DEFINIZIONE DEI QUATTRO ARCHETIPI DELL'EDGE E PRESENTAZIONE DEI RISPETTIVI REQUISITI TECNOLOGICI

Introduzione

Negli ultimi anni l'"edge computing" è diventato uno degli argomenti più dibattuti nel mondo IT per dei validi motivi. Grand Valley Research prevede un **CAGR (tasso di crescita annuale composto) del 41% per l'edge computing** tra il 2018 e il 2025. Quasi ogni settore riconosce i limiti nel supportare gli utenti e nelle tecnologie emergenti adottate per le infrastrutture IT centralizzate e sta spingendo per avvicinare storage e capacità elaborativa verso gli utenti e i dispositivi.

Le ragioni di questo cambiamento vanno ricercate nella maggiore connettività dei dispositivi e delle persone e negli enormi volumi di dati che vengono prodotti. Secondo il **Cisco Visual Networking Index** nel 2016 il traffico IP globale è stato di 1,2 zetabyte. Entro il 2021 quasi triplicherà, raggiungendo 3,3 zetabyte. Sempre entro il 2021, Cisco prevede che il numero di dispositivi connessi alle reti IP sarà tre volte la popolazione mondiale. Stiamo parlando di qualcosa come oltre 23 miliardi di dispositivi connessi in appena tre anni. **Altre società hanno previsioni simili**: entro il 2020 Gartner prevede 20,8 miliardi di dispositivi connessi, IDC 28,1 miliardi e IHS Markit 30,7 miliardi.

Una grande percentuale di questi dati IoT (Internet of Things) sarà costituita dai dati dei sensori mobili che devono essere trasmessi su reti wireless o mobili invece che tramite connessioni Internet cablate, fatto che mette sotto pressione l'infrastruttura della rete mobile. **Si prevede che il traffico IP mobile aumenti di sette volte entro il 2021**, raddoppiando il ritmo della crescita del traffico IP fisso.

I cambiamenti dell'infrastruttura di computer e di storage richiesti per supportare un futuro smart e connesso, specie a livello locale, saranno importanti.

Tuttavia, se si analizzano le informazioni disponibili oggi sull'edge computing, si scopre che nella migliore delle ipotesi sono poche le risorse che offrono una visione completa dell'ecosistema edge. Un approfondimento del mercato rivela una grande varietà di casi d'uso correnti ed emergenti e, pur condividendo alcuni aspetti comuni basati sull'ampia definizione dell'edge computing, le differenze non sono trascurabili.

Vertiv ha analizzato i casi d'uso che comprendono l'ecosistema edge, per sviluppare una migliore comprensione di queste differenze e delle loro implicazioni per l'infrastruttura di supporto. Come risultato di questa analisi, sono stati identificati quattro principali archetipi per le applicazioni edge:

- Ad Uso Intensivo di Dati
- Sensibili alla Latenza Umana
- Sensibili alla Latenza da Macchina a Macchina
- Life Critical

Questo documento presenta una descrizione di ciascun archetipo con esempi dei casi d'uso di maggior impatto, insieme a una panoramica dei loro requisiti di connettività agli hub locali, urbani e regionali, che rappresentano il livello e il core della trasmissione edge e sono a volte differenziati come edge, fog e cloud computing.

Comprensione dei casi d'uso dell'edge

Per identificare i quattro archetipi, è stato prima necessario comprendere i casi d'uso della tecnologia edge. Il team di ricerca Vertiv ha identificato ed esaminato più di 100 casi d'uso per la tecnologia edge e, per un'analisi più dettagliata, ha ristretto questo elenco ai 24 che avranno il maggiore impatto sull'infrastruttura IT.

L'analisi ha esaminato i requisiti di performance di ogni caso d'uso in termini di latenza, disponibilità e crescita prevista, nonché i requisiti di sicurezza come la necessità di crittografia, autenticazione e conformità alla normativa. È stata inoltre valutata la necessità di integrarsi con applicazioni esistenti o legacy e altre fonti di dati e il numero di potenziali location richieste per supportare il caso d'uso.

Il team ha studiato soprattutto le caratteristiche dei dati dei casi d'uso e ha scoperto che le applicazioni che sostengono ciascuno, oltre ai requisiti di disponibilità e sicurezza, hanno un insieme di requisiti di workload incentrato sui dati. Ad esempio, il volume dei dati, la modalità di accesso ai dati, i requisiti di trasmissione dei dati, l'integrità dei dati e l'analisi dei dati. Questo approccio incentrato sui dati, filtrato attraverso i requisiti di disponibilità e sicurezza, è fondamentale per comprendere e classificare i requisiti dei vari casi d'uso.

Nella Figura 1 è possibile trovare un elenco dei 24 casi d'uso, organizzati in base all'archetipo.

L'ecosistema edge

USO INTENSIVO DI DATI	SENSIBILI ALLA LATENZA DA MACCHINA A MACCHINA	LIFE CRITICAL	SENSIBILI ALLA LATENZA UMANA
• Connettività ristretta • Città intelligenti • Fabbriche intelligenti • Case ed edifici intelligenti • Distribuzione di contenuti HD • Informatica ad alte prestazioni (HPC) • Realtà Virtuale • Digitalizzazione di Petrolio e Gas	• Sicurezza intelligente • Smart Grid (reti elettriche intelligenti) • Distr. di contenuti a bassa latenza • Mercato dell'Arbitraggio • Analisi in tempo reale • Simulazione delle Forze Armate	• Sanità Digitale • Veicoli Connessi / Autonomi • Droni • Trasporto intelligente • Robot autonomi	• Ottimizzazione del sito web • Realtà Aumentata • Vendita al dettaglio Intelligente • Elaborazione della Lingua Naturale (NLP)

Figura 1: *Archetipo*

Archetipo 1: Uso Intensivo di Dati

Larghezza di banda	Latenza	Disponibilità	Sicurezza
Alta	Media	Alta	Media

L'archetipo Uso Intensivo di Dati rappresenta i casi d'uso in cui la quantità dei dati rende poco pratico il trasferimento direttamente nel cloud tramite la rete, o dal cloud al punto d'uso, a causa del volume dei dati, del costo o della larghezza di banda.

Probabilmente l'esempio più discusso di un'applicazione edge ad alta intensità di dati è il recapito di contenuti ad alta definizione. Nel **2016, i video hanno rappresentano il 73% di tutto il traffico IP e si prevede che arriveranno all'82% entro il 2021** dato che il video streaming e la realtà virtuale continuano a crescere. I principali fornitori di contenuti, come Amazon e Netflix, stanno attivamente collaborando con i colocation provider per espandere le proprie reti di fornitura e portare il video streaming ad alto consumo di dati più vicino agli utenti, per ridurre costi e latenza.

Già il, **35% dei contenuti visti da un utente Internet nel Nord America viene inviato dall'area locale in cui l'utente si trova.** Dato che i fornitori di contenuti continuano ad estendere le proprie reti verso l'edge, si prevede che tale percentuale arriverà al 51% entro il 2021. E tuttavia questa è solo la prima ondata di computing dal core all'edge. In presenza della crescita continua della domanda di video ad alta definizione, gli hub locali aumenteranno il supporto degli hub urbani attuali per ridurre ulteriormente i costi della larghezza di banda e i problemi di latenza.

Un altro ottimo esempio dell'archetipo Uso Intensivo di Dati è l'uso delle reti IoT per costruire case, edifici, stabilimenti e città smart. Da un sondaggio del 2018 condotto da 451 Research, Vertiv ha scoperto che mentre solo il 33% delle 700 organizzazioni intervistate aveva implementato strumenti IoT su larga scala, il 56% dichiarava che meno del 25% della loro attuale capacità IT poteva supportare le applicazioni IoT. Malgrado le soluzioni IoT stiano ancora muovendo i primi passi, le organizzazioni stanno già lottando con la gestione del volume di dati generati.

In questo caso, la sfida è l'opposto di quella presentata dalla fornitura di contenuti ad alta definizione. Invece di spostare i dati più vicino agli utenti, queste applicazioni devono portare le enormi quantità di dati generati dai dispositivi e dai sistemi dal punto di origine verso una posizione centrale affinché vengano elaborati. Ciò richiederà l'evoluzione di un'architettura di rete dall'edge al core.

L'IoT e l'Internet delle cose industriale (IIoT) rappresentano una rete di sensori che ogni ora generano volumi enormi di dati. Questi dati supportano un loop di "rilevamento-deduzione-reazione" che permette di vedere e controllare qualsiasi cosa, dagli elettrodomestici alle apparecchiature industriali. Solo un sottoinsieme di questi dati viene trasmesso ad un data center locale, regionale o su cloud per essere ulteriormente elaborato, il che significa che serviranno ingenti calcoli all'estremità dell'edge per permettere ai dispositivi e ai sistemi di prendere decisioni e agire sulla base dei dati forniti dai sensori.

La più semplice di queste applicazioni, la casa smart, deve supportare più dispositivi e sistemi ad uso intensivo di dati, compresi quelli di intrattenimento, HVAC e di sicurezza.

Uso Intensivo di Dati
Secondo IHS Markit, **il mercato mondiale per i dispositivi domestici connessi passerà dagli oltre 100 milioni di unità nel 2017 a circa 600 milioni nel 2021.**

Le città e le fabbriche intelligenti raccolgono le sfide insite nelle case intelligenti e le amplificano. Molte città stanno già sperimentando o testando la tecnologia delle città smart per migliorare i flussi del traffico, supportare i servizi di emergenza e ridurre i costi.

Le fabbriche intelligenti, che sfruttano la convergenza di IoT, sistemi cyber-fisici e cloud computing per consentire ai produttori di utilizzare dati in tempo reale per aumentare l'efficienza, ridurre i costi e adattarsi ai cambiamenti della domanda, vengono promosse come la prossima rivoluzione industriale. Secondo McKinsey, le fabbriche e gli altri ambienti di produzione hanno il potenziale per realizzare il più grande impatto finanziario dall'applicazione di IoT. Prevedono che l'IoT genererà un **valore economico tra $1,2 e $3.7 trilioni** entro il 2025. Questo valore verrà da nuove efficienze energetiche, produttività della manodopera, ottimizzazione dell'inventario e maggiore sicurezza dei lavoratori. Ma ottenerlo richiederà un'infrastruttura locale robusta.

Nell'industria del petrolio e del gas, la digitalizzazione ha già creato un enorme miglioramento nell'efficienza dei processi di esplorazione ed estrazione, ma ha anche introdotto enormi sfide nella gestione dei dati. Un singolo impianto di perforazione può generare terabyte di dati al giorno.

Altri casi d'uso che rientrano nell'archetipo Uso Intensivo di Dati includono realtà virtuale, calcolo ad alte prestazioni e ambienti con connettività limitata, come le aree in cui si

svolgono le operazioni di recupero in seguito a un disastro naturale o a un attacco informatico.

Ciò che accomuna tutti queste situazioni reali è la necessità di spostare grandi volumi di dati verso gli utenti che possono utilizzarli, o da dispositivi e sistemi in cui vengono generati verso un repository centrale.

Archetipo 2: Sensibili alla Latenza Umana

Larghezza di banda	Latenza	Disponibilità	Sicurezza
Media	Alta	Media	Alta

L'archetipo Sensibili alla Latenza Umana copre i casi d'uso in cui i servizi sono ottimizzati per il consumo umano. Come suggerisce il nome, la caratteristica distintiva di questo archetipo è la velocità.

La sfida della Latenza Umana può essere vista nel caso d'uso di ottimizzazione dell'esperienza del cliente. In applicazioni come l'e-commerce, la velocità ha un impatto diretto sull'esperienza dell'utente; i siti web ottimizzati per la velocità utilizzando l'infrastruttura locale si traducono direttamente in maggiori visualizzazioni delle pagine e in una crescita delle vendite.

Sensibili alla Latenza Umana
Google ha scoperto che l'aggiunta di un ritardo di 500 millisecondi ai tempi di risposta delle pagine comportava un calo del traffico del 20%, mentre Yahoo ha rilevato che un ritardo di 400 millisecondi portava a una diminuzione del traffico tra il 5 e il 9%.

Questo effetto si estende anche all'elaborazione dei pagamenti. Amazon ha rilevato che un ritardo di 10 millisecondi nell'elaborazione dei pagamenti causava una diminuzione dell'1% degli incassi. L'approvazione centralizzata tramite password richiedeva, in media, 7 secondi. Un passaggio all'elaborazione locale ha abbassato i tempi a 600 millisecondi, un miglioramento di 6.400 millisecondi in cui ogni 100 millisecondi potenzialmente risultava in un ulteriore 1% di incassi.

Un altro esempio emergente di un'applicazione Sensibile alla Latenza Umana è l'elaborazione del linguaggio naturale. La voce è probabilmente destinata a diventare la forma principale di interazione con le applicazioni IT quotidiane. Attualmente l'elaborazione del linguaggio naturale per Alexa e Siri avviene nel cloud. Tuttavia, poiché il volume di utenti, applicazioni e lingue supportate aumenta, sarà necessario migrare queste funzionalità più vicino agli utenti.

Altri casi d'uso della Latenza Umana identificati includono la vendita al dettaglio intelligente, come i negozi Amazon Go senza cassiere e tecnologie immersive come la realtà aumentata, dove piccoli ritardi di latenza possono fare la differenza tra divertimento e fastidio. In ogni caso, i ritardi nell'erogazione dei dati incidono direttamente sull'esperienza tecnologica di un utente, come con l'elaborazione del linguaggio e la realtà aumentata, o le vendite e la redditività di un rivenditore come accade con l'ottimizzazione del sito web e la vendita al dettaglio intelligente. Man mano che questi casi d'uso aumentano, aumenterà anche la necessità di hub di elaborazione dei dati locali.

Archetipo 3: Sensibili alla Latenza da Macchina a Macchina

Larghezza di banda	Latenza	Disponibilità	Sicurezza
Media	Alta	Alta	Alta

L'archetipo Sensibili alla Latenza da Macchina a Macchina copre i casi d'uso in cui i servizi sono ottimizzati per il consumo da macchina a macchina. Poiché le macchine possono elaborare i dati molto più velocemente dell'uomo, la caratteristica distintiva di questo archetipo è la velocità. Le conseguenze di un mancato recapito dei dati alle velocità richieste possono essere anche maggiori in questo caso rispetto all'archetipo Sensibili alla Latenza Umana.

Ad esempio, i sistemi utilizzati nelle transazioni finanziarie automatizzate, come il trading su materie prime e azioni, sono sensibili alla latenza. In questi casi, i prezzi possono cambiare in millisecondi e i sistemi che non dispongono dei dati più recenti quando è necessario non possono ottimizzare le transazioni, trasformando potenziali guadagni in perdite.

Latenza M2M
Secondo uno studio del Tabb Group, un broker potrebbe perdere **anche $4 milioni per millisecondo** se la sua piattaforma di trading elettronica fosse 5 millisecondi indietro rispetto alla concorrenza.

Anche la tecnologia delle reti intelligenti rientra in questo archetipo. Questa tecnologia viene implementata nella rete di distribuzione elettrica per bilanciare l'offerta e la domanda e gestire l'uso di elettricità in modo sostenibile, affidabile ed economico. Consente alle reti di distribuzione di

"auto-guarirsi", ottimizzare i costi e gestire le fonti energetiche intermittenti, presumendo che i dati giusti siano disponibili al momento opportuno.

Altre applicazioni Sensibili alle Latenze da Macchina a Macchina includono sistemi di sicurezza intelligenti basati sul riconoscimento di immagini, simulazioni di guerra militare e analisi in tempo reale.

Archetipo 4: Life Critical

Larghezza di banda	Latenza	Disponibilità	Sicurezza
Media	Alta	Alta	Alta

L'archetipo Life Critical comprende casi d'uso che hanno un impatto diretto sulla salute e sulla sicurezza umana. In questi casi d'uso, la velocità e l'affidabilità sono imprescindibili.

Probabilmente i migliori esempi di archetipo Life Critical sono i veicoli autonomi e i droni, che offrono grandi vantaggi quando funzionano come previsto. Tuttavia, se prendono decisioni sbagliate, possono mettere in pericolo la salute umana.

I progressi nei veicoli autonomi sono stati più veloci di quanto molti si aspettassero, e diverse aziende automobilistiche e tecnologiche li stanno già collaudando su strada. Nella maggior parte di questi veicoli, per ridurre al minimo il rischio per la salute umana, al posto di guida siede una persona pronta a escludere i controlli automatici in caso di problemi. Ma, nel prossimo futuro, su strada circoleranno veicoli di consegna e sistemi di trasporto senza conducente. Se questi sistemi non hanno i dati di cui hanno bisogno nel momento in cui servono, le conseguenze potrebbero essere disastrose.

Lo stesso vale per i droni. Non è difficile immaginare un futuro in cui centinaia di droni di consegna sorvoleranno le città.

Life Critical
Grandi aziende di e-commerce e consegna pacchi, come Amazon e DHL, stanno già sperimentando i droni per il recapito.

Anche l'aumento dell'uso della tecnologia nel settore sanitario rappresenta un archetipo Life Critical. Le cartelle cliniche elettroniche, la medicina cibernetica, la medicina personalizzata (mappatura del genoma) e i dispositivi di auto-monitoraggio, stanno ridisegnando l'assistenza sanitaria e generando enormi volumi di dati.

Altri esempi includono il trasporto intelligente e i robot autonomi. I settori dei trasporti e della logistica stanno prendendo in considerazione soluzioni incentrate sui dati per migliorare la sicurezza dei conducenti e dei passeggeri, l'efficienza dei consumi e la gestione delle risorse. La tecnologia in questo spazio includerà sistemi di trasporto intelligenti, gestione della flotta e telematica; sistemi di guida e controllo, applicazioni di intrattenimento dei passeggeri e di commercio, sistemi di prenotazione, pedaggio e biglietteria e sistemi di sicurezza e sorveglianza.

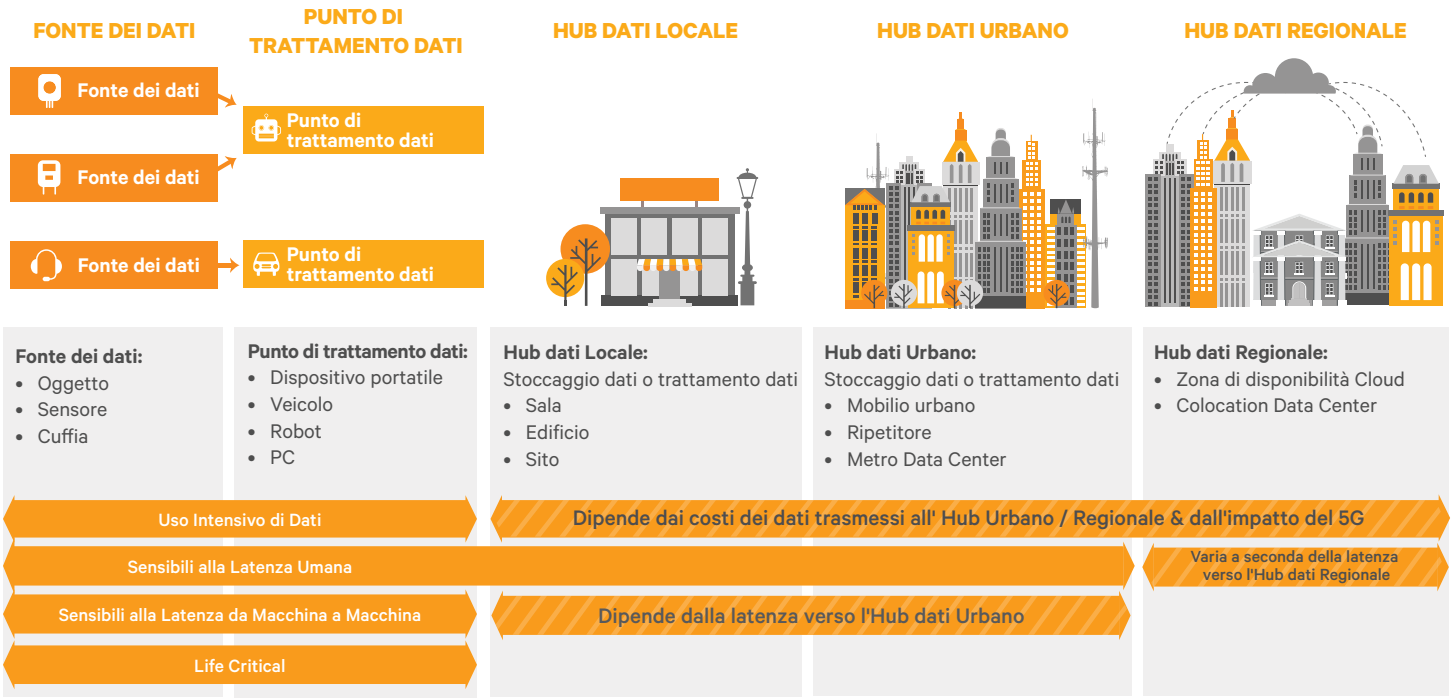
Requisiti tecnologici per Hub Locali e Regionali

L'infrastruttura necessaria per supportare questi casi d'uso attuali e consolidati consiste in quattro livelli di archiviazione ed elaborazione, oltre all'infrastruttura di comunicazione necessaria per spostare i dati tra i livelli.

Alla base, tipicamente troviamo il dispositivo che genera o consuma dati e un endpoint di elaborazione. Il dispositivo potrebbe essere un sensore che monitora qualsiasi cosa, dallo stato di alimentazione di una lampada all'accesso a una porta, dalla temperatura di una stanza ad altre informazioni desiderate. L'endpoint di elaborazione può essere semplice come il PC o il tablet a cui un utente trasmette video o potrebbe essere costituito dai microprocessori incorporati in automobili, robot o dispositivi indossabili. Questi componenti dipendono dall'applicazione e sono generalmente progettati dal produttore dell'apparecchiatura o adattati a dispositivi esistenti.

Ogni archetipo, ad eccezione di quello Life Critical, potrebbe essere inserito nell'hub dati locale a seconda dell'applicazione. L'hub dati locale fornisce archiviazione ed elaborazione in prossimità della sorgente. In alcuni casi, l'hub locale può essere un data center indipendente. Più comunemente, sarà un sistema basato su rack o fila di rack con una capacità di 30-300 kW in un armadio integrato che può essere installato in qualsiasi ambiente.

Questi sistemi di armadi basati su rack e su file integrano comunicazioni, elaborazione e archiviazione con un'adeguata protezione dell'alimentazione, controlli ambientali e sicurezza fisica. Per gli archetipi che richiedono un alto grado di disponibilità, come quelli Sensibili alla Latenza da Macchina a Macchina e Life Critical, l'hub locale dovrebbe includere sistemi di alimentazione ridondanti ausiliari ed essere in grado



di consentire la gestione e il monitoraggio da remoto. Molti casi d'uso richiedono anche la crittografia dei dati e altre funzionalità di sicurezza all'interno dell'hub locale.

Per tutti gli archetipi eccetto Life Critical, l'hub urbano e/o regionale potrebbe essere utilizzato per supportare i casi d'uso in base ai costi di trasferimento dei dati, alla larghezza di banda consentita dalla distribuzione 5G e alla latenza fino alla posizione fisica del data center. L'hub urbano sfrutta l'infrastruttura di telecomunicazioni esistente per fornire funzionalità di calcolo e infrastruttura. Sarà progettato per gli standard di telecomunicazione, compresi la corrente DC e il raffreddamento ad aria, con supporto per un intervallo di temperatura e umidità molto più ampio rispetto a quello dei data center tradizionali. L'hub regionale sarà probabilmente un data center su cloud o di colocation che opera nella stessa regione dell'hub locale e urbano.

Sia per gli hub urbani che per quelli regionali, i progetti modulari in grado di scalare facilmente oltre le specifiche di progettazione iniziali dovrebbero tener conto di imprevisti picchi della domanda. Queste strutture dovrebbero inoltre essere progettate per scalare in termini di densità. Le applicazioni ad alta intensità di immagini, come la realtà virtuale e le applicazioni ad alta intensità di elaborazione, ad esempio analisi e apprendimento automatico, richiederanno probabilmente densità di rack superiori alla tipica specifica di progetto di 10 kW. In quasi tutti i casi, questi hub dovrebbero fornire almeno lo stesso livello di ridondanza e sicurezza dell'hub locale.

Lo sguardo rivolto al futuro

Identificando le esigenze di workload per i ventiquattro casi d'uso discussi, sono emersi i quattro archetipi principali che possono guidare le decisioni riguardanti i requisiti di infrastruttura e configurazione per i casi d'uso analizzati e per quelli che emergeranno nei prossimi anni. Vertiv partirà da questo lavoro iniziale sugli archetipi per definire ulteriormente i requisiti e le configurazioni tecnologiche specifiche per ciascun archetipo.

VertivCo.it | Vertiv Srl, Via Leonardo da Vinci 16 -18, 35028 Piove di Sacco (PD), Italia, CF- P. IVA IT00230510281

SL-11490-IT (R04/18)