



**UNIVERSITÀ DEGLI STUDI DI PADOVA**

---

DIPARTIMENTO DI MATEMATICA “TULLIO LEVI-CIVITA”

*CORSO DI LAUREA IN INFORMATICA*

**DATA INGESTION E ANOMALY DETECTION: IL  
CICLO DI VITA DEL DATO**

*TUTOR INTERNO*

LAMBERTO BALLAN  
UNIVERSITÀ DI PADOVA

*TUTOR ESTERNO*

MICHELE GIUSTO

*STUDENTE CANDIDATO*

GIOVANNI SORICE

**ANNO ACCADEMICO 2018 - 2019**



THIS IS THE DEDICATION.



# Sommario

LOREM IPSUM DOLOR SIT AMET, consectetur adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetur. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.



# Indice

ELENCO DELLE FIGURE	ix
ELENCO DELLE TABELLE	xi
<b>I INTRODUZIONE</b>	<b>I</b>
1.1 L'azienda . . . . .	I
1.1.1 Le tecnologie da esplorare . . . . .	2
1.2 Processi del progetto . . . . .	2
1.2.1 Metodologia di sviluppo . . . . .	2
1.2.2 Strumenti di supporto . . . . .	2
1.2.3 Propensione alla modernizzazione . . . . .	2
<b>2 EVOLUZIONE DELLO STAGE</b>	<b>3</b>
2.1 La proposta di stage . . . . .	3
2.2 Il progetto . . . . .	3
2.2.1 Obiettivi . . . . .	4
2.2.2 Vincoli . . . . .	5
2.3 Motivi della scelta . . . . .	5
<b>3 ANALISI E PROGETTAZIONE</b>	<b>7</b>
3.1 Metodo di lavoro . . . . .	7
3.2 Problemi affrontati . . . . .	7
3.3 Pianificazione del lavoro . . . . .	7
3.4 Risultati . . . . .	9
<b>4 SVILUPPO</b>	<b>II</b>
4.1 Tecnologie . . . . .	II
4.1.1 Google Cloud . . . . .	II
4.1.2 Apache NiFi . . . . .	15
4.1.3 Apache Beam . . . . .	15
4.1.4 Apache AirFlow . . . . .	16
4.2 Integrazione . . . . .	17
<b>5 CONCLUSIONE</b>	<b>19</b>
REFERENCES	20





# Elenco delle figure

1.1	Short figure name. . . . .	I
4.1	Short figure name. . . . .	12
4.2	Short figure name. . . . .	15
4.3	Short figure name. . . . .	16
4.4	Short figure name. . . . .	17



## Elenco delle tabelle



# 1

## Introduzione

Dove? parte introduttiva, spiegazione cosa leggeremo

### 1.1 L'AZIENDA

Data Reply S.r.l fa parte delle aziende del gruppo Reply S.p.a e si occupa del mondo Big Data, Data Science e Artificial Intelligence. L'azienda è giovane sia nella sua creazione, dato che è stata fondata nel 2013, sia dal punto di vista della composizione del personale. Questo le permette di avere flessibilità e freschezza mentale (in termini di idee) ma allo stesso tempo know how e conoscenza del settore portata dagli elementi con maggiore esperienza.

Il mercato di Data Reply è quello della consulenza, un mondo spesso ostico e molto complesso ma che ha visto un importante crescita degli investimenti in Italia negli ultimi anni.



Figura 1.1: Logo Data Reply S.r.l.

### 1.1.1 LE TECNOLOGIE DA ESPLORARE

Le principali tecnologie utilizzate dall'azienda riguardano i mondi mondo Big Data, Data Science e Artificial Intelligence. Troviamo un importante utilizzo delle suite di prodotti cloud, come AWS e Google Cloud, e di prodotti da poter utilizzare in locale, come Cloudera. Infine, anche l'utilizzo di framework per il calcolo distribuito, come Apache Spark, sono di fondamentale importanza.

## 1.2 PROCESSI DEL PROGETTO

### 1.2.1 METODOLOGIA DI SVILUPPO

All'interno di Data Reply vengono utilizzati diversi metodologie di sviluppo, che vanno dalle metodologie agile, nella quale vediamo tra le più utilizzate Scrum e Kanban. Essendo un'azienda di consulenza, spesso si trova a lavorare anche con altre aziende allo stesso progetto e per questo motivo non è raro trovare metodologie più rigide ai cambiamenti come le metodologie a cascata.

### 1.2.2 STRUMENTI DI SUPPORTO

Gli strumenti di supporto, come le metodologie di sviluppo, utilizzati sono decisi in base al progetto, al cliente e in accordo con le altre aziende che lavoreranno al progetto stesso. Di consueto, l'azienda propone l'utilizzo di GitLab come strumento di versionamento e GitLab Mattermost per le comunicazioni informali tra i componenti del progetto e dell'azienda.

### 1.2.3 PROPENSIONE ALLA MODERNIZZAZIONE

L'azienda cerca sempre di rinnovarsi e rimanere al passo con le ultime tecnologie in modo da poter proporre ai propri clienti il giusto compromesso tra novità e affidabilità. Spesso infatti, vi sono progetti mirati allo studio ed utilizzo di tecnologie e prodotti da poco sul mercato, così da garantire un vantaggio competitivo sulle altre aziende.

*Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo.*

Quoteauthor Lastname

# 2

## Evoluzione dello stage

Perchè?

### 2.1 LA PROPOSTA DI STAGE

La proposta di stage mi è stata fatta nei mesi successivi alla visita all'azienda svolta attraverso l'Università di Padova. Fin da subito il complesso e lo spirito aziendale mi sono sembrati adatti ad ospitare degli studenti volenterosi di sperimentare nuove tecnologie. Negli ultimi anni, tra le tecnologie che hanno riscosso più successo e fama, troviamo i Big Data e tecniche di intelligenza artificiale con particolare attenzione al machine learning.

### 2.2 IL PROGETTO

Il progetto si prefissava l'obiettivo di analizzare e costruire l'intero ciclo di vita del dato, secondo metodi, strumenti e tecniche innovative e decise dal cliente. Quindi, per essere maggiormente concreti, il progetto prevedeva l'implementazione di un sistema di ingestion dei dati aziendali del cliente con successiva classificazione degli utenti che interagivano con il sistema stesso, in modo da intercettare eventuali comportamenti sospetti.

Per svolgere questo progetto, prima del mio arrivo e dell'effettiva proposta, l'azienda ospitante ha svolto uno studio di fattibilità in modo da sottolineare eventuali carenze di informazioni, criticità e punti a favore del progetto. Durante il mio periodo di permanenza, il progetto

è stato portato al termine seguendo le indicazioni iniziali e con alcuni aggiustamenti durante il percorso. Inoltre, sono state rispettate le fondamentali indicazioni del tutor aziendale che ha affiancato durante lo stage. Terminato il mio stage il prodotto risultava conforme alle aspettative e facilmente manutenibile grazie alla documentazione redatta durante il periodo di studio del problema, sviluppo e test. Spiegarlo ad ampio raggio (prima di me, con me e dopo di me)

## 2.2.1 OBIETTIVI

Durante i colloqui di esposizione del progetto di stage, sono sorti alcuni obiettivi da soddisfare al termine dello stesso.

Essi si dividono in:

- Tutti quegli obiettivi che sono fondamentali per la corretta riuscita del progetto, anche detti, obiettivi obbligatori:
  - Il sistema deve essere in grado di aggiornare periodicamente (es. 1 volta al gg) la concezione di normalità, per adattarsi al contesto dinamico;
  - Il sistema deve essere in grado di ricevere i flussi batch ed archivarli dove desiderato;
  - Il sistema deve essere in grado di processare i flussi batch producendo i dataset di output richiesti;
  - I processi devono essere schedulati per avviarsi periodicamente;
- Obiettivi desiderabili, cioè, quegli obiettivi che porterebbero ad una maggiore completezza del progetto di stage ma che non sono fondamentali per la sua buona riuscita:
  - Il sistema deve riuscire ad aggiornarsi continuamente senza interrompere l'erogazione del servizio;
  - Il sistema deve essere in grado di scrivere i risultati su una struttura interrogabile via SQL;
  - Il sistema deve poter scalare le risorse in base al volume di dati;
- Tutti quegli obiettivi che sono marginali al progetto ma che porterebbero comunque valore aggiunto allo stesso, anche detti, obiettivi facoltativi,;



- Il sistema deve poter lavorare su dati archiviati su Elasticsearch;
- Il sistema deve poter essere in grado di scrivere i propri output su DB NoSQL;

### 2.2.2 VINCOLI

Durante i colloqui di esposizione del progetto di stage, sono sorti alcuni vincoli da soddisfare per svolgere al meglio lo stesso.

#### VINCOLI TEMPORALI

La durata dello stage doveva essere al minimo di 300 ore e al massimo di 320, in modo da rispettare i crediti stabiliti all'interno del piano di studi, svolte tutte presso la sede in via Robert Koch 1/4 a Milano negli orari 9.00-13.00, 14.00-18.00.

#### VINCOLI TECNOLOGICI

Utilizzo della suite Google Cloud per lo sviluppo dell'ingestion dei dati.

### 2.3 MOTIVI DELLA SCELTA

Perchè Reply?



*This is some random quote to start off the chapter.*

Firstname lastname

# 3

## Analisi e progettazione

### 3.1 METODO DI LAVORO

Dopo un'attenta analisi delle esigenze di questo progetto, la metodologia che abbiamo deciso di utilizzare è Agile Kanban. Questa metodologia, ci ha permesso di avere un'alta flessibilità sui task da svolgere, una visione ampia del progetto e il focus su una singola attività alla volta. La metodologia Kanban è stata scelta anche per affrontare l'esiguo numero del team, composto da sole 2 persone, con effettivamente solo una che lavorava a pieno regime su di esso. Questo perché rendeva un'ottima visione dell'insieme delle attività da svolgere, in progresso e già svolte, oltre a non necessitare di ruoli predefiniti come nell'Agile Scrum

### 3.2 PROBLEMI AFFRONTATI

### 3.3 PIANIFICAZIONE DEL LAVORO

Durante i colloqui svolti con il tutor aziendale, è stato redatto il piano di lavoro. Ciò ha portato la suddivisione dello stage in 8 parti, ognuna della durata di una settimana.

#### **Prima Settimana (40 ore)**

- Incontro con persone coinvolte nel progetto per discutere i requisiti e le richieste relativamente al sistema da sviluppare;
- Verifica credenziali e strumenti di lavoro assegnati;

- Presa visione dell'infrastruttura esistente;
- Formazione sulle tecnologie adottate;

#### **Seconda Settimana (40 ore)**

- Comprensione strumenti di storicizzazione;
- Analisi funzionamento filesystem distribuito;
- Produzione script Python per produzione dati simulati;

#### **Terza Settimana (40 ore)**

- Caricamento dati su filesystem distribuito;
- Securitizzazione accesso ai dati;
- Studio teorico approccio statistico al problema;

#### **Quarta Settimana (40 ore)**

- Comprensione strumenti di processamento;
- Analisi funzionamento RDD;
- Implementazione processo batch con Python e Spark;

#### **Quinta Settimana (40 ore)**

- Analisi funzionamento Dataframe;
- Implementazione processo di elaborazione Spark;
- Ingegnerizzazione soluzione batch;

#### **Sesta Settimana (40 ore)**

- Test prestazionale processo spark;
- Tuning prestazionale;
- Studio applicazione real-time;

### **Settima Settimana (40 ore)**

- Storizzazione dati su storage persistente;
- Studio funzionamento Kafka;
- Implementazione processo real-time con Spark Streaming;

### **Ottava Settimana - Conclusione (40 ore)**

- Costruzione layer accesso al dato persistente;
- Ingegnerizzazione soluzione Spark Streaming;
- Integrazione storage persistente;

possibili diagrammi

## **3.4 RISULTATI**



*Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo.*

Quoteauthor Lastname

# 4

## Sviluppo

### 4.1 TECNOLOGIE

In questa sezione, verranno elencate e introdotte le tecnologie che sono state utilizzate nel progetto di stage.

#### 4.1.1 GOOGLE CLOUD

Il principale servizio utilizzato durante il progetto è stato Google Cloud, in esso troviamo una moltitudine di strumenti utili allo sviluppo, all'esecuzione e all'interazione di progetti riguardanti Big Data e Intelligenza Artificiale. Inoltre, importante risulta la possibilità di consultare, manipolare e presentare i dati. Il progetto Google Cloud (anche detto Google Cloud Platform), viene lanciato nel 2008 con uno dei suoi servizi di punta ancora oggi App Engine. Nell'arco degli anni sia il bacino di utenti, sia i servizi offerti sono via via aumentati. Oggigiorno i suoi **clienti** sono molti, tra cui grandi aziende, e le funzionalità di spicco sono molteplici, i campi di maggiore interesse riguardano strumenti di Compute, Storage & Databases, Big Data, Cloud AI e IoT. Per quanto riguarda l'esperienza fatta da me personalmente, ho trovato che Google Cloud fosse un buon servizio, con dei problemi da risolvere sulla completezza della documentazioni e a volte con grafiche non belle da vedere, ma comunque un ottimo compromesso tra qualità e prezzo. Il resto della sezione introduce il prodotti e i servizi GCP utilizzati durante lo stage.



Figura 4.1: Logo Data Google Cloud Platform

## GOOGLE STORAGE

Lanciato nel 2010, **Google Storage** è il servizio di hosting dei file della piattaforma. All'interno ogni utente può scegliere in che modalità salvare i dati (zona del mondo, carico di lavoro e frequenza d'accesso), così da trovare la soluzione maggiormente adatta ai propri bisogni. I suoi punti di forza sono:

- Interoperabilità: possibilità di interfacciarsi con strumenti e servizi di aziende terze (come Amazon S3);
- Consistenza: tutte le operazioni sono svolte secondo il principio di atomicità;
- Controllo di accesso: possibilità di definire l'accesso ad un singolo oggetto o bucket a seconda della categoria di appartenenza degli utenti;
- Resumable Uploads: in caso di fallimento di un upload, questo può essere continuato/ricominciato.

Per quanto riguarda l'esperienza d'utilizzo, si è rivelato un attimo servizio di storage, l'abilitazione di un'unica API rende disponibile l'intero pacchetto di funzionalità e ciò l'ho trovato molto comodo.

## GOOGLE BIGQUERY

Servizio lanciato insieme a Google Storage nel 2010, **Google BigQuery**, è lo strumento che permette di interagire e analizzare grandi moli di dati, definito come un data warehouse.



All'interno di GCP, vi è un'apposita interfaccia utente per utilizzarlo. Si basa su tabelle codificate in formato JSON e sulla quale vengono processate query in **standard SQL** tramite avanzati algoritmi di MapReduce.

I principali punti di forza sono:

- La piattaforma rende disponibile la creazione e distruzione di tabelle basate su schemi JSON, inoltre rende disponibile l'import e l'export di dati in formati come CSV e JSON;
- La semplicità del linguaggio SQL con scalabilità e prestazioni che solo i servizi cloud riescono a dare;
- Un'ottima integrazione con servizi e API esterni, per la raccolta e l'utilizzo dei dati;
- Possibilità di garantire l'accesso ad uno o più categorie di utenti.
- Possibilità di interagire con strumenti di Machine Learning nativi di Google e non (come TensorFlow).

L'opinione del team di sviluppo riguardo lo strumento è molto positiva soprattutto grazie alla scalabilità automatica che rende semplice e veloce l'utilizzo delle query su grandi moli di dati, un possibile punto di debolezza può essere trovato nell'utilizzo in sola lettura dei dati esistenti.

## GOOGLE DATAFLOW

Rilasciato al pubblico nel 2015, **Google Dataflow** è lo strumento che permette di processare pipeline per l'integrazione, la preparazione e l'analisi di grandi quantità di dati. Il modello di pipeline consigliato è quello di Apache Beam che vedremo in seguito.

Tra i maggiori punti di forza troviamo:

- Creazione di pipeline secondo modelli già esistenti o personalizzati dall'utente;
- Composizione di pipeline, e quindi job specifici, direttamente da codice;
- Possibilità di utilizzo di dati da molteplici fonti, non solo da strumenti Google;
- Ottima scalabilità interna in base agli effettivi bisogni del lavoro in esecuzione;
- Buona granularità della specificazione dei **parametri di esecuzione** della pipeline;
- Visualizzazione grafica della pipeline (Possibile immagine????).

Per quanto concerne l'utilizzo di questo prodotto, il team si esprime con un giudizio più che positivo dato anche la grande quantità di documentazione disponibile.

## GOOGLE DATA STUDIO

Lanciato nel 2016, **Google Data Studio** è lo strumento di creazione e visualizzazione dei dati attraverso report e dashboard. L'idea che sta alla base di questo tool è la condivisione dei dati che sono presenti all'interno di GCP attraverso tabelle e grafici.

I maggiori vantaggi nell'utilizzarlo sono:

- Accesso ai dati personali di GCP;
- Possibilità di integrare i dati fra di loro con query standard SQL;
- Condivisione a gruppi di utenti.

Durante l'esperienza lo strumento si è rivelato molto acerbo nelle sue funzionalità e nell'utilizzo, oltre a dover attendere troppo tempo per il caricamento di dati e grafici. L'idea alla base è ottima, dato che vi è la possibilità di utilizzare i dati all'interno di GCP anche in continuo aggiornamento, d'altro canto la documentazione e l'interfaccia utente non rendono semplice quanto potrebbero la creazione di un report aziendale da presentare al cliente.

## GOOGLE FUNCTION

Rilasciato in beta nel 2017, **Google Functions** è il perfetto intermediario tra gli eventi che vengono scaturiti nel cloud e l'invocazione di servizi API. Le origini degli eventi supportati sono:

- Cloud Pub/Sub;
- Cloud Storage;
- HTTP;
- Stackdriver Logging;
- Firebase.

Inoltre, per ogni tipologia di evento vi sono ulteriori specifiche che non verranno trattate tutte in modo estensivo in questo documento. Per quanto concerne l'utilizzo svolto durante il progetto, il team ritiene questo strumento valido e con ottime potenzialità e margini di miglioramento. Uno svantaggio molto importante da sottolineare è la scarsa documentazione presente nelle fonti ufficiali.

GOOGLE COMPOSER

<https://cloud.google.com/composer/?hl=it>

#### 4.1.2 APACHE NIFI

An easy to use, powerful, and reliable system to process and distribute data.

Quoteauthor Lastname

La descrizione che viene data nella documentazione di **NiFi** riassume alla perfezione le caratteristiche di questo strumento e ciò che ci permette di fare.

Innanzitutto, le sue funzioni sono proprio quelle di automatizzare lo "scorrere" del flusso dati all'interno del sistema in cui viene utilizzato. Le funzionalità sono molteplici e ben rodute, vanno dal semplice spostamento dei file, al filtraggio, composizione e conversione in diversi formati. La documentazione è molto esauriente, spiegando anche più del necessario tutti i passaggi da seguire per utilizzare un determinato componente. La semplicità e l'intuitività nel costruire il flusso di dati, rendono l'esperienza utente decisamente piacevole, anche grazie all'interfaccia grafica semplice e intuitiva. Altro punto a favore è l'affidabilità del prodotto e la grande quantità di componenti disponibili orientati alla sicurezza, così da rendere il passaggio di informazioni in più sicuro possibile.



Figura 4.2: Logo Data Google Cloud Platform

#### 4.1.3 APACHE BEAM

**Apache Beam** è un implementazione del **modello Dataflow**, esso infatti fissa come definire ed eseguire pipeline sui dati di tipo ETL, batch e streaming continui. Questo progetto nasce dopo il rilascio di un SDK contenente un implementazione del modello Dataflow da parte di Google nel 2014, nel 2016 Google decise di concedere il core del proprio SDK per costituire il progetto OpenSource Apache Beam.

I concetti alla base di questa implementazione, e quindi del modello Dataflow, possono essere riassunti nel cercare un modello il quale riesca a definire ed eseguire pipeline di tipo batch e streaming allo stesso tempo, senza dover differenziare le pipeline come nel modello **Lambda architecture**.

Un tipico utilizzo di una pipeline può essere il seguente:

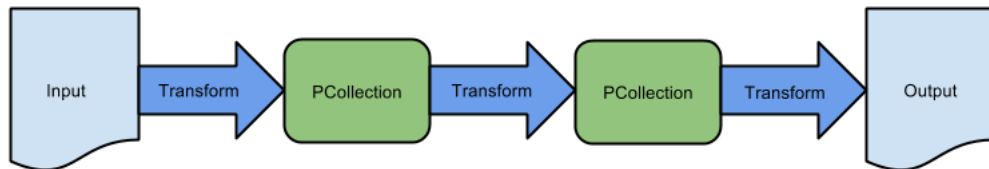


Figura 4.3: Logo Data Google Cloud Platform

- Creazione della Pipeline;
- Creazione di una PCollection da dei dati di input;
- Applicazione di PTrasforms sulla PCollection;
- Scrittura dei dati all'interno di un file.

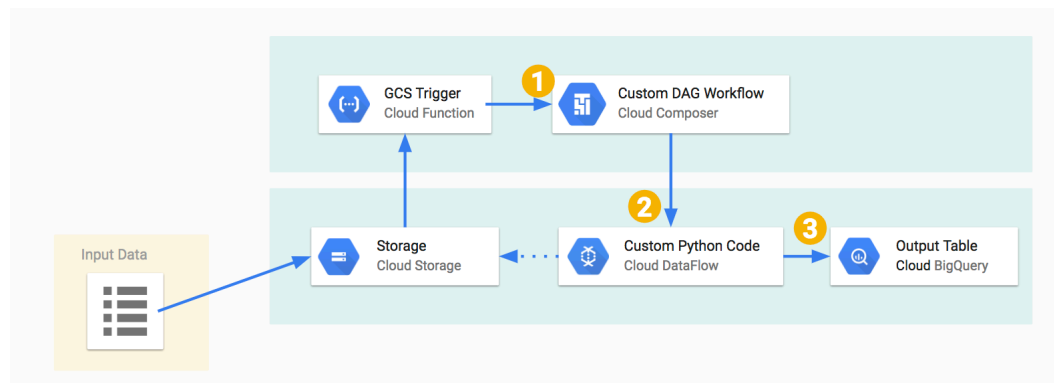
Le azioni di maggiore interesse stanno nelle PTrasforms, infatti rende possibile processare ogni singolo elemento delle PCollection, effettuare raggruppamenti per uno o più elementi, unire e partizionare le PCollections. La documentazione a disposizione è decisamente idonea e precisa rispetto alla potenzialità di utilizzo. Inoltre, essendo un progetto OpenSource vi è la possibilità di collaborare al continuo sviluppo, e quindi al successo, che questa tecnologia sta avendo.

#### 4.1.4 APACHE AIRFLOW

**Apache AirFlow** è una piattaforma per creare, schedulare e monitorare data warehouse. Airflow, inoltre, cerca la soluzione più performante per l'esecuzione dei task, creando un Grafo Aciclico Diretto (meglio conosciuto come "data warehouse" cioè "Directed Acyclic Graphs") in modo da riconoscere le dipendenze tra i compiti stessi ed eseguire i task con meno dipendenze. La scelta di utilizzare questa tecnologia è stata presa sia per la già precedente esperienza aziendale con essa, sia perché completamente compatibile e già installata all'interno di

Google Cloud Composer. Data la sua notevole compatibilità ci ha permesso di soddisfare un requisito del cliente riguardante l'ingestion in tempo reale dei dati inviati.

#### 4.2 INTEGRAZIONE



**Figura 4.4:** Logo Data Google Cloud Platform

Come?

Durante l'esperienza lo strumento si è rilevato...



# 5

## Conclusione

Analisi retrospettiva, cosa cambierei





## Bibliografia



# Acknowledgments

LOREM IPSUM DOLOR SIT AMET, consectetur adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetur. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.