è l'**Express Route** tramite la quale Genesys comunica con Azure; questo in quanto è emerso un problema relativo a dei timeout dove il tempo totale della transazione E2E non coincideva con la somma delle singole transazioni dei vari nodi attraversati quindi è stato ipotizzato un sovraccarico del networking.

Timeout/Errori Genesys verso Cognitive Services (speech to text, text to speech)

Genesys comunica con il BOT solo in modalità "testo" di conseguenza quando si passa per canali telefonici è necesario fare una conversione **Spech2Text/Text2Speech**. Questa viene realizzata grazie aduna integrazione ai Cognitive Service.

Trattandosi di un passaggio fondamentale è necessario considerare i casi in cui possa fallire. Questa integrazione potrebbe fallire per 2 motivi:

- Eventuali tempi di risposta più alti della media da parte dei Cognitive Service.
- Richiesta di **Text2Speech** (che rappresenta la risposta del BOT) formattata male visto che viene configurato manualmente un modello SSML dai vari team che gestiscono I BOT

Action Plan

L'action plan proposto per questo ambito si compone di un'azione su due fronti: **Genesys** e **Azure Cognitive Services**

- Genesys: Le chiamate da Genesys verso Azure vengono effettuate tratmie la **"Unimrcp"** quindi va sicuramente spinto il monitoraggio dei log del prodotto (vedi punto precedente)
- Cognitive Services: è possibile monitorare quanto messo a disposizione:
  - Resource health
  - Latency (tempo di risposta)
  - Client Errors
  - Server Errors

Inoltre durante il meeting è stata menzionata anche una possibile implementazione di una validazione del markup SSML digitato dagli operatori in fase di configurazione del flusso sul Conversational Designer, si tratterebbe di un intervento di modifica del software.

Loop infinito in flussi conversazionali

I flussi conversazionali sono progettati attraverso uno strumento visuale denominato **"Conversational Designer"** a volte nella progettazione potrebbero esserci degli errori logici che portano ad una ricorsione di "branch" per cui si verifica una fase di "stallo" nell'elaborazione della richieste dall'utente

Action Plan

L'action plan proposto prevede di identificare/realizzare una query sulla telemetria applicativa che riesca ad intercettare le seguenti casistiche:

- Tutte le conversazioni non chiuse entro una finestra temporale accettabile (parametrizzabile, es. 1 ora); oltre la finestra definita si considra visto che potrebbe essere in loop
- Estraendo le conversazioni con un numero alto di iterazioni rispetto allo stesso "messaggio"

Dato che questi loop infiniti causano impatti sulle performance del BOT (basti pensare all'utilizzo CPU) vanno indagate queste metriche, e monitorarle al fine di poter riportare l'ambiente in situazioni ottimali; un esempio potrebbe essere il riavvio della singola istanza AppService tramite l'uso di runbook.

Monitoraggio delle singole iniziative

Come anticipato ci sono più **"iniziative"** che utilizzano un'unica infrastruttura BOT comune quindi talvolta il SOM ha la necessità di dover rendicontare lo stato di salute dei servizi sottesi alle richieste della singola iniziativa. È necessario preparare un documento che evidenzi l'elenco delle **iniziative esistenti** ed il **mapping** dei servizi di terze parti associati alla singola iniziativa; tale documento conterrà quindi i dettagli di mapping tra gli endpoin richiamati e gli ambiti di gestione Bot. Ad ogni rilascio successivo verrà condivisa l'informazione dell'ambito di afferenza dei nuovi endpoint utilizzati così da mantenere questa informazioni aggiornata.

Action Plan

La redazione del documento/elenco è in carico ad **Accenture**.
La singola iniziativa deve poter essere identificata da almeno un codice che viene propagato su tutta la telemetria, in tal modo è possibile raggruppare le metriche descritte in precedenza rispetto a questo valore; a partire da questa condizione diventa quindi possibile:

- Creare delle dashboard ed alert per singola iniziativa che vadano a monitorare i rispettivi servizi di terze parti ed identificare:
    - Numero di chiamate in errore (HTTP Status Code)
    - Numero di chiamate con duration elevata
    - Numero di chiamate che hanno generato Exception