**CAMEO**

**Architecture for Manipulating Earth Observation Data**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**D2.5**

**WORK PACKAGE: WP2 Vertice Cloud**

**DELIVERABLE: 2.5 Design Intelligent Data Analytics Services**

**SUBMISSION DATE: 06 04 2023**

**NATURE: WORK PACKAGE OUTPUT REPORT (INTERMEDIATE)**

| **Document Control** | |
| --- | --- |
| **Creator** | **A Giles-Haigh, P Godfrey** |
| **Description** | **The intended design of the Earth Observation Intelligent Data Analytics Services for CAMEO project, as part of deliverable 2.5.** |
| **Publisher** | **Vertice Cloud** |
| **Audience** | **CAMEO Project** |
| **Status** | **Draft** |

| **Version** | **Issue date** | Stage | Changes | Contributor |
| --- | --- | --- | --- | --- |
| 0.1 | 17/02/2023 | Draft | No previous document | P Godfrey |
| 0.2 | 01/03/2023 | Draft | Intelligent Data Analytics Services definition | P Godfrey |
| 0.3 | 08/03/2023 | Draft | Added Data analytics techniques | P Godfrey / A Giles Haigh |
| 0.4 | 16/03/2023 | Draft | Added Design Plan for Data Analytics Services | P Godfrey / A Giles Haigh |
| 0.5 | 30/03/2023 | Draft | Internal Review | A Giles Haigh |
| 1.0 | 06/04/2023 | Draft | Final draft for review with wider CAMEO group | P Godfrey / A Giles Haigh |

**PROJECT IDENTIFICATION**

**Subject: Report on design of the Earth Observation Intelligent Data Analytics Services**

**Domain:**

**REF.:**

**Activity No.**

**ITEM NO.:**

**CONTRACT NO:**

**PRIME CONTRACTOR: Vertice Cloud**

**EXTERNAL SUPPLIER:**

**COPYRIGHT: Vertice Cloud**

**Contents**

[1](#_heading=h.gjdgxs) Introduction 4

[1.1](#_heading=h.30j0zll) Purpose of this document 4

[1.2](#_heading=h.1fob9te) Definitions, acronyms, and abbreviations 4

[1.3](#_heading=h.3znysh7) Reference Documents 4

[2](#_heading=h.2et92p0) Intelligent Data Analytics Services 5

[2.1](#_heading=h.tyjcwt) Introduction 5

[2.2](#_heading=h.3dy6vkm) What are the benefits of implementing an IDAS for CAMEO? 5

[2.3](#_heading=h.1t3h5sf) What steps are involved in implementing an IDAS for CAMEO? 6

[3. Data Analytics for CAMEO 7](#_heading=h.4d34og8)

[3.1](#_heading=h.2s8eyo1) Types of Data Analytics 7

[3.2](#_heading=h.17dp8vu) Data Analytics in Earth Observation 7

[3.3](#_heading=h.3rdcrjn) Data Analytics Techniques 9

[3.4](#_heading=h.26in1rg) Data Analytics Toolkits/Packages 9

[3.5](#_heading=h.lnxbz9) Identifying appropriate Data Analytics Services 10

[3.5.1](#_heading=h.35nkun2) Data Analytics Services backlog 11

[4. Conclusion 13](#_heading=h.1ksv4uv)

# 1 Introduction

## 1.1 Purpose of this document

This document describes the intended design of the Earth Observation Intelligent Data Analytics Services for CAMEO project, as part of deliverable 2.5.

It aims to provide:

* A general description of Data Analytics Services
* The benefits of implementing Data Analytics Services on CAMEO
* A design plan of Data Analytics Services for CAMEO including:
  + Data ingestion from the Data Warehouse.
  + Pre-processing steps required.
  + Analytics Services plan and timeline.

## 1.2 Definitions, acronyms, and abbreviations

| **Acronym** | **Acronym meaning** |
| --- | --- |
| ML | Machine Learning |
| MAMS | Multi-Agent Microservice |
| IDAS | Intelligent Data Analytics Service |
| VPN | Virtual Private Network |
| API | Application Programming Interface |

## 1.3 Reference Documents

None

# 2 Intelligent Data Analytics Services

## 2.1 Introduction

As a term, data analytics enables organizations to analyse all their data (real-time, historical, unstructured, structured, qualitative) to identify patterns and generate insights to inform and, in some cases, automate decisions, connecting intelligence and action.[[1]](#footnote-0)

Data analytic services are the wrapper, providing a suite of analytics (as a service) for a particular question. Common areas of Data Analytic Services are:

* Financial
* Customers
* Supply Chain
* Etc

All the above have specific analytical use cases for their own area. For the CAMEO project, the area is focused on earth observation data, including satellite imagery, and related analytical use cases.

In each case, the common goal is to ensure that users of the system can fully benefit from business intelligence, regardless of skill level, ensuring they can manage, transform, visualise, and implement data-driven innovation at scale.

As part of the Design element of the Intelligent Data Analytics Service, these use cases need to be identified, evaluated, and scored in terms of **importance** to the projectversus **time** to deliver.

## What are the benefits of implementing an IDAS for CAMEO?

It ensures all data analytic processes and requirements, as well as applied analytics are captured in a single pipeline, including:

* **Understanding what data is utilized** – it easier for new and existing clients to develop an understanding of an enterprise’s services.
* **Built-in data preparation –** any pre-processing steps are applied to the data/images returned from the Data Warehouse.
* **A single platform to provide self-service analytics** – each analytic process is pre-packaged and ready to run.

## What steps are involved in implementing an IDAS for CAMEO?

There are several steps required to create an intelligent data analytics service, all of which are outlined below and will be considered in the design and implementation stages of the build within deliverable 2.5.

The outline below will flow from source data, all the way through to the end-user being presented a result on screen. There will be cross-over with other work packages within this service, particularly WP1 (Data Warehouse) and WP3 (Data Quality), with input required from the SME’s and UCD.

1. **Search imagery data** from the data warehouse, using the VPN

2. Images would be returned using the relevant API

3. The images could be **stored** in the analytics engine (tbc where) such as an S3 bucket/object storage.

4. During storage, as well as accessing the analytics services, the design must consider segregation of users within the analytics engine. This is to ensure intellectual property is protected throughout the user journeys within the analytics engine.

5. Any **pre-processing** would be applied at this point (cleaning, aggregating etc). Depending on the analytics required, different pre-processing may be applied.

6. Analytical “tool” will then be selected, which could be pre-built by WP2/UCD e.g.

a. Cropping images  
b. Image classification (River, Road, Body of water, Forestry etc)  
c. Data Analytics (Visualisations/Machine Learning use-cases)

The identification of suitable data analytics will be covered in Data Analytics Options (Section 3).

7. Longer term, and in later iterations of this deliverable, the data analytic services will become microservices, available to Novice users through pre-baked pipelines, as well as Advanced users through MAMS (Multi-Agent Microservice).

# 3. Data Analytics for CAMEO

## 3.1 Types of Data Analytics

Data analytics is broken down into four basic types.

* **Descriptive analytics**: This describes what has happened over a given period. Have the number of views gone up? Has anything changed since last time the images were taken?
* **Diagnostic analytics:** This focuses more on why something happened, this involves more diverse data inputs and posing a business problem.
* **Predictive analytics:** This moves to what is likely going to happen in the near term.
* **Prescriptive analytics:** This suggests a course of action based on the history of the data.

## 3.2 Data Analytics in Earth Observation

To ensure appropriateness for CAMEO, we have researched common earth data analytical functions. These range in complexity from simple numerical functions to raster and vector operations, visualization and exploration, and machine learning.[[2]](#footnote-1)

They can be broken down into the below key areas:

3.2.1 Data Pre-processing:

As an important early phase, pre-processing consumes around 50–80% of the entire time for data analytics (Kempler & Mathews, [2017](https://www.tandfonline.com/doi/full/10.1080/20964471.2019.1611175)).[[3]](#footnote-2)

Graphical user interface, text, application

Description automatically generated

*Table 1: Examples of Earth Observation data pre-processing analytical tools*

3.2.2 Data Analytics Methods:

After pre-processing, the focus of data analytics is to reveal hidden patterns, unknown correlations, and other useful information from a large volume of heterogeneous data to facilitate Earth science study.

Graphical user interface, application

Description automatically generated

*Table 2: Examples of Earth Observation data analytical methods tools*

## 3.3 Data Analytics Techniques

There are several different analytical methods and techniques data analysts can use to process data and extract information. Some example of data analytical techniques:

* Regression analysis
* Factor analysis
* Cohort analysis
* Monte Carlo simulations
* Time series analysis
* Image analysis – deep learning
* Image analysis – classification

## 3.4 Data Analytics Toolkits/Packages

As CAMEO is a cloud-agnostic and open-source solution, there are several toolkits/packages which can be leveraged to help understand Earth Observation data.

The below are some examples of data analytical toolkits[[4]](#footnote-3) which could be included as part of the Data Analytics / Machine Learning pipelines:

| Toolkit | Description |
| --- | --- |
| EO-Learn | EO-Learn is a Python package that links closely with the data science and machine learning python ecosystem to the remote sensing/earth observation community. With eo-learn, even non-experts can use to extract and derive valuable information from satellite images. |
| EarthPy | Python package that makes it easier to plot and work with spatial raster and vector data using open source tools. EarthPy bridges the gap between raster and vector data so you can work effectively between the two different data types. |
| Rasterio | Essential, lightweight and flexible Python package for remote sensing image reading and writing. |
| GDAL | A translator library for raster and vector geospatial data formats and provides an extensive list of satellite image processing tools. |

*Table 3: Examples of earth observation Python packages for consideration*

## 3.5 Identifying appropriate Data Analytics Services

Research has suggested some common earth observation data analytics requirements are below. For the appropriateness of CAMEO, we also wanted to consider data analytics opportunities that can apply to all user journeys (Novice and Advanced) to make use of.

The three data analytics services we will look to provision in the first iteration is:

1. **Flipping / Rotating / Cropping** (Section: Augmentation)
2. **Image classification** (Section: Reduction)
3. **Anomaly detection** (Section: Machine Learning)

WP2 will begin to build out these data analytics services in silos, with the help of UCD and other relevant work packages. Longer term, we will look to integrate these into a packaged suite of microservices as the project progresses.

To make these data analytics available to the CAMEO platform, and in later iterations of this deliverable, the data analytic services will become microservices.

These will be available to Novice users through pre-baked pipelines, as well as Advanced users through MAMS (Multi-Agent Microservices).

## 3.5.1 Data Analytics Services backlog

For the initial phase, we will look to create the 3 data analytics services identified in 3.5, however there will be a requirement to build out additional data analytics services.

To assist with this, the creation of a Data Analytics Services backlog will be proposed.

To identify suitable data analytics options required, WP2 will establish these requirements through stakeholder engagement with WP7/UCD and other stakeholders, over various sessions to ensure analytical services are appropriate for CAMEO.

The aim of the initial sessions will be to gather requirements for additional data analytics services. Once all data analytic options have been identified, we will look to score each one. This exercise will help consider effort required to create the data analytics versus importance to the project.

This will subsequently help to provide an order for WP2, and relevant teams, to plan each data analytics option from the backlog.

The initial view of how this will be captured is below:

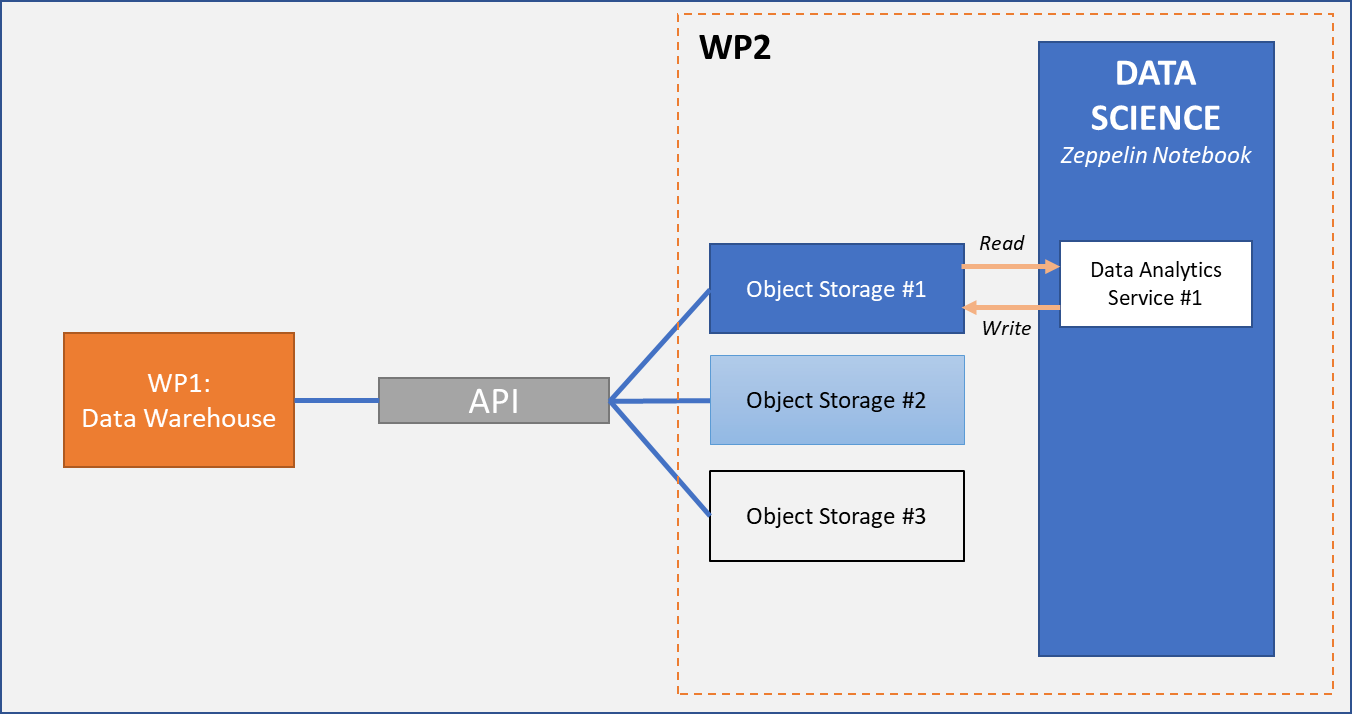
Chart, bubble chart

Description automatically generated

*Fig 1: Example chart for scoring Data Analytics Services (Importance vs Effort) within the Data Analytics Services backlog*

3.6 Design Plan for Data Analytics Service

As described in 2.3, below is a visual representation of how the data analytics services could be created/accessed.



*Fig 2: Design of Data Analytics Services accessed by CAMEO end-users.*

The outline below will flow from source data, all the way through to the end-user being presented a result on screen. There will be cross-over with other work packages within this service, particularly WP1 (Data Warehouse) and WP3 (Data Quality), with input required from the SME’s and UCD.

1. **Search imagery data** from WP1: Data Warehouse, using the VPN/API.

2. Images would be queried from Data Warehouse and returned using the relevant API.

3. The images would be **stored** in the analytics engine (object storage/MySQL database). Each SME would have access to their own object storage/MySQL database e.g., Object Storage #1, #2 and #3 will be separate SME’s. This is to ensure intellectual property is protected throughout the user journeys within the analytics engine.

4. Any **pre-processing** would be applied at this point (cleaning, aggregating etc). Depending on the analytics required, different pre-processing may be applied.

5. Image would be read in to Zeppelin Notebook, this will allow the end-user to apply Data Analytics Service. Each Data Analytics Service could be its own Notebook.

6. After Data Analytics service has been applied, additional data science techniques could then be applied, such as Machine Learning etc.

7. Once satisfied, the user could then save the image back to the Object Store / MySQL database.

As described in 2.3, longer term, and in later iterations of this deliverable, the data analytic services will become microservices, available to Novice users through pre-baked pipelines, as well as Advanced users through MAMS (Multi-Agent Microservice).

# 4. Conclusion

Implementing Intelligent Data Analytics Services within CAMEO will help ensure all data analytic processes and requirements, as well as applied analytics are captured in a single pipeline.

In the first iteration of Data Analytics Services, there are three services that WP2 will be provisioning for users to access.

1. **Flipping / Rotating / Cropping** (Section: Augmentation)
2. **Image classification** (Section: Reduction)
3. **Anomaly detection** (Section: Machine Learning)

Initially these services will be accessed via Python scripts, within a Zeppelin Notebook. Longer-term, these will be packaged up as microservice pipelines as work continues with WP2 & UCD and other relevant work packages.

A Data Analytics Services backlog will look to be created, to help identify additional Data Analytics Services to provision. WP2 will establish this backlog through stakeholder engagement with WP7/UCD and other stakeholders, over various sessions to ensure analytical services are appropriate for CAMEO.

1. https://www.tibco.com/reference-center/what-is-data-analytics#:~:text=Data%20analytics%20enables%20organizations%20to,decisions%2C%20connecting%20intelligence%20and%20action. [↑](#footnote-ref-0)
2. https://www.tandfonline.com/doi/full/10.1080/20964471.2019.1611175 [↑](#footnote-ref-1)
3. https://www.tandfonline.com/doi/full/10.1080/20964471.2019.1611175 [↑](#footnote-ref-2)
4. <https://towardsdatascience.com/the-best-earth-observation-data-science-toolkits-a51d867343a0> [↑](#footnote-ref-3)