People's democratic republic of algeria
Ministry of Higher Education and Scientific Research
University of Mohamed Boudiaf - M'sila-

The international colloquial:
The degital transformation of enterprises & predective models on big data
12 & 13 November 2017

Title of the Article:

Data scientist: new necessity for the 21st century –Case study of Mosaic Data Science-

Dr. ABDAOUI Nawel* - University of Biskra-

Abstract:

This study aimed to precise the identification of Data scientist as an urgent necessity for the 21st century. In order to answer the problematic study we chose Mosaic Data Science as a case study. The study concluded at the last a set of results, the most important are: Data scientist are people who can mix between differente skills and Mosaic Data Science is considered as a big Data Scientist which helps organizations gain insight from their data and use those insights to change the way a decision is made, through analysing data and helping them to distinguish between useful data and the not useful.

The added value of this research is that it highlighted a spot on an important subject which we need to know more about by treating it through differente sides.

Keywords: Big Data, Data Science, Data Sceintist.

الملخص:

هدفت هذه الدراسة إلى التعرف على مفهوم "عالم المعطيات" كونه أضحى ضرورة ملحة في القرن الواحد والعشرين. وللإجابة على إشكالية الدراسة اخترنا مؤسسة "Mosaic Data Science" كدراسة للحالة

* **nawel_abdaoui@yahoo.fr. phone n° : 0675646730 . field : Management economics.**

باعتبارها مركزا كبيرا منتجا لعلماء المعطيات. توصلت الدراسة في الأخير إلى مجموعة من النتائج أهمها: أن عالم المعطيات هو شخص يمكنه المزج بين العديد من المهارات، كما أن مؤسسة" Mosaic Data Science" تعتبر مؤسسة مهمة تساعد بقية للمؤسسات في تحسين عملية اتخاذ القرارات من خلال تحليل البيانات ومساعدة المنظمات على استغلال البيانات عن طريق تمييز البيانات المفيدة من غيرها.

أما القيمة المضافة للبحث فتتمثل في تسليط الضوء على موضوع مهم من المواضيع الراهنة والتي لا تزال تحتاج منا التعرف اكثر على مضامينها.

الكلمات المفتاحية: *المعطيات الكبيرة، علم المعطيات، عالم المعطيات.*

I. Introduction:

Nowadays companies are starting to realize the importance of using more data in order to support decision for their strategies. It was said and proved through study cases that "More data usually beats better algorithms". With this statement companies started to realize that they can chose to invest more in processing larger sets of data rather than investing in expensive algorithms. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations.

. With more and more companies using big data, the demand for data analytic specialists,—sometimes called data scientists, who know how to manage the tsunami of information, spot patterns within it and draw conclusions and insights—is nearing a frenzy.

I.1. Research Problematic

What is the importance of Data Scientists? And which role does Mosaic Data Science play in analysing Big Data?

To address this problem and take action on the aspects that make up the themes of this subject, we have divided it in the form of sub-questions, which are:

- What is Big data?
- What is Data Science?
- Who is Data scientist? And what is their role?
- What is Mosaic Data Science and how does it work?

I.2. Importance of the study

nowadays the Internet represents a big space where great amounts of information are added every day. The IBM Big Data Flood Infographic shows that 2.7 Zettabytes of data exist in the digital universe today. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) eguals 10^{21} bytes, meaning 10^{12} GB.

this study came to illustrate the growth of the importance of data scientits in analysing the huge number of data got by the companies. By highlighting a spot light on "Mosaic Data Science" as a big Data Scientist.

II. Theorical framework

II.1. Big Data:

"You can't manage what you don't measure", There's much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker, and it explains why the recent explosion of digital data is so important. Simply put, because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance.[1]

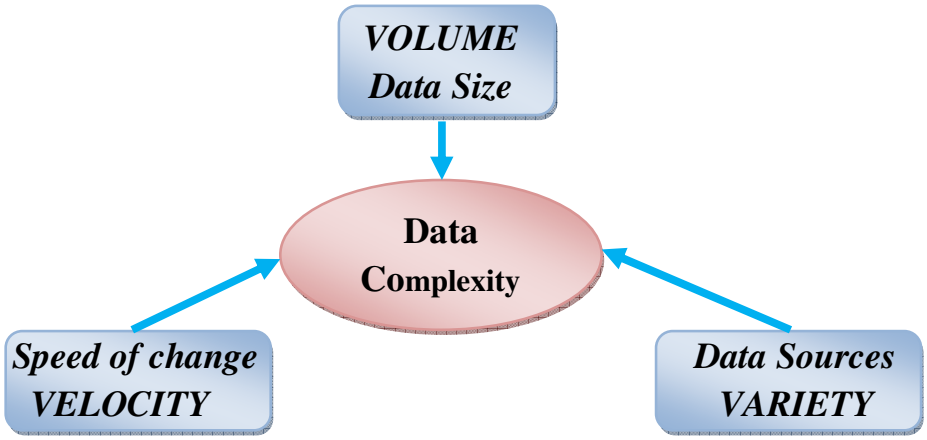
The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data. A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been *Perspectives on Big Data and Big Data Analytics* comprised in publications in 2008. Nowadays the Big Data concept is treated from different points of view covering its implications in many fields.[2]

II.1.1. Defenition of Big Data:

As one of the most "hyped" terms in the market today, there is no consensus as to how to define big data. The term is often used synonymously with related concept such as *Business Intelligence* (BI) and *data mining*. It is true that all three terms is about analyzing data and in many cases advanced analytics . But big data concept is different from the two others when data volumes, number of transactions and the number of data sources are so big and complex that they require special methods and technologies in order to draw insight out of data (for instance, traditional data warehouse solutions may fall short when dealing with big data). This also forms the

basic for the most used definition of big data, the three V: *Volume*, *Velocity* and *Variety (as shown in figure1)*

Firure1: the three V of Big Data



Source: Xiaomeng Su, **Introduction to Big Data**, Norwegian University of science and Technology, P2, Available on the web site: https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf visited on 10/09/2017 at 10.18 am)

As shown in the previous figure the definition of Big Data is based on three principals V_s which are:[3]

- Volume: Large amounts of data, from datasets with sizes of terabytes to zettabyte. As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the internet every second than were stored in the entire internet just 20 years ago. This gives companies an opportunity to work with many petabyes of data in a single data set—and not just from the internet. For instance, it is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions. A petabyte is one quadrillion bytes, or the equivalent of about 20 million filing cabinets' worth of text. An exabyte is 1,000 times that amount, or one billion gigabytes.
- Velocity: Large amounts of data from transactions with high refresh rate resulting in data streams coming at great speed and the time to act on the basis of these data streams will often be very short . There is a shift from batch processing to real time streaming. For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors.

- Variety: Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications , structured data as database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two. Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Many of the most important sources of big data are relatively new. The huge amounts of information from social networks, for example, are only as old as the networks themselves; Facebook was launched in 2004, Twitter in 2006. The same holds for smartphones and the other mobile devices that now provide enormous streams of data tied to people, activities, and locations. Because these devices are ubiquitous, it's easy to forget that the iPhone was unveiled only five years ago, and the iPad in 2010. Thus the structured databases that stored most corporate information until recently are ill suited to storing and processing big data. At the same time, the steadily declining costs of all the elements of computing-storage, memory, processing, bandwidth, and so on -mean that previously expensive data-intensive approaches are quickly becoming economical.

II.1.2 The importance of Big Data:

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth. Big Data can be used effectively in the following areas:[4]

- ✓ In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- ✓ In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- ✓ In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people;
- ✓ In the detection of fraud in the online transactions for any industry;

- ✓ In risk assessment by analyzing information from the transactions on the financial market.

II.2. data science:

Data science is the application of scientific experimentation (hypothesis testing, model generation, statistical analysis) in problemagnostic ways.[5]

The Data science is the intersection of Hacking skills, math & statistics knowledge and substantive expertise, so those are the pillars of data science: computing, statistics, mathematics and quantitative disciplines combined to analyze data for better decision making as shown in figure2:

Figure2: Data science

Source: Alejandro Correa Bahnsen, **Medern Data Science**, June 2016, P18 available on the web site: https://fr.slideshare.net/albahnsen/modern-data-science, visited on 10/09/2017 at 13.11pm.

II.3. Data Scientist:

II.3.1- Who is Data Scientist?

The people that does the big data analytic job are called data scientist nowadays[6]. And he is someone who extract insights from messy data.[7]

"Data scientists are inquisitive: exploring, asking questions, doing "what if" analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommenda"[8]

II.3.2. Data scientist's tasks:

Here are some tasks of the Data Scientist:[9]

- ❖ Interface with analytics, product management, and operations teams;

- ❖ Perform large-scale data analysisand develop effective statistical modelsfor segmentation, classification, optimization, time series, etc;
- ❖ Design and implement reporting dashboards that track key business metrics and provide actionable insights;
- ❖ Identify actionable insights, suggest recommendations and influence the direction of the business by effectively communicating results to cross functional groups;
- ❖ Work closely with Product or Engineering & Operations teams to proactivelycreate rule and manage decisions;
- ❖ Suggest improvements in the tools and techniques to help scale the team;
- ❖ Apply data-mining, machine learning and/or graph analysis techniques for a variety of modeling and relevance problems involving users, their relationships, their tweets and their interests;
- ❖ Design and evaluate novel approaches for handling high-volume real-time data streams;
- ❖ Code using primarily Java, Scala, and scripting languages such as Python or Ruby;
- ❖ Conduct design and code reviews;
- ❖ Work with large unstructured and structured data sets;
- ❖ Utilize data science and quantitative methodologies to help shape clinical care and long-term planning.

II.3.3 difference between Data Scientist and data analyst:

Data Analyst focus on the movement and interpretation of data, typically focus on the past and present. Where Data Scientist focus on summarizing data and to provide forecasting based on pattern identified from past and current data. define and differentiate between Data Scientist and Data Analyst as describe in table 1.

Table1: Data Scientist vs Data analyst

Data Scientist	Data Analyst
Building statistical models that make decisions based on data	Writing custom queries to answer complex business questions
Conductiong causality experiments that attemps to attribute the root cause of an observed phenomenon	Conceiving and implementing new metrics on capturing previsouly poorly understood parts of the business/ product
Identifying new products or feature that come from unlocking the value of data; being a thought leader on the value of data. A good example of that is the product recommendation feature that Amazon first made available to a mass audience.	Addressing data quality issues, such as data gaps or biases in data acquisition. Working with the rest of e,gineering to instrument incremental new data acquisition

Source:Wardah.Z Abidin & Nur.A Ismail, **Data Scientist Skills**, IOSR Journal of Mobile Computing & Application, Vol.3, N°3, July-August 2016, P53.

III. Case study of Mosaic Data Science

III.1. What is Mosaic Data Science?

Mosaic Data Science is a limited liability company founded in 2004. With 51 200 employees. Its headquarters are located near Washington D.C. and they have consultants across the United States. The entreprise logo is formed of a big red "M" and three arrows (see appendix 1)

Mosaic is considered as a big Data Scientist, it is specializes in helping organizations gain insight from their data, and use those insights to change the way a decision is made. they do this through a blend of statistical analysis, data visualization, machine learning, optimization, and data engineering techniques. Mosaic employs data science consultants who enjoy doing the hard mathematical work behind the scenes and using these insights to change the way decisions are made.

Mosaic Data Science is a dynamic, growing data science consultancy solving some of the most complex and interesting problems in industry. At Mosaic you'll contribute to data science projects requiring a variety of analytical and business skills in many verticals, especially logistics, telecommunications and media, manufacturing, and medicine. You'll work with extremely talented, self-motivated people who hold themselves to very high standards of professionalism and quality, with very little guidance from management. You'll probably telecommute, and you'll probably travel part-time to customer sites, especially at the start of engagements.

You can expect your responsibilities to grow as quickly as you're willing and able to receive them. You'll enjoy a fantastic benefits package that includes profit sharing and generous personal leave, as well as first-rate healthcare and retirement benefits. Most folks who join Mosaic stay here for a long time.

It has a decade of data science consulting experience. As one of the premier analytics consulting firms, it is steeped in delivering real world solutions, focusing their efforts on solving business challenges and delivering results, not just doing a fancy scientific study. Many of thier customers tell them that they have spent millions on collecting and storing data, but have not gotten the return-on-investment they thought they would due to a lack of expertise running analytics and data science projects.

III.2. Mosaic Data Science's team

The enterprise has ,as we cited above, 51200 employees. Including a very talented leadership team of Data Scientist which are sited in the table bellow:

Table2: some of leadership team of Mosaic Data Science

Name of leader	Post	Experience
Chris Brinton	*President*	More than two decades of R&D experience in aviation logistics. Founded Mosaic ATM in 2004. Led numerous projects improving dramatically the efficiency and safety of the National Airspace System as a consultant to NASA and the FAA. Formerly Chief Operating Officer and VP of R&D at Metron Aviation, and VP of Engineering at Wyndemere. Numerous scientific publications…
Stephen Atkins	*Executive Vice President*	Two decades of air traffic management R&D. Industry leader in airport surface traffic management and metroplex operations. Led numerous and extensive field surveys of airport operations, air traffic control operations, air carrier operations, and control facilities…
Mark Thomas	*Vice President, Strategy & Marketing*	Over two decades in strategy, marketing, business intelligence consulting and business development experience in multiple industries. Previously developed and managed multiple government client relationships including Air Force, Army, US Census, SEC and others. Masters of Science in Integrated Marketing Communications from Northwestern University…
Chris Provan	*Chief Data Scientist*	Nearly a decade of experience leading projects that apply data science and mathematical modeling methods to complex strategic planning and operations management optimization problems, producing collaborative decision-making decision-support tools used by facility managers and operators…

Mike Shumpert	*Managing Director of Data Science*	Over two decades of product/service development across diverse industries. Solid record of applying data science and technology to produce double-digit revenue growth and cost savings in both small and large companies. Skilled at taking solutions from concept to reality and delivering tangible, measurable results. Lived and worked in the US, UK, France, and Australia…
George Hunter	*Principal Data Scientist*	Nearly thirty years of experience in engineering and the life sciences. Helped to build a Silicon Valley consulting firm and has extensive experience in leading and executing a wide variety of data modeling, analysis and software projects. 60 conference and journal papers…
Lara Shisler	*Principal Data Scientist and Traffic-Flow Manager*	Nearly two decades of software engineering and operations research experience. Leads projects that model traffic flows in the presence of weather, capacity, and demand uncertainties. Formerly VP of Programs and Special Projects, and VP for Deployment and Operations, at Metron Aviation.

Source: http://www.mosaicdatascience.com/about/leadership-team

III.3. Mosaic Data Science's costumer and tools

Mosaic has a wide base of international customers like *(NASA, PEPSI, Us Bank, Ferrari, Exonmobil, Microsoft,…)* and it is always carefal for its relation with them. The relation often starts with identifying and analyzing a data science opportunity, working with customer stakeholders to develop a business case for pursuing it, and presenting the business case to key decision makers. This can be a Proof of Concept, Rent a Data Scientist, or Analytics Assessment engagement.

the consultants bring cutting edge methods from statistics, operations research, machine learning, decision science, cognitive science, software engineering, business intelligence, human factors, and organizational behavior to produce the most appropriate model or suite of models to bring value, not just a pretty bar chart. All of that by using a lot of types of tools like: (*SPSS,SAS, SAP, Microsoft Azure, Java,…)*

III.4. Mosaic Data Science's tasks:

"A lot of pain and confusion in the analytics marketplace is around focusing on insights rather than optimization," says Chris Brinton, President and Principal Analyst, Mosaic Data Science. Many data science consultancies of today merely focus on prediction and classification of data. But businesses need to go beyond just insights. Understanding this, Mosaic Data Science has developed expertise around solving complex optimization problems that range from improving operational efficiencies to safety of air traffic, and deeper into verticals such as medicine, finance, media and telecommunications.

Mosaic's range of services includes assessments, projects, and coaching. Mosaic's team of experts first assesses the data science capability, project scope and planning, or data-related opportunities. In the subsequent phases, they execute or lead several flavors of data-centric projects which include analytics, Big Data, Business Intelligence, decision analysis, and risk analysis. Apart from that, the company offers coaching to organizations that are trying to build and develop a solid data science function. In a nutshell, Mosaic's offerings find traction on every business problem involves optimization in the presence of multiple kinds of uncertainty.

Mosaic's transparency and reliability coupled with their practice of going out of their way to deliver extraordinary value, while maintaining very positive and professional relationships with customers, contributes to their amazing customer retention rate. The company also frequently writes white papers addressing specific business and technical questions, at their prospects' request.

Led by Brinton's decades of data science experience, Mosaic has been solving complex and large-scale optimization problems for the likes of NASA, the FAA, FedEx, Boeing, Northrop Grumman, UPS and Lockheed Martin for a decade now. "Not many data science consultancies can say they do rocket science for rocket scientists or logistics for the company that turned logistics into a brand. We were data science before data science was cool.[10]

III.5. How does mosaic Data Science work? (case study)

To know little bit of how does Mosaic Data Science work, we're going to introduce a case study of "Daily Store-Order Prediction":[11]

1- Background

A prominent soft drink manufacturer and distributor required assistance from a top analytics firm. The manufacturer owns and operates more than fifty brands, controlling six of the top ten non-cola soft drinks. Additionally, thirteen of its fourteen leading brands are number one or two in their flavor categories.

The company's business intelligence and analytics team needed an order optimization capability that would support their account managers both in providing better service to customer accounts and in driving increased revenue from existing accounts. This desired capability would provide recommendations for upsell, cross-sell, and promotional orders to increase total product sales for each account. It would also automate the order process for mainline SKUs(Stock keeping units – unique products/packaging configurations manufactured by the company), enabling account managers to focus on value added activities during store visits.

Predicting store demand accurately is a major industry challenge. Mosaic Data Science, a consultancy specializing in advanced analytics, is uniquely positioned to tackle and create applicable solutions. Mosaic's data scientists came at the problem in two phases: Phase One explored relevant issues, tested hypotheses, transformed and prepared the data for analysis, engaged in the business processes and technology, and experimented with various predictive analysis approaches. Using the information gathered in Phase One, Phase Two involved implementing and building the predictive model shown to produce the most superior results.

1- Phase One: Analysis:

In collaboration with the soft drink company, Mosaic data scientists developed a short-horizon predictive model that forecasts the case-quantity orders at a SKU-store level on the day of an account manager's store visit. The model in Phase One became a proof of concept (PoC), enabling commitment and sponsorship from company executives. Corporate leaders need evidence that their data can be optimized in useful ways, and Mosaic's model clearly provided that evidence. Predicted order quantities were compared against actual orders entered by account managers to demonstrate the precision of the model. Accuracy analyses were then distributed to business and IT stakeholders. Mosaic's effective PoC model was capable of generating a predicted-order quantity for a subset of SKU-store combinations covering the SKU-stores with sufficient order history.

The model incorporated historical order data, store and SKU characteristics, account category (type of store), and other properties determined to be of use to the customer. It was trained on the historical data and applied to new data points representing an upcoming order opportunity. The model's output was a daily prediction of the case-order quantities for each individual SKU at each store that an account manager would visit on that day.

Mosaic integrated and performed exploratory predictive analyses on three years of historical order data, store and SKU characteristics, account manager schedules, promotional data, and other sources. Mosaic's data scientists built an initial integration infrastructure to generate analyzable data sets from the different sources. Mosaic analyzed samples of integrated data to identify basic trends, correlations, and clusters of similarly performing customers/stores/SKUs.

Mosaic then focused on feature engineering and model development for the PoC predictive model. Mosaic generated modeling data sets with elements built from input data sources and evaluated and competed multiple models using cross-validation. Time series models, including ARIMA and ARIMAX, random forest regression, and

simple linear regression were all evaluated. Our data scientists also performed model and feature selection and consulted extensively with the client team, ensuring that the model was performing optimally and effectively addressing the customer's requirements.

2- **Phase Two: Results**

After testing different algorithmic approaches, Mosaic's data scientists determined that three different models were required to appropriately account for the different order patterns observed in the data. A unique dynamic selection model was implemented to select the optimally performing of the three models based recent model performance for each individual SKU-store combination.

In the second phase of work, Mosaic developed an automated tool outputting predictions and performance metrics based on dates of training and testing datasets. This tool runs behind the predictive model developed in Phase One. The forecasts underlying the automation are at the day-location-SKU level, but are aggregated to groups of SKUs, groups of locations, and periods of time to permit improved decision making.

The following sections examine the different models used:

❖ *Day-of-Week Model:*

This model captures the case-quantity trend by day of week (DOW). For case quantities that are very regular (i.e., consistently the same or similar values) or follow a consistent weekly pattern, this model describes behavior well. A significant proportion of store-SKU combinations follow this behavior.

❖ *Store-SKU Regression Model*

This model captures the behavior at the store-SKU level in a regression model using various features created from the historical case quantities. A separate regression model is implemented for each store-SKU combination, meaning that a dataset with 5000 store-SKU combinations will result in 5000 separate regression models.

The features included in the model are the case-quantity average per day of week, the case-quantity average for the previous 1 and 2 weeks, and indicator variables for each day of the week.

❖ *Normalized Store Regression Model*

This model captures the behavior of case quantities at the store level that are common across SKUs at that store. There is a separate regression model for each store that is applied to each SKU at that store. Input and output values for each SKU are scaled based on overall order volume for that SKU to account for the wide variability

in demand across SKUs at a single store. The features included are the same as those in the Store-SKU regression model.

❖ *Dynamic Selection*

This architecture combines the predictions of the above three models and selects the one performing best based on historical model performance. Because the three models complement each other in characterizing certain behavior of the case quantities, this model generally provides the most intelligent information and can adjust to pattern changes over time such as when a SKU moves from mainline to secondary status in a store or when a new account manager is assigned to a store. The model looks at the average forecast error for the previous 10 orders for each store-SKU, selecting the model that minimizes that error as the source of the next prediction for that store-SKU.

Summary performance metrics such as the MAEN (Mean Absolute Error, Normalized) and the MedianAEN (Median Absolute Error, Normalized) allow for the errors for different store-SKUs can be compared to one another on the same scale. Summary statistics for the errors are included at the daily and weekly level.

3- **Outcome**

Mosaic Data Science's prediction model and automated insights more accurately predict correct store orders, providing efficient and intelligent results. They optimize account managers' time in the store, allowing them to direct their focus to upselling and promotions, not on matching up orders. By closely collaborating with our clients, Mosaic delivers valuable insight into business's most pressing problems.

III.6.future goals of Mosaic Data Science

Over the next three years, Mosaic plans to grow their commercial data science consulting practice beyond transportation and logistics, where they already have a strong foothold. The company is also intent on continuing to maintain a high quality of living for their employees, rather than just focusing on maximizing revenue, profit, or growth. "Our employees respond to that by maintaining extraordinarily high standards in our customer relationships, and we intend to zealously preserve those values while growing our commercial consulting practice methodically"[12].

Conclusion

Through the theoretical and the empirical study, we determined number of results in order to ensure the importance of Data scientist as an attractive job in the 21st century, and we present these results in the following elements:

- Every day, 2.5 quintillion bytes of data are created — so much that 90% of the data in the world today has been created in the last two years alone;
- This data comes from various sources such as: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies and genomics. This data is referred to as big data;
- Analysing Big Data need special tools and methods but more than that, it needs talented people;
- The people who are specelized in collecting and analysing the hudge number of data called Data Scientist;
- Data scientist are people who can mix between differente skills;
- Mosaic Data Science is a company which helps organizations to ameliorate decision making through analysing data and helping them to distinguish between useful data and the not useful;

References:

1- Andrew McAfee & Erik Brynjolfsson, **Big data: the management revolution**, Harvard bussiness Review, UK, October 2012 Issus.
2- Elena Geanina ULARY &al, **Perspective on Big Data and Big Data Analytics**, Database System Journal, Vol.3, N°.4, 2012, P4.
3- Andrew McAfee & Erik Brynjolfsson, Op.Cit.
4- Elena Geanina ULARY &al, Op.Cit, P4
5- William Cukierski & Ben Hamner, **Just the basics: Core data science skills**, O'Reilly Strata conference, SANTA CLARA, CA, February 2014, P18
6- Xiaomeng Su, **Introduction to Big Data**, Norwegian University of science and Technology, P2, Available on the web site: https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf visited on 10/09/2017 at 10.18 am.
7- Joel Grus, **Data science from scratch**, O'Reilly, USA, April 2015, P3.
8- Anjul Bhambhri, Disponible on the web site: http://www-01.ibm.com/software/data/infosphere/data-scientist/ , visited on 15/09/2017 at 14.40pm.
9- Richard.F, **Data Scientist: A career for 2015 and Beyond**, Mitre corporation, USA, March 2015, P.P 5.6

10- Chris Brinton, **Mosaic Data Science: Saving Companies Millions through Data Science,** available on the web site: https://bigdata.cioreview.com/vendor/2014/mosaic_data_science, visited in 10/09/2017 at 19.49pm

11- http://www.mosaicdatascience.com/resources/case-studies/daily-store-order-prediction-case-study/

12- Chris Brinton, Op.Cit.

Appendix:

Appendix1: Mosaic Data Science logo

Source: http://www.mosaicdatascience.com