February 13, 2019

CASE STUDY: DOMINO DATA LAB AND THE CLIMATE CORPORATION

Climate Corp. Scales Up Data Science to Power Precision Agriculture

Platform Speeds Modeling for Farmer-Specific Recommendations, from What to Plant to When to Harvest

Doug Henschen

Vice President and Principal Analyst

Copy Editor: Jim Donahue

Layout Editor: Aubrey Coggins

TABLE OF CONTENTS

AT A GLANCE

The Climate Corporation (Climate) describes itself as "innovating at the intersection of agriculture and technology," with a mission to "help all the world's farmers sustainably increase their productivity with digital tools." The Climate FieldView™ (FieldView) platform helps farmers manage their crops using data on their acreage, soil type, elevation, average precipitation, crop and yield histories and other measures.

Climate's data science team builds models based on this data and each farmer's goals to deliver personalized recommendations on what seeds and hybrids to plant, how to plant them, how to care for the crops throughout the growing season and when to harvest to maximize yields.

The Climate Corporation

- **Headquarters:** San Francisco
- **2017 Revenue:** Climate's revenues are not disclosed. The company is a subsidiary of Bayer AG, which had 2017 revenues of €35.0 billion.
- **No. of Employees:** 600 at Climate
- **Industry:** Provides data-driven recommendations to help farmers maximize their return on every acre.
- **Founded:** 2006 as weather insurance company WeatherBill; repositioned as Climate in 2013.

Business Themes


Data to Decisions


Technology Optimization

AT A GLANCE

Problems	· Climate's homegrown system for spinning up data science environments on Amazon Web Services was hard to customize. · The homegrown system also lacked collaborative features and documentation of workloads needed to repeat experiments. · Data engineers struggled with inconsistent or, worse, sparse documentation when bringing models into production.
Solutions	· Climate deployed Domino Data Science Platform in 2016 as the firm's standards-based environment for model development. · The platform supports collaboration among Climate's team of approximately 200 data scientists and data engineers who bring models into production.
Benefits	· Domino has raised the productivity of Climate's data science team, driving a 100 percent annual increase in new models delivered. · The platform has standardized development and promotes collaboration to ease and speed repeatable data science. · The platform facilitated the 2018 launch of Seed Advisor, a new predictive seed selection and placement service.

The company has been practicing deep weather and agricultural data science for more than a decade. Two former Google employees founded it as a weather insurance company in 2006, and Climate pivoted to its agricultural-advisory focus in 2013. Soon after the launch of FieldView in 2015, Climate recognized that it needed to scale up its data-modeling capabilities to support and sustain more recommendations.

This report details Climate's deployment of Domino Data Lab's Domino Data Science Platform to speed, simplify and standardize the iterative process of model development. The collaborative platform supports Climate's team of approximately 200 data scientists and data engineers, and it has helped the company double its capacity to build models based on machine learning, deep learning and other statistical and advanced analytic methods.

THE COMPANY

Climate helps farmers around the world improve their operations through data-driven recommendations. Modern farmers are increasingly sophisticated, and they're embracing cutting-edge technology to maximize yield. The company's Climate FieldView product, introduced in 2015, combines agronomic modeling and high-resolution weather simulation to deliver precision agricultural recommendations on when and what to plant, when to fertilize, how to combat disease and when to harvest.

Climate FieldView runs on data that powers predictive models. The data includes geospatial, soil-type and elevation profiling of each farmer's land. To this collection, the FieldView platform can add historical data, including information from its tractor-mounted FieldView Drive. The Drive integrates with an array of sensors aboard the tractor as well as on attached equipment, including planters, liquid applicators and combines. For planting, FieldView can capture sensor-based measurements, including the speed and direction of the planter, the downforce used to plant seeds, how deeply seeds are planted, the space between seeds and rows, and whether seeds were skipped. It captures this data every second.

As per-customer data accumulates by the terabyte, Climate draws on in-season and season-to-season history to drive increasingly accurate models and predictions specific to each farmer. Even when new customers lack historical data, they can benefit from recommendations based on aggregated data and government and third-party data.

FieldView is used by farmers to help manage more than 60 million acres across the United States, Canada, Brazil and Europe. Climate intends to extend that footprint both in current markets and in new regions around the globe. But FieldView's success has presented challenges: With each new customer, data source, plant hybrid and type of recommendation, Climate must develop and maintain yet more models. By 2016, Climate recognized that it needed to scale up its model-development environment.

"We built a lot of custom infrastructure to spin up and tear down model-development instances on Amazon Web Services, but we didn't have a lot of insight into what was being done [in the model-development phase] on the cloud," explains Erich Hochmuth, director of engineering, data analytics group, at Climate. "We also weren't able to respond quickly enough to scientists who had requests for specific types of hardware capacity."

This case study explores how Climate more than doubled its model-development capacity and gained flexibility without adding significantly to its team of data scientists.

THE CHALLENGES

By 2016, Climate already managed data by the petabyte, storing everything in Amazon Simple Storage Service (S3) and using Amazon Web Services (AWS) Spark and Yarn services on EC2 compute infrastructure to process the data and run analyses. Climate had already moved its data scientists away from doing desktop analyses to spinning up development instances on AWS, but the homegrown infrastructure the company built to support cloud-based modeling work had limitations:

- Climate lacked insight into the number, type, duration, cost and other workload parameters associated with each cloud-based job and project.
- Data engineers had trouble quickly responding to data scientist requests for special hardware capacity, such as graphical processing unit (GPU) instances, because nonstandard compute options weren't supported.
- When jobs failed, data scientists (also known internally at Climate as researchers) would lose their analyses before they had a chance to extract vital workload information from the AWS Spark/S3 environment.

- Available metadata and documentation on the data, languages, dependencies and requirements associated with each model varied, often making it difficult for data scientists to repeat analyses and for engineers to bring models into production.

THE SOLUTION

With data and modeling demands continuing to grow at Climate, the challenges detailed above became more pressing. Climate initially considered upgrading its homegrown solution, but in 2016 it began a search for available commercial software options.

"It really wasn't worth our time to build and maintain a custom solution when [commercial vendors] were doing it," says Hochmuth.

After a brief search, Climate quickly settled on the Domino Data Science Platform from Domino Data Lab, attracted by the following features and capabilities:

- Uses configurable Docker containers to create and document data science environments that can be shared and versioned for reuse and repeatability.
- Supports discovery and reuse of data sources, including databases and distributed platforms deployed on-premises or in virtual or public-cloud environments.

The Technologies

- Amazon Web Services offerings, including Amazon S3, a variety of EC2 compute instances, Spark and Yarn services, and Kinesis Analytics.
- Domino Data Science Platform supports modeling work at scale through automation, enhancing self-service and productivity, and through documentation of work-loads, promoting collaboration and repeatable data science.
- Python and R support a range of data science modeling approaches, while QGIS, an open source geographic information system, is used for geospatial analysis.

- Documents jobs and resources used to provide insight into data and languages used, dependencies and costs by project, model, user and other parameters.
- Eases collaboration among data scientists, data engineers and business users.

Domino Data Science Platform was deployed within a matter of two days in Climate's AWS environment with configuration and management support from Domino Data Lab. Climate tailored Domino to make it easy for data scientists to securely access Climate's data on S3 and to deploy a variety of preconfigured development environments instantiated in Docker containers.

"Domino is now the standard development environment for all of our data scientists, and it also supports collaboration with our data engineers and business stakeholders," says Hochmuth. Climate has roughly 200 data scientists and data engineers, and the Domino platform now typically sustains 75 active users each day.

"It's easy for the scientists to sit down with engineers and share their projects in Domino," says Hochmuth. "The engineers can see all the requirements and the work that has been done, and they can fork executing projects and run them for themselves."

Models built in Domino aren't called directly from Climate's production environment, but the company is considering changes to that process. It currently uses an internally developed deployment service architecture for model deployment, but it is investigating other options for production deployment. The standardization and documentation supported by Domino has made it easier for data engineers to export models and expose them through its customer-facing production environment on AWS.

"Because we tailor the Docker containers that data scientists use within Domino, we're then able to align that environment with our production services; the containerization and configuration is standardized," says Hochmuth.

THE IMPACT

The benefits of the Domino deployment start with the data scientists, but they extend to the data engineering team, business stakeholders and Climate FieldView customers. Where it used to take Climate data engineers anywhere from one day to one week to set up special-request development environments, Domino has made it a self-service proposition for data scientists.

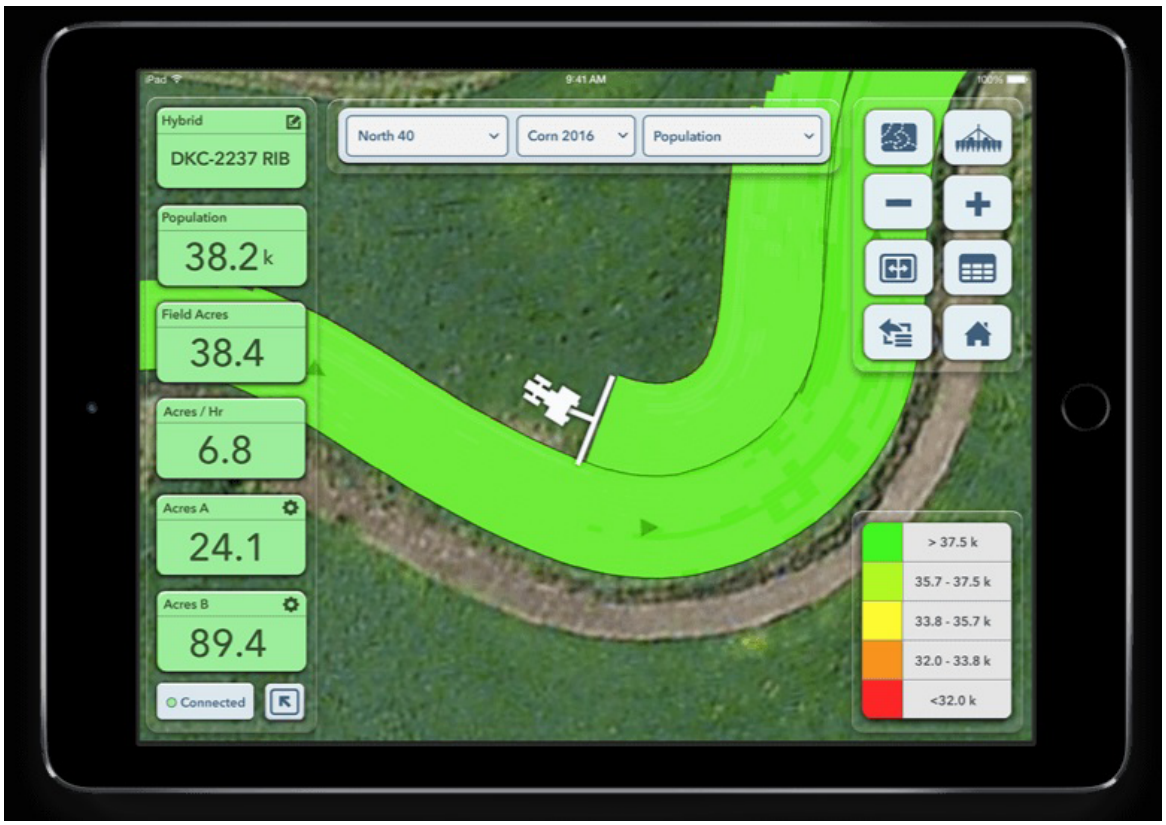
"If scientists want to try running a memory-bound model or a CPU-bound model on different hardware, now it takes just a matter of minutes to reconfigure, and scientists can do it on their own," says Hochmuth.

The platform has also set standards around code management, data management and cloud instances that have brought significant productivity gains. Climate estimates that its data science team has doubled its annual model output since Domino was deployed.

Climate's productivity gains have been driven in large part by Domino's support for collaboration, repeatability and reuse. "It's so much easier to share models and see what everybody is working on," says Hochmuth. "In our homegrown environment, it was up to the data scientist to figure out what artifacts and data they needed to save, but in Domino, that's all built into the platform. You can select any model run or experiment and see what version of the model and what data were used, the dependencies and what packages of software were used."

Climate's scaled-up modeling capabilities are driving new services as part of the FieldView platform. For example, Seed Advisor (see Figure 1), a new predictive seed selection and placement technology for corn, was introduced last August. The service provides recommendations on what seeds to plant in specific fields, how many seeds to plant per acre, the exact row spacing to use and how deep to plant the seeds. A Bluetooth-synchronized FieldView mobile application can even carry out these recommendations, instructing sensor-equipped tractors and attached equipment exactly how to move across specific fields and plant the crop.

Figure 1. Seed Advisor Recommendations on Planting Are Carried Out in Automated Fashion by Tractors Through the FieldView Mobile App



Source: Climate

According to Climate, tests of Seed Advisor conducted across 100,000 acres in Iowa, Illinois and Minnesota during the 2018 growing season demonstrated an average yield advantage of 9.1 bushels per acre compared with what the farmer would have planted without Climate's recommendations.

Predictive research is already underway to extend Seed Advisor to soybeans, and Climate has plenty of other work to do that will keep scaling up modeling demands. Climate has extensive data and physical models on how specific hybrids grow in specific climates and soil types and their nitrogen-uptake characteristics. These algorithms drive fertility prescriptions that help farmers optimize the timing and rates of fertilizer application. Disease risk and identification models based on three years of data successfully forecasted the occurrence of disease more than 80 percent of the time in 2018, giving farmers forewarning and recommendations on whether and when to apply fungicides.

Of course, once a new recommendation feature is added to Climate FieldView, data scientists face the ongoing task of keeping the underlying models optimized. "We have a continuous cycle of data collection and analysis because we're always using new data to retrain our models," explains Hochmuth.

THE TAKEAWAYS

Climate FieldView is a great example of the types of breakthrough, data-driven services that are emerging, thanks to the availability of high-scale data and flexible, rapidly scalable cloud computing capacity. Climate is an insights-as-a-service business, so its product is deep analytics. There are business users within Climate that market and sell its services, but this is not a case study about enabling internal business users with self-service analytics. Predictive recommendations are the very foundation of Climate's business, and the "business users" are Climate's farmer customers—a sophisticated and demanding group.

"We do extensive A-B testing before we bring models into production, and we don't use champion/challenger approaches because we can't just swap models because customers are very sensitive to changes in outputs," Hochmuth explains. "Farmers are highly educated on the physical aspects of their plants, how much nitrogen they need and how quickly they should be growing. If you're providing a recommendation and it suddenly changes, they would want to know specifically what additional data became available that changed the recommendation."

> *"Most data scientists spend 80 percent of their time figuring out where data exists, how to bring it together, how to set up data infrastructure and how to execute an analysis.... Now 20 percent of our data scientists' time might involve particulars around data and execution and 80 percent of their time is focused on model development."*
>
> —Erich Hochmuth, Director of Engineering, Data Analytics Group, Climate

Given the scarcity and cost of data scientists, productivity was the key deliverable Climate needed to not only sustain FieldView but to support new recommendations, new crops and new geographic markets.

"Most data scientists spend 80 percent of their time figuring out where data exists, how to bring it together, how to set up data infrastructure and how to execute an analysis," says Hochmuth. "For somebody with a Ph.D., that's not the best use of their time. Using Domino, now 20 percent of data scientists' time might involve particulars around data and execution and 80 percent of their time is focused on doing model development."

For companies using traditional methods, obstacles to data modeling speed and scale include iterative, manual work steps; data-movement and data-sampling requirements; waterfall development approaches; and on-premises storage constraints. With such constraints, cycles of development drag on for weeks if not months.

The benefits of standardization, automation and documentation supported by Domino have streamlined Climate's model-development workflow and doubled data scientist productivity. The platform gives data scientists self-service control over their development environments while easing collaboration with data engineers for ongoing optimization work and the handoff of models for production deployment.

THE RECOMMENDATIONS

- **Take a team approach.** Organizations that are sustaining model development and deployment at scale have the right people, processes and technologies in place and have evolved to a team approach. Industry leaders have mastered collaboration across large teams and promote shared responsibility for successful delivery and ongoing execution.

- **Look for a supporting platform.** Leaders are scaling up with the aid of technology that supports standardization and automation. Move on from uncoordinated, ad hoc experimentation. Look for a collaborative platform that will promote consistent standards and repeatable data science.

- **Ensure trust and transparency.** Machine-learning- and deep-learning-based recommendations can't be black boxes. That's a matter of regulation in the banking and insurance industries, where certain types of predictive decisions must be explainable and proven to be unbiased. Even when it's not a matter of regulation, customers and employees in any industry will more readily trust recommendations that are transparent, explainable and consistent.

  "We're very conscious of explaining to the customers what models and data are used," says Hochmuth. "The FieldView user interface was designed to provide descriptive statistics about the data and the features that were fed into the model. From a development perspective, we give customers descriptive statistics first and then add on machine-learning and statistical modeling capabilities."

- **Focus on business outcomes.** Stay focused on what predictive capabilities can do for your business rather than getting too wrapped up in the component technologies. Constellation discusses AI and underlying machine learning and deep learning capabilities in the context of seven outcomes:

  1. **Perception** of what's happening now
  2. **Notification** of defined conditions, through alerts, workflows or reminders
  3. **Recommendation** of suggested actions, based on past history, behaviors and outcomes
  4. **Automation** of desired actions based on defined business conditions
  5. **Prediction** of what to expect from people or assets based on detected behaviors
  6. **Prevention** of bad outcomes, through notifications or automated actions
  7. **Situational** awareness, meaning holistic understanding of what's happening now and which recommendations, predictions, automation steps and preventive steps must be acted upon right now

By focusing on the spectrum of outcomes above, organizations are more likely to deliver transformative, differentiated and disruptive capabilities.

RELATED RESEARCH

For related research, see: Doug Henschen, "AI Imperative: Advance from Experimentation to Deployment at Scale," Constellation Research, August 29, 2018. *https://www.constellationr.com/research/ai-imperative-advance-experimentation-deployment-scale*

Doug Henschen

Vice President and Principal Analyst

Doug Henschen is Vice President and Principal Analyst at Constellation Research, Inc., focusing on data-driven decision making. His Data-to-Decisions research examines how organizations employ data analysis to reimagine their business models and gain a deeper understanding of their customers. Data insights also figure into tech optimization and innovation in human-to-machine and machine-to-machine business processes in manufacturing, retailing and services industries.

Henschen's research acknowledges the fact that innovative applications of data analysis require a multi-disciplinary approach, starting with information and orchestration technologies, continuing through business intelligence, data visualization, and analytics, and moving into NoSQL and big data analysis, third-party data enrichment, and decision management technologies. Insight-driven business models and innovations are of interest to the entire C-suite.

Previously, Henschen led analytics, big data, business intelligence, optimization, and smart applications research and news coverage at *InformationWeek*. His experiences include leadership in analytics, business intelligence, database, data warehousing, and decision-support research and analysis for *Intelligent Enterprise*. Further, Henschen led business process management and enterprise content management research and analysis at *Transform* magazine. At *DM News*, he led the coverage of database marketing and digital marketing trends and news.

@DHenschen **constellationr.com/users/doug-henschen** **linkedin.com/in/doughenschen**

ABOUT CONSTELLATION RESEARCH

Constellation Research is an award-winning, Silicon Valley-based research and advisory firm that helps organizations navigate the challenges of digital disruption through business models transformation and the judicious application of disruptive technologies. Unlike the legacy analyst firms, Constellation Research is disrupting how research is accessed, what topics are covered and how clients can partner with a research firm to achieve success. Over 350 clients have joined from an ecosystem of buyers, partners, solution providers, C-suite, boards of directors and vendor clients. Our mission is to identify, validate and share insights with our clients.

Organizational Highlights

- Named Institute of Industry Analyst Relations (IIAR) New Analyst Firm of the Year in 2011 and #1 Independent Analyst Firm for 2014 and 2015.
- Experienced research team with an average of 25 years of practitioner, management and industry experience.
- Organizers of the Constellation Connected Enterprise—an innovation summit and best practices knowledge-sharing retreat for business leaders.
- Founders of Constellation Executive Network, a membership organization for digital leaders seeking to learn from market leaders and fast followers.

www.ConstellationR.com

@ConstellationR

info@ConstellationR.com

sales@ConstellationR.com

San Francisco Bay Area | Boston | Colorado Springs | Ft . Lauderdale | Los Angeles | New York Metro
Northern Virginia | Portland | Pune | Sacramento | San Diego | Sydney | Toronto | Washington, D.C.