Unifying SEEK's Complex Data Ecosystem for Business Transformation

July 2024

SEEK is a market leader in online employment marketplaces with a multinational presence spanning Australia, New Zealand, Hong Kong and South East Asia. In addition, SEEK has minority investments in China, South Korea and several other countries. In 2014, SEEK acquired leading employment marketplaces in South East Asia, JobsDB and Jobstreet.

We partnered with Seek to implement the Unification programme - an ambitious plan to merge diverse systems from various geographies into its existing Australia and New Zealand platforms. A lean team of consultants from Sahaj joined forces with Seek's multiple teams scattered across the globe to move the core online, CRM, data and finance platforms onto a single unified platform for eight unique markets.

The platform offers a range of new AI-powered enhancements to SEEK's Jobstreet and JobsDB customers. It further enables SEEK to develop and deliver new products to customers more efficiently and sustainably. The build once-deploy anywhere capability, gives SEEK the potential to impact over 500 million candidates and 5 million employers.

The Ask

The program's core initiative was to unify candidate and hirer data in a single system to transition to the unified platform seamlessly. A critical dependency of this process was the ability to live-sync data from the source to the destination to validate the integrity of the data continuously and to test and examine the behaviour in our new system.

After six months of testing, Sahaj came on-board to help overcome challenges related to the scale and scope of the data migration and support delivery to meet strict timelines.

The Challenge

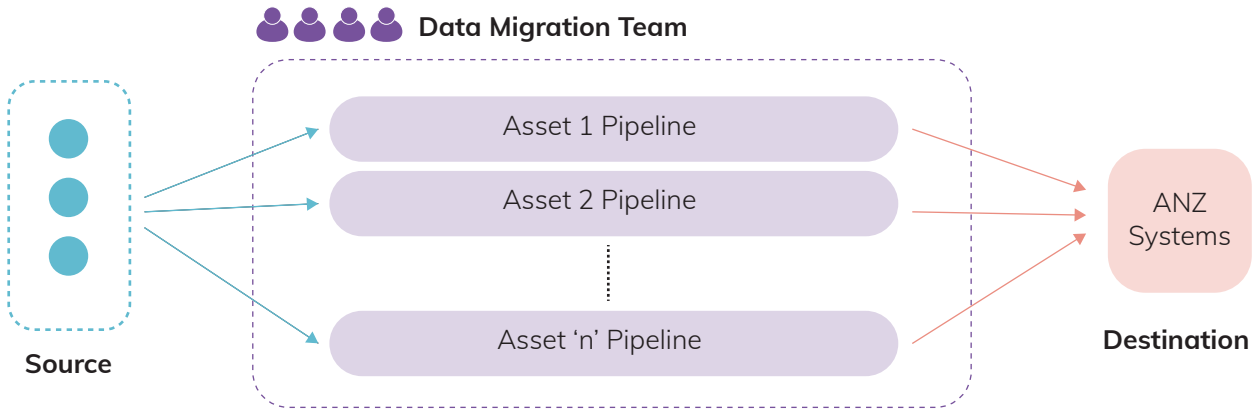
The initial plan was to have a single Data Migration team that would be solely responsible for migrating all the data assets in the right order. Sahaj joined this team and built the first data pipeline in 3 months. They encountered a range of challenges along the way:

- Understanding the domain and business rules as a central team and the order of precedence in parent and child objects would be time-consuming. Moreover, meeting deadlines would demand a much bigger team.

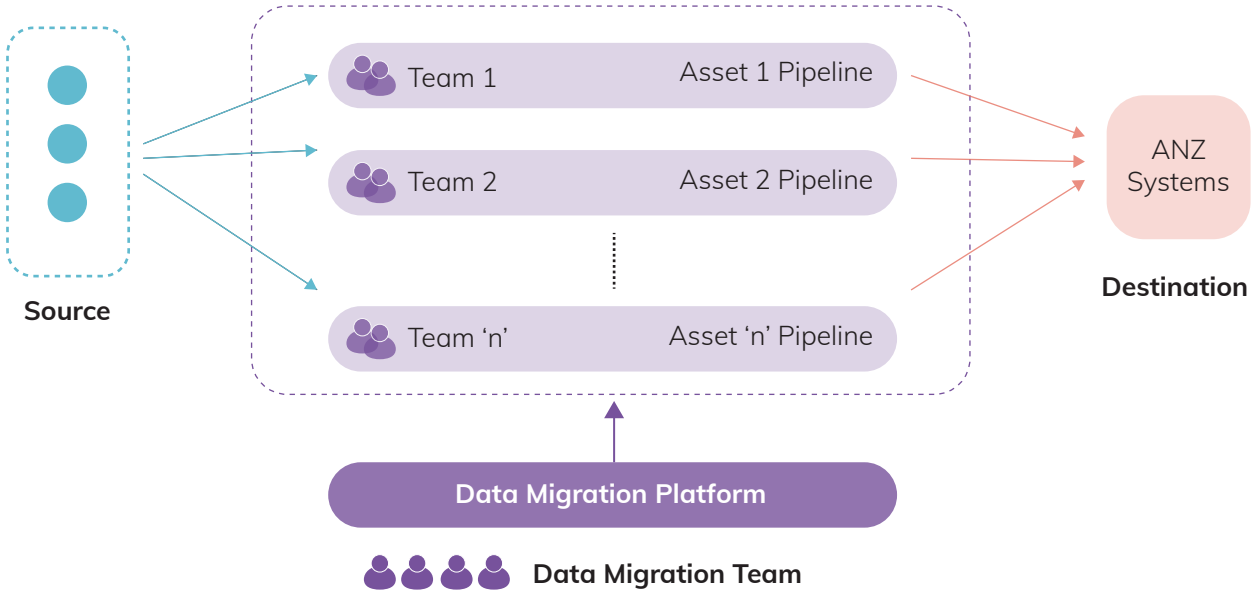
- Massive amounts of data spanning 6 countries had to be migrated keeping in mind the data quality issues, observability of progress, dependency between child and parent entities. This included
  - 30+ business entities or assets to be migrated.
  - Capacity to seamlessly run 18+ pipelines simultaneously.
  - More than 2.7 billion events to be migrated as live-sync and historical data.
  - 15+ teams with 80+ developers to be able to work simultaneously with different domains, deadlines and dependencies.

- Live-sync migration was from and to live systems which needed to trigger certain business rules as data was migrated as well as deferring the sync to a later date.

Initial Plan



Structure after Sahaj joined the team



The Solution

Along with our partners in the Data Migration team, Seek made the following recommendations which consultants from Sahaj adapted and executed:

- Pivot away from a central team executing data migration towards a platform team that provides tooling and guidance to enable data migration.
- Provide a coherent and standard interface for all existing custodian teams to take responsibility for writing their custom domain mappings and applying business rules – essentially, a standardised ETL process.
- The standardised ETL process should support both live sync and bulk import.
- The standard interface must support all SEEK data sources (S3, DynamoDB, SQL, JSON) and abstract error handling, retry mechanism and order-of-precedence challenges from our custodian teams.
- The Sahaj team would still take responsibility for a couple of assets to ensure that we used our platform.
- Work closely with an additional partner team who had the remit of governance, training and program management of the data migration delivery.

With their expertise in platform engineering, consultants from Sahaj built a platform that provided the following features to accelerate each domain team's journey:

- A CLI to bootstrap the domain team's code base with infra in AWS CDK, Dynamic CI/CD pipeline generation, basic app skeleton of the pipeline, observability and smoke tests in a couple of hours.
- Typescript-based monorepo used across app and infra code.
- Automated/Manual batch retrial processes to handle data inconsistencies.
- Automated completeness reports using athena to track progress.
- Opinionated handling of out of order events to ensure data integrity with the option to opt out.
- Framework for long-running batch jobs using AWS step functions.
- Platform engineering rituals like fortnightly dev sessions, slack support channel and extensive documentation on the developer portal.

Outcome

With the proposed initial plan, the first pipeline, built within 8 months, lacked observability, self-healing and automated historical migration jobs. After the platform initiative approach, we built 17~ additional data pipelines owned by 14 different teams, each built within a matter of weeks. During the project, 83 developers from different teams contributed to the platform codebase.

More than 850 million records across 23 distinct data assets were migrated with the platform, all while retaining data hierarchy and integrity with a real-time, one-way sync process running between source and destination systems.

This data migration platform was critical to the technical implementation and played a pivotal role in the successful outcome of the overall unification project.

The Data Migration Platform was built to be scalable and flexible to accommodate the diverse ecosystem of the systems being unified. Therefore, it removed the need to build multiple systems to achieve the same. It was able to handle heavy loads during peak migrations without any issues while maintaining low costs. It was built-in with various tools and frameworks which enabled the teams to set up a working pipeline in minutes.

Impact

- Due to the platform approach, the time taken to create a data pipeline to migrate a data asset reduced from months to weeks.
- The built-in observability and monitoring in the platform provided the developers with greater visibility into failure and helped them to take remedial measures in time.
- The platform provided inbuilt completeness reports, which enabled the stakeholders to have greater visibility into the completeness and correctness about data sync and migrations and enabled them to make better decisions around the Unification program.

Key Technologies Used

Typescript
Reusable infrastructure components using AWS CDK
Monorepo
Dynamic build pipeline generation using Buildtike API's
Core AWS Services: Kinesis, AWS Lambda, SNS/SQS, Athena, Dynamodb, AWS DMS, Step functions

Key Project Metrics

During the peak 5-6 deploys to prod a day.
279 PRs in a month
80+ platform users
More than 2.7 billion events moved through the pipelines.
More than 2000 AWS resources created using the platform.
Monthly running AWS cost - 6k+ AUD

Works Cited

(n.d.). Retrieved from https://www.prnewswire.com/apac/news-releases/seek-jobsdb-and-jobstreet-unify-marketplace-platforms-transforming-apac-employment-landscape-302068589.html