Spark è un motore generale e veloce per il processo di dati su larga scala che disaccoppia completamente il concetto di storage da quello di calcolo distribuito, consentendo di avere software esterni per il file system distribuito come Hadoop.

0.1 Resilient Distributed Dataced RDD

Ho la stessa garanzia di **Hadoop** per quanto riguarda la computazione, partendo da file distribuiti su un cluster sulla home/RDD.

0.1.1 Trasformazioni

Una funzione che prende un dataset e lo modifica. Esempi di questa categoria sono **map**, **filter**, **flatMap** e **groupByKey**.

0.1.2 Actions

Una funzione che compie un'azione sul dataset. Esempi di questa categoria sono **reduce**, **count**, **collect** e **take**.

0.1.3 Variabile broadcast

Variabile a sola lettura condivisa tra tutti i nodi.

0.1.4 Accumulatori

Un accumulatore può essere usato solo per operazioni strettamente associative, come conteggi.

0.2 Metodo di tassazione

Fino ad ora lo abbiamo visto come risoluzione per i problemi di convergenza dell'algoritmo di pagerank, ma è utilizzabile anche in altri casi come la personalizzazione del risultato del vettore prodotto, e viene utilizzato in algoritmi come **topic-sensitive pagerank**, dove per esempio se volessi personalizzare i risultati di un utente appassionato di sport farei teletrasportare il crawler a pagine sicure che so che parlano di sport.

0.2.1 Pagerank vs Trustrank

Se chiamiamo pagerank p e trustrank t, proviamo a calcolare $\frac{p-t}{p}$. Se questa quantità è negativa o vicina a zero, sono contento, se invece si avvicina all'uno probabilmente su quel nodo sono attive delle tecniche che falsano il pagerank e riducono il Trustrank.

0.2.2 HITS, chiamato anche Hub&Authorities

Definisce le pagine "buone" come fa pagerank, ma utilizza una ricorsione indiretta. Divide tutte le pagine web in due categorie, **hub** e **authorities**. Una pagina è **hub** quando linka delle buone authorities, mentre chiamata **authority** quando è linkata da un buon hub.