CGI GeoData360

CGI GeoData360

Industrialisation of geospatial and big data processing for CGI projects

Did you know that it's estimated that 87% of data-science projects fail to reach production and a major reason is the ability to scale proofs-of-concept to big data volumes. GeoData360 was built to solve this challenge.

CGI GeoData360[1] is a cloud-based scalable production platform designed to run complex processing workflows on large volumes of geospatial and Earth Observation (EO) data. It streamlines and manages various aspects of large data processing and leverages key benefits of the cloud to provide flexible and efficient on-demand provisioning of computing resources.

To support such large-scale processing CGI GeoData360 overcomes many of the common challenges related to provision of production-ready, data driven solutions making it an ideal platform to provide the processing backbone in the operationalisation of big data processing workflows.

While giving an overview of CGI GeoData360 and its capabilities, this paper will describe its primary goals, core design concepts and the technologies it utilises in doing so. To date, CGI GeoData360 has been adopted in several CGI solutions, proving its applicability to client requirements through scalable, complex processing workflows for geospatial and non-geospatial solutions alike.

1 https://www.cgi.com/uk/en-gb/geodata360

Contents

Why is CGI GeoData360 needed?

With the significant expansion of EO data in recent times, CGI GeoData360 was developed to meet the need for a greater scale of EO data processing and information extraction capability. The advent of on-demand, scalable cloud-computing enables a more efficient approach to user-driven data processing.

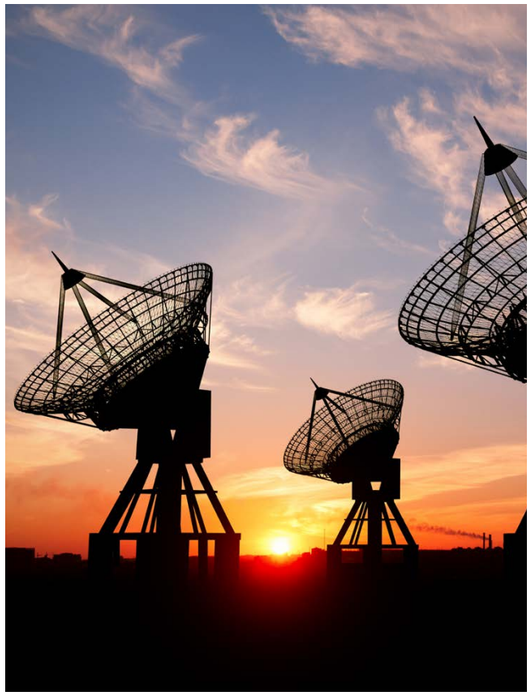
Over the last few decades, with the advancement of programmes such as the European Commission's Copernicus programme and its Sentinel satellites, the vast increase in available EO data has led to big data-driven opportunities spanning multiple industry sectors. More and more organisations are looking to capitalise on these opportunities with application examples including: intelligent crop production and agricultural land management; providing resilience for utilities infrastructures; or facilitating wide-spread monitoring for disaster management, such as oil spill response. While opportunities presented by this data are significant and diverse, consistently turning this data into global information layers for value extraction has been an on-going challenge in the EO community. Analysts' traditional workflows are confronted with a range of challenges to continuously generate higher-level products from large data volumes.
Such challenges include:

- **Scalability:** Due to significant variation in data volume and complexity of processing pipelines, there is a need for dynamic resource allocation to provide efficiency in computation time and memory occupation.
- **Reliability & traceability:** Workflows and services must provide expected outputs every time, independent of data volume or processing demand. This means having effective and efficient access to large data sources and a suitable level of monitored automation.
- **Availability:** Since the advent of the internet age there has been an increasing need for more streamlined ways of sharing data and information across organisations, communities and countries. This can range from delivering products and services around the world to collaboration through sharing of technical expertise and solutions.
A cloud-based solution such as CGI GeoData360 enables users from anywhere in the world to access and run services on the platform and obtain results.
- **Repeatability & automation:** Execution of a given process must provide consistency with every run, allowing reliable comparison of results. In addition, with the varying level of data to be processed automation where possible will streamline the processing activities and support scalability.
- **Monitoring:** To maintain operational reliability continuous monitoring solutions are vital. However, monitoring solutions that run efficiently at scale require substantial ICT resources that can be complex and costly. In many cloud-based systems the ability to reliably monitor environments is often limited by not having direct access to underlying hardware or associated log files.

In recent years, there has been a step-change in the capabilities of cloud-based systems for the collection and processing of large amounts of data, providing the means to tackle these common technical and operational problems. Large quantities of processing, memory, network and storage capability are now available on-demand in the cloud for efficient access to required ICT resources. In addition, significant quantities of Open Source Software (OSS) have been developed to support data-agnostic activities with methods of containerisation and orchestration that allow for a flexible ecosystem of solutions to be integrated together to form something new. This also includes OSS components specifically developed for EO processing and value extraction.

This wealth of available raw data and technological tools has laid the groundwork to build a framework capable of extracting value from large volumes of EO data efficiently and cost effectively, leading to the creation of CGI GeoData360. Such concepts have been explored and developed by CGI in earlier activities, most significantly the various Thematic Exploitation Platform (TEP) projects commissioned by the European Space Agency (ESA), which have also helped to inform the development of CGI GeoData360.

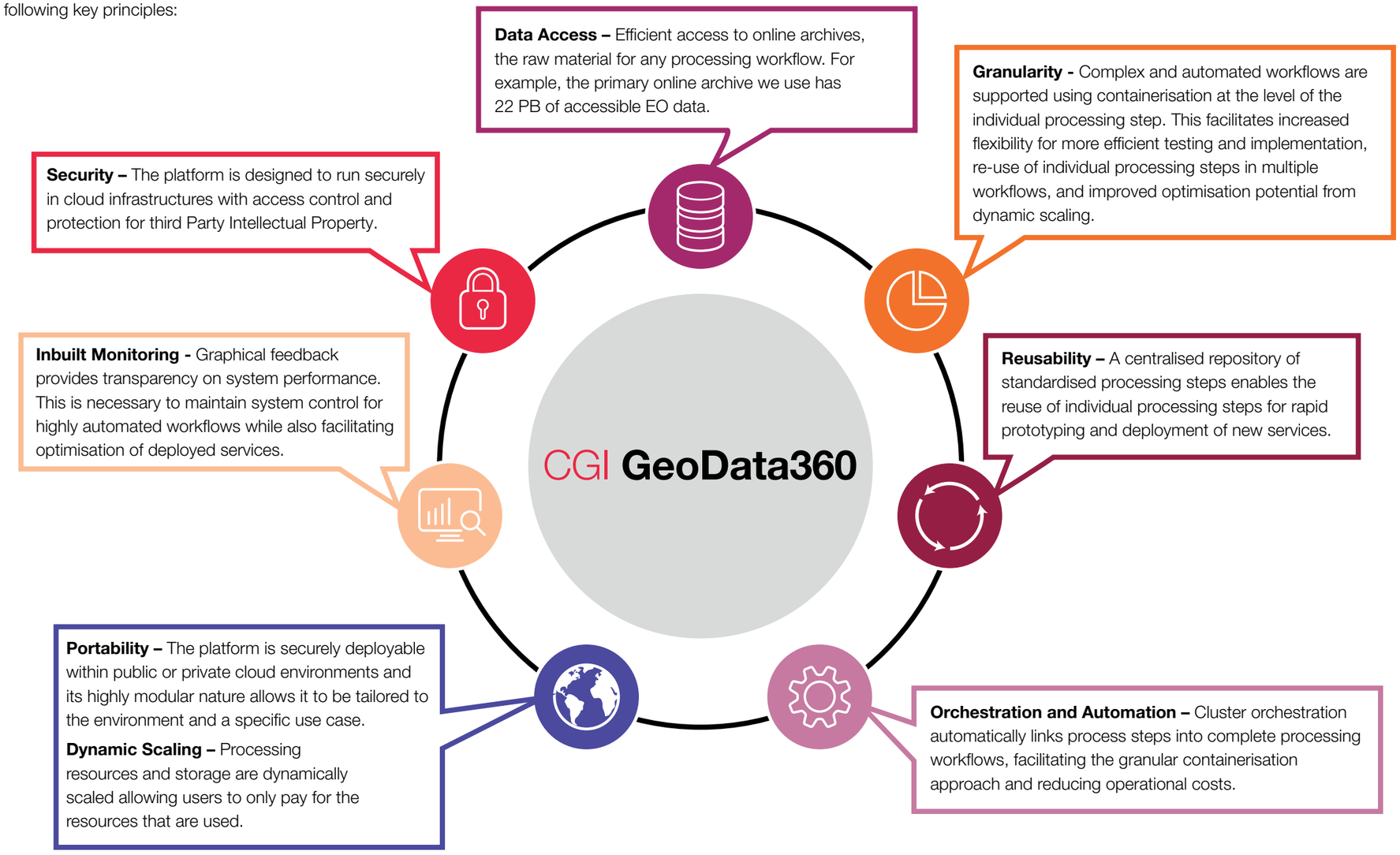
Although CGI GeoData360 has been designed to support geospatial data processing, it is also notable that many of challenges associated with large-scale EO data processing are applicable to any other domain dealing with big data processing needs. This has influenced the CGI Geodata360 design, which follows a modular approach to support deep customisation and extension, independent of domain or data type to meet user use cases.



What is CGI GeoData360?

To contend with the needs of data driven processing in production, the CGI GeoData360 framework was developed to offer an operationally focussed approach to big data processing. It is based on sophisticated orchestration of OSS components, supplying the reliable and consistent performance needed in the processing pipelines of commercial services. Within the environment, algorithms and workflows for extracting information from EO and geospatial data can be deployed within containers. This allows for increased flexibility of processing workflows and dynamic scaling to optimise the use of infrastructure resources from commercial cloud providers, whilst also reducing processing times. Furthermore, while initially developed to solve issues inherent in 'Big Earth Data' for geospatial applications, this containerised framework provides the flexibility to easily deploy any user developed processing algorithms within workflows and be run at industrial scale. This deployment of users developed services can either be done individually by an experienced user or with consultation from CGI.

To deliver a robust, scalable platform for industrial scale processing, CGI GeoData360 is built on the following key principles:



Key features

CGI GeoData360 hides the complexity of the data processing behind an easy to use application that lets users choose the data and services they want to use, specify if one-off or regular systematic processing, and the output format required. It offers the following key features:

Data ingestion: Flexibility of data ingestion – CGI GeoData360 provides in-platform access to online data archives for direct ingestion into configured workflows, bypassing the need for users to download and process large datasets on the own infrastructure. The default data source is via CREODIAS, an EU sponsored data repository for accessing a range of EO data, most significantly the EU's Copernicus Sentinel programme. CREODIAS currently hosts over 22PB of EO data, with data focusing on a variety of thematic areas hence meeting a range of user needs. CGI GeoData360 also has the flexibility to ingest data from various other data sources as needed by the client. This is not limited to EO or geospatial data. The ingestion and comparison of multiple data sources facilitates enhanced output interpretation.

Workflow configuration – Users are able to select desired processing services and configure them using parameters published by the developer of the service. These could be existing services or their own bespoke algorithms deployed within the platform. The CGI GeoData360 team offers Onboarding-as-a-Service (OaaS), to support clients in containerising, orchestrating and optimising their workflows for deployment in their CGI GeoData360 instance.

Scalable Batch Processing – For enhanced processing efficiency, CGI GeoData360 services can be batch processed to apply parallel processing functionality to linear workflows with many inputs. All processing is run on a shared Kubernetes cluster to auto-scale resources according to demand. Multiple parallel-containerised processing activities are created, and as more CPU and memory resources are requested, additional Virtual Machines are added or removed from the cluster as demand changes. Processing steps and workflows are versioned and stored in the repository for re-use, giving users the ability to reliably repeat previous workflows. The overall process is lean in terms of its use of resources.

Systematic workflows – Within the environment users have the ability to set up and configure systematic workflows, allowing data to be processed as it becomes available. This specifically useful in supporting the creation of long-term services, for instance providing automatic and frequent updates on a desired physical index over a specified region to support continuous monitoring. Further to this, being able to achieve this high degree of automation drives down operational costs which is passed on the customer.

Product delivery: CGI GeoData360 uses industry standards, particularly the Open Geospatial Consortium (OGC) standard interfaces to access and publish data and generated products. This minimises effort needed to configure ingestion of CGI GeoData360 products by a user's standard tools and systems, yet delivery channels remain highly configurable to meet client needs. This capability is currently being enhanced with the ability to publish generated products and maps to INSPIRE compliant metadata catalogues, with full text and spatial query capabilities for ease in discovering and collaborating on published products.

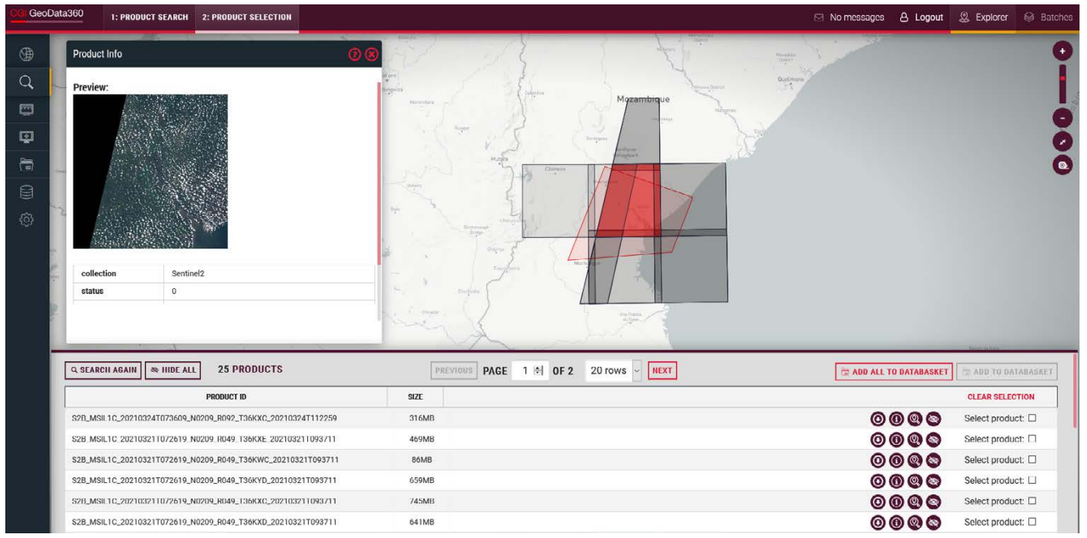


Figure 1: Searching for EO data in CGI GeoData360

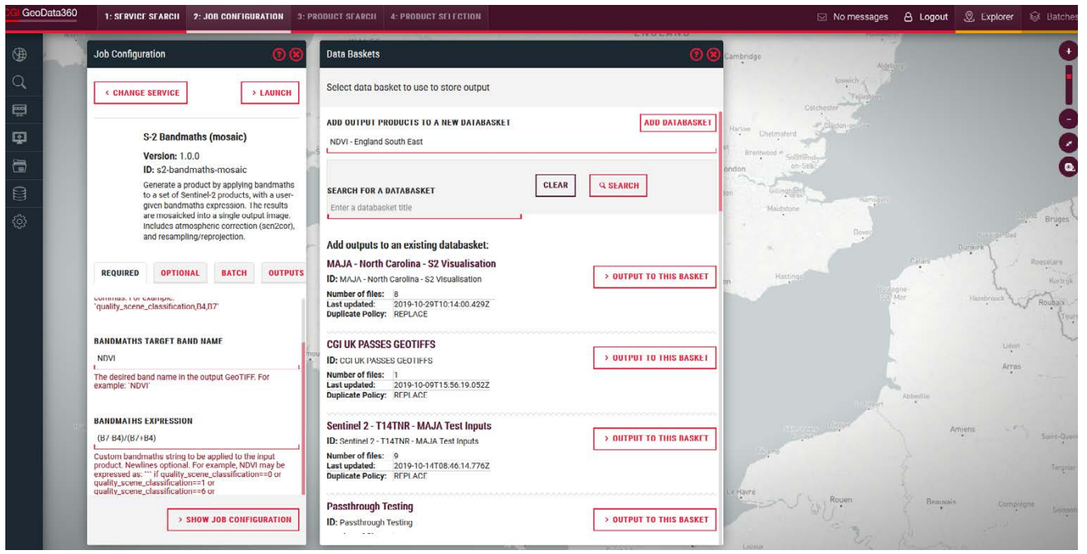
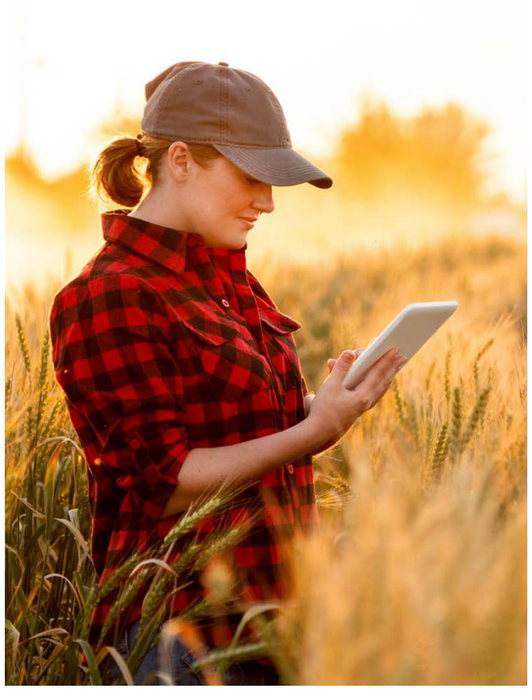


Figure 2: Configuring a service job and selecting data to be ingested from data baskets

Visualisation – EO and geospatial data visualisation capability is provided through the ‘GeoPortal’. This implements OGC Web Map Service (WMS) standards to display data. Layers can be stacked and transparency adjusted for direct comparison and enhanced interpretation of generated products. Due to its modular nature, the visualisation capabilities are not an intrinsic part of the platform and can therefore be interfaced with other OGC compliant visualisation systems if required. Visualisation capabilities are currently being further improved allowing users to create and style multi-layer interactive maps from any data published in the catalogue.

Monitoring and logging capabilities – CGI GeoData360 offers extensive monitoring and logging capabilities. With its high level of automation and scalability, sophisticated monitoring helps provide system control. Admin users can access graphical feedback on metrics such as CPU and RAM load as well as centralised logging capabilities to monitor system performance at each step of their workflows. Traceability also allows for monitoring of requests between components.

Security – CGI GeoData360 has been designed to ensure protection of user’s services, code, data and personal information. This is achieved through access control, allowing platform administrators to manage a user’s access to various platform aspects. To adhere with recent legalisation on data privacy, such as GDPR, GeoData360 uses common industry systems such as KeyCloak to securely manage any personal information (of which the minimum required to manage users is collected).



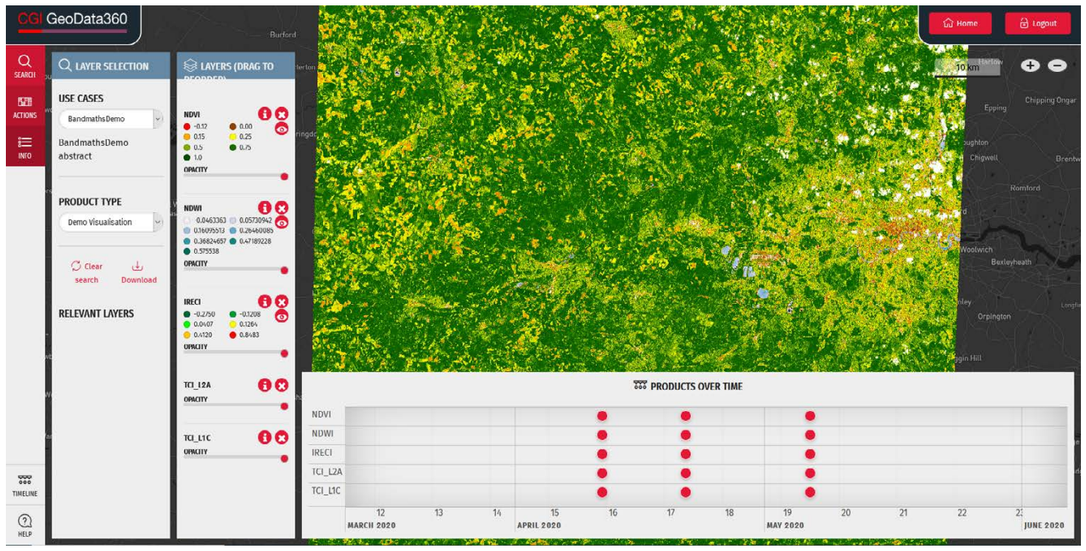


Figure 3: Visualising service outputs across a timeseries in the GeoPortal

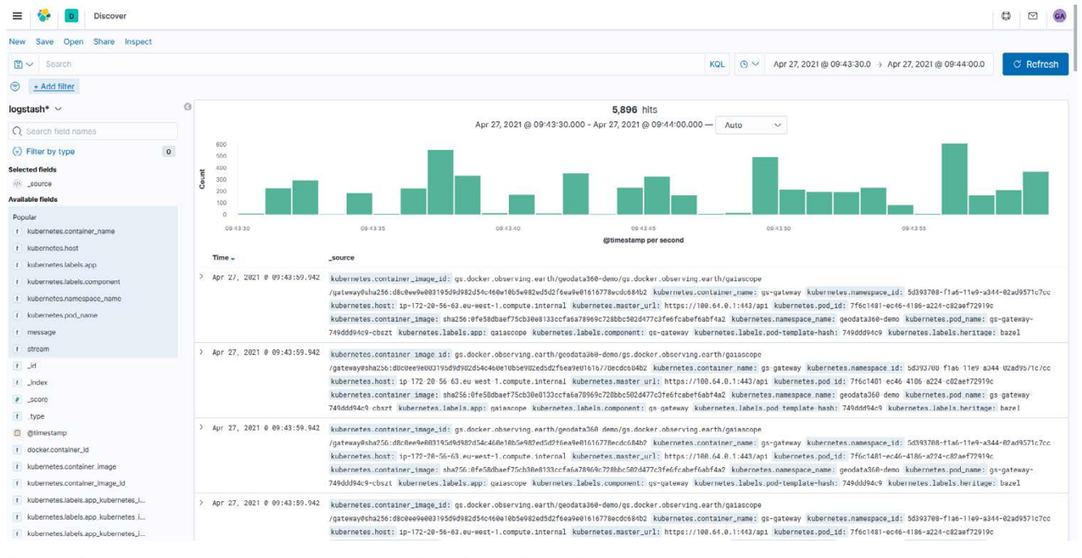


Figure 4: Checking container logs in-platform with the ElasticSearch/Fluentd/Kibana stack

CGI GeoData360 technology stack

CGI GeoData360 functionality is provided by a combination of configured 3rd party OSS applications and bespoke CGI code. By implementing containerised microservices of trusted industry leaders this ensures each component of the system has a high development standard that is translated to the client at a low cost.

To meet its primary use case as an EO and geospatial data processor, CGI GeoData360 includes a number of standard geospatial data management solutions. This includes GeoServer - an open source server for managing, sharing and editing geospatial data along with other elements that support visualisation such as OpenLayers (that provides the mapping interface). Our 'GeoPortal' visualisation component connects to a deployed GeoServer instance to render products in a map view via WFS and WMS APIs. A package of commonly used EO processing applications are also provided, including GeoTools, SNAP as well as commonly used programming languages such as R and Python. Further to this, other OS components provide standard platform functionality, such as Keycloak to provide user management, authentication and authorization services (along with OAuth2 single sign on service) Prometheus for metrics collection across a distributed system and an ElasticSearch/ Fluentd/Kibana stack for logging functionality.

Where CGI GeoData360 adds greater value is through its robust container orchestration of these services using the infrastructure abstraction layer, Kubernetes, to manage services across the virtual cloud to deploy a clustered, distributed application, which can scale out to different servers in the cloud (e.g. on AWS or CREODIAS) or scale down to a single machine.

This granular containerisation and orchestration approach is not only taken for the deployment of modules that make up the CGI GeoData360 functionality but for the bespoke processing algorithms deployed as services on the system. Containers themselves are created using the industry leading technology Docker. This allows for code and dependencies to be packaged up such that an application runs quickly and reliably from one computing environment to another and allows distinct processing activities to be instigated and executed independent of other activities. By wrapping user developed algorithms in isolated containers they can be deployed with ease in the language they were developed in. Further to this, with the granular approach to containerisation, processing steps of a workflow are divided between containers. These steps can therefore be re-used in multiple services if desired, and be configured in a range of workflow processing structures to unlock greater flexibility for building new EO and geospatial product workflows while facilitating efficient and rapid prototyping capabilities.

GeoData360's technology stack is shared with another CGI geospatial product: GeoApp[2], a dedicated platform for deploying geospatial portals and applications. This includes the OSS component GeoNode, a geospatial content management system specifically designed to make geospatial data findable, accessible, interoperable and re-usable.

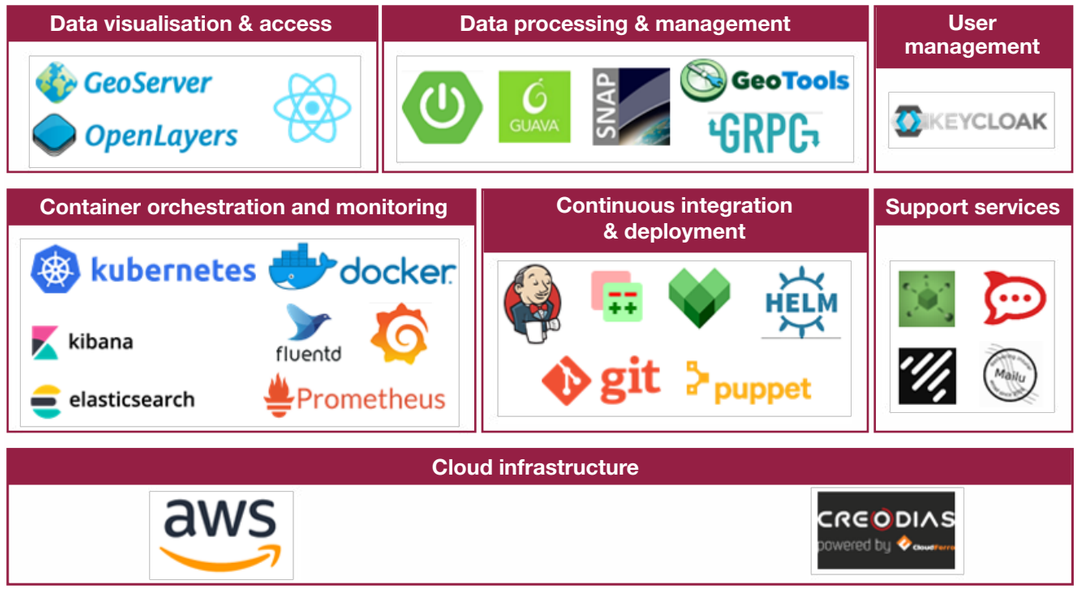


Figure 5: The CGI GeoData360 technology stack

2 https://www.cgi.com/uk/en-gb/solution/CGI-GeoApp

CGI GeoData360 use cases

CGI GeoData360 has the flexibility to operationalise a variety of applications. It is ideal for those who want to run applications cost effectively and at scale to grow their business, or for those who lack the infrastructure to meet customer service level expectations.

Since CGI GeoData360 can be used in private and public clouds it allows users to have a common approach to data stored in both locations whilst maintaining an air-gap. In the EO domain, typical use cases could include facilitating continuous crop monitoring for enhancement of agricultural practices, or extracting flood maps from SAR imagery as an information source to insurers and public health actors. As CGI GeoData360 is designed for deployment of bespoke algorithms and ingestion from varied data sources, there is no limit to the number of potential use cases. CGI has already proven CGI GeoData360's applicability as a stand-alone platform as well as using key platform features as enabling technology in multiple EO and non-EO initiatives. Some of these are provided below:

HiVaCroM

The **High Value Crop Monitoring project**, utilised CGI GeoData360 in the deployment of a demonstration version of an integrated crop yield service aimed at producers and commodity traders. The HiVaCroM concept was for an automated system to identify, collect and process satellite data to create a parameter, namely canopy cover value, for fields of potatoes. A time-series of such data would be collected over the growing season and would be inputted into a potato yield model both forecast crop yield and recommend harvest time. CGI GeoData360 enabled data intensive production workflows to be automated and run at national and regional scales. From this, HiVaCroM was able to create reliable and frequent estimates of canopy cover through the configuration of systematic workflows and hence produce time series of observations describing crop development.

CARNOT-SAT[3]

The **Computer-Aided Radio Network Optimisation Tool**, gives mobile and satellite operators powerful agives mobile and satellite operators powerful and accurate tools for planning, designing and optimizing radio networks using both terrestrial and non-terrestrial technologies. The tool will help open up the market for satellites in the backhaul network by providing backhaul network planning functionality to highlight situations where their deployment will be effective in terms of cost, performance and time to deploy. CGI GeoData360 has laid the foundations for the visualisation of Line-of-Sight Propagation and visibility coverage analysis for LEO, MEO and GEO Satellites allowing operators to make more accurate planning decisions. The CGI GeoData360 platform provides the software architecture basis of the Carnot-Sat toolset ensuring the running of the application is reliable, secure, highly available and stable across multiple infrastructures.

AUTSS[4]

The **Autonomous Satcom Solutions** programme is developing an advanced, sat-com focussed, secure AI platform. AUTSS will explore multiple use cases, from optimising and de-noising satcom transmission to fusing maritime AIS data with NOAA weather data to predict maritime satcom demand. CGI GeoData360 components are being used to manage the data

3 https://www.cgi.com/uk/en-gb/news/space/european-space-agency-awards-contract-for-terrestrial-satellite-5g-tool-kit-to-cgi
4 https://www.cgi.com/uk/en-gb/news/space/cgi-develop-innovative-artificial-intelligence-enabled-platform-global-satellite-communications-marketplace

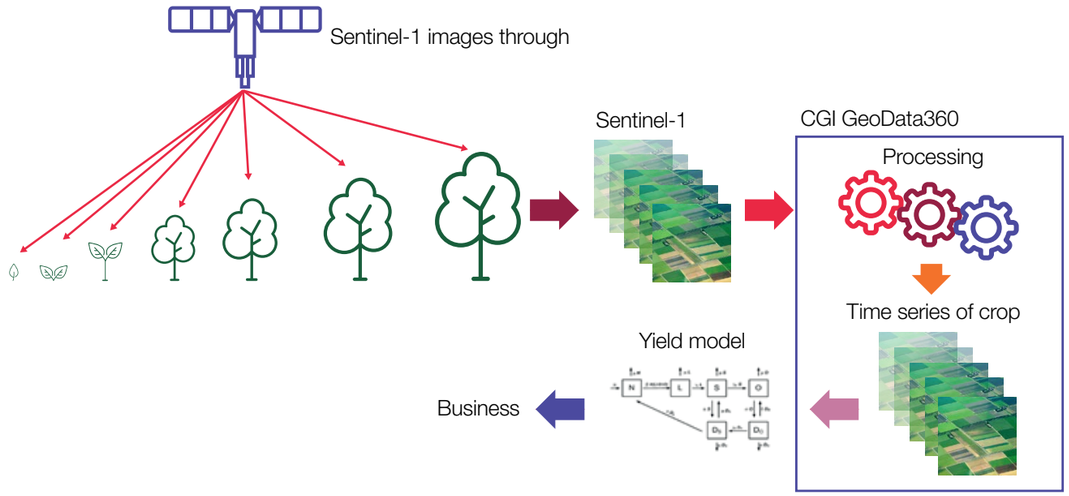


Figure 6: CGI GeoData360 enabled data intensive production workflows in the HiVaCroM project.

pre—processing, enabling automated data extraction, ingestion and transformation of the uploaded raw data to normalise and sanitise before being fed into the AI pipeline. CGI GeoData360's flexible ingestion capabilities facilitate the use of multi-variate data to maintain the AI platform applicable to many use cases.

Further to this, the scalability of CGI GeoData360 accommodates pre-processing of large volumes of training data and therefore permits high utilisation of training and experimentation within the rest of the machine learning pipeline.

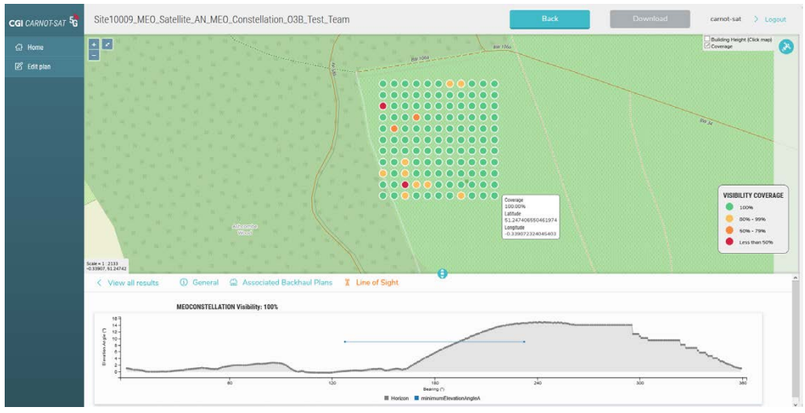


Figure 7: CGI GeoData360 provides infrastructure for the Carnot-Sat project to provide Line-of-Sight propagation and visibility coverage analysis.

How can CGI GeoData360 be provided?

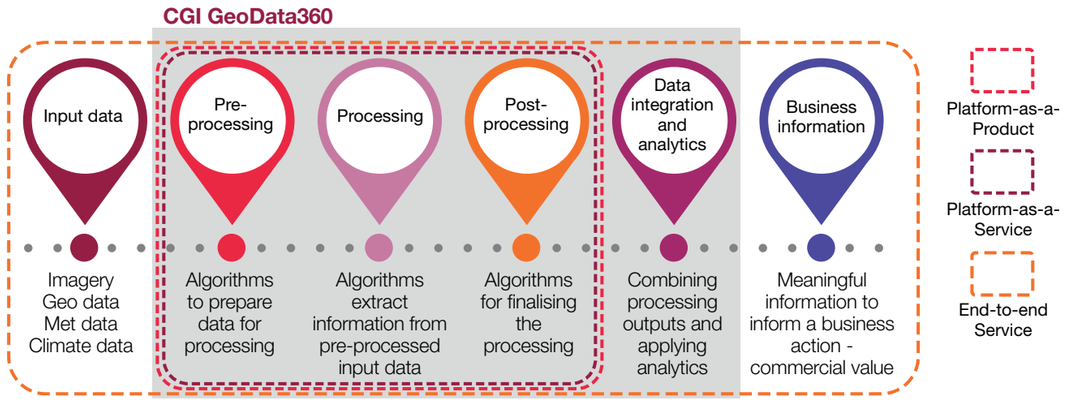
GeoData360 can be provided in a range of delivery models to cover the specific needs of an interested client. Important factors in determining the most suitable delivery model include available ICT, security needs, range of users planning to access the platform and the key use case (either provision of a platform for various data processing activities or a dedicated service for large volume data processing). To meet these needs GeoData360 is provided through the following models:

A. Platform as a Product (PaaP) – in which a dedicated instance of CGI GeoData360 is deployed on a client's infrastructure. The client will become the primary platform administrator / operator with CGI providing platform level support & maintenance. This allows clients to manage the users and their access to particular data and services. This can also of benefit to users who for security reasons want to deployed the platform within their own infrastructure.

B. Platform as a Service (PaaS) – where the client has access to a dedicated instance of GeoData360 that will be deployed and managed on a cloud environment by CGI. This is useful where the client may not have the capability or interest in managing the whole technology stack and as PaaS can be offered on a monthly basis it gives more flexibility.

C. End-to-End services – where the client orders access to a specific service on the platform, either their own service that they want to run at larger scale or one of the existing CGI developed services. This is most useful when a client is interested primarily in the generation of particular data products.

Note that, unlike traditional-off-the-shelf products, GeoData360 is quite complex and any new users will need to undergo initial consultancy to understand their primary use case and to determine the most suitable delivery model. After deployment, on-boarding would be needed to make the client familiar with the platform and key features as well as supporting deployment of any user developed algorithms onto the platform with OaaS. Such support will be factored into the overall licensing approach.



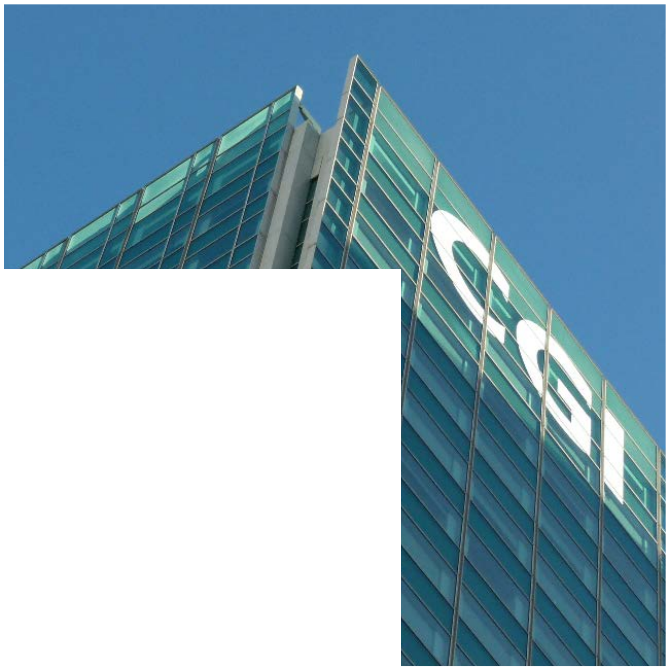
Summary

In recent years, there has been a significant increase of commercial interest in EO and geospatial technologies, with higher quality, more frequent and greater coverage of EO data now accessible, along with greater analytical capability in the form of AI and modelling, providing intuitive and unique methods to extract valuable business information.

To support this growing market, GeoData360 provides a cloud-based processing platform to support user-defined processing activities on geospatial data, particularly at a larger scale.

GeoData360 provides users with a range of opportunities that can be tailored to meet their specific needs. This can range from a platform offering the EO data and processing tools needed for development of concepts and algorithms into fully-fledged services, to a dedicated processing environment for large scale processing. Additionally, GeoData360 is not limited to EO or geospatial processing, but can support other data types for use in multiple industry sectors.

Looking to the future, GeoData360's aim is to expand its portfolio of use cases both in the EO and non-EO domains, within CGI and with users who are looking to deploy their bespoke processing algorithms at industrial scale to facilitate long running, reliable services.



About CGI

Founded in 1976, CGI is among the largest IT and business consulting services firms in the world.

We are insights-driven and outcomes-based to help accelerate returns on your investments. Across 21 industry sectors in 400 locations worldwide, our 76,000 professionals provide comprehensive, scalable and sustainable IT and business consulting services that are informed globally and delivered locally. We are an IT Systems Integrator working to advise, build and operate bespoke, technically complex, mission-critical information systems. Bringing innovation to our clients using proven and emerging technologies, agile delivery processes and our expertise across space, defence, intelligence, aerospace and maritime, all underpinned by our end-to-end cyber capability.

For more information about CGI,
visit **cgi.com/uk/space**,
or email us at **enquiry.UK@cgi.com**

CGI