



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di Ingegneria

Corso di Laurea in Ingegneria Informatica

Software Architecture and Methodologies

&

Quantitative Evaluation of Stochastic Models

Quarkus Car Rental

Edoardo Sarri

7173337

Settembre 2025

Indice

1	Introduzione	7
2	Car Rental Application	8
2.1	Architettura	8
2.2	Servizi Esterni	9
2.3	Comunicazione	10
3	Deployment	12
3.1	Environment	12
3.2	Servizi Esterni	14
3.2.1	Helm	14
3.2.2	Deployment con Helm	15
3.3	Health	15
3.3.1	MicroProfile Health	15
3.3.2	Implementazione	16
4	Tracing and Metrics	17
4.1	Tracing	17
4.1.1	OpenTelemetry	17
4.1.2	Collector	18
4.1.3	Jaeger	18
4.1.4	OTel workflow	19
4.2	Metrics	19
4.2.1	Prometheus	20
4.2.2	Grafana	20
5	Workflow	22
5.1	Blocchi	22
5.1.1	XOR	23
5.1.2	AND	23
5.2	Implementazione	23
5.2.1	XOR	23
5.2.2	AND	24
5.3	Nuovo workflow	25

6	Analisi E2E	27
6.1	Eulero	27
6.1.1	Workflow complessi	27
6.1.2	Tool e analisi bottom-up	28
6.1.3	Dati necessari	28
6.2	Tracce	28
6.2.1	Generazione	29
6.2.2	Elaborazione	29
6.3	Distribuzioni	29
6.3.1	Approssimanti di Whitt	30
6.3.2	Risultati	30
6.4	Distribuzione E2E	31

Elenco delle figure

2.1	<i>car-rental</i> users use cases [9].	8
2.2	<i>car-rental</i> communication [9].	11
4.1	OpenTelemetry workflow.	19
4.2	Prometheus-Grafana Flow.	20
5.1	Workflow iniziale di Car-Rental.	22
5.2	Nuovo workflow di reserve mostrato da Jaeger.	25
5.3	Sequence diagram dei blocchi XOR e AND.	26
6.1	Statistiche dell'execution time di circa 1500 chiamate del workflow. . . .	30
6.2	Approssimazioni di Whitt che meglio fittano i dati del nostro workflow.	31

Elenco delle tabelle

Listings

3.1	Kubernetes and Docker Configuration	13
3.2	Reference in <i>users-service</i>	14
3.3	MySQL Helm chart	15
3.4	JDBC configuration.	16
4.1	Jaeger configuration for <i>users-service</i>	18
4.2	Install Jaeger chart for <i>users-service</i>	18
4.3	Prometheus configuration for <i>users-service</i>	20

1 Introduzione

- nel libro si spiega tutto ma per la dev mode. raramente ci sono esempi per la distribuzione dell'app.

2 Car Rental Application

L'applicazione utilizzata durante l'intero progetto è stata *acme-car-rental*. Si tratta di un'applicazione a micro servizi sviluppata all'interno del libro *Quarkus in Action* [9]. Esso mostra le caratteristiche principali di Quarkus e sviluppa le varie funzionalità dell'applicazione capitolo per capitolo, rilasciando su GitHub la versione finale e completa alla fine.

Lato utente l'applicazione permette, come si vede in Figura 2.1, di svolgere le attività di base che ogni applicazione di noleggio auto dovrebbe implementare: ricerca di auto libere in un determinato intervallo di tempo; prenotazione di un'auto definito tale intervallo; ricerca delle auto prenotate; aggiunta e rimozione di una'auto dall'inventario della auto disponibili.

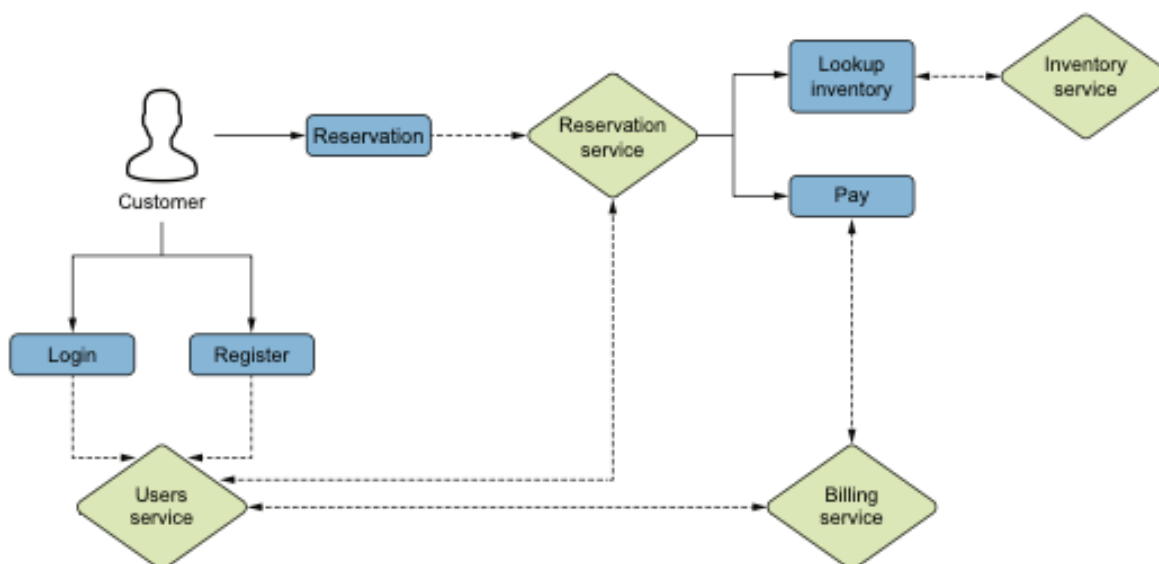


Figura 2.1: *car-rental* users use cases [9].

Oltre a quanto specificato in questo report e a quanto si trova nella *mia repo*, la versione originale rilasciata conteneva anche degli elementi che sono stati eliminati visto che per il nostro scopo non erano necessari.

2.1 Architettura

Vediamo come prima cosa l'architettura dell'applicazione. I micro servizi che *car-rental* definisce sono cinque:

- **Billing-service:**
Gestisce i pagamenti e le fatture. Viene chiamato da *reservation-service* quando un utente effettua una prenotazione e da *rental-service* quando un'auto viene noleggiata.
- **Inventory-service:**
Gestisce l'inventario delle auto. Fornisce l'elenco delle auto disponibili e permette agli impiegati di aggiungere o rimuovere veicoli.
- **Rental-service:**
Gestisce il processo di noleggio effettivo. Un impiegato può avviare un noleggio attraverso questo servizio, che a sua volta interagisce con il *billing-service*.
- **Reservation-service:**
Gestisce le prenotazioni delle auto. Verifica la disponibilità dei veicoli tramite l'*inventory-service* e, in caso di prenotazione, avvia il processo di pagamento tramite il *billing-service*.
- **Users-service:**
Fornisce una semplice interfaccia per la gestione delle prenotazioni.
Nel codice originale questo servizio si appoggiava a *Keycloak* per la gestione degli utenti. Per problemi di deployment e gestione, ho preferito rimuovere questa dipendenza; l'applicazione adesso permette a un solo utente, *guest*, di effettuare prenotazioni.

2.2 Servizi Esterni

Oltre ai servizi business dell'applicazione, descritti nella Sezione 2.1, essa si basa anche su dei servizi esterni, non sviluppati dai produttori ma che devono essere configurati e da cui i servizi principali dipendono. Questi sono:

- **Kafka:**
È una piattaforma di streaming di eventi distribuita, utilizzata per la comunicazione asincrona e mantenuta dalla Apache Software Foundation.
Gli eventi in Kafka vengono aggiunti in fondo a una coda e sono i subscriber che si devono occupare di gestire i messaggi già letti e quelli da leggere.
È ideale per streaming di dati di grandi dimensioni.
- **RabbitMQ:**
È un sistema di message broker che implementa il protocollo AMQP (Advanced Message Queuing Protocol), usato per la comunicazione asincrona.

Rispetto a Kafka, in RabbitMQ è il publisher che si occupa di distribuire i messaggi e che si assicura della loro ricezione.

È ideale per creare code di lavoro, dove ogni messaggio è un compito da eseguire.

- **MongoDB:**

È un database NoSQL orientato ai documenti, che memorizza i dati in file simili ai JSON (detti BSON).

È ideale per quelle applicazioni i cui requisiti sui dati non sono già stati ben definiti e che vogliono scalare orizzontalmente.

- **MySQL:**

È un database relazionale open-source.

È ideale per la memorizzazione di dati strutturati, per la rapidità e la semplicità d'uso.

- **PostgreSQL:**

È un database relazionale object-oriented open-source.

È più conforme agli standard SQL e offre un set di funzionalità per la gestione di query complesse più ricco rispetto a MySQL.

2.3 Comunicazione

Vediamo infine come i micro servizi, sia di business che esterni, comunicano tra loro. La Figura 2.2 illustra i vari flussi di comunicazione; da notare che *inventory-CLI* è uno di questi servizi rimossi rispetto alla versione rilasciata da Quarkus in Action, vista la sua inutilità nel nostro contesto.

- **REST:**

È lo stile architetturale di comunicazione più utilizzato nel web e si basa sul protocollo HTTP. È ideale quando si cerca una comunicazione sincrona tra servizi, dove un client invia una richiesta e attende una risposta.

Nell'applicazione viene usato ad esempio da *reservation-service* per interrogare *inventory-service* sulla disponibilità dei veicoli.

- **GraphQL:**

È un'alternativa più flessibile a REST, basata sempre sul principio client-server, che consente ai client di richiedere esattamente i dati di cui hanno bisogno; in questo modo il client non deve filtrare tutto quello che riceve.

Nell'applicazione viene usato da *users-service* per aggregare dati da diversi servizi con una singola richiesta, semplificando l'interazione dal lato client.

- **gRPC:**

Protocollo di comunicazione che si basa su HTTP/2, utilizza il formato binario, che permette di inviare più richieste contemporaneamente sullo stesso canale e di ricevere le relative risposte contemporaneamente. È ottimo nelle situazioni in cui si richiede una latenza minima con una comunicazione molto efficiente.

In questo progetto, *inventory-service* espone un endpoint gRPC che viene utilizzato da *reservation-service* per ottenere informazioni sulle auto in modo performante.

- **Kafka:**

Nell'applicazione viene usata per notificare eventi come l'inizio o la fine di un noleggio tra il *rental-service* e il *billing-service*. Questo garantisce una comunicazione resiliente anche in caso di fallimenti temporanei di uno dei due.

- **RabbitMQ:**

Viene usato da *reservation-service* per comunicare la creazione di una nuova prenotazione a *billing-service*, che si occuperà di creare la fattura.

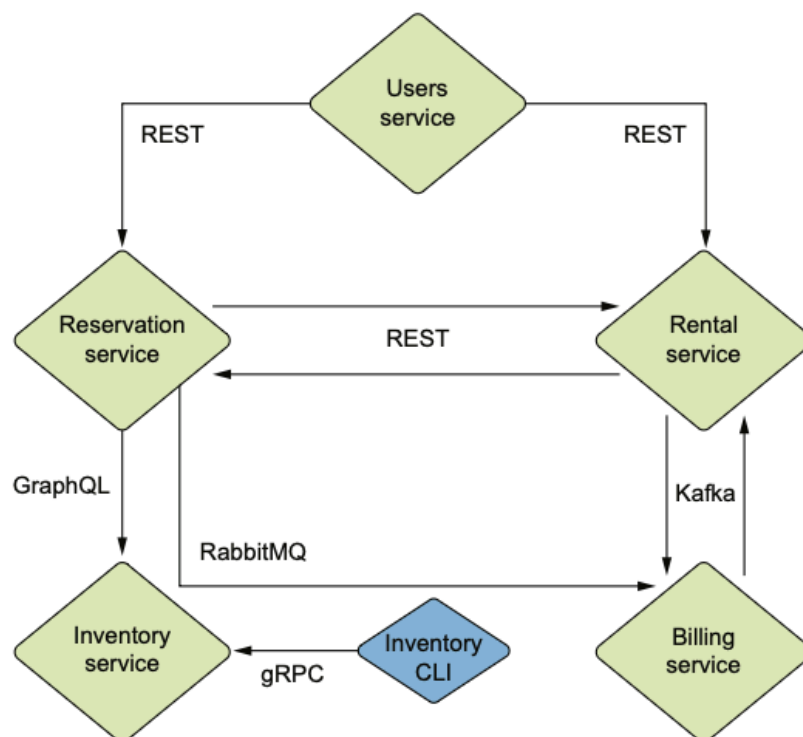


Figura 2.2: *car-rental* communication [9].

3 Deployment

In questo capitolo descriviamo il processo di deployment, più complesso di quello che ci aspettavamo. Nonostante infatti l'applicazione fosse già sviluppata in ogni suo aspetto, essa era predisposta per essere rilasciata su OpenShift e usare Quay.io come registry per le immagini.

Per questi motivi il file di configurazione Quarkus *application.properties* sono stati completamente rivisti e adattati alla nostra configurazione, che come abbiamo detto nel Capitolo 1 comprende Minikube come ambiente di deployment e Docker come runtime per i container.

Per automatizzare il deployment, nella root directory del progetto è presente una cartella *exec*, all'interno della quale è presente il file *deployment.sh*. Se eseguito a partire dalla root directory e con il Docker demon in esecuzione, allora verrà creato un cluster Minikube con tutti i micro servizi e le loro dipendenze. Il corretto funzionamento di questo script è garantito su MacOS con ARM64.

All'interno dello script viene usato, per ogni micro servizio, il comando *quarkus build*. Questo permette di costruire l'immagine Docker in automatico, senza scrivere a mano il manifesto Kubernetes; l'unico requisito è l'utilizzo dell'estensione *quarkus-container-image-docker*. Inoltre, cosa che in Minikube non è permesso, sarebbe consentito anche eseguire il deployment della nuova immagine appena costruita nell'ambiente di Deployment; questo può essere fatto tramite la configurazione *quarkus.container-image.push=true*

3.1 Environment

L'ambiente per il rilascio dell'applicazione è stato Minikube, un ambiente locale di sviluppo basato su Kubernetes. In questo modo è stato possibile testare il deployment in un ambiente simile a quello di produzione senza la necessità di risorse hardware dedicate. Come abbiamo detto, nonostante l'applicazione rilasciata su GitHub sia completa e configurata per un tipo di deployment, ci sono stati vari aspetti da considerare e che hanno richiesto attenzione e tempo:

- **OpenShift**

L'applicazione era stata pensata per essere distribuita su OpenShift, la piattaforma per container basata su Kubernetes e sviluppata da Red Hat. Nonostante essa sia appunto Kubernetes-based, in Quarkus si richiede di specificare quale ambiente di deployment si intende utilizzare all'interno del file di configurazio-

ne *application.properties*.

Un esempio di configurazione corretta per Kubernetes, usando Docker come runtime di container, è quella mostrata nel Listing 3.1: questo è il contenuto del file di configurazione Quarkus per *inventory-service*, ma generalizzando è lo stesso usato in tutti gli altri microservizi. Oltre ai comandi che sono naturalmente interpretabili, possiamo specificare alcuni concetti più complessi: `quarkus.container-image.build=true` permette di costruire l'immagine in automatico quando eseguiamo il comando `quarkus build`; `quarkus.kubernetes.service-type=NodePort` istruisce Kubernetes a creare un Service di tipo *NodePort* (di default è *ClusterIP*) e quindi accessibile anche dall'esterno del cluster (meno corretto in produzione di *LoadBalancer*, ma giusto per test locali); `quarkus.kubernetes.image-pull-policy=Never` istruisce Kubernetes a far fallire l'avvio del pod se l'immagine Docker su cui si basa non è presente localmente (a differenza di *Always* che la scarica sempre da un registry).

- **Registry**

Una volta costruita l'immagine docker, il file di configurazione era progettato per eseguire il push su un registry, Quay.io, registry di Red Hat che fa parte dello stesso ecosistema di OpenShift. Siccome il nostro ambiente per il deployment era Minikube, sono state fatte delle modifiche anche in questo senso in modo da rilasciare l'immagine del microservizio all'interno di Minikube stesso.

In questo scenario è fondamentale la configurazione `quarkus.container-image.push=false` all'interno del Listing 3.1.

- **Riferimenti**

Usando un ambiente di produzione diversi (OpenShift vs Minikube) rispetto a quello per cui l'applicazione era stata pensata, si sono dovuti cambiare gli URL con cui un microservizio identificava le proprie dipendenze. Un esempio di quello che è stato fatto si trova nel Listing 3.2, relativo al micro servizio *users-service*: i riferimenti per l'ambiente di sviluppo sono stati lasciati invariati, mentre sono state configurate correttamente le variabili d'ambiente all'interno del container in modo che sia presente il riferimento alla dipendenza usata.

```
# container
quarkus.container-image.build=true
quarkus.container-image.push=false
quarkus.container-image.group=edoardosarri
quarkus.container-image.name=inventory-service
quarkus.container-image.tag=2.0
# kubernetes
quarkus.kubernetes.name=inventory-service
```

```
quarkus.kubernetes.deployment-target=kubernetes
quarkus.kubernetes.service-type=NodePort
quarkus.kubernetes.image-pull-policy=Never
```

Listing 3.1: Kubernetes and Docker Configuration

```
# reference
%dev.quarkus.rest-client.reservations.url=http://localhost
:8081
%prod.quarkus.rest-client.reservations.url=http://reservation-
service
```

Listing 3.2: Reference in *users-service*

3.2 Servizi Esterni

Come abbiamo già detto nella Sezione 2.2, i micro servizi business dell'applicazione utilizzano al loro interno altri servizi ausiliari. Questi devono eseguire all'interno dello stesso cluster Minikube e quindi devono essere in un qualche modo rilasciati.

La prima soluzione esplorata è stata farsi scrivere da un LLM il manifesto Kubernetes per l'ambiente Minikube. Nonostante questa fosse una soluzione funzionante, mi sembrava chiaro che ci potesse essere un'alternativa off-the-shelf che mi permettesse di non dover fare configurazioni manuali; l'alternativa adottata è stata Helm.

L'unico servizio esterno per cui è stato generato (da un LLM, in particolare Gemini) il manifesto invece che usare un chart è stato MongoDB. Il problema, che non sono riuscito a risolvere, è stato quella della compatibilità tra la versione di MongoDB e quella di Minikube.

3.2.1 Helm

Helm è un gestore di pacchetti Kubernetes che semplifica il ciclo di vita delle applicazioni all'interno di Kubernetes.

Tramite i Chart possiamo definire, installare e aggiornare anche le applicazioni Kubernetes più complesse. Essi permettono infatti di non generare manifesti complessi e ridondanti a mano, ma di utilizzare una struttura predefinita e riutilizzabile. Nello stesso modo in cui esiste un database di immagini Docker, esiste un repository di Chart Helm, ArtifactHUB; oltre a fornirci molte applicazioni, il repository ci fornisce tutte le istruzioni necessarie per configurazioni avanzate [1].

3.2.2 Deployment con Helm

Per capire come Helm semplifica il deployment di servizi terzi da cui un nostro micro servizio dipende, prendiamo come esempio l'installazione del database MySQL, necessario per *inventory-service*. Tramite un comando, seguito da pochi parametri, come si vede nel Listing 3.3, viene rilasciato un pod MySQL nel cluster Minikube.

```
helm install mysql-inventory bitnami/mysql \
  --set auth.rootPassword=root-pass \
  --set auth.database=mysql-inventory \
  --set auth.username=user \
  --set auth.password=pass
```

Listing 3.3: MySQL Helm chart

3.3 Health

Durante la fase di deployment, e in particolare durante il deployment automatizzato con *deployment.sh*, è stato notato che il servizio *reservation-service* non funzionava correttamente. In particolare il pod partiva, andava in stato *running* e *ready*, ma non si riusciva a connettersi al suo database PostgreSQL. Dopo numerose e lunghe investigazioni, analizzando i log dei pod, il problema è stato risolto eliminando manualmente il pod: Kubernetes riavvia una nuova istanza e i problemi di connessione vengono risolti.

Questo ci può portare ad affermare che, per qualche motivo, il pod necessita che il database PostgreSQL sia disponibile e pronto per accettare connessioni prima di avviarsi correttamente.

3.3.1 MicroProfile Health

Per evitare di riavviare il pod manualmente o di introdurre nello script del deployment una *wait* è stato deciso di forzare il pod relativo a *reservation-service* ad attendere la disponibilità del database prima di avviarsi.

Questo viene fatto tramite la specifica Health di MicroProfile implementata poi da SmallRye; essa estende le specifiche di Jakarta per adattare quest'ultima all'architettura dei micro servizi. Permette di esporre lo stato di un'applicazione, cioè un valore binario (a differenza delle metriche che sono valori numerici qualunque), che viene solitamente controllato dall'orchestratore che gestisce il ciclo di vita delle applicazioni (e.g., Kubernetes) per decidere se un'applicazione è sana e può essere usata.

I tre maggiori controlli di salute che MicroProfile mette a disposizione sono *Startup*, *Readiness* e *Liveness* [5]: la prima verifica che l'applicazione sia stata avviata correttamente; la seconda controlla se l'applicazione è pronta a ricevere traffico; la terza monitora lo stato dell'applicazione durante il suo funzionamento. Se uno di questi controlli fallisce solitamente il pod viene riavviato.

3.3.2 Implementazione

Considerando che il nostro problema viene risolto facendo ripartire il pod del micro servizio una volta che il database è disponibile, l'Health Check utile è lo *Startup Check*. È stata definita la classe `DatabaseConnectionHealthCheck.java`, annotata con l'annotazione `org.eclipse.microprofile.health.Startup`, che implementa una `SELECT` sul database: se il comando ha successo allora il check termina con successo e il pod è in esecuzione correttamente, altrimenti il pod viene fatto ripartire.

Nella classe la connessione viene gestita da un oggetto di tipo `AgroalDataSource`^[7]: *DataSource* è un'interfaccia che rappresenta una connessione a un database standard in Java (JDBC); *Agroal* è un'implementazione di questa interfaccia ed è quella che funziona meglio con Quarkus.

Quarkus permette con molta facilità di gestire le metriche di Health con le apposite estensioni ^[8]. Basta infatti aggiungere *quarkus-smallrye-health*, *quarkus-agroal* e *quarkus-jdbc-postgresql* e il tutto funziona completamente. L'unica cosa aggiuntiva a cui dobbiamo pensare è a fornire l'url a cui l'oggetto `AgroalDataSource` deve connettersi; questo è facilmente configurabile nell'*application.properties*, come si vede dal Listing 3.4. È stata aggiunta anche la configurazione ^[6] per gestire dopo quanti tentativi falliti riavviare il pod.

```
# health
quarkus.kubernetes.startup-probe.failure-threshold=1
quarkus.datasource.jdbc.url=jdbc:postgresql://postgresql-
reservation:5432/reservation
```

Listing 3.4: JDBC configuration.

4 Tracing and Metrics

4.1 Tracing

In applicazioni con migliaia di micro servizi, seguire il workflow, cioè la cascata di chiamate, o valutare in che punto si è verificato un errore può essere molto complesso. Tramite il tracing non solo possiamo capire quale micro servizio viene chiamato e da chi, ma possiamo anche analizzare i tempi con cui ogni richiesta è stata servita, sia E2E che all'interno del singolo servizio.

Gli elementi principali che compongono un sistema di tracing sono: ogni singola unità di lavoro (i.e., chiamata a un microservizio) rilevata dal sistema è detta *span*; l'intera operazione E2E è composta da più span ed è detta *traccia*; ad ogni span è associato un ID univoco che rappresenta la traccia a cui appartiene; i diversi span che compongono una traccia hanno una relazione gerarchia, per cui si parla di span padre e/o figlio. Il funzionamento solitamente è abbastanza semplice: durante le varie chiamate nel workflow viene trasmesso l'ID della traccia; il trasporto di questo ID avviene tramite una qualche funzionalità del protocollo di trasporto, che in HTTP è il relativo header; quando la chiamata ritorna allora vengono unite gli span con lo stesso ID e viene generata la traccia.

4.1.1 OpenTelemetry

OpenTelemetry, noto anche come OTel, è un insieme di tool e API che permette di collezionare ed esportare le telemetrie di un'applicazione a microservizi. Oltre alle telemetrie in realtà possono anche essere gestite metriche e logs, ma in questo campo non è stabile e si preferisce usare altro (e.g., MicroMeter).

La vera forza di OpenTelemetry però non è tanto la raccolta di tracce, quando la definizione di OTLP, un protocollo standard che permette di esportare le telemetrie verso vari e diversi backend (e.g., Jaeger per la visualizzazione).

Per permettere il funzionamento di Opentelemetry con un qualche backend ci sono due possibili:

- **Direttamente:** OpenTelemetry una volta raccolte le tracce le invia al backend in questione. È una strategia più semplice, ma non scala quando i backend aumentano e se dobbiamo fare un filtraggio delle tracce.
- **Collector:** I dati vanno all'Otel-Collector, che è il responsabile dell'invio e di operazioni intermedie (e.g., filtraggio, il batching e la gestione dei fallimenti). È la strategia più complessa, però in applicazioni che devono scalare è necessario.

4.1.2 Collector

Nella nostra applicazione è stato usato il collector perché si voleva inviare le tracce al backend di Jaeger e anche salvarle in un file per eseguire ulteriori analisi in un secondo momento.

La raccolta delle tracce con Quarkus diventa estremamente semplice: aggiungendo l'estensione *quarkus-opentelemetry* il framework instrumenta il codice e in automatico le telemetrie vengono collezionate. Una volta che le tracce sono raccolte vengono inviate al collector, dopo la configurazione ne Listing 4.1 (relativo a *users-service*), dove l'unica configurazione da spiegare è forse `quarkus.otel.traces.sampler=always_on`: essa permette di non fare nessun campionamento delle chiamate, ma di considerarle tutte.

Il collector è un pod all'interno dell'ambiente Minikube che è stato deployato tramite una **configurazione helm**. Quando riceve le tracce le invia a due destinazioni: il pod di Jaeger per la visualizzazione e il file `/traces/traces.json` che si trova nel PVC (Persistent Volume Claim) *pvc-traces* sempre all'interno del cluster. Quando eseguiremo l'analisi dei dati allora dovremmo eseguire l'analyser in un pod a cui è collegato questo PVC; in questo modo lo script esegue sui propri dati locali.

```
# jaeger
quarkus.otel.service.name=users-service
quarkus.otel.exporter.otlp.traces.endpoint=http://otel-collector:4317
quarkus.otel.traces.sampler=always_on
```

Listing 4.1: Jaeger configuration for *users-service*

4.1.3 Jaeger

Come abbiamo detto Jaeger è un backend per la visualizzazione del work di una chiamata in un'applicazione a microservizi.

Come per i servizi terzi elencati nella Sezione 2.2, anche il backend Jaeger deve essere rilasciato nello stesso cluster Minikube di tutta l'applicazione. Anche in questo caso è stato utilizzato il chart Helm Jaeger [4]. Si può vedere il comando nel Listing 4.2.

```
helm install jaeger jaegertracing/jaeger \
  --set allInOne.enabled=true \
  --set agent.enabled=false \
  --set collector.enabled=false \
  --set query.enabled=false \
  --set provisionDataStore.cassandra=false \
  --set storage.type=memory
```

Listing 4.2: Install Jaeger chart for *users-service*

4.1.4 OTel workflow

Fino a questo momento in questa sezione abbiamo parlato dei componenti in modo separato; vediamo adesso di mettere tutto insieme per capire come i vari pod interagiscono nella nostra applicazione.

Analizziamo il flusso mostrato in Figura 4.1:

- Viene chiamata una qualche API dell'applicazione.
- OpenTelemetry raccoglie le telemetrie e invia le tracce al collector, che eventualmente fa qualche operazione.
- Il collector invia le tracce ai vari backend. Nel nostro caso Jaeger per la visualizzazione e il file system per l'analisi successiva.

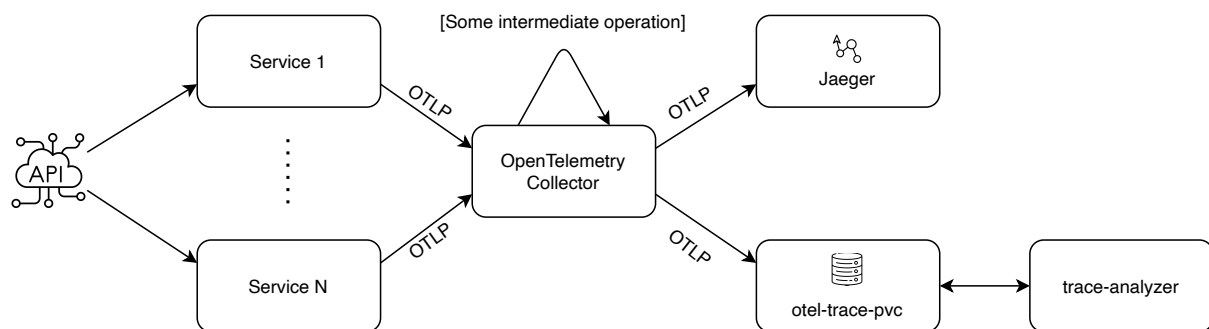


Figura 4.1: OpenTelemetry workflow.

4.2 Metrics

Oltre a tracciare la sequenza delle varie chiamate, come abbiamo mostrato nel Capitolo 4.1, in un'applicazione a microservizi è importante anche essere in grado di catturare metriche, valori reali che esprimono una qualche informazione.

In generale possiamo esporre metriche framework o business: le prime sono standard fornite da un qualche framework e sono generali e valide in un qualunque ambito (e.g., carico della cpu, uso della memoria o richieste per secondo); le seconde sono metriche specifiche per il dominio dell'applicazione.

La specifica MicroProfile per le metriche si chiama Metrics. Nonostante questo solitamente viene usata MicroMeter, un'implementazione che non segue la specifica MicroProfile. Quarkus permette in ogni caso di usare sia implementazioni della specifica Metrics che MicroMeter, nonostante la seconda sia quella preferibile.

4.2.1 Prometheus

Promethues è un sistema open-source per il monitoraggio e allerta di applicazioni. Svolge tre compiti principali: raccoglie metriche, le memorizza e fornisce un endpoint utile per fare querying.

In Quarkus utilizzare MicroMetrics e Prometheus insieme è molto semplice: dobbiamo solo aggiungere l'estensione *quarkus-micrometer-registry-prometheus*: senza strumentalizzare il codice, MicroMeter inizia a raccogliere metriche standard (e.g., utilizzo CPU e della memoria) nel formato richiesto da Prometheus, mentre quest'ultimo periodicamente interroga gli endpoint creati, raccogliendo e salvando tali metriche.

Dobbiamo a questo punto definire un modo per dire a Prometheus di raccogliere le metriche che MicroMetrics definisce. In Kubernetes questo metodo è detto *ServiceMonitor* e può essere semplicemente implementato nel file di configurazione, come vediamo nel Listing 4.3

```
# prometheus and grafana
quarkus.kubernetes.prometheus.generate-service-monitor=true
quarkus.kubernetes.labels.release=prometheus
```

Listing 4.3: Prometheus configuration for *users-service*

Infine serve installare nel cluster Minikube un container in cui esegue il server Prometheus. Questo è stato fatto, come nel resto del progetto, tramite Helm e il suo chart c. Questo non solo installa Prometheus, ma è un singolo pacchetto che comprende anche Grafana.

4.2.2 Grafana

Il sistema di query di Prometheus permette di visualizzare le metriche raccolte in forma testuale tramite terminale. Quando le metriche sono complesse, ma anche in generale, questo può non essere molto comodo.

Grafana è un backend che permette di visualizzare in modo grafico le metriche raccolte da Prometheus, come si può vedere nella Figura 4.2

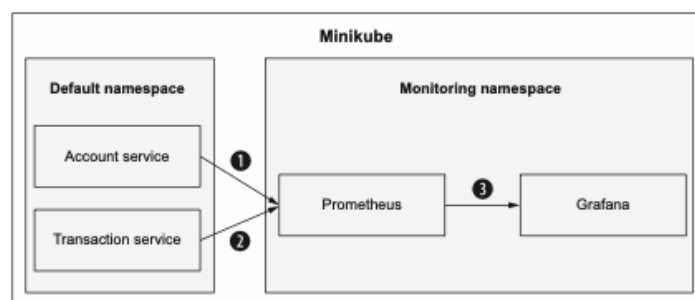


Figura 4.2: Prometheus-Grafana Flow.

Come dette nella Sezione 4.2.1 l'installazione all'interno del cluster Minikube è stata fatta con il chart di Helm *kube-prometheus-stack* [3], che comprende sia Prometheus che Grafana.

5 Workflow

Nei Capitoli 3 e 4 abbiamo visto come l'applicazione car rental, descritta nel Capitolo 2, viene rilasciata su Minikube e come possiamo analizzare metriche e workflow. In questo capitolo ci concentriamo invece sul workflow, cioè su come i microservizi si chiamano durante l'esecuzione di una funzionalità; in particolare la funzionalità che andremo ad esaminare sarà la prenotazione di una macchina.

Prima di qualunque modifica, come si nota dalla Figura 5.1, *Car-Rental* aveva un workflow esclusivamente sequenziale: quando un utente invocava una qualunque operazione, e in particolare *reserve*, tramite *users-service*, la sequenza di chiamate agli altri microservizio era puramente sincrona e deterministica, cioè era sempre la stessa data l'operazione chiamata; ovviamente i tempi di esecuzione variano di chiamata in chiamata.

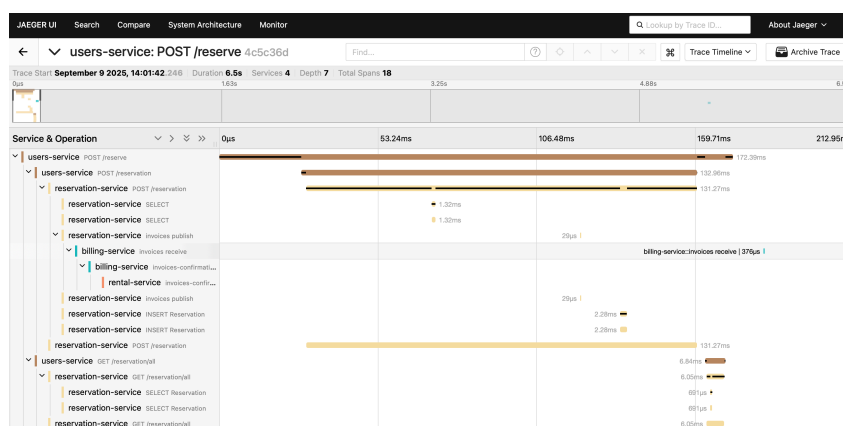


Figura 5.1: Workflow iniziale di Car-Rental.

Il nostro obiettivo è, facendo riferimento sempre all'operazione *reserve*, complicare questo workflow. Facendo riferimento alla terminologia usata nell'articolo del tool Eule-ro [2], che useremo più avanti, l'applicazione prima della modifica combina solamente blocchi di tipo sequenziale (SEQ), mentre a noi interessa aggiungere degli blocchi XOR e blocchi AND, che sono descritti nella Sezione 5.1 e implementati nella Sezione 5.2.

5.1 Blocchi

Vediamo come prima cosa quali blocchi sono stati aggiunti e in che punto dell'attuale workflow sono stati eseguiti.

5.1.1 XOR

I blocchi XOR rappresentano scelte casuali esclusive definite da una certa probabilità. Questo vuol dire che se abbiamo n possibili strade che il flusso può prendere, ne sarà scelta una e una soltanto una in base alla probabilità assegnata.

5.1.2 AND

I blocchi AND rappresentano esecuzioni indipendenti che possono essere eseguite in parallelo. Questo vuol dire che se abbiamo n strade, tutte saranno eseguite contemporaneamente e si sincronizzeranno alla fine.

5.2 Implementazione

Adesso che abbiamo visto quali blocchi vogliamo aggiungere, vediamo come sono stati implementati e in che punto del workflow vengono eseguiti.

La cosa importante da osservare è che l'obiettivo di questa modifica non è quello di aggiungere funzionalità all'applicazione, ma solamente quello di complicare il workflow. Per questo motivo ogni blocco aggiunto sarà composto da microservizi che non fanno altro che eseguire una busy wait per un tempo campionato da una distribuzione esponenziale; la busy wait è implementata con un ciclo `while` che controlla il tempo trascorso.

5.2.1 XOR

Il blocco XOR, che si trova nella cartella **XOR** su GitHub, implementa la scelta esclusiva tra tre microservizi. Questo blocco è inserito nel flusso di *users-service*: quest'ultimo microservizio chiama quello responsabile della scelta randomica e una volta che il blocco XOR termina riprende il controllo; in questo modo al codice dell'applicazione precedente è stata solo aggiunta la chiamata a questo blocco.

Il flusso del blocco XOR implementato è il seguente:

- L'inizio del blocco è definito da *start-choice-service*. Questo esegue una busy wait per un tempo campionato da una distribuzione esponenziale con tasso 1. Esegue poi una scelta randomica tra tre microservizio, dove: il primo ha probabilità 0.2, il secondo 0.5 e il terzo 0.3. Questo momento è utile per osservare un altro vantaggio di Quarkus: le probabilità con cui sono definite le scelte non sono build-in nel codice, ma sono definite nel file *application.properties*; questo permette di modificarle senza dover mettere mano al codice stesso.
- I tre microservizi chiamati sono *first-choice-service*, *second-choice-service* e *third-choice-service*. Ognuno di questi esegue una busy wait per un tempo campionato

da una distribuzione esponenziale con tasso 1. Al termine di uno di questi il controllo ritorna a *users-service*

5.2.2 AND

Il blocco AND, che si trova nella cartella **AND** su GitHub, implementa l'esecuzione parallela di due servizi chiamati dallo stesso microservizio client; i due server eseguono contemporaneamente e si sincronizzano nel client.

Anche questo blocco è inserito nel flusso di *users-service*, esattamente dopo la chiamata al servizio *start-choice*: il microservizio chiama *start-parallel* che è il responsabile del fork; una volta che il blocco AND termina riprende il controllo; in questo modo al codice dell'applicazione precedente è stata solo aggiunta la chiamata a questo blocco.

Per l'implementazione del blocco AND è stato usato gRPC, già descritto nella Sezione 2.3. In questo caso ha fatto la differenza l'utilizzo del framework Quarkus, che fornisce l'integrazione nativa con gRPC e permette di definire servizi asincroni in modo semplice.

La pipeline usata per implementare la comunicazione asincrona di gRPC con Quarkus è la seguente:

- Il server dichiara un file *.proto*, che definisce i servizi offerti e i messaggi scambiati. Questo permette di lavorare in modo API-first: si definisce cosa il server espone e si lavora a partire da questo.
- Con l'estensione *quarkus-grpc*, quando si chiama `quarkus build`, il framework genera automaticamente tutto quello che ci serve per implementare il server. In particolare la classe che espone il servizio deve implementare un'interfaccia e sovrascrivere i metodi che nel file *.proto* sono i messaggi scambiati.
- Il client deve memorizzare il file *.proto*, compilarlo in modo che Quarkus generi lo stub del server e intrumentare il codice in modo da poter chiamare i metodi esposti dal server come se fossero locali.

gRPC consente quindi di implementare una comunicazione asincrona tra client e server tramite canali asincroni. Questo si può fare usando l'interfaccia nativa di gRPC o le API di SmallRye.

In gRPC Java classico questi canali sono gestiti dall'interfaccia `StreamObserver<T>`, dove T è il tipo del messaggio; ogni comunicazione è quindi un oggetto che implementa tale interfaccia. Quando si invia un messaggio tramite il canale (usando il metodo `onNext(T message)`), gRPC avvia un thread del suo pool di thread dedicati, il quale si occupa della gestione. Questa interfaccia ha però numerosi limiti: è imperativa e quindi tutto deve essere gestito a mano; si basa su callback, cioè molte funzioni

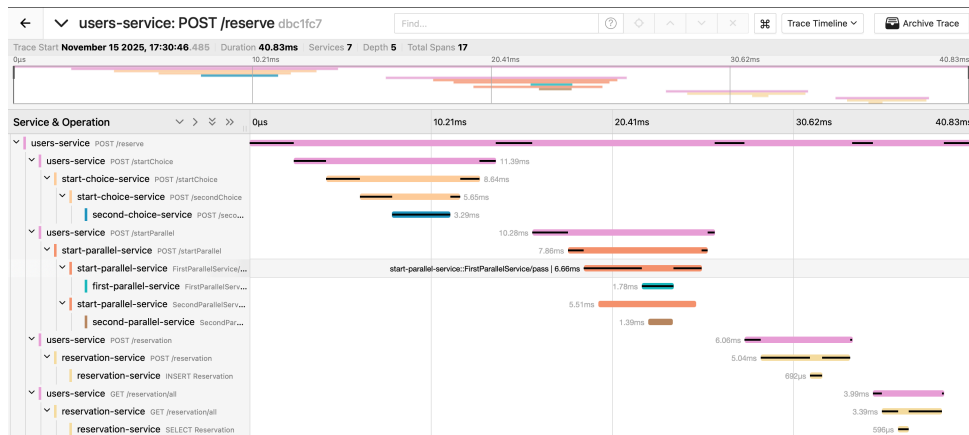


Figura 5.2: Nuovo workflow di reserve mostrato da Jaeger.

vengono passate come parametro (uno *StreamObserver* deve implementare tre metodi) e saranno chiamate in un secondo momento; è difficile da testare; non permette di gestire più risposte come un unico oggetto.

Quarkus, ovviamente, non utilizza questa interfaccia, ma sfrutta quella di *SmallRye*, *Mutiny*. Si tratta di una libreria reattiva e dichiarativa: si deve definire come il sistema si comporta quando i dati arriveranno; non dobbiamo dire come fare le cose ma solo cosa vogliamo ottenere (e.g., nello stesso modo delle *lambda*). L'interfaccia di *Mutiny* si basa su due tipi, *Uni<T>* e *Multi<T>*: il primo permette l'invio di un solo messaggio, mentre il secondo consente gli stream. I vantaggi che si ottengono sono: codice più leggibile; composizione di chiamate diverse facile da implementare; facile da testare e debuggare.

5.3 Nuovo workflow

Nel nuovo workflow, oltre ad includere i blocchi XOR e AND, è stata eliminata la parte di fatturazione. Il motivo di questa scelta è in dovuto al nostro obiettivo che ci ha portato a complicare questo workflow: vogliamo valutare il tempo E2E di una prenotazione. La prenotazione e la fatturazione sono in car-rental due operazioni indipendenti: una prenotazione ha successo, e quindi ritorna all'utente, indipendentemente dal completamento della chiamata di *reservation-service* a *billing-service*.

Usando ancora Jaeger per visualizzare la traccia della chiamata a *reserve*, possiamo notare dalla Figura 5.2 la presenza dei blocchi XOR e AND e l'esclusione dei microservizi *billing-service* e *rental-service*. In Figura 5.3 è mostrato invece il sequence diagram UML.

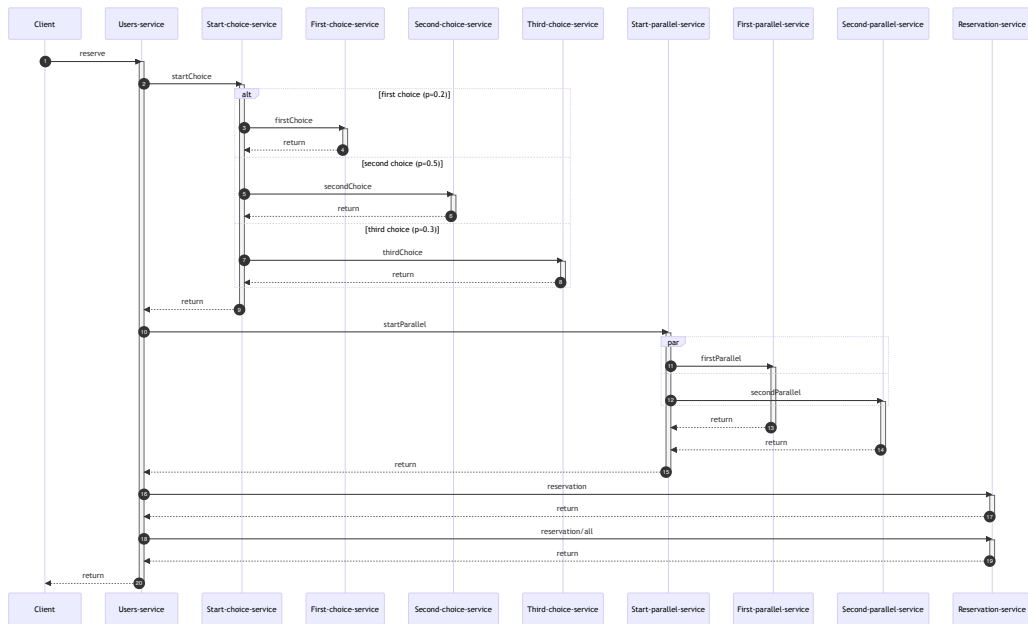


Figura 5.3: Sequence diagram dei blocchi XOR e AND.

6 Analisi E2E

In questo capitolo andremo ad eseguire un'analisi del tempo end to end (E2E) di una prenotazione da parte dell'utente, cioè dell'intervallo di tempo che trascorre dall'inizio della chiamata della funzionalità fino al suo termine, usando un approccio data-driven: supporremo di non avere il modello del workflow e usaremos i dati raccolti tramite il tracing per arrivare a delle conclusioni. Questo approccio è molto utile in quelle situazioni in cui il nostro workflow è composto da moltissimi microservizi, ma non abbiamo un'idea precisa dei tempi di esecuzione e delle relazioni tra di essi.

Nei casi in cui si voglia garantire un tempo di risposta medio di una funzionalità, l'analisi E2E può essere fatta semplicemente osservando media e varianza dei tempi di esecuzione dei servizi che compongono tale funzionalità. Ci sono delle situazioni in cui però questo non è sufficiente e si vuole garantire un determinato Service Level Agreements (SLA), cioè si vuole lavorare con un upper-bound sul tempo di esecuzione. Il nostro obiettivo è quindi più stringente rispetto al tempo medio: ogni chiamata a una qualche funzionalità deve sempre ritornare entro un tempo all'interno di una soglia. Questo richiede non solo di avere dati statistici descrittivi, ma di essere in possesso dell'intera Probability Density Function (PDF) o della Cumulative Density Function (CDF) del workflow che ci interessa analizzare.

Un esempio di metrica che possiamo ottenere seguendo questo approccio e basandoci sulla PDF è la probabilità che il tempo di esecuzione sfiori una data deadline T_{max} : semplicemente si ottiene integrando la PDF in $[T_{max}, +\infty]$. Questo è un buon esempio di osservazione che non potremmo fare con le sole statistiche descrittive.

6.1 Eulero

Per ottenere una distribuzione del tempo E2E di un workflow ci sono due tecniche: la prima è un'analisi numerica esatta, fattibile quando il workflow non è troppo complesso e i tempi di esecuzione seguono distribuzioni markoviane (i.e., esponenziali); la seconda è un'approssimazione basata su un approccio compositivo, che permette di analizzare anche workflow complessi con distribuzioni generali (GEN). Eulero utilizza questo secondo approccio bottom-up.

6.1.1 Workflow complessi

Quando un workflow è complesso dobbiamo ricorrere a tecniche approssimate. La complessità di un workflow nella vita reale deriva dal fatto che i tempi di esecuzione dei microservizi non seguono distribuzioni esponenziali, cioè markoviane.

Quando i tempi di esecuzione sono esponenziali, l'unica cosa che condiziona il futuro del sistema è lo stato attuale. I possibili percorsi che il sistema può prendere da un certo punto in poi non considerano il passato e non considerano il tempo trascorso in quello stato. Quando però introduciamo distribuzioni non markoviane, il futuro viene condizionato anche dal passato. Questo provoca un'esplosione dello spazio degli stati e rende l'analisi numerica esatta impossibile.

6.1.2 Tool e analisi bottom-up

Visto quanto detto sopra, per ottenere la PDF del tempo di esecuzione di un servizio abbiamo usato Eulero[2], un tool dell'STLAB dell'università di Firenze che esegue un'analisi bottom-up tramite un metodo composizionale: si divide il workflow in sub-workflow, fino ad unità elementari; si calcolano le PDF di queste unità; si compongono le varie PDF per ottenere la PDF del workflow completo.

Il vantaggio di questo approccio è la riduzione della complessità: si elimina infatti il condizionamento del futuro dagli stati precedenti anche per distribuzione *GEN* e questo porta a una riduzione del numero degli stati che devono essere analizzati. Il lato negativo è che si introduce un'errore di approssimazione nel tempo E2E calcolato.

6.1.3 Dati necessari

Senza scendere nel dettaglio, Eulero supporta workflow con strutture di controllo sequenziali (SEQ), parallele (AND), condizionali (XOR) e modellate con Direct Acyclic Graph (DAG). Le strutture parallele e condizionali sono quelle descritte nella Sezione 5.1 e che sono state poi implementate in car-rental.

Quello che dobbiamo fornire al tool per ottenere la distribuzione del tempo E2E è: la struttura del workflow modellata tramite blocchi, eventualmente annidati; la distribuzione che rappresenta il tempo di esecuzione di ogni blocco elementare. Osserviamo che i blocchi e la loro composizione possono essere osservati dalla Figura 5.2 e dalla Figura 5.3; le distribuzioni sono invece ottenute nella Sezione 6.3

6.2 Tracce

Nella Sezione 4.1 abbiamo descritto il funzionamento del collector di OpenTelemetry e come questo lavora all'interno della nostra applicazione. In questo momento ci interessa lavorare sulle tracce che il collector salva nel Persistent Volume Claim (PVC) *otel-traces-pvc* e che sono accessibili dal pod *analyzer*; l'obiettivo è eseguire molti esperimenti, ottenere i tempi di esecuzione delle chiamate ai microservizi che poi possiamo usare per estrarre i dati necessari per Eulero.

6.2.1 Generazione

Per il nostro obiettivo ci servono molti dati e quindi molte tracce da analizzare. Siccome generare queste tracce a mano non è molto pratico, è stato usato K6 come generatore di carico.

Il generatore di carico K6 può essere installato localmente su una macchina Linux, MacOS o Windows, ma ha anche il grande vantaggio di poter essere eseguito all'interno di un'infrastruttura cloud. Nel nostro progetto è utile e sensato distribuire K6 all'interno del cluster Minikube dove è in esecuzione la nostra applicazione; in questo modo può chiamare ogni risorsa di ogni servizio senza esporre tale servizio all'esterno del cluster. Per distribuire K6 in un cluster Kubernetes ci sono diversi modi, ma nel nostro caso è stato usato, come in gran parte del progetto, Helm. Su GitHub è possibile vedere l'installazione del **job di K6**: Helm utilizza la cartella *templates/* per generare le risorse Kubernetes; il file *templates/configmap.yaml* serve per iniettare lo script che il job dovrà eseguire; il file *job.yaml* crea una Job Kubernetes che esegue tale script.

L'implementazione questo job è molto semplice: logicamente si tratta di un ciclo *for* che esegue per un certo intervallo di tempo specificato chiamando l'entry point del servizio di prenotazione. Prenota sempre la stessa macchina spostando l'intervallo della prenotazione in avanti di un giorno a ogni iterazione.

6.2.2 Elaborazione

Le tracce sono salvate in un singolo file *traces.json*, che all'interno del pod *analyzer* si trova nella directory *data*. Si tratta di un JSON Lines, cioè un file dove ogni linea è un oggetto JSON indipendente.

L'elaborazione avviene portando i dati dal pod in memoria locale e poi eseguendo uno **script** costruito in modo da leggere ogni riga del file e analizzarla come json. L'output dell'analisi, come possiamo vedere in Figura 6.1, sono delle statistiche descrittive (i.e., minimo, massimo, media, varianza e coefficiente di variazione) dei tempi di esecuzione di ogni blocco. Queste statistiche saranno poi necessarie per ottenere le distribuzioni del tempo di esecuzione di ogni blocco del workflow.

6.3 Distribuzioni

Per eseguire Eulero e ottenere la nostra PDF (o CDF) del tempo E2E del workflow ci servono le distribuzioni dei tempi di esecuzione di ogni blocco di tale workflow. Al momento, come descritto nella Sezione 6.2, abbiamo le statistiche descrittive dei tempi di esecuzione di ogni operazione che compone il workflow. Dobbiamo a questo punto utilizzare questi dati per ottenere delle distribuzioni che abbiano il miglior fitting sui nostri dati osservati tramite gli esperimenti.

6.3.1 Approssimanti di Whitt

Se abbiamo una distribuzione del tempo di esecuzione di un servizio, allora possiamo ottenere i dati semplicemente campionando da tale distribuzione. La nostra situazione è però l'opposta: abbiamo i tempi di esecuzione e ci serve una distribuzione che fitti bene i dati.

Una prima approssimazione può essere usando una distribuzione esponenziale, che inoltre, come abbiamo detto nella Sezione 6.1.1, ci semplificherebbe la vita da un punto di vista dell'analisi. Sappiamo che una distribuzione esponenziale ha un coefficiente di variazione (CV) pari a 1; possiamo però vedere che dalle informazioni mostrate nella Figura 6.1 che nessuno dei blocchi del workflow ha un CV pari a 1, ma sono tutte maggiori di uno tranne la prima chiamata; questo implica che i nostri microservizi hanno una varianza maggiore di quella di una distribuzione esponenziale.

Service	Operation	Min (ms)	Max (ms)	Mean (ms)	Variance	Std Dev	CV
users-service	POST /reserve	12.888	786.823	30.005	848.737	29.133	0.971
start-choice-service	POST /startChoice	1.735	680.265	7.623	358.586	18.936	2.484
third-choice-service	POST /thirdChoice	0.387	66.357	2.257	20.595	4.538	2.011
users-service	POST /startParallel	1.380	99.777	5.424	60.994	7.810	1.440
start-parallel-service	POST /startParallel	2.836	242.163	11.773	302.806	17.401	1.478
first-parallel-service	FirstParallelService/pass	0.177	76.629	2.053	22.980	4.794	2.335
second-parallel-service	SecondParallelService/pass	0.163	62.763	1.955	17.488	4.182	2.139
users-service	POST /reservation	1.281	90.195	3.892	47.608	6.900	1.773
reservation-service	POST /reservation	1.307	167.913	4.022	69.307	8.325	2.070
users-service	GET /reservation/all	1.422	86.108	5.220	44.638	6.681	1.280
reservation-service	GET /reservation/all	1.574	159.313	5.765	77.672	8.813	1.529
second-choice-service	POST /secondChoice	0.359	73.874	2.127	13.521	3.677	1.729
first-choice-service	POST /firstChoice	0.394	84.829	2.296	31.756	5.635	2.455

Figura 6.1: Statistiche dell'execution time di circa 1500 chiamate del workflow.

In queste situazioni per trovare una distribuzione che approssimi bene i nostri dati si possono usare le approssimazioni di Whitt. Nel nostro scenario gli approssimanti che eseremo sono:

- HyperExp

Si usa per dati che hanno un coefficiente di variazione $C_V > 1$. Si tratta di uno XOR di due esponenziali, cioè di una distribuzione definita come $X = \begin{cases} Exp(\lambda_1) & \text{con prob } p \\ Exp(\lambda_2) & \text{con prob } 1 - p \end{cases}$; osserviamo che se $\lambda_1 = \lambda_2$ allora abbiamo un'esponenziale.

- HypoExp

Si usa per dati che hanno un coefficiente di variazione $C_V \in [\frac{1}{\sqrt{2}}, 1]$. Si tratta di una sequenza di esponenziali con tassi diversi, cioè di una distribuzione definita $X = Exp(\lambda_1) + \dots + Exp(\lambda_n)$, con $\lambda_i \neq \lambda_j$.

6.3.2 Risultati

I risultati, cioè le distribuzioni che meglio fittano i dati ottenuti in precedenza, possono essere calcolate in funzione del valore atteso e del coefficiente di variazione. Dopo aver

eseguito due **funzioni**, i risultati ottenuti sono quelli mostrati in Figura 6.2.

Service	Operation	distribution	p1	Lambda 1	p2	Lambda 2
users-service	POST /reserve	hypoExp	N/A	0.0343	N/A	1.1294
start-choice-service	POST /startChoice	hyperExp	0.9246	0.2426	0.0754	0.0198
third-choice-service	POST /thirdChoice	hyperExp	0.8884	0.7874	0.1116	0.0989
users-service	POST /startParallel	hyperExp	0.7955	0.2933	0.2045	0.0754
start-parallel-service	POST /startParallel	hyperExp	0.8050	0.1368	0.1950	0.0331
first-parallel-service	FirstParallelService/pass	hyperExp	0.9153	0.8916	0.0847	0.0825
second-parallel-service	SecondParallelService/pass	hyperExp	0.9004	0.9210	0.0996	0.1019
users-service	POST /reservation	hyperExp	0.8596	0.4417	0.1404	0.0721
reservation-service	POST /reservation	hyperExp	0.8942	0.4447	0.1058	0.0526
users-service	GET /reservation/all	hyperExp	0.7459	0.2858	0.2541	0.0974
reservation-service	GET /reservation/all	hyperExp	0.8165	0.2833	0.1835	0.0637
second-choice-service	POST /secondChoice	hyperExp	0.8530	0.8021	0.1470	0.1382
first-choice-service	POST /firstChoice	hyperExp	0.9229	0.8041	0.0771	0.0672

Figura 6.2: Approssimazioni di Whitt che meglio fittano i dati del nostro workflow.

A questo punto abbiamo tutto per eseguire l'analisi finale e calcolare la distribuzione dei tempi E2E del nostro workflow.

6.4 Distribuzione E2E

Bibliografia

- [1] *Bitnami Helm Charts for MySQL*. URL: <https://artifacthub.io/packages/helm/bitnami/mysql> (visitato il giorno 03/09/2025).
- [2] Laura Carnevali et al. «Compositional safe approximation of response time probability density function of complex workflows». In: *ACM Transactions on Modeling and Computer Simulation* 33.4 (2023), pp. 1–26.
- [3] Prometheus Community. *Kube Prometheus Stack*. 2025. URL: <https://artifacthub.io/packages/helm/prometheus-community/kube-prometheus-stack#configuration> (visitato il giorno 04/09/2025).
- [4] JaegerTracing. *Jaeger Helm chart*. URL: <https://artifacthub.io/packages/helm/jaegertracing/jaeger> (visitato il giorno 03/09/2025).
- [5] Kubernetes Authors. *Liveness, Readiness and Startup Probes*. 2025. URL: <https://kubernetes.io/docs/concepts/configuration/liveness-readiness-startup-probes/> (visitato il giorno 03/09/2025).
- [6] Quarkus - *All configuration options*. URL: <https://quarkus.io/guides/all-config> (visitato il giorno 04/09/2025).
- [7] Quarkus Team. *Configure data sources in Quarkus*. URL: <https://quarkus.io/guides/datasource> (visitato il giorno 03/09/2025).
- [8] Quarkus Team. *SmallRye Health - Quarkus*. URL: <https://quarkus.io/guides/smallrye-health> (visitato il giorno 03/09/2025).
- [9] Martin Štefanko e Jan Martiška. *Quarkus in Action*. <https://www.manning.com/books/quarkus-in-action>. Data ultima visualizzazione: 2026-07-30. Manning Publications, 2025.
- [10] Burr Sutter et al. *Kubernetes Native Microservices*. O'Reilly Media, 2020.