

ALLEGATO B

P.O.R FSE 2007-2013 OBIETTIVO COMPETITIVITÀ REGIONALE E OCCUPAZIONE

Asse IV Capitale umano

Linee di Attività I.1.1

BANDO BORSE DI RICERCA,, DI DURATA BIENNALE, A FAVORE DI DOTTORI DI RICERCA/RICERCATORI.

PRIMA FASE: MANIFESTAZIONE DI INTERESSE DA PARTE DI IMPRESE CON SEDE LEGALE E/O OPERATIVA NEL TERRITORIO REGIONALE

SCHEDA TECNICA RELATIVA ALLA MANIFESTAZIONE DI INTERESSE DA PARTE DI IMPRESE CON SEDE LEGALE E/O OPERATIVA NEL TERRITORIO REGIONALE

La scheda tecnica non può essere modificata, se non in termini di descrizione delle attività e degli obiettivi da perseguire.

*Anche nel caso di rete di imprese, dovrà essere redatta un'unica **scheda tecnica**, sottoscritta dai rappresentanti legali di ciascuna impresa partner, nonché dall'impresa con funzioni di capofila.*

DESCRIZIONE GENERALE DELLA SCHEDA TECNICA

Ambito di riferimento (art 6, punto 7.3 del Bando)	A. Tecnologie dell'informazione e della comunic...
Titolo della scheda tecnica	Sviluppo di una piattaforma di calcolo parallelo per analisi di reti complesse con utilizzo di sistemi di cloud computing per lo studio e il monitoraggio di network socio-economici e biologici

B.1 Situazione attuale

(Illustrare le attività caratterizzanti l'impresa/e. Nel caso di rete di imprese dovranno essere illustrate le attività caratterizzanti ciascuna impresa appartenente alla rete – Massimo 5.000 caratteri)

Il Laboratorio di ricerca Linkalab è stato fondato nel gennaio 2008 da ricercatori delle Università di Cagliari e Sassari e opera nell'ambito delle Teorie dei Sistemi e delle Reti Complesse che negli ultimi dieci anni hanno rivoluzionato l'approccio ai più diversi fenomeni scientifici, sociali e tecnologici e hanno fornito le basi per nuove applicazioni, utili per la ricerca, ma anche per l'impresa.

Quasi ogni problema può avere una valida formalizzazione come rete: utilizzando gli strumenti di indagine della Teoria delle Reti Complesse oggi si possono ottenere informazioni non accessibili con i classici strumenti statistici. L'esempio più noto in campo applicativo è il PageRank, che ha decretato il successo di Google, ma le applicazioni spaziano dallo studio della diffusione delle epidemie (sanitarie o informatiche), allo studio delle reti infrastrutturali (stradali, elettriche...), campo nel quale Linkalab ha pubblicato su riviste di rilievo internazionale (Federica Cerina, Vincenzo De Leo, Marc Bartélemy and Alessandro Chessa, "Spatial Correlations in Attribute Communities", PLoS ONE, Vol. 7, No. 5, 2012), per arrivare alla caratterizzazione delle reti genetiche (Nature Methods, "Wisdom of crowds for robust gene network inference" 2012 Jul 15;9(8):796-804. doi: 10.1038/nmeth.2016) e alle più recenti ricerche sui sistemi sociali di internet del Web 2.0. Linkalab è al centro di questa rivoluzione: ha sviluppato infatti gli strumenti teorici e di calcolo necessari, con standard qualitativi di livello internazionale. In questo senso si propone come service di ricerca rivolto alle problematiche di Complex Systems Analysis di realtà istituzionali e aziendali.

Linkalab è attivamente impegnata nel progetto FOC (Forecasting Of Crisis, <http://www.focproject.eu>) per il quale ha fornito supporto tecnologico e progettuale a vari livelli. Dal lato web, attraverso la creazione del frontend Drupal, e per ciò che riguarda gli aspetti di interaction design della parte editoriale con sistemi avanzati di gestione delle liste di abstract e pubblicazioni online. Infine Linkalab ha progettato e realizzato in dettaglio la parte di web visualization e user interaction per visualizzare i dati degli esperimenti condotti dai ricercatori del nodo ETH Zurigo di FOC. In questo momento la società sta attivamente lavorando per la realizzazione di una versione mobile (una mobile factory) di questi sistemi di visualizzazione interattiva degli esperimenti.

Un altro importante progetto che vede Linkalab coinvolto come protagonista è TETRIS (PON Nazionale su tematiche di Smart Objects e Smart Cities). All'interno di questo progetto Linkalab ha sviluppato un'attività di Ricerca Industriale volta allo studio delle reti infrastrutturali elettriche e del pendolarismo con tecniche di Reti Complesse e ha portato avanti una sperimentazione di monitoraggio territoriale tramite un'applicazione mobile (Green Steps) che, tracciando il comportamento dei pendolari, ha consentito di generare delle mappe territoriali con l'indicazione delle emissioni di CO2 e dei consumi energetico.

Sono inoltre stati svolti i seguenti progetti applicativi, come sbocco delle attività di ricerca del Laboratorio:

- Collaborazione come terza nel progetto "Hamatole Web 2.0" dell'azienda Logitec Service Snc a valere sui Pacchetti Integrati di Agevolazioni Industria Artigianato e Servizi 2007 della Regione Sardegna. Il progetto prevedeva la presentazione di dati acquisiti da diverse fonti e dispositivi hardware verso device mobili e sul web utilizzando tecniche di interazione e coinvolgimento degli utenti basate sul paradigma del web 2.0. All'interno del progetto sono state sviluppate tecniche di calcolo su cloud (es. transcodifica, georeferenziazione, ricerca e indicizzazione) i cui risultati venivano presentati su dispositivi connessi con basso potere computazionale.

- Collaborazione come terza nel progetto "Hamatole Web Service" dell'azienda Interbusiness Italia Srl a valere sui Pacchetti Integrati di Agevolazioni Industria Artigianato e Servizi 2007 della Regione Sardegna. Il progetto prevedeva la creazione di un sistema di data integration in grado di acquisire dati da diverse fonti attraverso l'utilizzo di tecnologie basate su web service e tecniche di harvesting, elaborare tali informazioni e riproporle verso altri sistemi mediante l'utilizzo di tecnologie e protocolli standard.

B.2. Illustrazione della scheda tecnica

(Analisi e proposta delle nuove attività da realizzare rispetto agli obiettivi da perseguire. Illustrazione delle criticità attuali – Massimo 5.000 caratteri)

Nell'ultimo decennio l'esplosione dei blog, dei social network e dei siti d'informazione online ha fatto emergere una problematica legata ai cosiddetti "Big Data", cioè volumi di dati così grandi da non essere gestibili con database tradizionali, di natura molto diversa tra loro e non strutturati come testi, file audio, file video, flussi di click, segnali provenienti da cellulari, transazioni commerciali o altro e che affluiscono e necessitano di essere processati in tempo reale.

Twitter, ad esempio, nel 2010 generava un flusso di dati utente pari a 7 TB al giorno (vedi <http://www.slideshare.net/kevinweil/big-data-at-twitter-chirp-2010>) con la problematica che, alla velocità massima di 80 MB/s degli hard disk, è necessario più di un giorno per memorizzarli; dal punto di vista dell'analisi, inoltre, fare la stima delle amicizie nel grafo sociale di Twitter richiede, utilizzando gli strumenti classici, un calcolo tipo "self-join" su una tabella di diversi miliardi di righe che non è ovviamente fattibile in tempi accettabili.

Sempre nel 2010 Facebook aveva un sistema di memorizzazione dei dati con una dimensione pari a 21 PB e, con l'uso di nuovi strumenti di memorizzazione e analisi basati su Cloud Computing, era in grado di analizzare fino a 800 TB di dati compressi al giorno.

Anche in campo biologico si sta delineando una situazione simile. Vi è una produzione massiva di dati genetici dai sequenziatori di DNA e RNA di nuova generazione; inoltre i sistemi di regolazione biomolecolari, che consistono di centinaia di migliaia di specie di molecole interagenti di diversa natura chimica, rappresentano un esempio prototipale di sistema complesso. È sempre più chiaro che la comprensione di un sistema simile può essere raggiunta solo considerando il sistema nella sua interezza, adottando cioè un "approccio sistemico".

La presente congiuntura è caratterizzata dall'incrociarsi di vari fattori tecnologici e di tendenze scientifiche legate alle nuove infrastrutture di calcolo e ai più recenti paradigmi per la soluzione di problemi ad alto grado di complessità. Abbiamo da un lato l'affermarsi delle architetture di tipo Cloud Computing, che hanno definito nuovi standard per lo storage dei dati e per il loro processamento. Dall'altra si sta facendo strada una nuova metodologia di indagine dei sistemi caratterizzati dall'interazione complessa di molti elementi: la Teoria delle Reti Complesse.

La sfida dei "Big Data" quindi non è solo legata all'enorme dimensione dei dati ma anche al modo in cui devono essere analizzati. Da questo punto di vista la teoria delle Reti Complesse offre un nuovo paradigma efficace per mappare l'eterogeneità delle relazioni tra gli elementi di questi dati e fornisce un potente strumento per estrarre il maggior numero di informazioni possibili da essi. La chiave di una strategia computazionale di successo è nell'implementare gli strumenti e algoritmi tipici delle Reti Complesse in un'architettura di Cloud Computing, nativamente capace di gestire ed elaborare grandi quantità di dati.

L'obiettivo che ci proponiamo di realizzare in questo progetto è di ottimizzare e sviluppare varie categorie di

algoritmi delle Reti Complesse utilizzando la piattaforma Hadoop, una piattaforma open source per il calcolo su Cloud usata dai più grandi siti web e social network quali Facebook e Twitter per conservare e gestire i loro dati. L'utilizzo di questa piattaforma per questo tipo di analisi permetterebbe, ad esempio, di analizzare grosse quantità di dati salvati nel servizio di internet storage di Amazon "S3" (vedi <http://aws.amazon.com/s3/>), utilizzato dai più grandi social network come Facebook, senza dover fare una copia locale dei dati ma copiando il codice di analisi sul servizio di storage stesso.

Il prodotto finale di questo lavoro sarà un sistema che permetterà, in modo automatico, l'analisi periodica di campioni di dati relativi a particolari intervalli di osservazione (un giorno, una settimana...) con la produzione di tabelle e grafici facilmente esportabili come widget sia su siti web che su applicazioni per dispositivi mobili per il monitoraggio periodico del comportamento dei parametri di interesse per le analisi che un committente potrebbe richiedere di svolgere.

Gli obiettivi operativi sono sintetizzabili nei seguenti punti:

- ricostruire lo stato dell'arte e gli scenari di business sulle piattaforme di calcolo distribuito
- ripensare i principali algoritmi utilizzati nelle reti complesse secondo il paradigma map-reduce/BSP e progettare la piattaforma che dovrà ospitare i calcoli sotto il profilo infrastrutturale e applicativo
- realizzare l'infrastruttura di calcolo e testare su di essa gli algoritmi, studiandone le performance
- gestire i dati da analizzare e la loro raccolta nella piattaforma di calcolo attraverso opportuni strumenti informatici
- applicare gli algoritmi e la piattaforma ai problemi individuati

B.3 Innovazione e prospettive di mercato

(Illustrazione dei livelli di innovatività individuati rispetto alle nuove attività da realizzare anche in termini di acquisizione di livelli più competitivi e di inserimento nel mercato - – Massimo 5.000 caratteri)

Allo stato attuale non esiste nessuna libreria software che metta a disposizione degli utenti una serie di metodi già pronti che utilizzino il paradigma Map/Reduce (paradigma chiave del calcolo parallelo su Cloud) per risolvere problemi tipici delle reti complesse. Lo studio che si vuole portare avanti non solo permetterebbe all'utilizzatore di avere a disposizione un potentissimo e avanzato strumento di Business Intelligence, ma sarebbe anche di facile utilizzo grazie a una semplice interfaccia scritta con un linguaggio di programmazione dinamico come Python. Inoltre l'automazione del sistema permetterebbe l'aggiornamento automatico dei risultati dell'analisi (sia in termini di grafici che di tabelle) in seguito a modifiche sui dati analizzati (se ad esempio questi fossero immagazzinati in un database) come già si sta cercando di fare in importanti progetti internazionali come il progetto FOC (vedi <http://www.focproject.net/>).

Non essendo il presente progetto legato all'introduzione di nuovi prodotti/servizi o all'innovazione di processi produttivi, non è possibile introdurre degli indicatori di natura finanziaria o fisica che misurino eventuali risparmi o economie di scala nei processi produttivi. Piuttosto è opportuno introdurre degli indicatori di realizzazione progettuale che siano in diretta relazione con gli obiettivi operativi e che quindi possano misurare l'impatto degli stessi sulle attività progettuali e sugli asset aziendali che in un momento successivo possano avere un risultato di mercato. Gli indicatori individuati sono i presenti:

- Una migliore definizione del posizionamento strategico di Linkalab nel mercato dei servizi avanzati di data mining
- Una accresciuta potenzialità di costruire nuovi prodotti/servizi
- La capacità di dispiegare una piattaforma di test completa e quindi potenzialmente la capacità dell'azienda di erogare un servizio

Il risultato finale specifico di questo progetto consisterà in documenti, librerie software, configurazioni sistemiche che possano nel loro complesso dare luogo in una seconda fase a un prototipo che possa essere testato sul campo. Anche attraverso l'utilizzo di test all'interno del Laboratorio Linkalab, potremo verificare la praticabilità dell'erogazione di un servizio verso un potenziale cliente della piattaforma, raccogliendo le impressioni di utilizzo e i suggerimenti per una correzione della modalità di interazione e per un arricchimento delle caratteristiche di base del servizio.

Le potenzialità di una piattaforma come quella sopra descritta risiedono principalmente nella sua capacità di rispondere a una sfida concreta che sta emergendo prepotentemente in questi ultimi anni, facendo leva contemporaneamente su due settori innovativi: uno nel campo dell'analisi statistica, basato sulla Teoria delle Reti Complesse, che consente di trattare in maniera adeguata la complessità dell'approccio 'sistemico' utilizzato in diversi ambiti scientifici, l'altro che fa leva sulle nuove architetture di cloud computing, pensate per lo storage di grosse quantità di dati.

Sul fronte specifico dei risultati scientifici che sicuramente deriveranno dalle attività di ricerca industriale, prevediamo di diffondere i risultati tramite pubblicazioni che possano anche approdare su archivi e riviste ad accesso aperto (Arxiv, PlosOne...).

I risultati del progetto, oltre ad essere di sicura rilevanza da un punto di vista scientifico, proponendo soluzione ad alcuni problemi scientifici centrali, sono di grande interesse per l'industria, viste le possibili applicazioni in vari ambiti connessi alle scienze della vita.

B.4 Esperienza già acquisita

(Indicazione delle esperienze già maturate in termini di ricerca e sviluppo – Massimo 5.000 caratteri)

Il Laboratorio di ricerca Linkalab si è da subito occupato delle problematiche e delle nuove potenzialità offerte dal calcolo parallelo su Cloud Computing e nel settembre del 2010 ha discusso una presentazione, durante l'incontro satellite "High Throughput Humanities" (HTH) della Conferenza Europea sui Sistemi Complessi (ECCS 2010) tenutasi a Lisbona in Portogallo, nella quale ha presentato un prototipo di sistema di calcolo basato su Cloud Computing con una chiara e semplice interfaccia utente scritta in Python. Questo prototipo è stato sottoposto a diversi test di funzionalità ed è in grado di analizzare la distribuzione del grado dei nodi (cioè il numero di link di ogni nodo) di una rete di circa 8 miliardi di nodi e 300 miliardi di link in meno di un giorno.

Recentemente il Laboratorio Linkalab ha collaborato in un importante progetto per il monitoraggio dei social network per le analisi dello scenario politico in previsione delle elezioni del febbraio 2013. In questo progetto Linkalab fornisce il supporto tecnologico e scientifico per la piattaforma tecnologica denominata Tycho (in collaborazione con IMT, Alti Studi Lucca). Il progetto è volto ad analizzare con tecniche di Complex Networks Analysis la rete twitter ed è stato usato sul sito de La Stampa (<http://voto2013.lastampa.it/>) per le elezioni 2013. In questo sito vengono presentate le analisi delle reti dei politici italiani estratte attraverso le frequenze assolute delle citazioni degli utenti twitter, attraverso le co-occorrenze di due politici dentro uno stesso tweet ed attraverso le rappresentazioni elementari dell'informazione su twitter: gli hashtags. Il progetto è correlato da mappe geografiche esplorabili di alta qualità che geolocalizzano le citazioni sui politici regione per regione. Un widget javascript consente di visualizzare ed esplorare la rete dei politici italiani e di avere informazioni sugli "hottopics" che vengono ogni giorno pubblicati nella rete.

Linkalab ha così raggiunto l'obiettivo di introdurre ulteriori strumenti di analisi e di conoscenza dell'influenza dei politici in rete, oltre ad uno studio approfondito della rete degli utenti che seguono i politici e ne commentano il contenuto.

Altre attività di ricerca hanno riguardato una collaborazione con l'Ospedale Brotzu di Cagliari sul progetto 'Head Movement Patterns': in collaborazione col Centro regionale per i Disturbi Pervasivi dello Sviluppo guidato dal dott. Giuseppe Doneddu è stata avviata una ricerca per individuare i pattern spazio/temporali caratteristici nei movimenti del capo di un gruppo di soggetti autistici e un gruppo di controllo costituito da soggetti con sviluppo tipico. Lo scopo finale di questa sperimentazione era quello di introdurre una nuova metodologia di diagnosi del comportamento autistico, soprattutto nei primi anni di vita.

E' stato inoltre avviato uno studio del World Wide Web con l'introduzione di una nuova equazione per lo studio del PageRank che consente di vedere il WWW come una superficie di potenziale le cui buche sono associate ai siti più importanti. Questo studio ha portato a una pubblicazione su una prestigiosa rivista internazionale e a citazioni sulle più importanti quotidiani nazionali (Il Sole24ore, Corriere della Sera).

Infine è stata avviata una collaborazione con ENEL Ricerche e CNR, nella quale Linkalab offre la propria consulenza scientifica di applicazione della teoria delle reti complesse nell'ambito dello studio delle cosiddette "Smart Grids".

FIRMA del Rappresentante/i legale/iⁱ

Luogo e data

ⁱ La Scheda tecnica deve essere firmata e data dal rappresentante legale per le imprese singole, dal soggetto capofila e dai rappresentanti legali dei soggetti partner nel caso di Rete.