

Data integration and analysis for precision farming

1 Proposta di progetto

Le attività si svolgeranno nell'ambito del progetto *WeLaser* [1]. A seguito della pervasività di sistemi IoT e dell'avanzamento tecnologico, l'agricoltura sta vivendo un forte processo d'innovazione che coinvolge sistemi robotici, algoritmi di *pattern recognition*, e sistemi di predizione per la gestione di risorse e trattamenti agricoli. Solo in Europa vengono utilizzati circa 130 milioni di tonnellate di erbicidi all'anno che persistono nell'ambiente, distruggono piante e insetti benefici per il suolo e producono effetti sulla salute degli animali e degli esseri umani. La sostituzione degli erbicidi con sistemi robotici automatici è in fase di studio [1]. Tale innovazione si nutre di dati eterogenei provenienti da fonti diverse (ad esempio dati robotici, sensoriali, meteo e satellitari), il cui volume raddoppia ogni due anni [2]. I sistemi convenzionali di gestione e analisi dati convenzionali sono obsoleti, e lasciano spazio a sistemi *big data* e allo sviluppo di nuove tecniche analitiche. In questo contesto si aprono numerosi scenari di ricerca, tra cui il supporto alla gestione e all'analisi di dati legati all'agricoltura di precisione, sui quali si sono focalizzate precedenti attività di ricerca ([1, 3, 4]).

Le attività di seguito indicate sono proposte come prosecuzione dei precedenti lavori, allo scopo di integrare i risultati già ottenuti con nuovi contributi di ricerca. In particolare, gli obiettivi sono i seguenti.

- L'integrazione di dati *streaming* (robot, sensori IoT) e storici (*database* operazionali) e la loro memorizzazione in una *smart data platform*.
- Lo sviluppo di servizi di *analytics* abilitanti l'analisi e la supervisione (in modalità *streaming* e *batch*) a supporto di *data scientist* e *stakeholder* esperti di dominio.

2 Integrazione di dati

Introduzione. L'elaborazione dei dati basata su batch comporta la raccolta e la memorizzazione di una grande determinata quantità di dati, quindi l'elaborazione avviene sull'intero gruppo (batch) di dati. L'elaborazione in batch funziona bene per i dati storici che vengono archiviati e a cui si accede periodicamente. In aggiunta, nell'ambito del progetto *WeLaser* [1] è necessaria

l'elaborazione di dati in tempo reale. In altre parole, ogni dato viene elaborato non appena raccolto, con risultati disponibili a bassa latenza.

Attività precedenti. Nelle precedenti attività di ricerca, sono state presentate tecniche per l'orchestrazione e la profilazione di collezioni di dati all'interno di una smart data platform [5] ed è stato introdotto e ottimizzato un approccio per abilitare interrogazioni in un contesto poliglotta, con l'idea di sfruttare l'espressività e l'eterogeneità degli schemi di dati come valore aggiunto, sia in fase di interrogazione che di interpretazione del risultato [6].

Attività proposte. La gestione di robot in tempo reale richiede la creazione di servizi finalizzati alla raccolta/memorizzazione/analisi di dati IoT e di traiettoria. L'assenza di una gestione consistente e scalabile trasforma il *data lake* (i.e., repository di dati all'interno della data platform) in una *swamp* (palude), dove lo sforzo per la gestione e ricerca di informazione diventa maggiore dell'analisi stessa. L'obiettivo di ricerca è quello di (i) estendere il modello di *smart data platform* proposta in [5] per la gestione di dati in stream, e (ii) di estendere il processo di raccolta dei dati al fine di supportare l'integrazione di dati in stream (e.g., supportare l'allineamento di dati provenienti da robot e sorgenti IoT, arricchire i flussi di dati con dati storici con il fine di abilitare meccanismi di *anomaly detection* e *alerting*).

3 Servizi di analytics

Introduzione. Il concetto di smart data platform è stato introdotto per indicare un repository in cui far confluire dati eterogenei provenienti da molteplici fonti. Le sfide da affrontare in questo contesto sono legate all'adozione delle metodologie e delle tecniche quanto più automatiche e automatizzabili per l'estrazione di conoscenza in maniera efficiente.

Attività precedenti. Nella precedenti attività di ricerca [3, 4], progetti finalizzati a supportare il processo decisionale nell'ambito dell'agricoltura di precisione, è stata definita un'architettura adeguata a gestire carichi di lavoro big data e sono stati sviluppati servizi d'estrazione, di pulizia ed d'integrazione per rendere i dati fruibili ad un livello analitico.

Attività proposte. L'obiettivo di ricerca è quello di estendere l'architettura esistente compatibilmente con le funzionalità richieste dal progetto WeLaser [1], tra cui l'integrazione e l'allineamento di dati robotici e IoT su ecosistema FIWARE [7] e la possibilità di abilitare analisi di stream di dati. Tra tali funzionalità, è necessario gestire (i) la definizione di un modello di dati sufficientemente espressivo per rappresentare il dominio applicativo di [1], (ii) la definizione di *key performance indicator* per la valutazione delle performance dei robot e dei modelli di movimento ed estirpazione delle erbacce, e (iii) lo sviluppo di servizi atti alla pianificazione e monitoraggio del robot durante la sua missione.

4 Attività di progetto

Nell'ambito dei due anni di lavoro, l'assegnista sarà coinvolto nelle seguenti attività.

- Nel contesto di integrazione di dati:
 - Aggiornamento sulle soluzioni ad oggi proposte in letteratura.
 - Definizione delle modalità a supporto di una streaming smart data platform.
 - Valutazione sperimentale delle performance e della scalabilità dell'acquisizione e integrazione di dati IoT e robotici.
- Nel contesto di servizi di analytics:
 - Valutazione delle *best practice* e tecnologie streaming disponibili in letteratura.
 - Progettazione ed implementazione dei suddetti servizi di analisi.
 - *Testing* dei singoli componenti e del sistema complessivo in termini di *latency* e risorse computazionali necessarie.

References

- [1] WeLaser, 2020. H2020 “Sustainable weed management in agriculture with laser-based autonomous tool” — n. Grant Agreement 101000256 e n. Cup J32F20001250006. <https://welaser-project.eu/>.
- [2] SINTEF. “Big Data, for better or worse: 90% of world’s data generated over last two years.”. <https://www.sciencedaily.com/releases/2013/05/130522085217.htm>.
- [3] Agro.Big.Data.Science. <http://agrobigdatascience.it/>.
- [4] MoReFarming. <http://www.morefarming.it/>.
- [5] Matteo Francia, Enrico Gallinucci, Matteo Golfarelli, Anna Giulia Leoni, Stefano Rizzi, and Nicola Santolini. Making data platforms smarter with MOSES. *Future Gener. Comput. Syst.*, 125:299–313, 2021.
- [6] Chiara Forresi, Matteo Francia, Enrico Gallinucci, and Matteo Golfarelli. Optimizing execution plans in a multistore. In *European Conference on Advances in Databases and Information Systems*, pages 136–151. Springer, 2021.
- [7] FIWARE. <https://www.fiware.org/>.