Data Hackathons: Jumpstarting Your Test Organization’s Digital Transformation

|  |  |  |
| --- | --- | --- |
| Nathan Cook | Capt Troy Soileau | Maj Matthew McCormack |
| 96th Operations Group | 96th Cyberspace Group | USAF Test Pilot School |
|  | Air Force Test Center Department of the Air Force |  |

# 1. Abstract

The United States Department of the Air Force’s (DAF’s) Air Force Test Center (AFTC) completed its third iteration of the AFTC Data Hackathon in August of 2022. “Hackathons” are low risk, high return investments that can jumpstart your test organization’s digital transformation. Hackathons for software engineering have existed for more than two decades in the disciplines of cryptography, web development, and apps. With the advent of data science, big data, machine learning, and artificial intelligence, “Data Hackathons” bring the hackathon model to these data-centric disciplines. A Data Hackathon allows your organization to: explore data infrastructure options, expose “data hackers” to your organization’s test and management data, evolve third-party and in-house scripts and apps to solve real-world problems, and expand awareness of the state of the art digital technologies within your organization. The future of test will require ever increasing data volume, variety, and velocity. The pace of improvement in tools and techniques will continue to accelerate. Data Hackathons can focus your test teams and provide momentum for your organization’s digital transformation.

# 2. Acronyms, Abbreviations, Symbols

|  |  |
| --- | --- |
| AFTC | Air Force Test Center |
| API | Application Programming Interface |
| AWS | Amazon Web Services |
| CDAO | Chief Data and Artificial Intelligence Office |
| DAF | United States Department of the Air Force |
| DoD | Department of Defense |
| EC2 | Elastic Compute Cloud |
| GFE | Government Furnished Equipment |
| OEM | Original Equipment Manufacturer |
| S3 | Simple Storage Service |
| US | United States |
| VPN | Virtual Private Networks |

# 3. Disclaimer

Any mention of specific vendors, platforms, or tools is factual history of their use, and in no way an endorsement by the authors, the Department of the Air Force, the United States Government, or a sponsorship of the vendors, platforms, or tools themselves.

# 4. Introduction

The United States Department of the Air Force’s (DAF’s) Air Force Test Center (AFTC) completed its third iteration of the AFTC Data Hackathon in August of 2022 (“Data Hackathon” n.d.; Cloys 2022). “Hackathons” are low risk, high return investments that can jumpstart your test organization’s digital transformation. Hackathons for software engineering have existed for more than two decades in the disciplines of cryptography, web development, and apps. With the advent of data science, big data, machine learning, and artificial intelligence, “Data Hackathons” bring the hackathon model to these data-centric disciplines.

This paper will explain the benefits of implementing a Data Hackathon in your flight test organization. For each benefit, the approach of the AFTC Data Hackathon will be summarized. In addition, a framework for planning, provisioning, executing, and following up on a Data Hackathon is provided, which can then be modified as required to meet the opportunities and constraints of your organization.

# 5. Benefits

A Data Hackathon allows your organization to:

* explore data infrastructure options,
* expose “data hackers” to your organization’s test and management data,
* evolve third-party and in-house scripts and apps to solve real-world problems, and
* expand awareness of the state of the art digital technologies within your organization.

## 5.1 Explore

**Explore data infrastructure options.** Data infrastructure is a prerequisite for working with data. Monica Rogati, among others, made a comparison with Maslow’s hierarchy of needs (Maslow 1943) and created “The Data Science Hierarchy of Needs” (Rogati 2017), depicted as a triangle or pyramid, [Figure 1](#fig-dshierarchy). Collection, movement, and storage are at the base of the pyramid. This representation is chronological, following the flow of data from source to working location, followed by techniques to explore, understand, and organize, before generating algorithms and models for inference, prediction, and influencing decisions.

|  |
| --- |
| Figure 1: The Data Science Hierarchy of Needs |

The exact sequence, though, is not essential for the bottom two layers. The basic components of infrastructure are:

* Store (sometimes called “data at rest”)
* Compute
* Transport (sometimes called “data in motion”)

A Data Hackathon directly stresses all three components, by design. Your organization may have been considering adding or modifying aspects of each infrastructure component. The hackathon is an opportunity to put these updates in action and uncover pain points that would remain invisible without direct application of focused effort. The hacker teams may use the components in ways that were not anticipated by the originators, as well.

### AFTC Data Hackathon Infrastructure

The AFTC Data Hackathon made use of infrastructure already existing and available to DAF members, but which may not be used to their full potential. To date, the primary platforms used, in chronological order:

* DAF CDAO “VAULT” providing access to:
  + Amazon Web Services (AWS) Elastic Compute Cloud (EC2) compute and Simple Storage Service (S3) store in the cloud
  + Databricks notebooks running Python and R orchestrated with Apache Spark
  + Hue SQL assistant
  + Apache Zeppelin notebooks running Python orchestrated with Apache Spark
  + Plotly Dash dashboard front ends with Python visualizations
* Microsoft 365 including:
  + Microsoft Azure compute and store in the cloud
  + Excel
  + Teams
  + SharePoint
  + Power Automate
  + Power Apps
  + Power BI
* DAF CDAO “Envision” providing access to:
  + AWS EC2 and S3
  + Palantir Foundry
  + Python and R scripting
  + Data Application Programming Interfaces (APIs)

To the maximum extent, the hacker teams used government-furnished equipment (GFE), including computers, network connections, and virtual private networks (VPNs), to ensure any outcomes of the event could be implemented within the significant constraints of the DAF network environment.

## 5.2 Expose

**Expose “data hackers” to your organization’s test and management data.** Data silos, also known as information silos, are widely acknowledged as detrimental to organizational health, as evidenced by a quick web search for "data silos" OR "information silos". The disadvantages are many, including:

* *Duplication of data*, because everyone feels as if they need a local copy, or out of simple ignorance of the existence of the data elsewhere in the organization; results in loss of configuration management and authoritative sources
* *Lack of diversity*, because no fresh eyes see the data, no new ideas or different perspectives can be heard; results in stagnation of process and thinking
* *Increased friction*, because even the simplest data requests must be manually handled, sometimes literally so, making processes vulnerable to single points of failure if an individual is sick, on travel, or on leave; results in increased frustration and decreased efficiency and effectiveness

[Figure 2](#fig-silos) shows a conceptual example of duplicated information in different teams of the same organization, with no connections between them (http://factory.dev n.d.). Such an organization would conceivably suffer from all of the disadvantages above.

|  |
| --- |
| Figure 2: Examples of data silos with duplicated information and no connections |

Even without doing the heavy lifting of breaking down data silos and forging the infrastructure, policy, tooling, and more required for a sustainable solution, a Data Hackathon can:

* identify and highlight duplication
* bring in diversity of experience, skill, and even culture
* reduce friction by creating interpersonal connections and building trust across the organization

### AFTC Data Hackathon exposure to data

Team formation for hackathons, or even the choice between individual hackers or a team approach, can vary according to the objectives of the event. One of the explicit objectives of the AFTC Data Hackathon is to discover latent organic talent. The AFTC Data Hackathon organizers operate on the assumption that there exists latent talent within AFTC and sister organizations, defined as people who:

* are already highly skilled, but have not been tasked with data-centric activities in the course of their primary duties
* are looking for ways to increase their data-centric skills, and even to make a career “pivot” away from their current role and into a more data-centric career field
* are “data-curious” and interested in exploring data-centric activities, but have not been prepared or able to fully commit to pursuing building the skill set, which can be expensive in time, money, and attention, whether directly or via opportunity cost.

The AFTC Data Hackathon is also focused on discovering and developing organic talent, defined as:

* Military - active-duty, Reserve, or Guard military members
* Civilian - United States Department of Defense (US DoD) civil service employees
* Contractor - DoD support contractors, and to a lesser extent “prime” contractors who represent the original equipment manufacturers (OEM) of systems acquired by the DoD

The AFTC Data Hackathon is not focused on commercial sources of talent during the event, and the event is not used as a way to assess commercial offerings not already available to AFTC. Proofs-of-concept or prototypes developed by organic talent during the event may lend themselves to commercial support to bring capabilities to production and daily use, but this path is not assumed a priori.

## 5.3 Evolve

**Evolve third-party and in-house scripts and apps to solve real-world problems.** Evolution, as opposed to “creatio ex nihilo” creating something from nothing, implies some existing capability that is then brought to bear to provide a solution. Evolution has many advantages for a Data Hackathon:

* building on an existing foundation of code and tooling
* encouraging reuse or remixing of existing capabilities (Creative Commons n.d.)
* increasing the likelihood of an “out-of-the-box” solution that requires minimum installation or configuration when put into production

Real-world problems likewise have several advantages:

* A built-in value proposition, since the Problem Owner has an existing pain point, little additional justification is required
* Existing dataset, typically the problem is centered on the use of an existing dataset, reducing the need for “synthetic data” to emulate the use case (Laboratory for Information and Decision Systems 2020)

## 5.4 Expand

# 6. Lessons Learned

Knowledge learned along the way and of value to the next brave soul to work in this area.

# 7. Conclusions

Points to be drawn from the material and data provided.

# 8. Acknowledgement

To extend thanks in support of the paper.

# 9. References

Citations available on Zotero in the hackathon-paper Group, <https://www.zotero.org/groups/4761353/hackathon-paper>

“Apache SparkTM - Unified Engine for Large-Scale Data Analytics.” n.d. Accessed August 26, 2022. <https://spark.apache.org/>.

“Cloud Computing Services - Amazon Web Services (AWS).” n.d. Amazon Web Services, Inc. Accessed August 26, 2022. <https://aws.amazon.com/>.

“Cloud Computing Services | Microsoft Azure.” n.d. Accessed August 26, 2022. <https://azure.microsoft.com/en-us/>.

“Cloud Object Storage – Amazon S3 – Amazon Web Services.” n.d. Amazon Web Services, Inc. Accessed August 26, 2022. <https://aws.amazon.com/s3/>.

Cloys, Tech Sgt. Robert. 2022. “Third Iteration of AFTC Data Hackathon an Innovation Success.” Air Force Test Center. August 17, 2022. <https://www.aftc.af.mil/News/Article-Display/Article/3130340/third-iteration-of-aftc-data-hackathon-an-innovation-success>.

Corporation, Microsoft. n.d. “Business Application Platform | Microsoft Power Platform.” Accessed August 26, 2022. <https://powerplatform.microsoft.com/en-us/>.

Creative Commons. n.d. “When Is My Use Considered an Adaptation?” Frequently Asked Questions - Creative Commons. Accessed August 27, 2022. <https://creativecommons.org/faq/#When_is_my_use_considered_an_adaptation>.

“Dash Overview.” n.d. Accessed August 26, 2022. <https://plotly.com/dash>.

“Data Hackathon.” n.d. Accessed August 26, 2022. <https://www.aftc.af.mil/About-Us/Data-Hackathon/>.

“Data Lakehouse Architecture and AI Company.” n.d. Databricks. Accessed August 26, 2022. <https://www.databricks.com/>.

“Foundry | Palantir.” n.d. Accessed August 26, 2022. <https://www.palantir.com/platforms/foundry/>.

http://factory.dev. n.d. “Factory.dev Different Teams in Business.” Accessed August 26, 2022. <https://www.google.com/imgres?imgurl=https://d28hck101xajo9.cloudfront.net/asset/BlogPostContentImages/Dfferent-teams-in-business.png&imgrefurl=https://factory.dev/pimcore-knowledge-base/whitepaper/data-silos&tbnid=MWONI7kL5MbgjM&vet=1&docid=pMhOfYw8Ix9ElM&w=1140&h=696&source=sh/x/im>.

“Hue - The Open Source SQL Assistant for Data Warehouses.” n.d. Accessed August 26, 2022. <https://gethue.com/>.

*Innovation Connect | DAF Data Fabric | 09 May 2022*. 2022. DAF Chief Data and AI Office. <https://www.youtube.com/watch?v=00HftDEJ6cY>.

Laboratory for Information and Decision Systems. 2020. “The Real Promise of Synthetic Data.” MIT News | Massachusetts Institute of Technology. October 16, 2020. <https://news.mit.edu/2020/real-promise-synthetic-data-1016>.

Maslow, A. H. 1943. “A Theory of Human Motivation.” *Psychological Review* 50 (4): 370–96. <https://doi.org/10.1037/h0054346>.

“Microsoft 365 - Subscription for Office Apps | Microsoft 365.” n.d. Accessed August 26, 2022. <https://www.microsoft.com/en-us/microsoft-365>.

Rogati, Monica. 2017. “The AI Hierarchy of Needs | HackerNoon.” June 12, 2017. <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>.

Secretary of the Air Force Public Affairs. 2019. “Chief Data Office Announces Capabilities for the VAULT Data Platform.” Air Force. October 11, 2019. <https://www.af.mil/News/Article-Display/Article/1987254/chief-data-office-announces-capabilities-for-the-vault-data-platform>.

“Zeppelin.” n.d. Accessed August 26, 2022. <https://zeppelin.apache.org/>.

# 10. Biography/Photograph

A brief paragraph(s) of the author(s). A photograph is highly recommended, but of course optional.