Secondo Progetto

Corso di Big Data

Obiettivi

- Risolvere problemi attinenti alle tematiche Big Data
- Sperimentare metodologie di Big Data
- Analizzare e confrontare tecnologie di Big Data

Approccio

- Selezionare un problema di big data
- Individuare uno o più metodi per risolvere il problema
- Scegliere le tecnologie più adatta
- Implementare i metodi con le tecnologie scelte
- Fare delle sperimentazioni (e dei confronti)

Modalità di scelta

- Scelta di un progetto tra i vari argomenti proposti
- I progetti hanno specifiche piuttosto lasche
  - in modo da poter essere scelti da più gruppi
  - ma soprattutto per favorire la libertà di inventiva
- I progetti sono esclusivi
  - non possono esserci due progetti con stesse specifiche
  - non tardate nella scelta del vostro progetto!
- In ogni caso, per l'assegnazione è necessaria l'approvazione del docente
- Scelta tramite una form sul sito Web del corso

TOPIC 1: (ANY KIND OF) DATA ANALYTICS

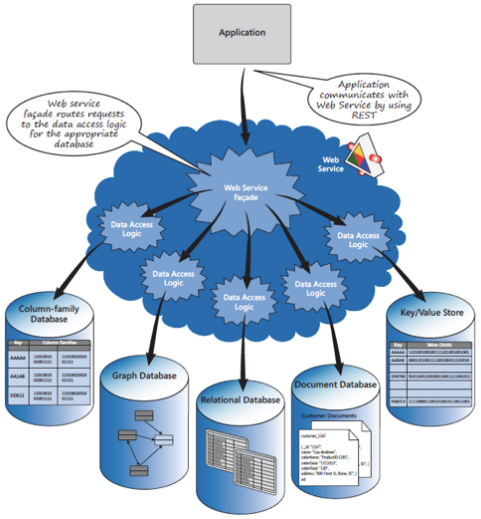
- Cercare autonomamente uno scenario applicativo e dei dati reali (possibili sorgenti in fondo alla presentazione)
- Individuare uno o più metodi di analisi dei dati individuati
- Scegliere la o le tecnologie più adatte (tra quelle viste a lezione ma anche altre, purché siano adatte alla gestione di "big data")
- Implementare i metodi con le tecnologie scelte (in ambiente distribuito)
- Fare delle sperimentazioni discutendo efficacia ed efficienza dei metodi

TOPIC 2: DATA LAKE MANAGEMENT

- Sistemi per gestire collezioni eterogenee di big data con diversi strumenti.
- Data Lake Management Systems
  - Strumenti per la preparazione dei dati (discovery, profiling, cleansing, validation and transformation) con riduzione del "time-to-action"
  - Seminario di Renè Miller ad ATCS

TOPIC 3: POLYGLOT DATA MANAGEMENT

- Creazione di uno o più scenari poliglotti (gestione di dati con sistemi di storage diversi)
- Approccio basato su:
  - “accoppiamento forte”: mediante middleware che offre un’interfaccia comune
  - "accoppiamento lasco": polystores con accesso diretto ai sistemi e query basate su tecniche di navigazione/espansione
- Implementazione e test di accesso alle risorse

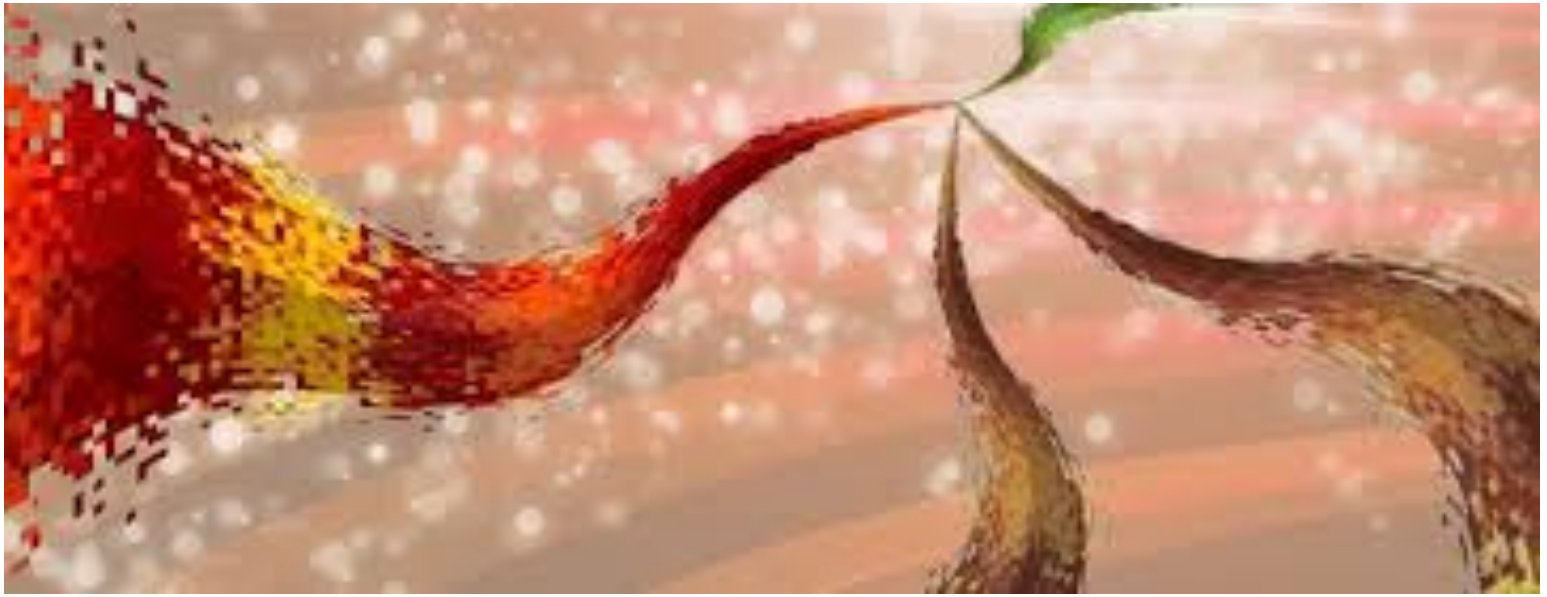


TOPIC 4: SQL vs NoSQL vs NewSQL

- Confronto sperimentale fra sistemi
- SQL
  - PostrgeSQL, MySQL, ..
- NoSQL
  - MongoDB, DynamoDB, Redis, Cassandra,..
- NewSQL
  - Google Spanner, CockroachDB, VoltDB, NuoD, ..

TOPIC 5: ARCHITETTURA LAMBDA

- Creazione di uno o più scenari applicativi (monitoraggio e analisi)
- Sperimentazione di una o più tecnologie che combinano:
  - Stream analysis (es. Apache Flink, Storm, Flume, Kafka, Spark Streaming, FuelDB, Apache Ignite, ecc.)
  - Batch analysis (es. Spark, Hive, Flink, Splunk, ecc.)



Topic 6: GRAPH DATA MANAGEMENT

- Sperimentazione di strumenti per la gestione di "big data graphs"
- Definizione di uno scenario di interesse
  - Knowledge graph,
  - Social network,
  - Recommendation system,
  - Network monitoring.

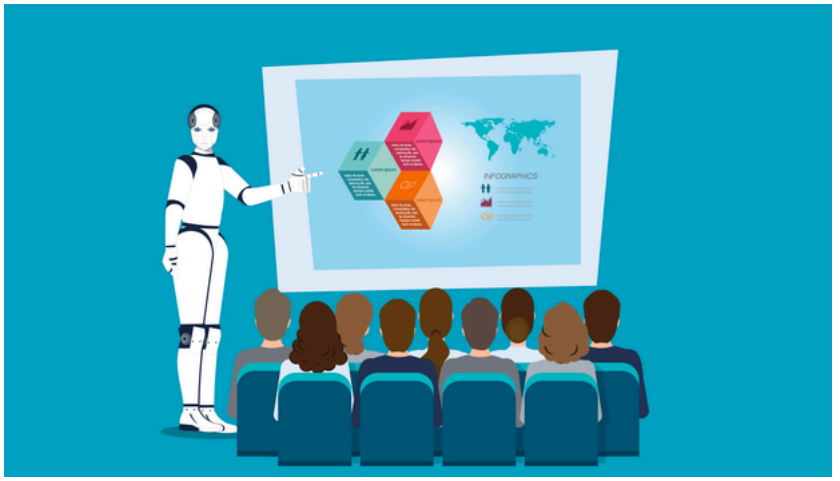
- Uso e confronto di GDBMS e strumenti di analisi su grafo
  - Neo4j, Titan, Cayley graph, Giraph, OrientDB, ArangoDB, …

TOPIC 7: ETHICS-AWARE DATA MANAGEMENT

- Metodi e tecniche per analizzare e garantire un trattamento etico dei dati
- Principali proprietà: Fairness, Trasparency, Diversity
- Problema allo studio:
  - Fairness nel trattamento di dati "soggettivi"
  - Caso di studio: ranking di recensioni (Yelp, TripAdvisor, Amazon, ..)

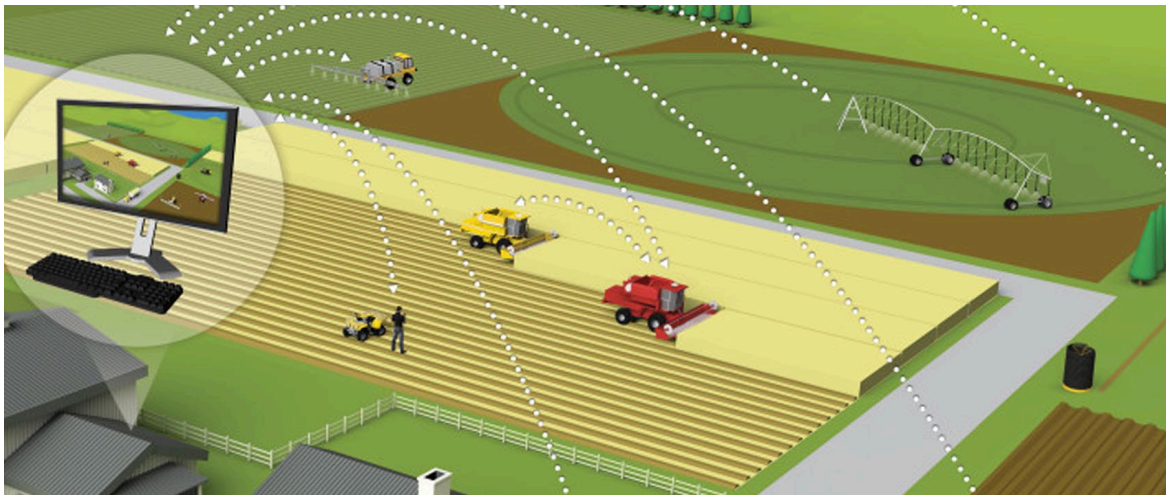
TOPIC 8: EXPLAINABLE AI FOR DATA INTEGRATION

- Problema: spiegare il comportamento dei modelli di apprendimento automatico dedicati all'Entity Recognition (per esempio BERT)
- Strumenti di Entity Recognition, Alaska benchmark
- Obiettivo: trovare comportamenti imprevisti e identificare le caratteristiche più importanti che guidano la predizione
- Dettagli: seminario di Donatella Firmani (disponibile sul sito Web del corso)



TOPIC 9: PRECISION AGRICULTURE

- Architettura per l'ingestion e la gestione di dati provenienti da sensori di vario genere nell'ambito dell'agricoltura di precisione
- Qualche dettaglio sul sito del progetto:
  - http://www.project-pantheon.eu/
- Dati: sensori, stazioni meterologiche, immagini da robot e droni.
- Tecnologie: Spark, MongDB, …
- Metodologie: analisi descrittive e predittive



Topic 10: GEMELLI GENERATOR

- Ontologie, Semantic Web o DB
  - definizione e implementazione di ontologie orientate alla ricerca clinica, basate su standard internazionali e ambienti open.
- Distributed learning e data hub su dati medici
  - sviluppo di algoritmi statistici distribuiti per la collaborazione fra più centri di ricerca senza scambio di dati sensitive
  - Data Mart clinici
- Richiede skills di Python oppure R
- Contesto: seminario del Policlinico Gemelli su sito Web del corso

TOPIC 11: DATA PROVENANCE

- Obiettivo: supporto alla "spiegazione" dei risultati di una elaborazione su big Data (debugger per I big data)
  - Utile in caso di risultati inaspettati e/o performance anomale
  - Esempi: Scorpion (https://github.com/sirrice/scorpion), Titian
- Tecnica: data provenance.
  - Partendo dal risultato di una pipeline di operazioni si genera:
    - L'origine dei dati nel risultato
    - Come e da chi sono stati modificati
- Problema allo studio:
  - Data provenace nel data processing per il Machine Learning

TOPIC 12: AtacMonitor

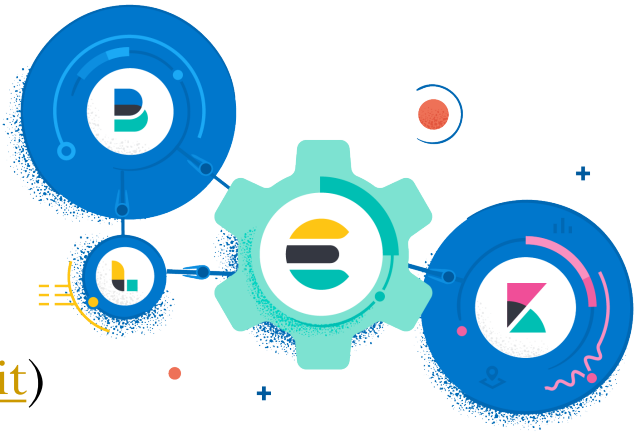
- Atacmonitor (www.atacmonitor.com) prestazioni del trasporto pubblico a Roma. Raccoglie ogni 3 minuti i tempi di attesa in più di 500 stazioni all'interno della città.
- Progetto open source su piattaforme AWS, MongoDB, Atlas cloud
- Attività:
  - data ingestion: estendere la raccolta dei dati
  - stream processing:avvisi in tempo reale basati su eventi (Twitter bot).
  - data mining: definire KPI e rilevare pattern
  - data analysis: combinazione di serie storiche e dati topografici

TOPIC 13: NewsOnWire

- Obiettivo: Archiviare le pagine di un sito di news con timestamp
- Realizzazione di un sistema non supervisionato
  - features per capire l'evoluzione di una pagina nel tempo
- Tecniche:
  - Metodi di crawling incrementale
  - Architettura "big data" per poter scalare su migliaia di siti su AWS
  - Pipeline con strumenti visti nel corso di Big Data (Spark in particolare)
- Dettagli:
  - Video
- Riferimenti:
  - Prof. Valter Crescenzi (crescenz@dia.uniroma3.it)
  - Ing. Valerio Cetorelli (valerio.cetorelli@uniroma3.it)

TOPIC 14: Web pipe

- Obiettivo: definizione di una pipeline di estrazione e analisi di dati da siti Web automatizzata e incrementale
- Utilizzo di tecnologie cloud e sistemi software all'avanguardia
- Tecniche:
  - Apache Lucene (motore di indicizzazione di ELasticSearch)
  - Il linguaggio KQL : per la realizzazione di analisi e dashboard interattive
  - Algoritmi di 'Machine Learning' con elastic.co
- Dettagli:
  - Video
- Riferimenti:
  - Prof. Valter Crescenzi (crescenz@dia.uniroma3.it)
  - Roger Voyat (roger.voyat@uniroma3.it)



Cosa Consegnare

- Relazione esaustiva ma sintetica che descrive: le attività svolte, le analisi fatte, le osservazioni conclusive.
  - Consegna a mano nel giorno delle presentazioni
- Materiale aggiuntivo per la riproducibilità delle attività svolte (es. codice, script, ecc.)
  - Su Moodle

Valutazione

- Il voto è unico per gruppo, commisurato al numero di membri del gruppo e basato sui seguenti fattori:
- Qualità presentazione e materiale consegnato (40%)
  - Esposizione finale
  - Riproducibilità
  - Chiarezza e sintesi della tesina
  - Capacità di evidenziare i risultati chiave
- Originalità e Proattività (30%)
  - Impostazione e significatività della sperimentazione
  - Capacità di problem solving in ambito Big Data
- Implementazione (30%)
  - Soddisfacimento requisiti
  - Completezza della sperimentazione
  - Difficoltà problemi affrontati

Deadlines (orientative)

- Prima presentazione (pitch): 12 giugno 2020
- Prova scritta: 7 luglio 2020
- Consegna progetti: 23 luglio 2020
- Presentazione finale: 24 luglio 2020
- Verbalizzazioni: entro il 31 luglio 2020