WHITE PAPER | 2015 | Emcien

REDEFINING ANALYTICS

This White Paper does the following:

- Examines current and emerging technologies.
- Proposes that rather than search through data, organizations need the ability to automate the process of analyzing, visualizing, and leveraging the insight within their data.
- Introduces an algorithmic approach that provides an efficient, sustainable, automated way to analyze data, detect patterns, and discover insights hidden within that data.

WWW.EMCIEN.COM

emcien

Table of Contents

emcien

Summary

The modern analyst's relationship today is with data. Today's most popular tools allow analysts and data scientists to play in a sandbox of data, but the explosion of data renders old tools and methodologies slow and difficult to use. In spite of this, the industry's focus is on making those same tools perform the same operations, but faster. More powerful processors and faster memory execute queries faster, but hardware is no longer the bottleneck in data. Instead of treating big data like small data, analysts must soon begin to leverage automation to understand the patterns that define the data.

Still, big data holds the promise of greater insight, competitive advantage, solutions to our toughest problems, and the possibility of solving problems yet to be imagined. These breakthroughs will come, not from the way that we store and search through data, but from software that automates the most difficult and time-consuming aspects of data analysis. While hardware is an important component, this paper focuses on the algorithms that will deliver the insights organizations need.

This paper will examine the current and emerging technologies, describing the benefits and drawbacks of each to create an overview of the big data landscape. It will also propose that on top of today's store and query approach, an automation layer will allow analysts to work directly with the patterns and connections that make up the data set.

The Promise of Data and the Search for Insight

Why is the world obsessed with data? Because the promise of data is insight. We have become exceptionally good at collecting data, and as the cost of storage has dropped companies are now drowning in that data. Analysts face an expanding gap as the amount of growing data far outpaces the human capacity to process it.

A New Way Forward

Humankind has always possessed a love for data. We can do remarkable things with data and have built tools to collect, store, manipulate, chart, report, and visualize it. Data can change the way we see the world and how we interact with it, but the world of data is changing. Data used to be scarce, and tiny bits of it were extremely valuable. Today, the size and variety of data are exploding. Organizations are recognizing the value of data and collecting increasingly more of it. Now manufacturers have realized the potential value for data and are adding sensors to machines, vehicles, and devices of all kinds. Shippers are collecting GPS and RFID data to better understand their businesses, automobiles can return log data to insurance companies and manufacturers. Businesses across industries are looking for new sources of data. Searching this data for tiny bits of it can still be valuable, but with so much data, finding the valuable bits can be difficult.

The growth of data demands that organizations change the way they interact with data. In the past, small amounts of data were sufficient to reveal answers and insights. By necessity, the data that had to be recorded by people could be understood by people. Modern technology allows automatic collection and aggregation of more data from multiple sources without much human input. Storage and processing technologies allow for the cheap and convenient holding of that data. We can pull data from stores across the United States or across the world in volumes that can be measured not against the Library of Congress, but Libraries of Congress. Data is everywhere, and all of this data comes with commensurate amounts of noise. Now analysts are searching for the important bits from more and more noise. In torrents of information, that task is overwhelming.

It's not uncommon for organizations to have analyzed less than ten percent of their data. The reason? Most technologies approach larger and more complex data with faster versions of traditional methods. What is needed is a completely new approach.

In this new approach, the only way to separate valuable data from noise is through automation. And automation requires algorithms that do the work of finding and quantifying patterns in data, allowing the analyst to work with the relationships in data, rather than the data itself. The role of the analyst will be changed to directing the algorithms and selecting the points of interest from the results of the analysis.

The Limitations of Search and Query

Current data analysis inevitably involves querying a database, or often several databases. Analysts mash up data across silos to discover the connections between data points. Marketers aggregate customer demographics, purchase data, and social media data, while purchasers aggregating supplier data with procurement and pricing data.

Data mashups often create a lot of columns, becoming unwieldy due to the number of rows and columns. High volume transactional data typically contains many rows or records. However, the millions of records is not the problem. The depth of the data, the number of rows, impacts processing time in a linear fashion and can be reduced with faster processors or parallel computing. Executing a query is simple enough. The problem is the width of the data. The complexity of a query increases exponentially based on the number of columns.

A simple query might ask for specific values within a subset of columns. The real question becomes, “How many queries will it take to answer even one of these questions?” And then, “How many questions must be asked before arriving at a satisfactory answer?”

A database with 100 columns and 6 choices per column yields more possible queries than there are atoms in the universe. Approaching complex data sets with queries will never result in a complete analysis, making laboring through thousands of queries to uncover a single answer inefficient. The real challenge of data analysis today is formulating the right queries through human intuition and a deep knowledge of the data. In order to get the right answer, one must already know the right questions to ask.

Landscape of Existing Analysis Tools

A high level categorization of data tools is critical to understanding the state of analytics. Over 85% of all data is unstructured. However, most existing tools are designed to analyze highly structured data.

Statistical Tool Kits

The history of statistical analysis has been to make inferences from samples of data, especially when data is scarce. With big data, however, scarcity is not the problem. Traditional statistical methods have significant limitations when analyzing big data:

1. Statistical methods break down as dimensionality increases.
2. In unstructured data, dimensions are not well defined.
3. Attempts to define dimension for unstructured data result in millions of dimensions.

Some vendors have created user-friendly analytics packages that incorporate traditional statistical practices with improvements in ease of use and extended capabilities.

Data Mining

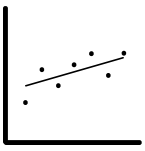
Data mining is a catchall phrase for a broad range of methods. Essentially, data mining consists of sifting through large amounts of data in attempt to find useful information. The name implies digging through tons of data to uncover patterns and relationships.

The primary limitation of data mining is that most forms of data mining require that the analyst knows what to look for. In classification and clustering analysis for example, the analyst is trying to find instances of known categories, such as people who have a high probability of defaulting on their mortgages. In anomaly detection, the analyst is looking for instances that do not match the known normal patterns or known suspicious patterns.

Machine Learning

Machine learning is a broad classification of algorithms that can learn from data. The approach to data analysis with machine learning typically involves inputting a training data set to classify the desired results before beginning the analysis.

Deep Learning



Deep learning is a relatively new field where a combination of hardware and new software form artificial neural networks to function more like the human brain than a traditional computer. These machines can be taught to recognize more complex concepts like images, sounds, and ideas. Efforts in deep learning have yielded impressive academic results, but deep learning projects remain difficult to conduct and out of reach for many organizations.

Data Visualization

Data visualization is the creation and study of the visual representation of data. Visualization tools help people comprehend and interpret data, and these tools are increasingly capable of dynamic interpretation of data. Visualization is a crucial tool in relating large volumes of data in a way that people can consume easily.

While visualization tools are helpful in conveying ideas about data, they rely on human intuition to extract insight and knowledge. These visuals are also limited in their ability to focus on two or three dimensions before the amount of information is overwhelming. The most common limitation of visualization is that while it is a good test for small samples, it is not a sustainable method to gain insight into large volumes of higher dimensionality data.

Business Intelligence & Reporting

Business Intelligence (BI) is a catchall phrase for reports created from a database. These are typically canned reports constructed on metrics geared towards specific users. References to “Analytics” frequently describes the computations performed for reporting. Many tools have added computational features to these reports, and many BI tools are now described as analytics or advanced BI. But at its core, BI was created as a simple way to extract data from the database. While it continues to serve that purpose, it remains focused on numeric data and reports guided by IT specialists.

Graph Databases

Graph databases are database structures comprised of heavily associated data. A graph database is structured so entities are recorded as a node on a graph and assigned edges connecting to another element, or node, on the graph.

Because each edge describes a node’s relationship with other nodes, the structure makes querying these interconnected databases much faster. These properties make graph databases a compelling approach to storing highly connected data sets. The format allows for faster searches, but does not reveal any insights without a data expert crafting the queries.

Emergence of Big Data Tools

Because big data includes data sets with sizes beyond the ability of common software tools to capture, curate, manage, and process within a tolerable time period, new technologies are emerging to address the challenges brought on by these large quantities of data. These technologies can be categorized into two groups: Hadoop-based solutions and In-Memory based solutions.

Hadoop and Hadoop-based Tools

While Hadoop is not an analytics tool, it is often mistaken for one. Apache Hadoop is an open-source software framework that supports data-intensive distributed applications running on large clusters of commodity hardware.

Hadoop aims to lower costs by dealing with data distributed across many inexpensive servers and processors. The software can help speed up certain calculations by sending queries to multiple machines at the same time. This has been a great advancement for querying large data sets. By breaking big tasks into smaller ones, they can be run in parallel to gain speed and efficiency. By operating on common hardware, these distributed systems allow Hadoop-style infrastructures to grow cheaply and easily as requirements increase. The technology has spawned new start-ups, including Hortonworks Inc. and Cloudera Inc., which help companies implement their own Hadoop implementations.

Hadoop makes it easy to store and manage large amounts of data, but doesn't provide new abilities to understanding information hidden within the data. The Hadoop ecosystem includes its own implementations of analysis technologies, including querying and machine learning. While Hadoop and similar solutions have democratized big data storage, implementations have fallen behind the hype. While Hadoop is easy and inexpensive to implement, finding the resources to ensure continued success means that quite often Hadoop installations become simply a mode of storage.

One result is that in-memory databases are gaining attention in an attempt to come closer to the goal of real-time business processing.

In-Memory Appliance

While traditional storage methods rely on disk storage, in-memory database solutions have delivered significant speed advancements using main memory, or RAM. An example is SAP's HANA (High Performance Analytical Appliance). In-memory storage has also been touted as the future database of big data.

The primary limitations of in-memory are cost and size. The cost of RAM continues to fall, but there is still a significant limit to the amount of data that can be held in memory. Once the investment has been made to build an in-memory system to scale, an analyst must then query a much larger, and likely much more diverse set of data. In-memory offers speed, but does not change the core processes of analysis.

Limitations of Existing Tools

The overwhelming shortcoming of these methods is that they are guided by queries. Big data offers an infinite number of queries. To date, advances in big data technologies have focused on either making those queries work faster or teaching machines to query a data set. Every tool approaches analytics from a mindset of asking questions of the data.

Although search remains the go-to information access interface, the reliance on search must be reconsidered. A new type of information-processing is needed.

An Overview of search-based analysis:

Because search helps you discover insights you already know about, it doesn't help you discover things about which you're completely un-aware.

Query-based tools are time-consuming because search-based ap-proaches require a virtually infinite number of queries.

These methods are largely limited to numerical data and over 85% of data is unstructured.

Emergence of Algorithms as a New Class of Big Data Software Tools

The size and speed of big data demands true automation, in which work is offloaded from human to machine. This automation is found in algorithms designed for calculation, data processing, and automated reasoning. Algorithms are ideal for tasks that are beyond human comprehension and require the speed of machines. This is the future of data.

A fundamental change in the role played by analysts, from data-miners to insight-evaluators.

Fast and efficient algorithms that automatically convert data to insight for evaluation.

Machine-readable outputs that drive not just visualizations, but actions.

Continual improvement of these algorithms to keep up with the speed of data and critical need for timely insights.

There are many technologies attempting to address the challenge of big data, but most remain time-consuming and rely on informed queries or are custom built for specific, repeatable situations.

Emcien's Approach to Big Data Analytics

Emcien's approach adds an automation layer between data and the user or output system. The algorithms leverage advanced mathematics to solve complex problems, putting users closer to the answer and enabling automation through APIs.

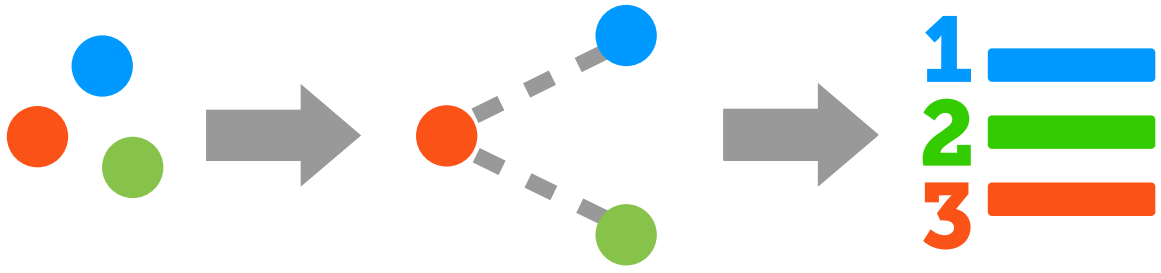
Emcien's innovation is the Emcien engine, algorithms designed to identify and rank the connections in data. These algorithms use a graph data model to capture the interconnectedness of the data elements and to create a representation of high volume data at a fraction of its original size.

Here is an outline of how the algorithm works:

Assign each unique element in the data a token.

Assess the data in order to identify and measure relationships of data points.

Rank the connections to identify patterns in the data. Differentiates noisy patterns from meaningful signals, isolating relevant connections.



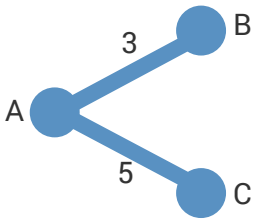
Outputs those patterns in a format that can be visualized, reported, or drive automated systems including recommendation engines, network analysis, etc.

emcien

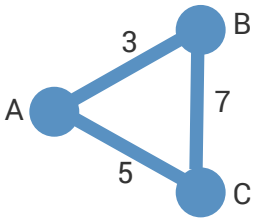
The Advantages of Emcien's Approach

The Emcien graph data model displays relationships and connectedness in a way that is not possible through other methods. Emcien's algorithms uncover these patterns and identify correlation across multiple variables. With a compact representation of large and complex data, the software can be run on commodity computing environments.

As entities continue to interact, big data becomes "bigger", with the number of events growing exponentially. The graph data model is ideal for big data because it creates a compact representation of the data. While common practice is to continue working with the raw data, Emcien creates a complete model of the relationships in the data set. In this graph data model, these interactions translate to connections on the graph, allowing the graph model to encapsulate data in much smaller structures.



Graph of Low Complexity Data Set



Graph of Higher Complexity Data Set

Understanding Data by Connections

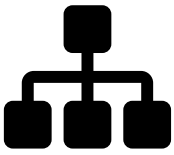
The graph data model is flexible and displays a distinct topography based on the density of connections. A cross sectional view of the graph data model will typically expose the layers outlined in the following table.

Noisy Connections	The most highly prevalent in most data sets, this layer is composed of high volume interactions that may be mundane or obvious.
Highly Connected	Lying just below the noise, this second layer is composed of the first signal that is interesting. This layer includes the stronger, but less obvious connections.
Weaker Connections	The third layer is a weaker signal and reveals the less-obvious connections. These relate to event correlations that are less frequent and may be connected in unexpected ways.
The Faint Signal	Composed of very weak connections and interactions, this last layer is of interest for security and surveillance. In many cases, this layer only emerges when the data is very rich in entities, causing connections to emerge in very non-obvious ways.

Not Just Theory: Solving Real-World Problems

The representation of complex networks as graphs helps to reveal timely and valuable information. One of the most important tasks in graph analysis is to identify closely connected network components that share similar properties. Detecting communities is of significant value in retail, healthcare, banking, and intelligence work - verticals where loosely federated communities deliver insight and intelligence into the profile of a customer base or any other group being analyzed.

How Can This Model Be Applied?



Internet of Things:

As more devices become connected there will be a need to monitor and understand the data that those devices create. Connected devices, from cars to medical implants, will reveal more about the way we interact with our world, but only for those who can extract information from that data.

Telecom:

Telcos create vital data in every stage of their operations. From data transmission to customer churn, this data can be leveraged to identify and address the levers that drive revenue.

Retail:

Customer analytics reveals insights in customer buying patterns, locations, demographics, loyalty, savings, lifestyle, and insurance. Pattern detection gives an in-depth breakdown of the relationships between customer characteristics and the items that they purchase.

Healthcare:

Analyzes massive volumes of clinical data on medications, allergies, medical claims, pharmacy therapies, lab results, medical records, clinician notes and more in order to reveal significant patterns. The engine is ideally suited to uncover relationships between thousands of numerical and non numerical data points.

Performance and Operations:

Analyzes raw information about performance and operations of every element of an organization. The complex patterns revealed in operations data can be interpreted to increase profitability or improve customer service.

Energy:

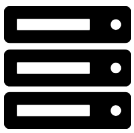
The discovery, extraction, transformation, and transmission of energy all create enormous amounts of data. The challenge in the industry is in identifying value from the noise created by the data that energy creates. Emcien keep these functions operating by identifying the events that predict equipment failure.

Fraud:

Builds patterns of behavior across millions or billions of transactions to identify activities that differ from the rest of the transactions on the graph. Patterns of money laundering and fraud can be identified by the unique characteristics that differentiate them from more common transactions. Data involving customers, locations, or transaction types that occur together in banking transactions can be analyzed to outline fraudulent activity from everyday behavior.

Intelligence:

Emcien software graphs and analyzes the connections between accounts and individuals, reveals the flow of interactions within groups, and identifies correlations between people that merit serious attention, determines key individuals in targeted social networks, and geo-locates persons of interest and their networks around the world, from gangs to terrorists.



Network Security:

Creates a representation of normal traffic to auto-detect intrusion patterns and reveal suspicious and anomalous behavior on the network. For example, Emcien analyzes millions to billions of transactions to identify patterns in source and destination-IP addresses, ports, days, times and activity to show you what you should be paying attention to. Emcien eliminates over 95% of the noise and identifies patterns that are “surprising” or that deviate from the norm.

Manufacturing:

Analyzing both sales and manufacturing data, manufacturers can analyze purchase patterns to discover the ideal build configurations to minimize build and carrying costs, maximize sales, and reduce the overall number of offerings.

Conclusion

The analytics revolution is underway. The expanding role of data in the business, government, and consumer sectors already impacts our lives. To fully harness the power of this data revolution, organizations must be able to do more than query large stores of data.

In the history of data analysis the objective of queries was to separate signals from noise. Until recently this has approach has been successful because we had clear-cut business questions, the size of the data was smaller, the data set was more complete, and we usually knew what we were looking for. In the new world of big data, it is now more important to know what to ignore because unless you know where to look, you'll never be sure of what's really important.

Extracting insight from big data requires methods of analysis that are fundamentally different from traditional querying, mining, and statistical analysis designed for small samples. Data is often noisy, dynamic, unstructured, inter-related, and untrustworthy. Using the traditional approaches to data analysis create opportunities for interpretation, bias, and overfitting. As sources of data continue to grow, analysts and data scientists will have to employ methods and tools like Emcien's that provide automated and repeatable analysis.