données

auteurs
Benjamin Croft, Director of Analytics Engineering, **Todd Schaefer**, IT Service Engineering, & **Brad Weiner**, Chief Data Officer, The University of Colorado Boulder

How The University of Colorado Boulder (UCB) is Delivering Data as a Product

Back from Educause #EDU22, a data-centric story

Préambule : En tant que membre de la délégation française au dernier congrès Educause, nous avons gardé contact avec un de nos hôtes visités, Brad Weiner, Chief Data Officer in The University of Colorado Boulder (UCB), et lui avons demandé de présenter une vision de cet établissement US sur une approche centrée sur la donnée. En accord avec les auteurs, cet article vous est présenté dans sa langue d'origine.

Institutions of higher education have substantial data assets yet little capacity for converting those assets into actionable intelligence. (Borgman & Brand, 2022). To solve this challenge, CU Boulder (UCB) is working to deliver curated data sets, as a product, to end users with appropriate training and governance safeguards. These data products span institutional domains including student success, admissions, retention, enrollment, teaching and learning, among others. This is a fundamental change from "analyses as a product" which limits analytic capacity by hiring a small core of data experts who act as information conduits to campus. These individuals are difficult to hire, expensive to train, and cannot scale to meet exponential campus data demands. In this paper, we will outline the philosophical and technical underpinnings of our process, so other campuses may learn from, and build upon, our efforts.

↘ BUILD FOR AVERAGE USERS, NOT EXPERTS.

Imagine a typical data user. Do you imagine people writing Python code or creating pivots in Excel? Although we often build for the former, at UCB, we are focusing efforts on the latter. We are doing this by:

- Augmenting relational databases with a secure, user-friendly, object storage layer.

- Focusing on "boring rectangles" that can be opened in Excel.

- Reducing knowledge barriers by standardizing analytic choices. Rather than including four variants of the same feature, we will include the most common version, document the rationale, and provide choice paths if different options are required.

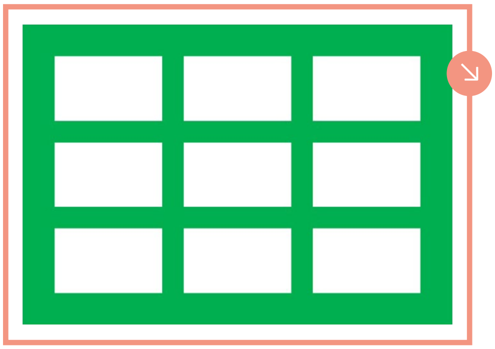

Figure 1 : A Boring Analytic Rectangle

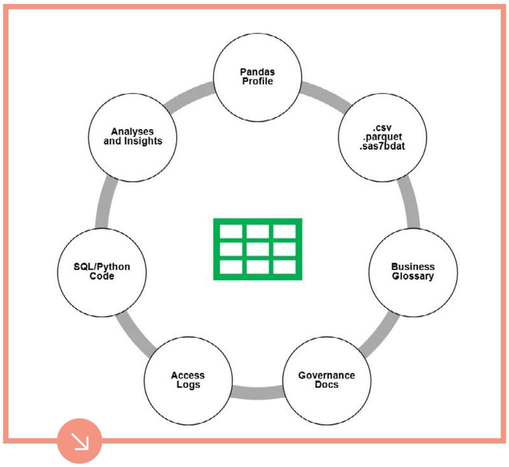

Figure 2 : "Planet and Moons" Model where artifacts are stored with data

↘ FOCUS ON CONTEXT AND DOCUMENTATION.

Earlier efforts have shown that the biggest barrier to data use is not access, it is understanding. Therefore, we have focused on making the data products as useful and easily understood as possible. To accomplish this, we are creating a “planet and moons” model where related artifacts are generated and stored in the same place as the data.

A few of the proposed “moons” include:

● An html data profile to surface data types, distributions, factor levels, and labels.

● Multiple file formats (csv/parquet/SAS7BDAT/Excel) to enable multiple toolkits.

● Business glossaries with human-readable descriptors.

● Governance documentation including data stewards, access policies, data lifecycle requirements, and contact information.

● Access logs to show who else is using the data. This mechanism will allow new product users to ask questions of people who have previously accessed the information.

● Completed analyses and results. This is intended to scale knowledge to expedite actionable interventions.

● Version-controlled code (SQL, Python) so advanced users can reproduce the product from source data systems.

↘ IMPLEMENT GOVERNANCE POLICIES THROUGH TECHNOLOGY.

To advance compliance and privacy, we must be cautious about use cases, build appropriate access controls, and maintain logs to audit compliance. That is why we are operationalizing data governance into the organizational and technical model from the start.

Some key features include:

● Generating *identified, de-identified*, and *synthetic* versions of each data product. This will enable policy-based access control (PBAC) as well as dynamic column and row masking. Only when a user can clearly demonstrate the need for an individual-level intervention identifiable records may be provided. Where group level aggregations will suffice, de-identified or synthetic data will suffice.

● Operationalizing data governance and cataloging through the Analytics Engineering team. With this structure, we can determine the "rules of use" for each data product, then automate workflows, logging, auditing, and validation of data life cycle policies.

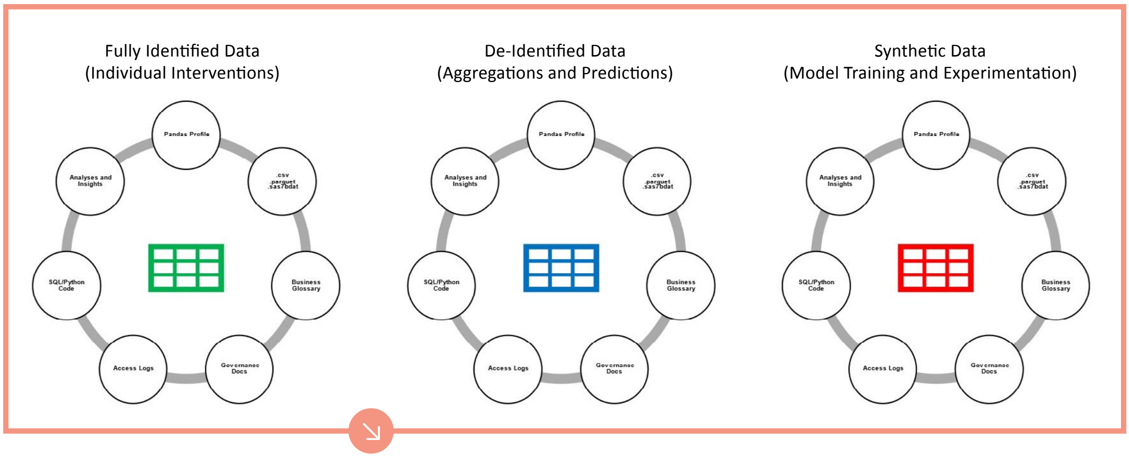

Figure 3 : "Planet and Moons" Model
Differential Access Based on Use Case

↘ ALIGN YOUR DATA ENGINEERING AND ANALYTIC TEAMS.

Making data useful at scale requires diffuse technical skills. Analytic teams may lack the capacity to build secure, reproducible data pipelines, while engineering teams may lack the institutional knowledge or domain expertise to create data models or documentation. In this situation, siloed efforts result in technical debt, conflicting sources of truth, and subpar products across both domains. To alleviate this problem UCB is:

● Creating a data engineering team comprised of analytic and data engineers. The former are responsible for interfacing with domain experts to model data sets. The latter are responsible for building and maintaining the data tech stack, and for putting data models into production. The teams converge on operationalizing analytic workflows, deployment of data science models, and integration of analytic outputs into downstream systems.

● Aligning the group with a common project charter, and unifying roadmaps, backlogs, and toolkits, agile leadership, and budgets. Eventually we aim to staff a "data desk" to act as a single point of entry for all campus data requests.

We believe that enabling data users will provide substantial benefits to our campus community, particularly in the areas of student success, faculty and staff retention, and operational efficiency while also maintaining strong governance, data management, data quality, and security standards.

References

Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. Science *(New York, N.Y.), 378*(6626), 1278–1281. https://doi.org/10.1126/science.add2734

amue
MUTUALISATION + SOLUTIONS

Numéro 6

L'ESR vu par le prisme de la donnée universitaire

Une autre vision de nos organisations

Photo by Dũng Trần Việt on

La collection numérique de l'Amue

NOVEMBRE 2019

la collection numérique
de l'Agence de mutualisation des universités et établissements d'enseignement supérieur ou de recherche et de support à l'enseignement supérieur ou à la recherche

décembre 2021

L'ESR vu par le prisme de la donnée universitaire

saison 2

amue
MUTUALISATION + SOLUTIONS

#18

« Retour sur…. »
N°06 - L'ESR vu par le prisme de la donnée universitaire, novembre 2019

« Retour sur…. »
N°18 - L'ESR vu par le prisme de la donnée universitaire - Saison2, décembre 2021

Pour aller plus loin :
Loi européenne sur la gouvernance des données
Un « acte » Européen, porté par la commission est à lire sur cette page. Ce « data governance act » est un des piliers de la stratégie européenne en matière de données.

Commission européenne