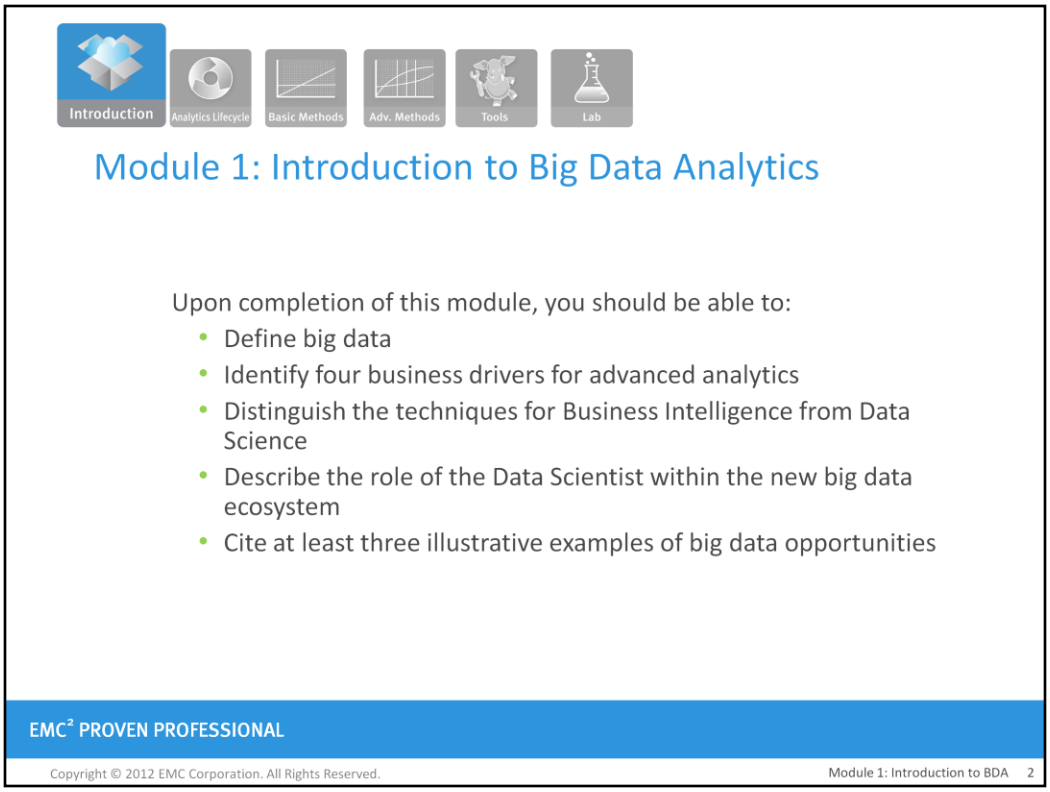
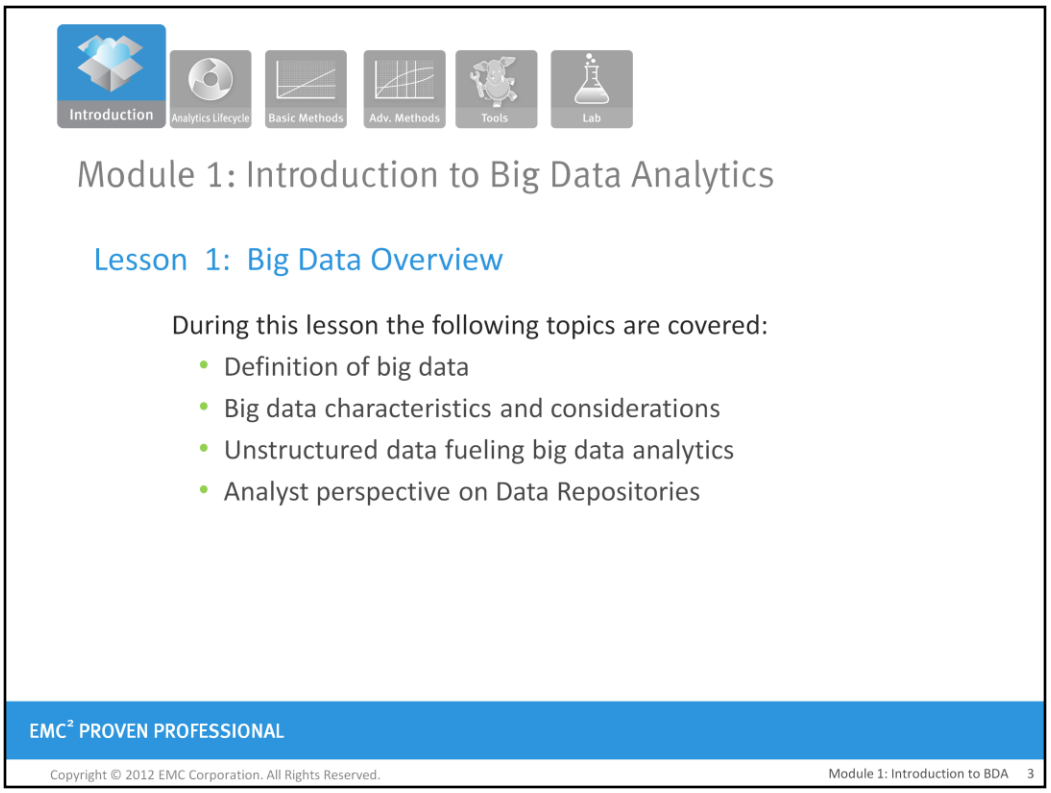
Introduction

Module 1 – Introduction to Big Data Analytics

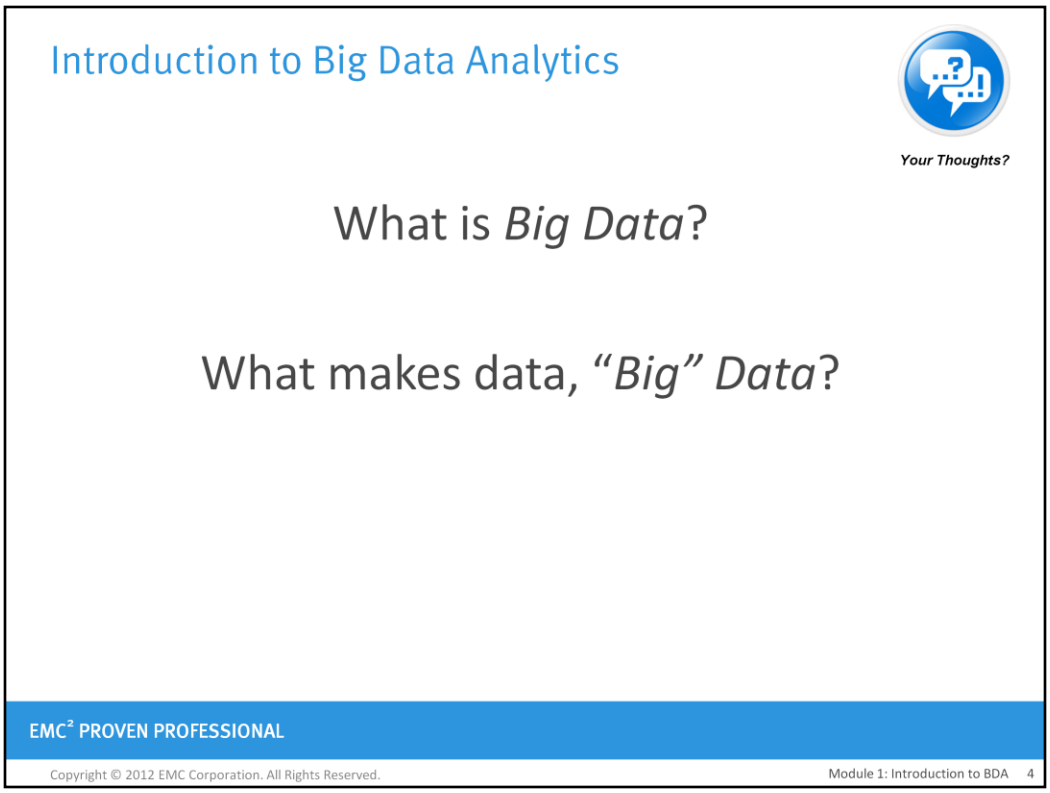
EMC² PROVEN PROFESSIONAL



This module provides an overview of today's big data environment, the rationale and opportunity for a new approach to analytics, the roles required, including the Data Scientist, and representative examples of big data analytics in industry verticals.



This lesson covers big data and its characteristics that are fueling big data analytics, and the evolution of data repositories.

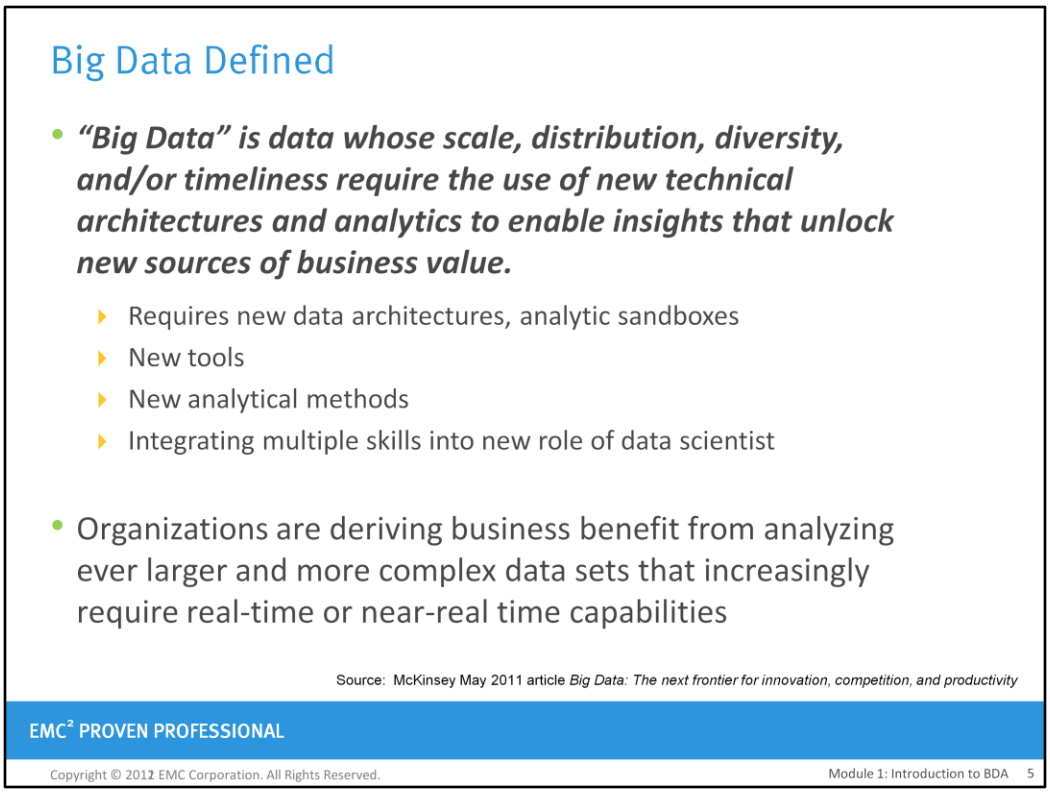


Think about what Big Data is for a moment. Share your thoughts with the group and write your notes in the space below.

Is there a size threshold over which data becomes Big Data?

How much does the complexity of its structure influence the designation as Big Data?

How new are the analytical techniques?



There are multiple characteristics of big data, but 3 stand out as defining Characteristics:

- **Huge volume of data** (for instance, tools that can manage billions of rows and billions of columns)
- **Complexity of data types and structures**, with an increasing volume of unstructured data (80-90% of the data in existence is unstructured)….part of the Digital Shadow or "Data Exhaust"
- **Speed or velocity of new data creation**

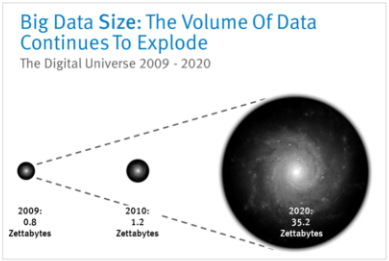
In addition, the data, due to its size or level of structure, **cannot be efficiently analyzed using only traditional databases or methods**.

There are many examples of emerging big data opportunities and solutions. Here are a few: Netflix suggesting your next movie rental, dynamic monitoring of embedded sensors in bridges to detect real-time stresses and longer-term erosion, and retailers analyzing digital video streams to optimize product and display layouts and promotional spaces on a store-by-store basis are a few real examples of how big data is involved in our lives today.

These kinds of big data problems require new tools/technologies to store, manage and realize the business benefit. The new architectures it necessitates are 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.

Key Characteristics of Big Data

1. **Data Volume**
   - 44x increase from 2010 to 2020 (1.2zettabytes to 35.2zb)



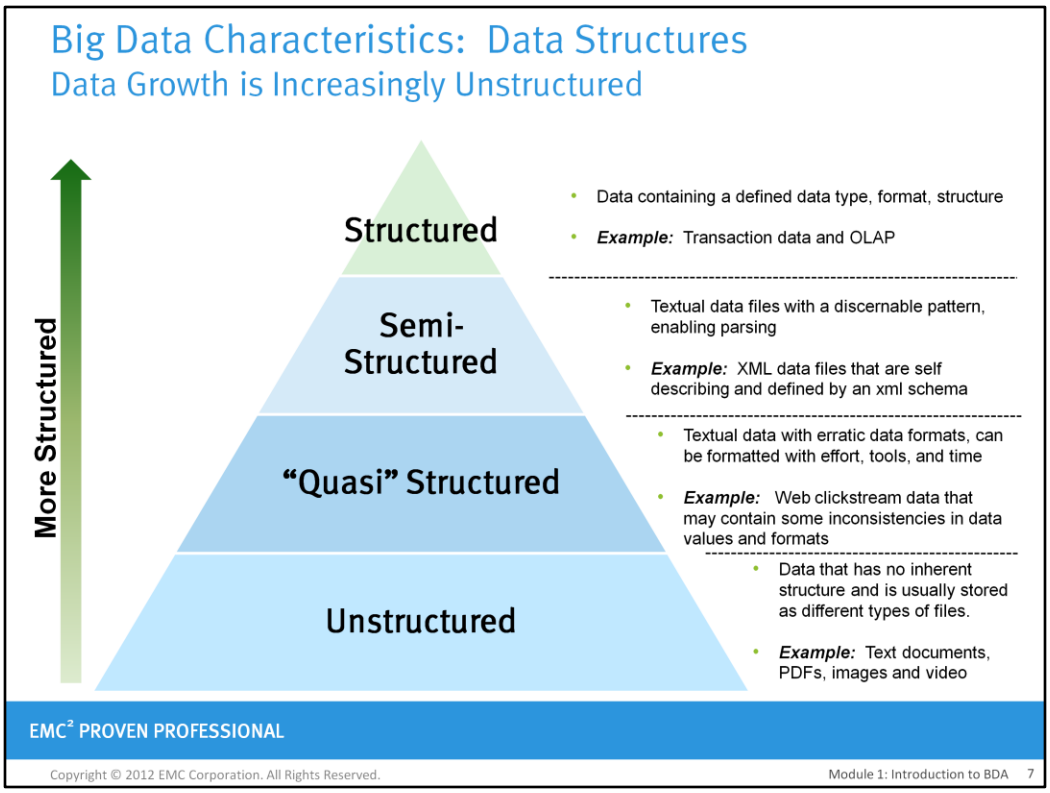
2. **Processing Complexity**
   - Changing data structures
   - Use cases warranting additional transformations and analytical techniques

3. **Data Structure**
   - Greater variety of data structures to mine and analyze

EMC² PROVEN PROFESSIONAL

Big data can come in multiple forms. Everything from highly structured financial data, to text files, to multi-media files and genetic mappings. The **high volume of the data is a consistent characteristic of big data**. As a corollary to this, because of the complexity of the data itself, the preferred approach for processing big data is in parallel computing environments and Massively Parallel Processing (MPP), which enable simultaneous, parallel ingest and data loading and analysis. As we will see in the next slide, **most of the big data is unstructured or semi-structured in nature**, which **requires different techniques and tools to process and analyze.**

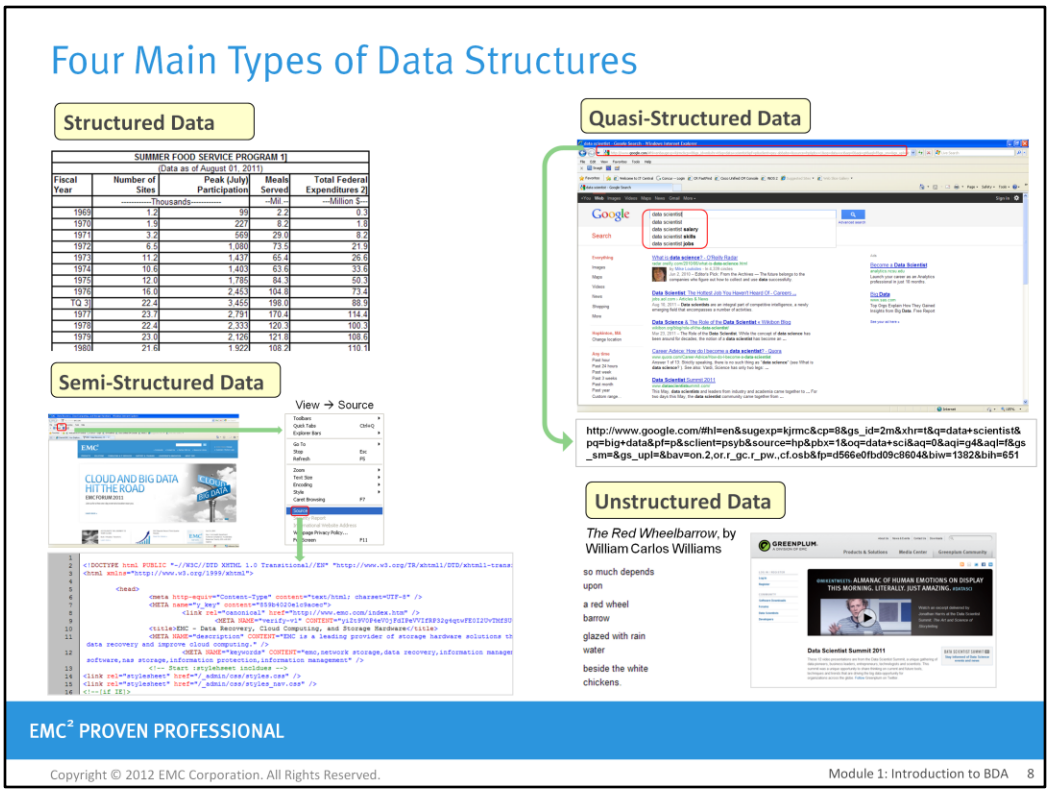
Let us examine the most prominent characteristic: its structure.



The graphic shows different types of data structures, with 80-90% of the future data growth coming from non structured data types (semi, quasi and unstructured).

Although the image shows four different, separate types of data, in reality, these can be mixed together at times. For instance, you may have a classic RDBMS storing call logs for a software support call center. In this case, you may have typical structured data such as date/time stamps, machine types, problem type, operating system, which were probably entered by the support desk person from a pull-down menu GUI.

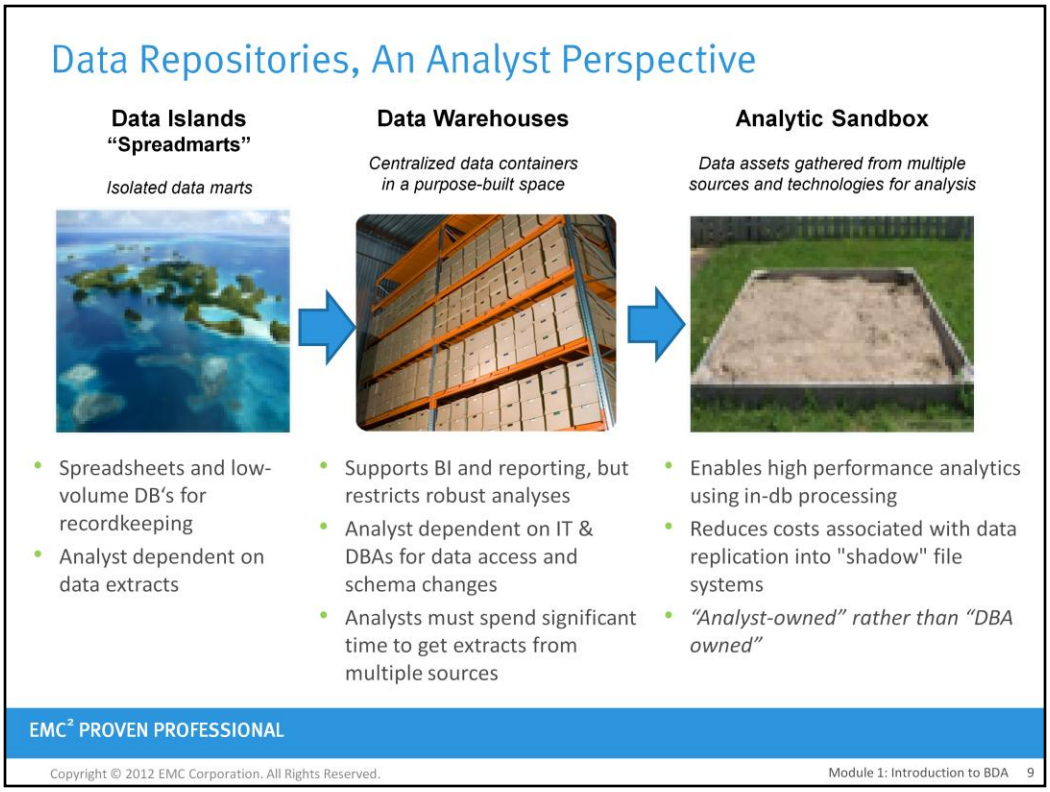
In addition, you will likely have unstructured or semi-structured data, such as free form call log information, taken from an email ticket of the problem or an actual phone call description of a technical problem and a solution. The most salient information is often hidden in there. Another possibility would be voice logs or audio transcripts of the actual call that might be associated with the structured data. Until recently, most analysts would **NOT** be able to analyze the most common and highly structured data in this call log history RDBMS, since the mining of the textual information is very labor intensive and could not be easily automated.



Here are examples of what each of the **4 main different types of data structures** may look like. People tend to be most familiar with analyzing structured data, while semi-structured data (shown as XML here), quasi-structured (shown as a clickstream string), and unstructured data present different challenges and require different techniques to analyze.

For each data type shown, answer these questions:

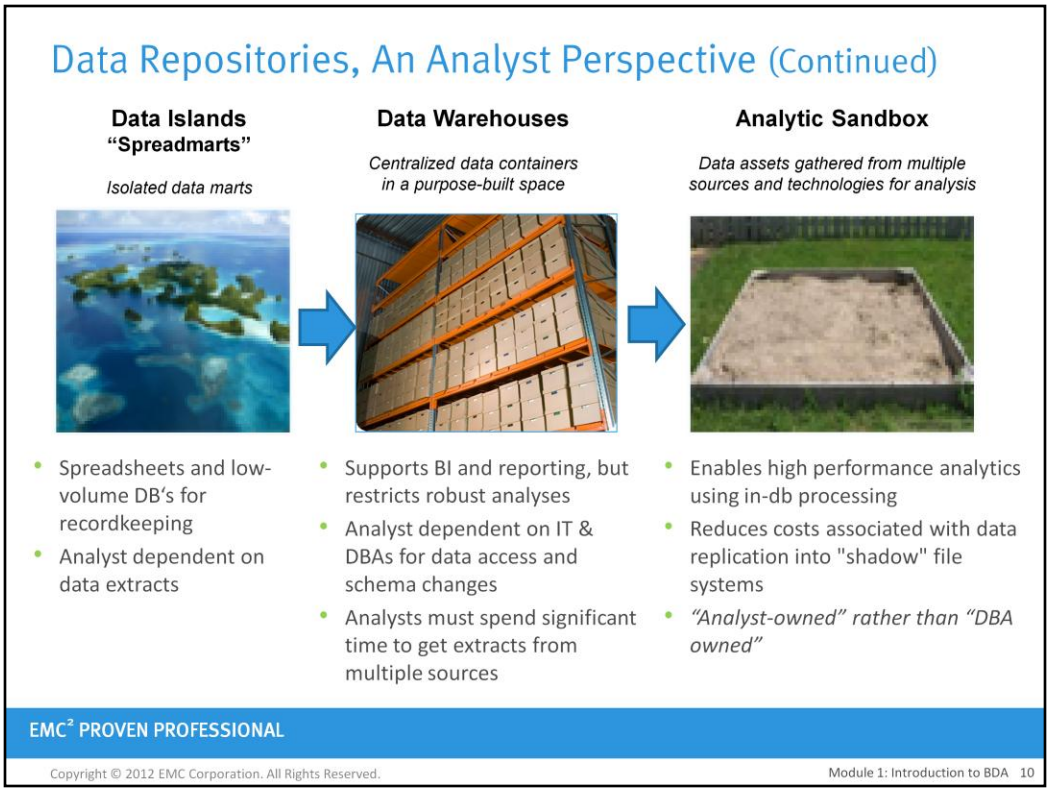
1) What type of analytics are performed on these data?
2) Who analyzes this kind of data?
3) What types of data repositories are suited for each, or requirements you may have for storing and cataloguing this kind of data?
4) Who consumes the data?
5) Who manages and owns the data?



People tend to both love and hate spreadsheets. With their introduction, business users were able to create simple logic on data structured in rows and columns and create their own analyses to business problems. Users do not need heavy training as a database administrator to create spreadsheets, meaning business users could set these up quickly and independent of IT groups. Two main spreadsheet benefits are that they are easy to share and that end users have control over the logic involved. However, their proliferation caused organizations to struggle with "many versions of the truth", i.e. it was impossible to determine if you had the right version of a spreadsheet, with the most current data and logic in it. Moreover, if a user lost a laptop or it became corrupted, that was the end of the data and its logic. Many organizations still suffer from this challenge (Excel is still on millions of PCs worldwide), which gave rise to the need for centralizing the data.

As data needs grew, companies such as Oracle, Teradata, and Microsoft (via SQL Server) offered more scalable data warehousing solutions. These technologies enabled the data to be managed centrally, providing benefits of security, failover, and a single repository where users could rely on getting an "official" source of data for financial reporting or other mission critical tasks. This structure also enabled the creation of OLAP cubes and business intelligence analytical tools, which provided users the ability to access dimensions within this RDBMS quickly, and find answers to streamline reporting needs. Some providers also packaged more advanced logic and the ability to perform more in-depth analytical techniques such as regression and neural networks.

<Continued>



Enterprise data warehouses (EDW) are critical for reporting and **Business Intelligence** (BI) tasks, although from an analyst perspective they tend to restrict the flexibility that a data analyst has for performing robust analysis or data exploration. In this model, data is managed and controlled by IT groups and DBAs, and analysts must depend on IT for access and changes to the data schemas. This tighter control and oversight also means longer lead times for analysts to get data, which generally must come from multiple sources. Another implication is that EDW rules restrict analysts from building data sets, which can cause shadow systems to emerge within organizations containing critical data for constructing analytic data sets, managed locally by power users.

Analytic sandboxes enable high performance computing using in-database processing. This approach **creates relationships to multiple data sources within an organization** and saves the analyst time of creating these data feeds on an individual basis. In-database processing for deep analytics enables faster turnaround time for developing and executing new analytic models, while reducing (though not eliminating) the cost associated with data stored in local, "shadow" file systems. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as webscale data, raw data, and unstructured data.

Introduction to Big Data Analytics: Mini-Case Study

Yoyodyne Bank Scenario

- Evolving from small community bank to a global bank
- Needs to move away from its legacy mainframes to an environment that supports more robust analytics
- Growing through mergers and acquisitions
- Subject to many new regulatory requirements
- Increasing customer base and increased product offerings

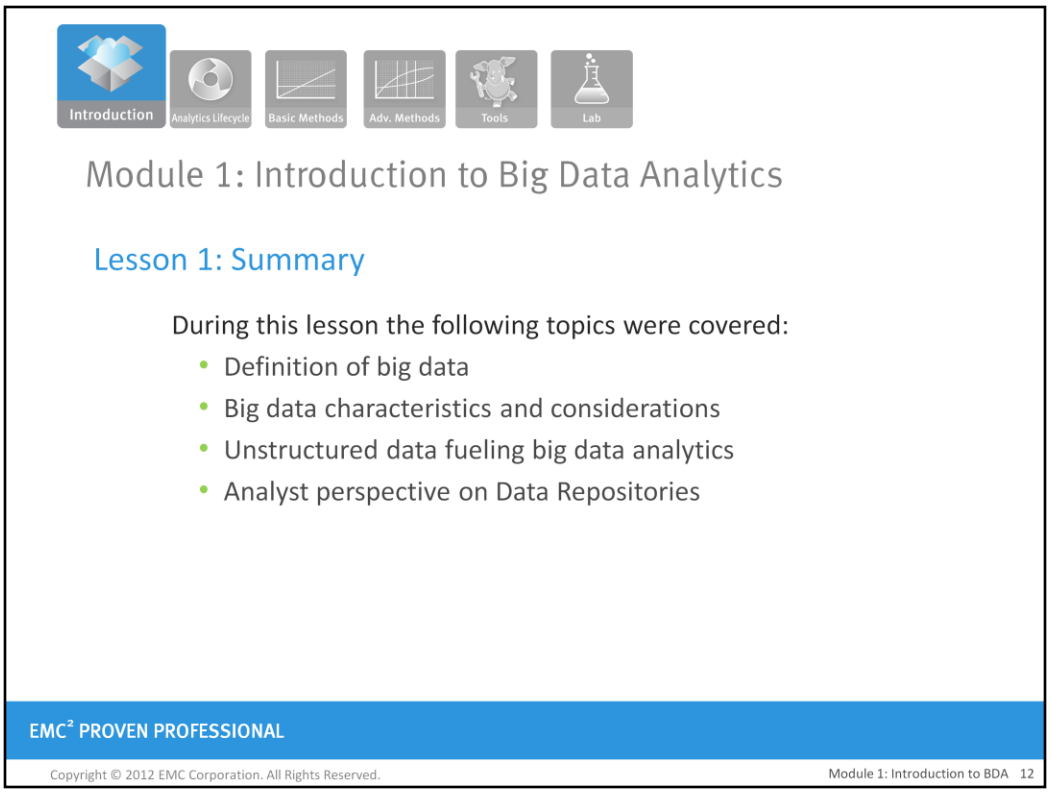
Your Thoughts?

Discussion Questions

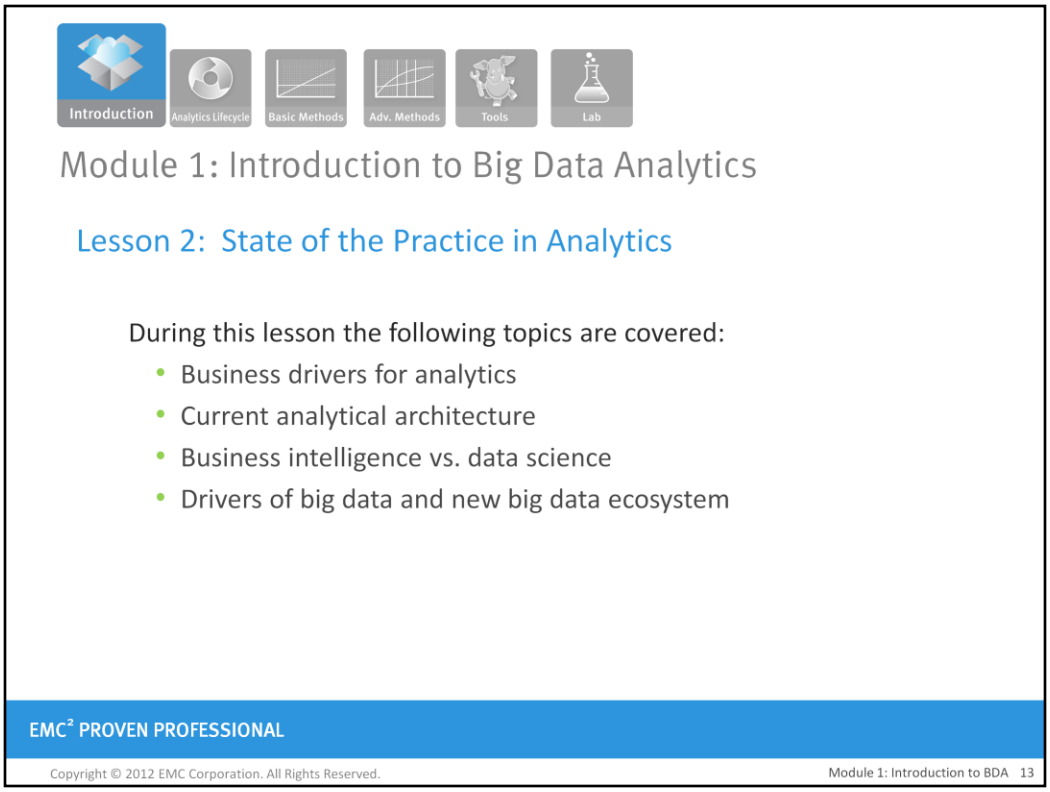
1. Discuss how the bank's data would change under these circumstances.
2. How are their needs changing with these business changes?
3. What do you need to consider from an analyst point of view? What are some things to consider implementing as the bank grows?

EMC² PROVEN PROFESSIONAL

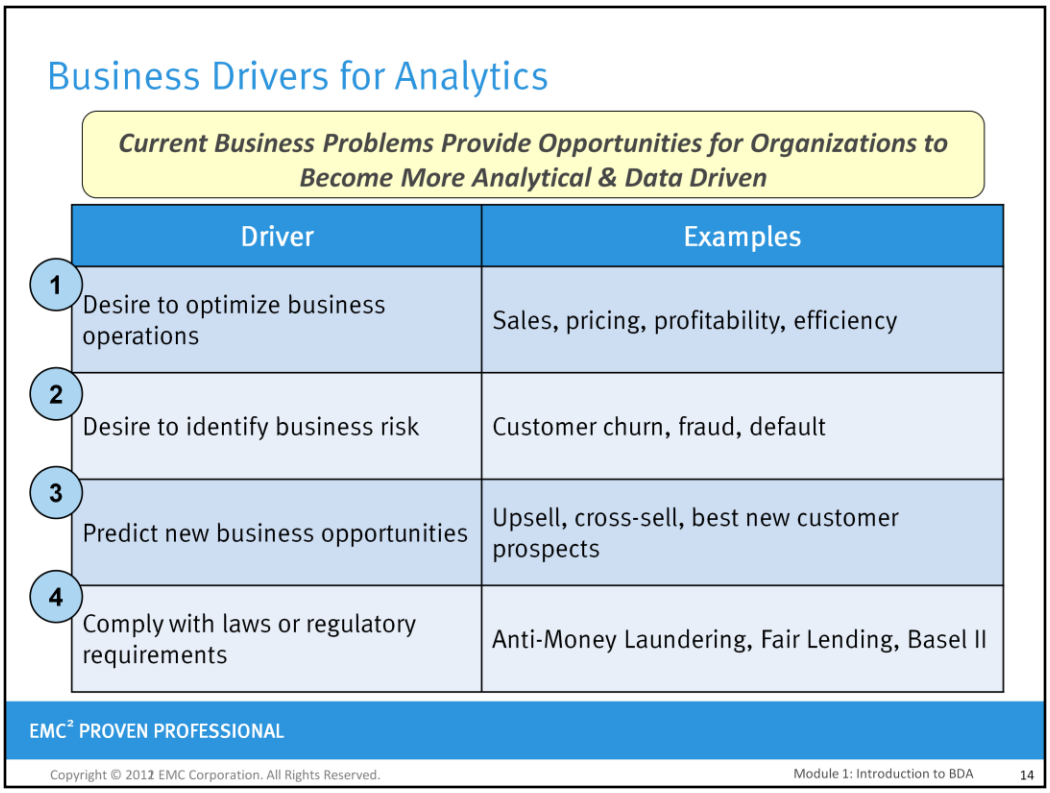
Outlined is a brief scenario for a mini-case study and the questions that will be discussed in the class. In the space below, please note your responses to the questions.



This lesson covered these topics.

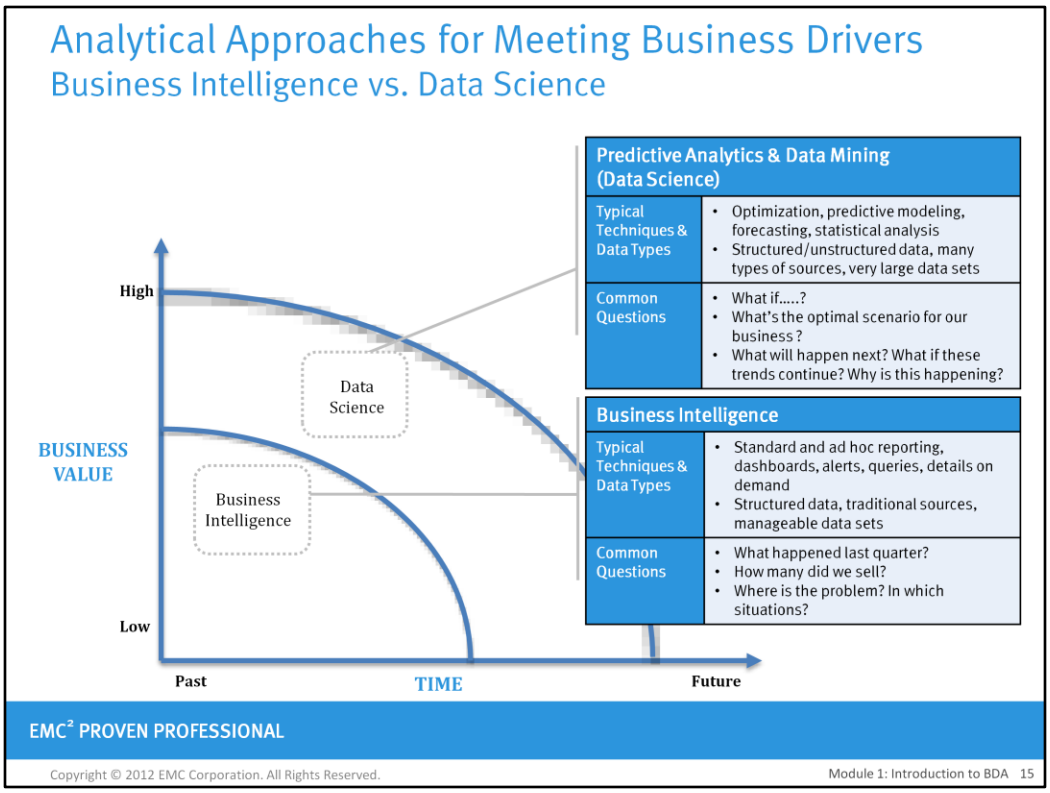


Here are the topics covered in this lesson.



Here are 4 examples of common business problems that organizations contend with today, where they have an opportunity to leverage advanced analytics to create competitive advantage. Rather than doing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these typical tasks.

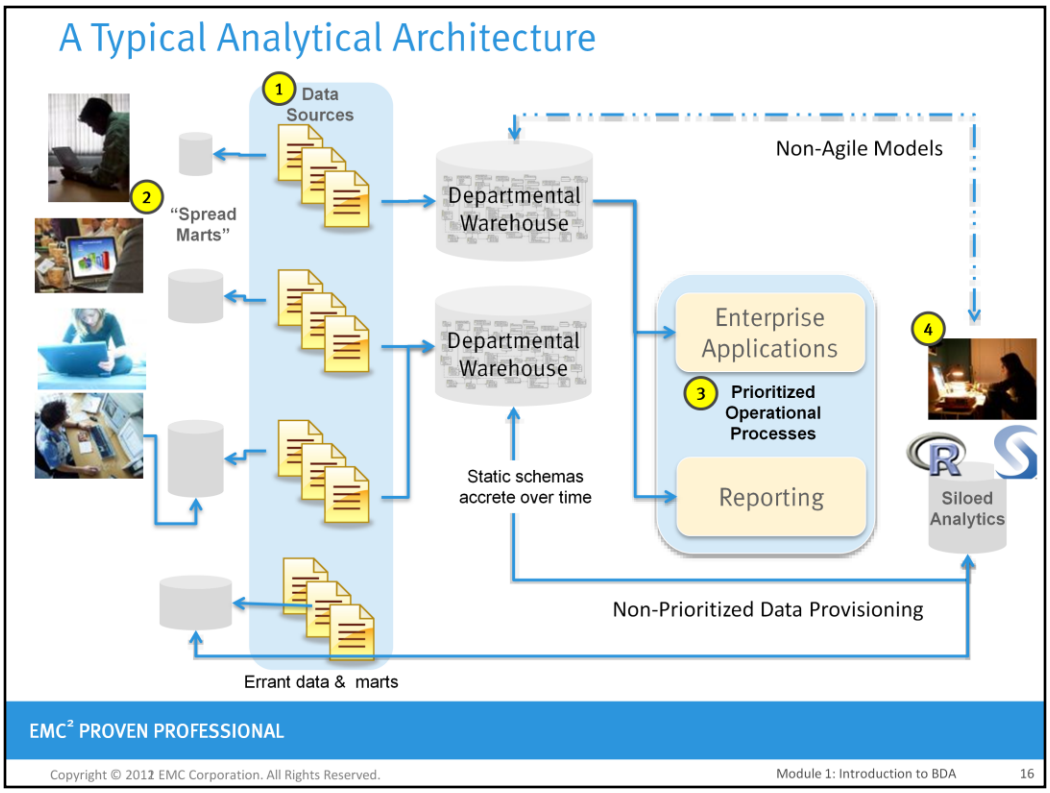
The first 3 examples listed above are not new problems – companies have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. **What's new is the opportunity to fuse advanced analytical techniques with big data to produce more impactful analyses for these old problems**. Example 4 listed above 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.



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 & 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 Mining are needed for organizations to meet these emerging business challenges successfully.



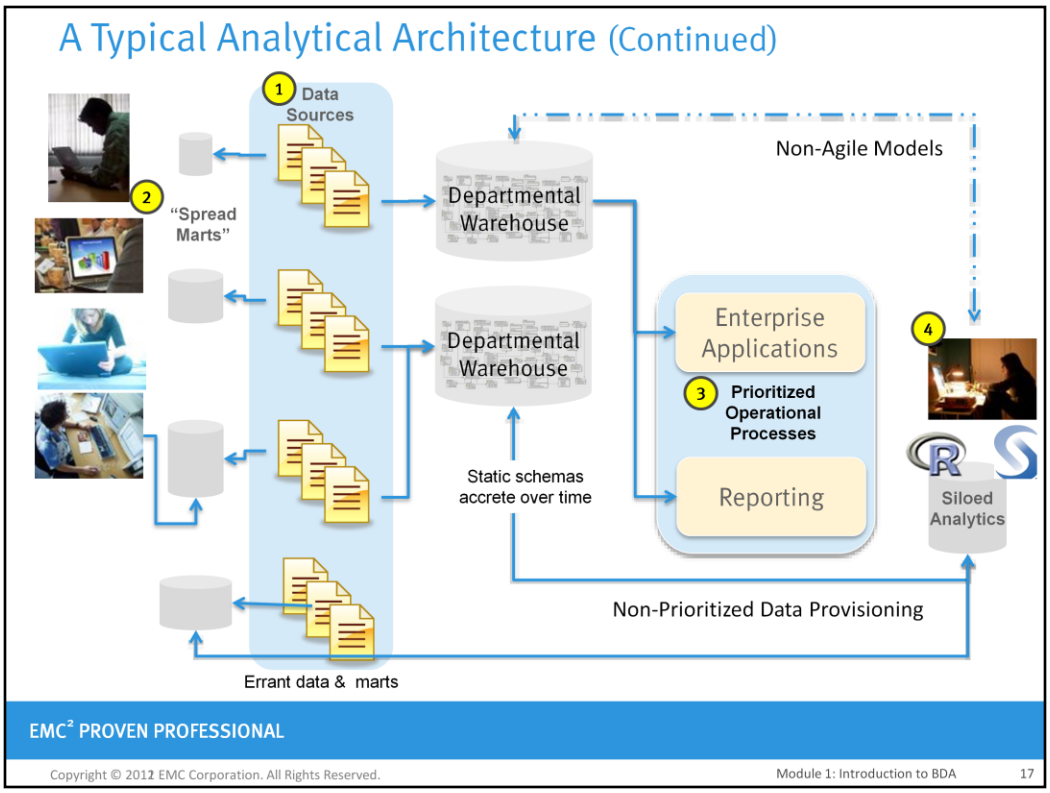
The graphic shows a typical data warehouse and some of the challenges that it presents.

For source data (1) to be loaded into the EDW, data needs to be well understood, structured and normalized with the appropriate data type definitions. While this kind of centralization enables organizations to enjoy the benefits of security, backup and failover of highly critical data, it also means that **data must go through significant pre-processing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.**

(2) As a result of this level of control on the EDW, shadow systems emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts do not have the same constraints for security and structure as the EDW does, and allow users across the enterprise to do some level of analysis. However, these one-off systems reside in isolation, often are not networked or connected to other data stores, and are generally not backed up.

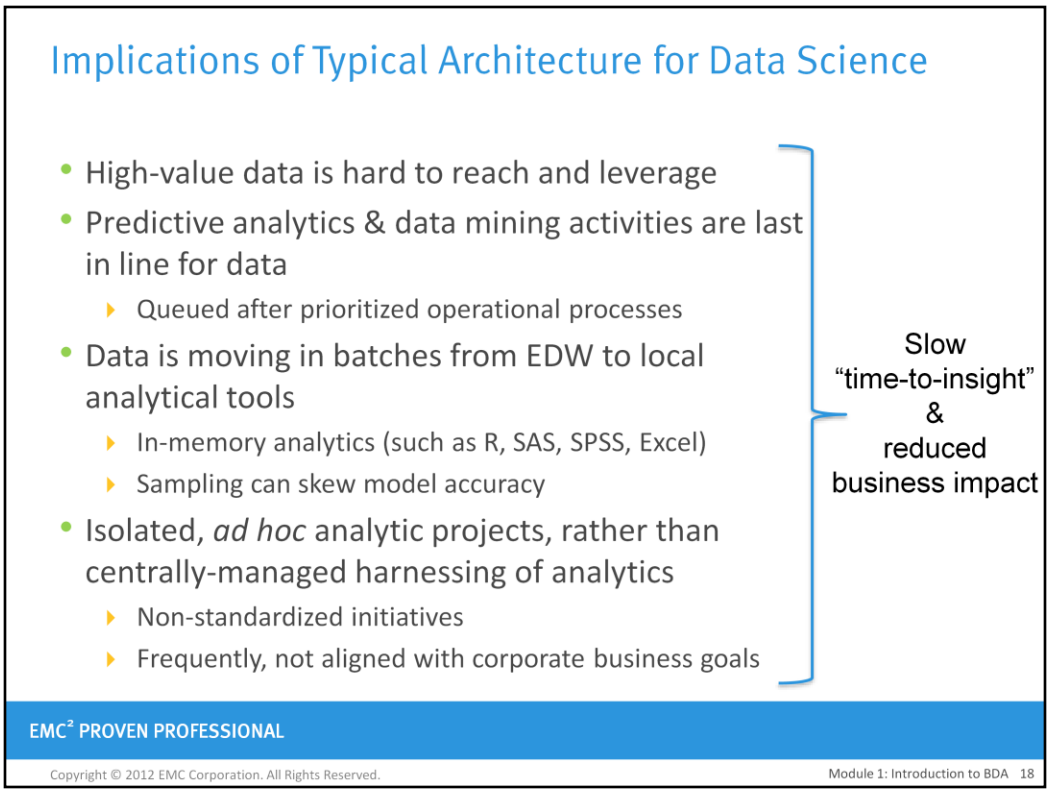
(3) Once in the data warehouse, data is fed to enterprise applications for business intelligence and reporting purposes. These are high priority operational processes getting critical data feeds from the EDW.

<Continued>

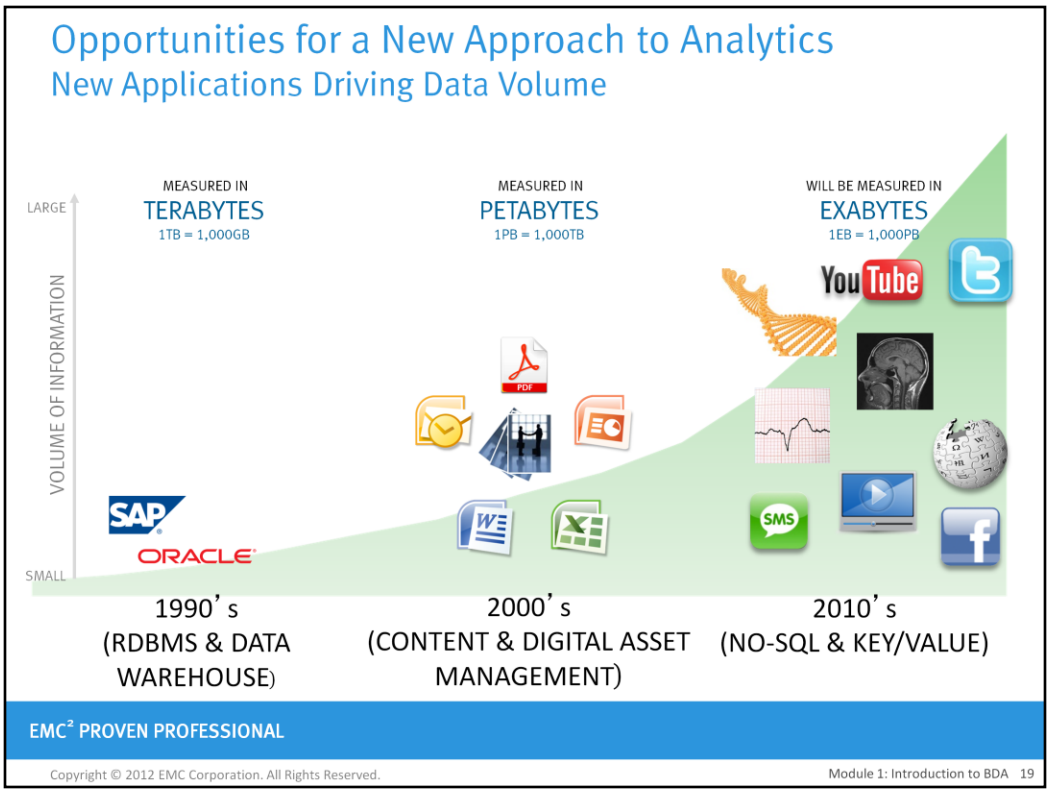


(4) At the end of this work flow, analysts get data provisioned for their downstream analytics. Since users cannot run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze offline in R or other local analytical tools. Many times these tools are limited to in-memory analytics with desktops analyzing samples of data, rather than the entire population of a data set. Because these analyses are based on data extracts, they live in a separate location and the results of the analysis – and any insights on the quality of the data or anomalies, rarely are fed back into the main EDW repository.

Lastly, because data slowly accumulates in the EDW due to the rigorous validation and data structuring process, **data is slow to move into the EDW and the schema is slow to change**. EDWs may have been originally designed for a specific purpose and set of business needs, but over time evolves to house more and more data and enables business intelligence and the creation of OLAP cubes for analysis and reporting. The EDWs provide limited means to accomplish these goals, achieving the objective of reporting, and sometimes the creation of dashboards, but generally limiting the ability of analysts to iterate on the data in an separate environment from the production environment where they can conduct in-depth analytics, or perform analysis on unstructured data.



Today's typical data architectures were designed for storing mission critical data, supporting enterprise applications, and enabling enterprise level reporting. These functions are still critical for organizations, although these architectures inhibit data exploration and more sophisticated analysis.



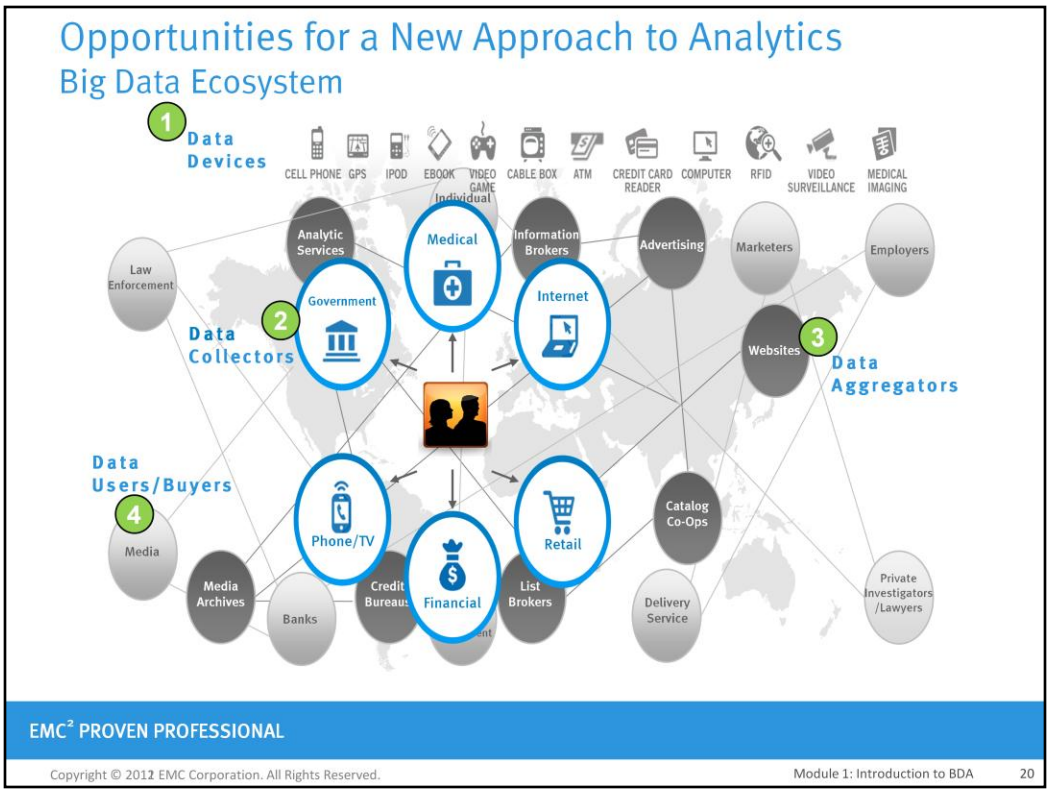
…..describe or refer to NO SQL and KVP

Everyone and everything is leaving a digital footprint. The graphic above provides a perspective on sources of big data generated by new applications and the scale and growth rate of the data. These applications provide opportunities for new analytics and driving value for organizations.

These data come from multiple sources, including:

- Medical Information, such as genomic sequencing and MRIs
- Increased use of broadband on the Web – including the 2 billion photos each month that Facebook users currently upload as well as the innumerable videos uploaded to YouTube and other multimedia sites
- Video surveillance
- Increased global use of mobile devices – the torrent of texting is not likely to cease
- Smart devices – sensor-based collection of information from smart electric grids, smart buildings and many other public and industry infrastructure
- Non-traditional IT devices – including the use of RFID readers, GPS navigation systems, and seismic processing

The Big Data trend is generating an enormous amount of information that requires advanced analytics and new market players to take advantage of it.



Key Concepts:

A) Significant opportunities exist to extract value from Big Data

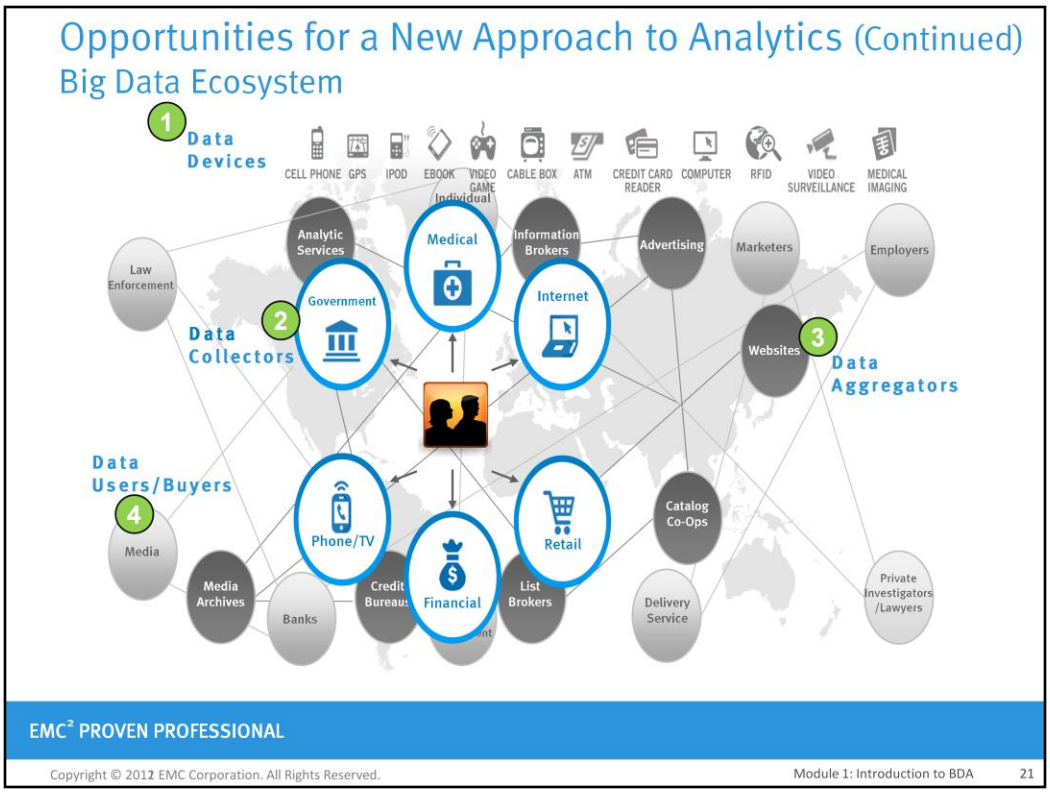
B) Entities are emerging throughout the new Big Data ecosystem to capitalize on these opportunities – from the Data Devices (1), Data Collectors (2), Data Aggregators (3), and Data Users / Buyers (4)

C) To accomplish this, these players will need to adopt a new analytic architectures and methods

Organizations and data collectors are realizing the data they can collect from you is really very valuable, and a new economy is arising around the collection of data because data is streaming off all of these computing and network devices. As this new economy continues to unfold, we are seeing the introduction of data vendors (such as Info Chimps), data cleaners to crowdsource the testing of machine learning techniques (such as Mechanical Turk and GalaxyZoo), as well as many who are value added providers repackaging open source tools in a simpler way to bring to market (such as Cloudera for Hadoop).

(1)Portrays Data Devices and the "Sensornet", which is collecting data from multiple locations and is continuously generating new data about this data. For each gigabyte of new data you create, a petabyte of data is created about that data. Take a simple example of someone playing video games in their home through their TV or their PC. Data is captured about the skill and levels attained by the player, but data is also created logging the dates and times and skill levels of people in the home playing certain games. In return, new games are offered to the player via recommendation engines for purchase and new levels of games are unlocked based on proficiency. Users can also purchase additional characters or items to enhance the games.

<Continued>

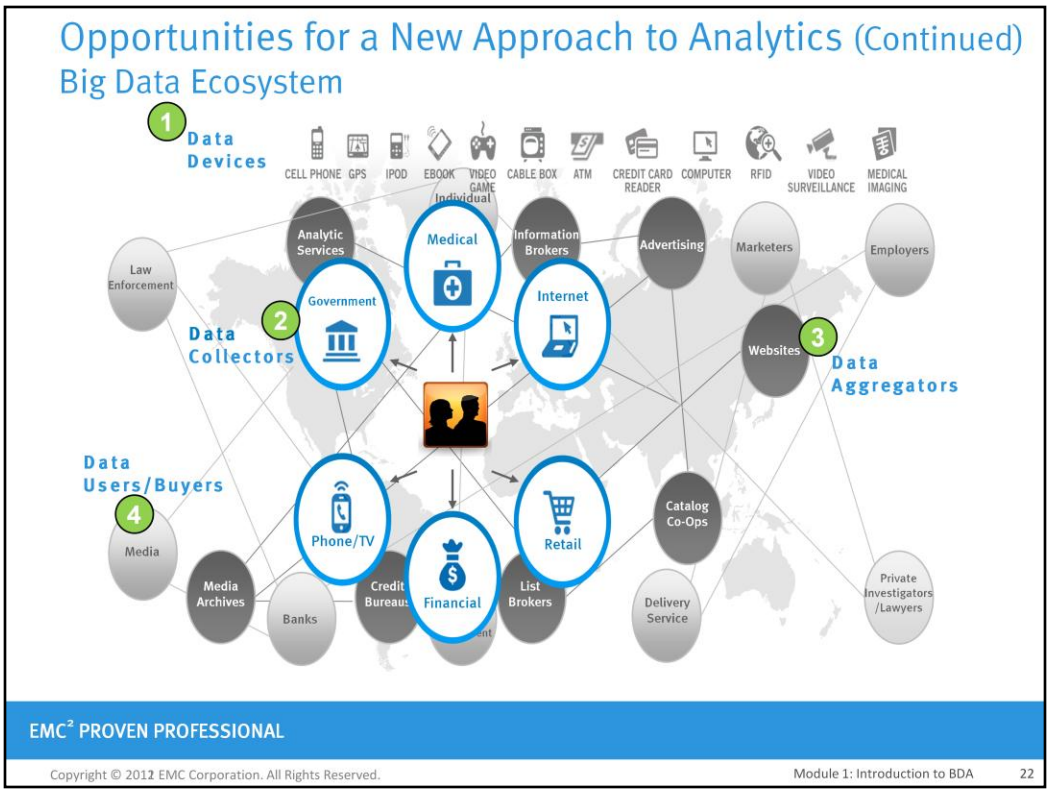


This information gets stored in the local consoles and also shared back with the manufactures who analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical profiles of specific kinds of users. Likewise, smart phones provide another rich source of data. In addition to the messaging and basic phone usage, they store and transmit data about your internet usage, SMS usage, and real time location. This can be used for analyzing traffic patterns, by scanning the density of smart phones in locations to track the speed of cars or the relative traffic congestion on busy roads. Lastly, consider shopping loyalty cards, which will record not just the amount you are spending, but the locations of stores you frequent, the sorts of products you tend to buy, the stores where you buy them, and the combinations of products you buy together, thus showing your shopping and travel habits, and your likelihood to be targeted for certain types of retail promotions in near real time.

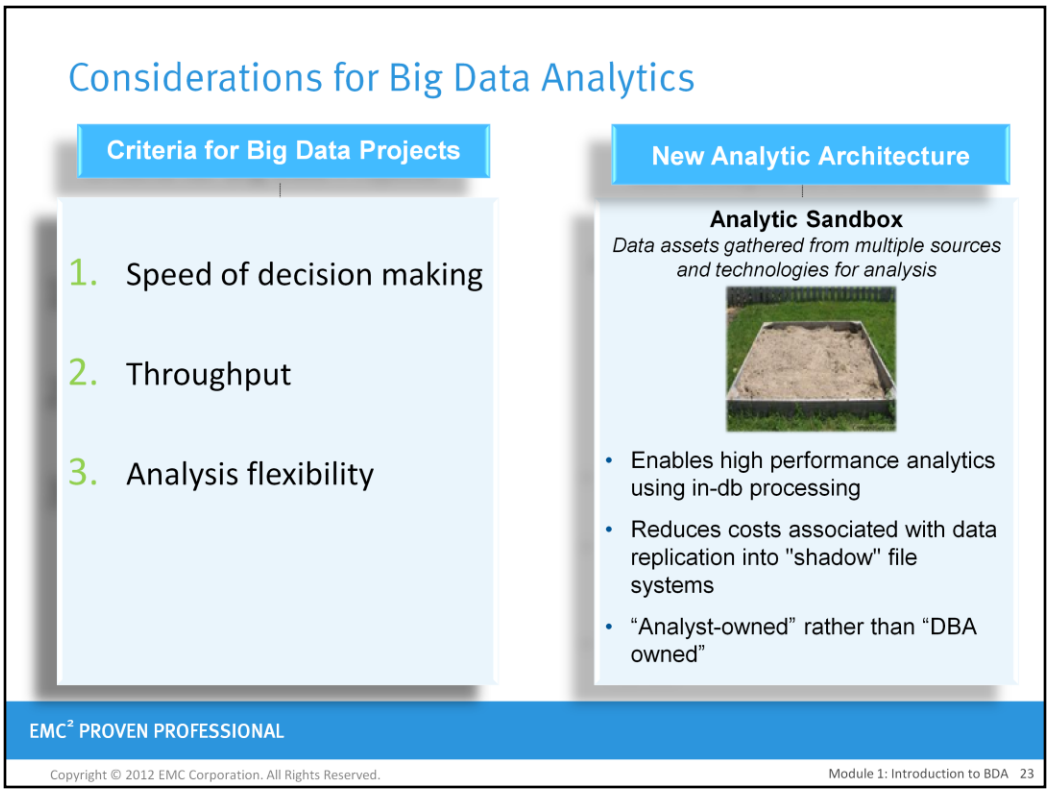
(2) Data Collectors (the blue circles) include entities who are collecting data from the device and users. This can range from your cable TV provider tracking the shows you watch, what you will and will not pay for to watch on demand and the prices you are willing to pay for premium TV content to retail stores tracking the path you take pushing a shopping cart with a Radio-frequency identification (RFID) chip in their store so they can gauge which products get the most foot traffic.

(3) Data Aggregators (the dark grey circles) make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things". These organizations compile data from the devices and usage patterns collected by government agencies, retail stores and websites. In turn, they can choose to transform and package these data as products to sell to list brokers, who may want marketing lists of people who may be good targets for specific ad campaigns or cross-sell opportunities to information brokers who store and aggregate data.

<Continued>

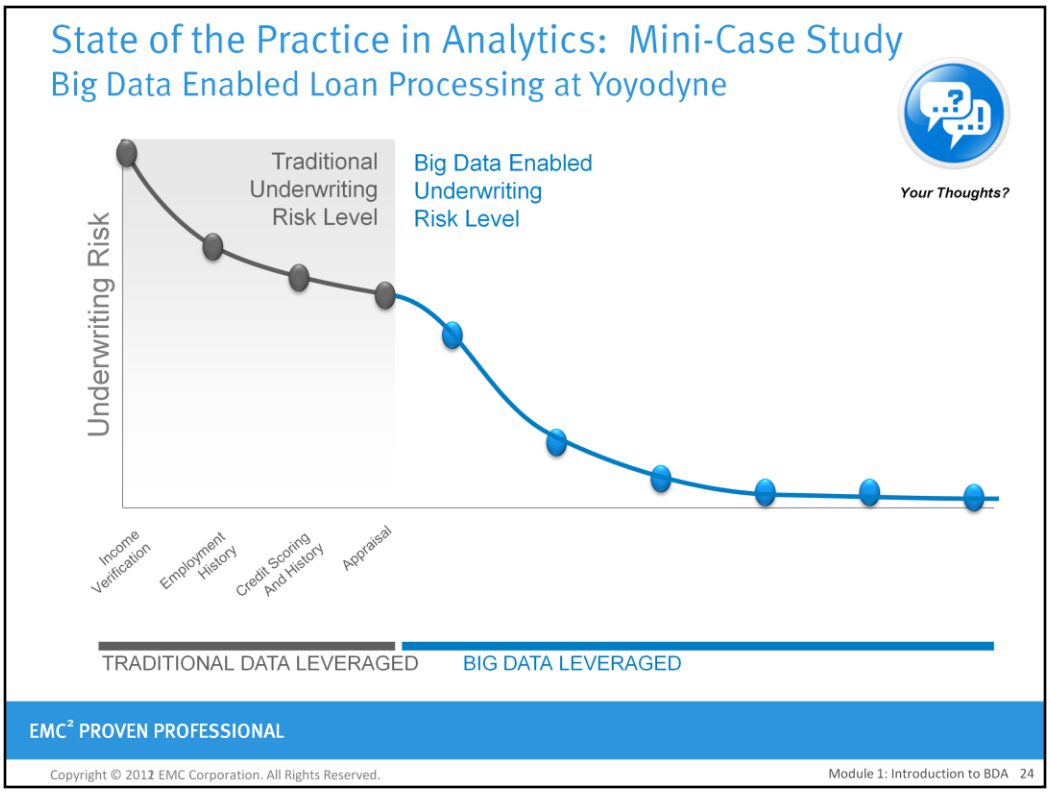


(4) At the outer edges of this web of the ecosystem are Data Users & Buyers. These groups directly benefit from the information collected and aggregated by others in the data value chain. For instance, retail banks may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit. To do this, retail banks may purchase data from a Data Aggregator showing demographics of people living in specific locations, those who seem to have a specific threshold of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit scores, and are searching the web for information about paying off debts or doing home remodeling projects. Obtaining data from these various sources and aggregators will enable a very targeted marketing campaign, which would not have been possible 5 or 10 years ago due to the lack of information. Another example would be using technologies such as Hadoop to conduct natural language processing and sentiment analysis on unstructured, textual data from social media websites to gauge the reaction to events such as presidential speeches (for example, to determine if the public reacted positively, negatively or neutral based on blogs and online comments) or natural disasters (for example, identify which areas the hurricane affects first, and how a hurricane moves based on which geographic areas are tweeting about it).



Big data projects carry with them several considerations that you need to keep in mind to ensure this approach fits with what you are trying to achieve. Due of the characteristics of big data, these projects lend themselves to decision support for high-value, strategic decision making with high processing complexity. The analytic techniques being used in this context need to be iterative and flexible (**analysis flexibility**), due to the high volume of data and its complexity. These conditions give rise to complex analytical projects (such as predicting customer churn rates) that can be performed with some latency (consider the **speed of decision making** needed), or by operationalizing these analytical techniques using a combination of advanced analytical methods, big data and machine learning algorithms to provide real time (requires high **throughput**) or near real time analysis, such as recommendation engines that look at your recent web history and purchasing behavior.

In addition, to be successful you will need a different approach to the data architecture than seen in today's typical EDWs. **Analysts need to partner with IT and DBAs to get the data they need within an analytic sandbox**, which contains raw data, aggregated data, and data with multiple kinds of structure. The sandbox requires a more savvy user to take advantage of it, and leverage it or exploring data in a more robust way.



The loan process has been honed to a science over the past several decades. Unfortunately today's realities require that lenders take more 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, including:

- Income data, such as pay 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 for which the loan is made (such as a home, boat, or car)

This model works but it's not perfect, in fact, the loan crisis in the US is proof that using only these data points may not be enough to gauge the risk associated with making sound lending decisions and pricing loans properly.

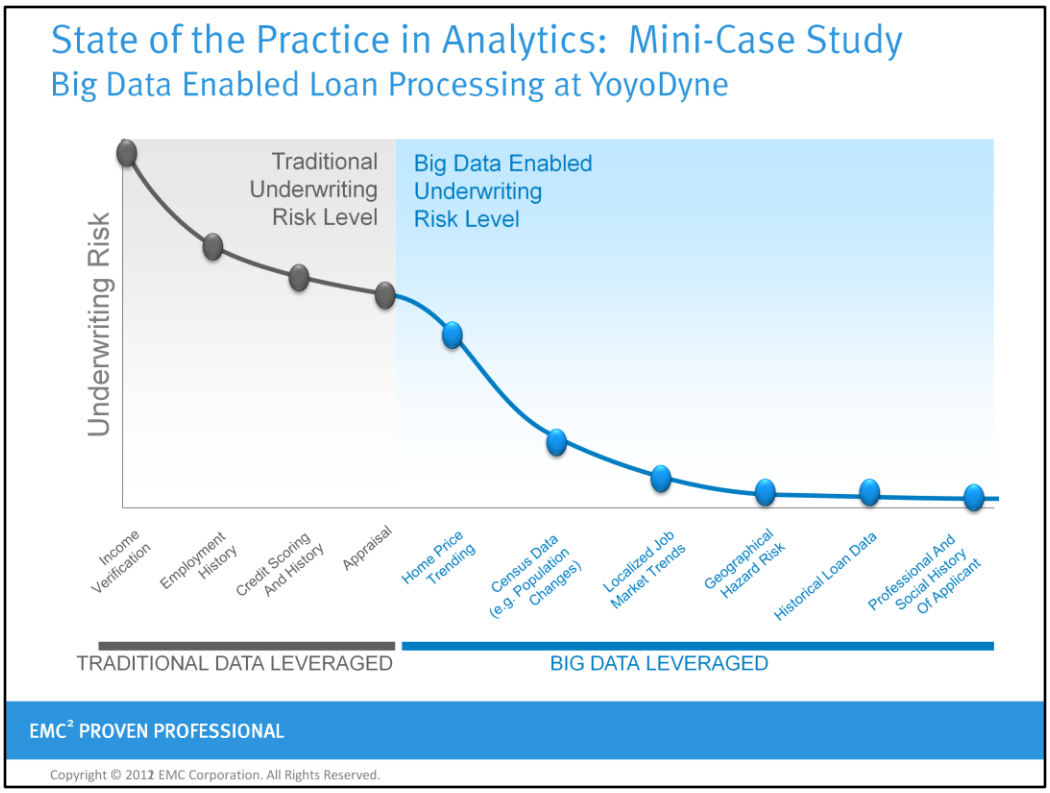
Case Study Exercise:

Objectives

1) Using additional data sources, dramatically improve the quality of the loan underwriting process
2) Streamline the process to yield results in less time

Directions

1) Suggest kinds of publicly available data (big data) that you can leverage to supplement the traditional lending process
2) Suggest types of analysis you would perform with the data to reduce the bank's risk and expedite the lending process



The loan process has been honed to a science over the past several decades

Unfortunately today's realities require that lenders take more 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; this includes:

- Income data – like 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 US is proof that using only these data points may not be enough to gauge the risk associate with loans

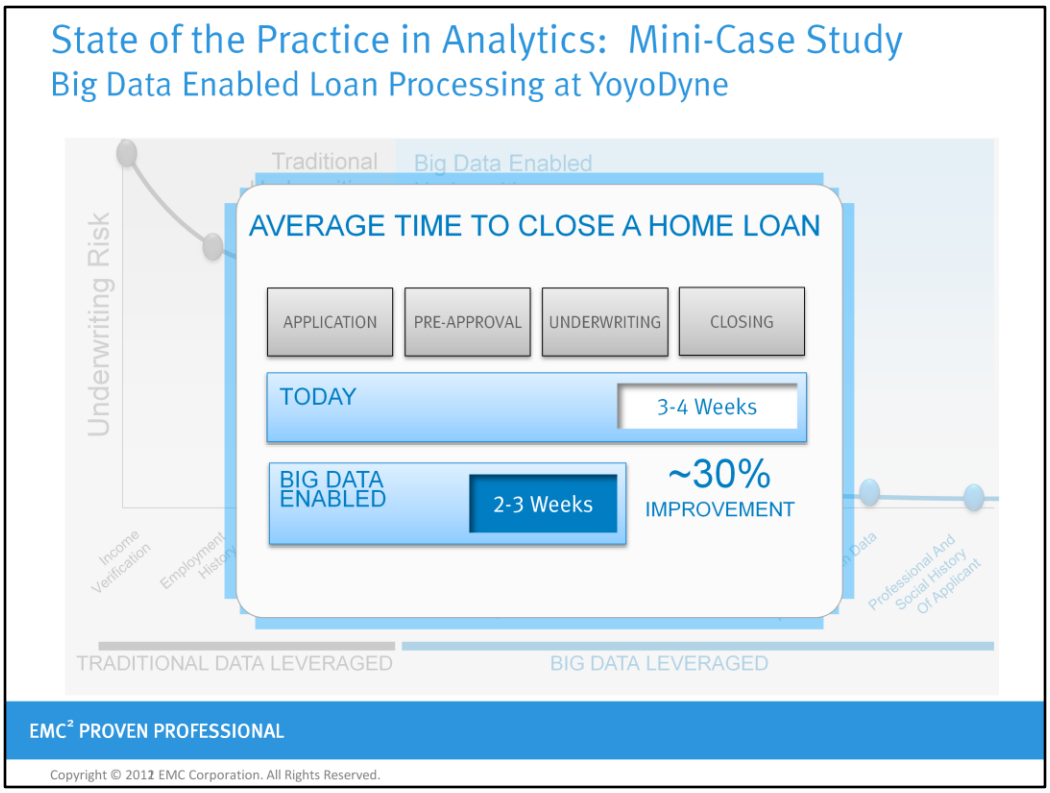
Using Big Data we can dramatically impact not only the quality of the loan underwriting process, but we can streamline the process to yield results in less time; data such as

- Home price trends from data sources like Zillow and eppraisal (eppraisal.com) can be mashed up with appraisal data to identify patterns or conflicting data
- Census data can be used to understand population migration trends and the potential impact on home values or geographic income variation
- Localized Job Trends to identify patterns in prosperity in a specific area
- Historic Loan Data from the lenders own records or purchased data from third-parties that may aggregate loan data from public records can be used to find patterns that point to more or less risky behavioral patterns
- Social and Professional History of Applications in order to correlate behavior in personal and professional life through sources like Twitter, Facebook, LinkedIn, etc. 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 be used to 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 3 to 4 weeks to 2 to 3 weeks a savings of over 30%. In some situations this benefit can cut the time to close in half if more data sources are available for analysis

Loan approvals are just the tip of the iceberg; far more transformative changes can be enabled with Big Data in



The loan process has been honed to a science over the past several decades

Unfortunately today's realities require that lenders take more 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; this includes:

- Income data – like 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 US is proof that using only these data points may not be enough to gauge the risk associate with loans

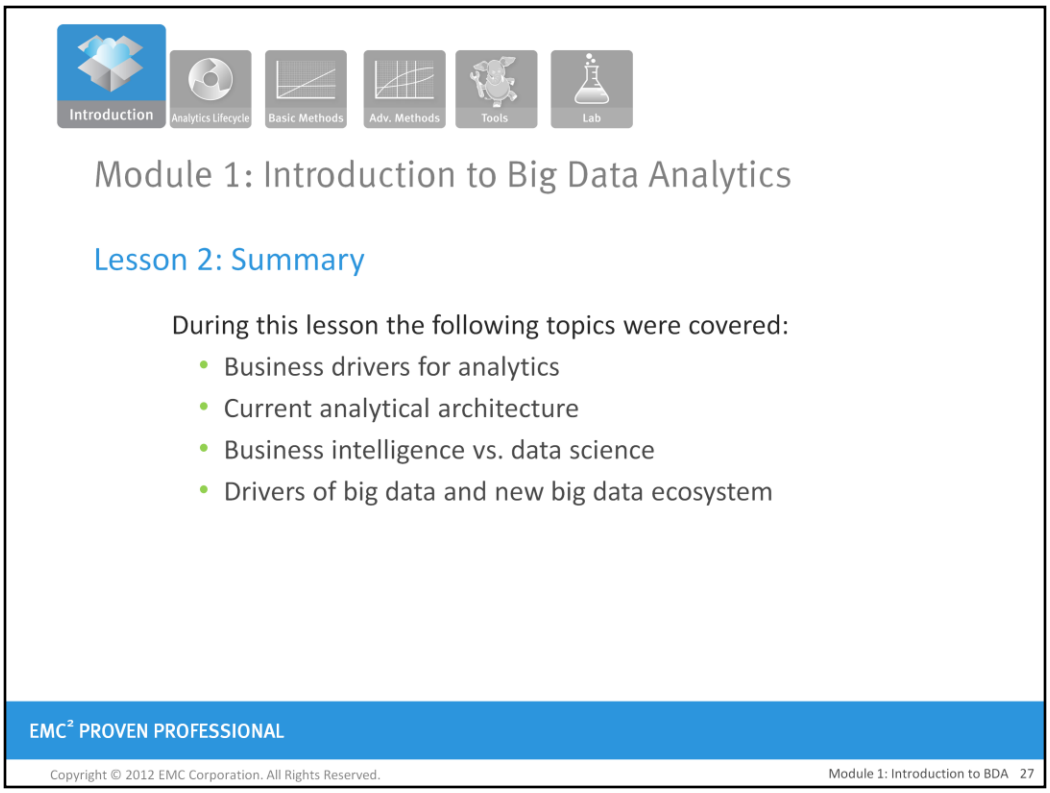
Using Big Data we can dramatically impact not only the quality of the loan underwriting process, but we can streamline the process to yield results in less time; data such as

- Home price trends from data sources like Zillow and eppraisal (eppraisal.com) can be mashed up with appraisal data to identify patterns or conflicting data
- Census data can be used to understand population migration trends and the potential impact on home values or geographic income variation
- Localized Job Trends to identify patterns in prosperity in a specific area
- Historic Loan Data from the lenders own records or purchased data from third-parties that may aggregate loan data from public records can be used to find patterns that point to more or less risky behavioral patterns
- Social and Professional History of Applications in order to correlate behavior in personal and professional life through sources like Twitter, Facebook, LinkedIn, etc. that may have an impact on the ability to pay or continue paying debts

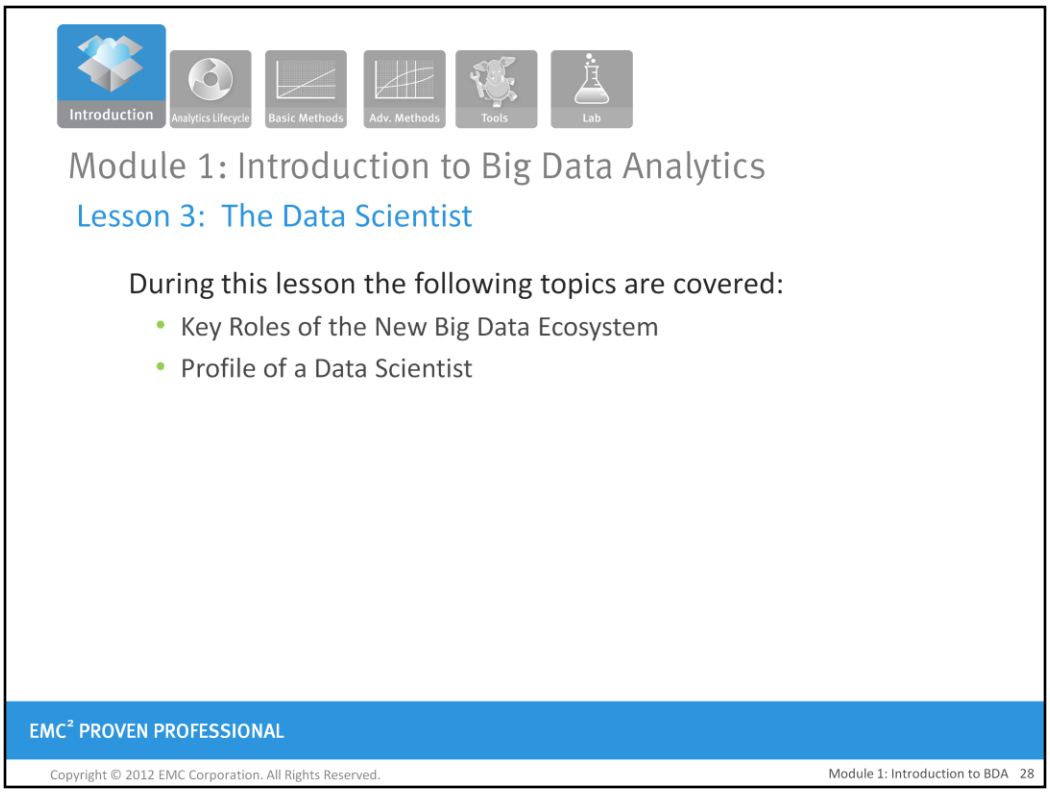
The end result is a much richer analysis and resulting insight which can be used to drive a process to either approve or deny a loan application with much lower risk than is possible today

The secondary benefit is that the approval process can be streamlined from an average of 3 to 4 weeks to 2 to 3 weeks a savings of over 30%. In some situations this benefit can cut the time to close in half if more data sources are available for analysis

Loan approvals are just the tip of the iceberg; far more transformative changes can be enabled with Big Data in



This lesson covered these topics.



This lesson covers the roles required to support the new big data ecosystem, including a profile of the Data Scientist.

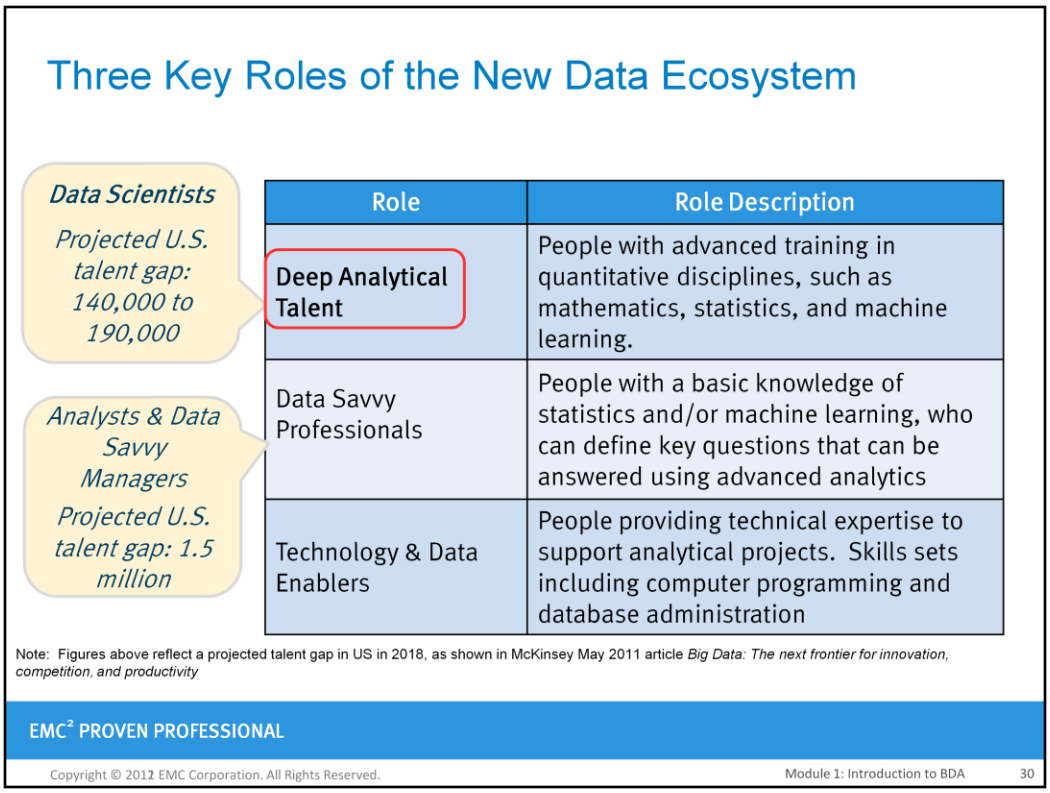
Skills Needed In the New Data Ecosystem

Your Thoughts?

- What new **skill sets** do you need to take advantage of the big data sets in the loan processing improvement case study?
- Do most large organizations have people with these **skill sets**?
- If so, **who are they**?

EMC² PROVEN PROFESSIONAL

In the space below, please note your responses to the questions.



The new data ecosystem driven by the arrival of big data will require 3 archetypical roles to provide services.

Here are some professions that represent illustrative examples of each of the 3 main categories.

Deep Analytical Talent

- Technically savvy, with strong analytical skills
- Combination of skills to handle raw data, unstructured data and complex analytical techniques at massive scales
- Needs access to magnetic, analytic sandbox

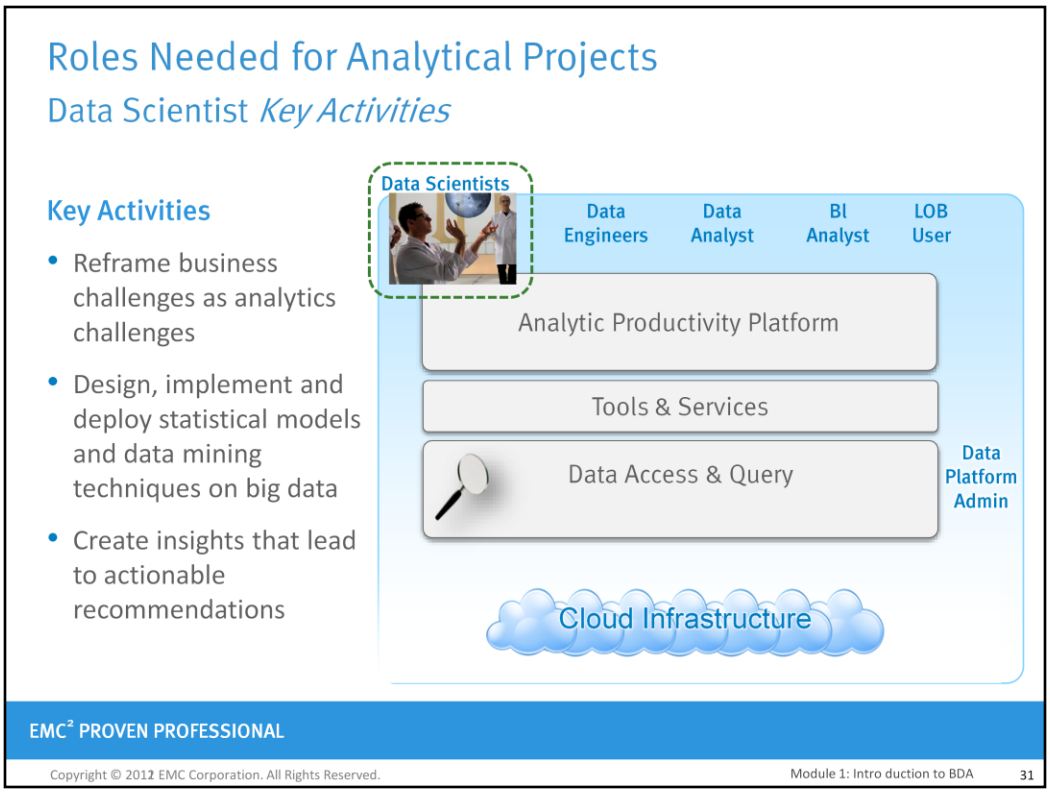
Examples of professions: Data Scientists, Statisticians, Economists, Mathematicians

Data Savvy Professionals

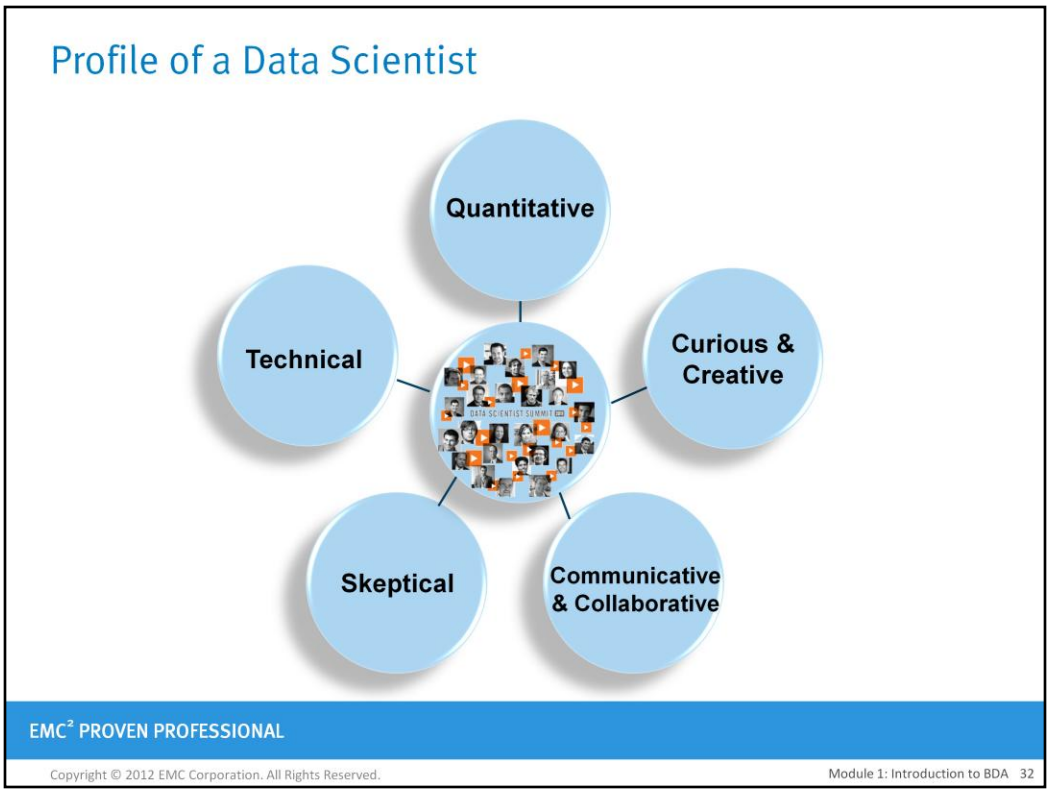
Examples of professions: Financial Analysts, Market Research Analysts, Life Scientists, Operations Managers, Business and Functional Managers

Technology & Data Enablers

Examples of professions: Computer programmers, database administrators, computer system analysts

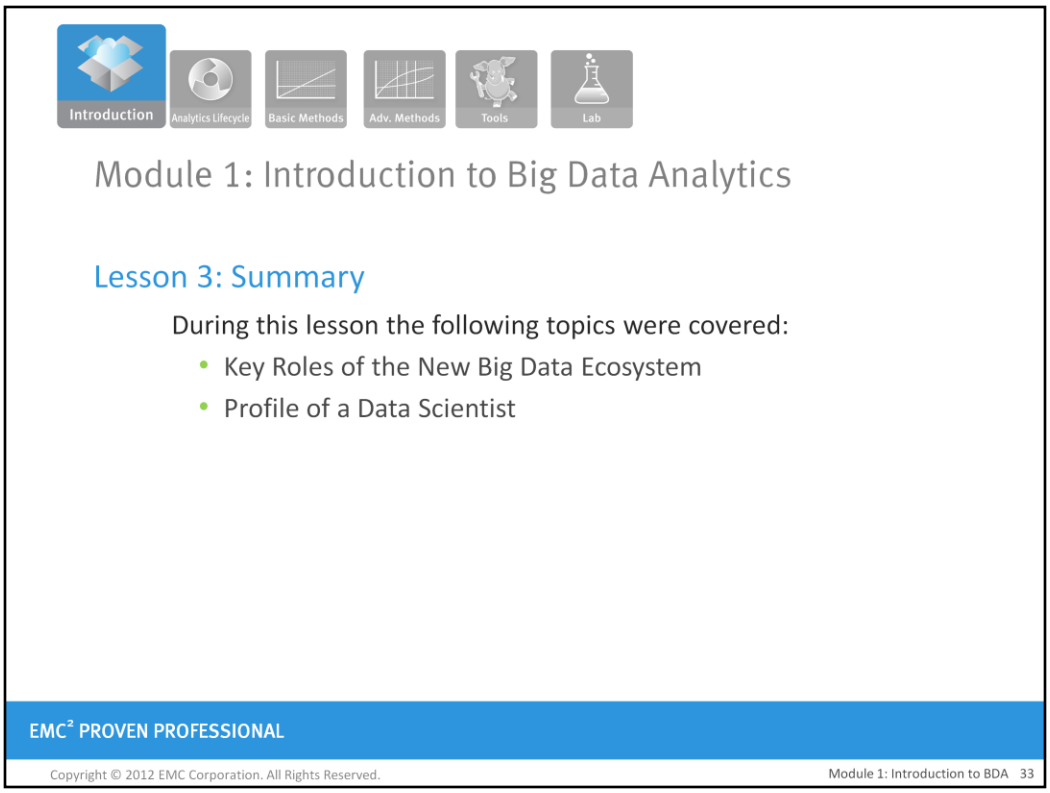


Here are several of the roles needed for typical analytical projects. Of note, the **Data Scientist shown combines several of the skill sets that were separated in the past into a single role**. Rather than having separate people for consultative aspects of the discovery phase of a project, a different person to deal with the end user in a line of business, another person with technical expertise and quantitative expertise, the Data Scientist is a combination of these aspects to help provide continuity throughout the analytical process.

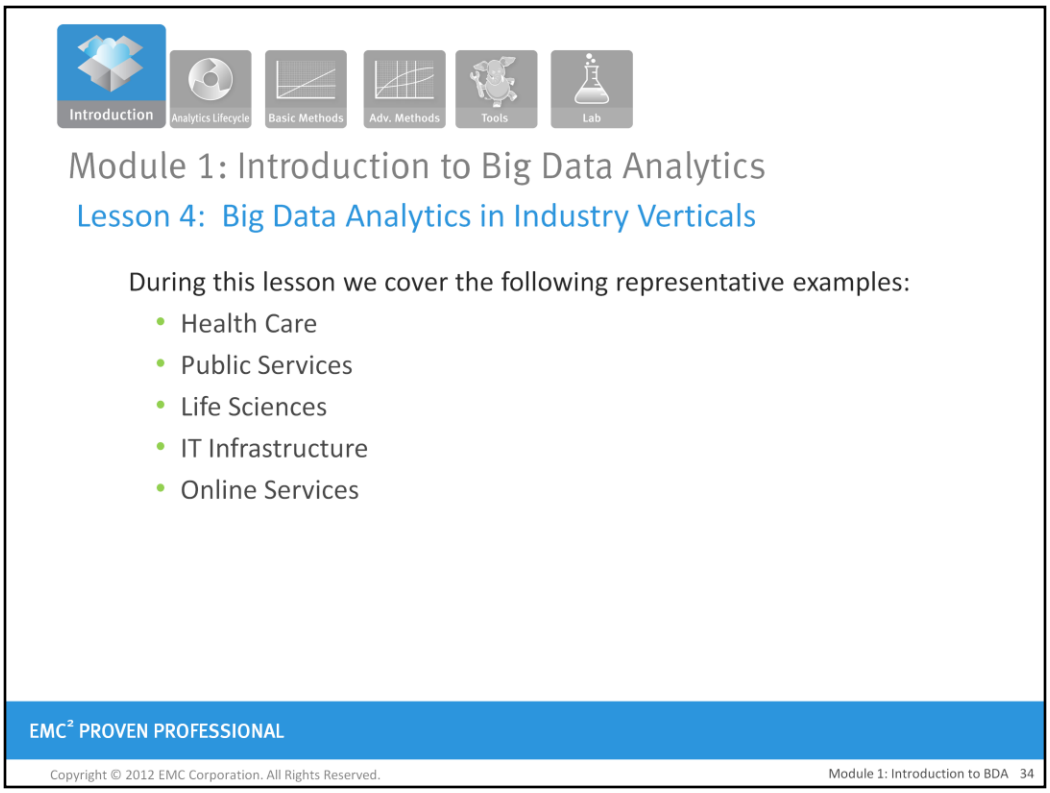


Here are 5 main competency and behavioral characteristics for Data Scientists.

1. **Quantitative** skills, such as mathematics or statistics
2. **Technical** aptitude, such as software engineering, machine learning, and programming skills.
3. **Skeptical**…..this may be a counterintuitive trait, although it is important that data scientists can examine their work critically rather than in a one-sided way.
4. **Curious & Creative**, data scientists must be passionate about data and finding creative ways to solve problems and portray information
5. **Communicative & Collaborative**: it is not enough to have strong quantitative skills or engineering skills. To make a project resonate, you must be able to articulate the business value in a clear way, and work collaboratively with project sponsors and key stakeholders.



This lesson covers these topics.



These examples are taken from the Data Hero awards at EMC World 2010.

Big Data Analytics: Industry Examples

1. Health Care
   - Reducing Cost of Care
2. Public Services
   - Preventing Pandemics
3. Life Sciences
   - Genomic Mapping
4. IT Infrastructure
   - Unstructured Data Analysis
5. Online Services
   - Social Media for Professionals

EMC² PROVEN PROFESSIONAL

1 Big Data Analytics: *Healthcare*

Situation

- Poor police response and problems with medical care, triggered by shooting of a Rutgers student
- The event drove local doctor to map crime data and examine local health care

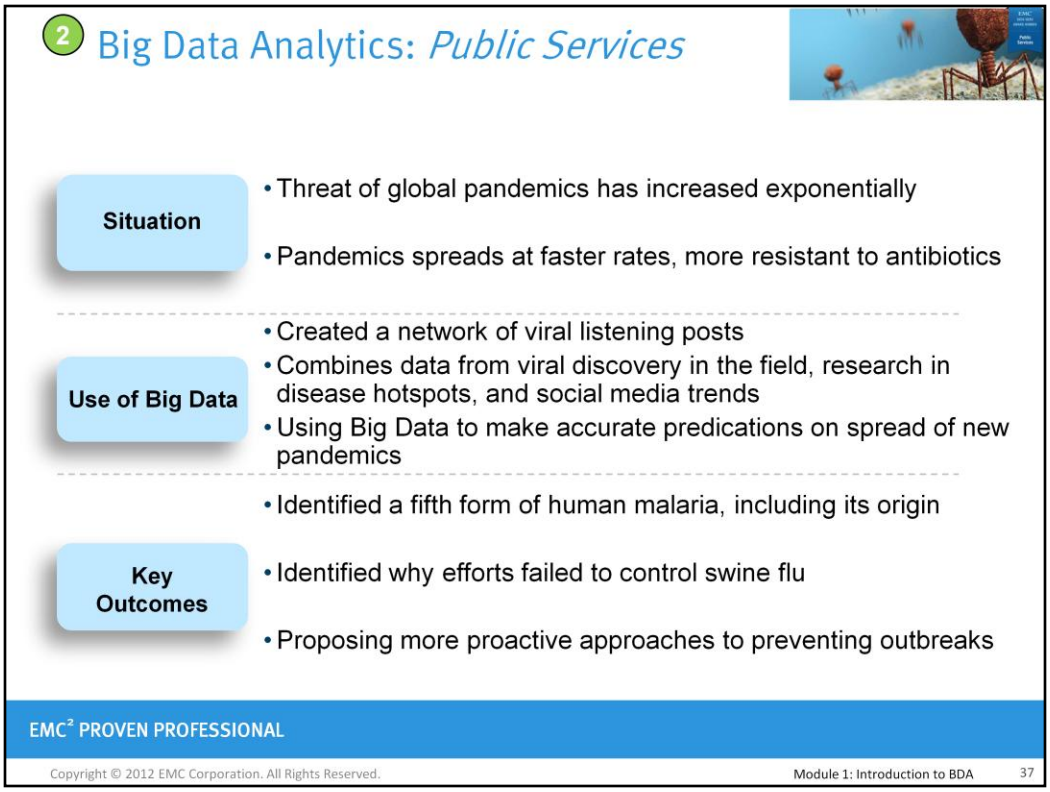
Use of Big Data

- Dr. Jeffrey Brenner generated his own crime maps from medical billing records of 3 hospitals

Key Outcomes

- City hospitals & ER's provided expensive care, low quality care
- Reduced hospital costs by 56% by realizing that 80% of city's medical costs came from 13% of its residents, mainly low-income or elderly
- Now offers preventative care over the phone or through home visits

EMC² PROVEN PROFESSIONAL



This is the standard format we will use for each representative example.

3 Big Data Analytics: *Life Sciences*

Situation

- Broad Institute (MIT & Harvard) mapping the Human Genome

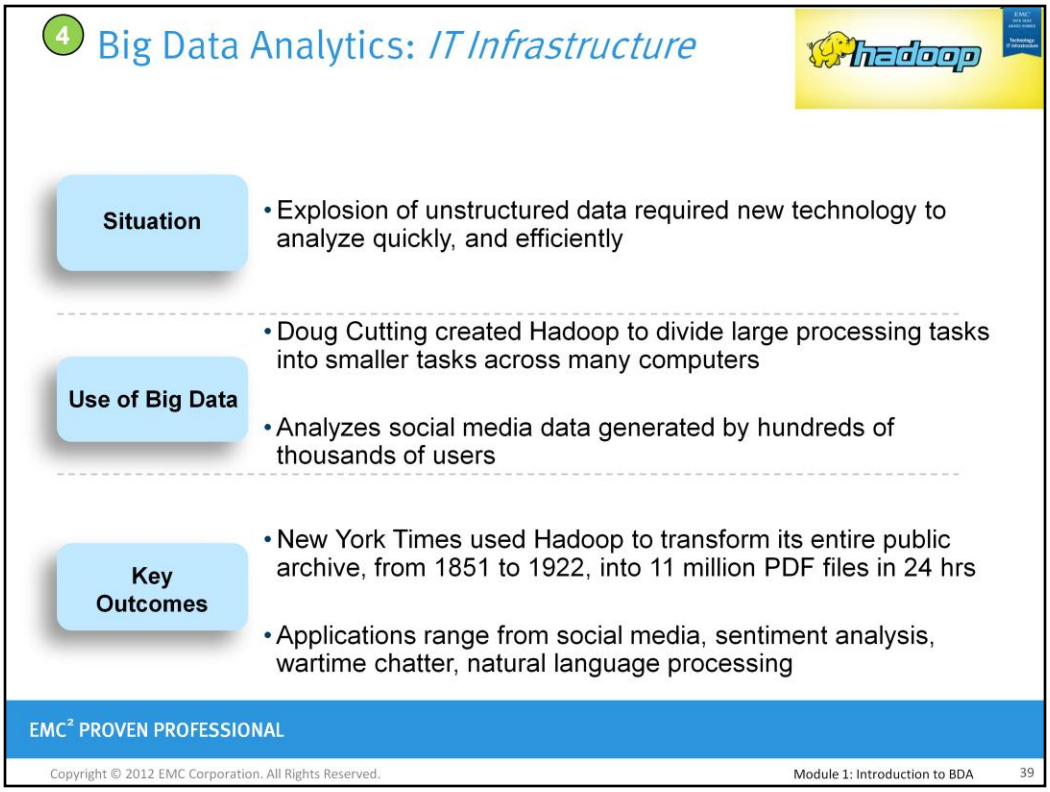
Use of Big Data

- In 13 yrs, mapped 3 billion genetic base pairs; 8 petabytes
- Developed 30+ software packages, now shared publicly, along with the genomic data

Key Outcomes

- Using genetic mappings to identify cellular mutations causing cancer and other serious diseases
- Innovating how genomic research informs new pharmaceutical drugs

EMC² PROVEN PROFESSIONAL



• Check http://wiki.apache.org/hadoop/PoweredBy for examples of how people are using Hadoop

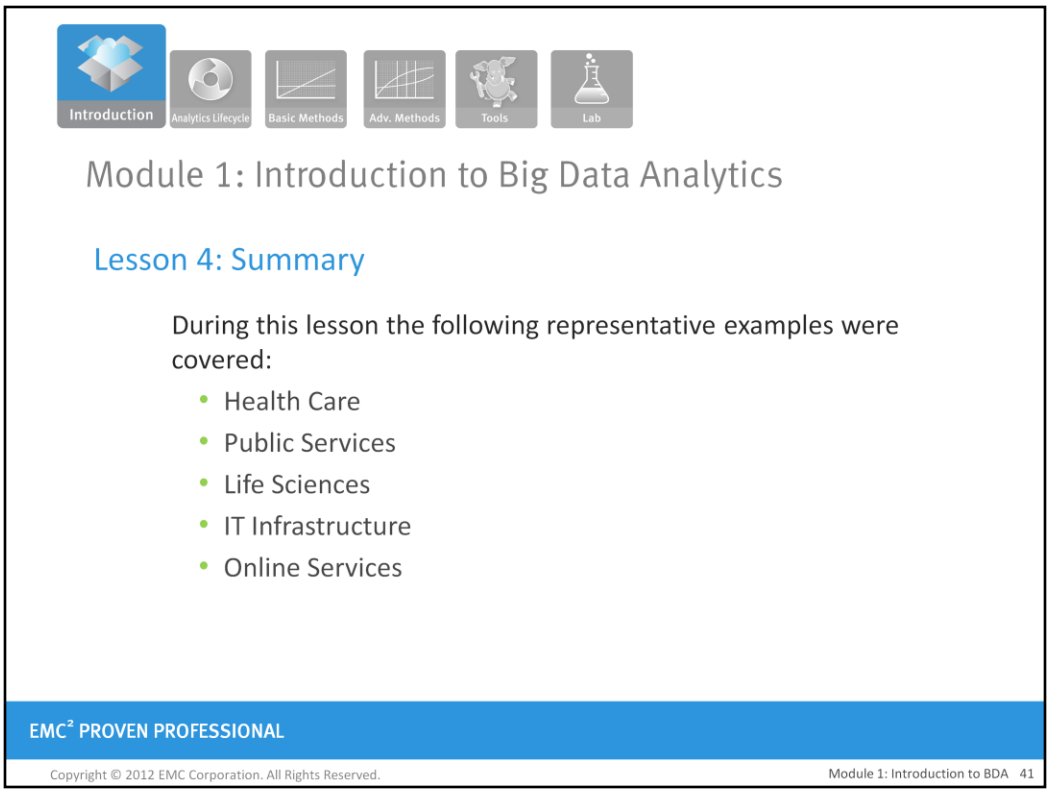
• Check this article on the large scale image conversion: http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/.

• Check this for an ad for a 'computer' from 1892…http://query.nytimes.com/mem/archive-free/pdf?res=9F07E0D81438E233A25751C0A9639C94639ED7CF

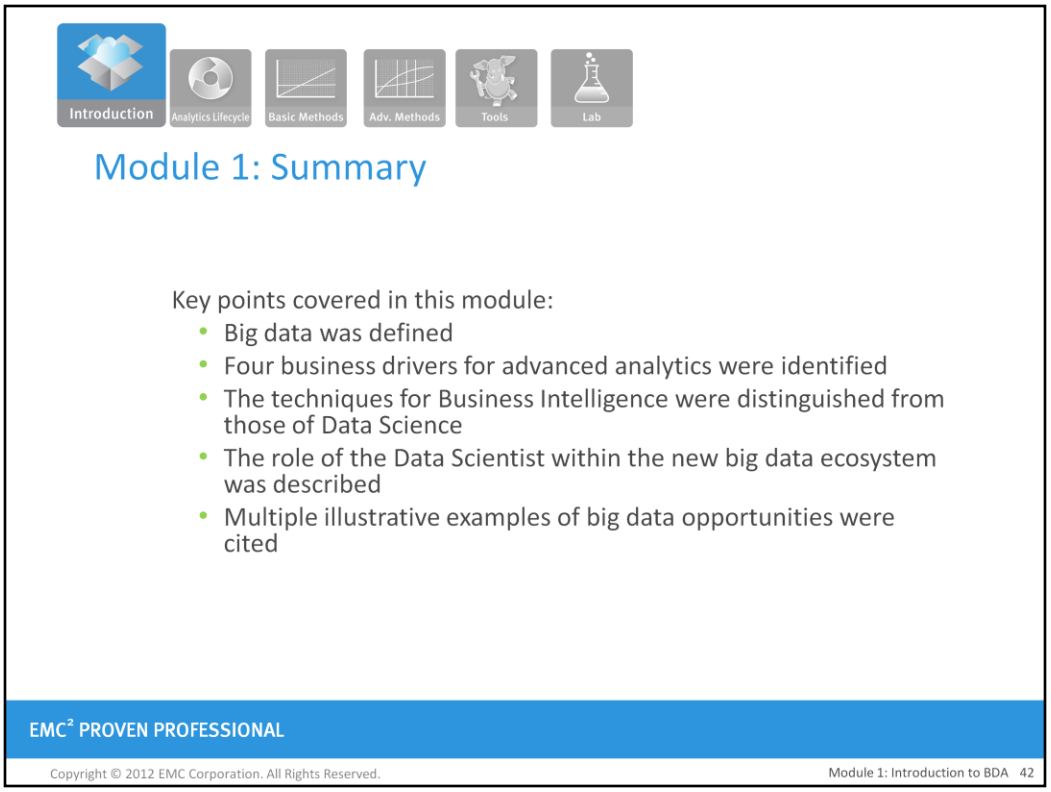
5 Big Data Analytics: *Online Services*

Situation	• Opportunity to create social media space for professionals
Use of Big Data	• Collects and analyzes data from over 100 million users • Adding 1 million new users per week
Key Outcomes	• LinkedIn Skills, InMaps, Job Recommendations, Recruiting • Established a diverse data scientist group, as founder believes this is the start of Big Data revolution

EMC² PROVEN PROFESSIONAL



This lesson covered these representative examples.



This summarizes the key points covered in this module.