JUNE 2023

GROWTH ACCELERATION PARTNERS

WHY OFF-THE-SHELF DATA ENGINEERING SOLUTIONS RARELY FIT (AND OTHER PITFALLS TO AVOID)

Is a Data Mesh Architecture Right for You? Get The Most Out of Your Data Engineering Architecture with GAPBuilt Accelerators

DATA ENGINEERING

If your organization is migrating to the cloud, then you will already know the myriad benefits, from agility and flexibility, to the ability to innovate quicker. The latter is perhaps the most alluring as, with data science projects for example, the ceiling is so high. But without data engineers providing the foundations of data management so the more abstract work can be done by the analysts and scientists, your exciting new project could be the proverbial house with foundations built on sand.

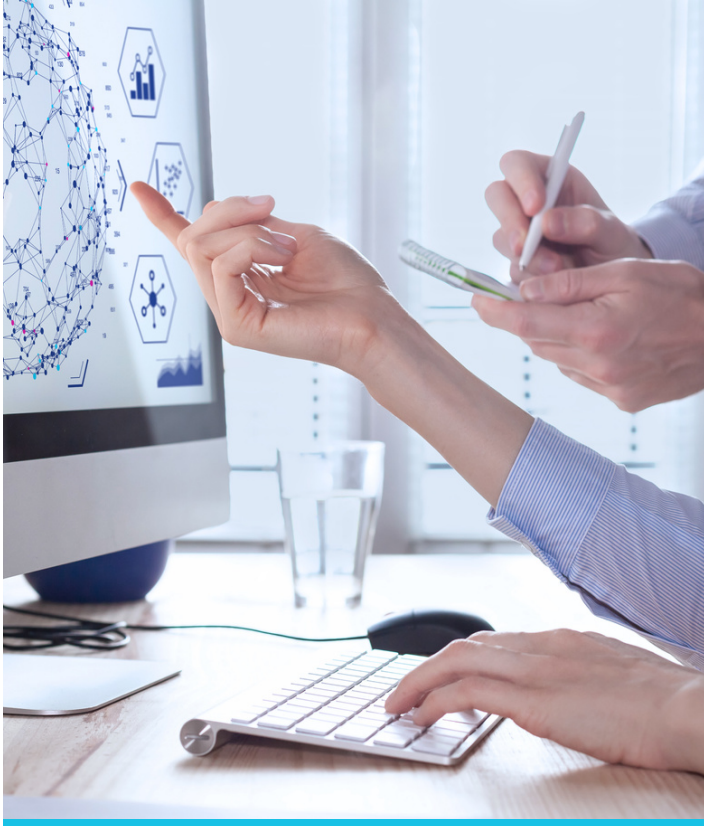
To continue in this vein, your house may not have had much work done on it in a while, if you have technical debt which potentially goes back as far as decades. This is perfectly understandable if your company has grown quickly and has been hyper-focused on time-to-market. But the upshot is that moving to new platforms, as well as new paradigms in data lifecycle and data governance, requires a significant restructuring of how an organization operates.

It is a major undertaking, which might explain why, like home improvements, people put it off for so long. But it needs to be done. Get it right, with the help of a trusted partner, and you can build to your heart's content. Get it wrong, and it will put your business back months — if not years.

Fundamentally, data engineers are the glue that holds an organization's data lifecycle together. They create repositories for data scientists and data analysts, and require knowledge of data warehouse architectures, as well as coding in languages including Python, SQL, Scala and R.

The ideal definition of a data engineer is one who manages the data lifecycle as it moves from the operational to the analytical plane. The operational plane relates to live data being used to fulfill your primary business objective. An example here is a sales website; your data will include records of inventory being offered; any transaction fulfilled; every customer that is currently active.

As these pieces of data age, they move into the analytical plan to become historical data where analytics can be applied. This can either be in a data lake, defined as containing an organization's raw unstructured data, or a data warehouse, with structured data that has been cleaned and processed. Depending on which plane is used, there are different requirements in terms of how readily the data needs to be available, storage, and what kind of actions will be implemented on them. These actions can be defined in three categories: descriptive, prescriptive and predictive analytics.



So what does a significant modernization and restructuring look like in more tangible terms? An emerging architecture that organizations may look to take advantage of is the data mesh. It is defined as an "analytical data architecture and operating model where data is treated as a product and owned by teams that most intimately know and consume the data." The theory goes that under a traditional structure, once the data are homogenized from the former to the latter, a lot of insight is removed that may be relevant to the business.

Take a music streaming company as an example. They will have different business departments: content, sales, client-side, etc. These departments would each have a team of software engineers, data engineers and analysts, and potentially a data scientist, working on building data products.

One other element of data mesh architecture is federated data governance. If data products are created in a decentralized and democratized way, then there needs to be structure to back it up. Even if each domain is its own unit, consumers may require data from multiple domains. Therefore each domain, if it is part of the mesh, needs to follow centrally-managed guidelines for data categorization, access and discovery. This includes schema structure, SLA definition and data contracts, alongside privacy and encryption.

The problem for organizations is that this, understandably, requires a significant architectural shift, and it is difficult to replace the foundations if it is not your specialization. This is where a partner like GAP can come in and take the strain from such a large-scale project.

Sometimes it can be a wake-up call for the business. One client felt they needed to change after they realized the company made various reports for the C-suite to make decisions that the executives did not trust. They may have looked at the reports, but they were not particularly used in the decision-making process.

Journeys such as this can sometimes take years, as the migration can just be the starting point. One particular cloud migration project saw a GAP client move from Microsoft SSIS (SQL Server Integration Services) to Microsoft Azure — specifically Azure Data Factory, a fully managed, serverless data integration offering. The client also opted for Profisee, an MDM (master data management) platform for data governance, as well as Snowflake for more advanced analysis.

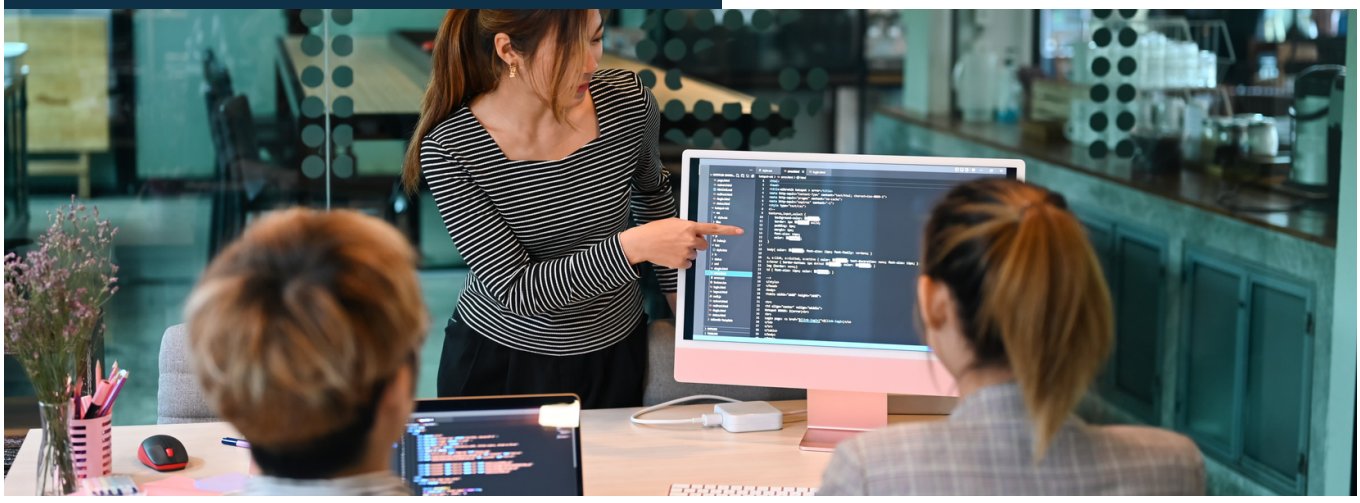
After the migration, the client needed to build up maturity on these platforms, so it meant a longer relationship between GAP and the client. As a result of this confidence, the client was subsequently able to launch insights, delve more into governance, and implement data quality automation and data governance tools.

One game-changing element is the use of services such as Azure API Management and Azure Functions in Python. Before the abstraction layer of APIs became commonplace, making a change in data models was a major problem, as it would require changing 10 different systems that were in production.

A frequent pitfall for organizations is when they are quoted for an off-the-shelf solution as opposed to a custom one. Organizations will have a lot of integrations with data vendors and different data sources to contend with — even if the data is heterogeneously formatted or documented within the company.

Think of it as a series of adapters that not only need to be plugged in for the data to be located in a data warehouse or data lake, but with each, one goes from the specific database into the data lake, another from an API, and another from a manual entry process. It is complicated and time-consuming to manage; yet nearly as complicated — if ultimately more streamlined — is understanding and implementing tools offered by the major cloud providers.

It can be difficult to justify doing a lot of foundational work without immediate benefits — and this is where GAP can come in to help. Utilizing the agnosticism of GAP data engineers and project managers, along with GAPBuilt Accelerators around data — which are scripts that can deploy infrastructure and an entire ingestion and data maintenance pipeline within Amazon Web Services (AWS) or Azure in minutes — can not only save time, but ensure the foundations of an important transformational project are rock solid.



To find out more, please visit WeAreGAP.com