ANACONDA

ANACONDA WHITEPAPER

THE JOURNEY TO OPEN DATA SCIENCE

By: Christine Doig, Data Scientist and Product Marketing Manager

March 2016

CONTINUUM® ANALYTICS

IN THIS WHITEPAPER

Migrating from traditional analytics to modern Open Data Science (ODS) is one of the most important trends of the decade. However, although this migration is critical to the core mission of many enterprises, practicing Data Science effectively has remained a thorny challenge. Part of this is due to the use of older proprietary data technologies that do not provide the flexibility and power needed to channel the full potential of a Data Science team — not just Data Scientists and Developers, but business experts and other stakeholders. These proprietary technologies are not innovative, transparent or community supported, and they typically cannot quickly adapt to changes as projects develop.

Fortunately, modern technologies, by leveraging developer talent from across the world through the open source paradigm, now provide comprehensive, best-of-breed and cost effective ecosystems. Not only do these technologies meet current challenges, but they also address potential future struggles, since they can rapidly adapt in ways that proprietary Data Science technologies cannot.

In this paper, you'll learn why Open Data Science is the foundation to modernizing data analytics, and:

- In what ways availability, interoperability, transparency and innovation are some of the most important benefits of the ODS approach
- The best path and important considerations when moving to ODS
- Why the Anaconda platform is the best solution for ODS
- How you can leverage your current investment in analytics while moving to ODS

OPEN DATA SCIENCE

Open Data Science is not a single technology, but a revolution within the Data Science community. ODS is an inclusive movement that makes open source tools for Data Science — data, analytics and computation — easily work together as a connected ecosystem. Data is everywhere, and ODS has emerged to meet the demands of modern analytics. The promise of ODS and the fundamental principles are:

- Availability
- Innovation
- Interoperability
- Transparency

Availability. ODS avoids the drawbacks of former approaches to Data Science that limited the access to holistic tools for all members of the Data Science team. Traditional approaches relied on proprietary technologies that were expensive (sometimes prohibitively so), designed for a single isolated purpose (such as visualization but not machine learning) and did not easily interoperate. For these reasons, data analytics teams were forced to commit to long evaluation periods and to work with monolithic tools that had to be integrated together and were less than optimal for their needs.

With ODS, all of these concerns evaporate. Technologies are nonproprietary, free and open source, making powerful tools available for both individuals and enterprise teams. This access allows anyone to use a large and flourishing interconnected ecosystem of analytic technologies, giving the Data Scientist the best tool for his or her problem today and tomorrow as technology advances. This future proofing is underscored with the open source guarantee — source code is available no matter what the circumstances — permitting the Data Science team to confidently build solutions that will endure the test of time.

Innovation. Data Science thrives on innovation, but traditional approaches tied to proprietary technologies are typically slow and rigid. When changes to proprietary products are made, they are on vendors' schedules. It sometimes took years for innovations to surface to the market.

Data scientists have become impatient with proprietary software that, because of its limitations, moves at a glacial pace and causes them to perform time-consuming, complex workarounds.

Meanwhile, the open source community delivers a simple, graceful way to deal with today's issues. With ODS, innovation comes from many different open source communities, including Python, R, Java, Hadoop, Scala, Julia and others. In these communities, a very broad based peer review brings fresh suggestions to quickly optimize the research trajectory, making the latest science instantly available through the open source paradigm.

In contrast to this slow moving proprietary approach, ODS — because of its hyper-flexible, customizable and inexpensive nature — invites experimentation, quick-to-fail approaches and rapid prototyping. This spreads the fast-paced innovation velocity of open source to the Data Science team. Ideas and contributors from across your Data Science team, including domain experts such as biologists, physicists, economists or others, are introduced to new ways of looking at the data. This leads to new models that deliver cutting-edge insights and unlock value that has been trapped in your data.

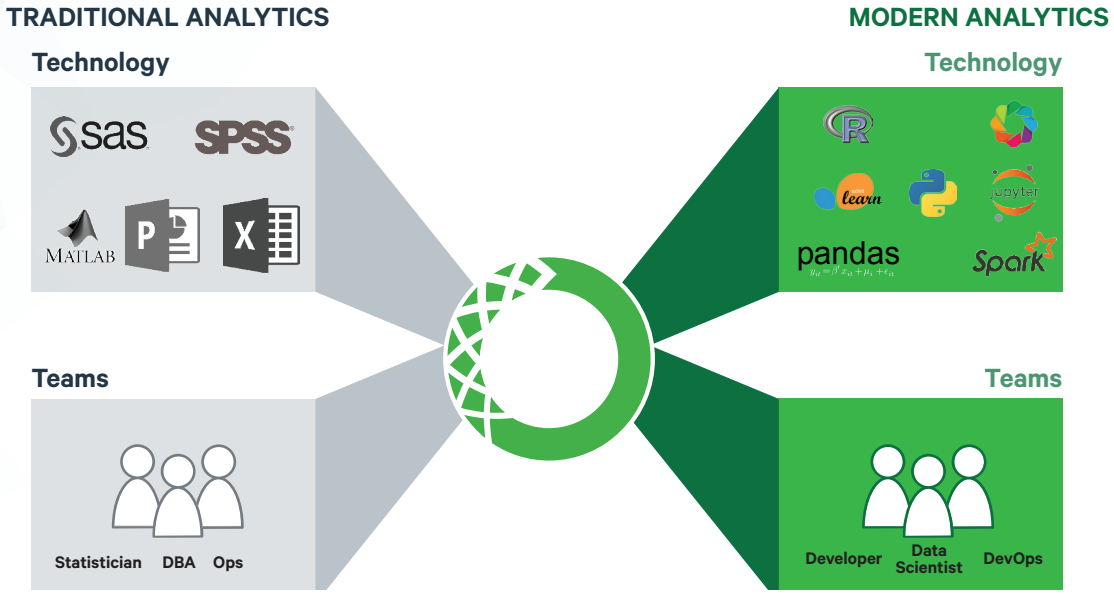
Interoperability. Monolithic tools typically integrate with their own suite but are either closed to integrating with outside tools or provide an inferior, often slower method when doing so. Many times the suite is a composite of tools acquired by the vendor over time, and these tools are clunky and sluggish when they do interoperate.

ODS excels at integrating tools from across the open source ecosystem. By the very nature of open source, contributors look to leverage existing know-how, code and tools. This type of approach eradicates the typical boundaries established by proprietary vendors that prevent interoperability. Python, in particular, is known as the "glue language" that makes it easy to create code to connect software components into a seamless ecosystem. This includes the ability to embrace new open source technologies, such as Hadoop and Julia, and to leverage legacy code, including Fortran and C/C++, into modern solutions.

Transparency. Proprietary software is a "black box" where the internal structure and processing is confidential. The algorithms and their implementations were encapsulated in the proprietary software. In the past this was acceptable, since the software technology era was in its infancy and there were few choices to solving Data Science problems. Thanks to decades of academic research, there is an abundance of ODS tools that disclose the algorithms and processing techniques to the public via open source, so that Data Scientists can ensure the technique is appropriate to solving the problem at hand. Additionally, Data Scientists can leverage open source and improve it to suit their problems and environments. This flexibility makes it easier and faster for the Data Science team to deliver higher value solutions.

Transparency also feeds back into innovation, since anyone from the community can check the aptness of the analytics. It also greatly stimulates future contributions, allowing students in universities to view and engage with open source. This results in new ideas, research and innovation continuously progressing the ODS ecosystem.

While proprietary vendors are aware of the sea change in Data Science and are trying to embrace it, their attempts to do so are typically out of sync with the market needs and appear overhyped and disappointing. In short, traditional proprietary analytic platforms are not meeting the needs of modern Data Science teams that expect four fundamental principles: 1) Availability 2) Innovation 3) Interoperability and 4) Transparency. As more businesses try to rapidly unlock the value of their data in modern architectures, ODS becomes essential to their strategy.



MOVING TO OPEN DATA SCIENCE

Moving to any new technology has an impact on your team, IT infrastructure, development process and workload. Because of this, proper planning is essential. The drivers for change are different in every organization, so the speed and approach to the transition will also vary.

Team. Shifting to an ODS paradigm requires changes. Successful projects begin with people, and ODS is no different. New organizational structures — e.g., a center of excellence, lab teams or emerging technology teams — are a way to dedicate personnel to jumpstart the changes. These groups are typically chartered with actively seeking out new ODS technologies and determining the fit and value to the organization. This facilitates adoption of ODS and bridges the gap between traditional IT and lines of business. Additionally, roles may shift — e.g., from statistician to Data Scientist and from database administrator to Data Engineer — and new roles, such as Computational Scientist, will emerge.

With these changes, the team will need additional training to become proficient in ODS tools. While instructor-led training is still the norm, there are also many learning opportunities for ODS available online where the team can self-teach using ODS tools. With ODS, recruiting knowledgeable resources is much easier across disciplines — scientists, mathematicians, engineers, business and more — as open source is the de facto used in most universities worldwide. This results in a new generation of talent that can be brought onboard for Data Science projects.

Whether trained at university or on-the-job, the Data Science team needs the ability to integrate multiple tools into their workflow quickly and easily in order to be effective and highly productive. Most of the skills-ready university graduates are very familiar with collaborating with colleagues across geographies in their university experience. Many are also familiar with Notebooks, an ODS tool that facilitates the sharing of code, data, narratives and visualizations. This familiarity is critical because collaboration in Data Science is crucial to its success.

Research shows that the highest indicator of success for Data Scientists is curiosity. ODS satisfies their curiosity and makes them happy as they are constantly learning new and innovative ways to deliver Data Science solutions. Moving to ODS increases morale as Data Scientists get to build on the shoulders of giants who created the foundation for modern analytics. They feel empowered by being able to use their choice of tools, algorithms and compute environments to get the job done in a productive and impactful way that satisfies their natural curiosity and desire to make a meaningful change and impact with their work.

Technology. Selecting technology with ODS is significantly easier than proprietary software, because the software is freely available for download. This allows the Data Science team to self-serve their own proof of concept, trying out the ODS technology to meet the specific needs of their organization. For Data Science, there is no shortage of choices. Open source languages such as Python, R, Scala and Julia are frontrunners in ODS, and each of these languages offers many different open source libraries for data analysis, mathematics and data presentation, such as NumPy, SciPy, pandas, matplotlib and others, available at no cost and with open source licensing. No matter what your goals are in Data Science, there will be an open source project that meets your needs.

Some open source software only works effectively on local client machines, while other open source software supports scale out architectures, such as Hadoop. Typically, a commercial vendor fills the gap on supporting a wider variety of modern architectures.

Migration. The strategy for migrating to ODS is determined to align with the business objectives and risk tolerance of the organization. It is not necessary to commit to a full recoding to ODS from the start. There is a range of strategies from completely risk averse (do nothing) to higher risk (recode), each with their own pros and cons.

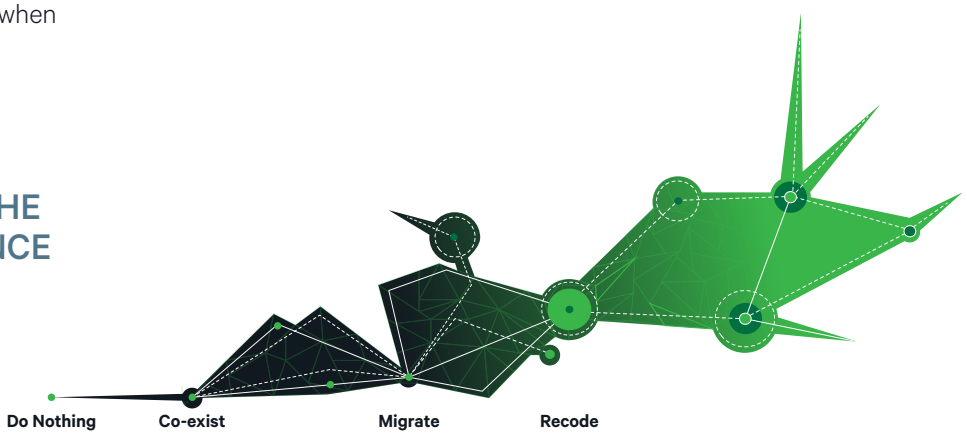
A coexistence strategy is fairly risk averse and allows the team to learn the new technology, typically on greenfield projects, while keeping the legacy proprietary technology in place. This minimizes disruption and, when the Data Science team is familiar and comfortable with the ODS tools, existing projects start to migrate to ODS when limits are reached with the proprietary technology. The proprietary technologies are then phased out over time.

A migration strategy is slightly riskier and moves existing solutions into ODS by reproducing the solution as-is with any and all limitations. This is often accomplished by outsourcing the migration to a knowledgeable third party who is proficient in the proprietary technology as well as ODS. A migration strategy can take place over time by targeting low-risk projects and limited scope until all existing Data Science code has been migrated to ODS. Migration strategies can also migrate all the legacy code via a “big bang” cutover. The Data Science solutions are improved to remove the legacy limitations over time, usually using a continuous integration continuous delivery (CICD) methodology.

A recoding strategy is higher risk and takes advantage of the entire modern analytics stack to reduce cost, streamline code efficiency, decrease maintenance and create higher impact business value often, through faster performance or from adding new data to drive better results and value. The objective of recoding is to remove limitations and constraints of legacy code by taking full advantage of ODS on modern compute infrastructure. With this strategy, oftentimes a full risk assessment is often completed to determine the prioritization of projects for recoding. The full risk assessment includes estimates for cost reduction and improved results to determine the risk.

The introduction of Big Data projects have become an ideal scenario for many companies to use a coexistence strategy - leaving legacy environments as-is and using ODS on Hadoop - for their Big Data projects.

MIGRATION STRATEGIES FOR THE JOURNEY TO OPEN DATA SCIENCE



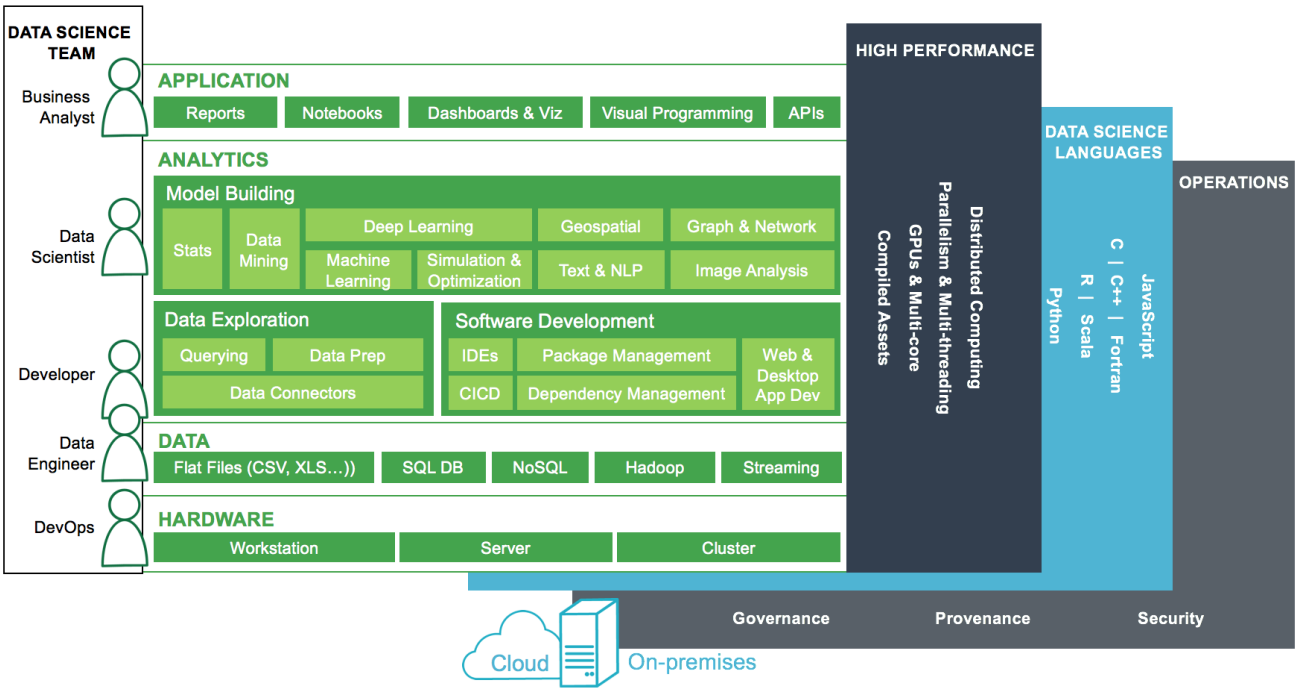
THE ANACONDA PLATFORM: OPEN DATA SCIENCE'S ONE-STOP SOLUTION

Anaconda is the leading modern open source analytics platform powered by Python. Anaconda enables modernization to Open Data Science as a platform that embraces the entire ODS ecosystem — Python, R, Java, Scala, Hadoop and more — across the entire modern analytics stack from desktop to server to clusters and cloud for enterprises. Anaconda makes it easy to:

- Create, collaborate and deploy Data Science solutions
- Leverage modern architectures and frameworks
- Set up and manage open source

Create, Collaborate and Deploy Data Science Solutions. Anaconda is an inclusive platform that brings all Data Science roles together to easily collaborate on high impact Data Science solutions. Data Science teams create models with their favorite languages and tools, including Jupyter Notebooks, IDEs, data exploration, data mining and Microsoft Excel, along with over 720 open source certified analytic Python and R packages. The Python and R packages are used for data prep, data mining, stats, machine learning, deep learning, simulation and optimization, text and natural language processing, geospatial, video/image/audio mining and graph and network optimization. Legacy analytics written in Fortran and C/C++ can also be leveraged into modern Data Science solutions with Anaconda. These solutions can be intelligent web apps or interactive visualizations, embedded into processes via RESTful APIs or dashboards for ad hoc analysis or production deployments. The breadth and depth of the Anaconda platform makes it feasible to move all your legacy analytics to ODS without making sacrifices.

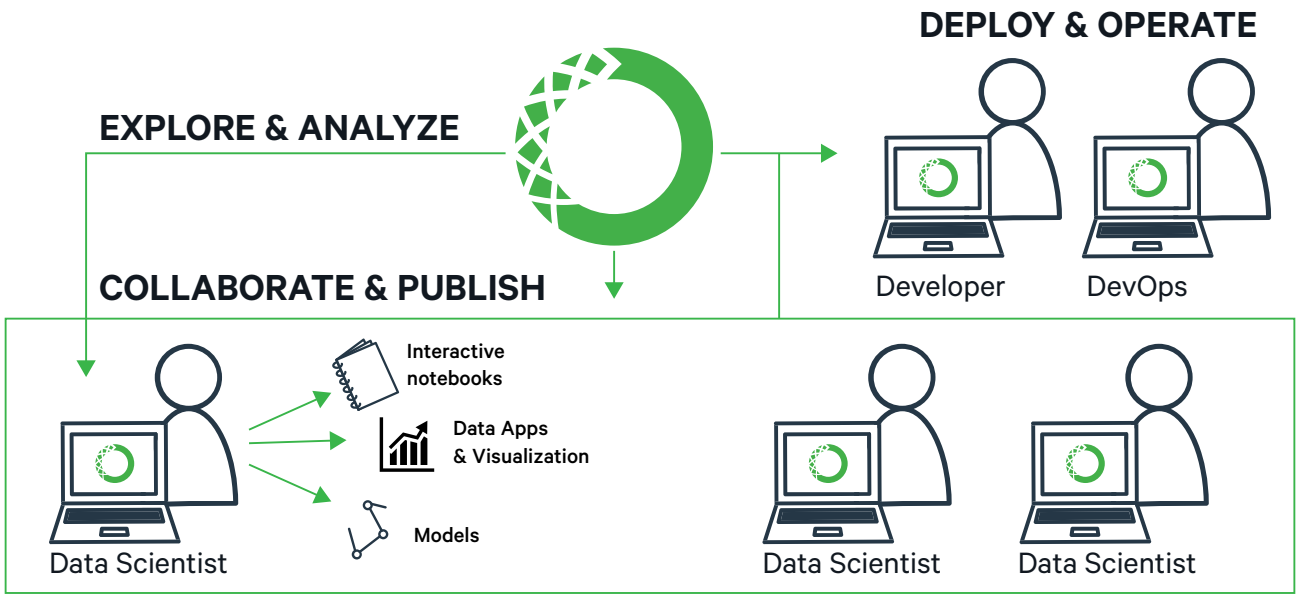
Leverage Modern Architectures and Frameworks. Anaconda is an enterprise-ready platform that delivers high performance, security and authentication to support the most demanding Data Science solutions. The compiler included in Anaconda delivers significant throughput that is comparable to the



performance of C while making it much easier to create and maintain the Data Science solutions. Anaconda exploits modern architectures for both scaling up and scaling out processing onto clusters and multi-core CPUs and GPUs to deliver cost effective Data Science solutions. Anaconda can co-exist alongside legacy and modern stacks, including Hadoop, Spark, Elasticsearch and others. Additionally, Anaconda can process inside Hadoop, eliminating data movement, easily parallelizing workloads to move computations to the data, and bypassing Hadoop overhead to read and write directly to HDFS files, all to deliver extremely high performance. These innovations allow migrated legacy analytics to achieve faster throughput while exploiting new data.

Setup and Manage Open Source. Anaconda includes conda, an open source, multilanguage (e.g. Python, R, Java, Fortran, C/C++), cross-platform (e.g. Windows, Linux, OS X) package, dependency and environment manager that makes open source convenient and feasible for enterprises.

Conda promotes innovation and collaboration by allowing teams to partake in the latest certified open source packages and makes installation a breeze. This supports innovation across the development cycle, as new packages or latest versions are needed. This same capability allows the Data Science team to package their own innovations to share with other teams. This fast and simple way to reproduce environments makes it easy to migrate Data Science solutions across the development, user acceptance testing (UAT) and production systems. With Anaconda, your Data Science solutions are effortlessly deployed into production. While migrating to open source may seem difficult and laborious, Anaconda makes it stress-free to setup and deploy ODS solutions.



SUMMARY

Open Data Science is the foundation to modernizing your data analytics. With Anaconda, you can now:

- Leverage the full power and innovation of open source for all your Data Science needs
- Collaborate in a secure fashion with teams across the globe
- Integrate with legacy analytics to maximize your investments
- Reduce costs while experiencing new freedom to select the best tool and architecture that fits your enterprise now and in the future

While moving to ODS can seem intimidating, Anaconda delivers an enterprise platform that makes the journey to ODS simple and painless.

In an Open Data Science world, Anaconda is your key to unlocking the value in your data.

ABOUT CONTINUUM ANALYTICS

Continuum Analytics is the creator and driving force behind Anaconda, the leading modern open source analytics platform powered by Python. We put superpowers into the hands of people who are changing the world.

With more than 2M downloads annually and growing, Anaconda is trusted by the world's leading businesses across industries – financial services, government, health & life sciences, technology, retail & CPG, oil & gas – to solve the world's most challenging problems. Anaconda does this by helping everyone in the Data Science team discover, analyze and collaborate by connecting their curiosity and experience with data. With Anaconda, teams manage their Open Data Science environments without any hassles to harness the power of the latest open source analytic and technology innovations.

Our community loves Anaconda because it empowers the entire Data Science team – Data Scientists, Developers, DevOps, Architects, and Business Analysts – to connect the dots in their data and accelerate the time-to-value that is required in today's world. To ensure our customers are successful, we offer comprehensive support, training and professional services.

Continuum Analytics' Founders and Developers have created or contribute to some of the most popular Open Data Science technologies, including NumPy, SciPy, Matplotlib, Pandas, Jupyter/IPython, Bokeh, Numba and many others. Continuum Analytics is venture-backed by General Catalyst and BuildGroup.

To learn more, visit http://www.continuum.io