A management console for parallelised AEROMOD/CALPUFF Batch Processing using AWS ECS and Docker for EMM consulting

1 Overview and Background

Sydney-based Environmental Consulting group, EMM Consulting Pty Ltd, utilise the CALPUFF and AERMOD air

pollution dispersion models to predict ground level air pollutant concentrations

across a given landscape or environment. It’s a highly compute intensive process often taking days to provide the results needed to finalise critical reports on behalf of their clients. In many cases, a simple adjustment to the model can require a repeat of the entire end-to-end process and EMM sought a solution that would allow this high yield step to be completed in a more accelerated manner eliminating the need to re-run the cycle end to end. 2pi Software successfully created a scripted solution that initially worked in a traditional big box ‘on-premise’ environment, but subsequently, in collaboration with the client, migrated the solution to the AWS cloud. Additionally, via a feature-rich user interface, the development team took advantage of AWS Batch (and AWS ECS) to enable an optimised parallel processing approach that in many cases reduced the overall compute times from multiple days to a handful of hours.

1.1 Client Testimonial

“Developing this tool with 2pi Software was a great process to be part of. From our initial concept on a page, Liam, Derek and the team at 2pi Software was able to develop a working protype tool to test on example datasets. These tests lead to a series of refinements of the processing steps. The current form of the tool, complete with a clean and easy to follow GUI, is actively being used on our current projects to great effect. Communication with 2pi was open and easy throughout and deliverables were on schedule and as described. We look forward to future collaborations with 2pi.” according to Scott Fishwick (Associate – National Technical Leader, Air Quality), EMM Consulting

2 EMM AERMOD/CALPUFF Workflow Requirements

EMM run a number of different air pollution dispersion models as a key part of

their service offering. Within the associated workflow steps, a compelling need for post-processing capability to speed up completion of the compute intensive tasks involved that would allow EMM to significantly improve the benefit of the modelling process had emerged.

The workflow typically involves setting up the models with a number of emission sources (varies by project) which are configured with a series of emission release data points. Additionally models are configured with calculation points (also varies by project) and these comprise of specific locations of interest (eg houses, sensitive developments etc) and a broader domain (typically featuring nested grids of increasing point spacing). The model works in hourly time steps for a specified period (typically a year) and predicts concentrations at each of the calculation points for each time step (e.g. for a full year, 8,760 individual values per calculation points).

As a general requirement, EMM needs to allow configuration of the data model such that actual emission rates can be plugged into each source and at the end of the modelling run, final results extracted. However, an issue that arises is that anytime a model needs to be revised (which happens regularly), the emission rates and model also need to be revised and a rerun can take hours or days per model.

To address this EMM sought a solution which allowed them to run each emission source as a unit release once and have the ability to scale those results through a post-processing tool.

As a later enhancement, and to facilitate productivity enhancements within the EMM team, management requested the creation of a Graphical User Interface(GUI) solution that would reduce the overall complexity and time to completion of the entire process end to end.

2.1 Handling both AERMOD and CALPUFF

Depending on the specific project context, EMM typically implements the two primary regulatory air pollution dispersion models, AERMOD and CALPUFF, in performing complex data analysis on behalf of their clients.

The AERMOD model, developed and maintained by the U.S. EPA, has become one of the most important and widely used air dispersion models in the world since the release in 2006. It is commonly utilised in short-range dispersion modelling – distances of 50KM.

CALPUFF is an advanced non-steady-state meteorological and air quality modeling system developed by scientists at Exponent, Inc. and is widely used for predictions up to 200km.

Applications of AERMOD and CALPUFF modelling generate results in slightly different output formats and developing the logic to parse and handle the large filesets produced in a convenient manner was identified as a core objective of the tool to be developed.

2.2 Key features of the Solution

The features of the post-processing tool representing greatest value to EMM include:

- Handling the AERMOD/CALPUFF models (i.e. essentially a simple menu selection option) with the ability to expand the tool to other models if

necessary in the future.

- Conformance to file formats (including specific columnar layouts) and input/output data stipulations by the EMM team
- Capable of accommodating a large number of calculation points and associated file sizes typically of between 100MB up to 5GB+ for the selected model.
- Outputs needed:
    - X,Y,Z statistic files (Excel viewable) listing the specified result by calculation point for a range of pollutant, statistic and averaging period combinations and example outputs of a maximum 24-hour average and 99th percentile 1-hour average
    - ASCII Grid file format for all outputs to allow generation of contour maps of concentrations
    - Generation of Excel-viewable time series files listing contribution to hourly concentration by emission source, for the subset of sensitive calculation points (i.e. certain houses etc).

3 De-risking the project – a phased approach

To some degree the project outcomes were unpredictable, and therefore, to ensure the experimental aspects of the work to be carried out did not unduly exhaust the available budget without the prospect of a meaningful result, a phased approach was taken. In particular, developing a scripted solution to prove the correctness of the proposed calculations prior to building a Graphical User Interface supported this risk minimisation approach. A discussion of these phases follows.

3.1 Phase I – Proving ‘scriptability’ of post-processing

Using the scripting capabilities of the PHP programming language, a scaling calculation mechanism for pollutant, statistic and averaging period settings was built.

The logic developed in this way enabled workflow to apply a linear scaling factor to an emission source provided as a unit release. The benefit of this is that it can be done as a post-processing step, tailored to the chosen model (i.e. AERMOD/CALPUFF model), subsequent to the generation of the unit releases.

A configuration file consumed by the programmatic scripts allowed permutations of settings associated with the pollutant type, as well as statistical and averaging period preferences.

3.2 Phase 2 Production rollout of a scripted solution

On completion of the basic scripting activities and related confirmation of the accuracy of the resulting processed data, a project to roll the solution out to EMM staff was undertaken.

To minimise costs, an ‘on-premises’ server computer within the EMM IT infrastructure network was set-aside as a shared resource amongst staff on which to conduct real-world evaluation of the toolset. A well considered file hierarchy and good system administration practices were applied to aid in the management of large data input and output sets involved in the workflow.

Simple tech choices for the application hosting environment were made including selection of the free and open source XAMP (an abbreviation for cross-platform, Apache, MySQL, PHP and Perl) platform to provide a natural emulation of the globally popular LAMP (Linux Apache MySQL PHP) stack within the Windows operating system environment.

All data processing activities were overseen using command line initiated (so-called CLI) processes. The skilled EMM team were comfortable in this non-GUI environment during this phase in keeping with the goal of avoiding any additional unnecessary costs until the ability of the solution to work in a day-to-day production context was established.

The cost-saving of building experimental solutions using free and open source tools was also appreciated by the EMM team.

3.3 Phase 3 Re-architecture for the AWS Cloud and utilisation of AWS Batch processing and Docker

After discussion with the EMM consulting division and IT administrators, progression to a modern cloud architecture on the AWS (Amazon Web Services) cloud platform was undertaken.

High-Level steps in this process included :-

- Completion of an intensive design effort to determine the most appropriate cloud server infrastructure assets and network to deploy on AWS making extensive use of the AWS Batch (using containers provisioned via AWS ECS) processing services.

- Evaluation of the **scaling and/or parallelization** requirements to ensure the solution load and performance requirements could be satisfactorily met – **the opportunity to significantly reduce the computation time for post-processing was a clear objective in this respect**
- Consideration of the system's multi-user security needs and data flows (Ingress, Transfer and Egress)
- Cost Estimations for ongoing cloud usage – an 'only pay for what you use' model
- Options for ongoing support and expansion of the solution

3.3.1 Selection of AWS CDK Infrastructure as Code – to automate cloud deployment

In keeping with 2pi Software's Infrastructure as code cloud best practice recommendations (refer blog post and linked YouTube video here) the EMM solution takes advantage of AWS CDK to effectively create an 'easy instantiate' sandbox. Additionally, this approach facilitated provision of cloud infrastructure using a **Continuous Integration/Continuous Deployment (CI/CD) pipeline** – a mechanism that can greatly speed up application of key feature additions or system patches and upgrades.

3.4 Phase 4 – Full feature GUI console for volume file uploads and compute intensive parallelised processing

To take full advantage of the powerful AWS cloud architecture, the proven production-grade scripted computational capability was transitioned to an environment where substantial processing activity could be completed by essentially activating multiple AWS Batch processing container instances (i.e. with AWS ECS), each hosting a suitably configured Docker image.

To optimally manage this environment running multiple batch jobs in parallel, a purpose-built GUI console was developed using the Cosine Business Framework (formerly Cmfive). This enabled the following features and capabilities :-

- Convenient handling of large file sets hosted on AWS S3 – an abstraction of the lower-level technical storage details allows users to navigate the folders and file locations in a logical and natural manner.
- Fine-grained secure User access control – individual workflows are siloed to avoid inadvertent impingement on running jobs and/or filesets of other users
- Enable real-time monitoring and dashboard-tracked status visualisation – additional complementary low-level monitoring is also provided by the AWS console.
- Adjustable computational parameters to shape/influence the calculation model

4 Business Benefits for EMM from the transition to the AWS Cloud

The business Benefits to EMM of migration to the AWS cloud include :-

- Scalability – given the large datasets and computational complexity of the modelling algorithms, the cloud offers an elastically expandable set of resources, which an on-premises server environment cannot provide
- Security can be more easily controlled and monitored in accordance with a tightly defined permissions model. The AWS platform supports sophisticated worlds best practice security oversight.
- Adoption of Infrastructure as code using AWS CDK – the cloud, in this way, becomes highly testable via scripts can be applied to verify that the live cloud environment confirms to key governance rules and matches the

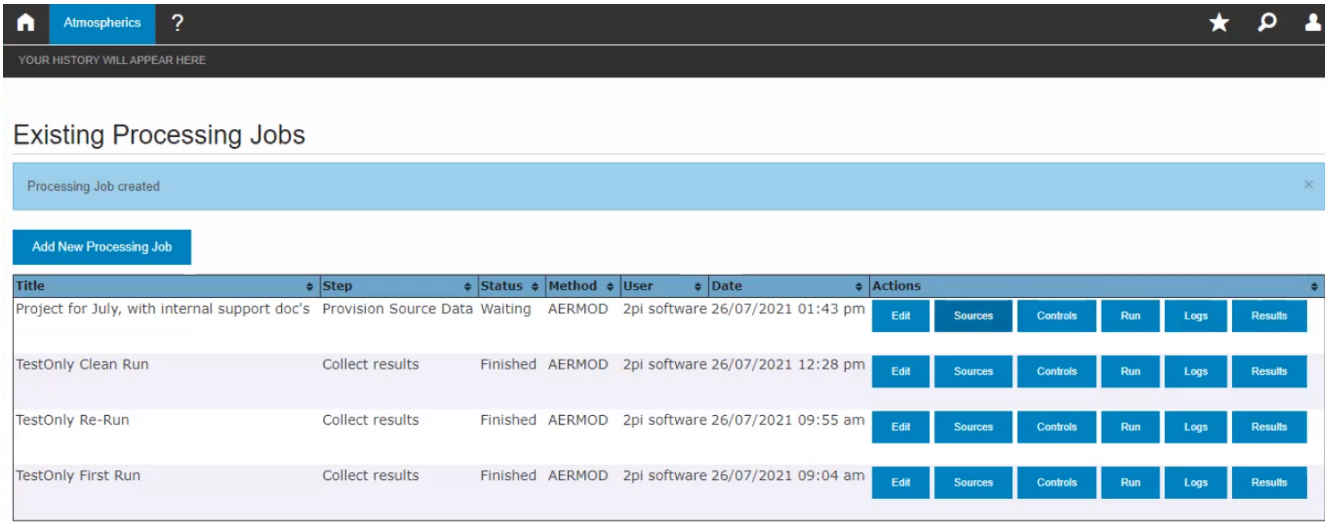
intentions of the Ops team
- Fast turnaround of solution and environment updates and fixes without impacting the existing EMM server network

5 ‘How it works’ summary (including example screenshots)

5.1 Mimicking traditionally scripted computational processes in a Graphical User Interface (GUI) – convenience and control

In essence, as a first step, the EMM team needed to verify that the post-processing enhancement they wished to implement, which involved a degree of experimentation, could be successfully verified as being algorithmically sound.

Once this had been completed and accepted, to avoid unnecessary costs, a path was taken to replicate some of the conveniences and intrinsic nature of the scripted computations in a modern graphical user interface.

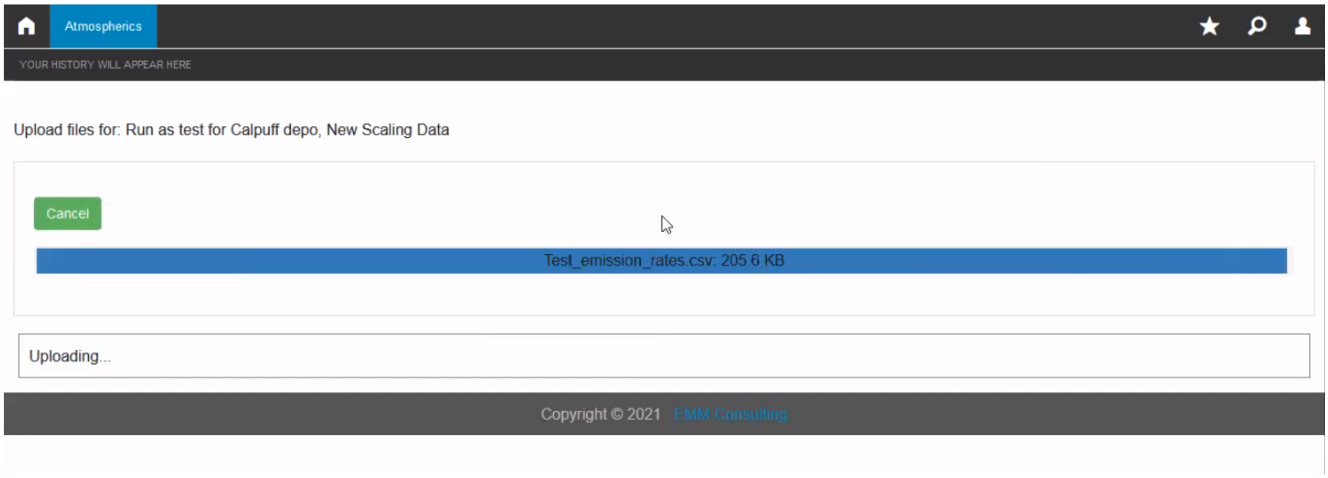


The above screen layout showcases some key aspects of this :-

- All processing jobs are named and listed along with key attributes such and Time-Stamp and model designation
- Each job is linked directly to the processing steps required, each one actionable via a button-click. These include :-
    - **Edit** – alter naming and configurational options for the job
    - **Sources** – this option allows selection of a suitable collection of files that will act as the input feedstock for the computational processing step – generally this fileset can be up to 12Gb in size.
    - **Controls** – this option allows selection of a suitable collection of configurational files and settings that will dictate preferred options regarding the content and format of the generated results data set
    - **Run/Go/Stop** – Initiates the job sequence step as a batch process that is spawned within the AWS Batch service. Termination of a job or reactivation of a paused job can be enabled via this screen control also.
    - **Logs** – viewability of logged status and job progress information allows the user to monitor progress of the compute activities and observe any matters that may prompt intervention or recommencement of the processing work
    - **Results –** a listing of the results fileset provides visual feedback on completion of the workflow steps.

Context sensitive Help is also provided within the application.

5.2 Upload Once, assign and use many times

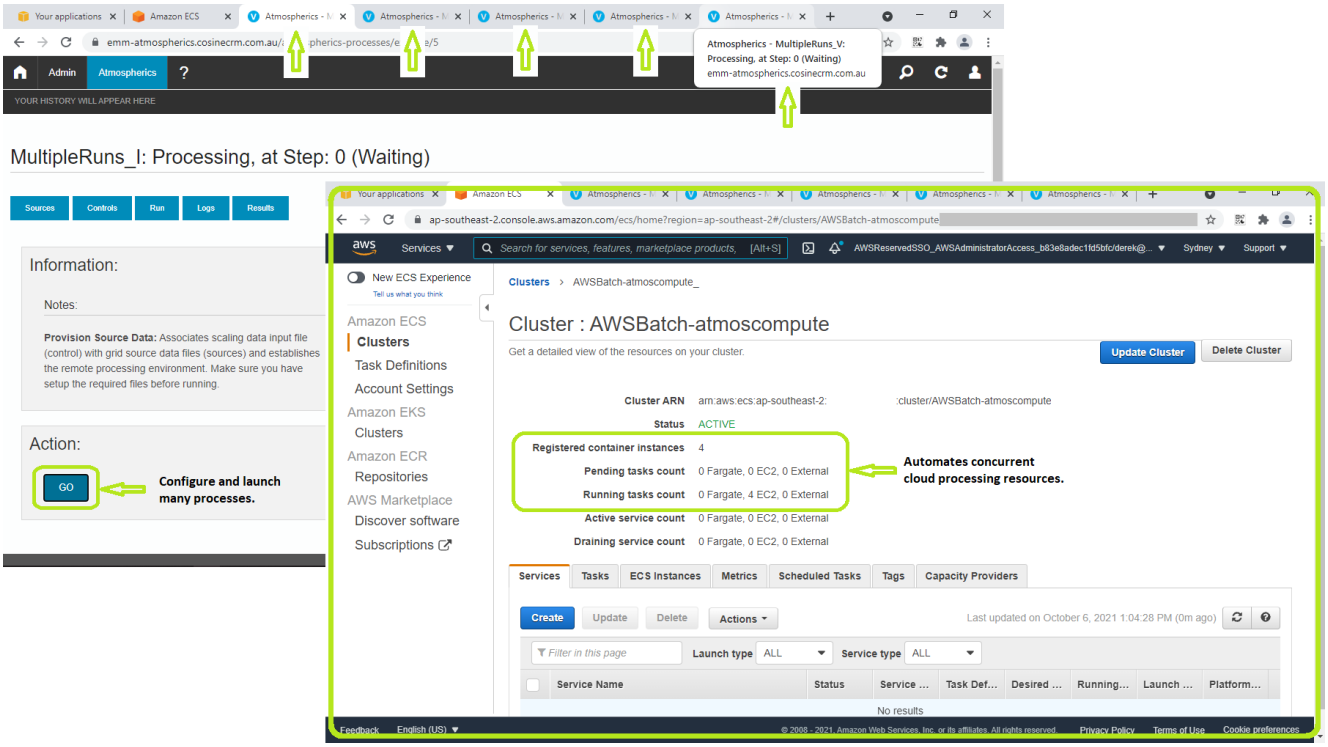


A specialist 2pi Software facility, the Cloud Uploader (*more on this in a later section of this document*) is provided as part of the EMM solution to allow secure and convenient uploading and assignment of both source and control data sets to the AWS cloud for later ingestion by the processing jobs.

Both Source and Control files uploaded in this way can be reused across multiple processing jobs running in parallel taking advantage of the notionally infinite processing and storage capacity available through the AWS cloud platform.

Read access at scale by multiple active systems is an intrinsic aspect of the AWS S3 storage 'bucket system' – file references to objects stored on AWS S3 system allow, effectively, unlimited processes to access and retrieve data simultaneously from this location. This attribute of the S3 system is facet of the inherent scalability of the AWS cloud platform.

5.3 Parrallelised ‘multiple’ running jobs – scales with ease



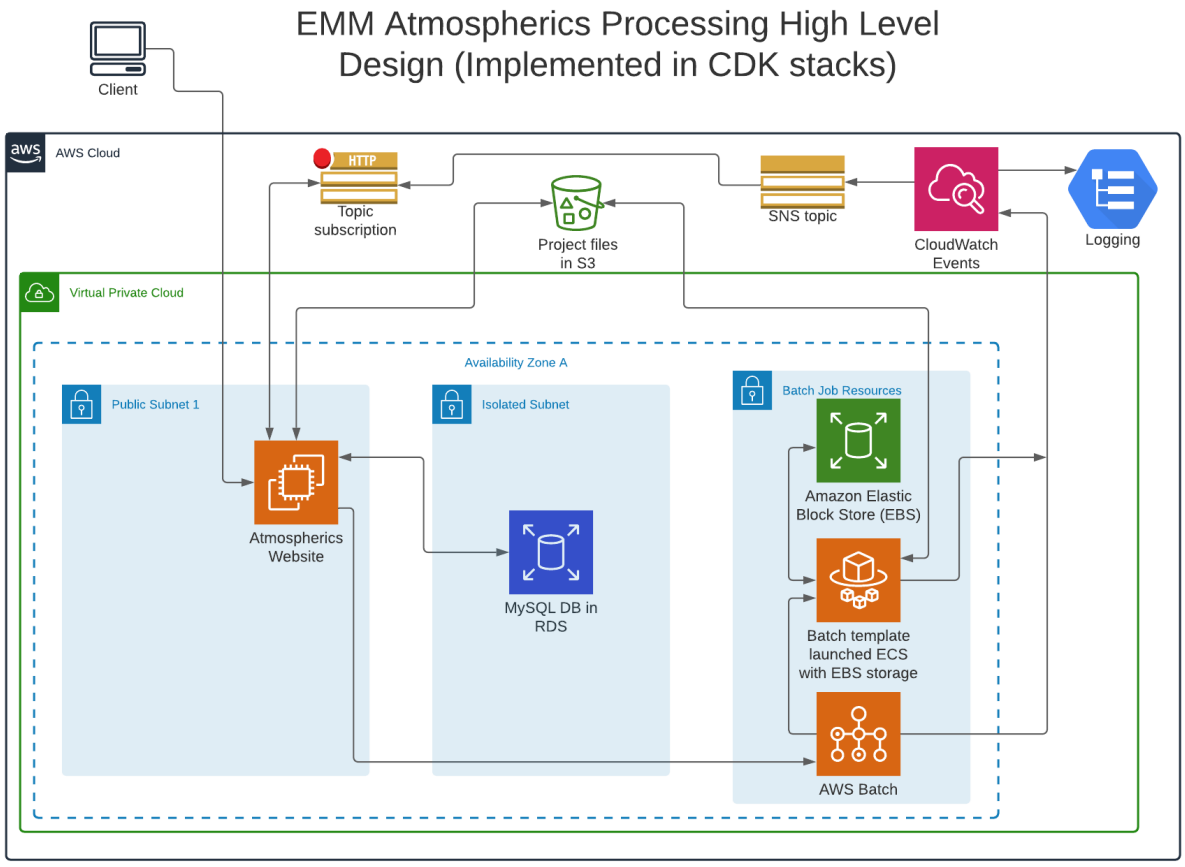
The system supports multiple processing runs on a common data set via the following steps :-

- the screen controls allows creation of multiple processing jobs, reusing a common set of source data uploads
- operators can individually attach distinct/appropriate control files to each job
- by launching all jobs in relatively quick succession, these will run concurrently within completely separate resource environments on the server side
- naming jobs by project activity and source data is recommended practice

6 Solution Architecture and Tech Tools

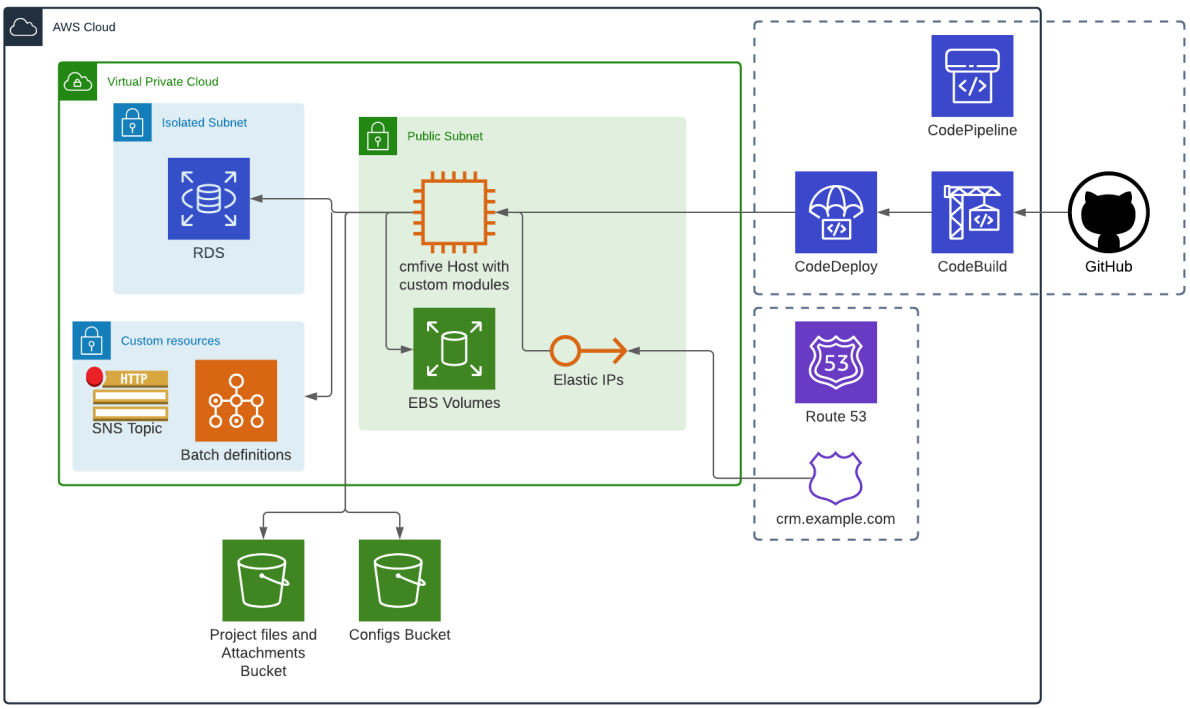
6.1 Use of AWS CDK to automate the entire AWS Infrastructure creation steps – a one-click set-up

The below diagram illustrates many of the key AWS services used to enable the EMM solution – all creatable via powerful AWS CDK infrastructure-as-code scripts.



6.2 Enablement of Continuous integration/Continuous Delivery (CI/CD) pipelines for fast feature-fix ‘updates to live’

The below diagram illustrates the manner in which modern so-called DevOps best practice has been applied to allow fast feature and fix updates to the EMM solution :-



6.3 The Tech Stack featuring AWS Batch, Docker and the Cosine Business Framework

6.3.1 Extensive use of a range of powerful AWS infrastructural services

The AWS environment greatly facilitates fast moving software development for secure scale environments with mature integration and inter-operability between cooperating sub-systems. The EMM solution takes advantage of a number of AWS features and services including :-

- Storage and Caching – S3, EBS, CloudFront
- Monitoring and Events – CloudTrail, CloudWatch
- VPC (Virtual Private Cloud) design and management
- Complex Account and Sub-Account management using AWS Control Tower

6.3.2 AWS services supporting modern scalable application deployment

From a software development perspective, a significant advantage of the AWS cloud environment is derived from the range of in-built features and services that reduce the amount and complexity of code that is required to enable robust scalable functionality. In the case of the EMM solution, the following AWS services were materially assistive :-

- High Availability Databases via AWS Relational Database Service – RDS
- Lambda functions (serverless computing) and API Gateway
- AWS SES to comprehensive email/messaging support
- AWS SQS message queuing supporting orchestration of key events and triggered processes
- AWS Eventbridge to coordinate activities within linked AWS Lambda functions

6.3.3 Modern Software Engineering Techniques in the cloud to handle complete Workflow

The EMM solution takes the fullest advantage of modern software development practices to offer an enterprise grade scaled-up platform that supports the implementation design.

Key architectural and technological approaches adopted in developing the EMM solution include :

- Adherence to Devops/Site Reliability Engineering principles and practise
- Agile Methodology (including SCRUM) governance of development cycles
- MicroServices – modular decoupled granular API services (e.g. AWS Lambda)
- Continuous integration/Continuous Deployment (of both application and infrastructure) – this allows faster feature updates of fixes to be applied to production software systems
- Infrastructure as Code (using the powerful AWS CDK toolkit)

6.4 Unlimited processing instances using Docker containers

Docker containers offer a modern era mechanism to manage and deploy multiple instances of computing power (i.e. essentially standalone servers). From a software engineering perspective these allow consistency of the application configurability across development, testing and production environments, even across different operating systems.

The flexibility of Docker containers to allow unlimited discrete instances of

computing power to launch simultaneously delivers significant benefits to the EMM solution including :-

- Reducing the overall post-processing elapsed time duration from days to hours
- Scalability to handle hitherto unprecedented computing/modelling needs
- Homogeneity across development, testing and production environments reducing the incidence of adverse configurational/set-up variations
- An architectural underpinning and software development environment that greatly conveniences feature and/or fix updates or solution extension to accommodate new EMM business requirements (eg noise and ground water modelling)

6.5 Custom components built using free and Open Source software toolsets (includes the Cosine Business Framework)

The globally popular LAMP stack (Linux, Apache, MySQL, PHP) provides the core software building blocks enabling the EMM solution user role and permissions features, in addition to the bulk file upload capability and asynchronous invocation of AWS Lambda functions.

Also underpinning the solution is the free and open source Cosine Business Framework (formerly Cmfive). More information about the Cosine Business Framework can be found here – http://cmfive.com. Features of the Cosine Business Framework include :

- User and group management and role based permission management
- Task and workflow management
- Modular architecture for easy extensibility and separation of concerns

- Restful API Integration with other software

6.6 Use of a fault tolerant multi-user secure Cloud-Uploader

As adoption of cloud usage grows globally, the ability to upload large data-sets in a logical and secure manner is an increasingly critical part of software system cloud deployments.

The EMM solution benefits from a fit-for-purpose modular Cloud Uploader facility built using Cmfive (the Cosine Business Framework) to allow bulk upload of large filesets by multiple parties to the AWS S3 bucket cloud storage environment.

Features of the Cloud Uploader applicable to the EMM solution include :-

- resilient uploads in low-bandwidth or extended duration circumstances
    - Retries logic in low bandwidth conditions
    - Management of partial or interrupted fileset cluster upload
- convenient bulk file naming and categorization
- Natural language metadata capture pertaining to uploaded data input that is easily provided even by non-technical users.
- User identification and access control via an associated administration console
- Temporary shareable web links (i.e. URLs) with configurable expiration

The Cloud Uploader module is built using the 2pi Software released Cosine Business Framework (formerly known as Cmfive).

7 About EMM

EMM is one of Australia’s leading planning and environment consulting specialists. The organisation’s integrated services include planning and environmental assessment, as well as specialist areas such as acoustics, air quality, contaminated land, ecology, heritage, water, social planning, soil and spatial solutions.

EMM is an employee owned business with offices in Sydney, Newcastle, Adelaide, Brisbane, Melbourne, Perth and Canberra. The organisation, headquartered in Sydney, has offices in a number of Australian cities, and offers a full suite of services nationally and internationally delivering outcomes across a wide range of project areas.

8 About 2pi Software

Rural/Remote Australian digital capability

2pi Software is a software engineering company based in the Bega Valley with ambitions to develop products and service markets Australia-wide.

2pi Software is staffed with people who passionately pursue excellence in the software engineering craft and consistently apply the intellectual rigour required to build long-lasting, supportable and well-documented systems.

Core Competencies

At its core, 2pi Software is a software engineering company specialising in cloud based systems and environments. Our problem-solving and solution development capabilities are well-honed over many years, and the company passionately pursues excellence in the software engineering craft, and applies the intellectual rigour needed to build long-lasting, supportable and well-documented systems.

In summary, our skills and experience cover :-

- High-end software engineering – this is the 2pi Software 'DNA'
- Cloud System implementation and interconnectivity. 2pi Software are a select AWS Consulting Partner
- ERP/CRM and practise management software including
    - Workflow – 'prospects to payments'
    - Project/task/time tracking
    - Invoice/quote auto-generation (including one-step bulk creation)
    - Data dashboard
    - Project management
    - API Integration (including Xero and Quickbooks)
    - Full-suite Reporting
    - Productivity, Salesforce and Workflow Automation
    - Document Management
    - Bespoke Software Extensions/Integrations
- Systems Integration (SI) – Interconnectivity of disparate systems – we provide the 'glue'
- B2B communication and automation
- GIS systems – training, support and development
- Promotion and adoption of Open Source and Open Standards
- Enterprise Web Design, Development and ongoing support and maintenance

The 2pi Software vision includes the following goals and initiatives :-

- Establishing software development services as a viable long-term business

activity in the region
- Promoting greater participation in software development as a career option for young people through regular ‘coding’ events, and liaison with local schools
- Advocacy of Intellectual Property asset creation as a potential key driver of job creation in our region
- Active organisation and involvement in entrepreneurial events such as the Annual StartUp Camp
- Continuous communication with local policy-makers about the possibilities of ICT as an economic growth enabler for the region
- Frequent networking with other professionals in the region to increase awareness of the potential impact that technological innovation can have in their respective domains, so-called Sector Seeding
- Maintaining strong business links to like-minded groups in nearby centres Cooma and Canberra

8.1 Community engagement for upskilling local people

Championing High Skill Jobs in Regional Australia from the Bushfire-affected Farming Community of the Bega Valley which is currently emerging from a lengthy period of drought.

The 2pi Software office is located on Carp St, the main street of Bega town, and the company management and staff live and work amongst the almost 60 farming families that currently make up the producers in the local Dairy industry.

2pi Software has a strong record of supporting the local community over many years, and through the recent tough times of bushfire and drought. In this respect that history of the organisation, in the following paragraphs, bears this out strongly :-

Since the company's inception, the 2pi Software team have promoted a community goal of creating 1000 tech sector jobs in our region by 2030.

To build a tech sector, the company has been at the heart of IntoIT Sapphire Coast (www.intoitsapphirecoast.com.au).

In 2014, 2pi Software funded the creation of Bega's first ever Digital Co-Working Space called 'CoWS Near The Coast' (visit cowsnearthecoast.com.au). The site was honoured in February 2015 to have been visited by the NSW Governor (now Governor General) General David Hurley who was very encouraging of the initiatives underway.

In May 2016 the 2pi Software team were the lead organisers behind Regional Innovation Week in the Bega Valley (begainnovation.com.au) a programme featuring 10 events celebrating a number of aspects of creativity, technology and entrepreneurship.

Coding was also a big part of Innovation week, and 2pi Software has played a big role in running hackathons since November 2014, and continue to do so

The company actively takes part in National Science week and have been instrumental in organising demonstration of scientific applications of drones/quadcopters, robotics-building workshops, synthesiser-making, 3D printing, virtual reality and coding.

In 2018, members of the 2pi Software team ran the highly popular Bega's AgTech Days programme (begaagtech.com.au) featuring a keynote address by Barry Irvin, Executive Chair of Bega Cheese, as well as invaluable showcasing of the relevance of tech to the local farming community.

In 2019 and 2020 2pi Software delivered a Federal Government sponsored Regional Employment Trials programme in the Bega Valley helping a number of Job Seekers learn Job Ready technology.