Gestione ed analisi di elevati volumi di dati con gli strumenti MongoDB, Spark e R

Contesto e obiettivi

Studiare elevati volumi di dati allo scopo di fare data exploration con R

- Superare i limiti di R
- Studio della libreria sparklyR
- Metodo di utilizzo congiunto tra MongoDB, Spark e R

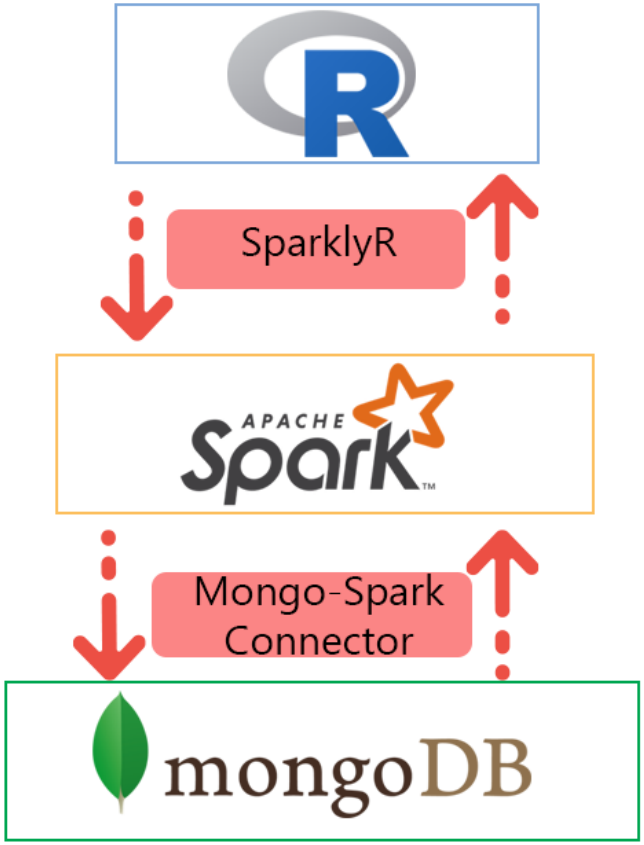
Soluzione testata

Uso congiunto di più strumenti per superare le limitazioni di R

- Spark: Framework dedicato alla computazione di dati su cluster
- MongoDB: Database NoSQL di tipo documentale

Configurazione utilizzata

- **sparklyR**: Nuova libreria sviluppata da Rstudio. Permette manipolazione dei dati appoggiandosi su Spark.
- **Mongo-Spark connector**: Scelta standard per la connessione tra Spark e MongoDB



Caso di studio: transazioni bitcoin

Si è preso in analisi un dataset contenente informazioni sulle transazioni tra il 2009 e il 2016

txID	unixtime	value
171	1231731025	5.0e+09
183	1231740133	4.0e+09
185	1231740736	3.0e+09
187	1231742062	2.9e+09
192	1231744600	1.0e+08
227	1231770060	1.0e+08
255	1231790660	2.8e+09

txin.txt	**txtime.txt**
txID	txID
addrID	unixtime
value	
65 M di righe	**30 M di righe**

txID	unixtime
2576	1233446920
2577	1233448111
2578	1233449296
2579	1233450262
2580	1233451816
2581	1233452942
2582	1233453846

Scenari di studio

Per ciascuno degli scenari di studio

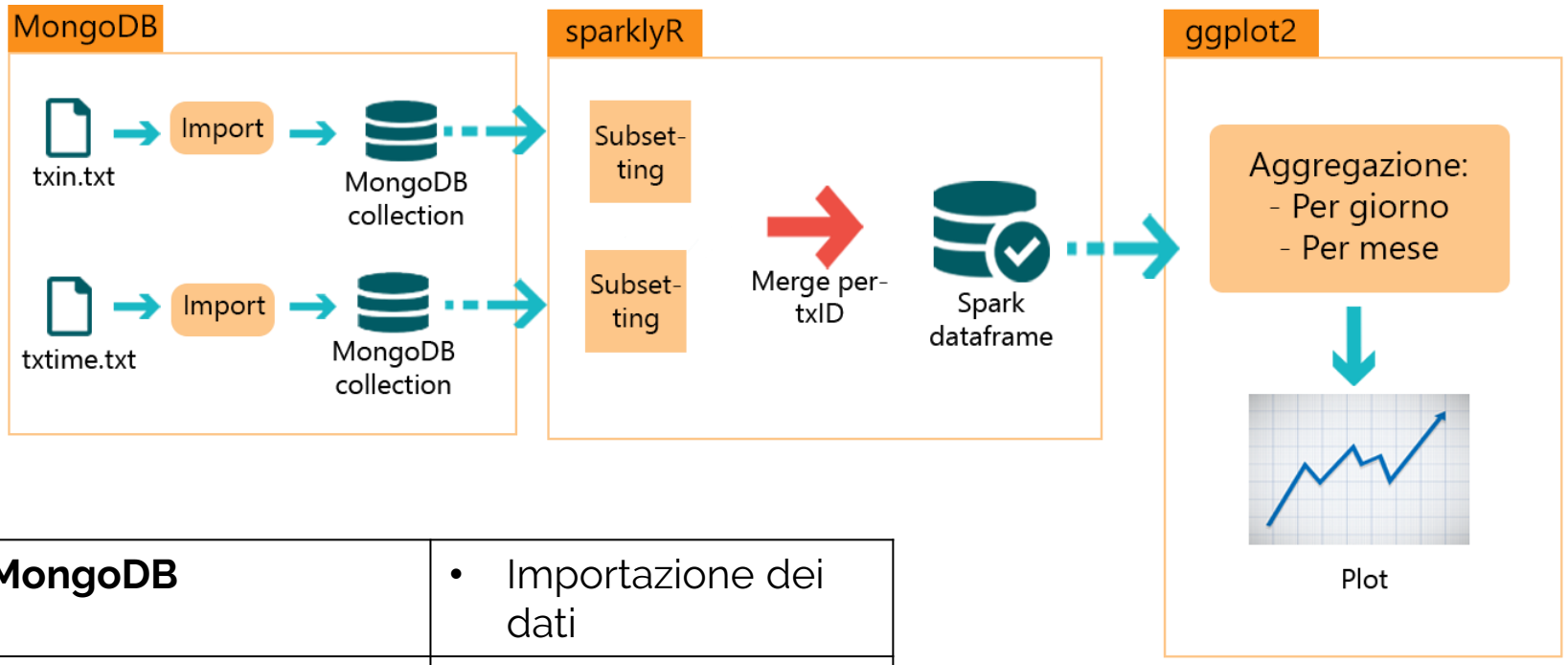
- Diversa distribuzione delle attività sulla configurazione di riferimento
- Valutazione delle prestazioni, vantaggi e svantaggi

Attività svolte sul dataset

Sono state necessarie le seguenti operazioni per lo studio in esame

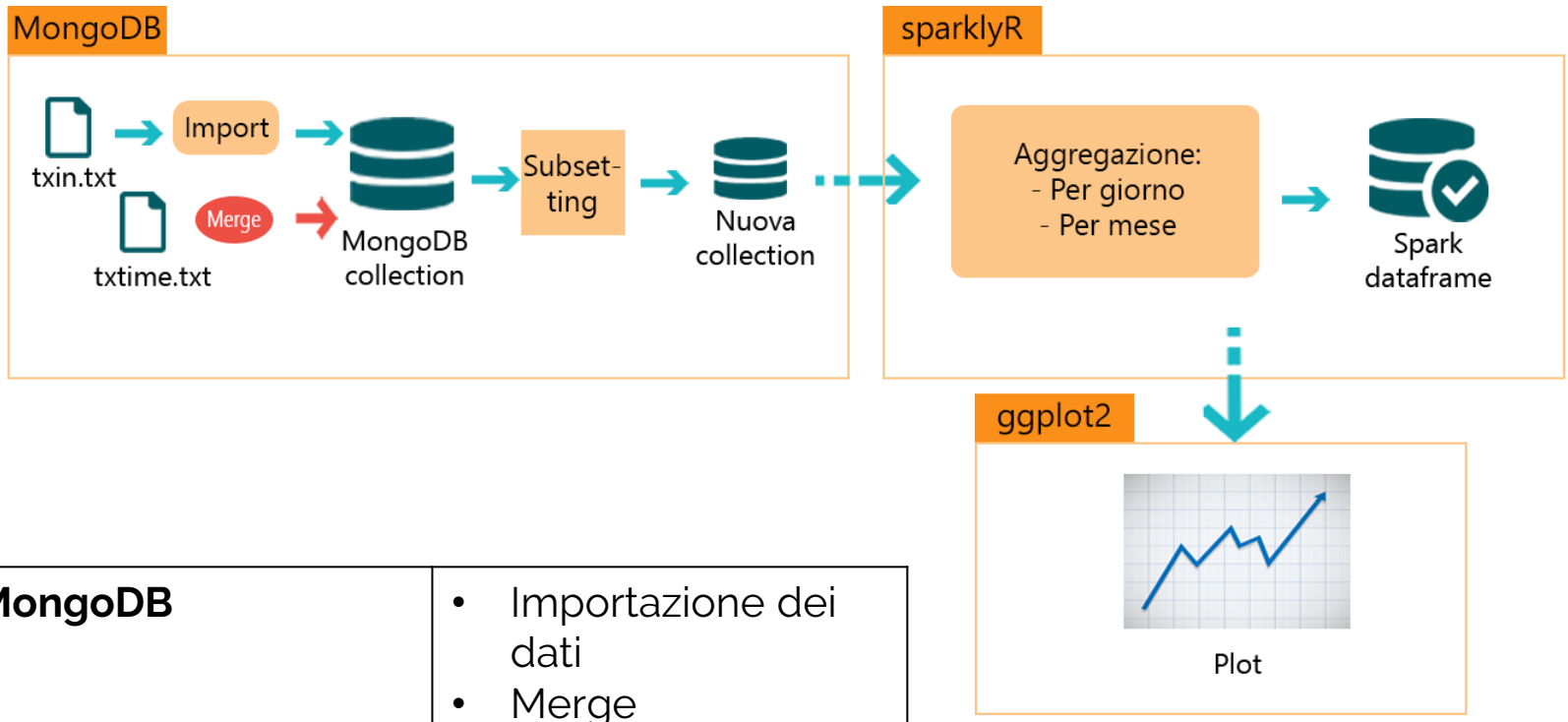
- Importazione dei file
- Subsetting + merge
- Aggregazione per giorno/mese
- Plotting, rappresentazione grafica

I Scenario



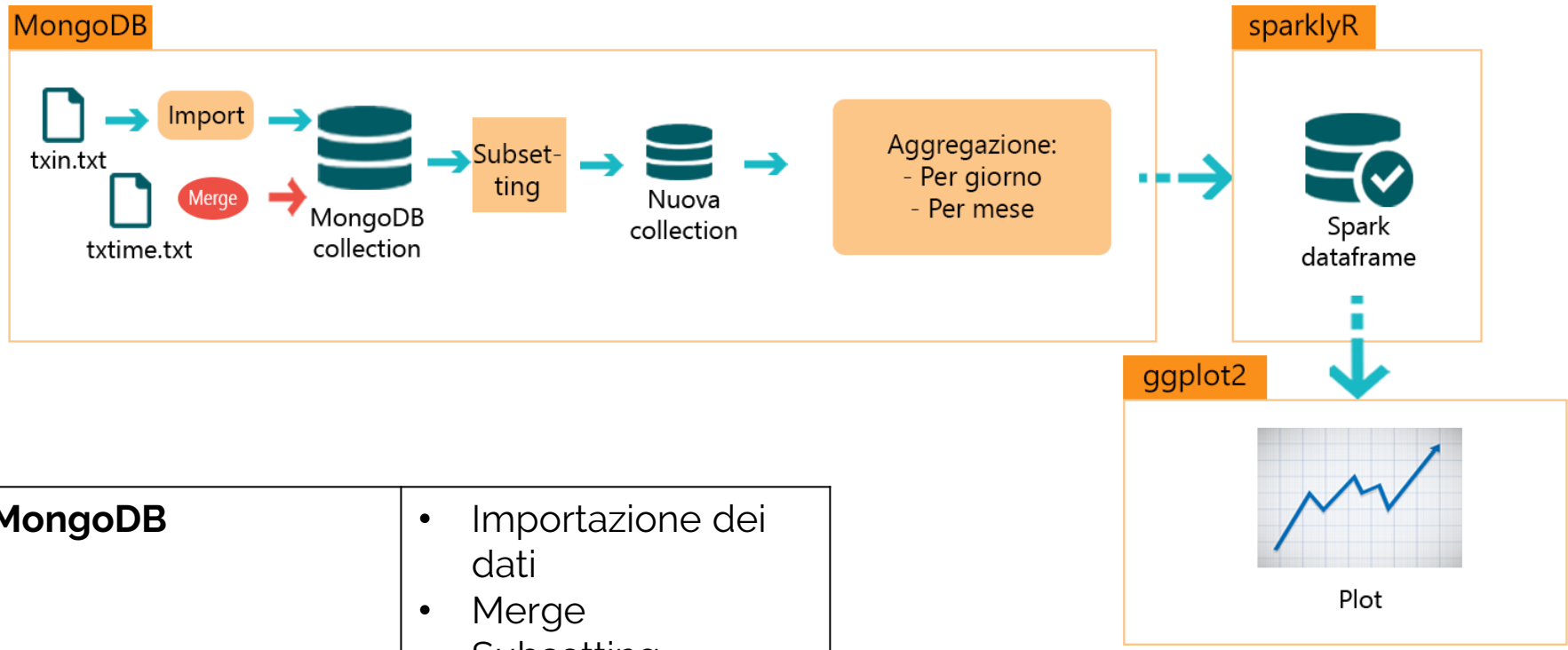
MongoDB	• Importazione dei dati
Spark	• Subsetting • Merge
R	• Aggregazione • Plotling

II Scenario



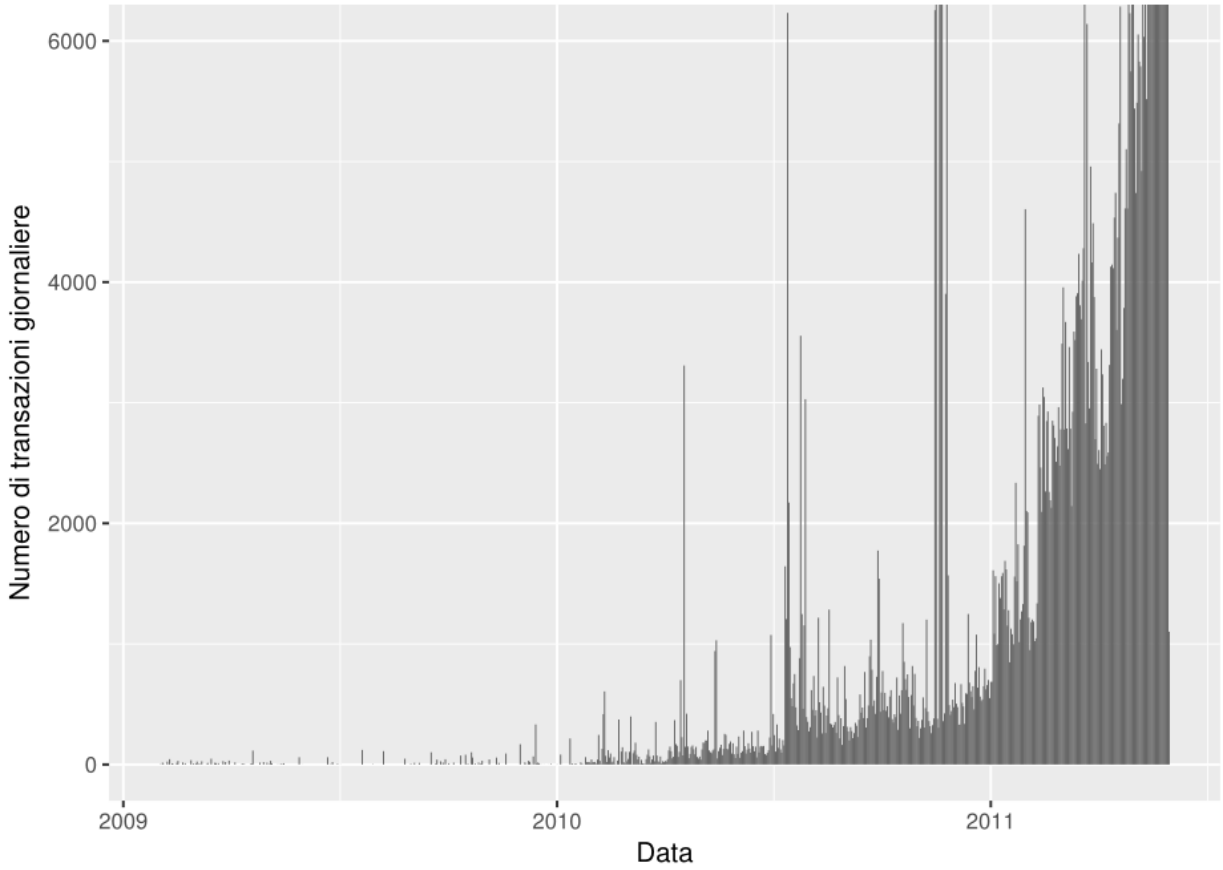
MongoDB	• Importazione dei dati • Merge • Subsetting
Spark	• Aggregazione
R	• Plotling

III Scenario



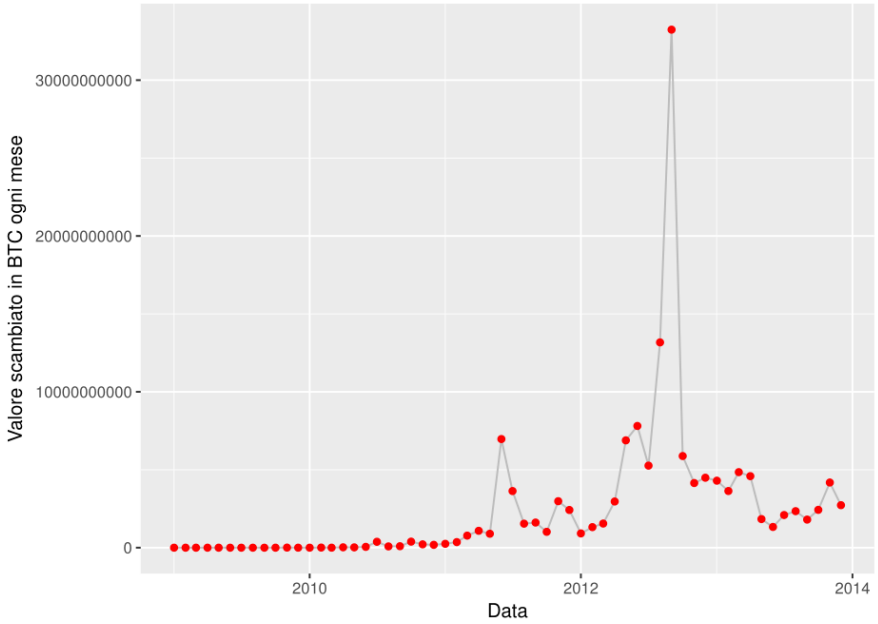
MongoDB	• Importazione dei dati • Merge • Subsetting • Aggregazione
Spark	/
R	• Plotting

Alcuni grafici ottenuti (1)

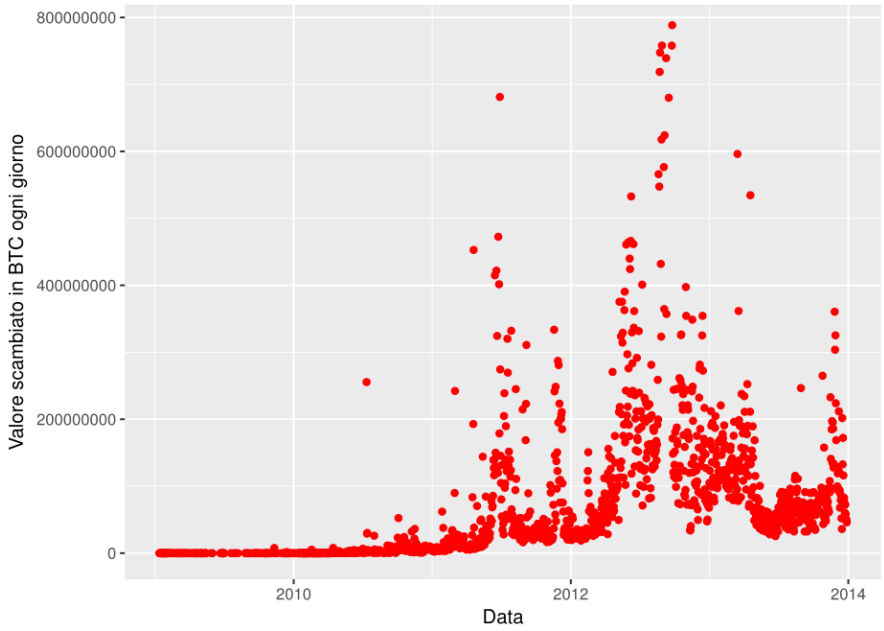


Numero di transazioni giornaliere dal Gennaio 2009 al Maggio 2011

Alcuni grafici ottenuti (2)



Valori raggruppati per mese



Valori raggruppati per giorno

Interessante osservazione del dettaglio giornaliero per comprendere il picco intravisto nel primo grafico.

Ambiente operativo

Tutti gli esempi sono stati svolti sulla stessa macchina caratterizzata dalle seguenti specifiche hardware

Processore	Intel core i5-2410M @ 2.30GHz
Memoria RAM	6,00 GB
Tipo di sistema	Sistema operativo a 64 bit

Risultati quantitativi

Tempistiche di ogni scenario analizzato

	MongoDB	**Spark**	**R**	**Totale**
I Scenario	• Importazione: ~35 min.	(12 min.) x 51*	5 min.	~612 min.
II Scenario	• Importazione: ~35 min. • Index: 8 min. • Merge: ~100 min.	(14 min.) x 5*	3 min.	~216 min.
III Scenario	• Importazione: ~35 min. • Index: 8 min. • Merge: ~100 min. • Aggregazione: 2 min.	~0	10 sec.	~145 min.

* Numero di subset per lo studio dei dati completi

Commenti conclusivi

Principali problemi riscontrati e soluzioni adottate

- Difficoltà nel merge dei dati fra le due tabelle considerate.
- Difficoltà nel passaggio della costante di tempo unixtime alla data in formato mese/giorno/anno.