INTRODUCING DATA SCIENCE
AND BIG DATA ANALYTICS FOR
BUSINESS TRANSFORMATION

Copyright © 2013 EMC Corporation. All Rights Reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

EMC², EMC, EMC Proven Professional and the EMC logo are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. Copyright © 2013 EMC Corporation. All rights reserved. Published in the USA.

Revision Date: 04-22-2013
Revision Number: 1.01 MR-1TN-NDSFBLIN
## Course Overview

<table>
<thead>
<tr>
<th>Description</th>
<th>Businesses are increasingly looking to take advantage of Big Data to be competitive. To do this, organizations need data-savvy business leaders who can identify opportunities to solve business problems using advanced analytics. This course provides business leaders with a baseline understanding of Data Science and Big Data. Specifically, it addresses: illustrative examples of Data Science and Big Data Analytics in three industry verticals, deriving business value from Big Data, characteristics of Big Data, and key elements to consider for driving the change toward Big Data projects.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience</td>
<td>The intended audience for this course includes:</td>
</tr>
<tr>
<td></td>
<td>• Executives who would like concrete examples of how organizations are taking advantage of Big Data and the related benefits of data-driven decision making</td>
</tr>
<tr>
<td></td>
<td>• Business leaders who want to develop new Data Science capabilities</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>There are no Prerequisites for this course.</td>
</tr>
</tbody>
</table>

This overview provides a description of the course and audience information.
This course provides an overview of some of the drivers of Big Data Analytics, the benefits of Big Data, and industry examples of how Big Data is being used today to make business decisions and leverage advanced analytics to create competitive advantage. This course also defines Big Data and provides an overview of some of the new approaches required to analyze it. Finally, this course presents information about new roles and processes that companies need to develop to take advantage of new Big Data opportunities.
“I’ll go into a company and say, ‘What data problems can we solve?’ We get blank looks,” he says.

...when he asks, instead, ‘what things can help a company lose money and make money, usually two out of three are problems that data can solve.’

Anthony Goldbloom, CEO of Kaggle

Data science is not about solving abstract data problems, rather it is about using data to solve pragmatic business problems. Many times the problems that data science methods can help to solve are problems that people do not realize can be addressed with data and analytics.
Applications of Big Data Analytics

This section covers healthcare, telecommunications, and financial services examples of how to use Big Data Analytics to improve business.
In this course, we will look at three use case examples of Big Data, spanning Health Care, Telecommunications, and Financial Services. Each of these use cases show how Big Data can be used to approach traditional problems in new ways to solve business problems in better and faster ways at potentially lower costs.
The first industry example we will explore is in health care.
To get started, let’s look at an example that incorporates Big Data Analytics with a traditional industry – healthcare.

When there is a pandemic, a disease control center needs to find ways to distribute vaccines to those in need. Traditionally, a center would distribute vaccines based on regional population or first-come-first-served, which do not necessarily represent the distribution of patients. The center may wait for reports from hospitals and statistics from authorities such as Center for Disease Control (CDC) which can take 3-6 months. During this time, the pandemic continues to spread and may become out of control before the disease control center can take actions.

CDC’s 2009 H1N1 vaccine allocation and distribution was simply based on population. More information can be found at: http://www.cdc.gov/h1n1flu/vaccination/statelocal/centralized_distribution_qa.htm.
By taking advantage of Big Data and advanced analytical techniques, the CDC and others like it can use social media to detect and monitor pandemic diseases in near real-time.

One approach for doing this is to query Twitter with keywords such as “flu”, “vaccine” and “immunization” to find potential patients. These queries will show people who are tweeting about the flu, about vaccines and symptoms they may have related to these terms. A disease control center can fetch the related tweets from Twitter to analyze the scope of a potential disease outbreak. The data contain text from individuals related to flu epidemics or related topics and from that Data Scientists can discern the severity of a potential epidemic in near real time.

After fetching the related tweets, data scientists can analyze the social networks within the data. This will tell you the likely spread of the disease within the social networks. In other words, if one person is having flu symptoms and has frequent contact with others in their social network, then these friends of theirs may more likely to get the illness. Moreover, other insights such as patient sentiments and the effectiveness of a vaccine can be obtained from analyzing the data.

Because the tweets contain information such as date, time, and geospatial location, they can be overlayed on a map to show where the highest concentration of people are who have an illness or where a disease is spreading the most rapidly. Instead of the three-six months needed for traditional reporting, this analysis can be done in near-real time. A disease control center therefore can quickly identify the disease spreading trend and make much better decisions about where to offer and distribute vaccines to meet the needs of people with a given illness of disease outbreak.

In addition to using social networks such as Twitter to detect and monitor pandemic diseases, they can also be used to track vaccine effectiveness, call for test subjects, or cloudsource volunteers to distribute vaccines, provide consultations, and perform emergency rescues and other assistances.
The next example focuses on new ways to look at churn prediction within the telecommunications industry.
There are many reasons why customers may ‘churn’, or decide to leave a mobile Telco company. Churn rate refers to the number of participants who discontinue their use of a service divided by the average number of total participants during a period. Some common reasons for this include customer dissatisfaction, the ease of switching providers, inadequate services, poor quality of service, or attractive competitive offers.

Instead of working with a small dataset as has been done traditionally, with Big Data we can now look at raw call history data for each customer, which contains 10 billions rows of data. Within this data, it is possible to predict which customers are the most likely to churn, or leave a Telco provider using conventional methods, such as looking for changes in a customer’s spending patterns or frequent phone calls complaining about service. However, this analysis can be made more robust and improve the ability to predict which customers may churn by storing all of the call history records in a database and constructing a social network from the call logs.

When viewed as a social network, we can infer features within the data, such as which callers call each other frequently, and therefore influence each others habits, such as leaving a mobile cell phone provider or joining a new one. Viewing this cell phone call history data as a social network of people enables much richer insights into the data.
Shown is a view of the cell phone call history data as a social graph, in the first month of the analysis.
Shown is the second month of call activity, with the first two people shown in red as canceling their cell phone contracts.
In the third month of call activity, two more people in the same social network have canceled their cell phone contracts.
In month four, three more people in the same social network have canceled their cell phone contracts.
From a Big Data perspective, the calling network can be examined as a social network and Big Data tools and techniques can be applied to extract more insights from it.

From this analysis, it is possible to determine that customers who received many incoming calls from churners were seven times more likely to churn. In addition, the Telco chose to become more proactive about churn prediction by systematically identifying customer value, planning a profitable marketing strategy, pointing out clients most likely to defect, and developing a win-back policy for worthwhile customers.

Rather than doing a traditional churn analysis once or twice per year, using the techniques described in this example, churn analysis can be performed in an automated fashion in-database in one hour. Due to the speed of the analysis, it can be done weekly or daily on large scales (10 billion records) in an hour. In this case, the Telco provider saved $40 million in the first year in lost customer revenue.
The third Big Data example is in the area of Financial Services. This example highlights new ways and new data sources for producing lending decisions.
The process of making lending decisions in the Financial Services industry has been honed to a science over the past several decades. Today’s realities require that lenders take care to make better decisions with fewer resources than they’ve had in the past. The typical loan process uses a set of data on which pre-approval and underwriting approval is based on, including:

- Income data, such as paystubs and income tax records
- Employment history to establish the ability to meet loan obligations
- Credit history including credit scores and outstanding debt
- Appraisal data associated with the asset that the loan is being taken out against (e.g., home, RV, boat, car, etc.)

This model works, but it’s not perfect. In fact, the loan crisis in the U.S. is proof that using only these data points may not be enough to gauge the risk associated with loans.
Using Big Data can improve load underwriting quality and streamline the process to yield results quicker. Consider leveraging publicly available data that to supplement the traditional lending process. Also, determine the types of analysis that can be performed with the data to reduce the bank’s risk and expedite the lending process. For example:

- Integrate home price trends from data sources like Zillow and eppraisal (eppraisal.com) with appraisal data to identify patterns or conflicting data.
- Use Census data to understand population migration trends and the potential impact on home values or geographic income variation.
- Use localized job trends to identify patterns in prosperity in a specific area.
- Use Historic Loan Data from the lenders’ own records or purchased data from 3rd parties that may aggregate loan data from public records to find patterns that point to more or less risky behavioral patterns.
- Use the social and professional history of applications to correlate behavior in personal and professional life through sources like Twitter, Facebook, and LinkedIn that may have an impact on the ability to pay or continue paying debts.

The end result is a much richer analysis and resulting insight, which can drive a process to either approve or deny a loan application with much lower risk than is possible today.
The secondary benefit is that the approval process can be streamlined from an average of three to four weeks to two to three weeks, which is a savings of over 30%. In some cases, if more data sources are available, the time can almost be cut in half.

Loan approvals are just the beginning; far more transformative changes can be enabled with Big Data in the financial services arena:

- What if life events such as a change in relationship status (for example, becoming single if they were originally married) in Facebook could generate an alert that kicks off a process to proactively check-ins in on a customer with a large mortgage loan to determine if the loan is at risk?

- What if travel insurance could be offered if credit card transaction data or a Tweet indicates an upcoming international vacation?

The possibilities are endless when insight is used to drive action with Big Data.
Now that we have examined several Big Data examples from three different industry verticals, let’s look at various types of business value that Big Data provides, the business drivers for advanced Analytics and the kinds of benefits that Big Data projects can bring.
Here are four examples of common business problems that organizations contend with today, where there is an opportunity to leverage advanced analytics to create competitive advantage. Rather than performing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these typical tasks.

The first three examples are not new problems—companies have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. What’s new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these old problems.

The last example portrays emerging regulatory requirements. Many compliance and regulatory laws have been in existence for decades, but additional requirements are added every year, which mean additional complexity and data requirements for organizations. These laws, such as anti-money laundering and fraud prevention, require advanced analytical techniques to manage well.
The successful implementation of a Data Science project results in financial benefits, such as increased revenue or improved efficiencies that lower costs. However, Data Science projects can provide improvements in areas that are harder to quantify. These areas include: decision quality, data quality, and collaboration.
Let’s start by looking at the financial benefits of Data Science projects. A November 2011 report by Nucleus Research found that analytics pays back $10.66 for every dollar spent. This report is based on 60 case studies, which found ROI in such areas as revenue, gross margin and expenses. Nucleus Research also noted that companies are slow to embrace analytical approaches and technologies.


A subsequent March 2012 Nucleus Research report stated specific ROIs for several big data projects in various industry verticals. The overall ROIs varied depending on how the analytical methods were implemented with an organization. The maximum ROI can be obtained when analytics are applied to automation, tactical, strategic, and predictive initiatives.


The October 2012 issue of the Harvard Business Review focused on topic of Big Data. Among other topics, Andrew McAfee and Erik Brynjolfsson’s article emphasized that to obtain the most benefit from Big Data and analytics, companies will need to implement data-driven decision making processes.

A key task in a Data Science project is to translate the business problem into an analytic problem statement. This process, by including the proper participation of the various sponsors and stakeholders, can provide clarity and consensus that the correct problem is being addressed by the Data Science project.

A second component of decision quality is knowing precisely how the decision was made. If the decision making process is well-defined, it can be repeated more easily and consistently. This repeatability allows further improvements and adjustments as the economic, demographic, and competitive conditions change.

Finally, a Data Science project will examine the uncertainty in the inputs and outputs of the decision making process. Inquiries into the data quality and the applicability of the assumptions are examined in a Data Science project. In many cases, the uncertainty of the decision can be quantified. For example, in determining whether or not a particular credit card transaction is fraudulent, some probability of such an event will be determined. Other examples of sample decisions enabled by Data Science are identifying items to recommend to a specific customer or determining which marketing promotions should be run together.
Just the simple act of looking at the data can lead to initiatives to prevent or correct bad data. However, when the data is being used to make decisions, the data goes under an even stricter level of scrutiny. Throughout a project lifecycle, the Data Science team will uncover data issues and take actions to address them appropriately. Data Scientists expect to find inconsistencies with the data. It would be very naïve to expect the data to be perfect.

In the short term, steps can be taken to address apparent outliers or to clean reference data and transactions to facilitate the model-building activities. In the longer term, controls and processes will need to be in place to ensure that any new data discrepancies are detected and addressed.

Thus, as organizations implement more and more analytical decision making into their operations, the need for formal Master Data Management (MDM) and Data Governance programs will become apparent. In simplest of terms, Data Governance can be considered as the top down policies, procedures and decision-making bodies that oversee how data is managed across an organization. Master Data Management is the bottom up, focused effort to ensure reference data is used consistently throughout an organization. A typical example of inconsistent master data occurs when separate customer databases exist within sales, manufacturing, marketing, and customer service, but there is no mechanism to keep the customer data in synch in a timely manner. In such a case, the customer account names will eventually differ between the systems.
There are several collaboration benefits that result from a Data Science project. For example, if a Data Science project is underway to reduce customer churn for a wireless service company, it is important for the Data Scientist to interact with the operational business users and their management to understand how the business processes operate and what kind of data may be available or required. The operational users may include customer support personnel who are taking the phone calls to drop the customer’s existing service. The marketing and sales groups may already have tried several activities and promotions to reduce churn. In sifting through the data, it will be important to determine what was attempted and when. Other groups may include IT data engineers and DBAs as well as business analysts who are already using the available data sources in their reporting.

A well defined and managed Data Science team will clearly identify the roles, responsibilities, and interdependencies, but will also develop a collaborative approach to reaching the desired outcome. The collaboration will provide a diverse set of ideas and opinions about how the existing processes work, what factors affect the process, how to improve the process, and even what the data elements mean.

Throughout the organization, the teamwork will result in better awareness of the project, its intended objectives, and eventually the implemented solution. This awareness will help to identify any implementation obstacles and pave the way for better acceptance of the proposed solution. For example, in the customer churn example, if customer service is to contact the customers with certain promotions intended to retain the customer, then customer service may be more receptive and effective in the calls being placed when customer service was involved in project. Related to awareness, lines of communication will be established to quickly address a project issue when it arises or additional help is required.

Also, initial Data Science projects help to establish a foundation for future projects and activities. Any data clean-up and integration can be leveraged in the next project. Finally, the team members become advocates for additional data integration and data integrity activities to continue to build out the foundation for future projects.
In the previous sections we’ve looked at how industries are using Big Data and discussed the benefits that are created from Big Data projects. This section focuses on areas that are common to these examples as they relate to Big Data, and the types of analytical methods being employed.
As we saw in the three Big Data industry examples covered earlier in this course, each of the use cases takes advantage of new kinds of data sets and analytical methods. Much of the data made use of is of a less-traditional nature, meaning it is varied in its structure and how it is being used to solve a specific industry problem.

With that in mind, consider that Big Data has many characteristics, but three stand out as the defining ones:

• Huge volume of data, rather than thousands or millions of rows, big data can be billions of rows and billions of columns.
• Complexity of data types and structures, driven by the variety of new data sources, formats, structures, and digital traces being left on the web for subsequent analysis.
• Speed or velocity of new data creation.

In addition, the data, due to its size or level of structure, cannot be efficiently analyzed using only traditional databases or methods.

There are many examples of emerging Big Data opportunities and solutions. Here are a few real examples of how Big Data impacts us:

• Netflix suggesting your next movie rental
• Dynamic monitoring of embedded sensors in bridges to detect real-time stresses and long-term erosion
• Retailers analyzing digital video streams to optimize product and display layouts and promotional spaces on a store-by-store basis

These kinds of Big Data problems require new tools/technologies to store, manage, and realize the business benefit. The new architectures it necessitates is supported by new tools, processes, and procedures that enable organizations to create, manipulate, and manage these very large data sets and the storage environments that house them.
Traditionally companies invested in data warehouses to harness the value in the data they generated and stored in databases as structured data. To derive business value from the big data which is predominantly non-structured new approaches have emerged to analyze them. The Data Science approach is more predictive in nature in terms of the business questions it answers compared to the Business Intelligence methods conventionally used with structured data that focused on the data to report on the past performances of the business.

- **Business Intelligence (BI)** - focuses on using a consistent set of metrics to measure past performance and inform business planning. This includes creating Key Performance Indicators (KPIs) that reflect the most essential metrics to measure your business. Measures and KPIs are commonly defined within the OLAP schema to enable BI reporting on defined metrics.

- **Predictive Analytics and Data Mining (Data Science)** - refers to a combination of analytical and machine learning techniques used for drawing inferences and insight out of data. These methods include approaches such as regression analysis, Association Rules (for example, Market Basket Analysis), optimization techniques, and simulations (for example, Monte Carlo simulation to model scenario outcomes). These are the more robust techniques for answering higher order questions and deriving greater value for an organization.

Both BI and Data Science are needed for organizations to meet these emerging business challenges successfully.
The amount of information created by mankind is soaring. We create data when we use our mobile phones, social media, or when imaging is used to determine a medical diagnosis. Merely keeping up with this flood of data is hard, but analyzing it to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life.

A few industries have led the way in their ability to gather and exploit data. Credit-card companies monitor every purchase and can identify fraudulent ones with a high degree of accuracy, using rules derived by crunching through billions of transactions.

Mobile-phone operators, analyze subscribers' calling patterns to determine, for example, whether most of their frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, he or she can then be offered an incentive to stay.

For companies like Linkedin and Facebook, data is their primary product. The company's valuations is entirely derived from the data they gather and host.
Over the course of the session, we have introduced examples of applying advanced analytics to industries ranging from financial services to health care to predicting customer attrition. After exploring these examples, refocus on the task that you are trying to solve and consider how to apply these concepts to your situation.

To look at more examples of influential Data Scientists who have shaped or created industries, look at this article, in which Tim O'Reilly highlights The World's 7 Most Powerful Data Scientists: [http://www.forbes.com/sites/nicoleperlroth/2011/11/02/tim-oreilly-the-worlds-7-most-powerful-data-scientists/](http://www.forbes.com/sites/nicoleperlroth/2011/11/02/tim-oreilly-the-worlds-7-most-powerful-data-scientists/)
After seeing multiple examples of ways that Big Data is transforming various industry verticals, learning the benefits of Big Data projects, and distinguishing between Business Intelligence and Data Science, it’s critical to understand next steps and consider how to begin taking advantage of Big Data Analytics in your organization.

Before evaluating tools and technologies, we recommend that you identify a champion for Big Data and advanced analytics for your organization. This is someone with a solid understanding of advanced analytics, business experience to draw on, and has an entrepreneurial spirit for trying new projects and driving change through analytics. This person can collaborate with you to develop a business case that will identify the right opportunity areas for your organization, help outline potential Big Data projects to undertake, and outline the costs and benefits of developing data science capabilities for your organization.

In addition, it is critical that the head of this new function receive training on how to be successful in their new role. EMC’s one-day course “Data Science and Big Data Analytics for Business Transformation” is designed for this purpose. This course provides an overview of the benefits of data-driven decision making, developing analytics teams, providing high-level understanding of several key data science methods, how to lead analytics projects using a structured lifecycle approach, and driving innovation and visionary thinking with analytics.

For people who join the new Data Science and Big Data Analytics team, we recommend having them attend EMC’s five-day course “Data Science & Big Data Analytics”. The objective of this course is to enable someone to participate immediately and contribute as a Data Science Team Member. This course content covers reframing a business challenge as an analytics challenge, deploying a Data Analytics Lifecycle for big data analytics projects, applying appropriate analytic techniques and tools to analyze Big Data, selecting appropriate data visualizations to clearly communicate analytic insights to business sponsors and analytic audiences, and hands-on usage of tools such as: R and RStudio, MapReduce/Hadoop, in-database analytics, and advanced SQL functions.
Summary

- Big Data has emerged from internet of things and variety of high-volume data sources
- New market drivers are providing urgency for advanced analytics
- Data Science techniques provide higher value than simple reporting and dashboards
- New technology architectures, roles, and processes are needed to take advantage of new Big Data opportunities.

This summarizes the key points covered in this module. As we saw through the industry use cases, Big Data provides tremendous opportunities for growth and innovation for businesses of all kinds. However, to take advantage of these opportunities one must be open-minded and creative to explore traditional problems and processes in new ways. For instance, we shared new ways to approach problems in Healthcare, Telecommunications and Financial Services. These problems (providing medicine, retaining customers, and making lending decisions) are not new, but with Big Data we have an opportunity to rethink many aspects of business and improve upon them. To take advantage of these opportunities, people will need to embrace new technology architectures, roles, and processes.