AUTOMATED DATA DISCOVERY

IN THIS EBOOK YOU WILL FIND:

- Executive Summary
- The What and Why of Data Discovery
- Discovery Today
- Where Does Data Discovery Fit?
- Automated Data Discovery
- What’s Next?
- Competing on Analytics
- Conclusion
- About Emcien
EXECUTIVE SUMMARY

Today, we are in the data economy. Companies are competing on data, analytics, and data-driven products. There is a growing tide of people who work with data on a regular basis, and this will only continue to grow. What we are seeing is the rise of the data worker. While more and more people work with data, there is a severe lack of tools and automation to improve their efficiency. Meanwhile, the pace of data growth has become overwhelming for many organizations striving to remain competitive in this new economy. What holds those data workers back is the ability to quickly navigate through data.

Compare this to a common and very data intensive mobile application. Google Maps gives users a dynamic, single-screen view of exactly the information they need to get where they’re going. It reduces tens of thousands of data points down to just the information that is helpful to the user, always available and up to date, right when it’s needed. If you find yourself in a strange and unknown place, you have an easy way to quickly orient yourself, discover what’s around you, and guide you to your destination. Now imagine you had that same capability, but for data. Connect to a data repository and immediately see all of the relevant and valuable information that data contains without having to search for it. That’s the power of automated data discovery.

Today IT is tasked with managing the data, keeping up with the high volume as it continues to grow and making it easily accessible to the business functions. Meanwhile the business functions are demanding access to the right data at speed and scale to remain competitive. For organizations managing gigabytes to petabytes of data, data management alone is a prohibitive task. But converting data to value requires understanding the data, and most business users lack the tools to quickly understand the content and the nature of data.

Even for the experienced data workers, grappling with truly big data requires very manually intensive work. Inside the data center or in the data lake there are terabytes and petabytes of data, and managing, cleaning, tracking, and moving the right enterprise data requires hours of querying column by column, which can take days in distributed file systems like Hadoop. These are the problems facing the data economy.

Emcien solves this problem with a suite of tools to automate data discovery and analysis. Through automated unsupervised machine learning, EmcienScan™ connects to your enterprise data to give you a bird’s eye view, a dynamic map of your data, so you always know what is in your data. Instead of requiring the user to query the data manually, Scan offers a continuous, automated process that helps you to understand your data. It gives you always available and always up to date visibility into your data repositories.
WHAT IS DISCOVERY?
Small data is easy to understand, but as data grows it just isn’t possible to comprehend an entire database or enterprise data warehouse. Data discovery is the ability to understand the data that has been collected. Without automation to drive this process, the only way to understand your data is with time consuming manual analysis and querying. There are exploration tools that require manual data tagging to set up a framework for searching the data, but this is time and labor intensive and certainly not sustainable as data continues to grow.

Attempting to understand the data by querying it creates only a snapshot of the data, a single point in time description of the data and only a small section of the larger data map. And because the user can only query for data he knows about, user bias is a constant threat. Two different analysts with the same data discovery task would render their own results.

SEARCH VS. TRUE DISCOVERY
The purpose of discovery is to understand what is in your data, evaluate quality, and find value wherever it exists. Organizations explore their data to discover what's possible with their data. In other words, where can the data take you? True automated data discovery is an unbiased approach to uncovering what is in the data. The result is that data discovery isn’t a process that users undertake, but is instead the capability of guiding the user to understand what is in the data. Users don't have to navigate data with queries, but are instead guided to discovery automatically.

DISCOVERY TODAY
Today data discovery is driven by manually intensive search. The methods for discovery are limited by the skills they require and the time they take, all to capture only a slice of the data and a small segment of the data map. Limited time and tools leave data workers to navigate the data with a very limited understanding of the data, and it’s not even clear that additional information exists unless the user is inspired to search for it.

When a business group creates a BI report to monitor some aspect of data, they typically only build the report around a static selection of available data. Discovery for IT is similar to the experience of business users, but at a much larger scale. Just like within the business units, exploration today is done ad hoc. Pressures of time and resources limit efforts to the specific needs of each project, again revealing only a slice of the data. IT is tasked with managing the data and making the right data available to the business in a timely and cost efficient manner.
WHERE DOES DATA DISCOVERY FIT?

A concept has emerged for managing and extracting value from data. Called the Data Factory, this process flow describes how data can be moved efficiently through the enterprise, from collection all the way to the end use.

- Data is collected and stored so that it can be accessed for a number of downstream uses. The data may come from the enterprise, external sources, or infrastructure sources like network, routers, etc. With the growth in data and the low cost of storage, collection and storage is where the modern enterprise excels. However, the costs of managing and maintaining this data escalate very quickly, particularly if not coupled with data evaluation and management to support the downstream steps.

- Data discovery enables evaluating, prioritizing, and organizing the data for downstream operations. It helps answer the question, “Is there any gold in my data?” Data discovery is the critical step to help drive efficiency in the data factory. Before any data prep or analysis, data discovery gets you to an immediate go/no go decision, reducing costs and risk of analytics projects.

- Data prep is the step where data is cleansed for downstream operations. Quality is typically poor, and the result is that most data requires cleaning and formatting before it can be pushed downstream. This is often called “data janitor work,” but is necessary for correct data results.

- This feeds into the downstream activities, including BI reporting, data analysis, or other data-driven systems.
AUTOMATED DATA DISCOVERY

The solution to the slow and manual process of data exploration is data discovery driven by unsupervised machine learning.

With Emcien Scan connected to your enterprise data, the change is transformative. Showing users only the relevant metadata, it becomes the lens through which your enterprise views all its data. It’s the 360 degree bird’s eye view of data that brings the most relevant information from across the data to the surface.

THE 360° BIRD’S EYE VIEW OF DATA

When we land in a new place, the first thing most of us do is look at the map to answer, “Where are we? What is the layout of the place? Where can I go from here?” A bird’s eye view gives you bearing and context. Just like your maps application combines GPS and thousands of data points, boiling everything down to the information that’s most relevant to the user, EmcienScan creates a single-screen visibility of into their data. For the end user this means a dashboard for all available data. It means getting to the right data in a matter of minutes, where it once meant hours or days.
For enterprise data administration, EmcienScan means visibility into every data repository. Connected to the enterprise, EmcienScan becomes an automatic mapping application for all of your data. Instead of analysts plotting a new course through queries and data tagging for each project, EmcienScan gives users a continuously updated map of all the scanned data. From this map, Scan guides users to relevant information without requiring them to search for it.

EmcienScan doesn’t just change the way users make discoveries in data, it transforms the organization’s relationship with data, helping them keep up with their data and accelerate monetization. No matter how large the data, questions about data don’t start with querying or coding but with a quick scan of the data.

DATA PROFILING
When you know the content and quality of your data, you have a deeper understanding of what that data can tell you. Is there a large spread of numeric values, or just a few unique variables? Are there data quality issues like outliers, anomalies, misspelling, missing values, etc.?

Data profiling answers these questions with the critical metadata that describes the nature of the data. This metadata includes minimum, maximum, standard deviation, frequency, variation, and other aggregates, to understand the distribution and nature of the data. Answering those questions is critical to maintaining valuable and usable data.

Without automation users must answer these questions manually, querying column by column. With automated profiling, EmcienScan becomes the fastest way for a user to get a complete profile of their data. It automatically guides the user to the most important content and the quality of the data, including data type, length, discrete values, uniqueness, null values, typical row length, and string pattern. But it goes a step further, identifying the distribution of each column, creating a definition of what is normal in the data and creating the queries to find and fix the outliers and anomalies in the data.

Switching from data profiling in coding intensive, distributed systems like HDFS to automated profiling reduces days of work to just minutes.
When profiling is automated with EmcienScan, this metadata is continually updated as the data changes over time. When the distribution of a column changes, or if there is an increase of outliers in the data, this will be reflected in the metadata and identified before the changes can affect the business.

**IS THE DATA DIRTY? WHAT IS DIRTY DATA? HOW DIRTY?**

Part of understanding what’s in the data is knowing the irregularities of the data. This is critical for quality control of your data. Common relationships and recurring events in data aren’t always interesting, but the odd and irregular data points represent potentially interesting events or errors that need to be addressed.

EmcienScan automatically detects the distribution of each column in the data, finding what’s normal for every column. With the normal data ranges discovered EmcienScan now knows what’s strange and dirty in the data, and will automatically create the queries for you to find and address the outliers in your source data.

With automated recurring scans, data can be monitored for surprising and interesting data and even alert users and administrators to errors and events. With a persistent understanding of enterprise data, it becomes possible to define what’s normal and identify new or strange data over time. Scan can be used to monitor continually for outliers and anomalies, even identifying data quality issues and connecting your data prep solution into the data factory.

**IS THERE VALUE IN MY DATA?**

Companies recognize the enormous potential of data and are collecting vast amounts of it, both from within their organization and externally. The main objective of all this collection is to understand how a system works and predict good and bad outcomes, but predicting these outcomes is only possible if there are connections in the data that link the outcomes to other data values.

EmcienScan gives users the ability to see the connections in data, showing them what can be predicted. When these connections are discovered automatically it suddenly becomes possible to see relationships across an entire table, dataset, or even the entire database. With EmcienScan users
instantly see how their data relates, and which parts of the data have no connection patterns, and no relevance for predictions.

Clicking on a column of interest immediately shows all of the data that is connected, telling the user exactly where to look and what data will be important. Additional metrics display the strength of the correlation between each column, showing not only that there is a relationship between columns, but also a comparable measure of the strength of each relationship.

KEEPING UP WITH THE VELOCITY OF DATA
Continuous data collection has become the hallmark of the data driven age. Data isn’t just collected and stored, but is constantly changing as a reflection of the business and the outside world. And while that data is changing, it’s still critical to the business infrastructure. Sales, marketing, operations, and human resources decisions are all made based on data.

Over time the underlying data changes, but without visibility the business units have no leading indicator of the shifting data or insight into what those changes might mean for the business. Sudden changes in sales data could signal a momentous shift in customer behavior. Slow changes to operations data might show how the basic functions of the enterprise are quietly being transformed. These changes might threaten the company, but without insight into the data they remain hidden until they are reflected in lagging indicators like revenues or quarterly profits.

With EmcienScan automating data discovery, these shifts are automatically brought to the forefront like accidents or delays in a traffic map. Metadata for everything from connection strength to outliers is tracked through time, and Scan’s APIs make it possible for triggered alerts to identify changes before they affect the business.
WHAT'S NEXT?

The most exciting thing about automated data discovery is what's possible when data everywhere becomes transparent and accessible. Just like Google Maps making sense of traffic data has spawned entirely new industries, the ability to see across so much enterprise data can be equally transformative.

- Connected to your CRM the persistent scans will let administrators see the evolution of their sales organization. Changes in the market will be discovered and tracked in the data long before they appear in quarterly revenues.

- EmcienScan will be critical to understanding the data of IoT, where sensors will be everywhere and changing data will signal shifts in real-world conditions.

- Gauge your organizational capacity. Know what is possible with the data you have, what data is available, and the quality of the data to become a more dynamic and data-driven organization. Rapidly assess how acquiring new data can enhance your competitive edge.

- Getting to know your customer? With EmcienScan you'll see the full 360 degree view of your customers and their actions.

- Day-to-day efficiency gains for all the analysts across your enterprise.
COMPETING ON ANALYTICS

The data driven economy is here and organizations are trying to keep up. The need for speed and efficiency is dire. Companies are striving for a competitive advantage in data, and the fastest way to achieve those gains is through automation.

This chart demonstrates where efficiencies are gained through automation.

<table>
<thead>
<tr>
<th>Annual T&amp;M Savings</th>
<th>EmcienScan ROI Model - Hard Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tech Staff</td>
</tr>
<tr>
<td>Description</td>
<td>Values</td>
</tr>
<tr>
<td>Technical staff count</td>
<td>1</td>
</tr>
<tr>
<td>Hourly burden rate</td>
<td>$85.00</td>
</tr>
<tr>
<td>% available time spent on data discovery, profiling &amp; identifying outliers</td>
<td>55%</td>
</tr>
<tr>
<td>Employee available Task work weeks/year</td>
<td>50</td>
</tr>
<tr>
<td>Total staff Hours worked on data exploration</td>
<td>1,100</td>
</tr>
<tr>
<td>Efficiency gain thru EmcienScan</td>
<td>50%</td>
</tr>
<tr>
<td>Net Total Annual hours to be re-purposed</td>
<td>550</td>
</tr>
<tr>
<td>Net Total Annual dollars to be re-purposed</td>
<td>$46,750</td>
</tr>
<tr>
<td>Total Annual hours to be re-purposed</td>
<td>496</td>
</tr>
<tr>
<td>Total Annual dollars to be re-purposed</td>
<td>$42,160</td>
</tr>
</tbody>
</table>
CONCLUSION

The single greatest barrier to realizing value from data is an organization’s ability to understand it’s content and quality. With EmcienScan, every data worker is enabled to assess, evaluate, and discover value in data.

Automated discovery embedded in the enterprise is transformative for anyone working with data. What has historically required a combination of technical skills and subject matter expertise becomes the domain of the business user, without any coding or querying.

Automating discovery doesn’t simply save time, but gives analysts and administrators information that was unattainable before. The old model of discovery meant waiting for the results of map-reduce jobs that could take more than 24 hours for a simple profile. Automated discovery gives users an always up to date profile of the data and a map of the relationships across the entire data set. EmcienScan becomes the lens through which users see their data.

ABOUT EMCIEN

Emcien is a pioneer in automated machine learning for data discovery and analytics. Our mission is to empower everyone to participate successfully in the data driven economy through powerful, easy to use software and embedded analytics. We do this by offering solutions that automate data analysis and delivers just the answers. No data skills required.

Learn more at http://www.emcien.com