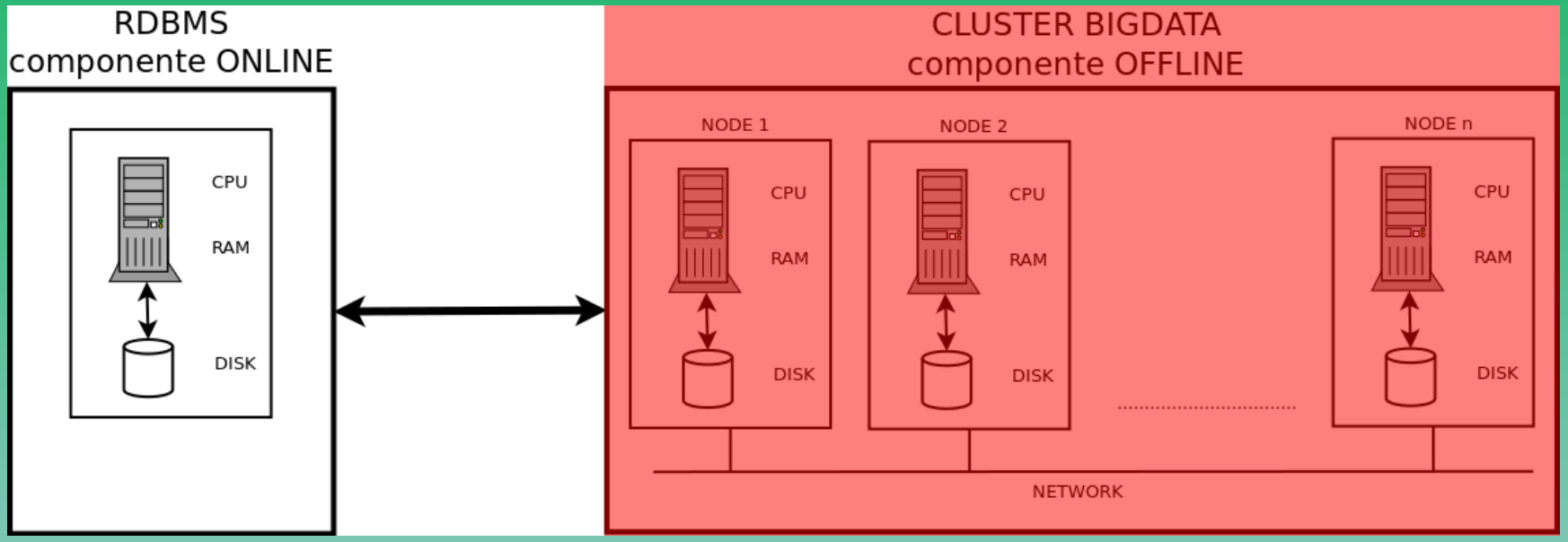
Politecnico di Torino
Facoltà di Ingegneria dell'Informazione
Tesi di Laurea in Ingegneria Informatica

Titolo
"Progettazione e sviluppo di un'architettura Big Data per dati geospaziali"

Relatori
Prof.ssa Elena Maria Baralis
Prof. Paolo Garza

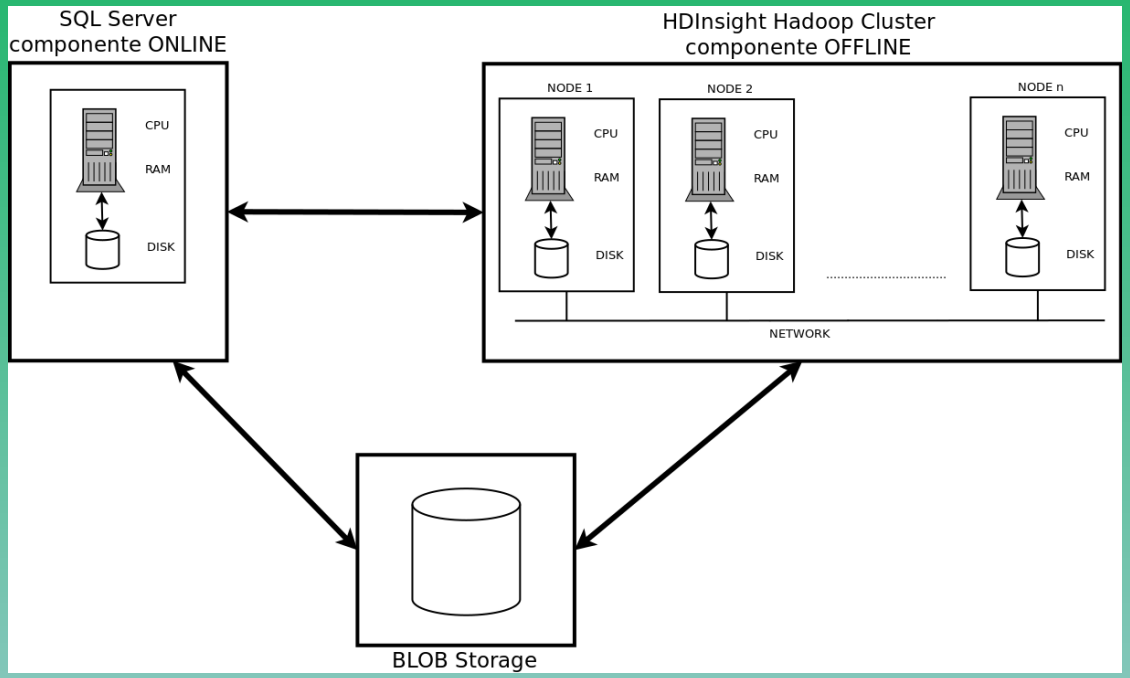
Candidato
Roberto Arcomano

Problema affrontato



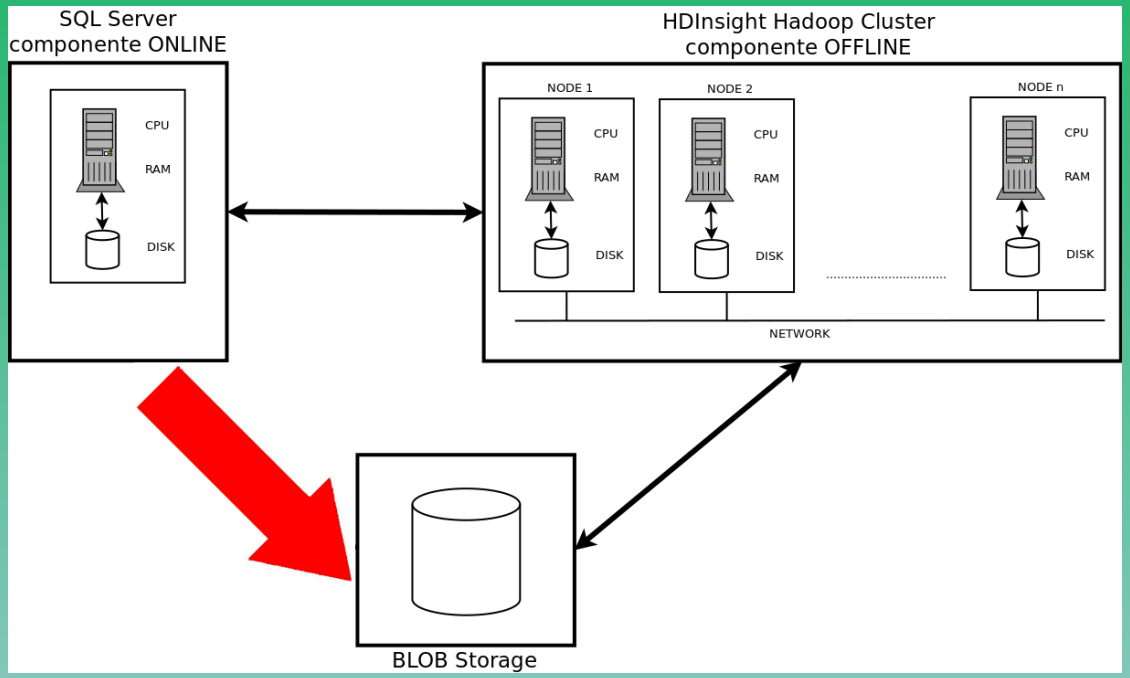
- Creazione infrastruttura Big Data OFFLINE (utilizzo: storico, analisi CPU intensive e raccolta dati statistici
- per la gestione di dati geospaziali
- che permetta la comunicazione con componente ONLINE basata su RDBMS, largamente diffusa e utilizzata ovunque

Architettura dettagliata



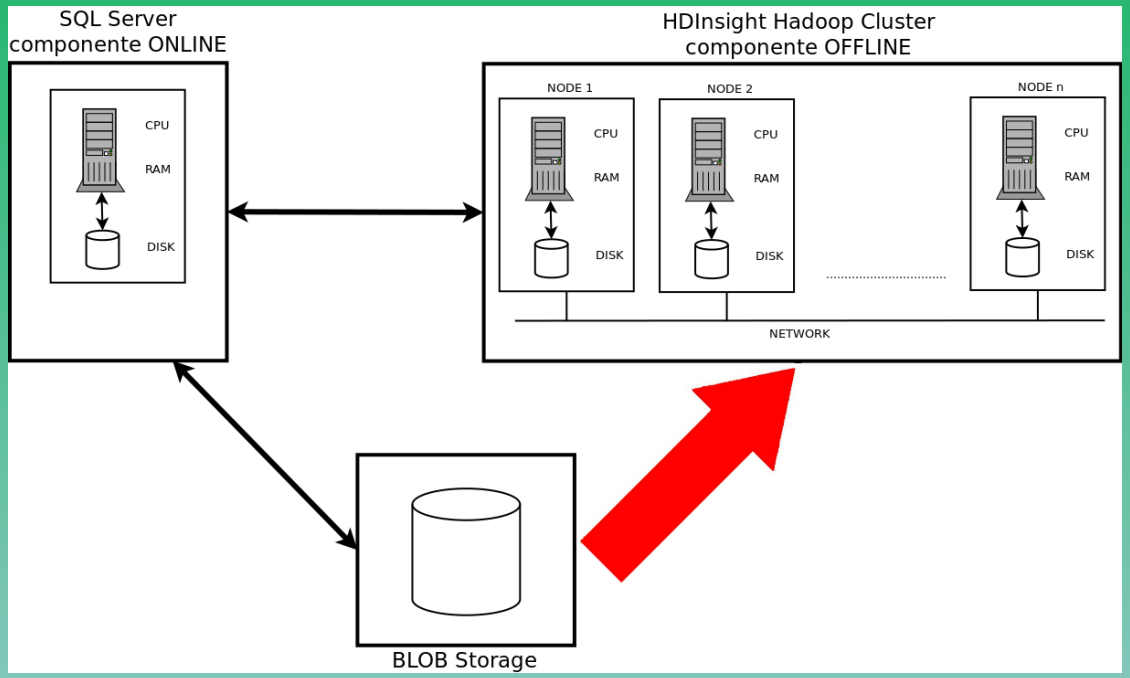
- Cloud Computing: Microsoft Azure
- RDBMS: Microsoft SQL Server
- Big Data: Cluster HDInsight, basato su Hadoop / Spark
- Supporto Storage: BLOB Storage a basso costo

Fase 1: Esportazione RDBMS -> CSV



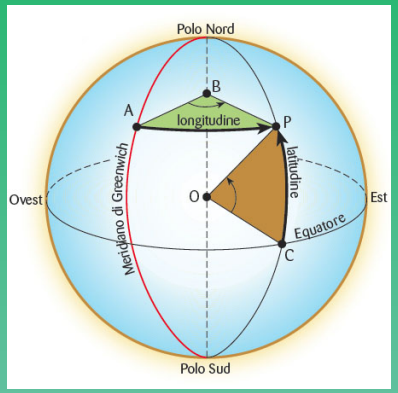
- Export RDBMS→Blob Storage (Data Factory, Apache Sqoop)
- Possibilità di filtrare i dati da esportare (where condiction)
- Schedulazione esportazione (job)
- Formato di output: CSV - Comma-separated values

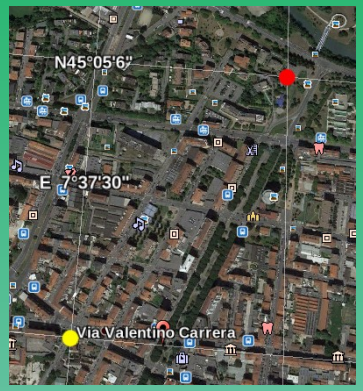
Fase 2: Analisi Big Data

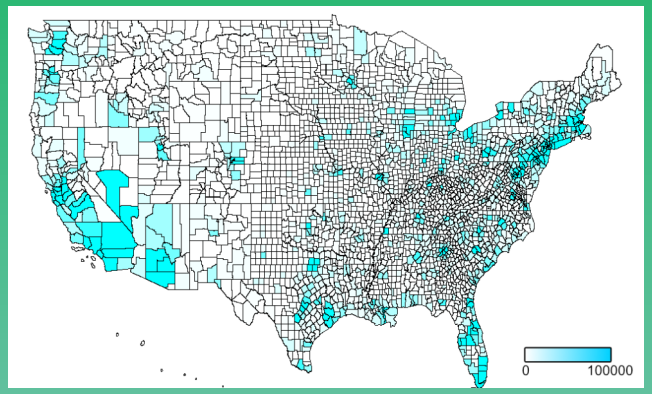


- Importazione formato CSV
- Utilizzo Frameworks a disposizione per l'elaborazione dati:
  - Spark
  - GeoSpatial Frameworks

GeoSpatial Frameworks

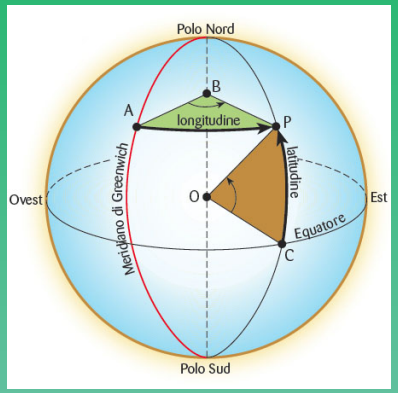


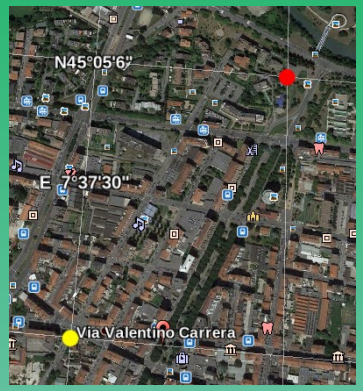


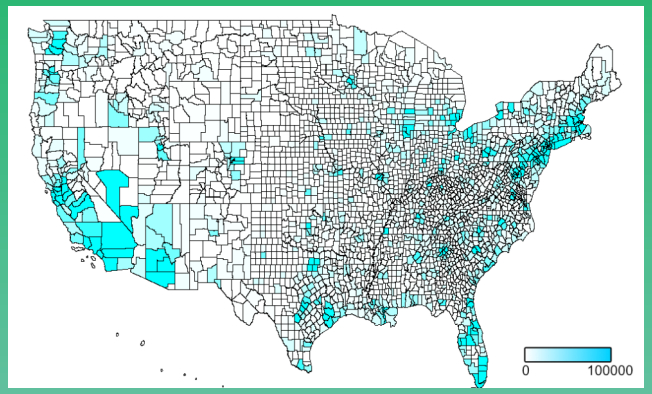


- Sono un'estensione delle librerie Spark
- Gestiscono dati geospaziali (punti, rettangoli, poligoni, …)
- Sfruttano le proprietà Big Data di suddividere l'elaborazione tra più nodi

GeoSpatial Frameworks: Tipi di dati

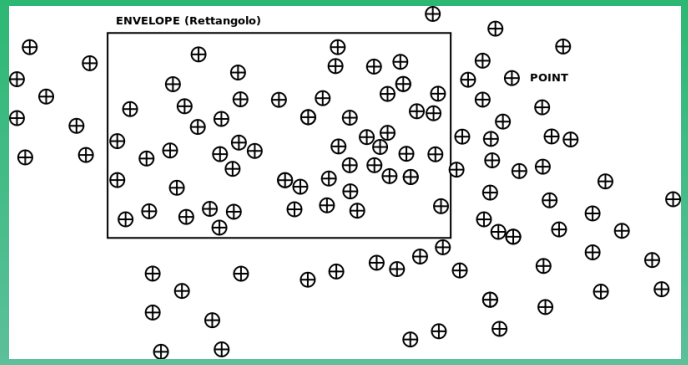


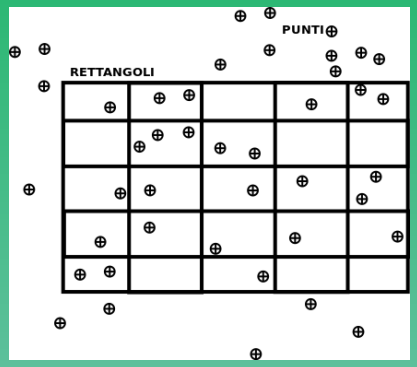




- Vividsolutions: Point, Geometry, Envelope
- Wololo: GeoJSON
- GeoSpark: PointRDD, RectangleRDD, PolygonRDD

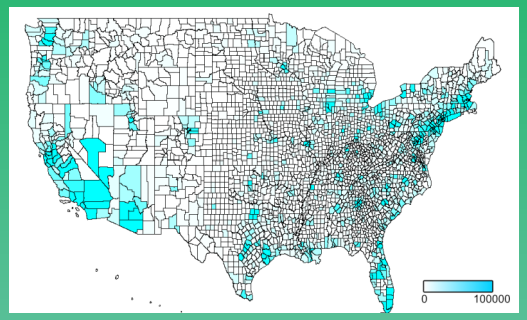
GeoSpatial Frameworks: Operazioni

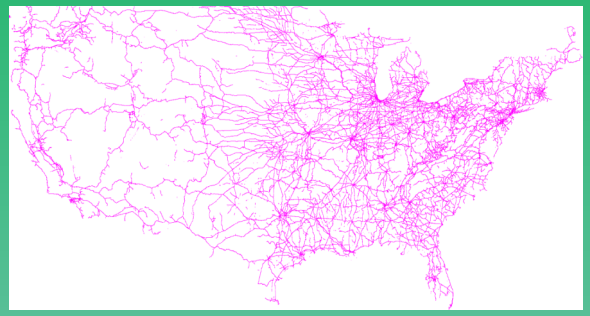


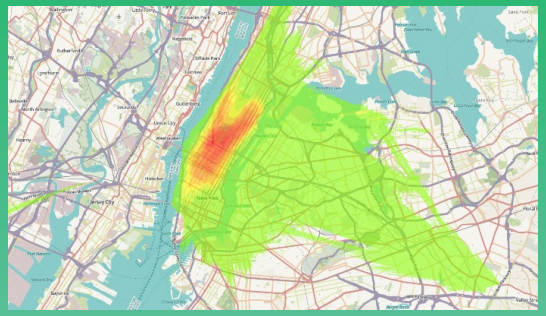


- Operazioni
  - Conteggio punti interni ad un rettangolo / poligono
  - Minimum Bounding Rectangle
  - Operazioni insiemistiche (union, intersection, subtract)

GeoSpatial Frameworks: Babylon

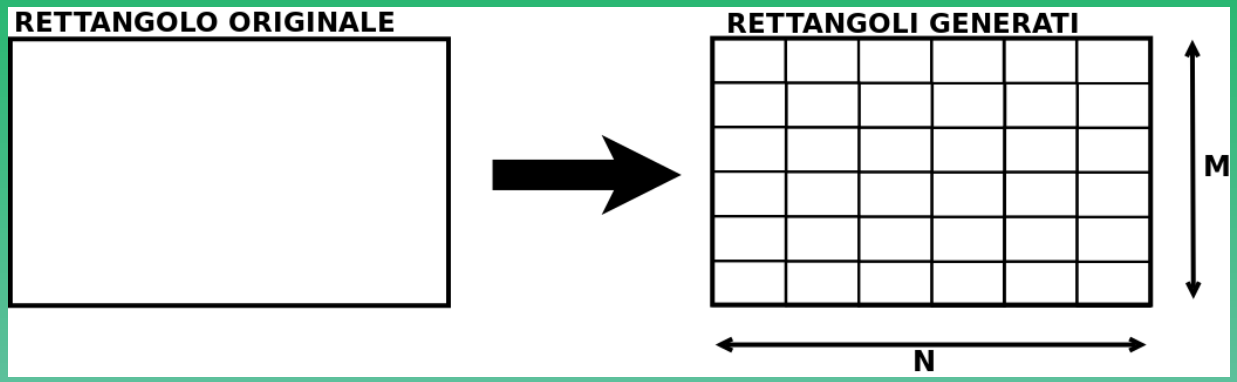






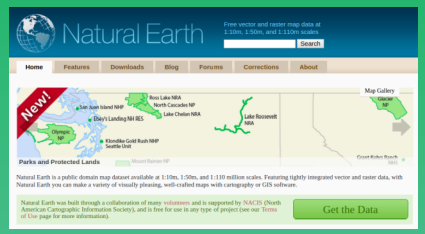
- Libreria Babylon di GeoSpark
- Prende in input dati di tipo PointRDD, PolygonRDD
- Genera immagini in diverse modalità
  - ChoroplethMap (matrice occorrenze)
  - ScatterPlot (adatto a linee)
  - HeatMap (mappa di calore)

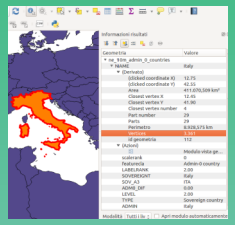
GeoSpatial Frameworks: Esempi

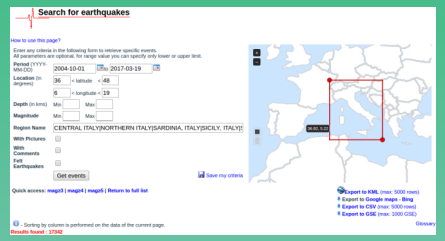


- Filtro file input CSV, es.: match String di un campo attributo
- Suddivisione di un rettangolo in più rettangolini
- FlatMap per aggiungere punti nell'intorno di una coordinata
  - Creazione zone partendo da un punto singolo
  - Creazione linee basate su 2 punti
  - Amplificazione zone o punti in base ad un parametro (es.: magnitudo per gli eventi sismici)

Open Data

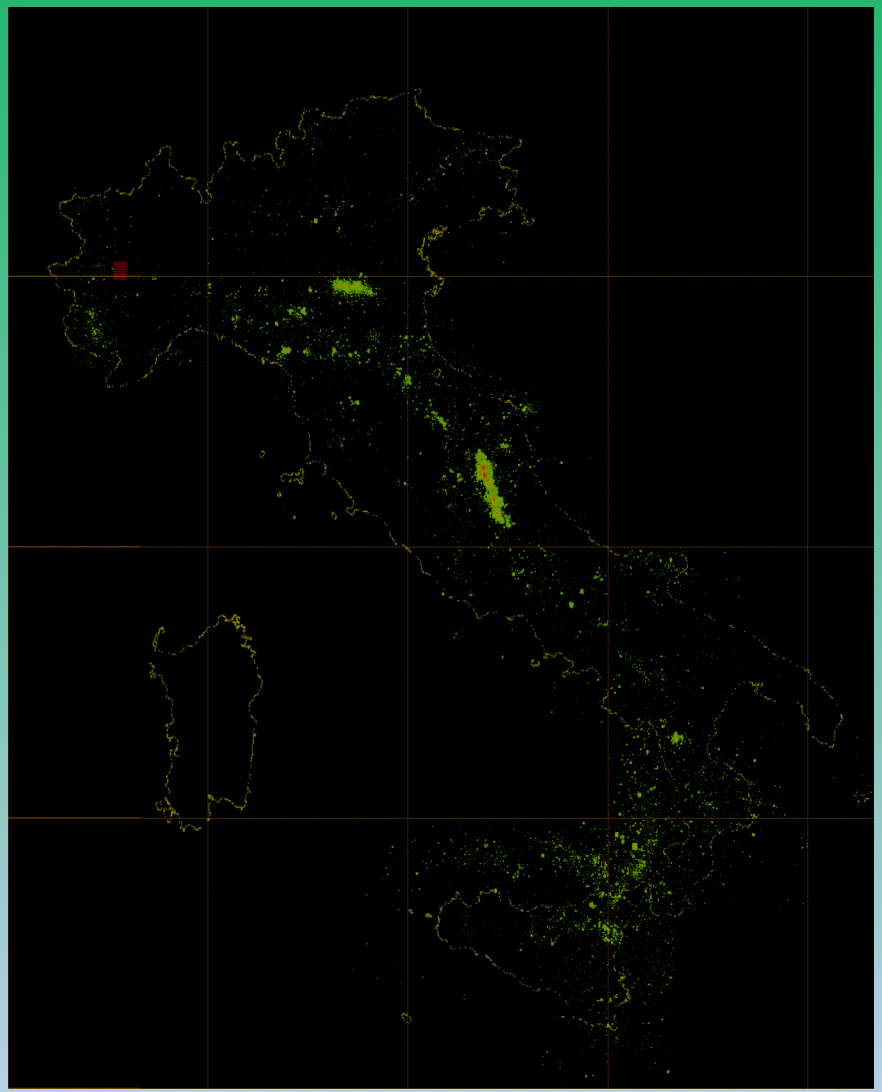






- Tipologie Dati: (strade, fiumi, edifici, confini e Fusi Orari)
- Dati su Eventi Sismici, Tsunami, Eruzioni
- Formato Dati usato: Shapefile Vector
- Esempi:
  - http://www.naturalearthdata.com (dati fisici e politici)
  - http://www.geofabrik.de (dati fisici e politici)
  - http://www.emsc-csem.org (eventi naturali)

Mappa di Esempio



- Intorno quadrato di Torino
- Confine Italiano
- Direttrici Longitudine e Latitudine
- Eventi sismici dal 2004 al 2017 (HeatMap)

Note sul Cloud Computing

- E' una infrastruttura comoda e flessibile
- Di tipo Pay for Use, on demand
- Conveniente per
  - Studi di fattibilità infrastrutturale
  - Tests
  - Scalabilità dinamica
- Servizi:
  - Macchine Virtuali
  - Memoria di massa
  - Clusters (Big Data)

Open Points

- Analisi prestazioni Standalone vs Cluster
- Indici Geospaziali
- Controllo formato shapefile vector su GeoSpark invece di usare software vector (es.: QGIS)
- Confronto tra diversi vendor Cloud Computing

Ringraziamenti