Marketing Knowledge Discovery and Big Data Analytics. Towards reducing technological entry barriers for marketing scientists

Michel CALCIU
Université Lille, RIME-Lab, France
mihai.calciu@univ-lille.fr

Jean-Louis MOULINS
Aix Marseille Université, CRETLOG, France
jean-louis.moulins@univ-amu.fr

Francis SALERNO
Université Lille, LEM, France
francis.salerno@univ-lille.fr

Abstract

Big Data Analytics (BDA) has become an important source of competitive advantage for companies. By its high operational and strategic potential it may substantially increase business efficiency and effectiveness, foster agility and change the competitive game. BDA are best implemented in the cloud in a servitized way. In Marketing who is at the forefront of Big Data generation and use, analytics need to follow this trend in order to cope with Big Data Volume, Velocity and Variety (3V).

We introduce an evolutionary view of the technological advances for marketing analytics in order to distinguish analytics in the Cloud as the strongly preferred pattern for dealing with Big Data. We insist on some new competences marketing scientist, need to acquire in order to deal with BDA in the Cloud. In contrast with analytic solutions that don't need more than one computer to be implemented, BDA that require clusters of computers come with additional orchestration, provisioning and deployment challenges. Apart virtualization as a technological foundation, upon which cloud computing relies, we show some benefits, facilities and simplifications brought over by containerization for BDA.

We finally suggest marketing analysts, that include academic model builders and data scientists, need to have a grasp of service oriented and cloud technologies as this is increasingly the way, and for Big Data it is probably the only way, analytic solutions can be brought to the market.

Keywords

Big-data, Marketing science, Cloud computing

Introduction

Big Data Analytics (BDA) has become an important source of competitive advantage for companies. More generally Business Analytics has been regularly reported, over the past decade, to be a top priority for many top-level managers (Holsapple, Lee-Post, & Pakath, 2014). The analytics process, including the deployment and use of BDA tools, has strategic potential and is seen by organizations as a means to improve operational efficiency, drive new revenue streams and gain competitive advantages over business rivals (Sivarajah & al, 2017). A study by Anderson (2015) showed that every $1.00 spent on analytics applications pays off $13.0.

In a narrow sense BigData (BD) are data that are too big to be dealt with using one computer. Besides *Volume*, there is *Variety,* because such data include textual content (i.e. structured, semistructured as well as unstructured), to multimedia content (e.g. Videos, images, audio) on a variety of platforms (e.g. Machine-to-machine communications, social media sites, sensors networks, cyber-physical systems, and Internet of Things [IoT]). This huge amount of complex and heterogeneous data pouring from any-where, any-time, and any-device resembling to a Data Deluge defines a new era of Big Data. It represents a datafication process (Mayer-Schönberger & Cukier, 2013) in which virtually anything and everything can be documented, measured, and captured digitally, and transformed into data.

In essence, BD is the artifact of individual and collective intelligence generated and shared mainly through the technological environment (Sivarajah & al, 2017). Some called BD as the New Raw Material of the 21st century (Berners-Lee & Shadbolt, 2011) it may substantially increase business efficiency and effectiveness, foster agility and change the competitive game. Therefore *Velocity* is a third defining element of BD, it demands appropriate data processing and management in order to expose new knowledge, and facilitate in responding to emerging opportunities and challenges in a timely manner (Chen et al., 2013). BD needs to be analyzed in a way that brings big *Value* (Sivarajah & al, 2017) and can be regarded as today's Digital Oil (Yi, Liu, Liu, & Jin, 2014). Besides the 3Vs (Volume, Variety and Velocity), Value, Variability, Veracity, Visualization are additional BD defining elements.

BD opportunities and marketing transformation

BD opportunities include value creation (Brown, Chui, & Manyika, 2011), rich business intelligence to support decisions (Chen & Zhang, 2014), and enhancing the visibility and flexibility of supply chain and resource allocation (Kumar, Niu, & Ré, 2013). Besides transforming marketing as technology always did, BD makes marketing "transformative". Kumar (2018, p. 2) suggests that "technological advancements and data-processing capabilities, make the marketing function ripe to enter a phase of transformative marketing". He offers a dialectic definition in which both marketing is transformed by the environment and the environment is transformed by marketing.

On one side the marketing function is being transformed by environment due to:

- a) societal changes (changing customer landscape which becomes more niche, more distinct)
- b) changes in the nature of markets (blurred geographical boundaries, technology acting as a powerful market integrator that affects competitive advantage generating locational or product differentiation difficulties.
- c) Multiplication of media outlets

On the other side the business environment is transformed by the marketing function through:

- data tracked (enabling employees and companies to offer relevant content and offerings),
- technology (agents, virtual reality, facial recognition, geofencing etc. enable realtime response, personalized messages and improve the customer experience at each touchpoint along the customer's journey),
- privacy factors (in an interconnected world privacy becomes an issue and for marketing it remains an essential ingredient in forming relationships)

We think that transformative marketing can be seen as the next 20 years phase in the evolution of marketing that continues the paradigmatic shift from transactional to relationship marketing as shown in table 1.

Table 1– Multi-decennial phases in the evolution of marketing

1960s	1980s	1990/2000s	Current	Next 20 years
Mass	Targeted	Personalized content	Engaged customization	Transformative
Transactional			Relationship	

Adapted from Kumar (2018)

Our approach focuses on marketing scientists or analysts as a key human resource and the need to operate, as regards their skills, a both technical and cultural paradigm shift from "data" to "big-data". Technical "big-data" skills refer to the know-how required to use new forms of technology to extract intelligence from big-data. Marketing scientists have a long tradition in developing and applying models and analytics to market and consumer data. Their valuable skills should constitute a powerful resource to build BDA capability.

Understanding of the technological advances that have launched the Big Data Revolution

We think that historical understanding of the technological advances that have launched the Big Data Revolution is fundamental for marketing scientists who want to acquire the additional skills needed for modeling and dealing with such data. We also think that acquiring such skills is essential in order to take advantage of the distributed and data-rich context provided by the Internet, cloud computing and High Performance Computing (HPC). Neglecting factors that enhance usability of models models and analytics developed by marketing scientists risks to make the latter irrelevant and limit their use.

The main points we develop and illustrate with marketing applications in this research are:

- Understanding key technological advances that launched the Big Data Revolution and the role of the MapReduce approach in democratizing BDA calculations
- Understanding the need for hardware, software and data scalability
- Understanding hardware virtualization and software servitization in the Cloud
- Understanding the role of containerization to facilitate orchestration, provisioning, and deployment of BigData calculation infrastructure and platforms.

Understanding key technological advances that launched the Big Data Revolution

We think that the main stages in the arousal of the BigData phenomenon were: Internet that allowed to potentially connect all computers in the world, the World Wide Web that democratized the use of the Internet and generated the "Data, data everywhere nightmare" and finally the MapReduce approach that democratized calculations on "monstrous" amounts of data.

Internet and WWW origin of BD revolution: The Internet (contraction of interconnected networks) as the global system of interconnected computer networks that use the Internet protocol suite (TCP/IP) to link devices worldwide is probably the origin of the BD revolution. This unimaginable network of networks, evolved almost silently during more than two decades, its services being mainly used by researchers until the beginning of the nineties, when Tim Berners Lee (1989) invented the World Wide Web (WWW).

Data, data everywhere: The democratization of the use of Internet, its continuously evolving web applications and the dominance of e-platforms and social-media came along, among others, with endless data flows, a kind of "data tsunami" that almost submerged businesses and institutions. Help was needed for "Shipwrecked" managers who saw "Data, data everywhere, and not a byte to use" which is a paraphrase of "Water, water, everywhere, Nor any drop to drink" from Coleridge's The Rime of the Ancient Mariner.

MapReduce to democratize BD calculations: High Performance Computing (HPC) most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business. It usually deals with supercomputers and/or computer clusters. Although under continuous development, HPC, dealing with huge amounts of data using supercomputers or computer clusters was regarded by most users, but also by most data scientists and in particular marketing scientists, as something not worth worrying about, or too expensive, or even as a myth or something that appears as unattainable. It is finally the MapReduce approach that helped democratize HPC. For more details see Calciu & al. (2016)

MapReduce defined: MapReduce is a high level programming model and an associated implementation for large scale parallel data processing. It has the merit to hide all complexities of parallel computing on distributed servers from users and to have contributed massively to democratize BigData processing. The name MapReduce originally referred to the proprietary Google technology (Dean and Ghemawat, 2004), but has since become a generic trademark. It is integrated in Apache's Hadoop , an open-source software framework, written in Java, for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. It is useful for marketing scientists to have a grasp of MapReduce in order to be able to adapt their models and algorithms to distributed parallel processing of massive datasets.

MapReduce is based on the observation that most Big Data computations can be expressed in terms of a Map() procedure that enables filtering and sorting and of a Reduce() procedure that performs an aggregating operation like counting, summing etc. Map and Reduce are common Higher-Order Functions in Functional Programming.

Understanding the need for hardware and software scalability

«Small is beautiful» is often used to praise small, appropriate technologies that are believed to

empower people more, and is opposed to "bigger is better". It is the title of a highly successful book on economics that reached millions (1973, 1999). E.F Schumacher, the author who took this phrase from his teacher L. Kohr, was interested not in smallness but in appropriateness of scale as a challenge to the 'too big to fail' showing premonition concerning the bank crisis we witnessed recently.

Scalability is the capability of a system to grow or to handle a growing amount of work. It is analogous with economic scalability of a company implying that the underlying business model offers the potential for economic growth within the company.

Marketing Decision Support Systems (MDSS), the classical way for delivering marketing analytics, have been first defined by Little (1979, p.11) as "a coordinated collection of *data, models, analytical tools and computing power* by which an organization gathers information and turns it into a basis for action." Although Little's decision calculus approach[1] defends the "small is beautiful" point of view while fostering usability his MDSS definition remains valid today and can be extended to include Big Data analytics with hardware and software scalability (Calciu & al. 2017).

Computing power or Hardware scalability means that computing power for BigData calculations, can be increased economically by regrouping several commodity computers in a network in order to form a computer cluster. Today's mainstream computer hardware is relatively cheap and almost infinitely replicable. It is much cheaper to purchase 8 off-the-shelf, "commodity" servers with eight processing cores and 128 GB of RAM each than to acquire a single system with 64 processors and a terabyte of RAM. While there exist other alternatives to grow computing power for analyzing very large datasets, the most successful, flexible and economic way is by distributed computing usually accessed though a cloud. Cloud computing provides the tools and technologies to build data/compute intensive parallel applications with much more affordable prices compared to traditional parallel computing techniques.(Hamdaqua & Tahvildari, 2012).

Data scalability is a subject that can be extended to data format, database, no-sql DB or even data-warehouse aspects. Here we focus on network data storage approaches that facilitate BigData calculations: Virtual volumes, HDFS and RDDs. *Virtual volumes* on a cloud are persistent storage devices that may be attached and detached from computer instances, like an external hard drive. They do not provide shared storage in the way a network file system (NFS) does. *Hadoop Distributed File System (HDFS)* which is the preferred way to store BigData. Besides providing shared storage, it is also distributed, fault-tolerant and scalable. It provides very high aggregate bandwidth across the cluster. Fault tolerance in this case is achieved through data replication (3 times). Fault tolerance is necessary in distributed calculations, because the probability of a fault in a cluster of computers increases exponentially with the size of that cluster. *Resilient Distributed Datasets (RDDs)* besides being the main idea behind Spark the BigData engine that increased calculation speeds up to 100 times compared to previous MapReduce solutions, achieve fault tolerance through a notion of lineage (see DAG further): if a partition of an RDD is lost, the RDD has enough information to rebuild just that partition. This removes the need for replication to achieve fault tolerance. Besides resilience which is defined as the ability (of the network) to provide and maintain an acceptable level of service in the face of various faults and challenges to normal operation, RDDs are ".. parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and ma-manipulate them using a rich set of operators" (Zaharia et al., 2012)

Computational scalability can be obtained by using the MapReduce approach and later developments

1 decision calculus models, in order to be used, had to be "simple, robust, easy to control, adaptive, as complete as possible and easy to communicate with (Little, 1970, p. 466).

like Apache Spark in order to scale and distribute calculations over computer clusters. They use Share-Nothing as the strongly preferred distributed-computing pattern in a cluster of commodity computers. At this stage of technology the alternative classical Share-Everything pattern is not a choice because network latency and congestion are the bottleneck relative to main memory and local storage. The key contributions of MapReduce implementations are scalability and fault tolerance achieved by optimizing the execution engine.

Software scalability is influenced by many factors, ranging from syntax details to component abstraction constructs, among which a subtle combination of object-oriented and functional programming is the most relevant. It is probably best embodied in a relatively new computer language called Scala. "..in Scala a function value is an object. Function types are classes that can be inherited by subclasses. This might seem nothing more than an academic nicety, but it has deep consequences for scalability". (Odersky & al, 2011, p.55). Also in order to better understand what recommends Scala as a platform for statistical computing and data science one could refer to Wilkinson (2017).

Understanding the virtualization of hardware and software servitization in the Cloud

Hardware virtualization as implemented in computer clouds can be seen as a de-materialization of "bare metal" computer clusters. The main enabling technology for cloud computing is virtualization. Virtualization software separates a physical computing device (bare metal environment) into one or more "virtual" devices, each of which can be easily used and managed to perform computing tasks. With operating system level virtualization essentially creating a scalable system of multiple independent computing devices, idle computing resources can be allocated and used more efficiently. *Virtualization* provides the agility required to speed up IT operations, and reduces cost by increasing infrastructure utilization.

The American National Institute of Standards and Technology (NIST) defines c*loud computing* "as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Clouds are composed of three service models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS)

Software servitization is another fundamental aspect of cloud computing. The shift from local computer toolbox-centered to web-based service-centered analytics software, has preceded cloud and big-data computing.

Web services (WS) are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web (Calciu & al., 2013). RESTful Web services are gaining increased attention because of their publishing and consumption simplicity (Vinowski, 2002). REST (REpresentational State Transfer) solution takes its inspiration from the web itself and shows that the same principles that have made the success of the World Wide Web can be used to solve enterprise application integration problems and to simplify service-oriented architectures (Calciu & al., 2013). These web service standards and approaches, and particularly REST are used in service-oriented (SO) solutions for the cloud that are preferred in Big Data calculations.

Implementing Big Data Analytics in the Cloud

Implementing Big Data Analytics in the Cloud comes with additional complexity. Apache Hadoop the today's preferred big-data solution has become an ecosystem formed by numerous software projects

that make up the services required by an enterprise to deal with Big Data in an agile way. For example Hortonworks, a major data analytics vendor, groups these open-source components into a Data Platform with 23 components[2] associated to five pillars: Data Management (2), Data Access (13), Data Governance & Integration (3), Security (2), Operations (3). For the computational aspects several of these Hadoop ecosystem components need to be *installed* on all Virtual Machines of a cluster. In contrast with analytic solutions that don't need more than one computer to be implemented, BDA on clusters of computers comes with additional orchestration challenges for marketing scientists that include :

- *Configuring* the local (client) computer to use the remote cloud
- *Provisioning the remote scalable BDA infrastructure*
- *Deploying the computational ecosystem components*

A more detailed discussion of these aspects can be found in Calciu & al. (2019).

Containerization a better way to deliver BDA

Containers, a new trend in virtualization, may substantially reduce the BDA orchestration burden. Containerization has gained ground as an alternative to virtualization. In fact containerization can be seen as a special kind of virtualization that occurs at operating system level while in Virtual Machines it occurs at hardware-level. Historically, in Unix systems, the first containers just provided isolation of the root file system (via chroot). Later FreeBSD jails extended this to additional namespaces such as process identifiers. A modern container is more than just an isolation mechanism: it also includes an image that contains the files of the application that runs inside the container (Burns & al., 2016). While much more lightweight than Virtual Machines (VMs), containers provide resource-management tools that make running applications efficient. They also provide robust kernel-level resource isolation to prevent the processes from interfering with one another. Docker is a tool that can package an application and its dependencies in a virtual container that can run on almost any server, out of the box, without installing any software. The resulting flexibility and portability allowing running applications everywhere. Kubernetes is an open source container-centric management environment developed at Google, that facilitates both declarative configuration and automation. It orchestrates computing, networking, and storage infrastructure on behalf of user workloads. This provides much of the simplicity of Platform as a Service (PaaS) with the flexibility of Infrastructure as a Service (IaaS), and enables portability across infrastructure providers. While container technology serves two key functions: software packaging and kernel privilege segmentation, Kubernetes extends on these key functionalities further to enables programmable, flexible, rapidly deployable environments. This is very useful for deploying BDA engines and their ecosystem.

Building the BDA infrastructure and software ecosystem in the cloud

As to our knowledge this is the first academic marketing attempt in France to apply BDA in the cloud. As acknowledged by some authors and reviewers in the latest special issue on Big Data of Marketing Science (Liu & al., 2016), academia lags behind industry in conducting cloud analytics. They state that cloud tools, like Spark, used by the industry go beyond the simple MapReduce programming model

2 Apache Hadoop YARN, HDFS, Apache Hive , Apache Pig, MapReduce, Apache Spark, Apache Storm, Apache Hbase, Apache Tez, Apache Kafka, Apache Hcatalog, Apache Slider, Apache Solr, Apache Mahout, Apache Accumulo, Workflow Management, Apache Flume, Apache Sqoop, Apache Knox, Apache Ranger, Apache Ambari, Apache Oozie, Apache ZooKeeper

and suggest that future marketing research that requires large-scale data analytics should consider adopting these tools. Our paper introduces Apache-Spark which is the todays most powerful and popular BDA computation engine.
A scalable BDA infrastructure has been orchestrated on the authors' university's private cloud[3]. We adapted a solution suggested by Borisenko & al. (2016) based upon Ansible to automate provisioning, configuration and deployment of this infrastructure and BDA software.
A computer cluster with 6 Virtual Machines (VMs) within the limits of the quota previously attributed by the cloud administrator has been created. For each VM we chose Ubuntu 16.04 LTS - 64 bit among the available operating system images with the biggest available flavor , 6 cores and 21GB RAM. In this way we scaled our computing power to 36 cores, 126 GB RAM.
On the same occasion the computational engine Apache Spark and additional components form the Apache Hadoop ecosystem have been deployed. Among them *Apache Torre* a kernel for the *Jupyter Notebook* platform providing interactive access to Apache Spark and *Ganglia* a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids.
We used HDFS commands in order to put the above mentioned datasets and some subsets of them on the "*data lake*" created by Hadoop by putting together the available disk space of the 6 VMs.

Two BDA case studies

To illustrate the way BDA are applied in the cloud we present two case studies each based upon a different dataset.

The first file we are using can be considered as BigData as it contains 343,766,402 transactions or rows (file size 9.57G) recorded during 78 weeks from 6,326,658 customers of a retail chain. It contains essentially three columns: customer identity followed by transaction date and amount. For confidentiality reasons the source of the data cannot be disclosed. We will call this dataset, the *purchase history dataset*. It is used to predict future customer purchase behavior based upon aggregated customer-level variables like Recency, Frequency and Monetary. Both the aggregation and prediction phases use cloud-computing based analytics.

Another file we use is the Amazon *customer reviews dataset* (curtesy McAuley et al., 2015, file size 58,3G) that contains 82.68 million reviews after deduplication (142.8 million reviews originally) spanning May 1996 - July 2014. The two first reviews in json (JavaScript Object Notation) format are given below:


```
{"reviewerID": "A00000262KYZUE4J55XGL", "asin": "B003UYU16G", "reviewerName": "Steven N Elich", "helpful": [0, 0], "reviewText": "It is and does exactly what the description said it would be and would do. Couldn't be happier with it.", "overall": 5.0, "summary": "Does what it's supposed to do", "unixReviewTime": 1353456000, "reviewTime": "11 21, 2012"}

{"reviewerID": "A000008615DZQRRI946FO", "asin": "B005FYPK9C", "reviewerName": "mj waldon", "helpful": [0, 0], "reviewText": "I was sketchy at first about these but once you wear them for a couple hours they break in they fit good on my board an have little wear from skating in them. They are a little heavy but won't get eaten up as bad by your grip tape like poser dc shoes.", "overall": 5.0, "summary": "great buy", "unixReviewTime": 1357603200, "reviewTime": "01 8, 2013"}
```


We will refer to this dataset as the *web reviews dataset*. It will be used to predict customer ratings from verbatim feedback.

3 Openstack cloud with 53 Tflops composed of 58 hypervisors pour un total de 1068 cors and 14 To of RAM.

Descriptive analytics with the RFM approach

MapReduce can be best understood when analyzing aggregation tasks in the *purchase history dataset* like computing transaction Frequency, Monetary value and Recency per customer. The Map phase implies sorting customer transactions by name, date or amount into queues, one queue for each name. The Reduce phase then performs aggregating operations such as counting the number of transactions in each queue, yielding customer purchase **F**requency, or retaining the maximum date, yielding customer **R**ecency or summing transactions amount, yielding customer **M**onetary amount.

Figure 1 – Aggregating Customer Transactions to obtain RFM Behavioral Variables and lazy evaluation

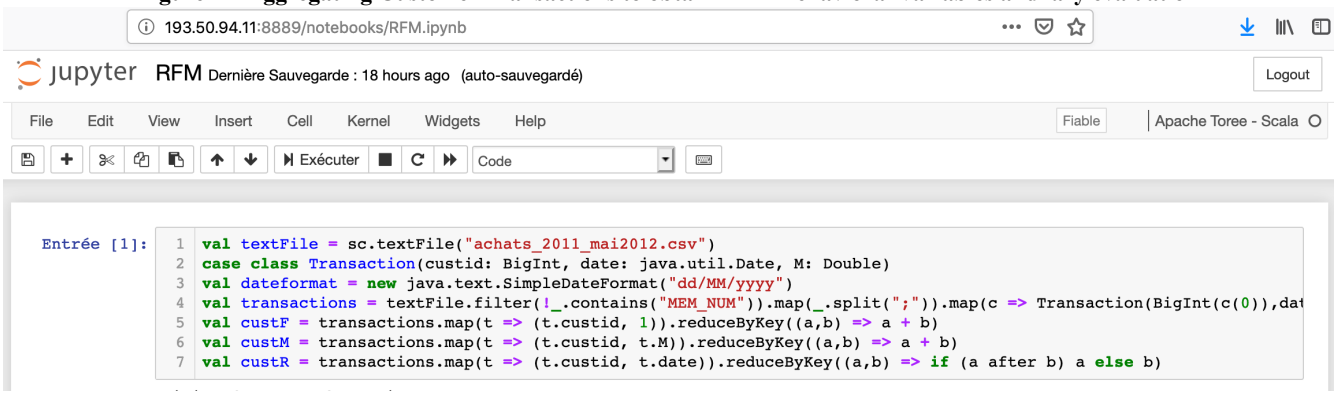


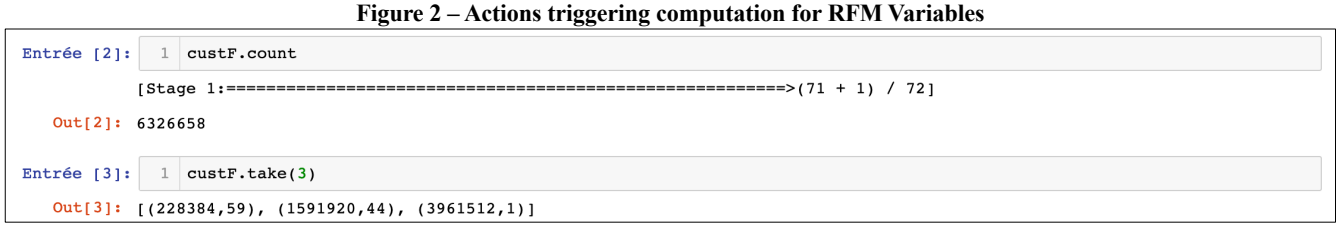
Figure 1 shows the first code entry in Jupyter Notebook which is the most popular way to access BDA in the cloud. The code mainly contains transformations. Transformations are simply ways of specifying different series of data manipulation.

Transformations in line 4 filter out the line containing headers from the *purchase history dataset* (raw text file) and then map several functions that split the rows using the ";" separator and associate appropriate data types to the three resulting fields (custid, date, M) as described in the Transactions class at line 2. Then, in lines 5 to 7, map and reduce transformations are performed in order to aggregate the about 350 million transactions to the about 6 million customers, defining customer Frequency (line 5), Monetary (line 6) and Recency (line 7) variables.

Transformations allow to build up a logical transformation plan. This leads us to a topic called lazy evaluation. Lazy evaluation means that Spark will wait until the very last moment to execute the graph of computation instructions.

To trigger the computation, we run an action. An action instructs Spark to compute a result from a series of transformations.

Figure 2 – Actions triggering computation for RFM Variables



The simplest action is "count" which gives the total number of records in the dataset. Entry 2 in the Notebook shows the count action that will trigger the whole transformation chain and finally, after several computing stages over the cluster, will give the exact number of customers (frequencies) . Entry 3 triggers another action that takes and displays the first three customer frequency records.

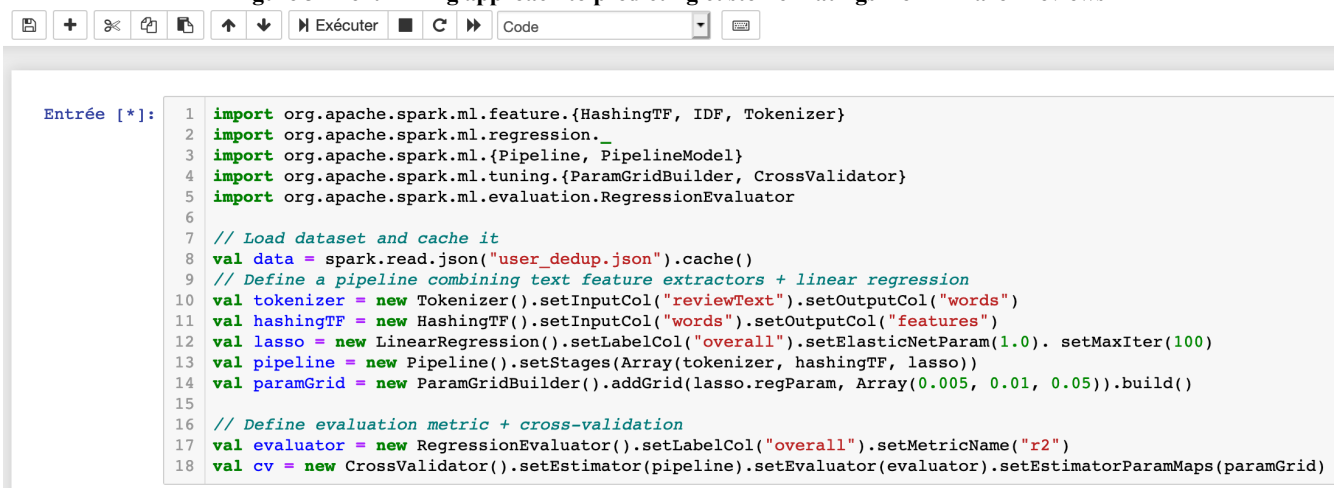
This shows that in Spark, instead of modifying the data immediately when we express some operation, we build up a plan of transformations that we would like to apply to our source data. Spark, by waiting until the last minute to execute the code, will compile this plan from the raw, dataset transformations, to an efficient physical plan that will run as efficiently as possible across the cluster.

RFM variables besides their descriptive, summarizing power can further be used in predictive analytics for targeting customers. For a more detailed discussion on such analytics one can read Calciu & Salerno (2005).

Predictive analytics as a text-mining exercise : Predicting Consumer Ratings from Amazon Reviews

Text mining methods to predict consumer ratings or sentiments from customer "voice" are heavily used in practice and constitute important subject for marketing research. Here we use the reviews dataset mentioned earlier in order to predict consumer ratings. The Lasso regression, applied here, uses a form of Regularized Least Squares that like Ridge regression is suited when the number of independent variables is big, and has the advantage over the latter to automatically select more relevant features and discard the others. This is another kind of BDA application in the cloud that is deemed to predict Amazon consumer ratings from reviews' unigrams by selecting a reduced number of more relevant features. It also illustrates the flexibility of Spark compared to the classical MapReduce approach.

Figure 3 - Text mining approach to predicting customer ratings from Amazon reviews



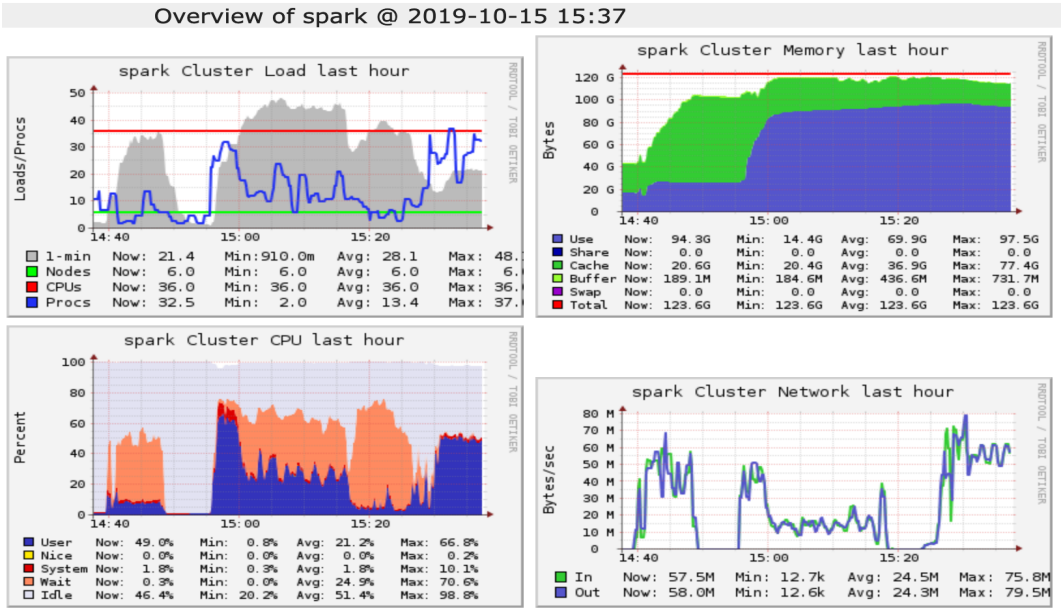
As before, we start by transforming the data into the correct format and create some valuable features. An important aspect here is the way a transformations pipeline is organized. A pipeline allows Advanced Analytics and Machine Learning, It can set up a data-flow of the relevant transformations, ending with an estimator. The tokenizer extracts and counts the individual words (line 11).A bag of words (BOW) representation of the review text is constructed. HashingTF (line 12) is a transformer which takes BOWs and converts them into fixed-length feature vectors. Each distinct word (token, unigram) defines a feature (independent variable) of each of the reviews. The final item this pipeline is the estimator of the lasso regression (line 13) that adds the "overall" customer rating as the dependent (explicative) variable. The next step will be evaluating the performance of this pipeline. Spark does this by setting up a parameter grid of all the combinations of the parameters that you specify (line 14). Here the evaluator's metric is the R2 (line 17). Finally the cross-validation procedure combines the pipeline as an estimator, the evaluator and the parameter grid (line 18)

Again, as in the previous exercise, up to this point the listing contains no actions. According to *lazy*

evaluation the engine will wait until the very last moment to execute the graph of computation instructions. Real calculations occur later when cross-validation is launched to fit over the training dataset. Once the model has been estimated over what should be a reviews training-set its performance can be evaluated over a test-set and R2 score and customer rating predictions can be given by using the calibrated model with customer reviews from the test set.

In order to show the potential computational burden we used the whole reviews dataset as a training-set as it is a good example of big data. We then analyzed (see figure 4) using *Ganglia* monitoring system the difference in terms of workload for the cluster processes mobilizing the available cores (CPUs), for the joint memory, for CPU resources mobilized by user, system or wait processes.

Figure 4



It appears clearly that the transformation or lazy evaluation phase that lasted about 10 minutes consumed much less resources than the action phase where the first model fitting phase lasted about 20 minutes and continued almost endlessly during the other phases of the cross-validation process.

Conclusion

While trying to give an overall picture of the role of BD and BDA in transforming the marketing function of a company, our approach focuses on marketing scientists or analysts as a key human resource and the need to operate, as regards their skills, a both technical and cultural paradigm shift from “data” to “big-data”. BDA solutions form rather complex software ecosystems and normally operate on clusters of computers in either “bare metal” or virtualized form. Marketing scientists have a long tradition in developing and applying models and analytics to market and consumer data. Their valuable skills should constitute a powerful resource to build BDA capability. By paraphrasing Clemenceau’s famous quote "War is too important a matter to be left to the military", we suggest that "Big Data are too important to be left to computer or data scientists". Marketing scientists should play an active role and contribute to the new approaches that lead to groundbreaking changes in data science. By trying to demystify BigData approaches and point out key technological aspects our paper invites marketing scientists to pay more attention to technologic evolutions, to become more involved in developing specific analytics.

We introduce an evolutionary view of the technical resources for marketing analytics in order to explain the shift from toolbox-centered to service-centered analytics and evaluate the impact on human and organizational resources. We then decline the several Service Oriented Approaches that have dominated service-centered analytics and decision support in order to distinguish analytics in the Cloud as the strongly preferred pattern for dealing with Big Data. We insist on some new competences human resources, that is marketing scientist, need to acquire in order to deal with BDA in the Cloud. In contrast with analytic solutions that don't need more than one computer to be implemented, BDA that require clusters of computers come with additional orchestration, provisioning and deployment challenges. Apart virtualization as a technological foundation, upon which cloud computing relies, we show some benefits, facilities and simplifications brought over by containerization for BDA.

We finally suggest marketing analysts, that include academic model builders and data scientists, need to have a grasp of service oriented and cloud technologies as this is increasingly the way, and for Big Data it is probably the only way, analytic solutions can be brought to the market.

References

Anderson, C. (2015). *Creating a data-driven organization*. O'Reilly Media, Inc

Berners-Lee, T. (1989) *Information Management: A Proposal*, CERN, Online Available at: https://www.w3.org/History/1989/proposal .html, [Accessed on 24th April 2019]

Berners-Lee, T., & Shadbolt, N. (2011). There's gold to be mined from all our data. The Times, London 1:1-2. Online Available at: http://www.thetimes.co.uk/tto/opinion/ columnists/article3272618.ece [Accessed on 24th April 2019].

Borisenko, O., Pastukhov R., Kuznetsov S. (2016), Deploying Apache Spark virtual clusters in cloud environments using orchestration technologies, *Proceedings of ISP RAS*, 28, 6,111–120

Burns B., Grant B., Oppenheimer D., Brewer E. & Wilkes B. (2016), Borg, Omega and Kubernetes: Lessons learned from three container management systems over a decade. *Communications of the ACM*, May, 59, 5, 50-57.

Brown, B., Chui, M., & Manyika, J. (2011), Are you ready for the era of Big Data? *TheMcKinsey Quarterly*, 4, 24–35.

Calciu M. & Salerno F. (2005), Modélisation prédictive de l'incidence et des montants d'achat en marketing direct: une comparaison a partir de variables RFM., *XXIème Congrès International de l'Association Française du Marketing*, Nancy.

Calciu M., Salerno F., Meyer-Waarden L., Willart S. (2013), Decision support for valuing customers as RESTful web services, *42nd EMAC Conference*, Istanbul, Turkey, June 4-6

Calciu M, Moulins J-L. & Salerno F. (2016) Big data and open-source computation solutions, opportunities and challenges for marketing scientists. Applications to customer base predictive modeling using RFM variables., *15th International Marketing Trends Conference*, Venice, January 21–23

Calciu M, Moulins J-L. & Salerno F. (2017), Small is beautiful but scalable is better. Scalable Marketing Decision Support Systems for BigData calculations in CRM, *16th International Marketing Trends Conference*, Madrid, January, 26-28

Calciu M, Moulins J-L. & Salerno F. (2019), Service Oriented Marketing Decision Support Systems(SOMDSS) for Big Data in the Cloud.Some orchestration, provisioning and deployment challenges for marketing scientists, *18th International Marketing Trends Conference*, Venice, January, 18-19

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data

management perspective. *Frontiers of Computer Science*, 7(2), 157–164.

Dean J. & Ghemawat S. (2004), MapReduce: Simplified Data Processing on Large Clusters, *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, December.

Hamdaqa, M., Tahvildari L. (2012) Cloud Computing Uncovered: A Research Landscape, *Advances in Computers*, 86, 41, 43-84, http://dx.doi.org/10.1016/B978-0-12-396535-6.00002-8

Kumar, V. (2018), Transformative Marketing: The Next 20 Years. *Journal of Marketing*. 82, 4, 1-12.

Little, J.D.C. (1979), Decision Support Systems for Marketing Managers, *Journal of Marketing*, 43, 3, 9-27.

Liu, X. Singh, P.V. & Srinivasan, K. (2016), A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing, *Marketing Science,* 35, 3 (May/Jun), 363-388.

Mayer-Schönberger, V., & Cukier, K. (2013), *Big data: A revolution that will transform how we live, work, and think*. Boston, MA: Eamon Dolan/Houghton Mifflin Harcour

McAuley, J., Pandey, R. & Leskovec J (2015) Inferring networks of substitutable and complementary products, *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Odersky, M., Spoon L., Venners B. (2011) P*rogramming in Scala, 2nd Edition: A comprehensive step-by-step guide*, 2 edition (January 4, 2011), Artima Inc

Schumacher E.F. (1973, 1999), *Small Is Beautiful: A Study of Economics As If People Mattered*, Blond & Briggs

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V. (2017), Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research*, 70, 263-286.

Vinoski, S. (2002), Putting the "Web" into Web services: Interaction models, part 2. *IEEE Internet Computing*, 6,4,90–92.

White, C. (2012), MapReduce and the Data Scientist Colin White, *BI Research,* January.

Wilkinson, D.J. (2017), *Statistical Computing with Scala: A functional approach to data science*, https://github.com/darrenjw/scala-course

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: architecture and challenges. *IEEE Network*, 28(4), 5–13.