sahaj software

SPRINGER NATURE

Driving Data-led Decision-making through Data Engineering for Superior Business Impact

Springer Nature is a global academic publishing company that advances discovery by publishing trusted research. Following the 2015 merger and subsequent growth with the acquisition of products, the company's workflows became inundated with multiple systems and different data models driving article submissions from authors - a key business process.

Data analysts across different teams would use several manual processes while navigating a complex ecosystem of multiple data stores to build an aggregate view of the business, identify trends and support data-driven decision-making.

This resulted in data silos, duplication of data due to multiple sources and difficulty in drawing insights to improve business processes. Data analysts struggled to derive business critical insights. This is where Sahaj stepped in.

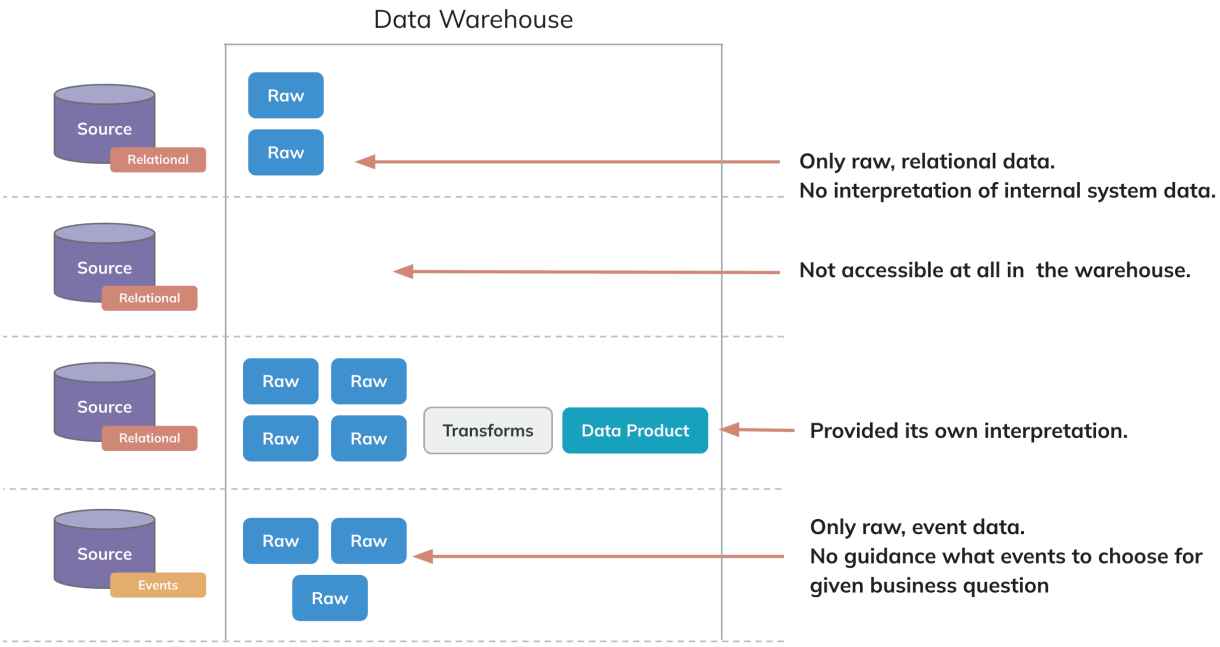
A team of consultants from Sahaj built an aggregated submissions data model from different peer review systems for publishing scientific research papers. The objective was to aid and boost business teams' efficiency while shielding them from changing data models or complexity in joining datasets, and eventually deliver better business outcomes.

Unifying the Springer Nature Universe with Data Engineering

Consultants from Sahaj brought a product-thinking mindset to the table. The objective was to put the user first and design the product for self-service with focus on outcomes rather than outputs. The approach was rooted in foundational data architecture principles including distributed data ownership, robust data governance, prioritised data quality and the adoption of open standards to mitigate vendor lock-in and foster seamless interoperability.

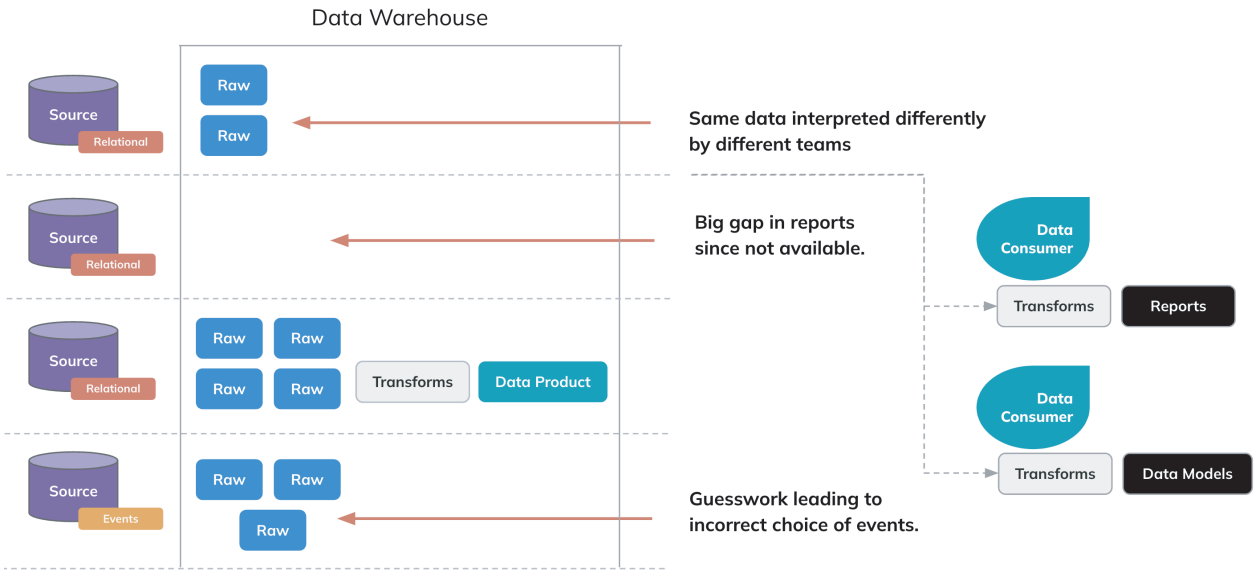
Below is a very high-level simplified view of the data landscape before Sahaj partnered with Springer Nature. Numerous data sources across the landscape made it tough for teams to compile consistent reports.

Data Landscape before Sahaj



Data Landscape after Sahaj

The team crafted a solution that would provide a single unified view of submissions across the business. This is how the data landscape evolved following Sahaj's partnership with Springer Nature:



Core components of the solution included:

- **Automated Data Pipelines** - to extract load and transform data from multiple submissions systems into a single data product with essential data elements that would allow several consumer-driven representations by combining with other available data products.

- **Data product** built with **embedded non-functional requirements** and **data quality attributes** such as freshness of data, data lineage, alerts, monitoring and self-service support for the consumers. Data pipelines have been in production for over a year now; earlier, there were 1-2 production incidents per month on an average, all of which were reported by automated alerts and monitoring solutions in place. Time to resolution was about 8 hours on an average to detect and release a fix into production.

- For the tech stack, we chose dbt for efficient, best-practice data transformation and model management. Apache Airflow was used to automate workflows for efficient data processing and scheduling, and providing a powerful, reliable solution for managing pipelines. re_data enabled upfront observability, helping us catch and rectify bad data in pipelines, ensuring a reliable end product.

Outcome & Impact

The Joint Submissions Data Product has resulted in a transformative impact on the business by unlocking new opportunities, improving decision-making and driving innovation across several business teams. Currently, the data product is used by over 32 teams across the business and this number is growing.

1. Analytics teams across the business are using joint submission datasets to build meaningful reports that did not exist before. These reports have over 13,000 monthly page views and continue to increase.

2. It serves as a critical data point that enables centralised strategic decision-making and aids increased content in collections. Collection reports built on joint submission dataset serve as a key metric to measure performance of collections and provide key insights to adapt new strategies for greater impact.

3. Business teams had a very complex data pipeline that was updated manually to understand rejections. They now update these reports daily. Step change is huge in what can be done with this data in driving content retention strategies.

4. With the rejection tracker report, business teams now understand the key areas of improvement for content retention.

5. Joint submissions data product has allowed the business intelligence teams to build a journal health dashboard to manage the articles pipeline and understand the bottlenecks in the process and provide actionable insights to the editorial teams that leads to reduced turnaround times. As a result, author experience is positive consistently.

6. Business had an unclear approach to data matching when extracting reports prior to joint submissions data product. They now have a streamlined approach which delivers quality data daily to the end users.

7. The Transfers part of the data product allows end-to-end visibility into the journeys of authors and manuscripts within Springer Nature, enabling daily situational awareness of performance vs targets as well as creating the foundation for developing sophisticated content retention strategies.

Joint Submissions Data Product has built a solid foundation that will deliver business impact in the future both in terms of content acquisition and content retention. It has enabled business teams to make strategic data-driven decisions. The availability of basic submissions data in one place in a structured manner enables better business opportunities and the product continues to evolve as business use-cases grow.