RightInsight: Open Source Architecture for Data Science*

Ahmet Bulut

Department of Computer Science
İstanbul Şehir University
Altunizade Mah. Oymaci Sok. No:15
34660 Uskudar, Istanbul
ahmetbulut@sehir.edu.tr

Abstract: We give the details of our reference architecture called *RightInsight* for enabling rapid data science. RightInsight is based purely on open source technologies. The data is stored in a standard distributed file system such as HDFS. The stored data is processed in Apache Spark, which provides an enhanced Map/Reduce programming environment. Its rich and powerful machine learning base makes it easy to construct descriptive, prescriptive, and predictive models. In addition to providing an agile environment for making sense of the data and the data science problem at hand, its Python-based middleware with a wide array of scientific libraries such as scipy, numpy, matplotlib, and pandas, enables interactive and exploratory data analysis. The ability to ask questions, especially the right questions, and to do what-if analysis is extremely important for any serious data science project. The results of such exploratory analyses are stored in a suitable format that is easily consumable in the Web tier. Using rich JavaScript libraries such as data driven documents and bootstrap, the formatted data can be visualised within a Web browser in creative ways for rapid insight discovery.

1 Introduction

Data scientists are inquisitive. They explore the problem space, ask questions, do what-if analysis, and while doing so they question their existing assumptions and processes. Rather than looking at data from a single source, they examine data from multiple and disparate data sources. All incoming data is sifted through with the goal of discovering hidden insights, which in turn can be used as a competitive business advantage or can be used towards coming up with solutions to pressing business problems. With the necessary analytics and tools support, a capable data scientist will be well equipped to communicate informed conclusions and recommendations across the whole organization [IBM14b].

The necessary analytics and tools support varies. From clustering and regression, to classification and probabilistic inference, and to data enrichment and visualisation, data scientists need to have a solid foundation in computer science and applications, modelling,

*This project is funded by Turkish National Science Foundation (Tubitak) under grant numbers 113E243 and 113E622.

statistics, analytics and mathematics. In order to explore exabytes of data and do what-if analysis, data scientists require powerful back-end systems so-called data science platforms in order to digest raw data. The platforms have to provide an interactive mode of data analysis required due to the iterative and inquisitive nature of data science projects.

Figure 1: The collaborative nature of data science projects, which require business analysts, business users, data scientists, and IT working together for a common objective.

A recent research study conducted by Ventana Research revealed that having the right tools are the most essential for predictive analytics [Res13]. Many organizations have made the necessary human capital investment by hiring experts, who have the core responsibility of ensuring success in predictive analytics. 70% of these companies with business foresight emphasize the need for tools with enhanced usability, being workflow-driven, and ability to handle any information source. The emphasis on usability reiterates the collaborative nature of data science projects, which require business analysts, business users, data scientists, and IT working together for a common objective as shown in Figure 1.[1]

[1]The depicted picture is drawn using the following source images at:
http://www.montel.com/en/markets/business/manufacturing/manuf-factor-it-department
http://blogs.sap.com/innovation/analytics/data-scientist-sexiest-job-century-01242806
http://www.computing.co.uk/IMG/768/231768/datascienceman-370x229.jpg
http://www.phoenixit.im/business_user.html
http://enterprise-dashboard.com/2009/07/
http://thebusyba.com/business-process-modeling-for-new-business-analysts/

1.1 Organization of the paper

The details of RightInsight architecture is discussed in Section 2. In Section 3, we present a representative data science project that exemplifies the primary uses cases for RightInsight. Since data science projects generally require collaboration of multiple stakeholders from multiple departments, i.e., being interdisciplinary in nature and scope, we present two models of interaction that describe how to collaborate in Section 4. Finally, we discuss our current agenda in Section 5 and emphasize key points to take away in Section 6.

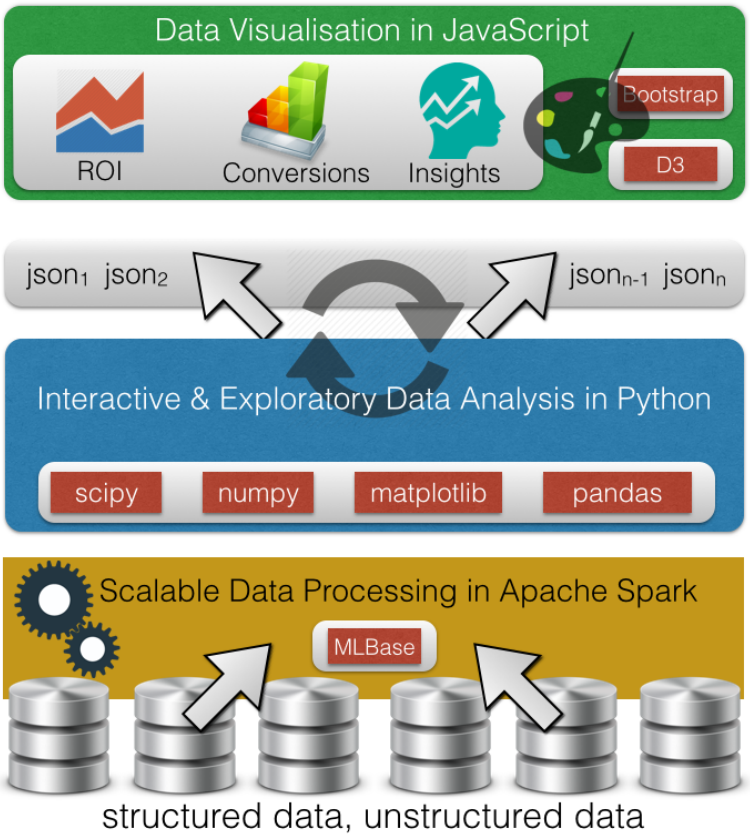

Figure 2: The RightInsight architecture for big data science depicted as a workflow.

2 RightInsight Architecture

RightInsight is architected for enabling rapid data science. The architecture is shown in Figure 2. RightInsight is built on open source technologies. Structured, semi-structured, or unstructured data are all stored in a standard distributed file system such as HDFS. The stored data is processed in Apache Spark [ZCF+10], which provides an enhanced Map/Reduce programming environment [DG08]. In order to construct descriptive, prescriptive, and predictive models using the available data, Spark's rich and powerful machine learning base called MLBase is utilised [KTD+13]. On top of it, a Python suite that includes such widely used scientific libraries as scipy, numpy, matplotlib, and pandas enables interactive and exploratory data analysis and the associated model analysis [vR97]. The ability to ask questions and do what-if analysis is extremely important for most data science projects. The agility of the underlying Python-based platform expedites the process of making sense of the data. The results of exploratory analyses made are stored in JavaScript object notation (json). Using a rich set of JavaScript libraries such as data driven documents [BOH11] and bootstrap [Ler12], the information in json can then be visualised inside a Web browser in creative ways for discovering insights and for furthering team communication and collaboration.

There are other computing platforms available for enabling data science. StarCluster is the closest to RightInsight. StarCluster is designed to automate and simplify the process of building, configuring, and managing compute clusters on Amazon's EC2 cloud [MIT13]. A StarCluster suits the computing needs of most distributed and parallel computing applications. The primary difference between RightInsight and StarCluster is the base platform used for scalable data processing. For a StarCluster, OpenMPI, Hadoop, and NFS are used as the base distributed data processing platform [GFB+04]. On the other hand for RightInsight, Spark ecosystem provides a superior performance in processing big data files. More specifically, Spark outperforms Hadoop by up to 80 times in iterative machine learning and graph applications [XRZ+13]. In both RightInsight and StarCluster, a Python suite is used as an interactive shell for agile data analysis.

In Spark, jobs are responsible for analytic or iterative computing on a large dataset. Each job first loads its working data into memory for enabling rapid data access. The main component of Spark is the construct of a resilient distributed dataset (RDD). An RDD provides granular fault tolerance and distribution of work among multiple worker nodes. The original input data is sliced into multiple chunks so that multiple jobs can be created for execution in parallel on each chunk. Fault tolerance is supported through the lineage information stored for a compute flow in the framework per RDD. In case of node failures, each compute step in the compute flow can be re-executed linearly in order to recover. In order to synchronise compute nodes when necessary for certain iterative end-user tasks, RightInsight uses broadcasting feature of Spark for in-flight data and uses Tachyon for data at-rest in contrast to the use of NFS in StarCluster. Tachyon is an in-memory based distributed file system, which enables memory-speed file sharing across cluster frameworks, and which acts as an intermediate layer between HDFS and Spark [LGZ+14].

IBM Watson Analytics is one of the most compelling data science enablers in the industry [IBM14a]. Watson Analytics brings together IBM Cognos Business Analytics capabilities

and IBM SPSS Modelling and Statistics functionalities in order to capture the most relevant facts, patterns and relationships in data, and presents them in an engaging analytics experience to the end users. The primary difference between RightInsight and Watson Analytics is that Watson is built on IBM's proprietary software stack; whereas RightInsight is built on top of an open source technology stack. Similar to the use of Hadoop in StarCluster, Watson uses a proprietary version of Hadoop called IBM BigInsights for improved performance. To the best of our knowledge, there are no benchmarks that compare Spark's performance with IBM BigInsights performance on competitive workloads.

3 Representative Data Science Application

In IBM's Smarter Planet initiative, RightInsight is being proposed as a candidate framework for holistic event analysis in a trillion giga big data ecosystem. Our planet continuously gets smarter and more complex. Heterogeneous information sources on the planet produce streams of events. These information sources could be external sensors (embedded in cars, home appliances, city roads, pipelines, medicine and livestock), internal systems in use, and we the people. Each event is associated with numeric or textual data, or other richer media. From 2005 to 2010, over 33 billion RFID tags were produced, which helped to grow the digital universe from 130 Exabyte (EB) to 800 EB. The digital universe will double every two years till 2020. In 2020, it will be 40 trillion gigabytes, which is more than 5,200 gigabytes for every man, woman, and child in 2020 [GR12].

In a trillion giga big data ecosystem, the collected events may not make sense when evaluated independently but they could indicate much larger developments and/or evolutions when evaluated holistically. Our overarching hypothesis is that there is a higher order generative process that produces the observed events, and our goal is to identify and estimate the parameters of that process through holistic event analysis. Once the inherent generative process is identified, it can be used for testing existing assumptions and doing what-if analysis. For instance, if a harmful development is at work, a security policy can be enforced in order to counteract it. The planned discovery process can be used for measuring the efficacy of the policies enacted and for driving new actionable insights. For holistic event analysis, streams of events have to be monitored for discovering associations, inferring generative themes, and capturing structural relationships. As an example, causal relationships can be captured using Conditional Random Fields (CRFs) [RH13] that are primarily used for structured prediction. Generative themes can be inferred using Latent Dirichlet Allocation (LDA) that is used for learning probability distributions [Bul14]. In this project, we will focus on news events available on the Web. Since news events could arise from disparate data sources, the data processing infrastructure have to be capable of handling the heterogeneity in data. Furthermore, having large amounts of data to work with prompts us to use Apache Spark on top of a Hadoop layer for scalable data analytics.

Holistic data analysis is extremely relevant to IBM's core business [TH11]. The capability to move from reaction to prediction is identified as one of the core business strengths for organizations. Organizations need to interpret and respond to streaming sentiments and dialogues among consumers for taking predictive actions rather than reactive actions.

That is, there is a shift towards increasing the information responsiveness of companies of all sizes. RightInsight can turn data into dollars through holistic event analysis. The generative process learned from the data can be used for predicting the near future or for simulating what can happen in the near future within a short look-ahead window.

At the beginning of 2014, we received funding from National Research Council of Turkey in order to develop a system called BOSS for inferring latent themes in dynamically evolving text collections. There is abundant textual data that evolves over time: Facebook status updates, product comments on e-commerce websites, articles on news sites, tweets, and business transactions to name a few. The goals in BOSS project is in line with the goals in holistic event analysis since inferred latent themes correspond to semantic information that is captured over raw data. And as such, the higher-level semantics helps decision makers drive actionable insights.

4 Collaboration Models for Data Science

Bringing together multiple stakeholders from multiple disciplines and backgrounds is not a straightforward undertaking. Viable collaboration environments and processes need to be crafted for increasing productivity and interaction while not hindering the participation of any stakeholder. We propose two candidate collaboration models for this purpose: (1) OODA and (2) Active Learning.

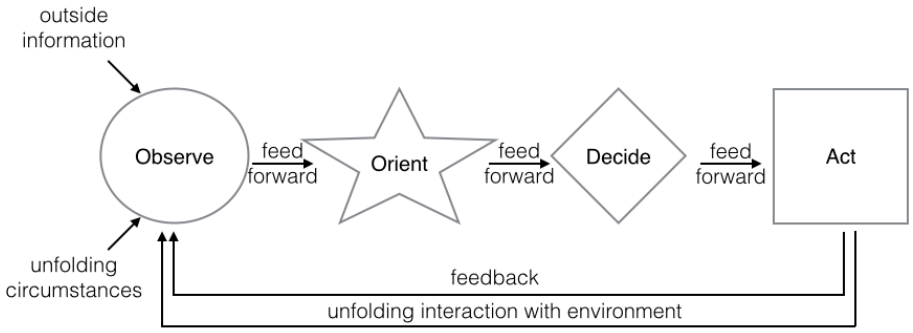

Figure 3: Observe, Orient, Decide, Act (OODA) Loop. (Adapted from http://upload.wikimedia.org/wikipedia/commons/3/3a/OODA.Boyd.svg).

4.1 Observe, Orient, Decide, and Act

In today's highly competitive business environment, all companies that want to stay ahead of the curve have to shorten the time it takes from the point of an observation being made to the point when an action is taken based up on that observation. This is how systems get

smarter, how organizations get more agile, and more importantly how people win.

The time it takes from the point of an observation being made to the point when an action is taken based up on this particular observation is the realization and one cycle in the Observation (O), Orientation (O), Decision (D), and Action (A) loop, i.e., OODA loop as shown in Figure 3 [Boy95]. For most types of decision-making processes in the business world, OODA presents a viable execution model.

RightInsight can facilitate the OODA collaboration model in data collection, analysis of the collected data, and in decision-making through controlled experimentation [KLSH09]. The process was proposed for enabling data-driven policy making in smart cities [Bul12]. It is depicted in Figure 4. The more loops are completed, the more patterns are revealed. Such patterns are catalysts for engineering automated drivers that correspond to these patterns and for crafting expert systems.

Figure 4: Six step process to orchestrate OODA loops.

Using RightInsight, we can detect anomalies using custom monitors, diagnose fault by comparing operational data with baselines, and finally apply control based on diagnosis [HY06]. This whole process describes a single OODA loop. The actions that are taken based on the decisions given at the last step affect the underlying social and physical processes. These processes give rise to a totally new set of observations. This in turn marks the beginning of the next OODA loop.

4.2 Active Learning

Imagine the following scenario that describes how a historian and a computer scientist collaborates. By using RightInsight, the computer scientist cultivates from a heterogeneous set of digital information sources prepares a portfolio of “interesting” events. The event

sources are spatio-temporal. At the beginning, the collected events are seemingly interesting events. Furthermore, the number of events presented to the historian is limited. The historian evaluates the suggested events and identifies which events are indeed interesting and may be worth delving deeper into. Armed with this insight, the computer scientist goes back to cultivation and drills in further for coming up with more interesting events to present. This process continues indefinitely with one caveat. Since the historian is a precious resource with limited time and energy, the computer scientist and the system in turn has to actively learn from the historian.

"Active Learning" defines the interdisciplinary collaboration between a computer scientist and a social scientist. Technically, active learning is used for training classifiers with less training data than needed during a regular supervised training [Set09]. The key idea behind active learning is that when the learning algorithm is allowed to choose the data from which it learns, then it can perform up to par with less training data. This is valuable in situations where unlabelled data is abundant but labelling them is expensive.

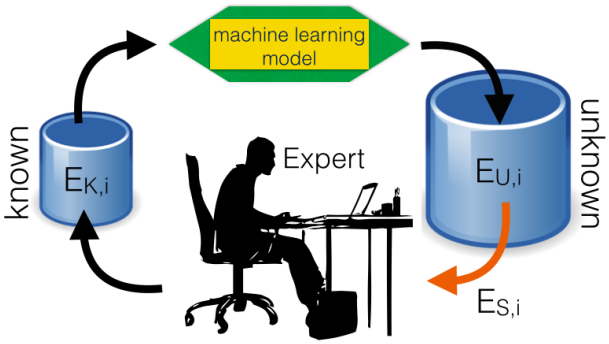

Figure 5: Illustration of an active learning cycle.

In our sample scenario, a custom data selector that is built based up on the historian's domain expertise can speed up the learning process with less training data. RightInsight provides the necessary tools suitable to interact with data and enables iterative machine learning for easing exploratory analysis outlined in the scenario. Let E be the total set of all events including those events that are known to have a certain interesting characteristic. During each active learning step i as depicted in Figure 5, the set E is broken up into three subsets as known events $\mathbf{E}_{K,i}$, unknown events $\mathbf{E}_{U,i}$, and $\mathbf{E}_{S,i} \subset \mathbf{E}_{U,i}$ that is selected for expert opinion. There is an active body of research on the best method to use for event selection. All events that carry a certain characteristic that are known to exist in previously known interesting events can be chosen for expert opinion. However this naive approach does not explore parts of the event space that is yet to be explored. The events to select is an interplay between the exploration and the exploitation over the event space representation. Thomson sampling, which is an implementation of exploration and exploitation tradeoff, has been shown to work well for display advertisement selection and news article recommendation [CL11]. Thomson sampling is used for event selection in RightInsight.

5 Discussion

RightInsight architecture is instrumental for us to set up a center of excellence in data science. The center stands on the shoulders of human resources and compute resources. Academicians and data scientists are the key human resources. A masters program in Data Science, which is a joint venture with IBM Turkey, sustains the human resources pipeline and supports the ongoing skills development. The compute resources are provisioned among various in-house data science projects. The initial infrastructure is an elastic private cloud; a hybrid or a purely public cloud will be considered when necessary.

6 Conclusions

RightInsight is an enabler of data science. For agility, it is built on top of open source technologies. Heterogenous data is stored in a standard distributed file system, and then processed iteratively in Apache Spark. Complex descriptive, prescriptive, and predictive models can be built on the collected data using a rich and powerful machine learning base called MLBase. On top of the base data processing platform, a Python suite enables interactive & exploratory data analysis, and the associated model analysis. The agility of the Python-based suite expedites the process of making sense of the data. The results of exploratory analyses made are visualised in a Web browser in creative ways for discovering insights and for furthering team collaboration.

References

[BOH11] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011.

[Boy95] J. Boyd. The Essence of Winning and Losing. In *United States Air Force*, 1995.

[Bul12] A. Bulut. *City Competitiveness and Improving Urban Subsystems: Technologies and Applications*, chapter City Competitiveness and Infrastructure, pages 1–20. IGI Global, November 2012.

[Bul14] A. Bulut. TopicMachine: Conversion Prediction in Search Advertising Using Latent Topic Models. *IEEE Transactions on Knowledge and Data Engineering*, 26:2846 – 2858, November 2014.

[CL11] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In *25th Neural Information Processing Systems (NIPS)*, pages 2249–2257, December 2011.

[DG08] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *ACM Communications*, 51(1):107–113, January 2008.

[GFB+04] E. Gabriel, G. E. Fagg, G. Bosilca, T. Angskun, J. J. Dongarra, J. M. Squyres, V. Sahay, P. Kambadur, B.Barrett, A. Lumsdaine, R. H. Castain, D. J. Daniel, R. L. Graham, and T. S. Woodall. Open MPI: Goals, Concept, and Design of a Next Generation MPI

Implementation. In *11th European PVM/MPI Users' Group Meeting*, pages 97–104, September 2004.

[GR12] J. Gantz and D. Reinsel. Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. In *IDC iView sponsored by EMC*, December 2012.

[HY06] T. Han and B. S. Yang. Development of e-maintenance system integrating advanced techniques. *Computers in Industry*, 57:569–580, 2006.

[IBM14a] IBM. Watson Analytics. In *http://www.ibm.com/analytics/watson-analytics/*, 2014.

[IBM14b] IBM. What's a data scientist. In *http://www-01.ibm.com/software/data/infosphere/data-scientist/*, 2014.

[KLSH09] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[KTD+13] T. Kraska, A. Talwalkar, J. Duchi, R. Griffith, M. J. Franklin, and M. Jordan. MLbase: A Distributed Machine-learning system. In *6th Biennial Conference on Innovative Data Systems*, 2013.

[Ler12] R. M. Lerner. At the Forge: Twitter Bootstrap. *Linux Journal*, 2012(218):6, 2012.

[LGZ+14] H. Li, A. Ghodsi, M. Zaharia, S. Shenker, and I. Stoica. Tachyon: Reliable, Memory Speed Storage for Cluster Computing Frameworks. In *ACM Symposium on Cloud Computing*, November 2014.

[MIT13] MIT. StarCluster. In *http://star.mit.edu/cluster/*, 2013.

[Res13] Ventana Research. Deploying Predictive Analytics for Competitive Advantage. In *http://www.slideshare.net/VentanaResearch/ventana-research-infographicpredictiveanalyticsibm2013final*, 2013.

[RH13] K. Radinsky and E. Horvitz. Mining the Web to Predict Future Events. In *6th ACM International Conference on Web Search and Data Mining*, pages 255–264, 2013.

[Set09] B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[TH11] M. Teerlink and M. Haydock. Customer analytics pay off. In *IBM Institute for Business Value*, 2011.

[vR97] G. van Rossum. Scripting the Web with Python. *World Wide Web Journal*, 2, 1997.

[XRZ+13] R. S. Xin, J. Rosen, M. Zaharia, M. J. Franklin, S. Shenker, and I. Stoica. Shark: SQL and Rich Analytics at Scale. In *ACM International Conference on Management of Data (SIGMOD)*, pages 13–24, 2013.

[ZCF+10] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *2nd USENIX conference on Hot topics in Cloud Computing*, pages 10–10, 2010.