ORACLE
Data Cloud

Scaling Data Science

A look at the tools, processes, and infrastructure your business needs to derive value from data.

WHITE PAPER / UPDATED JUNE 2019

ORACLE

INTRODUCTION

Though it has deep roots in academia, data science is now very much a business process. Just like accounting or marketing, there is a cost to doing data science—as well as great benefits that come from doing it well. In fact, 21 percent of executives report that investing in big data has been transformative for their firms.[1]

Scaling data science across the entirety of an organization requires extensive resources, a defined roadmap, and the right tools, processes, and infrastructure to put work into production.

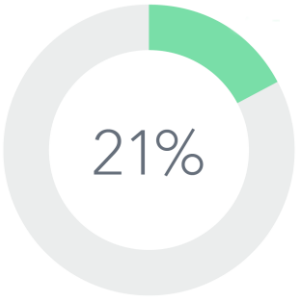
For businesses unable to realize transformative improvements from data-driven work, the roadblocks are varied: organizational impediments, lack of alignment between stakeholders, resistance to technology changes, and more can hold up progress.

Scaling data science across the entirety of an organization requires extensive resources and a defined roadmap, as well as the right tools, processes, infrastructure, and a plan for putting work into production.

WHAT DOES IT MEAN TO SCALE DATA SCIENCE?

The term *scaling data science* refers to making your data science team the engine that powers every decision with powerful predictive models and comprehensive analyses. For instance, a predictive customer churn model built by a data scientist could forecast the likelihood that individual customers will stop buying from your business in a given time period. That information could help you identify high-risk customers and intervene before they defect—but only if that information gets to the appropriate parties.

If you have the right processes in place, the outputs of that model can be integrated with your call center software so customer service reps will be able to view risk scores when they're on the phone with customers. Or they might be delivered to your marketing automation system enabling your marketers to create more targeted campaigns. This is just one example of how to scale your data science efforts. To make data science worth the investment, your company not only needs data scientists to do the high-value work in the first place, but the tools, processes, and infrastructure to support it.



21 percent of executives report that big data has been transformative for their firms.

1. NewVantage Partners, "Big Data Executive Survey 2017," January 2017.

1. TOOLS AND PROCESSES

Identifying the tools you need to scale is both essential and complex. No two data scientists possess the same skillsets or tool preferences, and tool sprawl—in which the volume of tools being used exceeds an organization's ability to effectively utilize them—is the number one problem data-driven companies face.[2] The following table provides examples of the wide variety of available tools, but many more exist and new ones enter the market frequently.

The typical enterprise data science project requires dozens of steps—from cleaning data and selecting model features to model validation and deployment. There are dozens of tools designed to cater to each part of the process.

The Data Science Technology Landscape

TECHNOLOGY	TOOLS
Programming languages	Python, R, Scala
Machine learning frameworks	Scikit-learn, R, Mllib, H2O, Turi
Data processing and compute resources	SQL, Spark, Hive, Pig, Cascading
Code editors and notebooks	Jupyter, Zeppelin, RStudio, Eclipse, IntelliJ
Data visualization libraries	Matplotlib, ggplot, Seaborn, reflect.io, D3
Package management	PyPI, CRAN, Maven
Build tools	pip, packrat, sbt
Data pipelines	Airflow, Luigi, Pinball, ML Pipelines
Model serialization	PMML, PFA, Parquet, JSON, Pickle
Model deployment	Palladium, Prediction.io, Oryx.io, Docker
Cloud computing	AWS, Google Cloud, Azure, IBM Bluemix

The typical enterprise data science project requires dozens of steps—from cleaning data and selecting model features to model validation and deployment. There are dozens of tools designed to cater to each part of the process. The tools and processes you have in place should support your data scientists' ability to:

Your tools and processes must support data scientists ability to

- Experiment
- Create work that is reproducible
- Collaborate and share across teams

EXPERIMENT

Not every data model is going to function perfectly out of the gate. Experimenting and iterating is part of a typical data science process. Therefore, the tools your team uses should support the deployment of different model versions and compile metrics to measure the success of those versions. Experimentation also requires the collection and storage of massive amounts of data. The more data your data scientists have, the more opportunities for analysis they can uncover.

2. Forrester Consulting, "Data Science Platforms Help Companies Turn Data Into Business Value," December 2016.

CREATE WORK THAT IS REPRODUCIBLE

If your data scientists are reinventing the wheel for every project, you're wasting valuable time and resources. Because data scientists often store their work locally, much of what has already been done goes unshared. Having a central location for files, models, and code will help your team find and reuse data science work that has already been battle tested.

COLLABORATE AND SHARE ACROSS TEAMS

Your processes and tools should encourage knowledge sharing, especially between technical and non-technical teams. Make sure it's easy for data scientists to publish analyses as shareable reports and integrate model results into the dashboards or real-time applications that stakeholders rely on.

ACCOMMODATE THE TOOLS YOUR DATA SCIENTISTS USE EVERY DAY

When it comes to mitigating tool sprawl, you'll find it's just not possible to dictate every tool your data science team uses. Data scientists embrace programming languages such as R, Python, SQL, Ruby, and Scala, and rely on integrated development environments (IDEs) and notebooks such as Apache Zeppelin, RStudio, and Jupyter. Rather than require your team to code in a specific language and environment, you can provide a data science platform that runs code written in any language in the notebooks or IDEs your data scientists prefer. That way, your team can work how they want and still share work, driving reproducibility and collaboration.

2. INFRASTRUCTURE AND ENVIRONMENTS

The infrastructure needed to support complex data analyses is a critical—if often unmentioned—component of data science. IT can spend a lot of time spinning up the necessary resources to support data science work or setting up environments with the right packages and then waiting for those packages to compile.

But with the right approach, your data science team can get the resources they need without costing you—and your IT team—an excess of money and time. Here are just a few elements to consider:

As you build your data science infrastructure, consider these elements

- On-demand computing resources
- Standardized data science environments
- Access control

ON-DEMAND COMPUTING RESOURCES

To reduce costs, companies are increasingly turning to on-demand computing—enabling compute resources to be spun up as needed. Working in a cloud environment is a great way to accommodate data science work in a way that doesn't require your computing resources to be always on.

STANDARDIZED DATA SCIENCE ENVIRONMENTS

Don't set up a new data science environment for every project. Containers such as Docker make it easy to download standard, repeatable environments with the tools and packages you need already installed.

ACCESS CONTROL

Two and a half quintillion bytes of data are created every day, and much of it comes from customer interactions with your business. It's unlikely that every member of your team needs access to every type of data or every analysis. The ability to delegate access by role and protect sensitive data is an important one.

GIVE YOUR IT TEAM THE CONTROL IT NEEDS

Scaling data science isn't just about giving your data scientists the tools they need. IT is an integral part of getting the resources your data science team needs up and running, and your IT department will likely be concerned with stability and data security. Giving them the ability to manage your pool of servers, connect data science tools reliably and securely to data sources, and grant or deny access to users in a few clicks is essential. A data science platform with an IT admin dashboard can make this process easier.

3. PRODUCTION

Getting data science work into production is arguably the most important step in any data science process. Until this point, you haven't truly scaled data science at your organization.

Putting data science work into production means building a pipeline in which data science analyses and models are continuously running in real time to power your business. Deploying a model—such as a recommendation engine—into production so that it can suggest products to shoppers on your site is just one part of the equation. That model will need to be retrained, A/B tested, and constantly fed new data. Making this work seamlessly is no small undertaking. The process should allow for:

The process for putting your data science system into production should allow for

- Automation of repetitive tasks
- Ongoing performance monitoring
- Constant improvement

AUTOMATION OF REPETITIVE TASKS

If your team is manually running the reports your stakeholders want regularly, cleaning data, and retraining models, you're not efficiently putting work into production. Currently, three out of every five data scientists spend the majority of their time cleaning and organizing data.[3] Setting up a system that automates many of these low-level tasks will give your data scientists more time to focus on building high-value analyses.

ONGOING MONITORING OF MODEL PERFORMANCE

Collecting data from your models to monitor their performance is essential for identifying issues and addressing them. Set up a pipeline to deliver relevant data from activities such as model API calls, training, and cross validation.

CONSTANT IMPROVEMENT

A data scientist's work is never done. Giving your data scientists the ability to deploy different versions of their predictive models to compare and iterate upon is a great way to constantly improve the results of your data science work.

TAKE THE BURDEN OF MODEL DEPLOYMENT OFF OF ENGINEERING

Rather than asking your data scientists to hand off their models to the engineering team for refactoring, testing, and deployment, consider providing a tool or platform that allows them to deploy models as APIs. With that capability, the models your data scientists build can be integrated anywhere instantly, like the dashboards and real-time applications your decision makers use — ultimately saving time and resources while providing real value.

3. CrowdFlower, "2016 Data Science Report," 2016.

HOW A DATA SCIENCE PLATFORM BRINGS IT ALL TOGETHER

Data science platforms are a relatively new technology that are becoming a must have for enterprise data science teams. In fact, platform adoption is expected to rise from 26 percent to 69 percent over the next two years as companies increasingly recognize the value of managing data science tools and processes in a centralized hub.[4]

In a nutshell, a data science platform is a software hub around which all data science work takes place. That work usually includes integrating and exploring data from various sources, coding and building models that leverage that data, deploying those models into production, and serving up results, whether that's through model-powered applications or reports. Platforms are designed to support this work to make scaling data science much more achievable.

Are you ready to start performing data science at scale? Oracle Cloud Infrastructure Data Science is a platform that provides integrations with the tools your data scientists already use and love—such as Jupyter notebooks and GitHub—intuitive project organization, easy report publishing, model deployment capabilities, and much more, backed by enterprise-grade security features and infrastructure.

A data science platform is a software hub around which all data science work takes place. That work usually includes

- Integrating and exploring data from various sources
- Coding and building models that leverage that data
- Deploying those models into production
- Delivering results through model-powered applications or reports

4. Forrester Consulting, "Data Science Platforms Help Companies Turn Data Into Business Value," December 2016.

ORACLE CORPORATION

Worldwide Headquarters
500 Oracle Parkway, Redwood Shores, CA 94065 USA

Worldwide Inquiries
TELE + 1.650.506.7000 + 1.800.ORACLE1
FAX + 1.650.506.7200
oracle.com

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com. Outside North America, find your local office at oracle.com/contact.

datascience.com/blog facebook.com/oracle twitter.com/oracle

Integrated Cloud Applications & Platform Services

White Paper / Scaling Data ScienceScaling data science across an organization requires extensive resources and a defined roadmap, as well as the right tools, processes, infrastructure, and a plan for putting work into productionScaling Data ScienceScaling Data Science
Updated May 2019