Abstract

ESG Lab recently completed testing of Dell EMC’s Analytic Insights Module, which is designed to enable organizations to analyze and extract value from big data more easily. Testing examined how Analytic Insights Module gathers, analyzes, and acts on data—with a focus on ease of use, collaboration, time savings, data security, and simplified data integration.

The Challenges

Big data analytics brings with it a unique set of challenges. ESG's 2016 Research Report, Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux, revealed that the data analytics challenges most-cited by surveyed organizations are weighted heavily toward data integration issues, including complexity, running analytics across different data sources and types, siloed IT systems, and structured and unstructured data sets. A shortage of the skills needed to properly manage and extract value from large data sets was also reported as a challenge.¹

Figure 1. Top Ten Data Analytics Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration is complex</td>
<td>34%</td>
</tr>
<tr>
<td>Data set sizes limit our ability to perform analytics</td>
<td>31%</td>
</tr>
<tr>
<td>Difficulty running analytics across disparate/siloed IT systems</td>
<td>25%</td>
</tr>
<tr>
<td>Difficulty running analytics across internal and external data sources</td>
<td>24%</td>
</tr>
<tr>
<td>Lack of skills necessary to properly manage large data sets and derive value from them</td>
<td>23%</td>
</tr>
<tr>
<td>Limited collaboration between IT, analysts, and/or line of business</td>
<td>23%</td>
</tr>
<tr>
<td>Difficulty running analytics across structured and unstructured data sets</td>
<td>23%</td>
</tr>
<tr>
<td>Current data analytics license costs are too expensive</td>
<td>23%</td>
</tr>
<tr>
<td>Storage requirements are too expensive</td>
<td>21%</td>
</tr>
<tr>
<td>Unable to complete analytics in a reasonable period of time</td>
<td>21%</td>
</tr>
</tbody>
</table>

As more organizations adopt big data analytics and data sets and big data clusters grow in both size and importance to the business, these challenges will intensify. What is needed is a solution that can simplify integration and enable organizations to extract value from these massive amounts of data quickly and easily.

The Solution: Dell EMC Analytic Insights Module

Dell EMC Analytic Insights Module is a solution that combines self-service data analytics with cloud-native application development into a single cloud platform, which is engineered to enable organizations to transform data into actionable insights. The Analytic Insights Module workflow consists of three primary components, as shown in Figure 2:

1. Gathering the data.
2. Analyzing the data.
3. Acting on the data.

Figure 2. The Dell EMC Analytic Insights Module Architecture

The Analytic Insights Module was engineered to make it easier for organizations to get real business value out of their big data quickly. Data analytics has traditionally required dedicated IT resources and personnel to perform data acquisition and analysis, application development to operationalize insights, and many other manual, time-consuming activities before value can be extracted. Analytic Insights Module is designed to enable organizations to rapidly transform data into actionable insights with high business value in the fastest time possible. Data analyst teams can rapidly find valuable insights on a self-service cloud platform where IT teams can apply policies to the environment for ensuring corporate governance.

Gathering the Data

Analytic Insights Module helps gather and index internal and external data from a variety of sources including databases, sensors, automated detection devices, social media, remote and branch offices, and public sources. Once a data source has been discovered, its data can be sampled to determine its value and completeness, then pulled into an Isilon data lake, which is part of the fully engineered Analytic Insights Module. A data lake is a large repository of data used for big data analytics that provides a single view of all discovered and indexed data. A data lake can include structured, semi-structured, unstructured, and streaming data in its original raw form, regardless of type or format.
Data gathering consists of more than just copying it into the data lake. The Data Curator provides a discovery engine that samples data and enriches it with metadata to provide context and help identify important data through semantic searching. The context provided by the metadata it produces helps an analyst differentiate between the nation of Turkey and the bird, for example. The result is a clear view into these data stores, enabling analytics teams access to more of the right data for their analysis. The Semantic Searching feature helps identify the important data that the analytics team will want to analyze without requiring them to take the time to scrub, cleanse, and tag the data. Analytics teams are valuable resources; automating these activities enables them to be 100% focused on revealing business value, rather than preparing data before analysis can even begin.

Discovered and ingested data is catalogued in the Data and Analytics Catalog, which provides one view of all data for the organization, regardless of where it resides—within the data lake, external to the data lake, or in public data sources. Additional data sources, whether structured or unstructured, publicly available, or leveraged from SAP Hana or enterprise data warehouses, can be blended with data internal to the data lake to produce insights and enhance the intelligence of cloud-native applications.

The Data Curator profiles, enriches, and unifies data across the enterprise, finding correlations and relationships across data stores. Because organizations have many data sets that are not useful to ingest into the Data and Analytics Catalog, it is important to discover the data and profile or sample it before making the decision to ingest. The Data Scientists would use Data Curator functionality to look at data samples and make informed decisions about quality.

Analytic Insights Module’s Data Governor allows the IT team to apply data-level role-based access policies for all users, applications, and data stores, regardless of file system, storage medium, or data type. Administrators apply precise access controls and policy enforcement, and the Data Governor continuously monitors incoming data to ensure proper, timely handling, and compliance with regulations like HIPAA and PCI.

**Figure 3. The Analytic Insights Module Workspace Management Screen**

**ESG Lab Tested**

Before gathering any data, users would create a data ingestion workflow and add data governance policies using the Data Governor. Next, a private workspace that serves as a user’s “sandbox” must be created. The workspace is a space where a data analyst can pull together private copies of the data to perform the desired analysis. Typically, IT administrators must be contacted to create a workspace, but with Analytic Insights Module, data analysts can do it themselves with a quota-based self-service experience in minutes. To create a private workspace, ESG Lab began from the Analytic Insights Module dashboard, as shown in Figure 3. Clicking on the plus sign initiates the workspace creation process with a pop-up asking for a name for the workspace; since the test workspace will be serving the role of a data consumer, the workspace was named consumer-ws-02. Almost immediately, a screen appeared indicating that a workspace was created and just a few more seconds were required for the workbench, a dedicated private virtual machine, to be cloned and for the workbench’s IP address to be displayed.
Next, services are selected from the Services Marketplace, as shown in Figure 4. Services are third-party applications including open source and commercial distributions of Hadoop that are selected and licensed by administrators and presented to users in a marketplace environment as seen in Figure 4. For this test, ESG Lab selected hortonworks_small. With one click, a full Hortonworks Hadoop cluster environment was created, with all the services that come with the Hortonworks distribution deployed and pre-configured.

**Figure 4. The Analytic Insights Module Services Marketplace**

Users can add new data sets to the Hadoop cluster by ingesting individual data sets one at a time through the Data Catalog, or by building a data mart within the cluster to create a pre-joined group of multiple data sets. A data mart is a working area within a big data cluster where users upload data, profile a data set, and perform some preliminary data manipulation, such as table joins. Users can only see the data sets to which they have been granted access, an important governance requirement.

ESG Lab initiated the process on the dashboard screen by clicking on the **Add Dataset** button and then on **Ingest New Data**, as shown in Figure 5. This initiated a login to Data Source Discovery, which indexes data, adds a semantic layer, and provides advanced search for data with automatic tags and recommendations.
ESG Lab clicked on **Build a Mart**—although marts can also be imported—named the mart **MyMart01**, entered a description, and clicked Next. The next screen presented an option for tags, which can be assigned to just about any object that Analytic Insights Module manages. The final screen asked about collaboration, identifying users who will share access to this data mart. In this case, none were selected. Once the **Next** button was clicked, the data mart was created. On the next screen was a list of data sets that had already been imported into the environment. Since Analytic Insights Module was designed with data-level role-based security in mind, each user might see a different list of imported data sets, or different parts of a data set appropriate to their role; one user might have access to full social security numbers while another sees only the last 4 digits, for example.

The Data Curator allows the user to manually upload a CSV, XLS, or ZIP file, or to select a new data source. ESG Lab clicked on **Add a Dataset to the Mart**, chose to upload a CSV file called **branch.csv**, and named the new data set **branch02**. Then we accepted the defaults on the Process screen (no workflows will be associated right now) and on the Scheduling Updates screen (update on demand), and clicked the **Create Source** button. It took about two seconds to connect the data source, and another few seconds to ingest all the data from the CSV file.

To complete the ingestion and get the data into the data mart, ESG Lab entered the data mart screen, chose the **Provision Mart** menu, and selected **Ingest** from the menu. The UI asked for the name of the workspace that the data would be imported to and the “containers” created by the services deployed in the workspace. For example, if a Hadoop service was deployed, we would have been given the option to import to the Hive table.

The creation of a workspace and the ingestion of data into that workspace was fast and easy. Analytic Insights Module automates much of the hard work, and the whole process is much faster and easier than if it were done manually. It took 15 minutes to set up a new Hadoop cluster and begin to import the data into the workspace. In contrast, when ESG deployed an eight-node open-source Hadoop Cluster on bare metal servers, setup and configuration consumed 56 hours of total active work time. Deploying a commercial distribution streamlined much of the configuration, but still took more than 24 work hours. This represents a reduction in time and effort of more than 99%.
Why This Matters

The value of big data is not simply in having it, but in understanding it and in gaining insights from it. But the size and quantity of the data makes it impractically slow to work with manually. ESG research indicates that one of the top-ten challenges facing organizations working with big data is that they are unable to complete analytics in a reasonable amount of time. Some customers report that it can take weeks or months to create a working environment.²

ESG Lab verified the ease and speed of bringing data into Analytic Insights Module. It took just 15 minutes to create a private workspace, deploy Hortonworks, import and index the CSV file, determine the automatic update schedule, and make the workspace ready for analytics. Dell EMC Analytic Insights Module makes it easy to import the data into workspaces, and can significantly reduce the amount of time it takes to get the environment ready for users, which in turn leads to improvements in the efficiency of IT administrators and data scientists, and in delivery times of new projects.

Analyzing the Data

Analytic Insights Module is focused on making the lives of data analysts easier. One way the solution achieves that is by maintaining an open platform. Dell EMC understands that data analysts have specific tools that they like to work with, and so the product has been designed as an open platform, permitting analysts to use their favorite tools to do their jobs. Once data has been pulled into the workspace, the analyst can work on the data on demand with complete self-service, without impact to the larger enterprise. Quotas ensure that multiple analytics jobs can run simultaneously, without disrupting the enterprise infrastructure.

When the analyst has identified trends and other insights from within the data set, those results can be easily published back into the Data and Analytics Catalog, enabling other analysts to build on this work without having to start from scratch. The result is a marked improvement in efficiency and collaboration among data analysts, and improved time to delivery.

Data security and compliance is, of course, a concern for any organization. Analytic Insights Module offers data-level security at a user and group level, so that sensitive data can be hidden from users. The default is for a user to have no access to published data at all; any permissions must be specifically granted to users to help prevent accidental breaches.

ESG Lab Tested

To test the user-level security in Analytic Insights Module, ESG Lab used the open source utility Zeppelin to view a database as three different users, Alice, Bob, and Charlie. Alice has full privileges and can see all data in this database. Bob can see all the entries, but his access to social security numbers is restricted through masking. Charlie has no access to this data at all. The three examples are depicted graphically using fictional data. Analytic Insights Module offers a great deal of flexibility in its ability to block or mask data at a granular level and simplifies governance of those security policies using the Data Governor.

Another useful feature that assists in data analysis is the Workflow and Transformation Tool, which allows users to create automated workflows to transform data as it is ingested, saving data analysts more time. ESG Lab performed tests using a combination of four different real-world data sets—unstructured Twitter text files, two sets of semi-structured weblogs, and structured transactional data. The 5TB data set generated 33.15 billion rows of post-parse data and ESG Lab recorded development time for data load and transform jobs from eight to 30 hours per data set.\(^3\)

The simple workflow example in Figure 7 depicts three steps being performed on incoming data. Step 1 is a Data Quality Action, where the tool verifies that the incoming data meets certain basic qualifications, such as having clean data in all required fields; if it does not meet them, then the data can be rejected or corrected as needed. Once data quality has been verified, a transformation is applied to it in Step 2. For example, changes may be needed if the name of the state is represented improperly. Instead of “Massachusetts,” perhaps an end-user entered “MA,” “Mass,” or a misspelling of the word. The transformation step can take that data and correct or normalize it automatically. In Step 3, the results are emailed to the data scientist who created the workflow.

Figure 7. The Workflow and Transformation Tool

---

\(^3\) Ibid.

© 2017 by The Enterprise Strategy Group, Inc. All Rights Reserved.
To create the simple transformation shown in Figure 7, ESG Lab dragged the task circles into place and connected them by dragging the mouse between them. Each task circle also has an associated data screen where specific information related to that task is managed and entered. For **Data Quality Action**, ESG Lab selected the fields that needed to be verified. For **Transformation**, there is a list of strings to look for, and a string that they should be transformed into. Under **Email Action**, there was a field to enter an email address and a template for the contents of the email message. Additional circles above the workflow represent other automated tasks that can be added to a workflow. In ESG Lab’s previous experience, developing the custom scripts and code to manually load and transform data for analytics is labor-intensive, consuming many hours per job. ESG Lab confirmed that Analytic Insights Module enables data scientists to create, manage, and maintain workflows quickly and easily, accomplishing in just under an hour what took up to 30 hours using Hadoop with native tools, a 96.7%-time savings in this example.

### Why This Matters

According to ESG research, the most often cited challenge facing organizations that use big data is the overall complexity of data integration. This challenge is closely related to several of the other reported challenges involving the analysis of siloed data, internal and external data, structured and unstructured data, and generally large data sets. Siloed data is a problem when one team works with data and internal security and privacy policies prohibit that team from sharing the data within the organization, or when data is stored in different data warehouses and databases that are not shared among business units, which leads to lack of access or insight into what data the company possesses. This leads to repeated work as popular analyses must be redone from scratch, leading to redundant programming work, and potentially the introduction of errors.

ESG Lab verified that Analytic Insights Module addresses analysis challenges in a couple of ways. The Data and Analytics Catalog provides a centralized repository where data scientists can publish their curated data sets and analytical models for others across the organization to reuse. This reuse of trusted data and analytical models increases productivity and reduces the time to reveal valuable insights for the business. At the same time, the data-level role-based security settings prevent inappropriate or private data from getting into the wrong hands. The result is efficient and secure data sharing throughout an organization.

### Acting on the Data

The final component of Analytic Insights Module helps organizations take the data and the insights that they’ve acquired and act on them. Analytic Insights Module is delivered on Native Hybrid Cloud, a turnkey platform based on Pivotal Cloud Foundry. This makes it easy to turn these insights into data-driven intelligent applications, visualization tools, and new business processes.

The key to this improved development experience is in the API calls that can bind applications directly to the analytics model. Developers can commit and push the app and analytics model to QA for performance and scaling testing, then push and scale to production with a simple command. Once in production, the application can feed data back into the analytics process within Analytic Insights Module, to enrich the intelligence in the next application version. The data can also be shared easily among different data scientists and projects, so data-driven applications can be built, tested, and deployed much more quickly than those based on traditional application development models.

Native Hybrid Cloud leverages the power of Pivotal Cloud Foundry utilities to manage the entire stack, which means that a locally developed application can be pushed out to the cloud without the developer or administrator having to worry about local compatibilities, and without IT intervention. Traditional IT provisioning for a multi-tiered app can take weeks or months

---

with multiple manual steps, including but not limited to specifying and acquiring hardware, validating OS and application software versions and compatibility, security and compliance reviews, QA, and actual installation.

Pivotal Cloud Foundry apps can be pushed out by the developer with a minimum of effort. Pivotal customers report that Pivotal Cloud Foundry delivers a two-times improvement in developer efficiency, and a ten-times improvement in overall operational efficiency, with product delivery cycles reduced by as much as 77%.

**ESG Lab Tested**

For the last part of the testing, ESG Lab deployed an already-written application called *marketing-dashboard-02* from the development environment into the Native Hybrid Cloud via Pivotal Cloud Foundry. The first step was to log into the Pivotal Cloud Foundry environment from either the Linux command line or a Windows Power Shell, and when asked, choose to connect to the cfapps organization and the development space to place the new application. The organization and space settings ensure that the right users have access to the application and that they can find it when they need it.

Basic instructions for deployment are stored in a small local manifest.yml file, as shown in Figure 8. In manifest.yml, *memory* represents the amount of memory on the VM that will be deployed. Should the developer require more than one identical instance of the application, all that’s required is to increase the value of *instances*. The *buildpack* contains the detailed instructions for unpacking and installing the application. A listing of all available apps, including marketing-dashboard-02, is shown in Figure 9.

**Figure 8. Example manifest.yml File**

```yaml
applications:
  - name: marketing-dashboard-02
    memory: 1G
    instances: 1
    buildpack: https://github.com/username/cf-buildpack-r.git
    path: .
```

**Figure 9. The Local Development Space in Native Hybrid Cloud**

The final step was to enter the command *cf push* on the command line, which initiated the deployment of the application into the Native Hybrid Cloud, in the cfapps organization and the development space. Fifteen minutes later, the application was ready to use.
Why This Matters

Organizations leverage big data analytics to use the insights gained to create and deliver new products, services, and support new business models. The process of piping data from data science groups to application developers, then pushing applications to production and updating them as needed can be extremely time consuming, especially when IT is burdened with facilitating this process at every step.

ESG Lab validated that Analytic Insights Module delivered on Native Hybrid Cloud automates and orchestrates the entire process from data collection and analytics to app development and deployment much faster and easier than conventional methods. Analytic Insights Module removes the friction of deriving valuable business insights and transforming them into intelligent applications through a self-service experience and infrastructure abstraction. Data scientists and data analysts can reduce the time to valuable business insights without having to wait for or rely on IT for assistance. We leveraged Native Hybrid Cloud and deployed a newly developed application into a production environment in just 15 minutes, orders of magnitude faster than possible with legacy methods.

In less than three hours, ESG Lab stood up a workspace, performed data set discovery, built data flows, ingested data, selected and implemented security and governance policies, and pushed an application to production, a process that could take multiple weeks, a time savings of up to 99% in some cases.
The Bigger Truth

Fully half of enterprise organizations (i.e., more than 1,000 employees) surveyed by ESG in 2016 expressed high levels of interest in big data and analytics initiatives, indicating that it was their most important IT priority, and another 31% ranked these initiatives among their top five. While IT organizations in the enterprise perceive strong financial benefits to come, there is some question over how soon these benefits can be realized.\(^5\)

Organizations must differentiate themselves and constantly evolve to remain competitive. This requires continuous data gathering, analysis, and action on the data. If an organization doesn’t have the requisite level of expertise to build a comprehensive solution, it will likely fail to move its data analytics projects beyond mere experimentation. In addition, siloes of data that cannot be shared across the organization are a challenge that will prevent them from mining data resources to reveal new business insights that can help accelerate growth—which is why they’re interested in analytics in the first place.

Therefore, build vs. buy is a key consideration. Organizations should be able to focus on their core business rather than spending valuable time and resources integrating hardware and software. With 78% of organizations reporting expected time to value for their big data and analytics initiatives of seven months or more,\(^6\) they should seek out a solution that accelerates the identification of valuable business insights and enables transformation of those insights into applications, new business processes, and embedded systems in the fastest time possible with the least effort. Dell EMC reports the average implementation of Analytic Insights Module is about 16 weeks—at least a 47% savings in time for organizations’ first implementation. The savings become even more pronounced over time. In less than three hours, ESG Lab stood up a workspace, performed data set discovery, built data flows, ingested data, selected and implemented security and governance policies, and pushed an existing application to production, a process that could take multiple weeks in a traditionally administered big data analytics environment, a time savings of 99%.

After the initial deployment, Dell EMC provides services and guidance for upgrading Analytic Insights Module, eliminating the need for the comprehensive testing required with individual component revisions in a “build your own” model. Support across the many pieces including hardware, Hadoop distributions and third party products is coordinated by Dell EMC, enabling speedy resolution of issues with one-call support. The Analytic Insights Module has been designed using hyperconverged infrastructure based on Intel Xeon E5-2600 v4 processors to offer scalability for data analytics compute and storage in an agile, efficient data center. The use of Isilon separates scaling of compute and storage to reduce the hardware base and potentially reduce license support costs that are often based on node counts.

Analytic Insights Module’s integration of components provides a platform that requires a lot less ongoing administration time and delivers results faster. ESG Lab testing verified data analysts can stand up analytics environments and allocate resources without IT intervention. ESG Lab testing demonstrated that data ingestion is fast, easy, and automated to ensure repeatable results, and that user-level security and governance protects shared data and analyses. Dell EMC includes Native Hybrid Cloud, a turnkey platform based on Pivotal Cloud Foundry, to accelerate and simplify cloud-native application delivery. The solution provides a complete cloud-native solution that can be supported and sustained as one product. In a radical departure from traditional analytics app delivery methodologies, which can take weeks to deploy, Analytic Insights Module includes Pivotal Cloud Foundry. ESG Lab pushed a pre-developed application to production in just a few minutes. Pivotal customers report that Pivotal Cloud Foundry delivers a two-times improvement in developer efficiency, and a ten-times improvement in overall operational efficiency, with product delivery cycles reduced by as much as 77%.

As organizations look to improve the understanding of their business and differentiate themselves in an ever more competitive world, big data analytics will become an even more important initiative. If you are looking to leverage insights from all your data sources—public and private—monetize new opportunities, enhance customer experiences, and optimize key business processes, ESG Lab recommends taking a close look at the Dell EMC Analytic Insights Module.


\(^6\) Ibid.