

CORSO DI BIG DATA
Primo Progetto
27 marzo 2015

Si supponga di avere a disposizione un file di testo generato da un sistema di billing di una catena di supermercati che contiene, per ciascuno scontrino, una riga con la data (in un formato a piacere) e la lista dei prodotti acquistati, separati da una virgola. Per esempio:

```
20150321,uova,latte,pane,vino
20150218,pesce,pane,insalata,formaggio
.....
```

(Il file può essere costruito autonomamente. E' disponibile un progetto Java per la generazione automatica di un file con tale formato: <http://torlone.dia.uniroma3.it/bigdata/DataGenerator.zip>).

Progettare e realizzare in MapReduce:

1. Un job in grado di generare, possibilmente in ordine decrescente, i prodotti acquistati seguiti dal numero complessivo di pezzi venduti. Per esempio:

```
pane 852
latte 753
carne 544
...
```
2. Un job in grado di generare, per ciascun prodotto, l'andamento delle vendite del primo trimestre 2015. Per esempio:

```
pane 1/2015:623 2/2015:634 3/2015:788
latte 1/2015:523 2/2015:488 3/2015:512
...
```
3. Un job in grado di generare le 10 coppie di prodotti che vengono più frequentemente venduti insieme indicando il numero di occorrenze, possibilmente in ordine decrescente. Per esempio:

```
pane,latte,322
latte,uova,223
...
```

Facoltativo: calcolare anche (a) per ciascuna coppia di prodotti (p_1, p_2), in quante transazioni nelle quali compare p_1 , compare anche p_2 (supporto della regola di associazione $p_1 \rightarrow p_2$); (b) la frequenza di insiemi di prodotti di cardinalità minore di 5, (c) tutti i job suddetti implementati con Pig e Hive.

Per ciascun job bisogna illustrare in un documento:

- Una possibile implementazione MapReduce (commentata) in pseudocodice (qualunque)
- Il relativo codice in Java (da allegare al documento)
- Un test di uso con file di input di piccole dimensioni e quello di output (da allegare)
- Log di esecuzione su computer locale
- Log di esecuzione su AWS con max sei istanze di tipo t2 o m3 di dimensione inferiore alla xlarge
- Tabella e grafico dei tempi di esecuzione in locale e su AWS possibilmente con dimensione dell'input crescente

Consegnare tutto **entro il 27 aprile 2015** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: <http://moodle2.ing.uniroma3.it/moodle/>.