Solution Brief

CTEC and Denodo

A Data Management Platform for Cloud and Hybrid Data Fabric Deployments

Introduction

Significant strides in cloud-based service delivery have been made in the last decade with infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS) deployment options. Cloud automation, implemented with continuous integration and continuous delivery (CI/CD) has increased the pace of solutions deployments. Cloud security models have also become more robust and have propelled many agencies to take on migration and modernization efforts to move legacy and on-premises applications to the cloud. For Federal Civilian and Defense agencies, this has become even more compelling, now that many FedRAMP Moderate and FedRAMP High certified options are readily available from AWS, Azure, Google, IBM, Oracle, Salesforce, and others, along with tools and security features for hybrid-cloud and multi-cloud deployments. CTEC's innovation center of excellence in cyber and cloud operations (iCyClOps™) brings leading-edge solutions to customers for their digital transformation journeys. We see there is an urgent need for agencies to embrace hybrid/multi-cloud data management approaches – it is projected that > 80% of agencies will leverage multi-cloud deployments by 2025, enabling them to increase IT productivity, enabled by AI/ML capabilities for faster data delivery options accelerated by automation. For federal agencies, the following requirements therefore need to be addressed efficiently and securely for the effective migration and modernization of platforms, capabilities, and tools:

- Data hosting – With on-premises repositories or through a cloud service provider (CSP) with required FedRAMP certifications and requisite FISMA Impact Level (IL) certification levels
- Security – With fully enabled on-premises or cloud security and governance features

- Master data management — To manage master data across disparate silos, maintain shared data sets, and ensure uniformity, accuracy, ownership, semantic consistency, and accountability

- Metadata management — Providing oversight of data associated with data sets that can be integrated, retrieved, shared within the organization, subject to security protocols setup, linked, and analyzed

- Document management — To track, manage, and store documents and records.

At a high level, the pillars of an enterprise data architecture are *data collection, data curation, data preparation, data provisioning, data consumption, data governance and security, and data monitoring* and operations. We highlight a leading data management solution from our technology partner Denodo Technologies, Inc. (Denodo), which uses a data virtualization platform recognized as a leader by Gartner and Forrester. This platform, called the Denodo Platform, provides a single master data abstraction layer that meets all the general and specific requirements for handling data both in the cloud and on-premises and provides an integrated dashboard for managing data across disparate sources with very low latencies, all without the need for data replication.

To manage all of these processes, both on-premises and in the cloud, agencies need to be well equipped to handle all aspects of data management including the ability to retrieve information on-demand - to enable transactions and business analytics with highly informative dashboard management. First, we highlight below some of the challenges with handling data on this transformation journey, and then we transition to discussing the attributes and benefits of the data virtualization platform from our iCyClOps™ technology partner Denodo. method for integrating data. Rather than replicating data to a central repository, data virtualization provides real-time, integrated views of the data across myriad heterogeneous sources, and these include on-premises and cloud repositories, and structured or unstructured sources.

Data Integration Challenge:

Most large enterprises are faced with data sprawl. The data ecosystem at agencies has, over time, developed and deployed many fit-for-purpose database systems. In the modern data landscape, data is no longer restricted to relational databases or line-of-business (LOB) systems. The typical agency profile is an enterprise with a heterogenous data landscape, which can include applications such as relational database management systems (RDBMS), Hadoop systems, data warehousing applications, and SAP implementations, along with legacy mainframe enterprise applications like DB2, SQL, and Adabas, which have limited integration points for the whole variety of data formats (structured, semi-structured, and unstructured data), data repositories, and platforms for different tasks and application requirements. Data exists in dozens of different file formats, hosted in hundreds of databases or servers, thousands of data messages exchanged per channel, and millions to billions of streaming devices.

Data Integration Challenge

A wide and heterogenous data landscape, typical of agencies, has data sources that host and serve many varieties and velocities of data. A data integration approach whereby data must be replicated, physically transported from one zone to another, and continuously transformed from one form to another is no longer a sustainable, scalable, or efficient approach. Users need the ability to integrate data from various sources, enrich and curate it using machine learning (ML) and artificial intelligence (AI) driven models, and access it from a unified interface that acts as a gateway to the enterprise data universe.

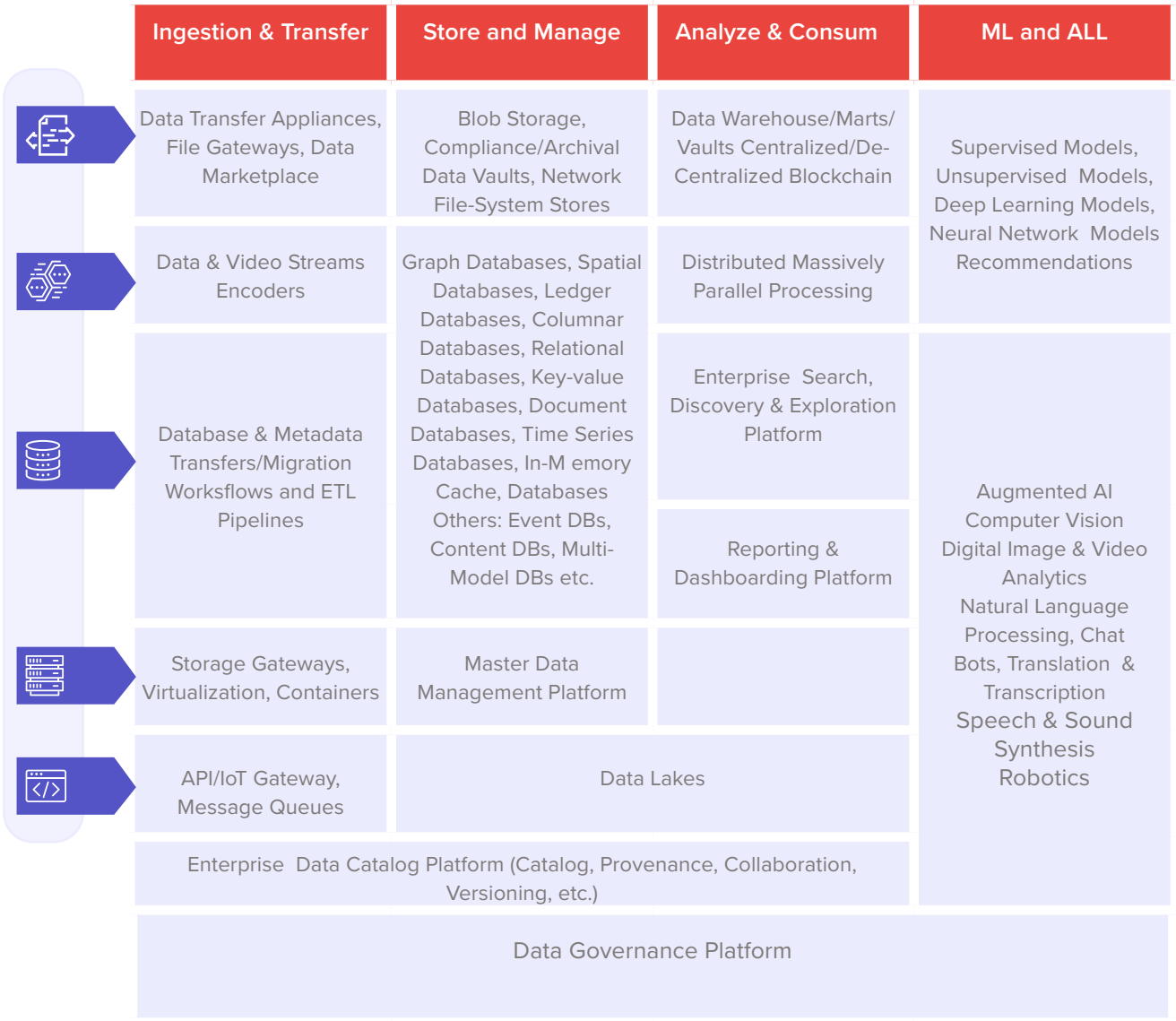


Figure 1: Elements of the Modern Data Journey Landscape.

This tremendous expansion in the data origination space has led to broader channels of data consumption and an increasing variety of systems and platforms that process the data to enrich and optimize it. Figure 1 above depicts an overview of the major components and platforms that make up a modern data landscape, in which data may originate in one form and pass through one or more of the components depicted in the diagram before ending its journey in the form of various consumption patterns [1]. There are multiple means of storing and managing the data, and multiple channels for consuming the data.

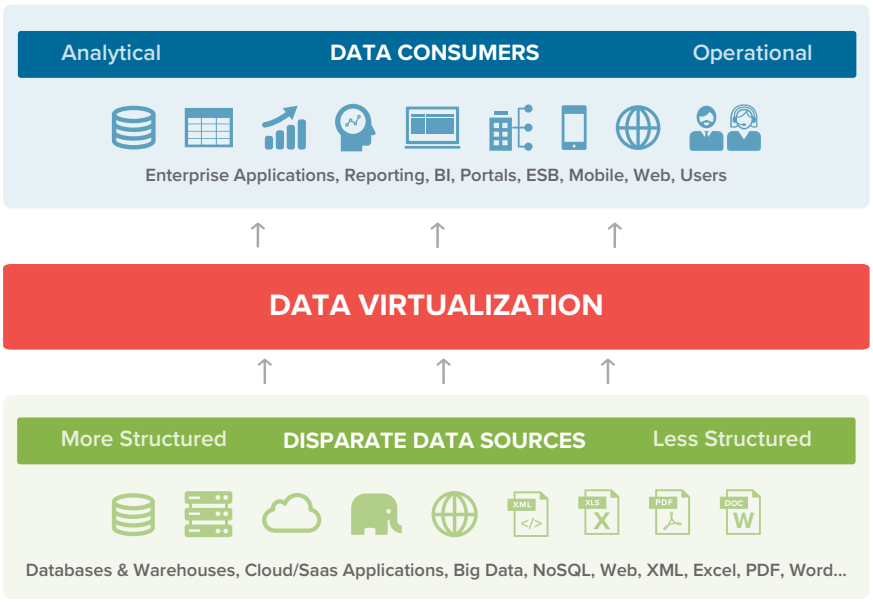
There are numerous challenges associated with typical data integration that result in significant costs and overhead to the agency, impacting business value and time-to-market. Data is typically hosted in a variety of fit-for-purpose data sources that may be located on-premises as well as in the cloud. When users want to dynamically integrate data from different siloed source systems, users often have to manually access these systems and integrate the data locally using various data analysis or exploration tools. Data integration teams supporting the IT departments are asked to take on the onerous task of integration. This is often a point-to-point data integration exercise which is neither scalable nor cost effective. It is difficult to integrate the disparate on-premises and cloud data sources. The common methodology of data integration involves several data hops from the data source, in which the data is physically transported, transformed, repeatedly replicated (this involves creating physical copies of data along with transformations), and fine-tuned or optimized for the method of consumption. Large and complex datasets can be in multiple different data formats, and of large volumes, from terabytes to petabytes in

scale. The traditional approach of transporting, transforming, and warehousing the entirety of the data in a one-size or one-type-fits-all data repository is time, resource, and cost-intensive. The ever-expanding volume, variety, and velocity of data poses some unique challenges. In addition, traditional tools cannot integrate streaming data and data-at-rest in real time. Consumption systems are tied to the last hop or copy of the data. In order to ensure that the data consumed is not stale, especially in cases where data needs to be made available in near real-time, a continuous or frequent data refresh cycle across all hops of data is required. This is resource intensive for large datasets, and often not scalable. Change data capture (CDC) is often the most complex and time-consuming process, and implementing CDC across all hops of data increases the cost and effort of implementation. With data getting replicated all over the data landscape, data governance becomes harder to manage, leading to risks related to missed governance. It is difficult and costly to maintain consistent data access and governance policies across data siloes. With ever-evolving data integration needs, it takes too long to get answers, which causes a direct and significant impact to mission and business agility. This can result in a significant impact to the efficiency of an agency's business operations.

Data Virtualization

Data virtualization presents a modern approach to data integration. It provides a logical data layer that integrates all enterprise data siloed across the disparate systems, manages the unified data for centralized security and governance, and delivers it to business users in real time. Unlike ETL solutions, which replicate data, data virtualization leaves the data in source systems, simply exposing an integrated view of all the data to data consumers. As business users drill down into reports, data virtualization fetches the data in real time from the underlying source systems, connecting to data in real time, which is far superior to collecting and replicating it. Data virtualization provides a 67% reduction in data preparation effort and a 65% decrease in delivery times over ETL [1]

Data virtualization uses a simple three-step process:



Consume data in any mode - reports, dashboards, portals, mobile, and Web applications

Combine any type of data - structured, streaming, social, and flat files

Connect to any data source - databases, data warehouses, cloud applications, big data repositories, Excel files (data wrangling to consolidate from different sources)

Figure 2: Data Virtualization Abstraction Layer

The table below summarizes the attributes of data virtualization:

Logical Data Layer – Data Abstraction	Data virtualization provides a virtual approach to accessing, managing, and delivering data without replicating it in a physical repository. It abstracts access to disparate data sources; acts as a single virtual repository – abstracts data attributes such as location, format, protocols – hiding data complexity for ease of use. It presents an end-user (or consuming application) with a uniform and abstracted data layer, so that the data sources can be logically combined and accessed in a uniform manner without accessing each siloed data source separately. From an end-user experience, this is similar to accessing a single local data source with the ability to create a virtual model by combining data objects from various data sources. It is not an ETL process. Because the underlying data is not replicated (zero replication, zero relocation – leaves data at its source, extracts only what is needed on demand), the need for effort-intensive ETL process steps is greatly reduced, significantly reducing data redundancy and its associated costs. Supporting data architecture components like the metadata catalog, user roles and policies, data format conversions, and data archival strategies are easier to implement, resulting in significant cost and resource savings through reduced development time and overall TCO.
Data integration	Data virtualization integrates data siloed across all enterprise systems, regardless of data format, location, or latency. It helps modernize business applications while replacing legacy systems; enables migration from on-premises applications to cloud applications; and facilitates the search for and discovery of interrelated data while adhering to governance protocols and complying with regulations that require the traceability of historical data
Data Management	Data virtualization provides a centralized secure layer to catalog, search, discover, and govern the unified data and its relationships - enables information discovery and self-service by facilitating access to all data (internal and external), enables the creation of universal semantic models reflecting business taxonomy, connects, and unlocks data from silos and legacy systems. Data virtualization platforms do not store data. They create virtual views of data on demand. Data virtualization is not a governance tool or a metadata management tool. However, data virtualization has many capabilities in the areas of governance, metadata management, and security, and works well with tools in those areas. It provides for centralized metadata, security, and governance by presenting data to the user via an abstracted layer in a unified view – it simplifies data security, privacy, and audit by abstracting data source security models and enabling single-point security and governance, extending single-point control across cloud and on-premises architectures, providing multiple forms of metadata (technical, business, operations) to facilitate the understanding of data. It is not an orchestration engine. Data virtualization compliments architectures built on services, APIs, or microservices.
Real-Time Delivery	Data virtualization provides real-time access to data across the enterprise, including data at rest or in-motion, arriving through streams. Data is accessible on demand without the need to engage an IT team for point-to-point or case-specific data integrations. Integrated information is delivered in real time to users and/or consuming applications, enabling the efficient execution of complex processes. This enables timely decision making and real-time provisioning of data to meet demands for analytics and transactional data. Real-time logical views of data can be created across many data sources, supporting transformations and quality functions without the latency, redundancy, and rigidity of legacy approaches.

Data Virtualization Reference Architecture

Figure 2 below depicts a high-level data virtualization reference architecture for big data and cloud, emphasizing operational use cases; how and where data virtualization can fit into the overall data journey, depending on different data use cases; reporting; and analytics. Data virtualization addresses the twin aspects of data-as-a-service, namely IT semantics – where data is stored and processed, and business semantics – how data is consumed and used.

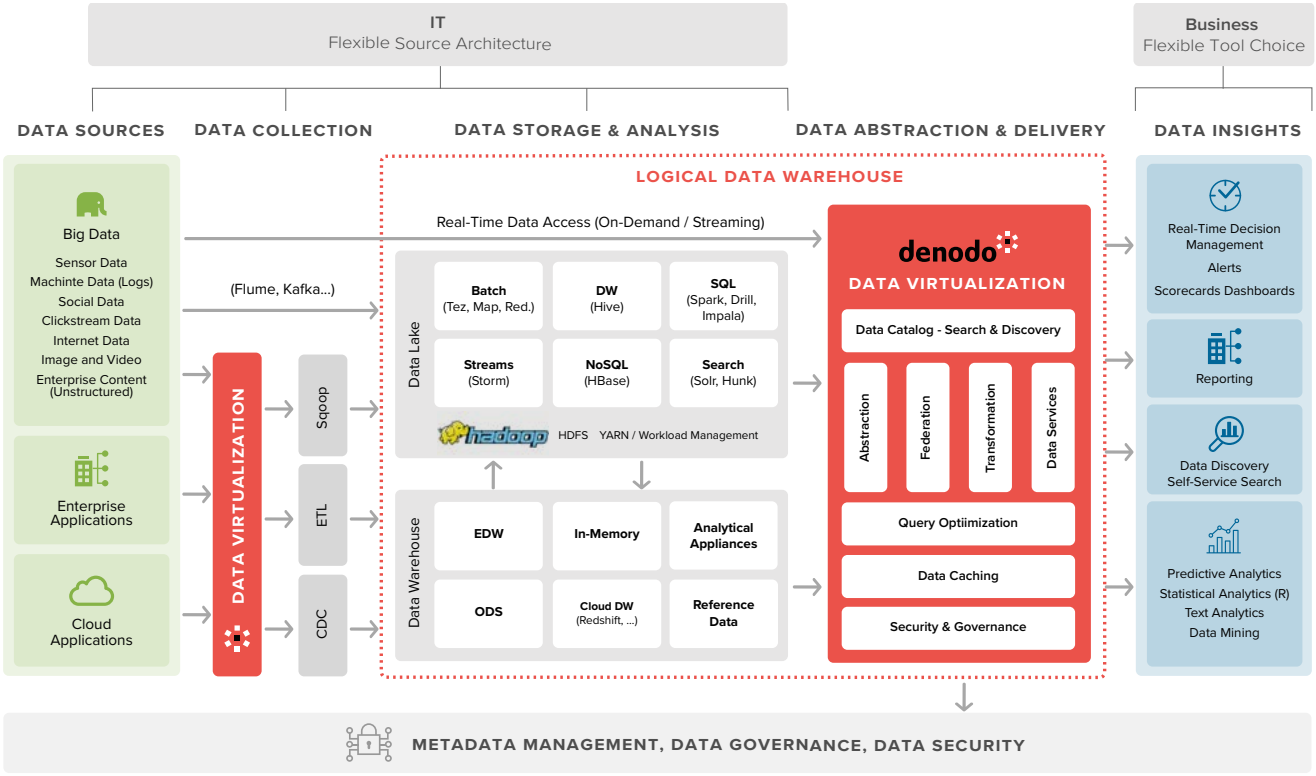


Figure 3: Reference Architecture for flexible business intelligence and analytics

Depicted in Figure 3 above are the elements associated with the data sources (data warehouse, enterprise applications, web, cloud, SaaS), data collection, data virtualization, and data analytics engines. The items in the blue boxes pertain to IT infrastructure, with a flexible choice of the hardware, software, and data enterprise tools, while those in the red boxes pertain to the data virtualization layer. Data virtualization decouples data sources from data consumers. This enables IT to change underlying systems without affecting users. This partition enables IT to move at a slower speed without affecting business processes, which can move at a faster pace, and make sophisticated decisions, as all data is accessible by any tool.

Why Denodo?

Some examples of commercial vendors in the data virtualization space are Tibco Data Virtualization, Denodo, Actifio, and Data Virtuality's Logical Data Warehouse. Denodo is the leader in data virtualization. With more than 20 years of innovation, the Denodo Platform offers enterprise-grade data virtualization with an easy-to-use interface [1-2, 4-8]. More than 900 customers across 35 industries trust it to conduct complex business operations including regulatory compliance, data-as-a-service, systems modernization, and others. The Denodo Platform provides a robust data abstraction layer: It abstracts access to myriad disparate data sources; acts as a single repository (virtual); and makes data available to consumers in real time. The Denodo Platform publishes the data to applications; combines related data into views; and connects to disparate data sources. Denodo makes it easy for customers to use the data virtualization platform by providing a packaged solution that addresses the following essential elements:

- A common user interface (UI) for end-user interaction.
- Builds a metadata catalog and registers all data sources with it. It integrates the UI with an existing metadata catalog.

- It features custom-built interfaces like APIs that accept end user requests and routes them to a query optimization engine.
- A full-fledged query optimization engine that accepts queries from the UI and translates them into data-source-specific queries, executes the queries, and collects the output.
- A caching layer that caches the data and returns it back to the calling interface.
- A security layer that talks to the security repository and checks for granular level permissions.

Denodo platform strengths

"Denodo's key strength lies in its unified data fabric platform that integrates all of the key data management components needed to support real-time and dynamic use cases...."

Denodo's AI and machine learning capabilities are expanding rapidly to focus on delivering a higher degree of automation at every layer of the big data stack."

– Analyst Noel Yuhanna, Forrester Research

In addition, the Denodo Platform is available on Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) for use with the BYOL license model, simplifying the set up steps required in the cloud environment and obviating the need for a cloud native deployment, which is complex and needs cloud resources to configure and launch the solution.

Summary:

Data virtualization is a key technology enabler of modern data architectures that helps agencies to quickly move forward on their digital transformation journeys. Data virtualization is easy to configure and it provides both flexibility and agility, reducing the time to deliver data to business by up to 10x. Delivery is significantly faster (For the FAA, 96% faster; for the U.S. Army, 97% faster). Data virtualization hides the complexity of a constantly changing data infrastructure from users, enabling the seamless introduction of new technologies, formats, and protocols without ever causing disruption to users.

About CTEC

Cybermedia Technologies Inc. (CTEC) is a CMMI L3 (DEV & SVC) company taking our customers through their IT migration and modernization journey with expertise ranging from legacy applications such as mainframe, client/server environment, to cutting edge technologies that are Cloud based. iCyCIOps™ is CTEC's innovation center of excellence in Cyber and Cloud Operations providing solutions that fit the Client's multi-missions: we focus on innovations in ITIL Service Desk, DevSecOps, Microservices, Middleware, Integration and Enterprise Architecture expertise. The lab provides sandbox environments to build rapid prototypes using AI/ML, RPA, Analytics technologies, and state-of-the-art tools. The lab's mission is solution'ize artificial-intelligence-powered IT operations, cloud migration factories, and analytics-based procurement tools and insights [9, 10]

For more information,
visit **www.ctec-corp.com**

About Denodo

Denodo is the leader in data virtualization providing agile, high performance data integration, data abstraction, and real-time data services across the broadest range of enterprise, cloud, big data, and unstructured data sources at half the cost of traditional approaches. Denodo's customers across every major industry have gained significant business agility and ROI by enabling faster and easier access to unified business information for agile BI, big data analytics, Web, cloud integration, single-view applications, and enterprise data services. Denodo is well-funded, profitable, and privately held.

For more information,
visit **www.denodo.com**
Email: info@denodo.com
or call +1 877 556 2531 / +44 (0) 20 7869 8053. [1-2 4-8]

References:
[1] https://www.denodo.com
[2] https://www.denodo.com/en/document/analyst-report/total-economic-impact-data-virtualization
[3] Cloud Native Data Virtualization on AWS, September 2020
[4] Data Virtualization for Big Data, Big Data | Denodo
[5] 2021 Gartner Magic Quadrant for Data Integration Tools - Denodo recognized as 'Leader'
[6] 2021 Gartner Peer Insights 'Voice of the Customer': Data Integration Tools - Denodo One of only two vendors to receive Customer's Choice for 2021
[7] Forrester Wave: Enterprise Data Fabric, Q2 2020 - Denodo recognized as 'Leader'
[8] API ecosystem and Data Virtualization – Complementary Technologies, Denodo White Paper, April 2021
[9] https://ctec-corp.com
[10] https://ctec-icyclops.com