New Perspectives in Small Data

Trino for small and medium enterprises

Executive Homes

BUILDING DISTINCTION®

Introductions

Benjamin Jeter

Staff Data Architect

ben@executivehomes.com

Tommy Zugibe

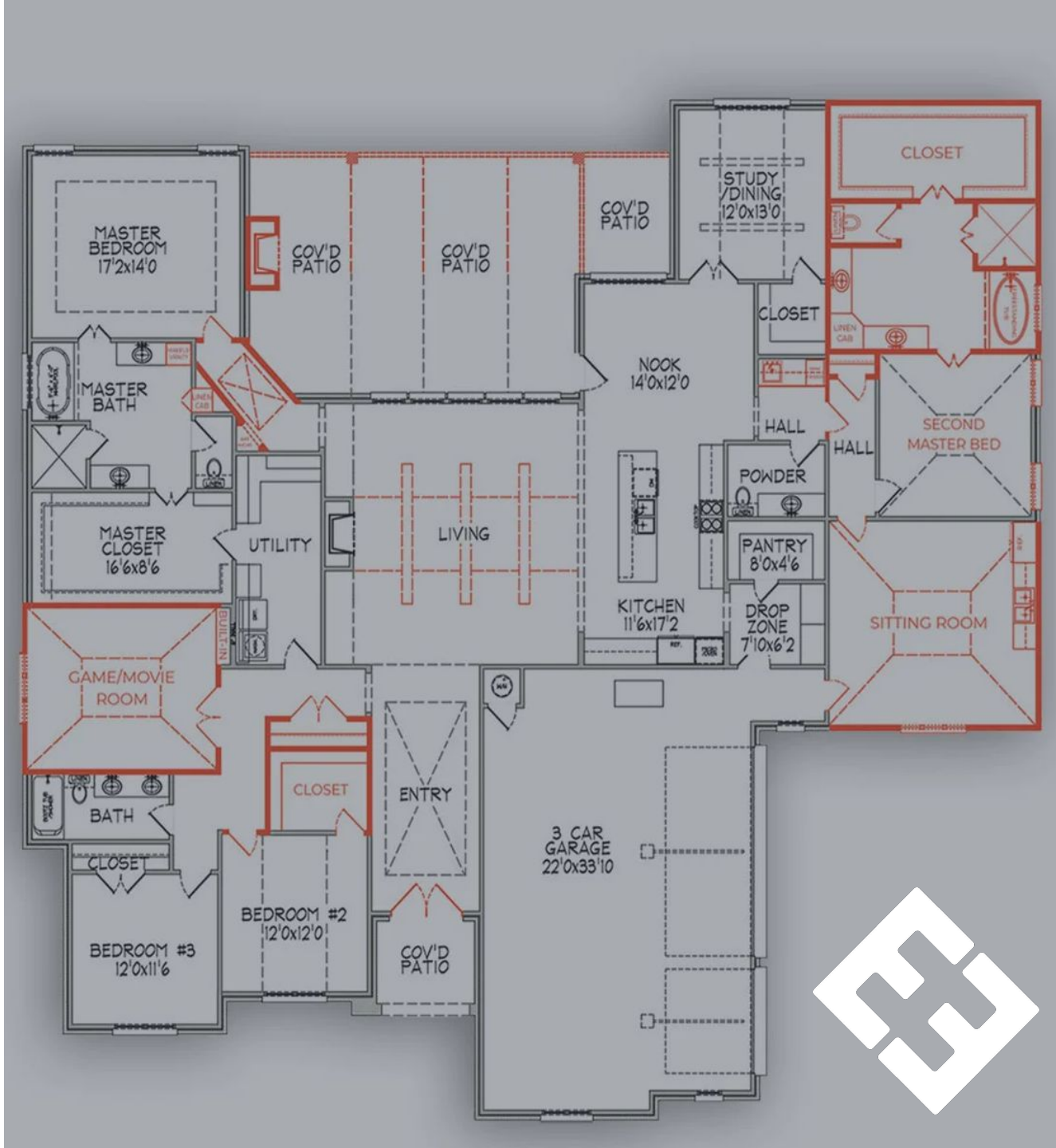
Lead Data Scientist

tommyz@executivehomes.com

Do the (obvious?) thing…

Agenda

- Who are we, and why should you care?
  - What does Executive Homes do?
- Why Trino, why here, why now?
- Why the connector architecture is awesome
- Data Integrations
- Any Questions?



Why Trino @Executive Homes

Maximizing for Flexibility With Connectors

What are we doing here?

01

Background & Why Trino?

02

Data Platform
Then

03

Data Platform
Leveled up

You're processing *how much data?*

Not much!

We're doing things a bit differently. Our largest table is less than ~2 million records, or about ~1.5 GB

Connectors, connectors everywhere!

The team at EH is optimizing for flexibility which drives cost savings. We can always scale later (clearly) if we need to.

Ergonomics

ANSI SQL is portable, simple, and broadly understood. We leverage external Hive tables, integrations with connections to app DBs, reporting infra, and everything in between.

A small narrative conceit…

Imagine if you will

A small-medium business with no “formal” data platform. There are multiple sources of truth (Google Sheets, Postgres, some data in S3, and maybe a vendor DB)

Sound Familiar?

Let’s see a show of hands for those of you that have dealt with this kind of messy “data platform”

Meet the business where it is.

Any solution we deploy for Data Science & Analytics needs to integrate with existing business processes, the less fuss the better.

Leveling up our Data Platform

We did the (maybe obvious) thing.

What tool is out there that integrates with 35+ systems, is easy to deploy, and provides a single point of entry? You guessed it…

It was cheap and easy, too.

We let Starburst do the work for us, and in about half an hour had deployed an X-Small cluster (which is frankly still overkill at our data volume) and integrated 3-4 data sources.

In the before time…

Things were a pain. In the now, things are awesome.

Data Integrations

If you can read it, you can ~~dream~~ process it

Doing the Obvious Thing

Working against a moving target (your data)

01

Read The Data

Set up your connection…

02

Manipulate the Data

Write some SQL

03

Write (or don’t) it out

GOTO 01

Let’s do something simple…

We have some data in Google Sheets

We’re tracking home inventory in a Google Sheet for our expansion into Stillwater, Oklahoma. It’s the source of truth for the business.

We need it in Postgres & S3…

For Machine Learning! We want to enrich this inventory data with some analytical data from a vendor in S3 and transactional data in PG. Then we’ll persist it to a second PG instance dedicated to Data Science.

Let’s spin up some infrastructure!

We’re going to need a few things:

- ~~Airflow, Python, Spark, …~~. ***Just say no!***
- ***A Trino (Galaxy) cluster***

How’s this going to all go down?

In the simplest way possible…

We’re moving thousands to tens of thousands of records around. We don’t need Spark or orchestration, just queries on a schedule (Thanks, Galaxy!)

Configure those connectors..

- Connect to Google Sheets & Postgres DB instances
- Create an unpartitioned external table in S3 for that vendor data
- Write a view combining all this stuff (exercise left for reader)

Persist your data and move on with life…

CREATE TABLE <PG_DSCI>.<ENRICHED_SCHEMA>.<FANCY_TABLE>

AS SELECT * FROM <SOME_CATALOG>.<VIEW_SCHEMA>.<FANCY_VIEW>

That's it?

Yes, thanks for coming to our talk!

It's obviously a bit more complicated like all things, but at our scale (and arguably up to quite a large scale) *you don't need more abstractions.*

You don't *actually* do this, right?

We certainly do! Trino, especially deployed on Starburst Galaxy, lets us do this with tons of flexibility & a minimum of infrastructure.

Some discussion..

This is a simplification, but not by much. Integration of small data is a fantastic use case for Trino. Have something as a DAG, *orchestrate your Trino cluster*

Thank you!

Contact:

ben@executivehomes.com

tommyz@executivehomes.com