

TECHNOLOGY TRANSFER PRESENTS

SHAKU ATRE

10 SKILLS

TO GET THE BEST

OF BIG DATA

APRIL 4-5, 2016
RESIDENZA DI RIPETTA - VIA DI RIPETTA, 231
ROME (ITALY)



info@technologytransfer.it
www.technologytransfer.it

ABOUT THIS SEMINAR

Big Data technology is new to most organizations and so is awareness of the skills needed to get the best out of Big Data. To “have” these skills overnight is wishful thinking. As a result, in most organizations a large percentage of Big Data skills need to be either learned or recruited, or a little bit of both.

Big volumes of data beg for analysis in order to glean correlations and inferences and to prove or disprove hypotheses. These methods point straight to Data Science. In the past, Data Science was practiced only in the academic world. Now, in order to be competitive in the marketplace, every business is expected to possess these academic skills. With one big difference - in academia, results typically did not need to be obtained very quickly, if the problems and the data were very complex. They could take their dear time - something businesses cannot afford to do; **Time to Results is of paramount importance for businesses to succeed.** That said, besides volume, the bigger problem is speed - meaning the velocity with which the data arrives, with which it is supposed to be worked on, and with which the insights are supposed to be provided to the decision makers. It is not only that the standard of “how much data” has changed but also “how soon” has changed dramatically as well.

Data goes mainly through four phases; the major problems with Big Data occur in Phases 2, 3, and 4:

- **Phase 1:** Data is generated by transactions (e.g., billing and reservations), interactions (e.g., shopping online), and observations (e.g., measuring carbon monoxide levels in different sections of a plane)
- **Phase 2:** Data is received by various recipients – Are the receiving systems fast enough to handle the output of the data-generating systems? Is it like multiple lanes of cars trying to get into one tunnel?
- **Phase 3:** Data is stored and processed - Is the storage capacity big enough and is the processing fast enough? (How many tunnels should there be? The number of cars on the road is increasing at a dizzying speed)
- **Phase 4:** Insights are created - has to be done fast enough to benefit the business’s bottom line. (Can instantaneous rerouting of the cars be done to avoid deadlock, or, even worse, a deadly embrace?)

Data Science’s main building blocks are mathematics and statistical analysis, skills which what you will learn

Analysts of Big Data should have the following strengths:

- Familiarity with newer statistical languages like **R**
- Understanding and use of **analytics modeling techniques**
- Outstanding familiarity with the **data to be analyzed**
- **Risk-taking** mentality to experiment with data (it is always a good idea to back up the data before it disappears in front of your eyes because you were trying something unusual with the data - and unusual is exactly what you are supposed to do)

Technical skills needed are, among others:

- Very good understanding and experience with **Open Source** Software
- Data architecting of databases with **terabytes** of data and growing every minute
- Experience managing software frameworks like **Hadoop**; expertise in databases like **noSQL**, **Cassandra**, and **HBase**
- Expertise with analytics programming languages and facilities such as very important languages **R** or **Pig**
- Ability to manage hardware with hundreds or **thousands of “small” CPUs**, for multiple terabytes of data

And, **soft skills** having not much to do with Big Data are needed in many organizations:

- Understanding of the **“ins and outs” of the business**
- Understanding of the **“bottom line”** of the business
- Ability to discern which analytics will answer **the bottom-line questions**
- **Communications skills** to explain the analytics results
- **Understanding** not only transactions (as we have been doing all along) but also interactions (such as people buying products on the web) and observations (such as machines or sensors measuring and reporting about happenings or not-happenings)

WHO SHOULD ATTEND?

This seminar has value for **various levels of audience:** CEOs, CFOs, CIOs, CTOs, middle Management, Project Managers, systems Analysts, Developers, system Programmers, database Administrators, Users at various levels, as well as individual Performers. During this two-day seminar, we will discuss the ten “must” skills needed for Big Data, complemented by small-group workshops and individual Q&A periods.

1. Open Source: Apache Hadoop

A Big Data processing software has to be able to disperse the data in “chunks” to a number of processors and reassemble it without losing anything in the process! The Hadoop platform is powerful, but it is a beast which requires tender loving care and appropriate feeding by skilled technicians because of its distributed storage and processing architecture. Skills with Hadoop stack-such as HDFS, MapReduce, Flume, Oozie, Hive, Pig, HBase, and YARN - are in high demand in the industry.

2. Open Source: Apache Spark- an alternative to MapReduce

In contrast to Hadoop’s two-stage disk-based MapReduce paradigm, Spark’s multi-stage in-memory primitives provide performance up to 100 times faster for certain applications by allowing user programs to load data into a cluster’s memory and query it repeatedly. Spark could be used either within a Hadoop framework or outside it. Spark requires technical expertise to program and run.

3. Some More Technologies: Python, Data Lake, NoSQL

Python

Is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. Python supports multiple programming paradigms, including object-oriented, imperative, and functional programming or procedural styles.

Data Lake

A **Data Lake** is a large storage repository that “holds data until it is needed.” The term was coined by the chief technology officer of Pentaho.

NoSQL

A NoSQL (originally referring to “non SQL” or “non-relational”) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational da-

tabases. Motivations for this approach include: simplicity of design, simpler “horizontal” scaling to clusters of machines, which is a problem for relational databases, and finer control over availability.

4. SQL

This is the old faithful of a programming language - almost 40 years old. It has been resurrected after a lull of the relational world. NoSQL is used in the more complex environment of humongous data, but SQL is used for “no brainer” simple applications. And because of the impetus of organizations such as Cloudera’s Impala, SQL is almost becoming the *lingua franca* for the next generation of Hadoop-scale Data Warehouses.

5. General-Purpose Programming Languages: Java, C, Python, Scala

General-purpose programming languages such as Java, C, Python, and Scala would be very useful for a person with an analytics background. Computer programmers with data analytics backgrounds are highly in demand. In computer software a general-purpose programming language is a programming language designed to be used for writing software in a wide variety of application domains.

6. Data Mining and Machine Learning

Data Mining

Data Mining is the computational process of discovering patterns in large **data sets** (“**Big Data**”) involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database systems**. It is the analysis of data with the intent to discover gems of hidden information in the vast quantity of data that has been captured in the normal course of running the business.

Machine Learning

Machine learning evolved from the study of **pattern recognition** and **computational learning theory** in **artificial intelligence**. Machine learning explores the study and construction of **algorithms** that can **learn** from and make predictions on **data**.

7. Statistical and Quantitative Analysis

This is the crux of what Big Data is all about, and its main purpose. If a person has a background in quantitative reasoning and a degree in a field like mathematics or statistics, the person is already halfway there. If you have worked with the language R, or have used statistical software, you are a number of notches up. Quantitative background is a BIG plus. It is analysis of a situation or event by means of complex mathematical and statistical modeling. It is translation of data into information and in turn into Predictive Insight.

8. Data Visualization

Big Data could be very complex to comprehend if one is looking only at numbers and letters. There is no comparison to comprehension by the human brain when your eyes see the “shape of your data.” Visualized representation is an interface that presents information in an easy-to-understand and easy-to-relate, often graphical way, providing users with a lot of meaningful information at a glance.

9. Creativity

Creativity is a phenomenon whereby something new and somehow valuable is formed. No matter what software and hardware you use, in whichever industry, your brain is invaluable. These tools listed here will be replaced with other ones in a few years. But the human brain has been developed over a few million years. The creativity potential of our brain cells is monumental. Curiosity is the key to creativity, leading to new ways of looking at Big Data. Can you tell stories based on the data and can you communicate to the appropriate audience? Do you like data and like to play with it?

10. Problem Solving and Subject Matter Expertise

If you are equipped with the subject matter expertise, such as health, finances, telecommunications, retail, etc., and have the ability to think out of the box (look at the data differently from the way everybody else is looking at it), are not afraid of swimming against the stream, and don't take the path of least resistance out of convenience, you are the best candidate for Big Data projects.

Have you thought about possibly moving the Business Analytics to business departments away from IT?

It is completely unrealistic to expect one person to have all the skills needed to handle Big Data, so staffing for the required strengths will likely be a “mix and match.” A manager should have, so to say, a “Yin Yang” balance. For example, you should have two people, one with “Hadoop” as area of major expertise and “Spark” as a minor expertise; the other person should have “Spark” as area of major expertise and “Hadoop” as a minor expertise. This way, if one of them leaves, which is bound to happen because these people are in high demand, the company is not completely at a loss.

INFORMATION

<p>PARTICIPATION FEE</p> <p>€ 1300</p> <p>The fee includes all seminar documentation, luncheon and coffee breaks.</p> <p>VENUE</p> <p>Residenza di Ripetta Via di Ripetta, 231 Rome (Italy)</p> <p>SEMINAR TIMETABLE</p> <p>9.30 am - 1.00 pm 2.00 pm - 5.00 pm</p>	<p>HOW TO REGISTER</p> <p>You must send the registration form with the receipt of the payment to: TECHNOLOGY TRANSFER S.r.l. Piazza Cavour, 3 - 00193 Rome (Italy) Fax +39-06-6871102</p> <p>within March 21, 2016</p> <p>PAYMENT</p> <p>Wire transfer to: Technology Transfer S.r.l. Banca: Cariparma Agenzia 1 di Roma IBAN Code: IT 03 W 06230 03202 000057031348 BIC/SWIFT: CRPPIT2P546</p>	<p>GENERAL CONDITIONS</p> <p>DISCOUNT</p> <p>The participants who will register 30 days before the seminar are entitled to a 5% discount.</p> <p>If a company registers 5 participants to the same seminar, it will pay only for 4.</p> <p>Those who benefit of this discount are not entitled to other discounts for the same seminar.</p> <p>CANCELLATION POLICY</p> <p>A full refund is given for any cancellation received more than 15 days before the seminar starts. Cancellations less than 15 days prior the event are liable for 50% of the fee. Cancellations less than one week prior to the event date will be liable for the full fee.</p> <p>CANCELLATION LIABILITY</p> <p>In the case of cancellation of an event for any reason, Technology Transfer's liability is limited to the return of the registration fee only.</p>
--	---	--

SHAKU ATRE 10 SKILLS TO GET THE BEST OF BIG DATA

April 4-5, 2016
Residenza di Ripetta
Via di Ripetta, 231
Rome (Italy)

Registration fee:
€ 1300

If registered participants are unable to attend, or in case of cancellation of the seminar, the general conditions mentioned before are applicable.

first name

surname

job title

organisation

address

postcode

city

country

telephone

fax

e-mail



Stamp and signature

Send your registration form with the receipt of the payment to:
Technology Transfer S.r.l.
Piazza Cavour, 3 - 00193 Rome (Italy)
Tel. +39-06-6832227 - Fax +39-06-6871102
info@technologytransfer.it
www.technologytransfer.it



Shaku Atre is an exceptional speaker, with the reputation of capturing the attention of audiences and maintaining their interest while guiding her listeners painlessly through sophisticated material. Ms. Atre is the President of Atre Group Inc. which is a leading consulting, training, and publishing company specializing in Business Intelligence (BI), Data Warehouses and Big Data. Before heading her present company, Ms. Atre was a Partner with Price Waterhouse Coopers. She also has fourteen years of experience in various fields with IBM. Ms. Atre is an acknowledged expert in the Data Warehousing and database field. She has extensive practical experience in database projects, has helped a number of clients in establishing successful Data Warehouses and client/server installations, and has taught at IBM's prestigious Systems Research Institute. She has lectured on the subject to professional organizations in the USA and Canada, as well as in more than 35 countries around the world. Ms. Atre is frequently quoted in reputable publications such as Computerworld and Information Week. She has written an award-winning outstanding book on database management systems that has become a classic on the subject: ***Database: Structured Techniques for Design, Performance and Management***, published by John Wiley and Sons, New York. The book has sold over 250,000 copies (not including its Spanish and Russian translations) and has been selected by several book clubs and leading universities including Harvard, Columbia, Cornell, MIT, New York University, Stanford, and U.C. Berkeley as well as by the Moscow University. Her book ***Information Center: Strategies and Case Studies*** published by Atre International Consultants Inc., has also been very well received by the industry. ***Database Management Systems*** is another successful book authored by Ms. Atre. Her fourth book, ***Distributed Databases, Cooperative Processing & Networking*** was published by McGraw-Hill. She has also authored a very well received book: ***Atre's Roadmap for Data Warehouse/Data Mart Implementations*** published by Gartner Group, and is co-author of her latest BI book ***Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*** published by Addison Wesley.