

# Progetto BD4M

*Valutazione di impatto per industrializzazione  
Piattaforme Kafka e Grafana*



**Andrea De Marco**

Manager BigData & Advanced Analytics

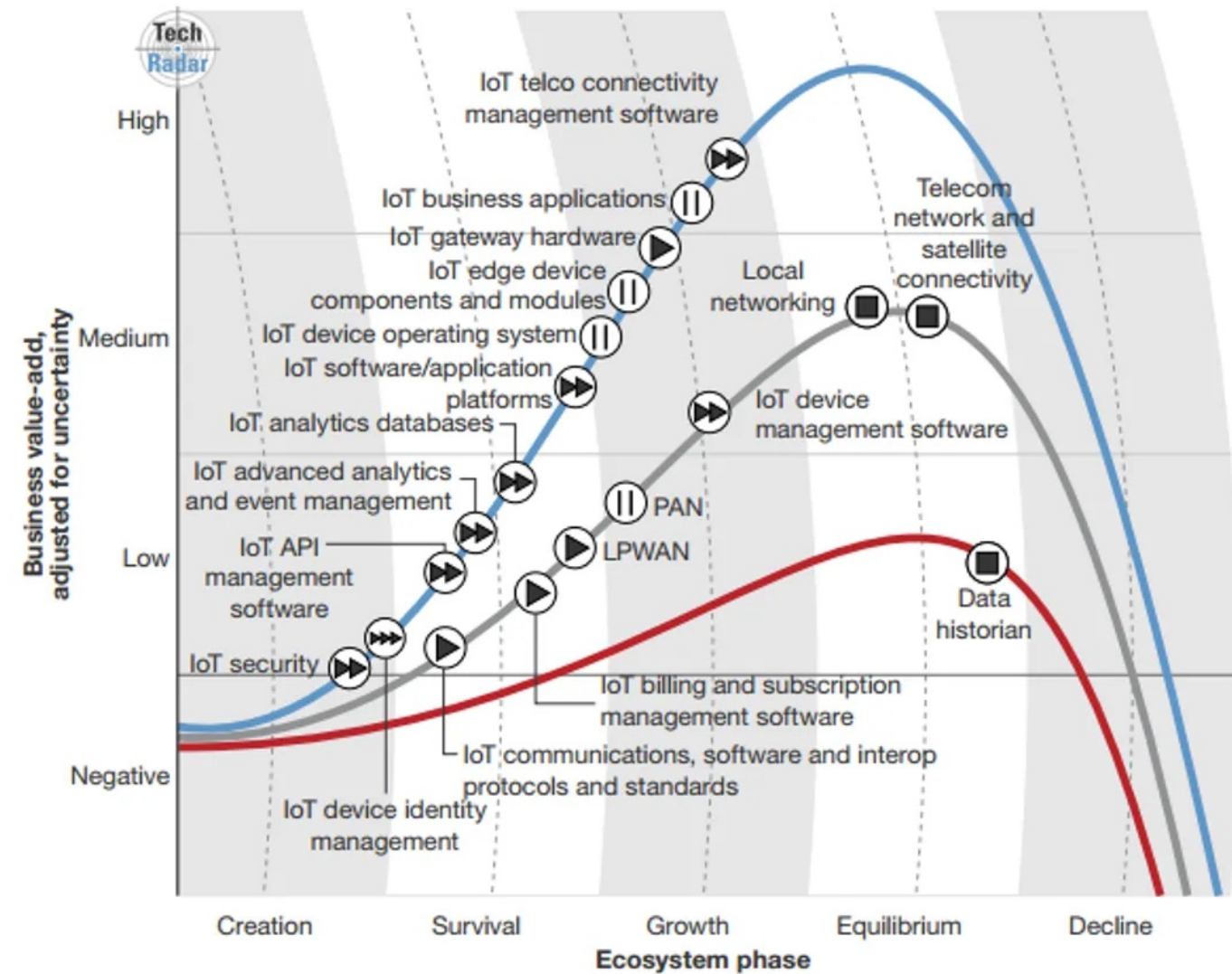
BitBang

# Industrializzazione Piattaforme

- Analizzare l'**impatto** di una scelta di piattaforma tecnologica in ambito industriale è una attività complessa
- Comporta una serie di **considerazioni** di ordine generale e alcune questioni che cambiano per ogni azienda
- In questa sede potremo solo fare riferimento alle considerazioni di ordine generale cercando di **generalizzare** il più possibile le valutazioni in modo da renderle utili per tutte le aziende interessate

Le questioni più importanti che approfondiremo sono:

1. Le pratiche virtuose di settore
2. I termini e le condizioni legali/commerciali
3. Il TCO

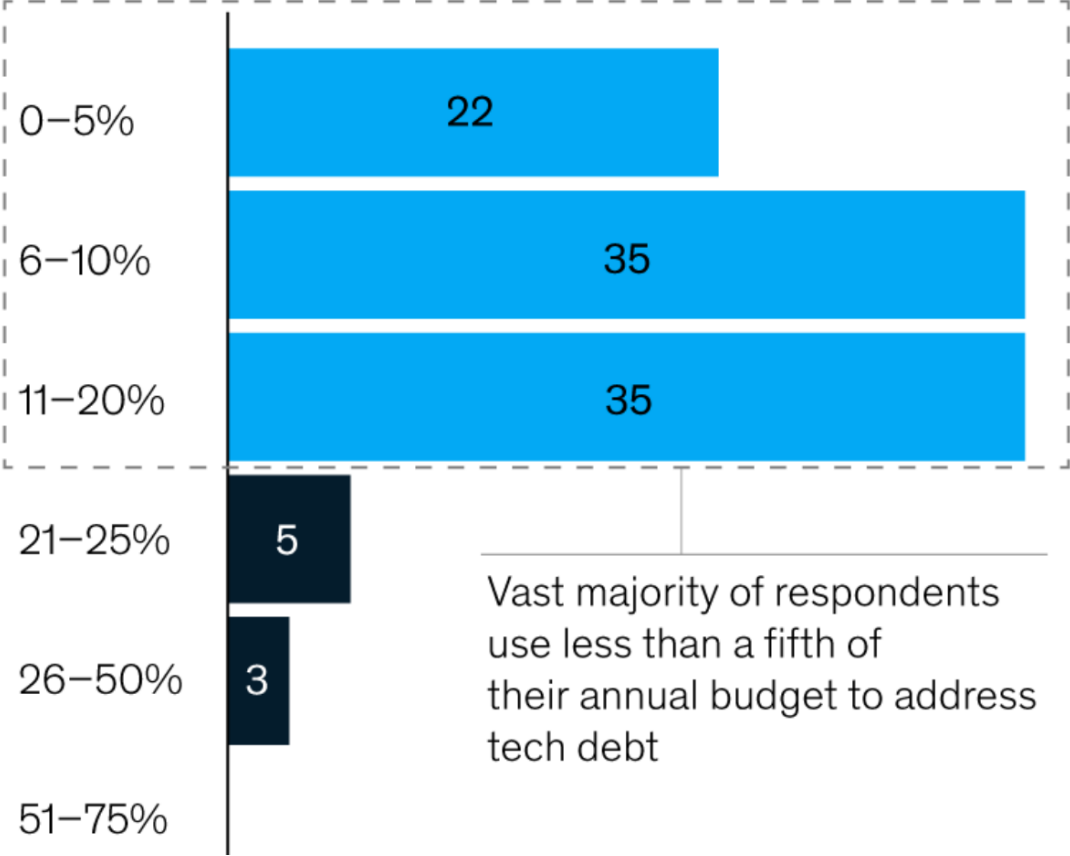


# Le pratiche virtuose di settore

Share of tech budget allocated to paying down tech debt

Share of IT budget

% of respondents



In 20 anni di esperienza, abbiamo svolto molte attività di studio di fattibilità, specie in ambito innovazione per le piattaforme di BigData.

Questo ci permette di riportare le pratiche virtuose del settore con il supporto degli studi che abbiamo certificato in passato con Forrester Research e Gartner.

Il fattori chiave che sono sicuramente interessanti per tutte le tipologie di azienda sono: l'analisi del **debito tecnico** e la road map tecnologica che permette di fare scelte che possiamo considerare **a prova di futuro**.

# Le pratiche virtuose in ambito Analisi

Negli ultimi anni abbiamo riscontrato alcune tendenze che hanno influenzato in maniera **irreversibile** le scelte di piattaforme per l'analisi dei dati in streaming.

Grazie alle storie di successo di aziende che hanno sperimentato con i diversi framework (early adopters) è più semplice puntare sulle **più promettenti** per chi arriva solo oggi ad affrontare queste scelte.

Molte delle pratiche virtuose sono **già state indicate** all'interno del progetto BD4M, ad esempio l'approccio ai micro-servizi grazie all'uso dei container e il sistema di orchestrazione.

Queste scelte permetteranno di adattarsi in maniera Agile agli scenari futuri specie di Dev/Ops e Data/Ops su sistemi distribuiti in cloud.

# Le pratiche virtuose in ambito Analisi avanzata di dati in Streaming

La scelta di Spark come ambiente di **DataLakeHouse** in grado di svolgere diversi compiti dal Machine Learning (specie in Python) ai workload in SQL è sicuramente una garanzia per il futuro.

Tutte le ricerche sono concordi nell'affermare che questo framework per il calcolo distribuito sarà sempre più al centro di tutte le architetture di analisi dei dati.

Inoltre, Databricks che lo sviluppa, è la società con in più alto finanziamento da venture della storia.

In particolare, per lo streaming è di sicuro **Kafka** la scelta che ha premiato maggiormente in questi anni i progetti di innovazione.

Tra i vari sistemi di monitoring dashboard, **Grafana** negli ultimi anni si è indubbiamente imposto come uno degli attori principali.

Nato originariamente come fork di Kibana, parte dell'ecosistema di Elasticsearch, pur avendo mantenuto forti legami con quest'ultimo, ha nel tempo sviluppato connettori a molteplici altre fonti dato, diventando di fatto uno dei principali tool di visualizzazione di dati in streaming.

# Kafka



# Event Stream Processing Platforms

Secondo Gartner, il settore dello streaming, per la gestione e l'analisi dei dati è cresciuto molto negli ultimi anni come volume di affari.

La crescita sarà sostenuta anche in futuro e per questo motivo molte software house continueranno ad investire su queste piattaforme rilasciando nuove funzionalità ed innalzando gli standard di sicurezza.

The event stream processing platform software market grew 19.6% in 2020 to \$719 million. Organizations are doing more stream processing because of the need for continuous intelligence, better situation awareness and faster, more precise business decisions.

# Confluent e le versioni Open di Kafka

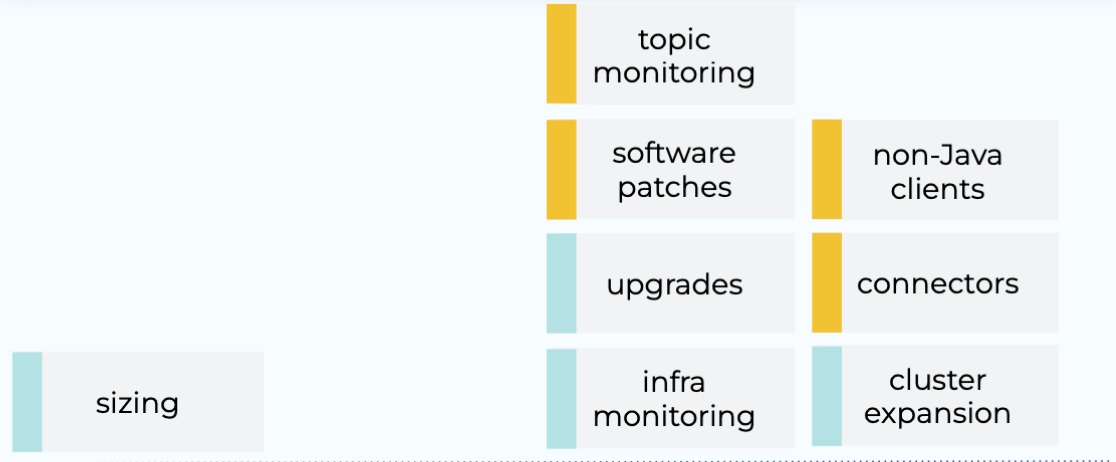
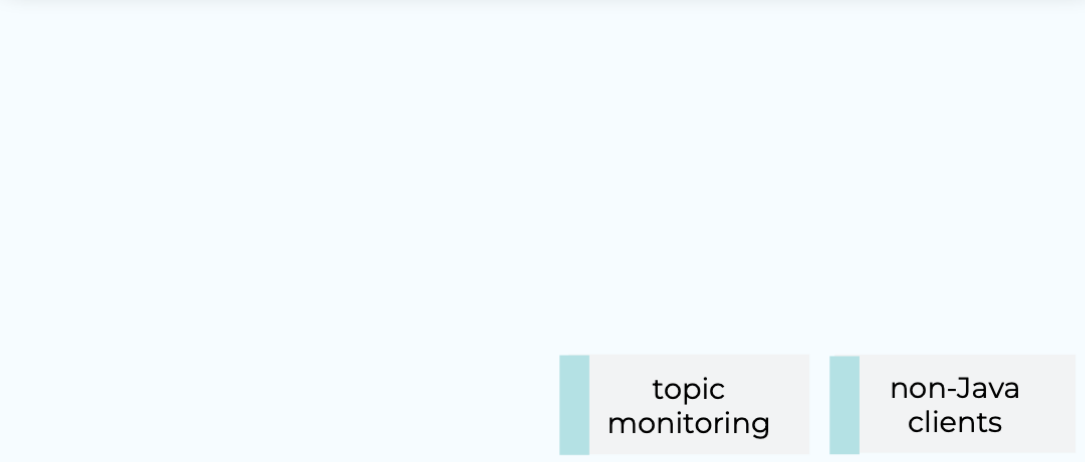
## CONFLUENT CLOUD

Automated operations enable efficient growth

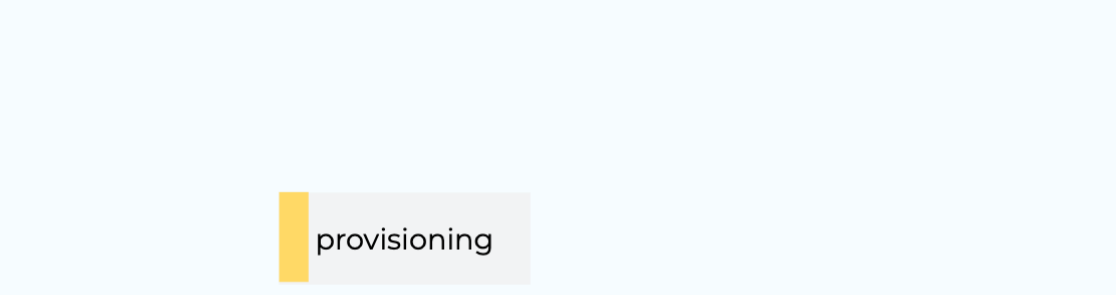
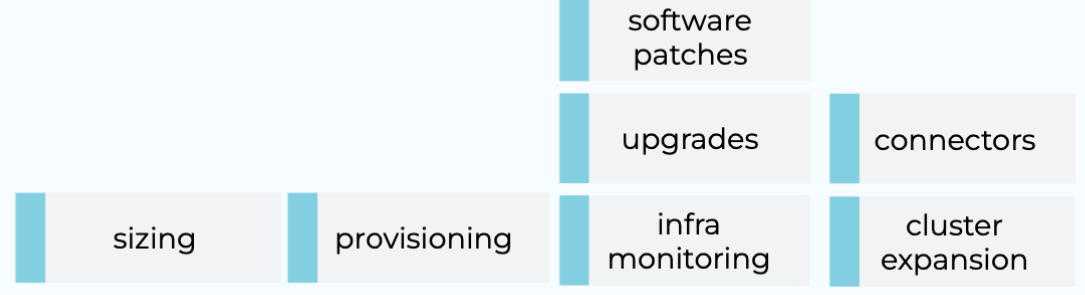
## Open Source Kafka

Operational burden at every step

manual operations  
(less is better)



automated product capabilities  
(more is better)



Freedom of Choice  
Committer-driven Kafka Expertise

Open Source Kafka  
Need in-house Kafka support expertise



# Considerazioni su Scalabilità Kafka on Premise

Configurazione suggerita per deployment su bare metal per produzione:

- RAM da 32 a 64 GB
- CPU da 12 a 24 core
- Storage dedicato, raccomandato RAID 10

Necessità di installazione di Zookeeper:

- Sebbene dalla versione 2.8 Kafka possa farne a meno, non è ancora raccomandato in produzione. <https://developer.confluent.io/learn/kraft/> "KRaft is in early access and should be used in development only. It' is not suitable for production."
- Nodi zookeeper necessari da 3 a 5.
- <https://github.com/apache/kafka/blob/trunk/config/kraft/README.md>

## Warning

KRaft mode in Kafka 3.1 is provided for testing only, *NOT* for production. We do not yet support upgrading existing ZooKeeper-based Kafka clusters into this mode.

There may be bugs, including serious ones. You should *assume that your data could be lost at any time* if you try the preview release of KRaft mode.

# Considerazioni sulla Sicurezza di Kafka Open Source

- Difficoltà set up
- Nessun tool monitoraggio out of the box
- Configurazioni manuali security TLS & Kerberos
- Fine tuning complesso
- Non scalabile
- Nessun supporto
- Difficoltà di patching

No SSO for single sign on

No managed connectors for out of box integrations to downstream systems

manual cluster rebalancing

Self-managed KSQLDB

No SLA based 24x7 support, rely on open source community

No proactive monitoring and troubleshooting

No security/vulnerability testing or patches

authorization: ACL's

no separation of compute and storage

no governance via schema registry

# Conclusioni

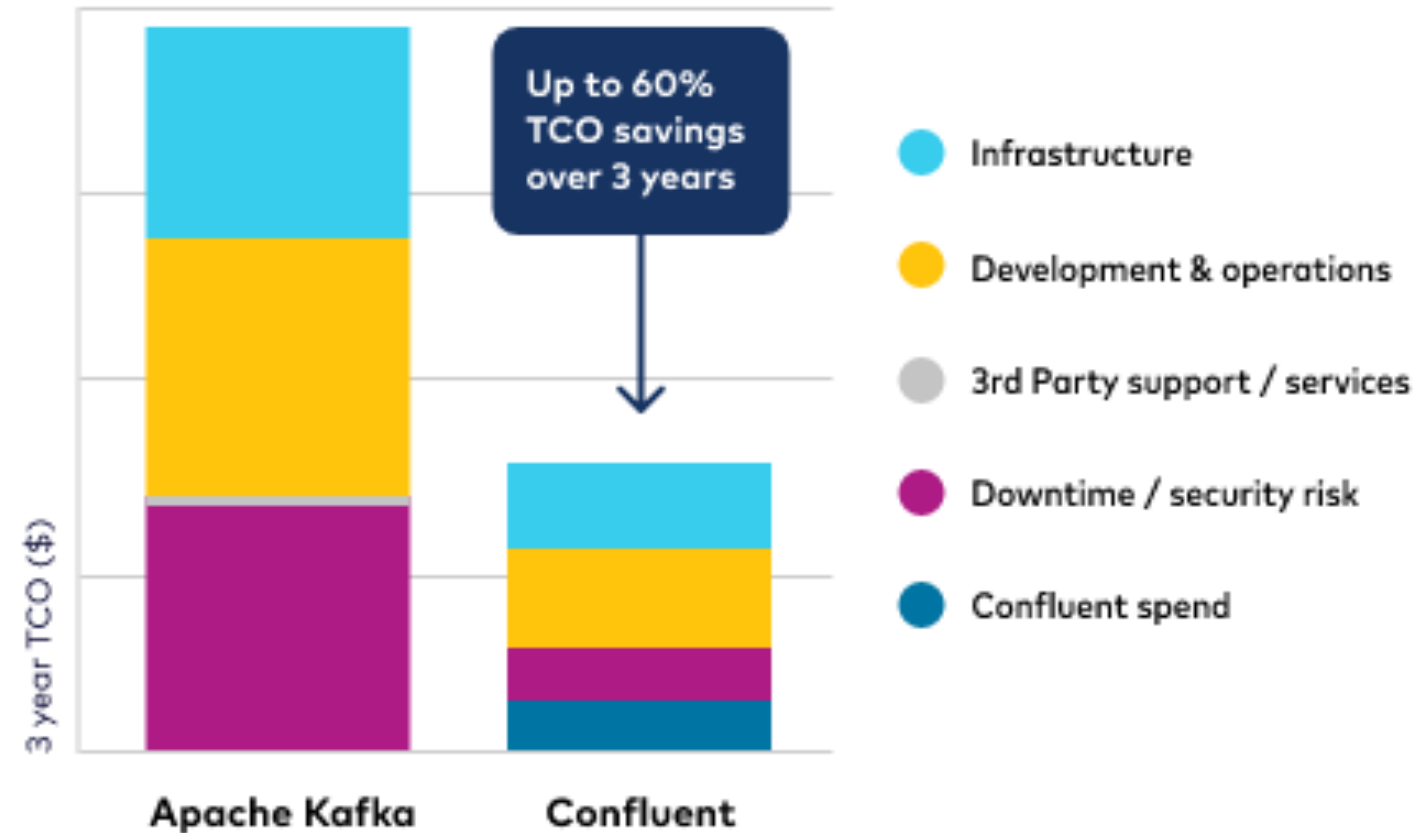
## Container

- Aggiunge complessità di setup e manutenzione di cluster Kubernetes
- Complessità di setup configurazioni DNS ad hoc per il discovery dei servizi
- Utilizzo di storage NAS, non raccomandato per degrado delle prestazioni

<https://www.confluent.io/kafka-summit-london18/kafka-in-containers-in-docker-in-kubernetes-in-the-cloud/>

<https://www.confluent.io/kafka-summit-lon19/running-kafka-in-kubernetes-practical-guide/>

## Apache Kafka vs. Confluent

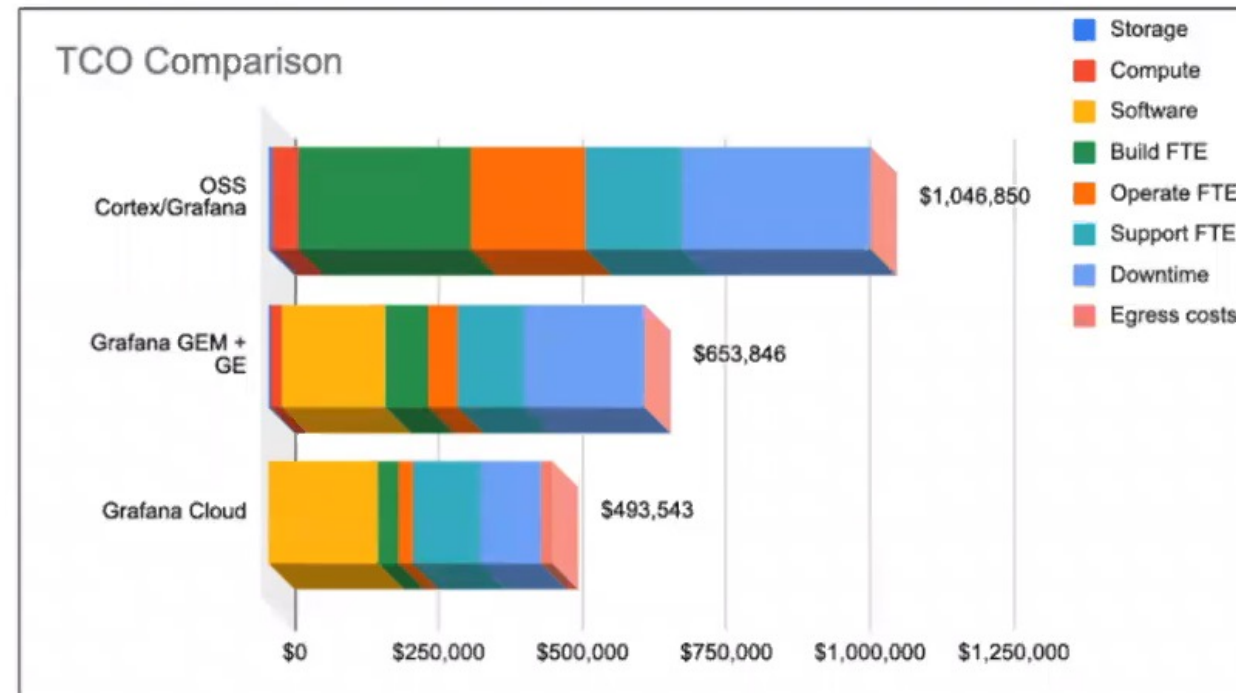


# Grafana



# Grafana TCO comparison

- Secondo Gartner: «TCO includes hardware and software acquisition, management and support, communications, end-user expenses and the opportunity cost of downtime, training and other productivity losses.»
- Le soluzioni On-Prem di Grafana OSS e Enterprise, rispetto a Grafana Cloud, prevedono costi sommersi, quali:
  - Hardware: storage, CPU cores (e relativa mancanza di scalabilità)
  - FTE (Full Time Employee):
    - installazione
    - sviluppo
    - gestione della sicurezza
    - maintenance
  - COD (Cost of Downtime)



# Grafana OSS Terms of Service



- L'utilizzo di software open source in ambito commerciale impone particolare attenzione nel non violare i termini di servizio.
- A partire da aprile 2021 Grafana OSS viene distribuito con licenza AGPL (Affero General Public License) v3, che rispetto alla precedente Apache License 2.0, implica maggiori restrizioni:  
<https://grafana.com/blog/2021/04/20/qa-with-our-ceo-on-relicensing/>



# Grafana Editions Features



## Grafana Open Source

- On-Premises
- Community & Labs free plugins
- Autenticazione LDAP
- Autenticazione Oauth

## Grafana Enterprise

- On-Premises
- Enterprise plugins (tutti inclusi)
- Autenticazione Enh. LDAP (sync)
- Autenticazione SAML
- Dashboard usage insights
- Query Caching
- White Labeling
- Supporto

## Grafana Cloud

- Cloud
- Enterprise Plugins (Cloud Advanced License per MongoDB)
- Autenticazione Enh. LDAP (sync)
- Autenticazione SAML
- Dashboard usage insights
- Alerting (Grafana OnCall)
- Query Caching
- White Labeling
- Supporto

# Grafana Pricing



## Grafana Enterprise

- Minimo 60 utenti: 40K\$/anno
- Ogni utente aggiuntivo: 300\$
- Tutti i plugins inclusi
- 3 stacks inclusi (dev-test-prod)

## Grafana Cloud Advanced

- Minimo 60 utenti: 25K\$/anno
- Necessario per plugin MongoDB
- Costo aggiuntivo per extra plugins

## Grafana Cloud Pro

- Minimo 10 utenti: 49\$/mese  
(per confronto con 60 utenti: 3.6K\$/anno)
- Ogni utente aggiuntivo: 5\$/mese
- 15K metriche, 100GB logs, 100GB tracce
- Billing extra su volume dati extra
- No plugin MongoDB

Non vi è più distinzione tra utenti “admins” e “viewers”: ora sono tutti utenti generici (per il pricing)

# Considerazioni sul deployment on-premise dei Container

Una importante questione da tenere in considerazione per i progetti che partono oggi in ambienti on premise basati su Docker riguarda i costi di licenza.

[Docker is Updating and Extending Our Product Subscriptions - Docker Blog](#)

Gartner ha chiaramente riportato in questa analisi i costi aggiuntivi (da aggiungere ad esempio al costo delle licenze del Sistema Operativo) da tenere in considerazione per progetti che ospitano all'interno dei container software open source.

<https://www.gartner.com/doc/4005890>

Docker will require a license for commercial use from 31 January 2022.  
Software engineering leaders must assess their teams' Docker usage and purchase the required licenses or move to alternative solutions immediately.



**Thank you**

[Info@bitbang.com](mailto:Info@bitbang.com)

**Bologna**

Via E. Mattei 102,  
40138, Bologna, BO  
+39 051 58 75 314

P.IVA 02329121202  
Capitale Soc: € 60.000  
R.E.A. BO 431512

**Milano**

Via S. Marco 21,  
20121, Milano, MI  
+39 02 12 41 23 352