Abstract
This solution guide introduces Analytic Insights Module, a highly scalable technology platform for deploying data-driven analytics and application development.

H15560R

This document is not intended for audiences in China, Hong Kong, and Taiwan.
# Contents

## Chapter 1  Executive Summary  
5
- Business case .......................................................... 6
- Solution overview ..................................................... 7
- Key benefits ............................................................ 8
- Document purpose .................................................... 8
- Audience ............................................................... 8
- Essential reading ...................................................... 8
- Terminology ............................................................ 9
- We value your feedback! ............................................. 10

## Chapter 2  Data Platform  
11
- Introduction ............................................................ 12
- Solution features and functionality ............................... 12
- Solution architecture ................................................ 13
- Key components ....................................................... 20
- Hardware resources .................................................. 20
- Software resources .................................................... 20

## Chapter 3  Data and Analytics Catalog and Workspaces  
22
- Data and Analytics Catalog ........................................ 23
- Working in workspaces ............................................. 26
- Managing data in workspaces .................................... 30
- Analyzing real-time data in workspaces ......................... 33
- Assisting data analysts with data analysis ..................... 34

## Chapter 4  Security  
35
- Introduction ............................................................ 36
- Security solution ...................................................... 36
- Securing Hadoop clusters .......................................... 38
- BlueTalon security infrastructure ................................. 40
- DAC security ........................................................... 44
- Workspace security .................................................. 48
- ODC security .......................................................... 52

## Chapter 5  Data Ingestion  
57
- Introduction ............................................................ 58
- Data ingestion workflows ........................................... 58
Chapter 1 Executive Summary

This chapter presents the following topics:

- Business case
- Solution overview
- Key benefits
- Document purpose
- Audience
- Essential reading
- Terminology
- We value your feedback!
Chapter 1: Executive Summary

Business case

Introduction

A new generation of analytics, applications, and the Internet of Things is creating a dramatic shift in business. Big Data drives new possibilities for businesses to gather, analyze, and act in real time and use data-generated insights to optimize key business processes, identify new business opportunities, and increase profitable customer engagements.

Businesses that are successfully adapting and exploiting the digital economy are doing so across the enterprise through software or insight-driven applications. They are collecting data from nontraditional sources to gain more insights and to create innovative products and solutions. These businesses must analyze collected data, build models, make predictions that are based on previously collected data, and then use the resulting insights to support business outcomes. The speed at which a business can capture and analyze data and then act on the resulting insights affects business competitiveness and growth.

In addition, with the advent of the third platform, businesses are asking IT to develop, deploy, support, and scale a new class of applications that operationalize the insights that are generated from Big Data. Insight-driven applications produce more data, creating newer insights. This new data and the resulting insights can then be integrated back into the data analytics process to enrich intelligent applications and provide greater insights. The speed at which the business can cycle through capturing data, analyze it, and upgrade the application or business processes enhances the ability to grow the business and create a competitive advantage.

Challenges

In capturing Big Data and deriving value from analytics, businesses face inherent challenges such as:

- Finding, keeping, and curating the data
- Identifying appropriate use cases
- Operationalizing insights to support business outcomes

Hadoop

With Hadoop and the Hadoop Distributed File System (HDFS), businesses can keep a far greater quantity of historical data online. They can also process the data together with realtime data that is streaming from sensors and systems. However, even with the advent and maturity of Hadoop platforms and the ability of IT to easily build physical Hadoop clusters, organizations struggle to take the next steps due to the following challenges:

- Emphasis on technical activities instead of business outcomes
- Difficulty in managing and deploying the various tools
- Inability to easily expand restrictive physical infrastructures, slowing analytics to a crawl
- Lack of necessary data governance
- Inability to keep up with the continual arrival of emerging technologies
Solution overview

Analytic Insights Module is designed to enable businesses to go from data discovery to actionable insights in the shortest time possible, reducing to days what used to require months. Combining data, analytics, and application development in a single platform, Analytic Insights Module provides the following key functions:

- **Gather**—Provides comprehensive data awareness, governance, and management, as well as scalable infrastructure and end-to-end security
- **Analyze**—Empowers data science teams with self-service, quota-driven workspaces for data analytics, boosting productivity for faster time to value
- **Act**—Simplifies and accelerates the ability to transform actionable insights into data-driven applications and business processes to deliver new products and services and to support new business models

Analytic Insights Module is a robust enterprise-grade solution that makes Big Data initiatives radically simple. Built with a combination of Dell EMC engineered components and third-party tools, Analytic Insights Module includes key features such as self-service, data model reuse, and security. The solution provides:

- Preconfigured building blocks and a converged infrastructure that let you focus on new capabilities instead of new infrastructure
- A flexible and self-service virtualized ecosystem that enables you to prove value quickly, scale out rapidly, and keep security and governance under confident control
- User-initiated analytics workspaces (self-service sandboxes) with assigned quotas per workspace
- Analytics tools that are deployed seamlessly to workspaces using the user interface (UI)
- The ability to provision published, prepared datasets into a workspace, which enables the data scientist to take advantage of the finding, blending, wrangling, and preparing of data from the prior efforts of other data scientists
- The ability to invite other consumers to collaborate in a workspace

With this solution, Dell EMC has integrated a data lake, flexible analytics, and an application development platform to provide a modern digital transformation platform as a service (PaaS). The Analytic Insights Module solution removes barriers and bottlenecks from analytics and development lifecycles, while enabling IT to maintain security, visibility, and control of applications, infrastructures, and services, all with single-contact support from Dell EMC.

Key benefits

Analytic Insights Module is a fully engineered solution that addresses the challenges that are listed on page 6 by providing the following features and benefits:

- **Full integration**—Integrates a data lake, self-service analytics, scalable storage, and an ecosystem of technologies to enhance the ability of businesses to move from insights to action
- **Fast business value delivery**—Provides instant access to data, analytics, and rapid application development, enabling you to go quickly from concept to realization
- **Reduced risk**—Eliminates risks that are prevalent for realization and production deployment by combining proven technologies and compatible software versions, from Dell EMC and other companies, that are engineered to work together
- **Ease of use**—Addresses the needs of all users across the organization, including data engineers, data scientists, application developers, and IT, while providing a seamless support and service experience across the entire platform
- **Enterprise-scale, open, and extensible platform**—Provides an infrastructure and capabilities that scale effortlessly to meet the Big Data analytics needs of the enterprise, both in terms of dataset sizes and system performance requirements

Document purpose

This solution guide describes Analytic Insights Module. The guide introduces the main features, functionality, and key components of the Analytic Insights Module solution, such as the Data and Analytics Catalog (DAC), and the validated hardware and software environment.

Audience

This guide is intended for executives, managers, architects, business analysts, data scientists, data engineers, and application developers who want to implement Analytic Insights Module. Readers should be familiar with Hadoop, analytics tools, and data-lake concepts.

Essential reading

The following documents describe the architecture, components, features, and functionality of the Analytic Insights Module solution:

- Analytic Insights Module Reference Architecture Guide
- Analytic Insights Module User Guide
- Analytic Insights Module Release Notes
This guide provides external references where applicable. Dell EMC recommends that users implementing this solution be familiar with the referenced documents. For details, see Chapter 8: References.

Terminology

Table 1 provides definitions for some of the terms that are used in this guide.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Electronically stored information representing a collection of structured or unstructured data that has been previously registered and published in Analytic Insights Module.</td>
</tr>
<tr>
<td>Data and Analytics Catalog (DAC)</td>
<td>A core component of Analytic Insights Module that includes a management interface that enables designated users to search, assess, and publish available datasets from customized workspaces, and to deploy available tools and applications to customized workspaces.</td>
</tr>
<tr>
<td>Data container</td>
<td>Container used to store data in a workspace. Containers can be Hadoop (Cloudera or Hortonworks), MongoDB, Attivio, MySQL, or Redis.</td>
</tr>
<tr>
<td>Data lake</td>
<td>Centralized repository for large amounts of structured and unstructured data. In Analytic Insights Module, the data lake is Isilon™ storage.</td>
</tr>
<tr>
<td>Data source</td>
<td>Object that represents an enterprise content source that contains or provides electronically stored information outside Analytic Insights Module.</td>
</tr>
<tr>
<td>Published Data Container (PDC)</td>
<td>A dedicated data lake holding dehydrated data assets that can be shared with workspaces, or other customer-specific data containers, which are defined by customer needs. The PDC is a subset of the Operational Data Container (ODC), but data in the PDC is known (published) to the DAC (indexed).</td>
</tr>
</tbody>
</table>
| Role                        | One of the following user groups in Analytic Insights Module:  
  - Consumer  
  - Data administrator  
  - Security administrator  
  - IT Administrator  
  Users or groups are assigned with roles that are defined in the role database, and each role maps to related permissions. (For information about role-based access control and definitions of each role, see DAC security.) |
| Workspace                   | Isolated, managed container of virtual machines, working sets, and analytic tools that are configured and deployed within Analytic Insights Module. A workspace is a private sandbox for a single user or a group of users. |
| Working set                 | Dataset that is deployed to a workspace or created within a workspace.                                                                 |

Table 1. Terminology
We value your feedback!

Dell EMC and the authors of this document welcome your feedback on the solution and the solution documentation. Contact EMC.Solution.Feedback@emc.com with your comments.

Authors: Dave O’Donoghue, Philip Edwards, Nihar Nanda, Colleen Jones
This chapter presents the following topics:

**Introduction** ........................................................................................................................................... 12

**Solution features and functionality** ...................................................................................................... 12

**Solution architecture** ............................................................................................................................. 13

**Key components** ...................................................................................................................................... 20

**Hardware resources** ............................................................................................................................... 20

**Software resources** ................................................................................................................................. 20
Chapter 2: Data Platform

Introduction

Analytic Insights Module enables organizations to implement a robust Big Data environment (storage, compute, and Big Data analytical tools) with minimal complexity, effort, and time. The infrastructure has been designed and tested to satisfy the current needs of the organization and to scale easily to future needs, whether that scale is for compute, storage, or analytics.

Solution features and functionality

Broad, flexible ecosystem

Analytic Insights Module helps businesses on their journey to become data-driven enterprises. In a new era of analytics capabilities, ushered in by emerging platforms and open-source frameworks, businesses can now evaluate all the internal and external data assets in the enterprise and from public data sources. Organizations are free from the traditional confines of large-scale business justifications that are based historically on cost, complexity, and compartmentalization. Data scientists have adopted an agile approach to data exploration, and nonlinear workflows have emerged.

This new culture presents businesses with a twofold challenge:

- How does the business rationalize and use the analytics for competitive advantage?
- How does the business integrate the new processes and technologies that are required to facilitate this revolution?

Analytic Insights Module helps you address this challenge. It simplifies the building of an infrastructure that can scale to meet the demands of data-driven enterprises by:

- Smoothing the transition of the analytics function from a narrow IT-based silo to a holistic corporate focus by providing well-established data management and data-governance workflows as part of a dedicated, easily deployed infrastructure
- Facilitating agile platform growth that mitigates the need for repetitive and heavyweight infrastructure planning for every new platform requirement by defining a reference architecture to simplify the view of storage, compute, and software resources
- Providing easy integration with popular tools, enabling you to continue to use familiar workflows and data visualization, data management, and data analytics tool sets
- Breaking Hadoop out of the IT department so businesses can fully use its analytical capabilities on an organization-wide level
- Facilitating mixed Big Data lifecycle workloads and data analytics

Virtualized platform

The VMware vSphere virtualization platform, which enables the compute components of this solution, provides an efficient, agile, and manageable compute infrastructure in a virtualized environment.
vSphere empowers users to virtualize any application with confidence, redefining availability and simplifying the virtual data center. The result is a highly available, resilient, on-demand infrastructure that is completely transparent to the data scientists and data engineers working within Analytic Insights Module.

Several physical storage platforms support data storage in Analytic Insights Module. Shared, private, or temporary stores are provisioned and used to encapsulate physical storage. Analytic Insights Module can then use the provisioned storage for storing analytics applications. This enables you to take advantage of Isilon, Dell EMC™ VNX™, and Dell EMC XtremIO™ capabilities in Analytic Insights Module.

This solution features monitoring capabilities to provide a comprehensive view of the Analytic Insights Module environment. These capabilities are based on a combination of matrixes from both the compute and storage components.

**Solution architecture**

**Architecture**

The open-architecture approach that is used in the solution implementation is a key enabler of rapid development and delivery of solutions and applications on the Analytic Insights Module platform.

Figure 1 depicts the deployed solution including optional components.
Figure 1. Analytic Insights Module deployment
## Logical view

Figure 2 depicts the logical layout of Analytic Insights Module.

![Figure 2. Analytic Insights Module logical view](image)

### Compute

The compute option for Analytic Insights Module uses one of the following:

- **Dell EMC VxRail™ appliances**—Fully integrated, preconfigured, and pre-tested hyperconverged infrastructure appliance that is powered by vSphere

- **Dell EMC Vblock™ Systems**—Converged infrastructure system that includes all the required resources to field a highly available, secure, and scalable compute infrastructure

**Note:** Analytic Insights Module is available on a VxRail appliance. For very large deployments, Analytic Insights Module can be deployed on Vblock Systems.

### Storage

This solution uses the following types of storage:

- **Hadoop Distributed File System (HDFS) storage**—The persistent data store for Analytic Insights Module is HDFS, which is facilitated on an Isilon system.

- **Block storage**—Storage to support virtualization high availability in the compute cluster or as a general service to support other containers such as MongoDB and other NoSQL applications. Supported systems are:
  - VxRail appliances
  - VMware vSAN
  - Vblock Systems
  - VNX array
Chapter 2: Data Platform

Network

The primary Analytic Insights Module network is Ethernet, which is used for general access and connectivity between components and for access to the HDFS storage. A Fibre Channel (FC) storage area network (SAN) can also be deployed for block storage support.

- **Ethernet**—Dell EMC recommends using a high-speed, 10 Gigabit Ethernet (GbE) network between the compute and storage elements of Analytic Insights Module. A multilink, 10 GbE fabric provides the best balance between cost, implementation complexity, and reliable standards.

- **FC**—FC SANs can also be present when block storage elements use FC connectivity. The Analytic Insights Module reference architecture uses Vblock, which includes FC to connect servers to the VNX array, presenting shared storage for virtualization high availability.

Native Hybrid Cloud foundation

Native Hybrid Cloud is an engineered solution that features an elastic infrastructure and monitoring tools to support an application production environment. IT professionals can easily monitor, manage, and scale the environment to ensure its fitness for current and future application production needs.

Native Hybrid Cloud provides a predictable, accurate, and protected Pivotal Cloud Foundry (PCF) deployment that is sized correctly and installed on a VxRail appliance, a hyperconverged infrastructure appliance powered by vSphere. VxRail hardware sizing configurations are based on expected PCF usage and system testing. The virtual infrastructure is sized based on PCF usage. The Native Hybrid Cloud Monitoring and Reporting dashboard monitors and reports consumption and billing information transparently.

Operational Data Container

The ODC is a core component of Analytic Insights Module. It is designed for ingesting, integrating, hosting, and performing analytics on data originating from multiple sources. The ODC is a multidata container storage environment that primarily comprises Hadoop clusters, database systems, and indexing and cataloging subsystems. It supports the storing, tracking, and processing of structured and unstructured data of various types and from various sources. Data and applications that are published by data scientists are also retained in a secured area of the ODC. Operational data resides in the ODC and is ingested by data engineers in support of business analytics and sourcing data for data scientists and business analysts.

Dataset

A dataset is an abstraction that represents a collection of similarly structured or unstructured data. Table 2 lists some common possibilities for dataset types. Some characterizations overlap. For example, a set of key-value pairs can be stored as a file.
Table 2. Dataset structures

<table>
<thead>
<tr>
<th>Dataset structure</th>
<th>Typical examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>MongoDB document, XML document, RSS feed, financial results (XBRL), website content</td>
</tr>
<tr>
<td>Files</td>
<td>Flat file, audio recording, photo, process output file, log file, binary large object (BLOB)</td>
</tr>
<tr>
<td>Graph</td>
<td>Employee skills, transit routing, network assets, access permissions, web browsing behavior, gene sequences, social media influence, recommendation engine internals</td>
</tr>
<tr>
<td>Index</td>
<td>Content index, federated content index</td>
</tr>
<tr>
<td>Key-value pairs</td>
<td>Twitter hashtag and counts, sensor attributes, and measured values</td>
</tr>
<tr>
<td>Query or search result</td>
<td>Email messages from an archive, medical test results, current product suppliers, product pricing at a specified time</td>
</tr>
<tr>
<td>Report</td>
<td>Sales by region, top blog topics in last 24 hours, marketing test results</td>
</tr>
<tr>
<td>Relational table</td>
<td>Customers, employees, transactions, resource access history, time series summary</td>
</tr>
<tr>
<td>Stream of elements</td>
<td>Market ticks, surveillance video, messages, RFIDs, sensor data</td>
</tr>
<tr>
<td>Web service</td>
<td>Customer discount by volume, shipment tracking</td>
</tr>
</tbody>
</table>

Metadata

Analytic Insights Module includes the following types of metadata:

- **Technical**—Includes creation date, creator (and other aspects of lineage), file format, structure, and size
- **Business**—Includes scope (for example, global versus geo-specific, enterprise versus division, everything versus time-based increment versus changed-data only), business tags, data encodings, security and privacy indications, quality assessments, prior usage, and recommendations

Policies

Policies that apply to a dataset can be used to govern its location, visibility, security, access, quality, and tracking, and the operations permitted on the dataset (based on requester role and context). Operations can include, for example, deploying to a workspace or publishing.

Subject to policies, datasets and services in the DAC can be provisioned to workspaces. This provisioning can mean providing a “working set” copy of a dataset known to the DAC or providing dataset access through a service.

A user can have access to multiple workspaces. A user can also request that individual working sets and applications be published from a workspace when they are deemed ready to be shared more broadly with other data scientists or with the business.
Chapter 2: Data Platform

**Controller**

The Analytic Insights Module controller is packaged as a virtual machine (VM) that orchestrates workflows within Analytic Insights Module. It automates infrastructure operations such as VM creation and teardown. It provides internal services such as Lightweight Directory Access Protocol (LDAP), Domain Name System (DNS), and Dynamic Host Configuration Protocol (DHCP), and it controls the integration of third-party tools.

The controller functions as a key component of the DAC to enable the on-demand deployment of additional tools and services that users might require within their workspaces.

**Services**

PCF is a scale-out, rapid development and deployment platform that provides tailored access to elements of Analytic Insights Module for organizations building their own data-driven applications on Analytic Insights Module.

PCF applications are 100 percent cloud native in their architecture and development style. This enables all the deployment speed and operational benefits of a cloud native approach when new applications are deployed in Analytic Insights Module. The 12-factor application-development methodology encapsulates the guiding principles for this deployment style.

The 12 factors describe design principles that ensure your application is cloud native. They encompass guidelines such as the following:

- Build applications as stateless processes
- Do not depend on local storage
- Do not inject configuration into an application on demand, instead of shipping with it

Adhering to these principles is important because they constitute a sort of contract between the cloud platform and your application. If your application obeys the contract, the cloud platform can correctly run and manage the application.

Many of the engineered components in Analytic Insights Module, including the DAC and UI, were developed and deployed using the 12-factor application methodology.

**Workspaces**

A workspace is an isolated, managed container of VMs, containers, and working sets that are configured and deployed on demand within Analytic Insights Module. A workspace establishes the tools and the data that are accessible to a user, providing a private sandbox that enables data exploration, transformation, enhancement, collaboration, and business analysis for data scientists and other specialized users. A workspace facilitates the use of tools, business applications, and services, and the development of new analytics tools and models.

**Working set**

A dataset that is deployed to a workspace or created within a workspace is a working set. A working set can be deployed from a published dataset into a workspace to be used for analysis by data scientists. Deployment can be through copying the dataset into a local workspace container (an instance of MongoDB or an HDFS cluster, for example) or by...
providing a data service (for example, to reference the latest version of product pricing). Working sets can be published to the ODC as datasets.

Figure 3 depicts the Analytic Insights Module data lifecycle.

Figure 3. Analytic Insights Module data lifecycle
Chapter 2: Data Platform

Key components

Analytic Insights Module can be rapidly and automatically provisioned. The analytics layer is completely virtualized with vSphere running on Vblock Systems with analytics tools and automated provisioning and configuration. The Isilon platform provides the storage foundation, delivering the ideal balance of capacity and performance. The solution offers a choice of the most popular Hadoop distributions. The result is a highly available, resilient, on-demand infrastructure that is completely transparent to data scientists and data engineers working with Analytic Insights Module.

The key components of the foundation Analytic Insights Module solution are listed in the following hardware and software resource tables and are described in detail in the Analytic Insights Module Reference Architecture Guide.

Hardware resources

Table 3 provides an overview of the solution hardware requirements.

Table 3. Solution hardware

<table>
<thead>
<tr>
<th>Category</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage</td>
<td>• Isilon</td>
</tr>
<tr>
<td></td>
<td>• Dell EMC VNX</td>
</tr>
<tr>
<td></td>
<td>• Dell EMC XtremIO (optional)</td>
</tr>
<tr>
<td>Compute services</td>
<td>• Dell EMC VxRail 200F</td>
</tr>
<tr>
<td></td>
<td>• Dell EMC VxRail 200</td>
</tr>
<tr>
<td></td>
<td>• Dell EMC VxRail System 540</td>
</tr>
<tr>
<td></td>
<td>• Dell EMC VxRail System 340</td>
</tr>
</tbody>
</table>

Software resources

Table 4 details the software resources for the platform configurations.

Table 4. Solution software

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage infrastructure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dell EMC OneFS™</td>
<td>8.0.0.3</td>
<td>Operating environment for Isilon with HDFS and Dell EMC SmartConnect™ licenses</td>
</tr>
<tr>
<td>Dell EMC VNX</td>
<td>05.33.009.5.155</td>
<td>Operating environment for VNX</td>
</tr>
<tr>
<td>Dell EMC XtremIO</td>
<td>4.0</td>
<td>Operating environment for XtremIO</td>
</tr>
<tr>
<td>Dell EMC VxRail 200F</td>
<td>3.5</td>
<td>Operating environment for VxRail</td>
</tr>
<tr>
<td>Dell EMC VxRail 200</td>
<td>3.5</td>
<td>Operating environment for VxRail</td>
</tr>
<tr>
<td>Dell EMC Vblock System 540</td>
<td>6.0.10</td>
<td>Operating environment for Vblock</td>
</tr>
<tr>
<td>Software</td>
<td>Version</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Dell EMC Vblock System 340</td>
<td>6.0.10</td>
<td>Operating environment for Vblock</td>
</tr>
<tr>
<td><strong>Virtualization</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMware vSphere ESXi</td>
<td>6.0</td>
<td>Server hypervisor</td>
</tr>
<tr>
<td>VMware vCenter Server</td>
<td>6.0</td>
<td>vSphere management server</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlueTalon</td>
<td>3.1</td>
<td>Tool that enables data security, policy management, and query audit capabilities</td>
</tr>
<tr>
<td><strong>Analytics applications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hortonworks Data Platform</td>
<td>2.5</td>
<td>Hadoop distribution from Hortonworks</td>
</tr>
<tr>
<td>Ambari</td>
<td>2.4</td>
<td>Tool for provisioning, managing, and monitoring Hortonworks</td>
</tr>
<tr>
<td>Cloudera Distribution Hadoop (CDH)</td>
<td>5.9</td>
<td>Hadoop distribution from Cloudera</td>
</tr>
<tr>
<td>Cloudera Manager</td>
<td>5.9</td>
<td>Administrative and management portal for deploying and administering Cloudera Hadoop clusters</td>
</tr>
<tr>
<td>MongoDB</td>
<td>3.2</td>
<td>Highly parallel NoSQL document-oriented data store</td>
</tr>
<tr>
<td><strong>Ingest applications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attivio Active Intelligence Engine (AIE)</td>
<td>5.1.9</td>
<td>Highly scalable text-indexing data store for search and analytics</td>
</tr>
<tr>
<td>Attivio Data Source Discovery (DSD)</td>
<td>5.1.9</td>
<td>Data discovery software</td>
</tr>
<tr>
<td>Zaloni Bedrock</td>
<td>4.1.2</td>
<td>Data ingestion workflow platform</td>
</tr>
<tr>
<td>Spring XD</td>
<td>1.3.0</td>
<td>Open source framework for easy data ingestion</td>
</tr>
<tr>
<td><strong>Application platform tools</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Hybrid Cloud</td>
<td>1.2</td>
<td>PaaS</td>
</tr>
<tr>
<td>Pivotal Cloud Foundry</td>
<td>1.8</td>
<td>Platform for developing cloud-scale applications using DevOps</td>
</tr>
<tr>
<td><strong>Analytic Insights Module</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data and Analytics Catalog</td>
<td>3.1.0.0</td>
<td>Engineered component that publishes data and analytics tools from Analytic Insights Module</td>
</tr>
</tbody>
</table>
Chapter 3  Data and Analytics Catalog and Workspaces

This chapter presents the following topics:

Data and Analytics Catalog ........................................................................................................23
Working in workspaces ................................................................................................................26
Managing data in workspaces ......................................................................................................30
Analyzing realtime data in workspaces .......................................................................................33
Assisting data analysts with data analysis ....................................................................................34
Data and Analytics Catalog

Overview

The DAC is a service that runs within Analytics Insights Module. The DAC responds to requests using an API across the network environment.

The catalog that is exposed through the DAC server and the DAC Client Services uses Attivio AIE as the primary repository where information about the DAC assets is stored and maintained. Logically, the DAC includes and exposes information about data, applications, and services that are known to or currently deployed within Analytic Insights Module. The DAC is the basis for identifying and searching for data assets within the Analytic Insights Module environment.

As shown in Figure 4, the DAC Client Services runs within the Analytic Insights Module environment. The DAC tracks all assets in the environment—from assets that have been published or deployed into the data lake to assets within the workspaces of individual data scientists.

Functional roles

The DAC is designed to deliver an asset-management solution that is extensible in the field by Dell EMC Professional Services, highly experienced customers, or third-party vendors.

The DAC includes a management interface that enables designated authorized users to do the following:

- Search, assess, and publish available datasets to workspaces
Chapter 3: Data and Analytics Catalog and Workspaces

- Search, assess, and deploy available tools and applications to workspaces

**DAC components**

**Asset**
An asset is a discernible digital instance of an item or entity that has the characteristics of being addressable (for example, a datafile, a directory, a database, an executable file, an indirect link or reference). It can be independently copied or relocated.

**Asset class**
Asset class is a reference to a specific asset classification that can exist and be tracked within the DAC. For example, an asset class can identify an asset as a flat file, a database, an indirect link, or an application.

**Asset type**
An asset type is a reference to the type of asset within the asset class. For example, an asset class can be “file” and the asset type can be “text”, indicating that the asset is a flat file containing only text (printable character data). The asset class might be “database” and the asset type “mysql” to identify the asset as a MySQL database.

**Manifest**
A manifest is an artifact for representing an instance of an asset as an object, meaning it provides an object-oriented representation of an instantiated asset. The manifest provides a common way of representing and interacting with any instance of an asset regardless of the underlying nature and characteristics of that asset.

**Template**
A template represents an entire asset class or asset type within an asset class. Templates provide a mechanism for creating a manifest for a specific asset.

**DAC Services manifest**
A DAC Services manifest represents an instance of an asset or an asset class. The manifest is a representation of an asset instance from the perspective of the asset, metadata about the asset, and actions that the asset supports and that are enabled by the manifest.

By treating assets as objects, DAC Services does not need to implement asset-specific logic to perform the actions on assets that are required by the DAC.

The manifest provides a mechanism for providing metadata about the instance of an asset, and for implementing the logic behind the actions as required to support the DAC Services. The DAC Services must understand specifics about assets. It also must understand how to parse and interpret the content of a manifest, store the content in the DAC and provide a means of running the actions that are defined by the manifest and supported by the asset.

**DAC Services template**
A DAC Services template is a "skeletal" and generic representation of an asset of a specific class or class and type. Templates establish the types of metadata that are used to identify and describe any instance of an asset of its class or class and type. This information includes the logic behind the DAC actions as required by the DAC.
While a manifest describes a specific instance of an asset, a template describes and defines what the manifest for an asset looks like and contains. Actual metadata about a specific instance of an asset of a class or class and type is applied to a template to generate the manifest for that instance.

A new template can be authored and subject to authorization and the **PUBLISH** review and governance process. You can register and publish new templates into the DAC, which adds support for new classes and types of assets throughout the Analytic Insights Module environment.

**DAC workflows**

The DAC supports a number of operations for interacting with the assets and asset classes. These operations are APIs that are available from the DAC command-line interface (CLI) and the DAC Client Services. Some of the end-user actions are integrated with the UI for the data scientist to access from the workspace. Also, the DAC supports a powerful text-search capability for identifying assets and asset classes.

The DAC has an asset manager that stores and changes the asset metadata in the Attivio AIE data container. The physical assets (data or tools or applications) are stored in the PDC.

Table 5 lists the DAC generic actions.

**Table 5. DAC generic actions**

<table>
<thead>
<tr>
<th>Action</th>
<th>Origination</th>
<th>Destination</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPY</td>
<td>PDC</td>
<td>Workspace</td>
<td>Moves the physical asset to a workspace in a temporary location. A <strong>DEPLOY</strong> action follows this action.</td>
</tr>
<tr>
<td>DELETE</td>
<td>Workspace or PDC</td>
<td>Workspace or PDC</td>
<td>Moves the asset to the archive state. <strong>Note</strong>: Full implementation of this action is not available in this release.</td>
</tr>
<tr>
<td>DEPLOY</td>
<td>PDC</td>
<td>Workspace</td>
<td>Materializes an asset at a target location. For example, a dataset is placed inside a data container, and a tool is installed on the workbench virtual machine.</td>
</tr>
<tr>
<td>PUBLISH</td>
<td>Workspace or PDC</td>
<td>ODC or PDC</td>
<td>Prepares the asset to be shared by the users of the data lake. The physical assets are copied to the ODC or PDC.</td>
</tr>
<tr>
<td>REGISTER</td>
<td>Workspace or PDC</td>
<td>-</td>
<td>Creates a DAC entry for the asset that is to be recognized. <strong>Note</strong>: Use the appropriate asset template for the asset class.</td>
</tr>
<tr>
<td>UNREGISTER</td>
<td>-</td>
<td>-</td>
<td>Removes the entry from the DAC. <strong>Note</strong>: The physical asset is not updated, changed, or removed.</td>
</tr>
</tbody>
</table>

The DAC controls the resource and optimization requirements for the applications that can be deployed in Analytic Insights Module. It stores information about the applications that
are already deployed, what resources they are consuming, and, as shown in Figure 5, what applications are available to be deployed. It provides a means of integrating new applications into Analytic Insights Module through an application store concept whereby new tools can be provisioned on demand and new functionality can be rapidly added to the data lake.

Figure 5. DAC published tools

The DAC is based on Dell EMC experience with deployments of the various applications in numerous customer situations. Eliminating the need for a lengthy trial and error process, the DAC ensures that effective resource utilization and tuning parameters are applied to the deployment from the moment of instantiation.

Working in workspaces

Overview

A data scientist or other user in the consumer role creates analytical models, performs data analysis and generates reports, uses realtime data to work with the Internet of Things, and so on.

Workspaces are the central concept of Analytic Insights Module, providing self-service capability for users to access and analyze data. Workspaces provide secure access to the
data, tools to build analytics, and compute to analyze the data. Workspaces are private and personal spaces for the individual user, as shown in Figure 6.

![Figure 6: Workspaces view](image)

Users customize their workspaces based on their needs, and they can create, modify, or delete workspaces as required.

The IT administrator manages the quotas that are associated with the workspaces for all users to monitor resource consumption.

**Note:** The solution allows one user per workspace. In a future release, the solution will allow for multiple users per workspace.

### Viewing catalogs

A workspace provides a catalog view of assets that are contained within that workspace. The catalogs options are:

- Data catalog
- Services catalog
- Workbenches catalog

Data for the catalogs is populated from the DAC.

### Data catalog

The data catalog, as shown in Figure 7, lists all the datasets that are published from the data lake and the registered datasets present in the workspace. You can choose to subscribe to datasets from the catalog, publish registered datasets to the data lake, or tag datasets to be searchable from the text search.
Chapter 3: Data and Analytics Catalog and Workspaces

Figure 7. Data catalog

**Services catalog**

The services catalog lists instances of the data containers, tools, or applications available for use in the workspace, as shown in Figure 8. You can add or remove services using the services catalog option.

Figure 8. Services catalog

**Collaborators catalog**

The collaborators catalog lists other consumers who have access to Analytic Insights Module. You can invite listed consumers, as shown in Figure 9, to the workspace to collaborate in your workspace.
Workbenches catalog

The workbenches catalog lists the workbench virtual machines that are attached to the workspace, as shown in Figure 10.

Creating workspaces

Use the UI hosted in PCF to deploy and manage assets in a workspace. The UI provides a self-service capability to administer your work area. The UI is a web-based tool that is securely accessed from your internal network. You authenticate to the UI with a user ID and password pre-configured by the security administrator.

Every new workspace includes a Linux 7.0 workbench VM, which has command-line accessibility. You can use the workbench VM to store datafiles or install your preferred analytical tools. It comes preinstalled with the DAC Client tools that are required to register datasets with the DAC. Each workbench is also preconfigured with an xRDP Server to provide greater ease of use for xRDP-compliant remote desktop clients.

A workspace is also attached to a space connected to PCF. The PCF space can be used to build and run the cloud applications for data analytics.

By default, the system assigns the creator of the workspace as the administrator. This enables the workspace creator to add datasets, tools, or applications to the workspace.

Adding data services

When you add a data service, the DAC installs it to the workspace, and the utilized quota that is associated with the workspace is updated to reflect the resources that are consumed from the data lake. Each group of artifacts behaves the same way, with additional parameters that must be provided during the setup.

Data containers are part of the data services that are used to store data in the workspace. Other services can include analytic tools, analytic applications, or data ingest tools.

Subscribing to published datasets from the DAC

To subscribe to a published dataset, select a target data container. If the data container does not exist, the published data is made available to the default workbench virtual machine file location. During the subscription process, you can add a data container to the
workspace and place data into it. The DAC deploys the new data container to the workspace and deploys the data into the container.

You can also deploy one or more empty data containers to a workspace. You have full access to the data containers after they are deployed in the workspace. Deployment of a data container is a long-running process depending on the container type you choose. During deployment, the status of the data container is updated and displayed. After the data container is fully installed, connection strings to the data container are provided so that you can access the container.

**Subscribing to tools**

You can deploy tools to a workspace by choosing from a list of available tools. By default, the chosen tools are deployed to the workbench virtual machine. You can use the tools to analyze data in the workspace.

**Adding a Windows virtual machine**

Analytic Insights Module facilitates the installation of a Microsoft Windows VM. This optional VM can be added during the installation of Analytic Insight Module.

The VM facilitates the installation of applications such as Tableau.

*Note:* Microsoft Windows and Tableau licenses are required.

**Deleting workspaces**

You can remove an artifact from a workspace or delete a workspace entirely. Before you use this option, Dell EMC recommends that you contact the IT administrator to create a backup of the workspace. The delete function does not automatically trigger a backup. When you request the deletion of a workspace, all the artifacts, including data, tools, data containers, and workbench virtual machines, are removed before the workspace is deleted.

**Managing data in workspaces**

You are in full control of managing the data within a workspace. Data that is manually added to or created in a workspace is unknown to the DAC unless you register the data in the DAC. If the dataset is registered but not published in the DAC, it is not visible to other workspaces. In this case, data remains private to a workspace.

**Registering data**

Data that is added to a workspace through Attivio DSD and related Zaloni Bedrock ingestion workflows is automatically registered in the DAC. You do not need to register the data.

**Storing data**

Workspaces store data locally in any of a number of supported data containers. The data containers in a workspace are hosted on local virtual instances of a data platform or within PCF data services.
Tracking data
Workspace data and data lineage are tracked from:
- Local private workspace copies of ODC data (under security governance)
- Data that is ingested through Attivio DSD and Zaloni Bedrock

Adding datasets
Import new datasets into a workspace in any of the following ways:
- Using the ingest tools available in the Hadoop clusters in the workspace
- Using the Attivio DSD and Zaloni Bedrock ingestion process to ingest data
- Using Spring XD
- Manually importing data into a data container
- Deploying a dataset from the data lake (ODC or PDC)
- Subscribing to a published dataset from the data lake (as shown in Figure 11)

![NCDC Weather Data Extracted](image)

Figure 11. Adding a dataset by subscription

All added datasets reside either in a data container within the workspace or in the file system of a workbench virtual machine.

Deleting datasets
Delete unused datasets from the workspace, as shown in Figure 12. Deleting a registered dataset in a workspace also removes the metadata entries from the DAC and deletes the data from the data container or the workbench file system. Deleting a subscribed dataset from a workspace also removes the subscription entries from the DAC and deletes the data but does not delete the source data.
Transforming datasets

Prepare data that is collected from various sources by combining datasets, transforming and normalizing the data, and so on. You can then use this data in the analytical models. The workspace provides tools and data containers to prepare new datasets to be tested against the models. To create the prepared sets, use the Hadoop-based Extract, Transform, and Load (ETL) tools or install Spring XD, R, or Python, all of which can be deployed to the workspace from the UI.

Registering datasets in the DAC

Registering a dataset in the DAC enables you to view the dataset from the data catalog in the workspace. However, dataset registration is not required unless you want to publish the dataset from the workspace. In that case, you must register the dataset in the DAC before you can complete the publication request. Use the DAC CLI to register the dataset from a workspace.

Registering data and other assets in a workspace is useful for tracking the data and tools in a workspace.

Publishing and deploying datasets

Publishing a registered dataset enables other users to be aware of and use the available datasets.

You register the dataset in the DAC, which makes the dataset visible in the workspace data catalog. You then submit a request to publish and deploy the dataset. A snapshot of the dataset is created for the data governance team to review, and the original version of the datasets is released for use in the workspace, as shown in Figure 13.
Tagging datasets
You can add text tags to any dataset in the workspace data catalog. These tags are free-form text fields that provide context to enhance the usability of the datasets. Tagged datasets can be discovered within a workspace data catalog through a text search.

Analyzing realtime data in workspaces

Overview
From the Internet of Things, you can derive actionable intelligence or hidden trends from the continuously streaming information sets. Workspaces are integral to enabling analytics in realtime over information streams. Dell EMC recommends that you ingest a selected stream directly into a workspace application instead of ingesting all the available data into the data lake. You then have time to assess and analyze the selected intelligence.

To analyze streaming data in realtime, you can create a workspace, add a stream-handling service available from the services catalog, and add a data container in which to perform the task. Analytic Insights Module supports several options to handle streaming inputs, including the following:

- Spring XD and Flow services provide a framework to ingest streams from a variety of different sources and formats. These sources and formats include file systems, email, Twitter streams, RabbitMQ integration, JMS, MQTT, Apache Kafka, MongoDB, Redis, Splunk, and Pivotal GemFire continuous queries. The streams are ingested in real time, and then analyzed through a set of online filters and classifiers to extract actionable values.
- Spark services from Cloudera and Hortonworks Hadoop distributions enable you to ingest and analyze streaming data in memory.
- Cloud-scale applications (12-factor applications) that you create and run in a PCF space in the workspace provide an efficient mechanism to perform realtime analytics.
Workspaces, which are the foundation for these complex data-analytics use cases, can host and run code ranging from as-needed to sustainable and scalable.

Assisting data analysts with data analysis

**Workspace tools**
The majority of enterprises employ data analysts, who might not be data scientists. These analysts might not have experience using SQL or NoSQL tools. They might prefer to use tools, such as Tableau or Business Objects, to create trending or reporting applications.

**Lightweight workspaces**
Data analysts can install lightweight workspaces with MySQL or MongoDB in Analytic Insights Module and provide a foundation for data analysts to generate analytic value for their business. Data analysts can quickly create these lightweight workspaces with published data from the PDC without any support from IT. They can also access and use the data container to develop reports or analyze data.
This chapter presents the following topics:

- **Introduction** .......................................................... 36
- **Security solution** ....................................................... 36
- **Securing Hadoop clusters** ......................................... 38
- **BlueTalon security infrastructure** ............................... 40
- **DAC security** ............................................................ 44
- **Workspace security** .................................................... 48
- **ODC security** ............................................................. 52
Chapter 4: Security

Introduction

Overview

A major concern as organizations begin to harness Big Data to derive actionable business insights is security of the data and protection of sensitive information. Advances in Big Data have overtaken traditional security methods in scale and in the need to be dynamic and real time. As data is transferred within the system to allow analysis and collaboration, data lineage must be preserved.

In the Analytic Insights Module, the Data Governor provides data access security for the solution regardless of whether the data is in a central repository (ODC) or in the user’s sandbox environment (workspace).

Security solution

Because of the complexity of this solution, we designed and built the security for Analytic Insights Module in layers, as shown in Figure 14.

Figure 14. Security layers for Analytic Insights Module

Layer 1: External security firewall
Layer 2: User and access security control
Layer 3: Data and data path security
Layer 4: Network security for workspaces (future release)
Layer 5: Audit and surveillance (future release)

Layer 1 (external firewall) is the customer’s responsibility to provide. Dell EMC can provide guidance regarding the rules that are required to protect the environment.

Layers 2 and 3 are currently provided in the solution and are described in this chapter.

Layers 4 and 5 will be addressed in a future release of Analytic Insights Module.

Security architecture

Figure 15 shows that security layers 2 and 3 are implemented on the solution components.
Analytic Insights Module is not directly accessed through customer networks. Users must go through the network-level security first. After clearing security, standard users assigned to the consumer role have unrestricted access to datasets, data containers, and virtual machines. This is currently a customer responsibility. An internal security firewall will be provided in a future release of the solution.

### Layer 2: User and access security control

#### DAC security

User access control is provided using a standard OAuth2 process on the UI. The OAuth2 authorization grant type is “authorization code”. When a user logs in to the DAC UI, OpenLDAP authenticates the user through the PCF User Account and Authentication Service (UAA).

#### Workspace security

User access control is provided based on workspace user credentials to access machines or data in their workspace.
DAC security
Data access control refers to the back-end interaction with the Attivio database that holds all the DAC assets (datasets, applications, services, and so on). The request for a dataset from the Attivio database is intercepted and security controls are applied.

Workspace and ODC security
BlueTalon provides data access control. Policy Enforcement Points (PEPs) intercept any data requests and modify the query that is based on identity. ODC PEPs are separate from the workspace PEPs because the ODC is on a different Hadoop cluster.

A user who wants to take action on a resource makes a request to a PEP. The PEP creates a request that is based on the requester's attributes, the resource being queried, the action, and other information pertaining to the request. The PEP then sends the request to a Policy Decision Point (PDP), which assesses the request and any policy that applies to the request, and then provides an answer about whether access should be granted. That answer is returned to the PEP, which then allows or denies access to the requester.

Note: The solution allows one user per workspace. In a future release, the solution will allow for multiple users per workspace.

Securing Hadoop clusters
Overview
Hadoop clusters are located in the following areas in Analytic Insights Module:

- ODC is the central data repository. This data is available to every user that has access to Analytic Insights Module. Security controls are based on the user's identity.
- Workspace is a sandbox environment for a specific user to perform analytics on data. The workspace consists of the following types of data:
  - New data that is introduced by the user
  - Data that is copied from the ODC

There are two approaches to authenticating to a Hadoop cluster:

- Simple (pseudo) authentication, which places trust in users' assertions about who they are.
- Secure authentication, in line with best practices, which Hadoop provides. Secure authentication also relies on widely accepted corporate user stores (such as LDAP or Active Directory) so that a single source of identity can be used for a credential catalog across Hadoop and existing systems.

In the solution, the standard approach is to provide the simple security model, as shown in Figure 16, for both the ODC and workspaces.
Figure 16. Simple security model

Figure 17 shows how the ODC provides the secure authentication by "kerberizing" the ODC Hadoop cluster with Kerberos. Kerberos is a third-party authentication mechanism in which users and services rely on the Kerberos server to authenticate each to the other. This is an optional feature for Analytic Insights Module and is provided through an Dell EMC Professional Services engagement.
Chapter 4: Security

Figure 17. Secure security model

Note: If Kerberos is used on Isilon storage, Kerberos restricts the number of Hadoop clusters that you can deploy on that cluster. Each workspace represents a new Hadoop cluster. The maximum number of Hadoop clusters that are allowed per Isilon cluster is 50.

BlueTalon security infrastructure

Overview

BlueTalon is the primary security infrastructure for data and data-path security control in the solution. The data governance team writes their policies around this security control. Analytic Insights Module uses a policy-enforcement-point mechanism. During user access, role-based access control (RBAC) and attribute-based access control (ABAC) security models are applied over the data path. Available data must go through this policy enforcement.

BlueTalon intercepts data during a data-path access request to:

- View the access control list (ACL) permissions that are based on the ABAC policies to determine if the user has access rights.
- Determine what elements, if any, the user has access to. The user only sees the authenticated data that they have permission to view.
The principles that support the data-access control are:

- **Strong coverage**
  - Secure data access against data stores and file systems (broad coverage)
  - Secure HDFS, relational database management systems (DBMS), and NoSQL stores (heterogeneous coverage)
  - Deploy a spectrum of techniques (encryption, query modification, tokenization, dynamic masking, obfuscation, and so on)

- **Policy-based access control**
  - Enforce policy at the point where data is accessed (low overhead)
  - Provide a unified and centrally managed view of policies (consistency, simplicity)
  - Extend RBAC to ABAC (flexible policies)

- **Integrated access control**
  - Integrate with PCF security (in the Analytic Insights Module on Native Hybrid Cloud co-deployment)
  - Integrate with partner software in Analytic Insights Module (for example, Attivio searches within the DAC)

- **Extensible controls**
  - Add and enforce policies for separation of duties
  - Create and enforce more general policies (for example, for dataset placement)

**XACML standard**

eXtensible Access Control Markup Language (XACML) is an OASIS standard that describes both a policy language and an access-control decision request/response language. The policy language describes general access-control requirements, and has standard extension points for defining new functions and data types, combining logic, and so on. The request/response language enables you query whether a particular action is allowed, and interpret the result. The response always indicates whether the request is allowed by using one of several values: Permit, Deny, Indeterminate, or Not Applicable.

XACML defines the following elements:

- **PEP**—The part of the architecture that intercept requests and send them to the PDP to enforce the defined policies from the Policy Administration Point (PAP).
  - HDFS and Hive
  - Relational and NoSQL data sources
  - Within Attivio processing – see DAC security

- **PAP**—The part of the architecture that manages policies. In the platform, in most cases, the PAP is in the Data Governor. A PAP can also be in an application outside the platform, for example, a surveillance application.

- **PDP**—The part of the architecture that evaluates access requests and makes access decisions; likely to be in the Data Governor.
BlueTalon is based around a subset of the XACML standard. The primary elements of the BlueTalon security solution are the:

- **Policy engine**—which provides runtime allow, deny, filter, and mask decisions to PEPs based on user identity, group, data, metadata, business criteria, locations, and so on.

- **Audit engine**—which provides a single point to visualize who is using what data.

- **PEP**—which transparently intercepts user requests for data to enforce security (using the policy engine for decisions) and send audits from across platforms to the audit engine.

The BlueTalon policy engine gives you the option to centrally manage the authorization policies and unify the handling of the policies.

**PEP types**

In Analytic Insights Module, the BlueTalon PEP types that are provided include Hive PEP and file system enforcement point (FSEP) for HDFS, with the following advantages:

- Security administrators can define policy through a single web interface, in a unified policy model. There is no requirement to log in to each Hadoop cluster, which can be running different scripts or CLIs for policy definitions.

- The BlueTalon policy engine provides an API interface to integrate high-level Analytic Insights Module orchestration workflows. Developers can disregard the details of how the policies are implemented in different Hadoop clusters.

The various BlueTalon PEPs intercept queries from applications and send them to the BlueTalon policy engine, as shown in Figure 18.

![Figure 18. BlueTalon architecture](image-url)
Policy attributes
BlueTalon enables precise authorization. Administrators can author and manage role, attribute- and business-based data-access policies from a single, easy-to-use console without the need to be familiar with any query logic.

An attribute is a variable for which a value can be determined at runtime and that can be used to make policy decision. An attribute is used in a record filter qualifier within a rule.

A session attribute is determined for each client session and is either specified by the end-application (specified session attribute) within a query comment, or determined automatically by the intercept (detected session attribute) within the protocol, client driver, network, and so on.

Policy engine
With the BlueTalon policy engine, shown in Figure 19, you can apply significantly more detailed control over data than with alternative solutions, including the ability to filter data at the column, row, cell, or partial cell levels. Combined with dynamic data masking and stealth analytics to further protect sensitive data, you get the most granular control over any data without the need for duplicating data or views.

![Figure 19. BlueTalon policy and audit engines](image)

The BlueTalon policy engine serves as a level of control between users, applications, and enterprise data repositories. BlueTalon controls and audits the data that can be seen and modified across various on-premises and cloud-data sources.

Audit engine
The BlueTalon audit engine records the query that is made and the modified policy-compliant query that defines the data that is consumed. The audit engine monitors who attempted to access data, when, and whether they were allowed or denied access. Requests made by users and the modified policy-compliant queries can be viewed in the BlueTalon audit console.

BlueTalon is transparent to queries and applications. The BlueTalon PEP intercepts queries from applications and sends them to the BlueTalon Policy Engine. The Policy Engine parses the queries according to defined policies and rules. A policy-compliant query is returned to the Enforcement Points and a policy-compliant dataset result is returned to the user.
The high-level query process is:

1. The application connects to the BlueTalon PEP by providing a valid identity (in a Java Database Connectivity (JDBC) connection string), which is addressed to HiveServer2.
2. The application requests data, and then the PEP extracts a query from the request session and passes it to the BlueTalon policy engine.
3. The policy engine validates the query and user ID against the defined policies, modifies the query according to the rules, and returns a modified query back to the PEP.
4. The PEP passes the modified query to HiveServer2, gets results, and passes the query back to the application. During this step, a superuser is used to connect to Hiveserver2 from the PEP.
5. The previous steps are repeated until the application disconnects.

DAC security

Overview

The DAC provides an interface for external applications and users to access data that is stored in Analytic Insights Module components such as Attivio. To ensure that Analytic Insights Module data is correctly accessed, you must protect the DAC service APIs and DAC UI.

The DAC security is designed to enforce unified security and data-governance policy across the DAC, providing granular access control and efficient operation and maintenance.

Identity management

Different components that are related to the DAC solution, for example, PCF or Attivio, have their own identity-management systems. To simplify identity management and to reduce the security risks during operation, you need a central identity-management system. This enables an administrator to manage the DAC user IDs, groups, and credentials from a single location.

In this solution, we use OpenLDAP to centrally store and manage the DAC identities.

Central authentication

DAC users can keep a single credential with one-time login to access the DAC. This method simplifies the DAC user-side operations and also reduces the need for users to keep multiple credentials for different components.

DAC users or external applications can access the DAC service from the UI, UI application, or service APIs. The OpenLDAP is used as a back-end authentication server.

Authorization

To ensure that the correct person has access to authorized DAC resources, we implemented an RBAC system to enforce granular access control. A role database centrally manages the RBAC system.
Currently, there are several defined roles:

- **Consumer**—Explores and examines data from multiple disparate sources, and runs analytics to address business problems. A consumer can build dashboards, reports, and applications.

  The consumer role is usually assigned to data scientists, business analysts, application developers, or any other customer who is served by the capabilities of the workspaces within Analytic Insights Module.

- **Data administrator**—Understands all the sources of data and creates a plan for integrating, centralizing, and maintaining all the data. A data administrator must be able to understand how the data relates to the current operations and the effects that any future process changes will have on the use of data in the organization.

  The data administrator:
  - Creates the data source connection that enables the consumer to access the data and run analytics
  - Reviews and approves or rejects consumer requests for additional workspace resources

- **Security administrator**—Adds users to Analytic Insights Module. The security administrator approves and provides access to the datasets that are defined by consumers.

- **IT administrator**—Configures the system hardware and software. The IT Admin installs database systems, writes complex queries, scales the systems to multiple machines, manages backups, and implements disaster recovery systems. The IT administrator also configures the solution for users in other roles.

The role database contains different tables that define the mapping relationships between roles, LDAP groups, and permissions. We use Spring Security, such as PCF UAA, OpenLDAP, and BlueTalon, to implement the authorization.

The UI is a service portal for DAC administrators and DAC users to perform operations. Generally, DAC administrators or users must be able to access different resources through the UI:

- Resources in PCF, for example, PCF service instance
- DAC service REST API, for example, publish dataset
- Web resource, for example, URL resource of the UI server

**Accessing resources through the UI**

**Authentication**

The UI authentication system has two components:

- PCF UAA server is integrated with the OpenLDAP server
- PCF UAA acts as an OAuth2 authorization server, while the OpenLDAP server is responsible for back-end user authentication and identity storage

The OAuth2 authorization grant type is “authorization code”, which determines the grant access process used. When a user logs in to the DAC UI, OpenLDAP authenticates the user through PCF UAA.
**Authorization**

After the user authentication, the OAuth2 authorization server grants JSON web token (JWT) access to the UI application (which acts as an OAuth2 client). The access token delegates authorization for that user.

The DAC resources authorization is based on the scope that is defined in the access token. There are predefined scopes that are known to the PCF UAA. Customized scopes that are defined in the role database can also be redefined to meet new requirements. These scopes work as permissions. Users (groups) are assigned with different scopes that are based on their roles.

Spring Security (and Spring Security OAuth2) options validate the access token and evaluate whether the user has permission to access claimed resources that are based on the scopes that are retrieved from the access token. Depending on the resource type being accessed, the authorization process might be slightly different.

When accessing a DAC REST API resource, authorization is based on customized scopes that are defined in the role database. The UI application sends a query to the DAC Service API with the access token. The DAC Client Services retrieves the scopes from the role database for authorization.

---

**Note:** To protect the role database, applications cannot access the role database directly. All authorization-related queries are initiated through a dedicated authorization service API.

---

**Accessing applications with the DAC Client**

The DAC Client Services provides public-service APIs to external applications. There are different types of applications that must be able to access DAC Client Services API resources:

- UI (web application), which calls the DAC Client REST APIs according to the user's operation on the UI
- UI application, which is a CLI interface to enable users to operate DAC resources
- Other external applications that may need to access REST API resources

**Authentication**

To meet the requirements of different access types, DAC Client provides multiple authentication modes:

- Basic user authentication by OpenLDAP user and password
- OAuth2 authentication through OAuth2 (PCF UAA) integrated with OpenLDAP
- Session authentication that is based on basic authentication
- Session authentication that is based on OAuth2 authentication

Session-based authentication makes the authentication process more efficient because basic authentication or OAuth2 authentication requires each request to be authenticated.

UI (web application) authentication is based on OAuth2 authentication (or session-based OAuth2 authentication). For details, refer to Accessing resources through the UI.
The UI application uses basic authentication (or session-based basic authentication). OpenLDAP acts as authentication server and ID store. The DAC Client service redirects the authentication request to the OpenLDAP server. With basic authentication, the DAC Client service can run independently without PCF or PCF UAA.

Other external applications can use both basic or OAuth2 authentication methods that are based on their requirements.

**Authorization**

Authorization is similar to what is described in *Accessing resources through the UI*. Users or groups are assigned with roles that are defined in the role database, and each role maps to related permissions. When a user logs in, the DAC Client Services queries the role database for the permission that is assigned to the role, and evaluates whether the user has permission to access the requested resource.

**Storing data with Attivio AIE**

Attivio AIE is a central repository of the DAC for storing all metadata about assets in workspaces, the ODC, and the PDC. The UI application or external applications must call the DAC Client Services API to access resources stored in Attivio AIE. They cannot access Attivio AIE directly. This design largely mitigates the security risk that is posed by exposing Attivio AIE to a public environment.

Attivio AIE has its own authentication and authorization system. To enforce the unified RBAC rule and data-governance policy, we added BlueTalon and OpenLDAP to build a granular-access control and audit system.

OpenLDAP acts as the identity store for the whole DAC system. Users and groups are centrally defined and used for access control rule definition.

The BlueTalon policy engine works as a PDP to centrally define the access-control rules that are based on LDAP users or groups. The BlueTalon policy engine provides a PDP for the DAC, ensuring that users only see what they can access when they query the catalog.

**Back-end security (Elasticsearch modification)**

Back-end security control is required to ensure that datasets that are transferred into the workspace are controlled. This security control is in addition to the finer-grained access control provided after data has been copied into a workspace.

The operations of this feature are:

1. The DAC gets requests from a DAC Client service (UI, CLI, and so on) and creates an Elasticsearch, which will extract the information from Attivio AIE that the DAC request needs.

2. Before sending the Elasticsearch payload to Attivio AIE, the DAC sends the search payload with the LDAP username to BlueTalon through an API that supports Elasticsearch.

3. BlueTalon modifies the Elasticsearch payload that is based on user-defined policies and returns a modified Elasticsearch payload (including filters that are based on any BlueTalon policies in the context of the LDAP user).
4. The DAC sends the modified Elasticsearch payload to Attivio AIE, which returns the results that are based on any modifications BlueTalon made to the original Elasticsearch payload.

The Elasticsearch modification process is shown in Figure 20.

Figure 20. Elasticsearch modification

**Workspace security**

**User access**

A workspace is a sandbox environment where a data scientist can use a set of preferred tools to analyze datasets. The elements of a workspace must be secured. Security access controls ensure that the user can access the elements of the workspace.

A workspace consists of a PCF element and non-PCF virtual-machine elements. For the PCF element, standard PCF authentication through UAA and LDAP is used to access the relevant parts of the workspace.

Applications that users deploy to PCF exist within a PCF space. Spaces exist within organizations. To view and access an organization or a space, a user must have permission. PCF uses RBAC with each role-granted permission to either an organization or a specified space. In the PCF element of a workspace, these controls regulate access.

The non-PCF elements of the solution are deployed by the Analytic Insights Module controller and are configured by the DAC.

A user (User A in Figure 21 and Figure 22) accesses the environment through the internal network of the organization. To use the workspace, User A must use remote desktop protocol (RDP) to access the workbench virtual machine. RDP access is controlled by an LDAP group that has been created specifically for the workspace.
The workbench virtual machine contains tools that require HDFS access. These tools can connect to HDFS through the FSEP using webHDFS. Connections strings are provided for access in the Analytic Insights Module UI.

User A might also want to use client tools that are available in the worker nodes or access a management component in the worker nodes. This is achieved by using SSH to connect to the workbench virtual machine. SSH access is also only possible if User A is a member of the relevant LDAP group.

Clients in the workbench nodes connect to HDFS through the FSEP using webHDFS. Connections strings are provided for access in the Analytic Insights Module UI.

Controls are added specifically for HDFS, as shown in Figure 21. These controls:

- Block protocols that are not necessary
- Change ports of PEPs and services to ensure that connections are made to the PEP, which proxies the service, instead of the service

![Figure 21. User access control—HDFS](image)

For Hive and Spark, access is achieved in the same way, as shown in Figure 22.

Additional controls are added specifically for Hive. These controls:

- Blocks protocols that are not necessary
- Changes ports of PEPs and services to ensure that connections are made to the PEP, which proxies the service, instead of the service
Chapter 4: Security

Data access control

The data in a workspace that resides in a Hadoop system comes from various sources:

- New data that is ingested from external sources into the workspace by the data scientist who created that workspace
- Data that is ingested from internal sources within Analytic Insights Module

Data can be accessed in several ways:

- HDFS
- Hive
- Spark Thrift Server

The HDFS structure is applied as data domains. Data resides in several areas:

- The root of the HDFS file structure, which is used for new data that is created by the workspace user. A rule in the BlueTalon policy server that allows access only by the workspace user is required.
- A subfolder to the root, which is used to copy files from the ODC. A default rule that denies access to every user of Analytic Insights Module is required.

Specific files are copied into the ODC folder. When they are transferred to the workspace, the rule governing access to these files is also added to the workspace data domain.
Figure 23 shows the rules and directories.

Allow all (workspace users in LDAP group for workspace) rule

Deny all (everyone) rule

Specific rule from ODC for specific users (policy mapping)

A similar approach is taken for Hive and Spark. A default database is used for tables that are transferred from the ODC. A new sandbox database is created for new tables. The rules are applied as shown in Figure 24.

Figure 23. HDFS BlueTalon policies in workspaces

Figure 24. Hive BlueTalon policies in workspaces
ODC security

**Data access control – simple security**

The ODC is part of the standard deployment of Analytic Insights Module and is deployed as an “unkeberizered” Hadoop cluster. This method relaxes the restriction that is associated with the number of Hadoop clusters.

In this solution, access to the data is protected by the Data Governor.

BlueTalon protects the data from unauthorized access by users and service users, including:

- DAC
- Zaloni Bedrock

Figure 25 shows how a user, a data administrator who is named Bob in this example, uses the ODC data access control for HDFS.

![Diagram of ODC data access control for HDFS](image)
ODC data access control for Hive and Spark is shown in Figure 26.

The ODC also can deploy Kerberos for secure data access control.

Figure 27 shows the logical architecture of the security system using Kerberos.
Hadoop clusters exist in the following areas of Analytic Insights Module:

- ODC
- PDC
- Workspaces

Users must be able to reliably identify themselves and have that identity propagated throughout the Hadoop cluster. Then, users can access resources (such as files or directories) or interact with the cluster (for example, running MapReduce jobs). Besides users, Hadoop cluster resources (such as hosts and services) must be able to authenticate with each other to avoid potential malicious systems or daemon’s posing as trusted components of the cluster to gain access to data.

Hadoop can use Kerberos as the basis for strong authentication and identity propagation for both user and services. The Kerberos server is known as the KDC. At a high level, the KDC has several parts:

- A database of the users and services ("principals") that the KDC knows about and their respective Kerberos passwords
- An authentication server that performs the initial authentication and issues a Ticket Granting Ticket (TGT)
A Ticket Granting Server (TGS) that issues subsequent service tickets that are based on the initial TGT

A principal user requests authentication from the authentication server. The authentication server returns a TGT that is encrypted using the principal user's Kerberos password. This TGT is known only to the principal user and the authentication server. The principal user decrypts the TGT locally using its Kerberos password, and until the ticket expires, the principal user can use the TGT to get service tickets from the TGS. Service tickets are what allow a principal to access various services.

Because cluster resources (hosts or services) cannot provide a password each time to decrypt the TGT, they use a special file, a “keytab”, which contains the principal resource's authentication credentials. The set of hosts, users, and services over which the Kerberos server has control is known as a “realm”.

**Note:** By default, in the current solution, the clusters in the workspace and the ODC are non-kerberized.

**Managing users with OpenLDAP**

LDAP is a method of organizing information and providing access to it. It is commonly used for user, service, and machine information. OpenLDAP is an open-source implementation of the LDAP, and is used as the central identity-management database. Linux uses OpenLDAP to manage authentication and information (login credentials, user ID, home-directory location, and so on).

**Kerberos KDC and OpenLDAP**

The Kerberos KDC is integrated with OpenLDAP. OpenLDAP servers work as the back-end database of the KDC. KDC data, such as principals and keys, are stored in OpenLDAP. This can map Kerberos principal users to OpenLDAP user entities (User ID (UID) and Common Name (CN)), and make it easy to define the authorization policy.

**Implementing Kerberos with OpenLDAP**

Kerberos is generally used with OpenLDAP. After a user is authenticated using Kerberos, OpenLDAP controls what the user can do. Replicating a Kerberos principal database between two servers can be complicated, and adds an additional user database to the network. Kerberos can be configured to use an LDAP directory as a principal database.

OpenLDAP implements most authentication options through the simple-authentication session layer (SASL), a network protocol designed for authentication.

By default, the LDAP server **slapd** runs as user and group LDAP, while the keytab file is readable by root only. Therefore, you must either change the LDAP configuration so the server runs as root or make the keytab file readable by the group LDAP.

For Hadoop clusters in Analytic Insights Module, Kerberos requires that the user has a valid Kerberos credential. This is not practical in many environments because every user who interacts with the Hadoop cluster must have a Kerberos principal configured. LDAP protocol is used where clients authenticate themselves using their username and password.
**HDFS with Isilon**

The HDFS access zones use OpenLDAP as external identity provider. There is no requirement to create users on Isilon one by one, because users are created on Isilon for all nodes.

**BlueTalon audit and policy engines**

The BlueTalon audit and policy engines integrate with OpenLDAP. Security administrators can define user- and group-based policies for granting access and authorization.

**Hadoop nodes**

Hadoop worker nodes (Linux OS) are integrated with OpenLDAP through a system-security-service daemon (SSSD). This method simplifies the user management on Hadoop nodes. There is no requirement to create users on Isilon one by one, because users are created on Isilon for all nodes.

---

**Using principals with Hadoop and Kerberos**

Each service and subservice in Hadoop must have its own principal. A principal name in a specified realm consists of a primary name and an instance name. The instance name is the fully qualified domain name (FQDN) of the host that runs that service. Because services do not log in with a password to acquire their tickets, their principal authentication credentials are stored in a keytab file. The keytab file is extracted from the Kerberos database and stored locally in a secured directory with the service principal on the service component host, as shown in Figure 28.

![Figure 28. Hortonworks Data Platform (HDP) and Kerberos keytabs and principals](image-url)
This chapter presents the following topics:

- **Introduction** ..................................................58
- **Data ingestion workflows** ....................................58
- **Working with unstructured data** .............................64
- **Working with structured data** .................................64
Chapter 5: Data Ingestion

Introduction

You can ingest data into Analytic Insights Module from different sources for immediate use or for storage in the ODC. Data can be ingested as soon as it is issued by a source or imported in discrete chunks at periodic intervals. It is more efficient to ingest data by prioritizing the data sources, validating individual files, and routing data items to the correct destination.

Big Data can comprise hundreds of sources in multiple formats, which makes ingesting data at a speed that ensures smooth performance difficult. Analytic Insights Module ingests data quickly and easily using automation and software to structure and organize the incoming data. The data can be analyzed immediately or stored and analyzed later using the built-in analytics programs.

It is best practice to process all unstructured data as it is ingested. However, individual data scientists might import data directly into their workspaces.

Data ingestion workflows

Figure 29 shows that Analytic Insights Module supports multiple methods of data ingestion, which can involve structured and unstructured data, from either internal or external (raw) data sources:

- External data ingestion into the ODC
- External data ingestion directly into a workspace (“dirt road”)
- Controlled external data ingestion into a workspace (“paved road”)
- Internal data ingestion
Data ingestion roles

Data ingestion is typically performed by the following:

- Data administrators can:
  - Ingest external data into the ODC
  - Refine raw ODC data (for example, clean, standardize, and summarize) or create ODC sample datasets for data scientists
  - Configure Attivio DSD to reference refined and sample ODC datasets as data sources for data scientists
- Consumers can:
  - Explore and mine data from external data sources or refined and sample datasets that are hosted in the ODC
Data administrators can bring external data into the ODC. The raw data may be refined, and can then be used later for analytics, as shown in Figure 30.

**Data administrators can:**
- Create a batch or realtime workflow in Zaloni Bedrock and run it to ingest data from external data sources
- Monitor and administer Zaloni Bedrock workflow runs
- Support complex data transformations in Zaloni Bedrock workflows
- Search the stage area for lookups or master data for enrichment
- Add ingestion rules or definitions to data elements
Consumers can bring external data into their workspace by processing it through tools such as Attivio DSD, as shown in Figure 31.

**Figure 31. Ingesting paved-road data into a workspace**

Consumers can:

- Access Attivio DSD from the workspace UI
- Browse discovered data sources, review sample data, run a semantic search of data sources
- Create datasets from available data sources
- Discover and create data models
- Ingest data into a data container of a workspace
- Run ingest processes on demand

**Note:** You do not need an operating knowledge of Zaloni Bedrock.
Chapter 5: Data Ingestion

Ingesting dirt-road data into a workspace

Consumers can manually bring external data into their workspace without processing it first, as shown in Figure 32.

Consumers can:
- Use ingestion applications from Hadoop clusters to source external data
- Build ingest-workflow pipelines to gather and transform data before loading it to workspace containers
- Support “bring your own device” (BYOD) use-case wrangling in realtime or batch data feeds
- Register the data sources in the DAC
- Develop machine-learning algorithms using the Spark platform on Hadoop

Figure 32. Ingesting dirt-road data into a workspace
Ingesting internal data

Data administrators can ingest data to create refined data for consumers, as shown in Figure 33.

**Figure 33. Ingesting internal data**

Data administrators can work directly with Zaloni Bedrock to:

- Define data sources and ODC targets
- Define ingestion workflows

Security administrators can directly access the BlueTalon policy UI to fine-tune ODC security policies as required.

With internal ingestion:

- Ingestions import data from external data sources, landing raw data in the ODC
- Metadata about new ODC data is cataloged as newly "published" datasets
- Raw data is refined to better support data analytics
- BlueTalon instantly secures access to any new data added to the ODC
Chapter 5: Data Ingestion

Working with unstructured data

Prerequisites
You need credentials to launch the DAC with sufficient unused quota to create workspaces and to use CDH, for example, to process data from outside Analytic Insights Module.

Unstructured data
Analytic Insights Module uses an Attivio DSD hub with Zaloni Bedrock, which provides enterprise data-lake management and governance software and services to monitor external sources. Zaloni Bedrock can ingest and process any volume or type of unstructured data based on a data-preparation trigger predefined by a data engineer. DSD can mark the data with tags and other metadata before storing the data in the ODC.

Analytic Insights Module eliminates the need to archive data by processing and storing any data type at massive volumes in active storage with no need to archive it. This enables quick and easy data access for end users and applications.

The ODC can be used for staging external datasets of unstructured data and for running analytics on the dataset.

You can create a workspace, deploy a Hadoop cluster, CDH for example, and select a service. After you create a service instance, you can copy data, for example, using Hadoop User Experience (HUE). HUE enables you to copy file to the share location. You must create a Hive table and a schema. You can add previously uploaded datasets to the Hive table. You can then run the Hive queries and send the results to a file for publishing back to the ODC.

Working with structured data

Prerequisites
To create the workspace, you need credentials to launch the DAC with sufficient unused quota.

Datasets must be registered in the DAC and published with appropriate tags in Analytic Insights Module before you can add them to the workspaces you create.

Structured data
You can run analytics on structured data that is already available and published within Analytic Insights Module. Analytics tools can be added to the workbench or through a virtual machine that is connected to Analytic Insights Module.

You can create a workspace, add a dataset, and deploy a working set. You can then analyze the new dataset using a program such as RStudio, and publish the resulting output in a format such as CSV back to the solution.

Note: You can ingest either unstructured or structured data into a workspace and later publish the dataset to the ODC or PDC.
This chapter presents the following topics:

Introduction ................................................................. 66
Use case overview ............................................................ 67
Use case data workflow ...................................................... 67
Collaborating in workspaces ............................................. 78
**Introduction**

Managers regularly ask data scientists or business analysts to find new insights in data. For example, a manager might ask a data scientist to analyze customer sales data to determine where the sales team can send its best sales representatives to help close deals for the quarter.

In a traditional scenario, the data scientist moves the data to an external unsecured environment in the public cloud. The data scientist can quickly set up and run the environment without having to log a request with IT. In only a few hours, the data scientist can be analyzing the data. However, because the data is sensitive and proprietary, and relevant to deals the sales team is trying to close, there is a serious risk of a data leak.

The data scientist spends many hours cleaning and analyzing the dataset, but can only save it in isolation in the external environment, which means that nobody else on the team can access and reuse it later.

This scenario creates several problems:

- Moving the data outside the internal network creates a security risk.
- Setting up a secure environment can take six to eight weeks for an IT administrator to build for a data scientist. The data scientist has already missed the end of quarter sales and therefore, the best sales representatives cannot win the critical deals.
- Processing data in isolation means that other data scientists might not know it is available or that it has already been curated. This method creates duplication of effort and means datasets are not as complete or useful as they can be.

Analytic Insights Module resolves these problems by:

- Creating a single, secure data lake environment for a high volume of different types of data from multiple sources
- Providing data scientists and other users with the ability to create their own workspaces within the Analytic Insights Module environment. They can then curate and analyze data and create applications.
- Making curated data available to all users by publishing it in the Analytic Insights Module environment.

**Note:** We use “data scientist” to refer to data scientists, business analysts, application developers, or any other person in an organization who explores and examines data from multiple disparate sources, and runs analytics to address business problems.
Use case overview

Manager Solange asks Jenna, a data scientist, to provide a data analysis for their newly launched eco coffee grinder as soon as possible. She asks Jenna to identify which of their customers on the west coast are most likely to buy the product, with the right incentive, before the end of the quarter. Solange wants to send their best sales representative, Mandeep, to the customers before the end of the quarter, which could result in winning the customers’ business.

Jenna logs in to the Analytic Insights Module environment and opens her workspace. She creates a new working set and enters the search criteria in the DAC to find relevant data, including all the west coast customers and their purchase histories of other coffee products. She finds a matching dataset that was published by her colleague three days earlier. This will save her time, because she can use the data that was already curated. Jenna imports this dataset to her workspace and imports data from a social media source that was flagged by the Attivio DSD with the hashtag #ILoveCoffee.

Jenna spends a day wrangling the raw data. When everything is cleaned up and saved, she runs the RStudio service so that she can begin her data analysis.

Jenna finishes her analysis and draws her conclusions. She highlights five customers who have previously purchased beans in bulk, and who have all expressed an interest in the eco coffee grinder. She sends this report to Solange, who assigns Mandeep to work with the sales representative for each customer. The sales team makes presentations to each of the five customers, resulting in three of them agreeing to purchase the product.

Jenna adds some additional details to her working set, then saves and publishes it as a new dataset in the DAC, so that it can be used again.

Use case data workflow

Data workflow summary

Bob, a data administrator, uses Zaloni Bedrock to ingest the data (as shown in the workflow in Figure 34), and performs minimal data transformation to provide clean data for downstream analysis. The company’s data scientists can then access the information in their own workspaces. Using tools of their choice, they can derive and publish actionable insights or develop applications that are based on the requirements of their end users (customers).
Figure 34. Data workflow summary

Data workflow details

Bob’s workspace includes customer transactional information in the ACME-DEProduction workspace, as shown in Figure 35. Bob can use tools such as RStudio and Spark, which are available on his workbench within the workspace, to make changes and aggregations to the data.

Figure 35. Viewing My Workspaces
Bob opens the **ACME-DEProduction** workspace, to see all datasets, services, and tools that are available, as shown in Figure 36.

![Figure 36. Viewing ACME-DEProduction workspace contents](image)

Bob adds additional datasets to the **ACME-DEProduction** workspace, as shown in Figure 37.

![Figure 37. Adding datasets](image)
Bob selects several datasets and clicks **Ingest New Data**, as shown in Figure 38.

![Published Datasets Table]

Figure 38. Ingesting datasets

Bob can then use RStudio (as shown in Figure 39) or Spark to pull the data into his workspace and aggregates it to two datasets:

- Customer profile data
- Aggregated transaction data

![RStudio Code]

Figure 39. Using RStudio to get data
In RStudio, Bob creates a general-purpose function to derive insights and import various data types, such as CSV, Spark Data Frame, or Hive, as shown in Figure 40.

Figure 40. Using RStudio coding to identify and import multiple data types

Bob deploys datasets that are addressable with the name that is given by the DAC. He opts to apply more descriptive names for the datasets. Bob also aggregates multiple data types, as shown in Figure 41, so they are cast correctly.

Figure 41. Using RStudio to aggregate data types

Bob performs multiple aggregations on the data to create a profile for each customer. The profile examines all past customer transactions, including:

- Total number of transactions
- Average value of transactions
Chapter 6: Use Case

- Total value of transactions
- Types of transactions, for example, brokerage, insurance, or banking
- Last transaction date
- Average time between transactions

Later, the Analytic Insights Module users on Bob’s team can use this information for downstream processing to determine details, such as the values of an individual customer, or which sales campaign might be more effective when presented to customers with a similar values profile.

Bob saves the datasets as `cusprofile1.csv` and `tr5data1.csv`. After Bob completes the aggregations, the two output files are available in the workbench of the ACME-DEProduction workspace, but not in the ODC.

Bob can use the Zaloni Bedrock workflow or the DAC CLI to publish the datasets to the ODC so that they are available for others to use.

After publishing, the datasets can be added to any other workspace, for example ACME-Sales as shown in Figure 42.

![My Workspaces](image)

**Figure 42. Adding datasets to a different workspace**

The same datasets are available to multiple departments from their open workspaces, including Sales, as shown in Figure 43.

In the Sales department, Preeta can deploy the data, run more aggregations or modeling, and make that available to her customer.
Preeta, working in the Sales workspace, can access the same cusprofile1.csv file in RStudio that Bob created from his workspace, as shown in Figure 44.

The application that Bob created in the ACME-DEProduction workspace can be run in the other workspaces. This enables Preeta to run the code in the Sales workspace, which reads the data and opens the application in her browser, as shown in Figure 45.
Figure 45. Running Bob’s custom application

Preeta then uses RStudio to display the aggregated data and provide a visual insight for the data. The initial window provides a view of the sum of the total value that is grouped by continent. Preeta can use other tools to perform this type of work but opts to use Bob’s custom application and RStudio to do her cluster analysis, as shown in Figure 46.

Figure 46. Running a cluster analysis
Preeta verifies the quality of the data by examining the cluster using pairwise plotting, trend analysis, and forecasts, as shown in Figure 47, Figure 48, and Figure 49.
Preeta looks at segments of the data, and applies filters. As shown in Figure 50, she can use filters for region, branches, and products. She can also filter out customers who do not want to be solicited.
When Preeta wants a more in-depth review of each customer, she uses the **Customer 360 – Profile** option, as shown in Figure 51.

![Customer 360 - Profile](image)

**Figure 51. Using the Customer 360—Profile option**

To save time and effort, Bob creates and publishes a simple **Customer Profile Analysis** application using PCF that enables Preeta to quickly and easily perform this set of tasks, as shown in Figure 52.
Collaborating in workspaces

Manager Solange asks Alice, a data scientist, to provide a data analysis for why their new “Hi Protein” drink product is not selling effectively in small markets. Solange asks Alice to identify demographic data about their customers in large markets versus the smaller markets. Larger markets in big cities are most likely to buy their product. Meta Drinks wants the marketing team to create campaigns for small markets to grow market share by the end of the quarter, which could result in growing business. Alice logs in to the Analytic Insights Module environment and opens her workspace, as shown in Figure 53.

Alice creates a new working set and enters the search criteria in the DAC to find relevant data, including all the small-market customers and their purchase histories of other competing protein drink products. She finds a matching dataset that was published by her colleague, Bob, a week ago. This will save her time, because she can use the data that he already curated.

Alice imports this dataset to her workspace and spends a day wrangling the raw data. She realizes she does not have enough business knowledge about small-market customers to
complete the work. She invites Bob into the workspace in the same division by clicking **Add Members** as shown in Figure 54.

**Figure 54. Inviting a collaborator to a workspace**

Alice selects Bob from the list of collaborators shown in Figure 55.

**Figure 55. Selecting a collaborator**
Alice’s workspace now has two members as shown in Figure 56.

![Alice’s workspace with multiple members](image)

**Figure 56.** Alice’s workspace with multiple members

Alice views the collaborators list, which now has two collaborators as shown in Figure 57. Alice can remove Bob from the workspace when their collaboration is complete.

![Alice’s workspace with multiple collaborators](image)

**Figure 57.** Alice’s workspace with multiple collaborators

Bob logs into Analytic Insight Module and navigates to his My Workspaces view. He now sees both his workspace and Alice’s workspace, as shown in Figure 58.

![Bob’s My Workspaces view](image)

**Figure 58.** Bob’s My Workspaces view
Bob can access Alice’s workspace and view the datasets that Alice is working on, as shown in Figure 59. Bob cannot invite or delete other collaborators to Alice’s workspace.

Figure 59. Bob’s view of Alice’s workspace

Bob adds and modifies the metadata of the working sets that Alice has been working on. He also connects to Alice’s workbench using the connection string that is provided, as shown in Figure 60.

Figure 60. Workbench connection string
Chapter 6: Use Case

After Alice and Bob clean and save all the data, Alice runs her preferred tools for the data analysis. Alice finishes the analysis and draws her conclusions. When the project is complete, Alice removes Bob as a collaborator in her workspace, as shown in Figure 61.

Figure 61. Deleting a collaborator from a workspace

Alice highlights five small markets where they can target young athletes as part of the campaign. She sends this report to Solange, who sends it to the market campaign team. As an experiment in the targeted small markets, the marketing team creates a campaign to target high-school football teams. The sales team tracks product sales and continues to expand the campaign based on early feedback.
This chapter presents the following topics:

Conclusion ...

...84
Conclusion

Businesses are recognizing that combining Big Data technologies with new data sources and analytic approaches generates new opportunities. However, many are struggling to identify the best opportunities and convert them into measurable business value.

Analytic Insights Module is a fully engineered solution that makes enabling Big Data initiatives radically simple. Preconfigured building blocks, core Dell EMC technologies, and an open ecosystem let you focus on building new capabilities instead of a new infrastructure. A flexible, self-service, virtualized ecosystem enables you to prove the value quickly, scale out rapidly, and keep security and governance under control. Removing the difficulties of integration enables you to quickly realize the value of analytics.

Analytic Insights Module solution:

- Delivers business value quickly by providing a complete engineered solution for managing and analyzing data and for rapidly developing applications
- Provides an intuitive platform to all users across the enterprise
- Reduces risk by using proven technologies from Dell EMC companies, and provides seamless support and service
- Features an open and extensible platform that scales effortlessly to meet the needs of the enterprise, both in terms of dataset sizes as well as system performance requirements
This chapter presents the following topics:

Dell EMC documentation ...........................................................................................................86
Dell EMC documentation

The following documentation on EMC.com or EMC Online Support provides additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell EMC representative.

- Analytic Insights Module Reference Architecture Guide
- Analytic Insights Module Data Protection Solution Guide
- Analytic Insights Module System Monitoring Solution Guide
- Analytic Insights Module User Guide
- Analytic Insights Module Administration Guide
- Analytic Insights Module Release Notes
- Native Hybrid Cloud Solution Overview
- Design and Implementation for Native Hybrid Cloud Service Overview