Modern Data Stack

Clouds Everywhere!

Information Assets

Agenda

- Who is this guy?
- What is this stack?
- Why should I care?
- What are the pieces?
  - Extract/Load
  - Store
  - Transform
  - Visualize
- Demo

Anything else?

How Did I Get Into This?

Different Stuff

- Flooring
- Electrical Engineering
- Energy Efficiency
- Aquaponics
- Data Analytics
- Solo Consulting

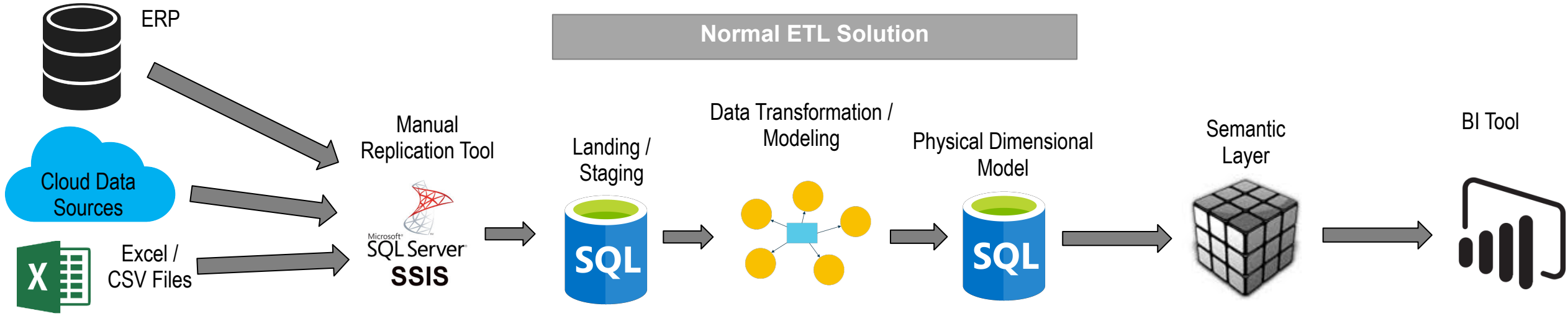
What is the “Modern” Data Stack story?

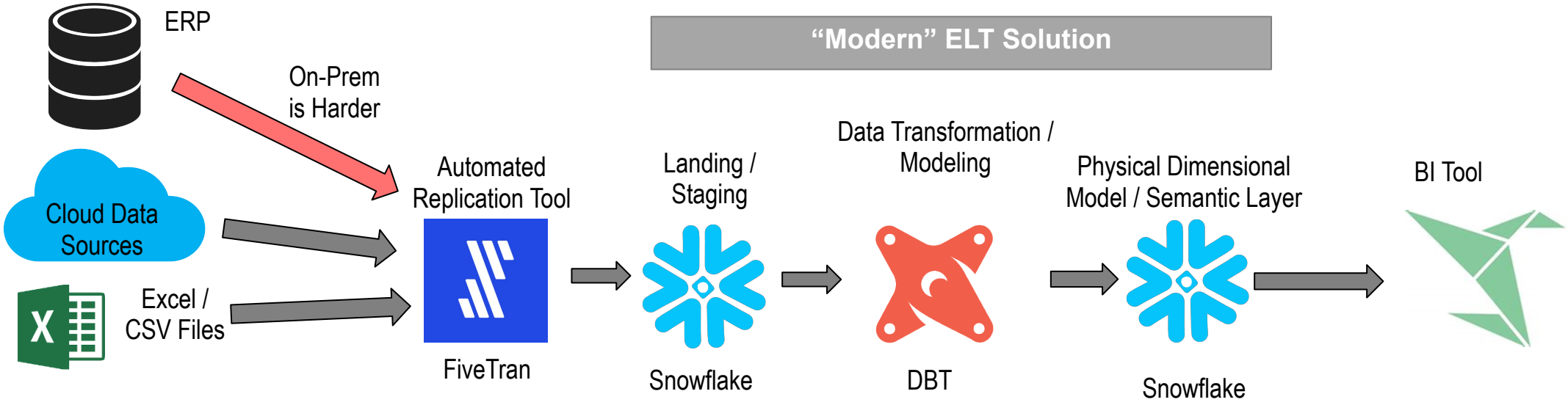
- **Cloud-Optimized Databases**
  - Auto-indexing
  - Auto-scaling
  - Must faster
- **Automation**
  - Replication
  - Testing
  - Deployments
- **Design Simplification**
  - Hashed Keys (avoids key lookups in some cases)
  - Semantic Layer
  - Expand complexity for larger data sizes

The Goal is Action

How does it compare?

Note that branded items are just examples. There are other brands who provide similar functions.





FiveTran

What is it?

- Automated Replication Tool

Advantages

- Idempotence
- Easy for native sources – a few clicks
- Handles CDC (assuming source keys)
- Includes orchestration tools
- Pricing at small scale
- Great documentation

Disadvantages

- On-prem data sources are harder to use
- Hard to set up non-standard APIs
- Limitations on real-time ingestion
- Pricing at scale

Snowflake

What is it?

- Columnar cloud database

Advantages

- Auto-indexing
- Auto-scaling (i.e. cheap to run)
- Handle simultaneous reads and writes
- Handles JSON data better
- Cloud Agnostic
- *Can reduce need for some Kimball best practices*

Disadvantages

- Query tools aren't as sophisticated
- Error reporting less complete

DBT (Fishtown Analytics)

What is it?

- Open source, code-driven data transformation engine

Advantages

- Can automate testing, version control, and deployments with Git
- Easier to create “models” (views), so analysts can participate
- Great user community and macro libraries

Disadvantages

- No auto-incrementing tables
- Limited experience at enterprise scale

Philosophy

1. code, not graphical user interfaces, is the best abstraction to express complex analytic logic.
2. data analysts should adapt similar practices and tools to software developers. (full post)
3. critical analytics infrastructure should be controlled by its users as open source software.
4. analytic code itself — not just analytics tools — will increasingly be open-source. (full post)

Sigma Computing

What is it?

- Spreadsheet-driven BI tool from same folks as Snowflake

Advantages

- Works kind-of like Excel does, using columns and aggregated columns as building blocks.
- Pushes SQL back to database.

Disadvantages

- Limited ETL functionality
- No small-customer pricing model.
- Less flexible than Tableau or PBI.

Demo Time

Contact Me

- jr@info-assets.com
- www.linkedin.com/in/jeremiah-robinson1

Demo

- FiveTran
- Snowflake
- DBT
- Sigma

Information Assets