

Unlocking Affordable Housing Demand Data for BAHFA

A new data pipeline gives BAHFA its first analytics-ready view of who is applying for affordable housing across the nine-county Bay Area.

By Patrick Crouch & Jackson Dawson – USF MSDS Practicum at MTC / BAHFA

Introduction

Affordable housing in the Bay Area is a supply-and-demand problem, but the two sides have never been equally visible. The supply side — units, developments, grants — is reasonably well understood, because developers must coordinate with regional agencies to get projects funded and built. The demand side has been much harder to parse. Who is applying, where, in what volume, and with what needs has historically been spread across hundreds of individual properties.

That asymmetry has consequences: it is difficult to plan well for demand you cannot measure. The Bay Area Housing Finance Authority (BAHFA) — established by the state legislature in 2019 as California's first regional housing finance authority — has been narrowing the gap since 2023, when it launched Doorway, the first regional affordable housing platform in the country. For the first time, listings and applications from across the region flow into a single system. Until now, however, that effort stopped at the point of collection. Our task was to move the demand data out of the storage where it was resting and into an analytics-ready database — one that can inform internal decision-making and, eventually, support public data transparency.

The Challenge

The encouraging part is that the data already existed. The difficulty was that Doorway's database is maintained by a third-party contractor, Exygy, and is optimized for the live application experience: fast, reliable sign-ups and submissions. It was not designed to answer analytical questions, and it holds all of the sensitive personal information that applicants provide. Producing a regional trend from it directly is impractical. What BAHFA needed was its own copy of the data — owned by MTC and BAHFA, organized for analysis, and safe to work with.

Privacy First, By Design

Before describing how the pipeline works, one point deserves emphasis: this is application data about real people seeking housing, and responsible privacy handling was central to the design from the first step rather than an adjustment made at the end.

The approach is most easily described as what we remove versus what we keep:

- **Removed:** names, phone numbers, exact birth dates, and street addresses. These identify a specific person and carry little analytical value, so they never reach MTC's systems.
- **Kept for insights:** age at time of application, location (at the zip-code or census-tract level), household size, reported income level, and demographics. These support analysis in aggregate without identifying any individual.

One field is treated as a deliberate exception. Email is intentionally retained, for a practical reason: when an applicant contacts support, email is the most reliable way to locate the correct account, and it is the most dependable key for eventually linking demand data to placement data. It is the one identifier kept on purpose, and a long-term plan for retiring or deleting it over time is part of the project's future work.

What We Built

At the highest level, the system performs a single function: it moves data from the Doorway platform into an MTC data warehouse, where it is ready for internal analysis.

The data moves in four steps:

1. It begins in Doorway's live database — fast for applicants, but not built for analysis, and containing the personal information that must be protected.
2. It is copied into a data lake, which captures a fresh snapshot every day and records every change over time.
3. It is de-duplicated into a staging copy — a clean mirror of the original database, with personal information already removed.
4. It is transformed into a final, analysis-ready warehouse that is straightforward to query.

The entire process refreshes automatically every night. No one has to run it or pull data by hand; each morning, the warehouse reflects where the data stood the day before.

Checks are built in at every step. As data moves from one stage to the next, the pipeline confirms that row counts and record keys match and enforces strict definitions of what each table should contain. The purpose is trust: when a dashboard reports a number, there are safeguards behind it confirming that nothing was silently dropped or altered along the way.

(For readers interested in the technical detail: the warehouse runs on AWS, with a data lake in Amazon S3, an Amazon Redshift warehouse, and nightly jobs orchestrated through AWS Glue.)

What This Unlocks

Questions that previously required a special data request — and often several rounds of back-and-forth — are now answerable through a dashboard or a single query. For example:

- What kind of housing is in demand, including how many bedrooms applicants are seeking?
- What household sizes are applying, and from where?

- What are the demographics of applicants for a given listing or jurisdiction?
- How much of the demand is local — applicants applying within the jurisdiction where they already live?

The early findings are concrete. As of data collected on May 27, 2026, roughly 45 percent of applicants are single-person households, and about 19 percent apply within their own jurisdiction.

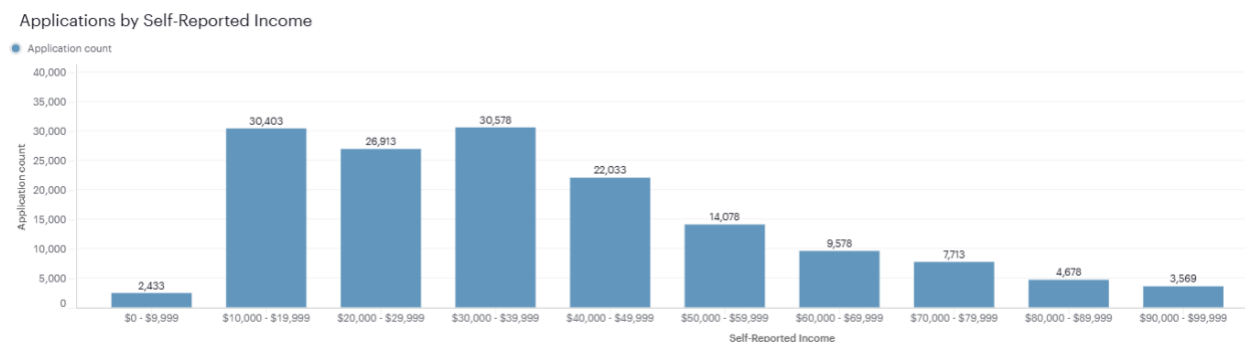
The clearest change is in everyday workflow. Previously, answering a question of this kind meant a request changing hands and a Technical Support Services staff member stepping away from their own work to assemble a one-off data pull. Now a BAHFA team member can consult a dashboard or answer the question directly — fewer steps, less waiting, and direct access to the data.

Where This Goes Next

A common framework describes data maturity in three stages: bronze, silver, and gold.

- **Bronze** is raw data, more or less as it lands from Doorway.
- **Silver** is data that has been cleaned and transformed — organized, reliable, and ready to power internal work and controlled dashboards. **This is where the project stands today.**
- **Gold** is an aggregated, fully privacy-safe dataset — pre-summarized and anonymized to a degree that allows it to be shared publicly with confidence.

The current system can safely power dashboards, a brief sample of which follows, because the team controls precisely what is displayed and can mask any group small enough to risk identifying someone. A fully public, downloadable dataset is a larger step that requires the gold layer first. That is the main work still ahead, and it is the path toward the longer-term goal: a public-facing affordable-housing indicator, potentially as part of MTC's Vital Signs, that lets anyone see how well supply is meeting demand across the region.



Example dashboard graphic: Sample of total Doorway applications by income bracket between May 27, 2025 and May 27, 2026

Thank You

We are grateful to have contributed to this effort, which has been years in the making. Particular thanks to Kaya Tollas (MTC), Barry Roeder (BAHFA), Glenn Brown (TSS), Roshen Sethna (external advisor), and the Exygy team for their guidance and partnership. We hope the project will be carried forward by next year's USF MSDS practicum team, and we look forward to seeing what they build on top of it and where BAHFA takes it from here.