



Dipartimento di Ingegneria Informatica, Modellistica,  
Elettronica e Sistemistica

---

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Sviluppo di una libreria parallela per Social Data  
Mining basata su PyCOMPSs

Relatori:

**Prof. Paolo Trunfio**  
**Ing. Fabrizio Marozzo**  
**Ing. Loris Belcastro**

Candidato:

**Domenico Costantino**  
Matricola 189168

Anno Accademico 2018/2019

# Indice

<b>Introduzione</b>	<b>1</b>
<b>1 Big Social Data Analytics</b>	<b>4</b>
1.1 Big data: proprietà e applicazioni . . . . .	4
1.1.1 Fonti e struttura dei dati . . . . .	8
1.1.2 Big data analytics . . . . .	9
1.2 Modelli di programmazione per big data . . . . .	12
1.2.1 Introduzione al Data Mining . . . . .	13
1.2.2 Requisiti e classificazione . . . . .	15
1.2.3 Modelli più usati . . . . .	17
1.3 Social Media Mining . . . . .	20
1.3.1 Social Network Sites e Analysis . . . . .	22
1.3.2 Social Media Big Data Analytics . . . . .	25
<b>2 La libreria ParSoDA</b>	<b>28</b>
2.1 Stato dell'arte e framework usati . . . . .	29
2.1.1 Apache Hadoop . . . . .	30
2.1.2 Apache Spark . . . . .	33
2.2 Struttura della libreria . . . . .	36
2.2.1 Architettura ed esecuzione . . . . .	38
2.3 Sviluppo applicazioni in ParSoDA . . . . .	39
<b>3 Il sistema COMPSSs</b>	<b>43</b>
3.1 Installazione, configurazione ed esecuzione di applicazioni COMPSSs	46
3.2 Il framework PyCOMPSSs . . . . .	50
3.2.1 Modello di programmazione . . . . .	51
3.2.2 Sintassi . . . . .	54
3.2.3 Modello di esecuzione . . . . .	59
3.2.4 Storage API e supporto a Jupyter Notebook . . . . .	60

3.3	Il package DDS di PyCOMPSSs . . . . .	61
<b>4</b>	<b>Implementazione di ParSoDA-PyCOMPSSs</b>	<b>65</b>
4.1	Python e le dipendenze richieste . . . . .	66
4.2	Schema logico di un'applicazione . . . . .	68
4.2.1	Configurazione delle applicazioni . . . . .	70
4.3	Struttura dei componenti . . . . .	72
4.3.1	Funzioni disponibili . . . . .	76
4.3.2	Dettaglio delle fasi e funzioni DDS usate . . . . .	80
<b>5</b>	<b>Caso di studio: applicazione di Trajectory Mining con ParSoDA-PyCOMPSSs</b>	<b>84</b>
5.1	Panoramica sul Trajectory Mining . . . . .	84
5.2	Scopo dell'applicazione e dati usati . . . . .	87
5.3	Schema di esecuzione . . . . .	89
5.3.1	Algoritmi di analisi . . . . .	92
5.4	Risultati . . . . .	99
5.4.1	Analisi delle traiettorie . . . . .	99
5.4.2	Prestazioni della libreria . . . . .	103
	<b>Conclusioni</b>	<b>106</b>
	<b>Bibliografia</b>	<b>108</b>

# Elenco delle figure

1.1	Dati generati sul Web in un minuto . . . . .	6
1.2	Modello delle 5 V per i big data . . . . .	7
1.3	Processo di estrazione di conoscenza dai big data . . . . .	10
1.4	Flow di esecuzione del modello MapReduce . . . . .	18
1.5	Flow di esecuzione del modello DAG . . . . .	19
1.6	Percentuale di utenti attivi sui social media in relazione alla popolazione totale . . . . .	22
1.7	Cronologia del lancio dei maggiori SNS dal 1997 al 2006 . . . . .	23
1.8	SNS e servizi di messaggistica più attivi in Italia nel 2018 . . . . .	24
2.1	Architettura dell'ecosistema Hadoop . . . . .	31
2.2	Architettura dei componenti Spark . . . . .	34
2.3	Architettura di ParSoDA . . . . .	39
2.4	Esecuzione di applicazioni ParSoDA su Hadoop e Spark . . . . .	40
3.1	Architettura del sistema di esecuzione di COMPSSs . . . . .	44
3.2	Grafo delle dipendenze di un'applicazione COMPSSs . . . . .	50
3.3	Codice di un'applicazione PyCOMPSSs d'esempio . . . . .	52
3.4	Grafo delle dipendenze dell'applicazione d'esempio . . . . .	53
3.5	Gestione degli oggetti in PyCOMPSSs . . . . .	60
3.6	Livello di Storage API in PyCOMPSSs . . . . .	60
4.1	Fasi di un'applicazione ParSoDA-PyCOMPSSs . . . . .	69
4.2	Codice di un'applicazione ParSoDA-PyCOMPSSs d'esempio . . . . .	71
4.3	Struttura dei componenti della libreria . . . . .	73
4.4	Diagramma dei package della libreria . . . . .	74
4.5	Diagramma delle classi usate nel passo di reduction . . . . .	75
4.6	Metodo <i>reduce</i> della classe <i>ReduceByTrajectories</i> . . . . .	82
4.7	Metodo <i>execute</i> della classe <i>SocialDataApp</i> . . . . .	83

5.1	Tassonomia Trajectory Data Mining . . . . .	86
5.2	Mappa di Roma con RoI . . . . .	90
5.3	Grafo delle dipendenze dell'applicazione con Gap-BIDE . . . . .	93
5.4	Grafo delle dipendenze dell'applicazione con PFP-Growth . . . . .	98
5.5	PoI più frequentati di Roma . . . . .	100
5.6	Traiettorie di lunghezza minima 2 più frequenti . . . . .	101
5.7	Traiettorie di lunghezza minima 3 più frequenti . . . . .	102
5.8	Ritardo e Speed-up dell'applicazione . . . . .	104
5.9	Tempi dei singoli passi della fase 2 . . . . .	104
5.10	Memoria occupata . . . . .	105

# Introduzione

Nella storia dell'umanità sono presenti dei momenti che hanno determinato un cambiamento radicale nel modo di vivere degli individui. Questi momenti vengono spesso indicati come "rivoluzioni" (agricola, industriale, ecc.). L'introduzione di Internet e la sua rapida diffusione, appunto, possono essere viste come la *rivoluzione digitale* dell'umanità. Negli ultimi decenni, infatti, è evidente che il progresso tecnologico abbia avuto un tasso di crescita esponenziale. Questo ha generato profondi cambiamenti nella società, che hanno introdotto la necessità di evolvere processi produttivi, servizi e studi per riuscire a stare al passo con i tempi e stimolare il miglioramento della condizione umana.

Oltre alla nascita di Internet, altri fattori, quali la miniaturizzazione dei dispositivi, l'abbattimento dei costi dell'hardware, lo sviluppo di nuovi linguaggi di programmazione e la nascita di nuovi campi di ricerca, hanno contribuito all'evoluzione digitale della società. La diffusione su larga scala dell'utilizzo del Web (e in generale dei dispositivi elettronici) ha portato in dote quelli che oggi chiamiamo *big data*, ovvero collezioni di dati di grandi dimensione, provenienti da diverse fonti, che possono essere generati continuamente e con un'elevata alta velocità. Per citare qualche dato, basti pensare che nel 2019, a fronte di una popolazione mondiale di 7.7 miliardi di persone, gli utenti di Internet sono 4.4 miliardi, sui social media sono presenti 3.5 miliardi di profili attivi e, mediamente, una persona spende 142 minuti della sua giornata su queste piattaforme [56]. Vari studi hanno cercato di stimare il valore economico dei dati generati da un utente social e, anche se non è facile individuare un valore preciso, si può stimare un guadagno annuale, per le diverse piattaforme, di circa 200 dollari nel 2018 che crescerà fino a circa 600 dollari nel 2022 [15]. Per cercare di dare una misura alla quantità di dati che circola sul Web, si può considerare che Google, il motore di ricerca più usato, elabora 100 miliardi di

ricerche ogni mese (circa 40 mila al secondo). Nel 2018, inoltre, il traffico dati che è transitato nei data center cloud dei principali competitor IT è stato nell'ordine dei 10 zettabyte, con la previsione che esso raddoppierà in un paio di anni [47].

Il processo di analisi e di estrazione di conoscenza da questa immensa mole di dati, di supporto nei processi decisionali, viene indicato con il termine *Big Data Analytics*. La Big Data Analytics ha generato, solo nell'ultimo anno, un giro di affari nell'ordine dei 190 miliardi di dollari nel mondo [48], ma risulta in continua crescita insieme alla quantità e al tipo di dati presenti. Le applicazioni di Big Data Analytics non hanno solo lo scopo di estrarre informazioni dai dati al fine di produrre un profitto diretto, ma possono anche essere usati in contesti quali l'analisi dei comportamenti umani, la lotta al terrorismo, le ricerche in campo medico e scientifico. Diversi esempi di applicazioni di questo tipo sono presenti in [33] e [51].

Per riuscire a manipolare i big data sono necessari strumenti specifici che permettano di gestire la complessità temporale e spaziale delle applicazioni. Da questo punto di vista, la comunità scientifica è stata molto prolifica, sviluppando modelli, framework e linguaggi utili allo scopo. Questi strumenti, però, sono spesso caratterizzati da una intrinseca difficoltà nell'utilizzo, in quanto richiedono un background di competenze informatiche non sempre posseduti da coloro che hanno necessità di analizzare questi grandi volumi di dati (es., economisti, analisti e scienziati, esperti di settori nell'ampio spettro dei contesti applicativi). Da qui nasce la necessità di strumenti con un elevato livello di astrazione, che semplifichino lo sviluppo di applicazioni per l'analisi dei dati, senza sacrificare l'efficienza computazionale. La libreria ParSoDA, appunto, sviluppata dallo *Scalable computing and Cloud Laboratory* del DICES, rientra in questa categoria e permette lo sviluppo di applicazioni di data mining parallele e distribuite, mantenendo una logica di sviluppo chiara ed accessibile, anche a chi non ha competenze di programmazione avanzate.

Lo scopo di questo lavoro di tesi è la riscrittura della libreria ParSoDA, che nella sua versione originale è scritta in Java ed utilizza i framework Apache Hadoop e Spark, per adattarla all'ambiente di distribuzione COMPSSs, sviluppato dal *Barcellona Supercomputing Center*, tramite le funzionalità esposte dal framework PyCOMPSSs. La libreria è stata interamente riscritta nel linguaggio Python, che ha permesso di aumentare ulteriormente la leggibilità e la semplicità del codice, favorendo l'ampliamento della libreria. L'integrazione di ParSoDA

con PyCOMPSS permette l'esecuzione di applicazioni ParSoDA su tutte le infrastrutture basate su PyCOMPSS, garantendo scalabilità, efficienza e trasparenza nella gestione delle risorse.

Questo elaborato è organizzato in 5 capitoli. Nel **primo capitolo** verrà effettuata una panoramica sui big data, analizzandone le caratteristiche ed elencando gli approcci e le tecnologie principalmente utilizzate per la loro analisi. In particolare, si discuterà dei dati provenienti dai social media e dei possibili contesti applicativi. Nel **secondo capitolo** verrà presentata la libreria ParSoDA nella sua versione originale, con un approfondimento dei framework di parallelizzazione usati. Il **terzo capitolo** è descritto l'ambiente COMPSS, le sue caratteristiche, il modello di programmazione ed esecuzione e le principali funzioni utilizzate nell'ambito di questo lavoro di tesi. Il **quarto capitolo** descriverà l'implementazione della libreria ParSoDA per l'ambiente di distribuzione COMPSS, approfondendo il linguaggio usato, lo schema concettuale delle applicazioni sviluppate e l'organizzazione dei componenti. Nel **quinto capitolo**, infine, verrà descritta un'applicazione di *Trajectory Mining* sviluppata usando la libreria sviluppata, al fine di testare l'utilizzabilità e la scalabilità.

# Bibliografia

- [1] Sihem Amer-Yahia, Noha Ibrahim, Christiane Kamdem Kengne, Federico Ulliana, and Marie-Christine Rousset. Socle: Towards a framework for data preparation in social applications. *Ingénierie des Systèmes d'Information*, 19:49–72, 2014.
- [2] Rosa M. Badia, Javier Conejero, Carlos Diaz, Jorge Ejarque, Daniele Lezzi, Francesc Lordan, Cristian Ramon-Cortes, and Raul Sirvent. Comp super-scalar, an interoperable programming framework. *SoftwareX*, 3-4:32 – 36, 2015.
- [3] Loris Belcastro, Fabrizio Marozzo, and Domenico Talia. Programming models and systems for big data analysis. *International Journal of Parallel, Emergent and Distributed Systems*, 0(0):1–21, 2018.
- [4] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Appraising spark on large-scale social media analysis. In *Euro-Par 2017: Parallel Processing Workshops*, pages 483–495, Cham, 2018. Springer International Publishing.
- [5] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. G-roi: Automatic region-of-interest detection driven by geotagged social media data. *ACM Transactions on Knowledge Discovery from Data*, 12(3):27:1–27:22, January 2018.
- [6] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Parsoda: High-level parallel programming for social data mining. *Social Network Analysis and Mining*, 9(1), 2019.

- [7] Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.
- [8] Marco Buttu. *Phyton: Guida completa*. LSWR, 2014.
- [9] BSC-CNS Barcellona Supercomputing Center. Installation and administration manual. [http://compss.bsc.es/releases/compss/latest/docs/COMPSSs\\_Installation\\_Manual.pdf](http://compss.bsc.es/releases/compss/latest/docs/COMPSSs_Installation_Manual.pdf), 2018.
- [10] BSC-CNS Barcellona Supercomputing Center. Pycompss distributed data set user manual. [http://compss.bsc.es/releases/compss/latest/docs/DDS\\_Manual.pdf](http://compss.bsc.es/releases/compss/latest/docs/DDS_Manual.pdf), 2018.
- [11] BSC-CNS Barcellona Supercomputing Center. User manual application development guide. [http://compss.bsc.es/releases/compss/latest/docs/COMPSSs\\_User\\_Manual\\_App\\_Development.pdf](http://compss.bsc.es/releases/compss/latest/docs/COMPSSs_User_Manual_App_Development.pdf), 2018.
- [12] BSC-CNS Barcellona Supercomputing Center. User manual, application execution guide. [http://compss.bsc.es/releases/compss/latest/docs/COMPSSs\\_User\\_Manual\\_App\\_Exec.pdf](http://compss.bsc.es/releases/compss/latest/docs/COMPSSs_User_Manual_App_Exec.pdf), 2018.
- [13] BSC-CNS Barcellona Supercomputing Center. Comp superscalar overview. <https://www.bsc.es/research-and-development/software-and-apps/software-list/comp-superscalar/>, 2019.
- [14] Carmela Comito, Deborah Falcone, and Domenico Talia. Mining popular travel routes from social network geo-tagged data. *Intelligent Interactive Multimedia Systems and Services*, 40:81–95, 01 2015.
- [15] Jamie Condliffe. The week in tech: Can you put a price on your personal data? <https://www.nytimes.com/2019/06/28/technology/data-price-big-tech.html>, Giugno 2019.
- [16] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.
- [17] Apache spark ecosystem – complete spark components guide. <https://data-flair.training/blogs/apache-spark-ecosystem-components/>.

- [18] Jianqing Fan, Fang Han, and Han Liu. Challenges of Big Data analysis. *National Science Review*, 1(2):293–314, 02 2014.
- [19] George Firican. The 10 vs of big data. <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx>, 2017.
- [20] Sources of big data. <http://www.hadoopadmin.co.in/sources-of-bigdata/>, 2018.
- [21] Facebook for developers. Api graph. <https://developers.facebook.com/docs/graph-api>, 2019.
- [22] The Apache Software Foundation. Apache flink® - stateful computations over data streams. <https://flink.apache.org/>, 2019.
- [23] The Apache Software Foundation. Apache hama. <https://hama.apache.org/>, 2019.
- [24] The Apache Software Foundation. Apache hive tm. <https://hive.apache.org/>, 2019.
- [25] The Apache Software Foundation. Apache oozie workflow scheduler for hadoop. <https://oozie.apache.org/>, 2019.
- [26] The Apache Software Foundation. Apache storm. <https://storm.apache.org/>, 2019.
- [27] The Apache Software Foundation. Getting started with ambari. <https://ambari.apache.org/>, 2019.
- [28] The Apache Software Foundation. Welcome to apache giraph! <https://giraph.apache.org/>, 2019.
- [29] The Apache Software Foundation. Welcome to apache pig! <https://pig.apache.org/>, 2019.
- [30] The Apache Software Foundation. Welcome to apache zookeeper™. <https://zookeeper.apache.org/>, 2019.
- [31] The Apache Software Foundation. What is cassandra? <http://cassandra.apache.org/>, 2019.

- [32] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *Int J. Information Management*, 35:137–144, 2015.
- [33] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. Social media big data analytics: A survey. *Computers in Human Behavior*, 2018.
- [34] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.
- [35] William Gropp. *MPI (Message Passing Interface)*, pages 1184–1190. Springer US, Boston, MA, 2011.
- [36] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.
- [37] Hadoop, che cos’è e perchè è importante? [https://www.sas.com/it\\_it/insights/big-data/hadoop.html](https://www.sas.com/it_it/insights/big-data/hadoop.html).
- [38] Hadoop ecosystem. <https://www.geeksforgeeks.org/hadoop-ecosystem/>.
- [39] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, Jan 2004.
- [40] Abid Hussain and Ravi Vatrapu. Social data analytics tool (sodato). In Monica Chiarini Tremblay, Debra VanderMeer, Marcus Rothenberger, Ashish Gupta, and Victoria Yoon, editors, *Advancing the Impact of Design Science: Moving from Theory to Practice*, pages 368–372, Cham, 2014. Springer International Publishing.
- [41] MongoDB Inc. The database for modern applications. <https://www.mongodb.com/>, 2019.
- [42] Twitter Inc. Twitter developer docs. <https://developer.twitter.com/en/docs.html>, 2019.

- [43] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2009.
- [44] Dominique LaSalle and George Karypis. Mpi for big data: New tricks for an old dog. *Parallel Computing*, 40(10):754 – 767, 2014.
- [45] Chun Li and Jianyong Wang. Efficiently mining closed subsequences with gap constraints. In *Proceedings of the SIAM International Conference on Data Mining*, pages 313–322, 04 2008.
- [46] Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang, and Edward Y. Chang. Pfp: Parallel fp-growth for query recommendation. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys ’08, pages 107–114, New York, NY, USA, 2008. ACM.
- [47] Shanhong Liu. Global cloud data center ip traffic from 2015 to 2021. <https://www.statista.com/statistics/227267/global-cloud-ip-traffic-growth-by-segment/>, Febbraio 2019.
- [48] Shanhong Liu. Revenue from big data and business analytics worldwide from 2015 to 2022. <https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/>, Agosto 2019.
- [49] Ezio Melotti. Guida python. "<https://www.html.it/guide/guida-python/>", 2017.
- [50] Microsoft. Servizio di azure machine learning. <https://azure.microsoft.com/it-it/services/machine-learning-service/>, 2019.
- [51] Jared Oliverio. A survey of social media, big data, data mining, and analytics. *Journal of Industrial Integration and Management*, 03(03):1850003, 2018.
- [52] OpenStack. Swift. <https://wiki.openstack.org/wiki/Swift>, 2019.
- [53] Evelien Otte and Ronald Rousseau. Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28:441–453, 12 2002.

- [54] Lucas M. Ponce, Walter dos Santos, Wagner Meira Jr., and Dorgival Guedes. Extensão de um ambiente de computação de alto desempenho para o processamento de dados massivos. In *Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Porto Alegre, RS, Brasil, 2018. SBC.
- [55] Python. Cos'è python. "<http://www.python.it/about>", 2019.
- [56] Kit Smith. 126 amazing social media statistics and facts. <https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/>, Giugno 2019.
- [57] Domenico Talia, Paolo Trunfio, and Fabrizio Marozzo. *Data Analysis in the Cloud: Models, Techniques and Applications*. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1st edition, 2015.
- [58] Enric Tejedor, Yolanda Becerra, Guillem Alomar, Anna Queralt, Rosa M Badia, Jordi Torres, Toni Cortes, and Jesús Labarta. Pycompss: Parallel computational workflows in python. *The International Journal of High Performance Computing Applications*, 31(1):66–82, 2017.
- [59] J. Wang and J. Han. Bide: efficient mining of frequent closed sequences. In *Proceedings. 20th International Conference on Data Engineering*, pages 79–90, April 2004.
- [60] Wikipedia. Social media — wikipedia, l'enciclopedia libera, 2019.
- [61] Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas S. Huang. Diversified trajectory pattern ranking in geo-tagged social media. pages 980–991, 04 2011.
- [62] L. You, G. Motta, D. Sacco, and T. Ma. Social data analysis framework in cloud and mobility analyzer for smarter cities. In *Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics*, pages 96–101, Oct 2014.
- [63] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA, 2014.
- [64] Yu Zheng. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.*, 6(3):29:1–29:41, May 2015.