Zsofia Vamos

Capstone Project Public Summary

Dataiku Usage Monitoring Dashboard

Abstract

This report aims to summarize a Capstone Project completed as part of MSc in Business Analytics at CEU, 2020. The client is the Analytics & Data department of a multinational company in the financial sector. The project's goal was to refine the existing methods of usage monitoring of one of the organization's BI tools by creating a more streamlined data flow and an interactive visualization. This would allow the client to have a better grasp on what is happening in the analytical environment in terms of data usage, user behavior or usage patterns. The resulting dashboard had two pages with three main components – one generic overview of usage patterns, a breakdown of usage data from a user perspective and a detailed view of usage patterns on the folder level. The dashboard is currently deployed in the internal Tableau server of the client.

Introduction

The client is the team responsible for driving the platform and enablement strategy of a major business unit within the company. This project focused on creating a simpler and better way of monitoring the usage of Dataiku, a BI tool used for data cleaning & analysis in the company's prototyping environment. Due to the successful adaptation of this tool, more and more users started exploring it and hence closer monitoring has become hard to perform.

The existing monitoring methods were tied to manual log extraction from the tool itself; however this method was time consuming and failed to provide an accurate picture without further data transformation. Some required metrics were not readily available in the tool's log as these were created by the organization for internal tracking purposes.

In short, the goal was to gather all the data available – logs, stats, table attributes, employee data – and create a Tableau dashboard that enables the client to see the bigger picture as close to real-time reporting as possible. The three main areas of interest have been identified as power-users, longest running workflows and usage patterns. Identifying power users and longest running

workflows are both crucial factors in terms of creating a strategy for technical training, while potentially sensitive data usage can also be tied to names. Having a visual history of usage patterns will help the team start making estimations in case they will decide to move on to a more scalable data storage solution in the future.

Tools & Methodology

The project had two key elements, a data model and the visualization. The data was built within Dataiku from the combination of three tables – internal metrics of the tool, employee dataset and a dataset focusing on current tables on the server.

The first two sources were readily available – given proper access - but the third one had to be retrieved through a Python script customized for the attributes I was looking for. The metrics dataset was transformed – eg. it was appended with a duration column and adjusted time zone columns, while a couple of irrelevant variables were removed. The data set resulting from the Python script got grouped and aggregated to a level were the server connection stats could be analyzed. The employee table was joined to both of these datasets separately, in one case to fill in user data, in the second case to provide information on the project owners.

The dashboard itself was built in Tableau 2019 and included two separate tabs. The first tab contained 10+ calculated fields that intended to shape the data source to fit the goal. The second tab used a bit less calculations and information, but meant to provide more of a drill-down experience to the user.

Dashboard

The final dashboard has two tabs and three key storylines. The first part focuses on the total usage metrics from the standpoint of the last 7 days. The calculation I used was a rolling window calculation comparing the last 7 days' total duration/run count to that of the 7 days preceding the current time range. I included both exact and relative changes in frequencies and assigned the accent colors to the two key metrics – duration and job count. Whenever these colors appear, the same metrics are being displayed but the level of granularity may change. In some cases, running workflows may fail and a higher rate of failures would indicate a systematic problem – so I also included the success rate with a detailed tooltip.

Belonging to the first key area, I created a line of three items that tend to illustrate the process of starting from the macro level and drilling into the micro. The first chart here is a line chart showing aggregated duration data for the last 30 days relative to the max date in the dataset. A reference line was also included to give an impression of how 'normal' the past 7 days have been compared to the past 30. Lines are split based on whether or not the job was an automated or manual operation. The two other graphs show the same distribution for the last 7 full days and the hourly distribution for the last full day in the data. With all the interactive tooltips included, these charts give the client a very good sense of extreme cases or extreme days.

The second piece visualizes the same metrics but on an individual user level, allowing us to see the top 10 users of the past 'week'. Here I included a line chart showing the users' 30-day pattern to have a sense of when a user is doing something out-of-ordinary compared to their own standard behavior.

Finally, the second tab of the dashboard shows a chart and a table – and is intended to visualize the volume of Dataiku projects and their place within the overall environment. These visualizations are somewhat correlated as one can filter the other, but both can provide a great drill down into the project level usage metrics.

The dashboard is currently deployed on the internal Tableau server and is used for the very purpose it was created for. Reproducibility should not be a problem as all steps in the data model were documented in Dataiku itself, and the 'skeleton' workbook is shared with the responsible team. This makes future upgrades easier too.

Lessons Learned

While the creation of a functional Tableau dashboard may not look like a lot of work, I have definitely encountered some obstacles along the way. Inconsistencies in internal definitions and labels within the tool made it hard to understand what my goal actually was and how to measure aspects of usage. Tableau is a great BI tool but it struggles with complex datasets and doesn't always behave well in a star-schema setup. I had to create the data model twice, once the way it looked like in my head and once the way Tableau could process it the way I wanted.

Creating a visualization that shows important data but does not overwhelm the audience is also something that I got to learn here – the resulting dashboard was preceded with 5 prototypes. That

part of the process taught me the importance of client feedback and I ended up valuing the agile framework a lot more. Throughout the process I learned that a very good dashboard should not attempt to convey more than three important messages – in that case it's better to move on to creating more in separate projects.

Throughout the initial stages I also learned that a prediction model is not always the golden key to everyone's problems. I experimented with including a GBM model to predict future peaks but at this point in the project it made no actual contribution to the content preferred. It is nice to have a data model where a model is easy to integrate, but it is better to focus on understanding the client's problem first and tailoring a solution for their needs.