ANACONDA
Powered by Continuum Analytics®

WHITEPAPER

Low-income countries
Enhanced investment scenario
Lower-middle-income countries
Enhanced investment scenario
UNDER-5 MORTALITY RATE
mortality rate (deaths per 1,000 livebirths)
CONVERGENCE TARGET
2013
Jan.
Febr.
März
April
Mai
Juni
Juli
Aug.
Sept.
Okt.

Productionizing and Deploying Secure and Scalable Data Science Projects

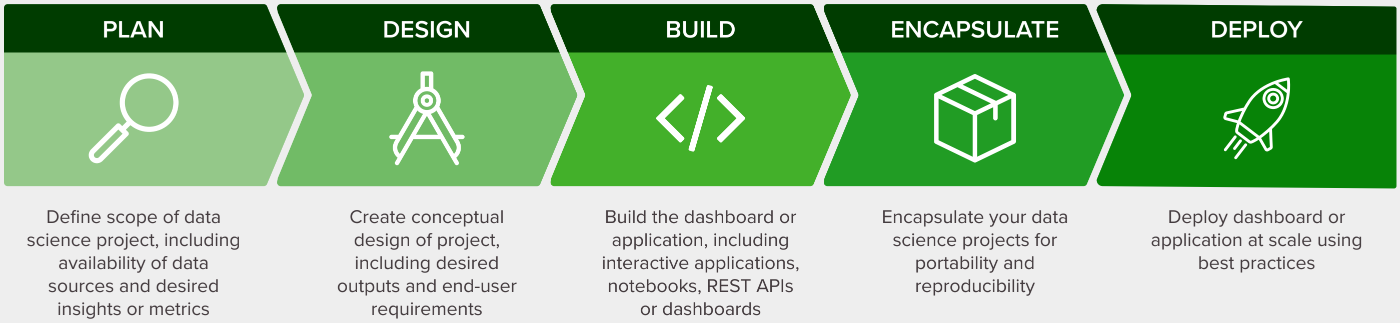
Christine Doig, Senior Data Scientist & Product Manager, Continuum Analytics
Kristopher Overholt, Product Manager, Continuum Analytics

In This Whitepaper

An end-to-end data science workflow includes stages for data preparation, exploratory analysis, predictive modeling and sharing/dissemination of the results. At the final stages of the workflow, or even during intermediate stages, data scientists within an organization need to be able to deploy and share the results of their work for other users (both internal analysts and external customers) to consume.

Traditional data science project deployments involve lengthy and complex processes to deliver secure and scalable applications in enterprise environments. The result is that data scientists must spend a nontrivial amount of time preparing every new data science asset for production as well as setting up, configuring and maintaining deployment infrastructure.

Data Science Deployment Process



PLAN

Define scope of data science project, including availability of data sources and desired insights or metrics

DESIGN

Create conceptual design of project, including desired outputs and end-user requirements

BUILD

Build the dashboard or application, including interactive applications, notebooks, REST APIs or dashboards

ENCAPSULATE

Encapsulate your data science projects for portability and reproducibility

DEPLOY

Deploy dashboard or application at scale using best practices

But why take valuable time away from data exploration and analysis workflows when Anaconda Enterprise can handle the process for you?

Through the power and flexibility of Anaconda Enterprise, any application, notebook or model can be encapsulated and securely deployed on a server or scalable cluster, and the deployed applications can be easily shared within your data science team or enterprise organization—all with the single click of a button.

In this whitepaper, we will examine:

- The traditional approach to preparing your data science projects for deployment
- How to leverage Anaconda Project to encapsulate your data science projects
- The unparalleled advantages of productionizing, encapsulating and deploying your data science projects through the power of Anaconda Enterprise

Traditional Approach to Preparing Your Data Science Projects for Deployment

First, let's take a look at the numerous steps involved in a typical data science project deployment. As you prepare your data science assets to be productionized, a number of considerations must be made to ensure that the deployed projects and applications are robust, performant, reliably accessible, secure and scalable, among other factors. At some intermediate or later stage in their workflow, data scientists want to encapsulate and deploy a portion or all of their analysis in the form of libraries, applications, dashboards or REST API endpoints that other members of the data science team can leverage to further extend, disseminate or collaborate on their results.

When starting an analysis from a Python script or Jupyter notebook, there are many different approaches that can be used to transform this code into an asset that can be leveraged and consumed by many different users and roles within a data science team. Depending on the desired output, different types of data science assets can include:

- Reports that can be deployed as hosted, static notebooks
- Code that can be encapsulated inside of a package and shared for reusability
- Dashboards and applications that can be deployed and used across an organization
- Machine learning models that can be embedded in web applications or queried via REST APIs

When defining a process for data scientists and users to productionize and deploy their own custom applications, notebooks or dashboards within your organization, you'll need to ensure that the deployed applications and compute infrastructure are robust, stable, secure and scalable. The many factors to consider when deploying any type of data science asset or project within your organization include:

- **Provisioning compute resources**
  Before you deploy your data science project, you'll need to reserve and allocate compute resources from a cloud-based provider or one or more bare-metal machines within your organization that will act as deployment servers. Once you've identified compute resources for your use case, you'll need to install and configure the production environments on the machines, including system-wide configuration, user management, network and security settings and the system libraries required for your applications.
- **Managing dependencies and environments**
  When you deploy a data science application, you'll want to ensure that it's running in an environment that has the appropriate version of Python, R and libraries that your application depends on, including numerical, visualization, machine learning and other data science packages and their dependencies.
- **Ensuring availability, uptime and monitoring status**
  Once you've deployed your data science application, you'll need to ensure that your application's runtime environment and processes are robust and reliably available for end-users. You might also want to set up log aggregation, uptime monitoring or logging alert systems such as Sentry or Elasticsearch/Logstash.

Factors to Consider when Deploying Data Science Assets or Projects

- Provisioning compute resources
- Managing dependencies and environments
- Ensuring availability, uptime and monitoring status
- Engineering for scalability
- Sharing compute resources
- Securing data and network connectivity
- Securing network communications via TLS/SSL
- Managing authentication and access control
- Scheduling regular execution of jobs

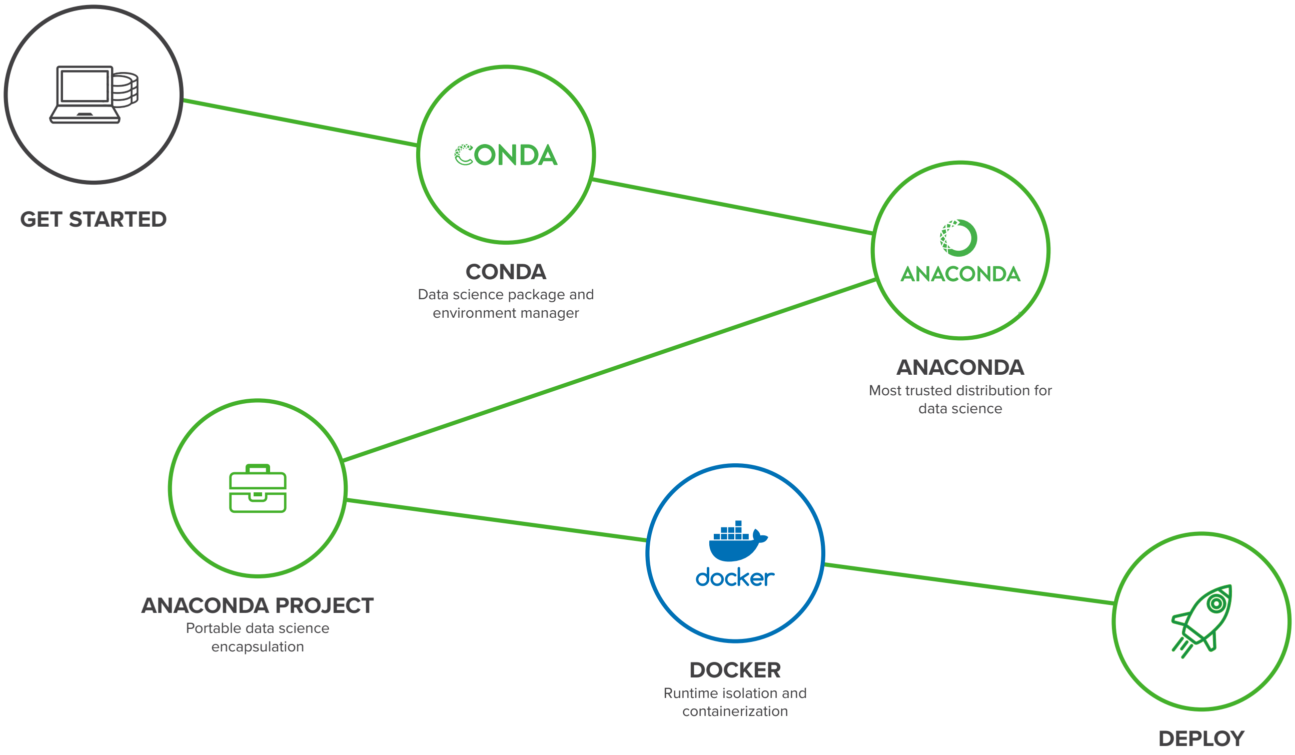
- **Engineering for scalability**
  Before you deploy your data science project, you'll need to estimate the scalability limits of the computational load and overhead of the applications. To allow for the scalability of demanding applications or large concurrent usage, you might need to implement and configure load balancing and reverse proxy functionality in your web application servers such as NGINX and Gunicorn so that your application is responsive and scalable under heavy load and peak usage conditions.
- **Sharing compute resources**
  When multiple users in your organization are running exploratory analyses, sharing and collaborating within notebooks, and deploying various data science applications, you'll need to ensure that the compute resources on your cluster can be reasonably shared between users and applications. This can be accomplished by using resource managers or job schedulers, which are typically installed and configured on a cluster and can be configured with job queues for different types of applications and needs.
- **Securing data and network connectivity**
  The data science applications deployed within your organization will likely be accessing data stored in files, a database, or a distributed/online file system. You will need to ensure that your deployment server(s) have the appropriate network/security configuration and credentials for your applications to securely access the data and file servers without exposing your data or compute resources to risky situations.
- **Securing network communications via TLS/SSL**
  When you deploy an application, dashboard or notebook, you will likely want to utilize end-to-end encryption for your network and API communication to ensure that traffic between your users and the application is secure. This might involve configuring your web application servers to use TLS/SSL certificates, certificate authorities and secure proxies as needed with the appropriate hooks and layers for your end-user applications.
- **Managing authentication and access control**
  If you're deploying a data science application to your own infrastructure and want to restrict access to a subset of users within your organization, you'll need to implement layers of authentication and access control that can be used on a per-application or per-project basis. This can be implemented via various HTTP authentication methods, a web framework such as Django or Flask (with various authentication back-ends), or a third-party authentication service/API.
- **Scheduling regular execution of jobs**
  After you deploy your data science project, you might want to incorporate scheduled execution intervals for a notebook, dashboard or model so that the end-users will always be viewing the most up-to-date information that incorporates the latest data and code changes. This can be accomplished via a job scheduler or workflow/pipeline manager. Using these tools, you can configure your data science application to run at regular intervals every few minutes, hours or days to perform tasks such as data ingestion, data cleaning, updated model runs, data visualization and saving output data.

As we can see from this list, setting up, configuring and maintaining deployment infrastructure can be a lengthy, time-consuming process, taking data scientists away from what they do best: exploring and analyzing data.

> *When defining a process for data scientists and users to productionize and deploy their own custom applications, notebooks or dashboards within your organization, you'll need to ensure that the deployed applications and compute infrastructure are robust, stable, secure and scalable.*

But what if Anaconda could handle the process instead, and provide your team with secure, scalable and reproducible data science deployments with the single click of a button?

Data Science Encapsulation Process


GET STARTED
CONDA
CONDA
Data science package and environment manager
ANACONDA
ANACONDA
Most trusted distribution for data science
ANACONDA PROJECT
Portable data science encapsulation
docker
DOCKER
Runtime isolation and containerization
DEPLOY

Leveraging Anaconda Project for Data Science Project Encapsulation

If you are already using Anaconda Distribution, then you probably know that Anaconda and Docker make a great combination to empower your development, testing, encapsulation and deployment workflows. But using Docker alone as a data science encapsulation strategy still requires coordination with your IT and DevOps teams to write your Docker files, install the required system libraries in your containers and orchestrate and deploy your Docker containers into production.

Making data scientists worry about infrastructure details and DevOps tooling takes precious time away from their most valuable skills: discovering insights in data, modeling and running experiments, and delivering consumable data-driven applications to their team and end-users. By working directly with our users and customers and listening to the needs of their data science teams, our Anaconda experts identified the need for a more convenient approach to data science project encapsulation.

Anaconda Project, leveraged by Anaconda Enterprise, enables you to easily encapsulate data science projects and make them fully portable and deployment-ready across different operating systems. It automates the configuration and setup of data science projects, such as installing the necessary packages and dependencies, downloading data sets and required files, setting environment variables for credentials or runtime configuration and running commands.

> *Anaconda Project, leveraged by Anaconda Enterprise, enables you to easily encapsulate data science projects and make them fully portable and deployment-ready across different operating systems.*

However, while these tasks are essential to any data science workflow, preparing your data science projects for production in addition to fully managing the security, scalability and availability needs of your organization will require a complete end-to-end data science platform.

What your organization truly needs is an enterprise-ready, secure and scalable data science platform that empowers teams to manage dependencies and data, securely govern and version control data science assets, share and collaborate, and deploy data science projects backed by enterprise scalable compute and data sources.

What your organization truly needs is Anaconda Enterprise.

Anaconda Enterprise: Secure, Scalable and Reproducible Data Science Deployments

The data science deployment and collaboration functionality in Anaconda Enterprise leverages Anaconda Project plus industry-standard containerization with Docker and enterprise-ready container orchestration technology with Kubernetes to provide secure isolation and scalability of user-deployed applications. This productionization and deployment strategy makes it easy to create and deploy data science projects with a single click for projects that use Python, R or anything else you can build with the 1,000+ data science packages in Anaconda.

All of this is possible without having to edit Docker files directly, install system packages in your Docker containers or manually deploy Docker containers into production. Anaconda Enterprise handles all of that for you, so you can focus your attention where it belongs: on data science analysis.

> *Anaconda Enterprise is a true end-to-end data science platform that integrates with all the most popular tools and platforms, offering you secure and scalable data science project collaboration and empowering your team to easily encapsulate, productionize and deploy their work with the single click of a button.*

The result is that any project a data scientist can create on their machine with Anaconda Distribution can then be deployed to an Anaconda Enterprise cluster in a secure, scalable and highly available manner with just a single click. Anaconda Enterprise uses Anaconda Project and Docker as its standard project encapsulation and deployment strategy to enable simple one-click deployments of secure and scalable data science applications for your entire data science team.

Anaconda Enterprise is a true end-to-end data science platform that integrates with all the most popular tools and platforms, offering you secure and scalable data science project collaboration and empowering your team to easily encapsulate, productionize and deploy their work with the single click of a button.

Data Science Deployment with Anaconda Enterprise

ANACONDA
ENTERPRISE

GET STARTED

ANACONDA ENTERPRISE

Secure and scalable data science collaboration, productionization and deployment platform

ONE-CLICK DEPLOY

Build, encapsulate and deploy data science projects with the single click of a button

With Anaconda Enterprise, your organization will be able to leverage a range of powerful functionality, including:

- The ability to deploy data science projects—including interactive applications, notebooks, dashboards and machine learning models with REST APIs—using the same 1,000+ libraries in Anaconda Distribution that you and your data science team already know and love
- Data science application encapsulation, containerization and cluster orchestration using industry standard tooling
- A scalable on-premises or cloud-based deployment server with configurable cluster sizes
- Single-click functionality for secure data science project deployments, complete with enterprise authentication/ authorization and secure end-to-end encryption
- The ability to share and collaborate on deployed applications that integrate with enterprise authentication and identity management protocols and services
- Centralized administration and control of deployed applications and cluster utilization across your organization
- Connectivity to various data storage back-ends, databases and formats

The data science deployment capability in Anaconda Enterprise builds on existing features in the Anaconda Distribution to enable powerful end-to-end data science workflows, and provides your data science team with on-premises package management and governance, secure enterprise notebook collaboration and project management, and scalable cluster workflows with Hadoop, Spark, Dask, machine learning, streaming analytics and much more.

Summary

Anaconda Enterprise is the go-to solution for any organization using Python or R to deploy data science projects into production. Through the power and flexibility of Anaconda Enterprise, any data science application, notebook or model can be encapsulated and deployed on a server or scalable cluster, and the deployed applications can be easily and securely shared within your data science team or enterprise organization.

The data science deployment and collaboration functionality in Anaconda Enterprise leverages Anaconda Project plus industry-standard containerization and enterprise-ready container orchestration technology, enabling users to deploy their data science applications into production with confidence.

> *Anaconda Enterprise is the go-to solution for any organization using Python or R to deploy data science projects into production.*

Anaconda Enterprise is a true end-to-end data science platform that integrates with all the most popular tools and platforms, and provides your data science team with an on-premises package repository, secure enterprise notebook collaboration, data science and analytics on Hadoop/Spark and secure and scalable data science deployment.

About Anaconda

With over 4.5 million users, Anaconda is the world's most popular and trusted data science ecosystem. We continue to innovate by leading development on open source projects that are the foundation of modern data science. We also offer products and services that help support, govern, scale, assure, customize and secure Anaconda for enterprises. Visit www.continuum.io.

Edited by Rory Merritt, Technical Marketing Copywriter, Continuum Analytics