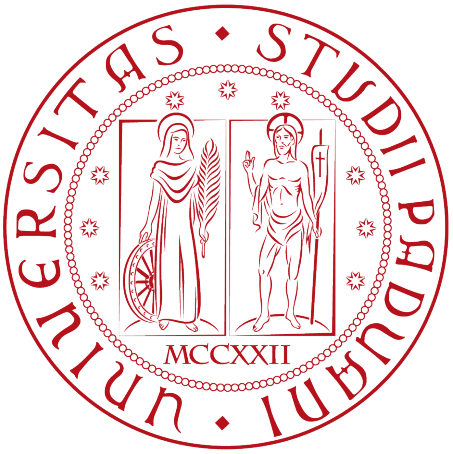
Università degli Studi di Padova

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

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.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Il contesto aziendale

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;

* **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 e 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

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

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 riguardanti 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, in particolare utilizzando le specifiche fornite dalla versione *Enterprise Edition*[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;
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;

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

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 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 del 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 di tecnici specializzati 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, 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

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

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 sul singolo prodotto 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.

Sviluppo nuovo prodotto

Nel caso il prodotto non sia precedentemente commissionato da un cliente, le attività che vengono seguite sono leggermente differenti. Il reparto di ricerca e sviluppo, il team *big data* ed il reparto marketing collaborano alla ricerca di un prodotto appetibile per un eventuale cliente a cui verrà in genere presentato solamente un *Minimum Viable Product*: in questo modo è possibile presentare all'interessato un prototipo del prodotto con alcune funzionalità essenziali e significative, senza perdite inutili di tempo e risorse per l'azienda. Le attività di implementazione, rilascio e manutenzione che sono eseguite in seguito all'ottenimento di un cliente interessato sono invece pari a quelle dello sviluppo su commissione.

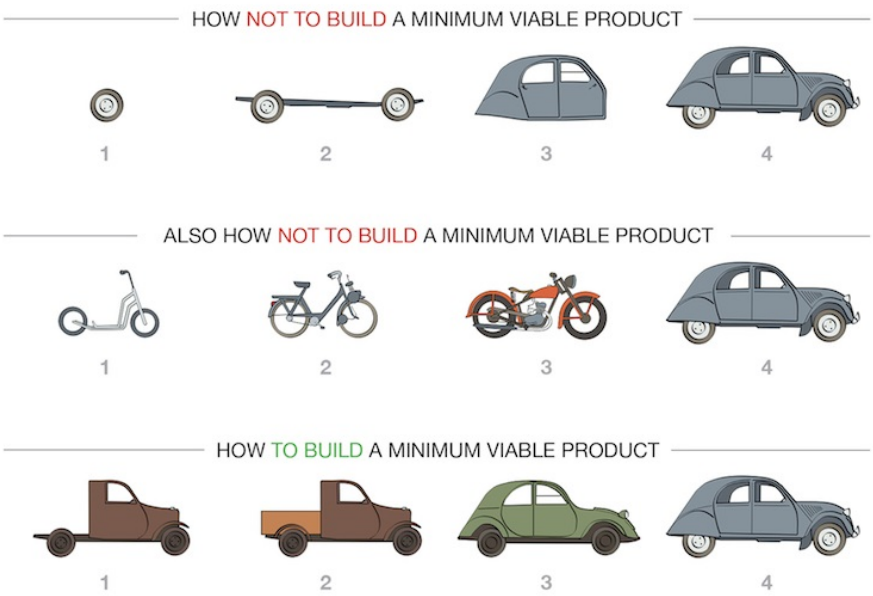


Figura 1.1: MVP: Minimum Viable Product

1.3.2 Controllo di versione

Come strumento di versionamento del codice, l'azienda utilizza Git[10] ed in particolare, per poter gestire tutto il codice derivante dai vari progetti, viene utilizzata la versione *enterprise* di GitLab[11], disponibile gratuitamente sotto licenza *open source*.
Alcuni dei vantaggi che ha spinto l'azienda ad utilizzare Git sono:

* **Ridondanza**: ogni sviluppatore possiede una copia dell'intera *repository*. Il rischio di perdita dei sorgenti del progetto è quindi inversamente proporzionale al numero di sviluppatori che ne possiedono in locale l'ultima versione; in caso di perdita del progetto di uno sviluppatore quindi si andrà incontro solamente alla perdita delle ultime modifiche personali fatte;
* **Disponibilità**: anche in assenza di connessione alla *repository* principale, è possibile continuare ad effettuare *commit* ed a salvare le modifiche fatte nel tempo. Una volta ripristinata la connessione, la *repository* locale può essere sincronizzata con quella remota rendendo le modifiche disponibili a tutti;
* ***Branch e Merge***: permette con molta facilità la creazione di *branch*, consentendo di creare quindi delle ramificazioni in cui sviluppare funzionalità non stabili, evitando di intaccare il ramo principale dove di norma risiede una versione stabile e testata del prodotto. Nel momento in cui si vogliono salvare le modifiche anche nel ramo principale, Git permette di effettuare l'operazione di *merge* del nuovo ramo testato con il ramo principale e considerare quindi le ultime modifiche come stabili;
* ***Fork e Pull request***: queste due operazioni permettono di clonare una *repository* (tramite *fork*) e, successivamente, proporre l'inclusione delle modifiche apportate all'interno del clone nella *repository* originale.

1.3.3 Ambiente di sviluppo

Gli strumenti utilizzati per lo sviluppo, divisi in categorie, sono i seguenti:

* **IntelliJ IDEA**[12]: è l'IDE utilizzato in prevalenza dagli sviluppatori in quanto supporta vari linguaggi di programmazione e vari tool utili;
* **Visual Studio Code**[13]: come il precedente, supporta vari linguaggi di programmazione e molteplici estensioni volte a migliorare lo sviluppo. A differenza del precedente, però, non supporta nativamente Java ed è quindi meno utilizzato;
* **Bash**: per effettuare l'analisi dei dati preliminari da parte del team *big data* viene utilizzata la *shell* di Linux per eseguire i tool interessati (ad esempio Hive, Impala, Spark o per visualizzare il *filesystem* HDFS).

[10] **https://git-scm.com/**.
[11] **https://about.gitlab.com/**.
[12] **https://www.jetbrains.com/idea/**.
[13] **https://code.visualstudio.com/**.

1.4 Tipo di clientela

I clienti principali di I.T. Euro Consulting sono aziende nazionali ed internazionali che lavorano nei seguenti settori:

* ***Banking***
* ***Insurance***
* ***Telecommunications***
* ***Media & Technology***
* ***Public Administration***
* ***Utilities & Energy***
* ***Manufacturing***

1.5 Propensione dell'azienda per l'innovazione

L'azienda è alla continua ricerca di nuove tecnologie e prodotti innovativi che possano soddisfare sempre più le esigenze del cliente. Nel primo caso l'azienda cerca di cogliere il meglio delle nuove tecnologie per poterne trarre il maggior beneficio possibile attraverso progetti sperimentali e, per la prima volta quest'anno, stage universitari. Essendo l'innovazione un importante fattore di crescita, I.T. Euro Consulting coltiva questo aspetto cercando personale che abbia attitudine al cambiamento e contemporaneamente esprima le proprie soluzioni ai problemi incontrati dando libero sfogo alla propria creatività. Nel secondo caso, invece, l'azienda si prefigge l'obiettivo di creare nuovi prodotti al passo con le esigenze di potenziali clienti, utilizzando le ultime tecnologie disponibili.
Un'ulteriore prova della propensione all'innovazione dell'azienda è la presenza di un team specifico ed in continua espansione per il segmento *big data*, cosa usuale all'estero ma ben più rara in Italia in aziende di medie dimensioni, nelle quali l'importanza dei dati e della potenzialità che essi possiedono non è ancora entrata pienamente nell'ideale di business delle aziende.

Glossario

Bash shell testuale del progetto GNU usata nei sistemi operativi Unix e Unix-like, specialmente in GNU/Linux. Si tratta di un interprete di comandi che permette all'utente di comunicare col sistema operativo attraverso una serie di funzioni predefinite, o di eseguire programmi e script.
Bash è in grado di eseguire i comandi che le vengono passati, utilizzando la redirezione dell'input e dell'output per eseguire più programmi in cascata in una pipeline software, passando l'output del comando precedente come input del comando successivo. Oltre a questo, essa mette a disposizione un semplice linguaggio di scripting nativo che permette di svolgere compiti più complessi, non solo raccogliendo in uno script una serie di comandi, ma anche utilizzando variabili, funzioni e strutture di controllo di flusso.. 5

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.. iii, 2, 3

Git Git è un software di controllo versione distribuito utilizzabile da interfaccia a riga di comando, creato da Linus Torvalds nel 2005 con lo scopo di essere un semplice strumento per facilitare lo sviluppo del kernel Linux, e diventato poi uno degli strumenti di controllo versione più diffusi al mondo.. 5

GitLab GitLab è un manager di *repository* Git basato su interfaccia web, che include anche funzioni quali una wiki per ogni progetto e un sistema di tracciamento issue. Esso è stato sviluppato da GitLab Inc. ed è distribuito gratuitamente con licenza *open source*.. 5

IDE (in lingua inglese *Integrated Development Environment* ovvero IDE, anche *integrated design environment* o *integrated debugging environment*, rispettivamente ambiente integrato di progettazione e ambiente integrato di *debugging*) è un software che, in fase di programmazione, aiuta i programmatori nello sviluppo del codice sorgente di un programma. Spesso l'IDE aiuta lo sviluppatore segnalando errori di sintassi del codice direttamente in fase di scrittura, oltre a tutta una serie di strumenti e funzionalità di supporto alla fase di sviluppo e *debugging*.. 5

Minimum Viable Product prototipo più semplificato possibile che è possibile presentare ad una cerchia di possibili clienti (early adopter). È il mezzo con cui testare e validare le idee e il prodotto stesso, senza sprecare tempo e soldi a sviluppare il prodotto completo, per poi constatare che quel prodotto non interessa alla clientela.. 4

Software Development Lifecycle processo di divisione del lavoro di sviluppo software in fasi distinte per migliorare la progettazione, la gestione del prodotto e la gestione del progetto.. 1

Single Customer View rappresentazione olistica del cliente che integra tutti i dati e gli eventi del cliente, e consente di arrivare ad un'interpretazione completa e contestuale dei suoi comportamenti indipendentemente dai canali utilizzati.. 2

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