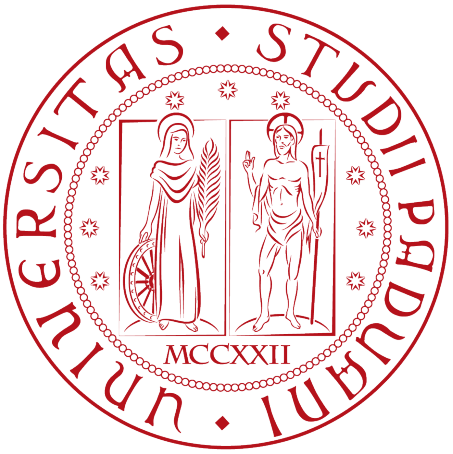
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea in Informatica



Introduzione al big data computing ed una sua applicazione

Tesi di laurea triennale

Relatore

Prof.Tullio Vardanega

Laureando

Stefano Panozzo

Anno Accademico 2017-2018

Perché sognare è importante, ma impegnarsi concretamente lo è ancora di più.

— Pan di Stelle - Mulino Bianco

Sommario

Il presente documento descrive il lavoro svolto durante il periodo di stage del laureando Stefano Panozzo presso l'azienda I.T. Euro Consulting S.r.l. di Padova. Lo stage è stato svolto alla conclusione del percorso di studi della Laurea Triennale ed è durato in totale 320 ore. Gli obiettivi da raggiungere erano molteplici.
La prima richiesta dell'azienda era analizzare la struttura del cluster in cui risiedevano i dati utilizzati in seguito. Successivamente era richiesta l'analisi e la trasformazione del dataset di interesse per estrarre ed ottenere nuove informazioni utili per creare un modello che prevedesse il target desiderato. Infine, era richiesta la progettazione e lo sviluppo di una web app per la rappresentazione dei risultati ottenuti in precedenza.
I primi due capitoli del presente documento hanno lo scopo di presentare il contesto aziendale in cui è stato sostenuto lo stage e di spiegare come il progetto di stage si renda utile all'interno della strategia aziendale. Il terzo capitolo documenta lo svolgimento dello stage descrivendo le attività che sono state portate a termine, i punti salienti del progetto stesso e le principali scelte progettuali. Il quarto ed ultimo capitolo presenta infine una valutazione dello svolgimento dello stage rispetto agli obiettivi aziendali e alle conoscenze acquisite dallo studente.

Ringraziamenti

Innanzitutto, vorrei esprimere la mia gratitudine al Prof. Tullio Vardanega, relatore della mia tesi, per l'aiuto e il sostegno fornitomi durante la stesura del lavoro.

Desidero ringraziare con affetto i miei genitori per il sostegno, il grande aiuto e per essermi stati vicini in ogni momento durante gli anni di studio.

Ho desiderio di ringraziare poi i miei amici per tutti i bellissimi anni passati insieme e le mille avventure vissute, che mi hanno fatto crescere e diventare la persona che sono ora.

Padova, Dicembre 2018 Stefano Panozzo

"Perché sognare è importante, ma impegnarsi concretamente lo è ancora di più."

— Pan di Stelle

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Il contesto aziendale

Esempio di utilizzo di un termine nel glossario
Application Program Interface (API).

Esempio di citazione in linea
site:agile-manifesto.

1.1 Il profilo aziendale

I.T. Euro Consulting S.r.l.[1] è un'azienda di medie dimensioni con sede legale a Padova, nata nel 2007 e facente parte del gruppo SCAI, presente su tutto il territorio italiano. Dalla sua nascita si è sempre occupata prevalentemente di consulenza, *System Integration* ed *Application Management*, in ambito ICT, operando in tutti i principali settori di mercato: bancario ed assicurativo, industria, pubblica amministrazione e servizi. Nel corso degli anni l'azienda ha consolidato le proprie conoscenze soprattutto nei seguenti ambiti, offrendo svariati servizi:

* **Big Data**: supporto alle aziende nel loro processo di crescita e cambiamento, tramite moderne soluzioni di *Business Intelligence* e la possibilità di prevedere scenari ed eventi futuri e prendere le più opportune decisioni operative o di business grazie all'analisi della gran mole di dati che ogni giorno vengono creati. Vengono quindi offerti servizi di *big data engineer*, *big data scientist*, *big data architect* e *big data administrator*;

* **Internet of Things**: soluzioni end to end, basate su tecnologie leader di mercato che consentono di indirizzare in modo efficace la realizzazione di sistemi IoT accelerando la realizzazione di componenti web e mobile per la raccolta, la visualizzazione e l'analisi dei dati;

* **Reference Architecture**: intesa come *best practice* e struttura di base per un insieme di domini applicativi all'interno di un'organizzazione, la quale agevola il continuo allineamento dei processi e delle strategie con le giuste soluzioni tecnologiche. Vengono quindi offerti servizi di *assessment*, design e consulenza;

[1] **https://www.itecons.it**.

- **DevOps**: automatizzazione delle attività manuali nelle diverse fasi del Software Development Lifecycle. Il modello DevOps non si concentra esclusivamente sull' introduzione di nuovi tool, ma è inteso come una combinazione di cultura, processi unita agli strumenti di automazione. Vengono quindi offerti servizi di *assessment* e consulenza;

- **System Integration**: servizi di consulenza o interventi progettuali per aiutare le aziende a gestire al meglio le proprie strutture tecnologiche complesse e soluzioni applicative per semplificare la coesione fra i vari sottosistemi che compongono la struttura;

- **Application Management**: servizi di manutenzione correttiva, adattativa ed evolutiva di soluzioni applicative durante il loro intero ciclo di vita;

- **Customer Relationship Management**: con l'obiettivo di ottenere una visione completa per perseguire uno scenario di *Single Customer View*, abilitante al dialogo one-to-one tra l'organizzazione ed il proprio cliente indipendentemente dalle canalità attraverso le quali avviene l'interazione;

- **System & Data Administration**: servizio consultivo svolto avvalendosi di un insieme di strategie, processi e regole che consentono di gestire i sistemi e trattare i dati fondamentali per lo sviluppo aziendale. Vengono quindi offerti servizi di *database administration*, *database security*, *data governance*, *data analysis* e *scheduling management*.

1.2 Tecnologie utilizzate

Le tecnologie utilizzate dall'azienda per la realizzazione dei propri prodotti si possono raggruppare in due macro-sezioni, ovvero riguardante lo sviluppo software e le attività inerenti ai *big data*. Per quanto concerne la prima, si può suddividere ulteriormente in due aree: *backend* e *frontend*.

1.2.1 Sviluppo software

Backend: buona parte del *backend* dei prodotti dell'azienda è scritta in linguaggio Java EE[2]. Questa scelta è dovuta al grande supporto offerto da questo linguaggio in fatto di controllo degli accessi e sicurezza di applicativi delicati come quelli in ambito bancario e assicurativo;

Frontend: per quanto riguarda il *frontend* è utilizzato prevalentemente il *framework* TypeScript Angular[3] in quanto offre una grande elasticità d'impiego e buone prestazioni.

1.2.2 Big Data

Hadoop[4]: *framework* utilizzato per la gestione del *cluster* e supporta applicazioni distribuite con elevato accesso ai dati, strutturati tramite il filesystem chiamato HDFS. Permette alle applicazioni di lavorare con migliaia di nodi e petabyte di dati;

[2] **https://www.oracle.com/technetwork/java/javaee**.
[3] **https://angular.io/**.
[4] **https://hadoop.apache.org/**.

Hive[5]: utilizzato per effettuare le *query* e l'analisi preliminare dei dati in *dataset* di grandi dimensioni;
Impala[6]: simile ad Hive, ma fornisce prestazioni leggermente migliori a discapito di una minor affidabilità e peggior gestione degli errori;
Spark[7]: *framework* per il calcolo distribuito di dati strutturati in un cluster. Supporta applicazioni scritte in molteplici linguaggi, quelli utilizzati in azienda sono principalmente Scala e Python;
R[8]: linguaggio di programmazione e un ambiente di sviluppo specifico per l'analisi statistica dei dati, utilizzato per la stima di modelli predittivi partendo dai dati ricavati utilizzando i precedenti strumenti;
Python[9]: grazie alla sua elasticità ed ai suoi svariati utilizzi, Python è utilizzato anche per scopi analoghi al precedente strumento. Essendo la sintassi di questo linguaggio molto semplice, è ultimamente preferito a R.

1.3 Processi aziendali

L'azienda svolge il suo operato in base alla tipologia di lavoro da effettuare. Oltre ad eseguire progetti per un possibile cliente, sono attivi progetti per lo sviluppo di nuovo prodotto, dopo aver effettuato le ricerche di mercato d'interesse, da consegnare poi al reparto marketing per trovare compratori interessati; spesso, inoltre, vengono attivati dei progetti in seguito all'interesse relativo ad alcune gare d'appalto, principalmente per il settore privato ma in passato anche per quello pubblico.

1.3.1 Metodologia di sviluppo

Sviluppo su commissione

Per quanto riguarda i progetti relativi allo sviluppo di un nuovo prodotto a seguito di una richiesta da parte di un cliente, vengono eseguite le seguenti attività, supervisionate durante tutta la durata del progetto da un *project manager* per coordinare i lavori ed interfacciarsi con il *management* dell'azienda:

1. **Analisi**: svolta in contemporanea dagli analisti e dal reparto marketing per il contatto con il cliente. Solitamente si cerca di partire da prodotti già sviluppati in azienda per poi personalizzarli in base alle richieste del cliente, così da poter offrire soluzioni già di base consolidate e testate in molteplici situazioni. Questo è preferibile soprattutto se le applicazioni sottostanti hanno bisogno di una maggior sicurezza, com'è il caso in ambito finanziario, anche per poter effettuare una manutenzione rapida in caso di malfunzionamenti.
   Il reparto marketing mantiene la maggior parte delle relazioni con il cliente finale all'inizio del rapporto, anche se un colloquio diretto è ovviamente necessario dopo le prime interazioni per poter adottare soluzioni più specifiche e tecnicamente più complesse in base alle necessità del cliente;

2. **Implementazione**: dopo aver identificato i requisiti assieme al cliente, il team incaricato si occupa dell'implementazione del prodotto concordato. Solitamente,

[5] **https://hive.apache.org/**.
[6] **https://impala.apache.org/**.
[7] **https://spark.apache.org/**.
[8] **https://www.r-project.org/**.
[9] **https://www.python.org/**.

per ogni progetto, sono assegnate un certo numero di persone e risorse per occuparsi del *backend* e del *frontend* in base alla complessità ed alle tempistiche del progetto; in base alla tipologia ed alla necessità, questi saranno affiancati anche dal *team* che si occupa di *big data* all'inizio dei lavori per le analisi sui dati necessari. Ogni team è supervisionato da un *team leader* che mantiene il controllo sull'andamento dei lavori e le relazioni con gli altri team di sviluppo ed il *project manager*;

3. **Rilascio**: dopo un'attenta attività di *testing* interna all'azienda e di collaudo con il cliente, viene rilasciata una versione stabile del prodotto;

4. **Manutenzione**: con il passare del tempo, il prodotto viene mantenuto e aggiornato secondo nuove specifiche del cliente o in seguito a problemi riscontrati, sia da parte sua sia in caso di problemi in prodotti che condividono la stessa base di partenza e quindi passibili degli stessi errori che potrebbero compromettere la stabilità e la sicurezza del prodotto.

Sviluppo nuovo prodotto

Nel caso il prodotto non sia precedente commissionato da un cliente, le attività che vengono seguite sono leggermente differenti.

1.4 Tipo di clientela

1.5 Propensione dell'azienda per l'innovazione

Glossario

cluster (indicato anche come computer cluster) insieme di macchine connesse tra loro che lavorano in parallelo. L'utilizzo di questi sistemi permette di distribuire un'elaborazione molto complessa tra le varie macchine, aumentando la potenza di calcolo del sistema e/o garantendo una maggiore disponibilità di servizio, a prezzo di un maggior costo e complessità di gestione dell'infrastruttura: per essere risolto, il problema che richiede molte elaborazioni, viene infatti scomposto in sottoproblemi separati i quali vengono risolti ciascuno in parallelo su tutti i nodi che compongono il cluster.. v

dataset . v

web app sistema di tipo client-server in cui l'interfaccia utente e la logica client-side viene eseguita in un browser web.. v

Acronimi

API Application Program Interface. 1

Bibliografia