Starburst

dbt Cloud & Starburst Galaxy hands-on workshop

Starburst Academy
v1.0.0

Connection Before Content

Lester Martin - https://about.me/lestermartin

- Educational Engineer @ Starburst
  - Build the content
  - Teach the class
  - Repeat
- 30 years of technology experience
  - Started my journey on a TRS-80 Model III
  - Played most every role, but consider myself a programmer at my core
  - Half of career in transactional systems and the second half in analytical processing
  - A DECADE of “big data” experience to include
    - Trino/Starburst, Hadoop, Hive, Spark
    - NiFi, Kafka, Storm, Flink
    - HBase, MongoDB

Workshop objectives

- Introduce dbt Cloud and Starburst Galaxy
- Review the data lakehouse reference architecture
- Present a data pipeline scenario
- Help YOU build it out across several labs
    - Create Starburst Galaxy account and data catalogs
    - Discover & review the land zone datasets
    - Create dbt Cloud account and connect to Starburst Galaxy
    - Build the structure zone with dbt Cloud models
    - Materialize the consume zone with a joined & aggregated dbt Cloud model
    - Define a Starburst Galaxy data product
    - Commit changes and create a production environment

https://www.getdbt.com/

What is dbt?

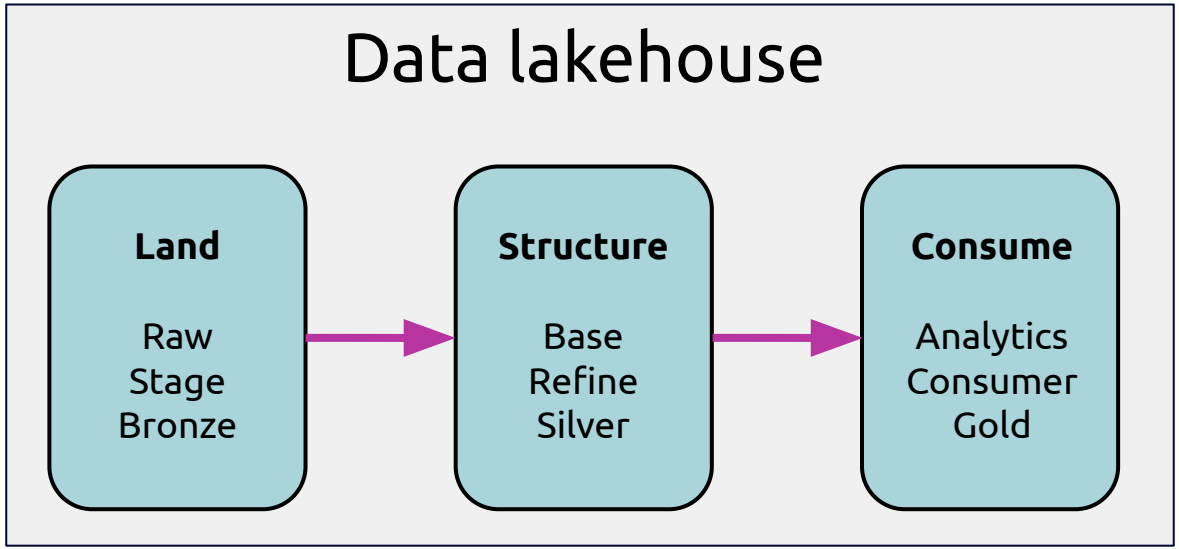
dbt™ is a SQL-first transformation workflow that lets teams quickly and collaboratively deploy analytics code following software engineering best practices like modularity, portability, CI/CD, and documentation. Now anyone on the data team can safely contribute to production-grade data pipelines.

Starburst Galaxy

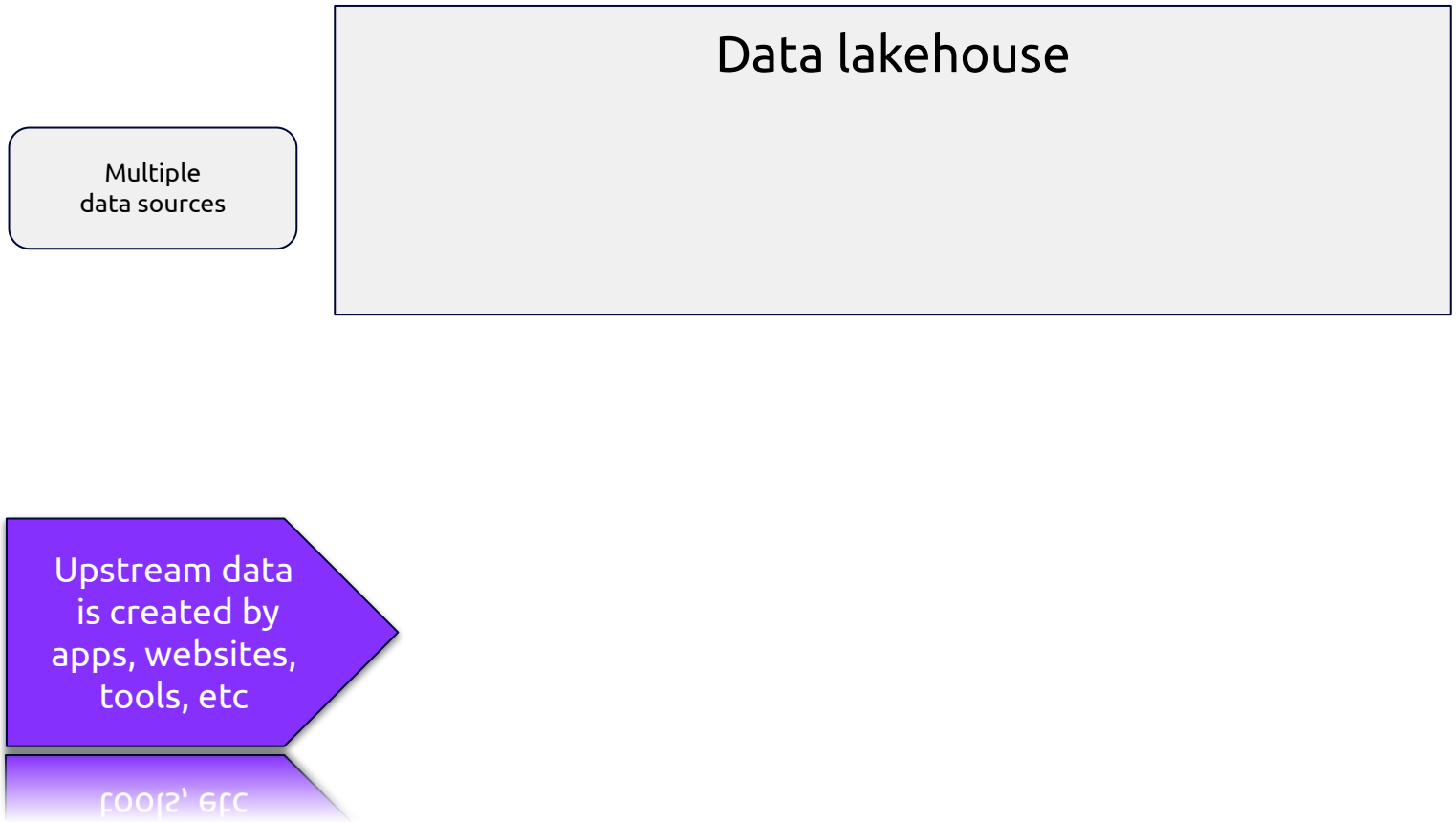
https://www.starburst.io/platform/starburst-galaxy/

Reference architecture

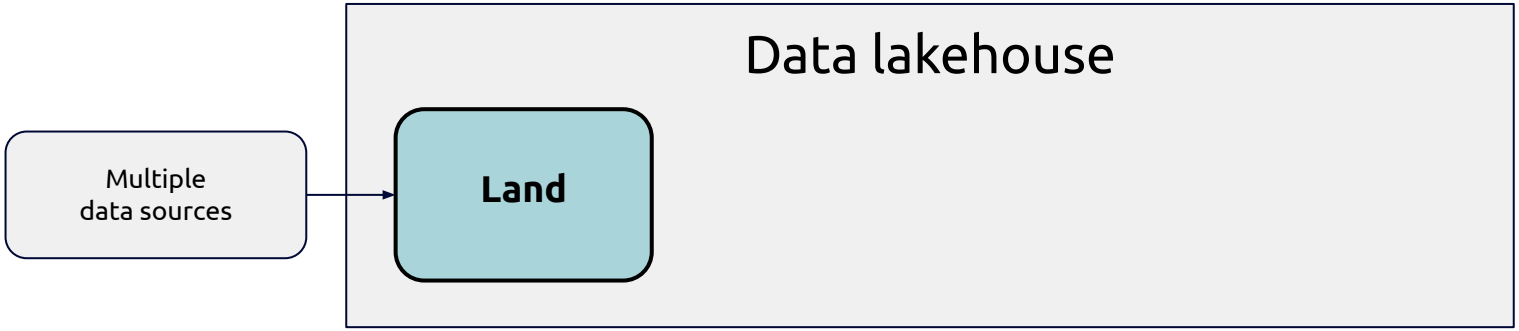
The reference architecture centers around the data lakehouse and how we classify our data assets into distinct zones. Data pipelines populate the zones.



Activities across the architecture

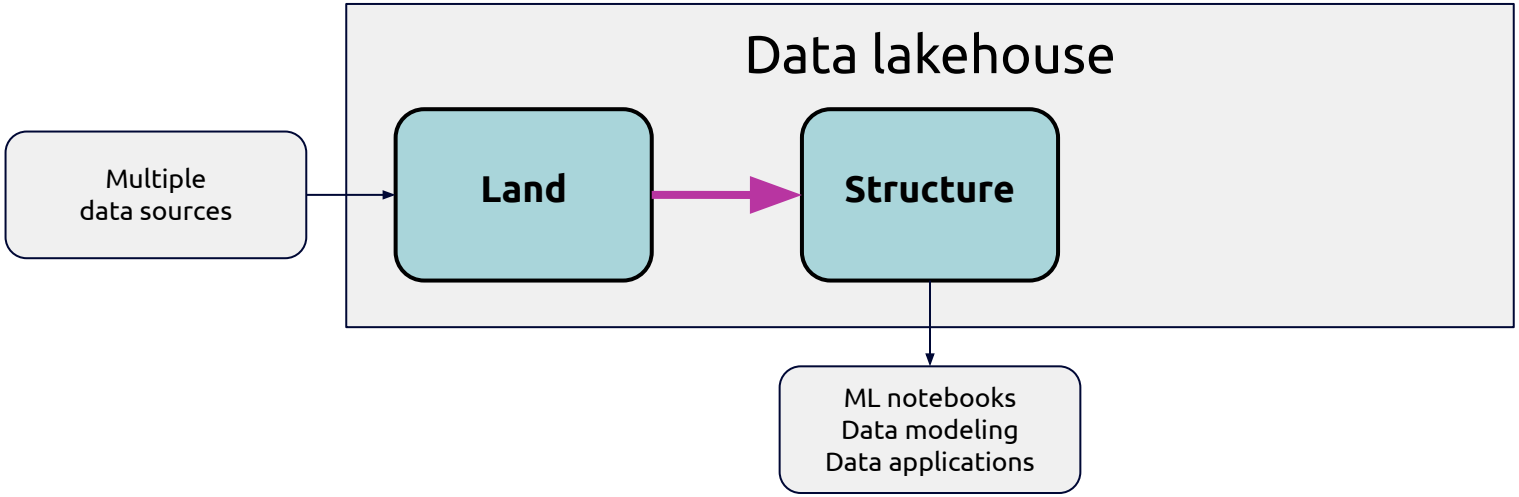


Activities across the architecture



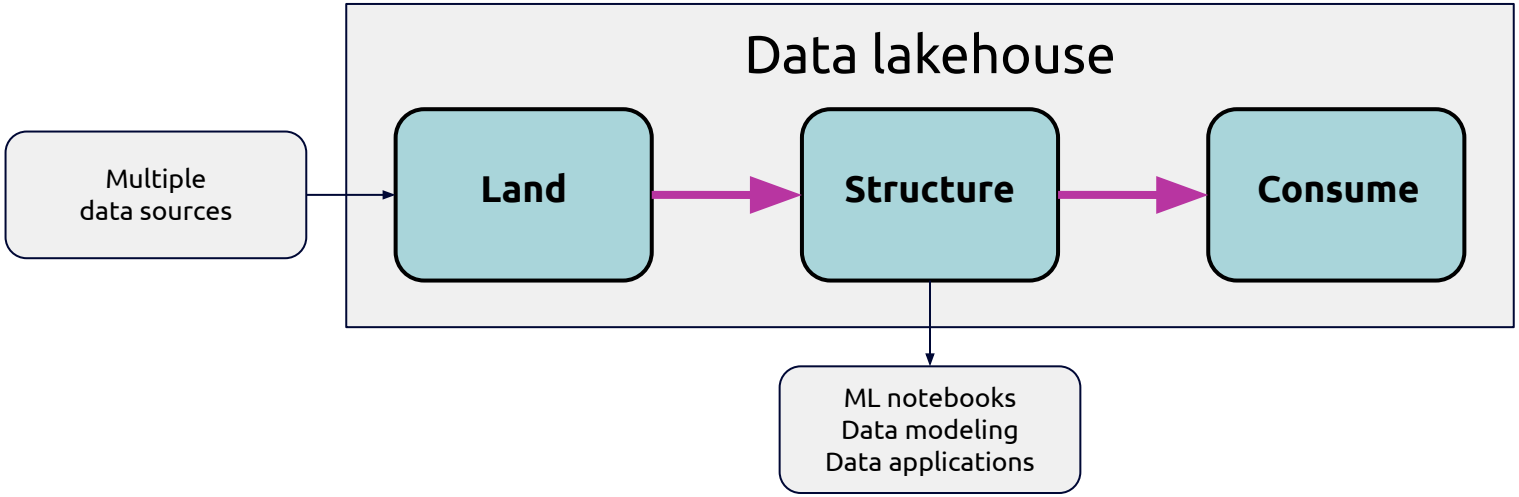
Starburst

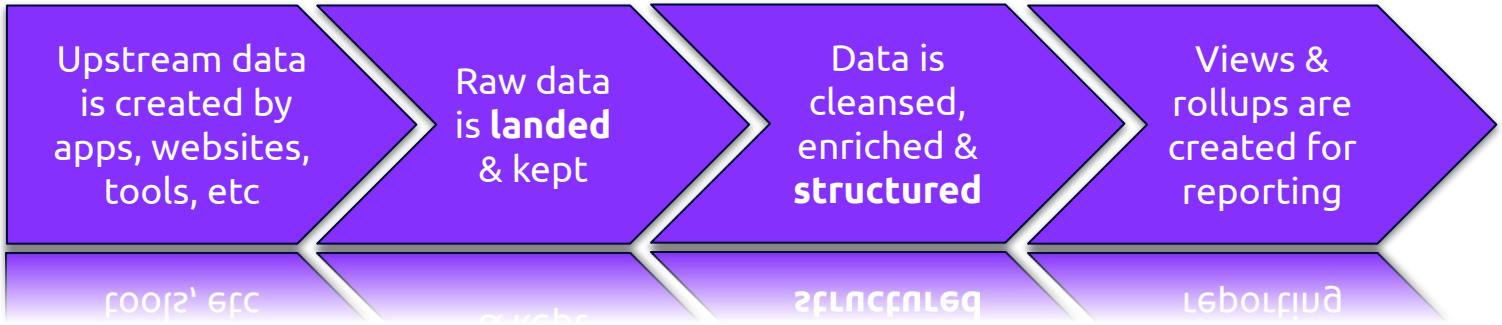
Activities across the architecture



Starburst

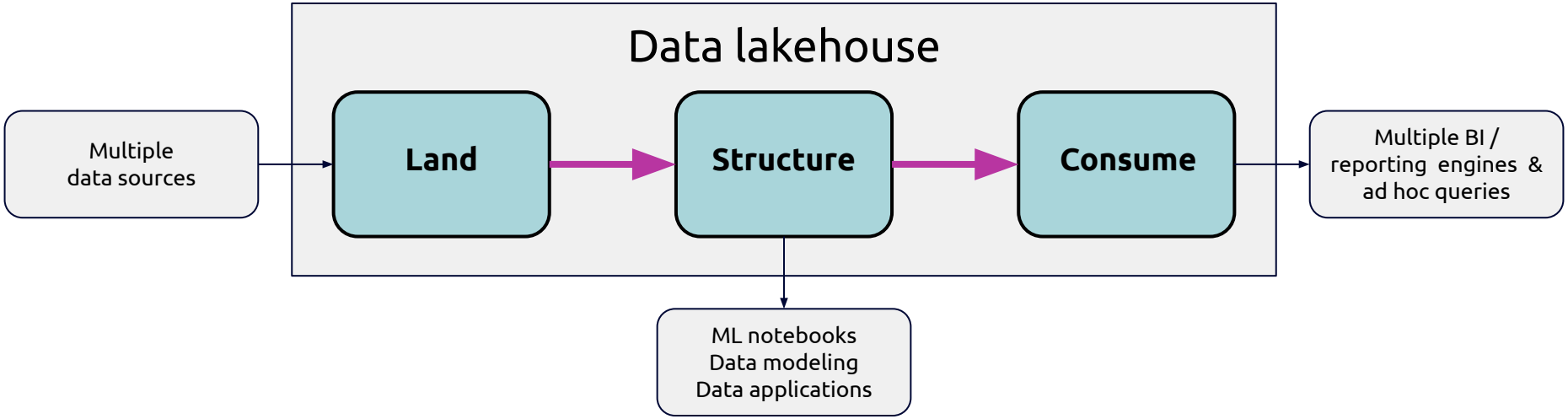
Activities across the architecture





Starburst

Activities across the architecture



Starburst

Workshop tools & technologies

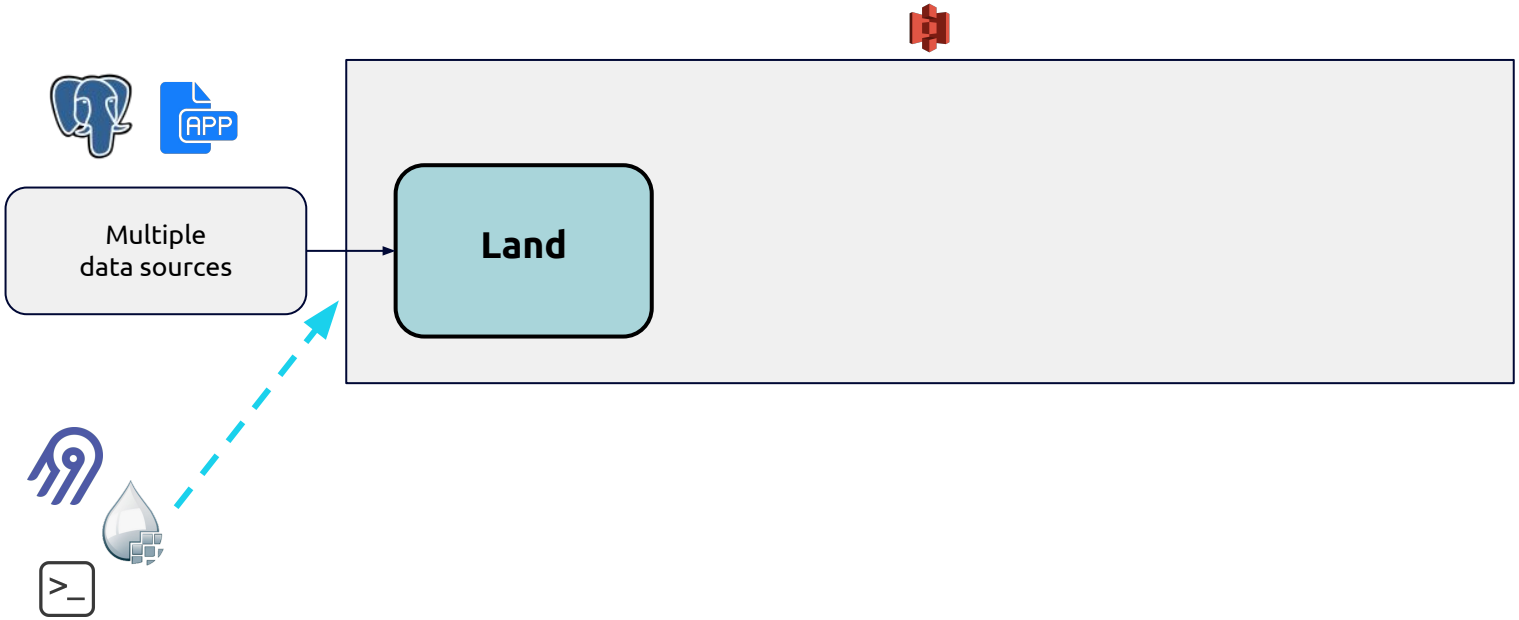
Data lakehouse

Starburst

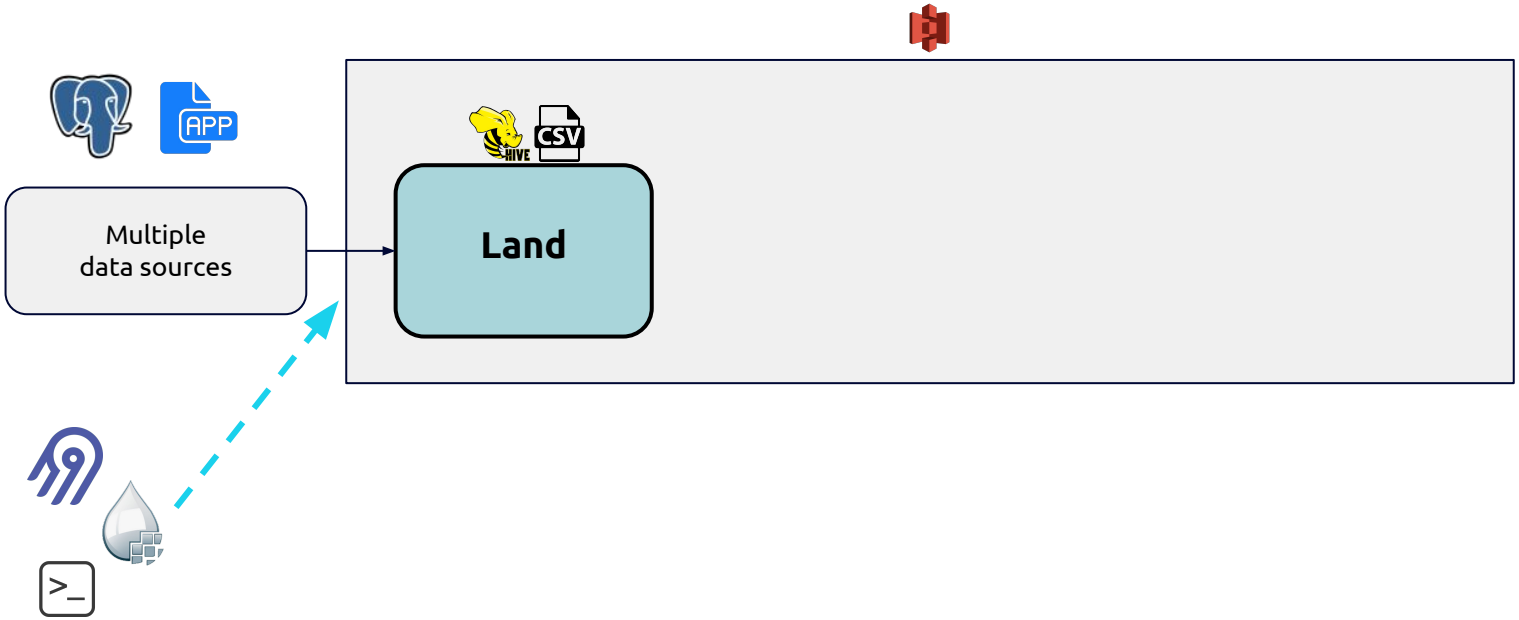
Workshop tools & technologies

Multiple
data sources

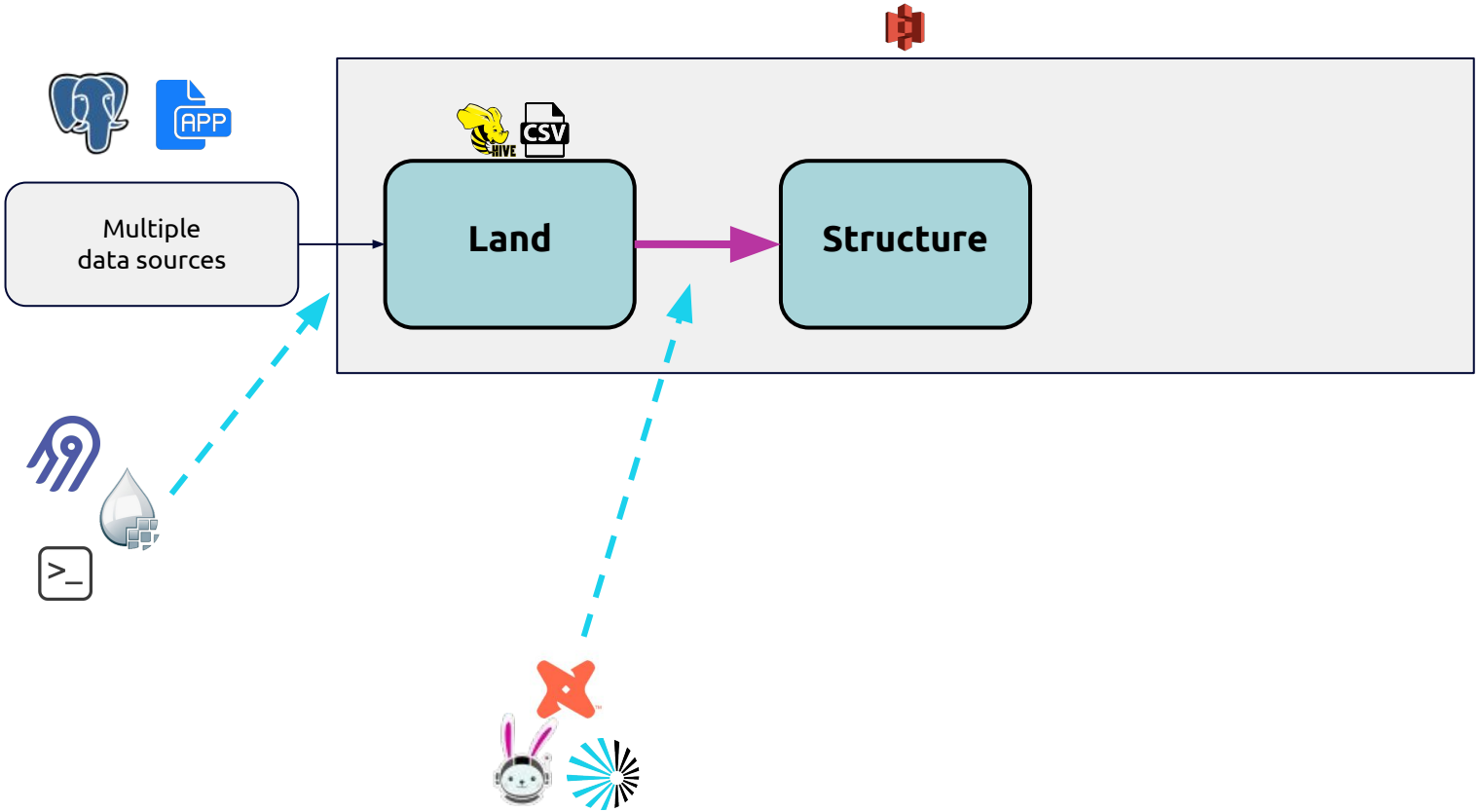
Workshop tools & technologies



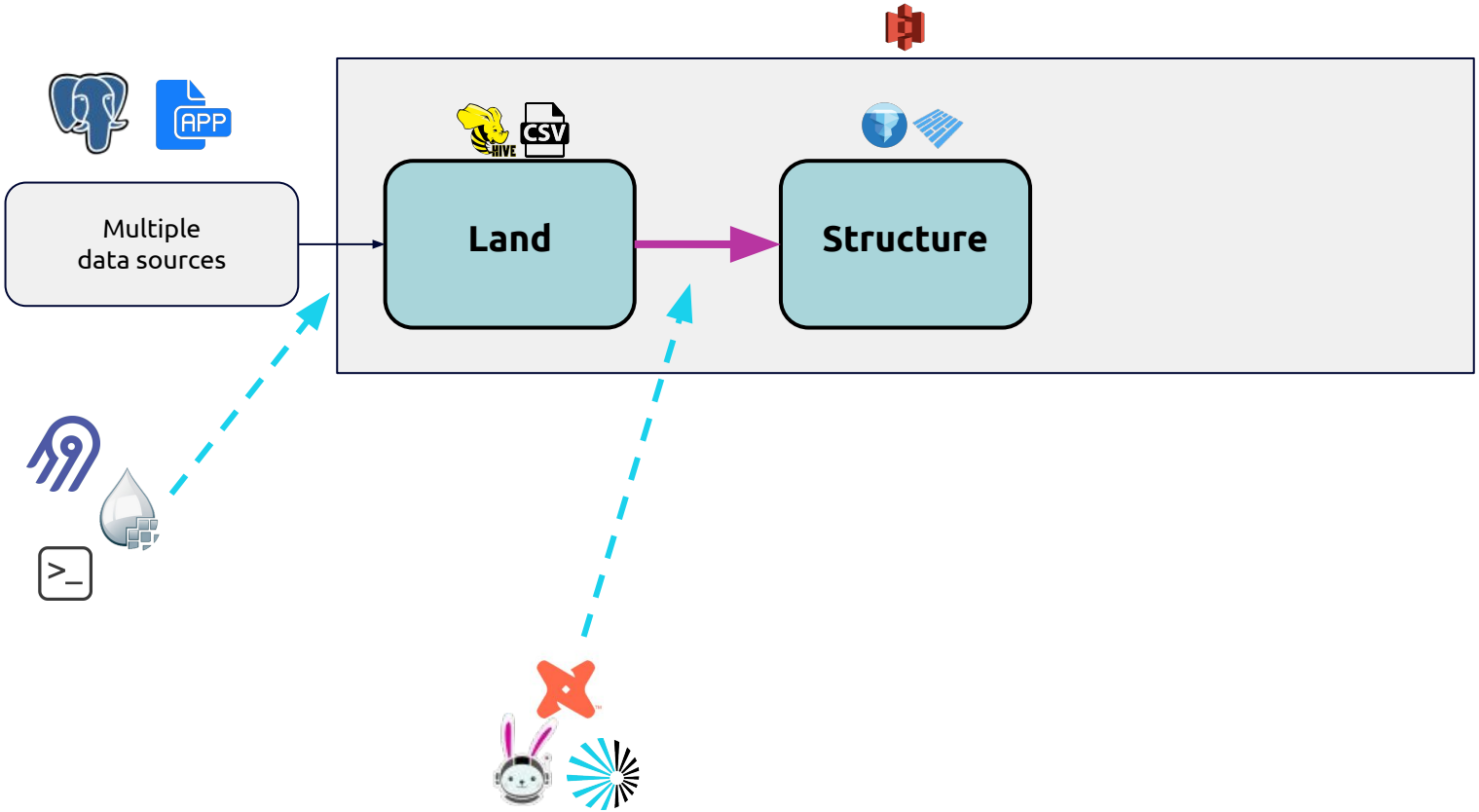
Workshop tools & technologies



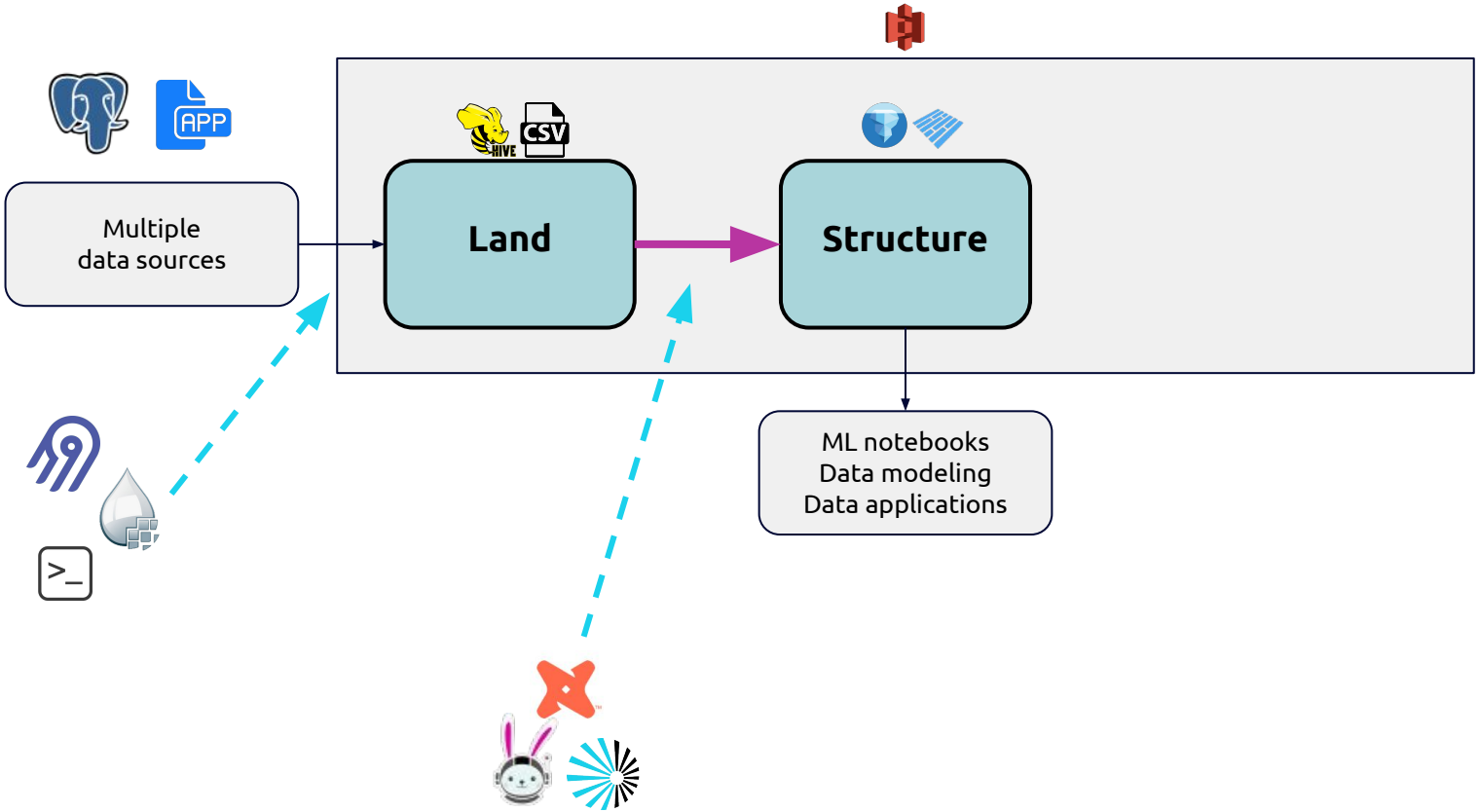
Workshop tools & technologies



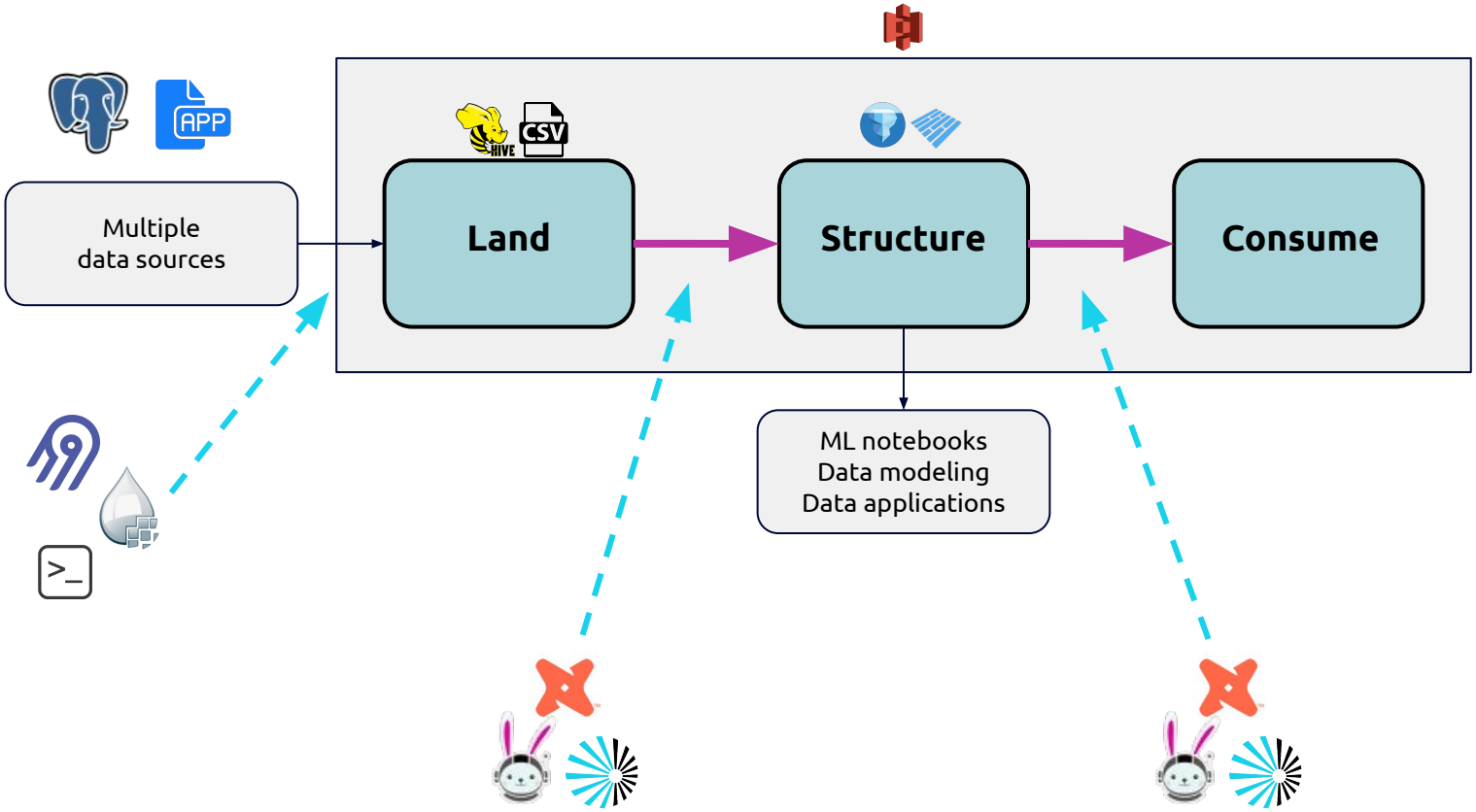
Workshop tools & technologies



Workshop tools & technologies

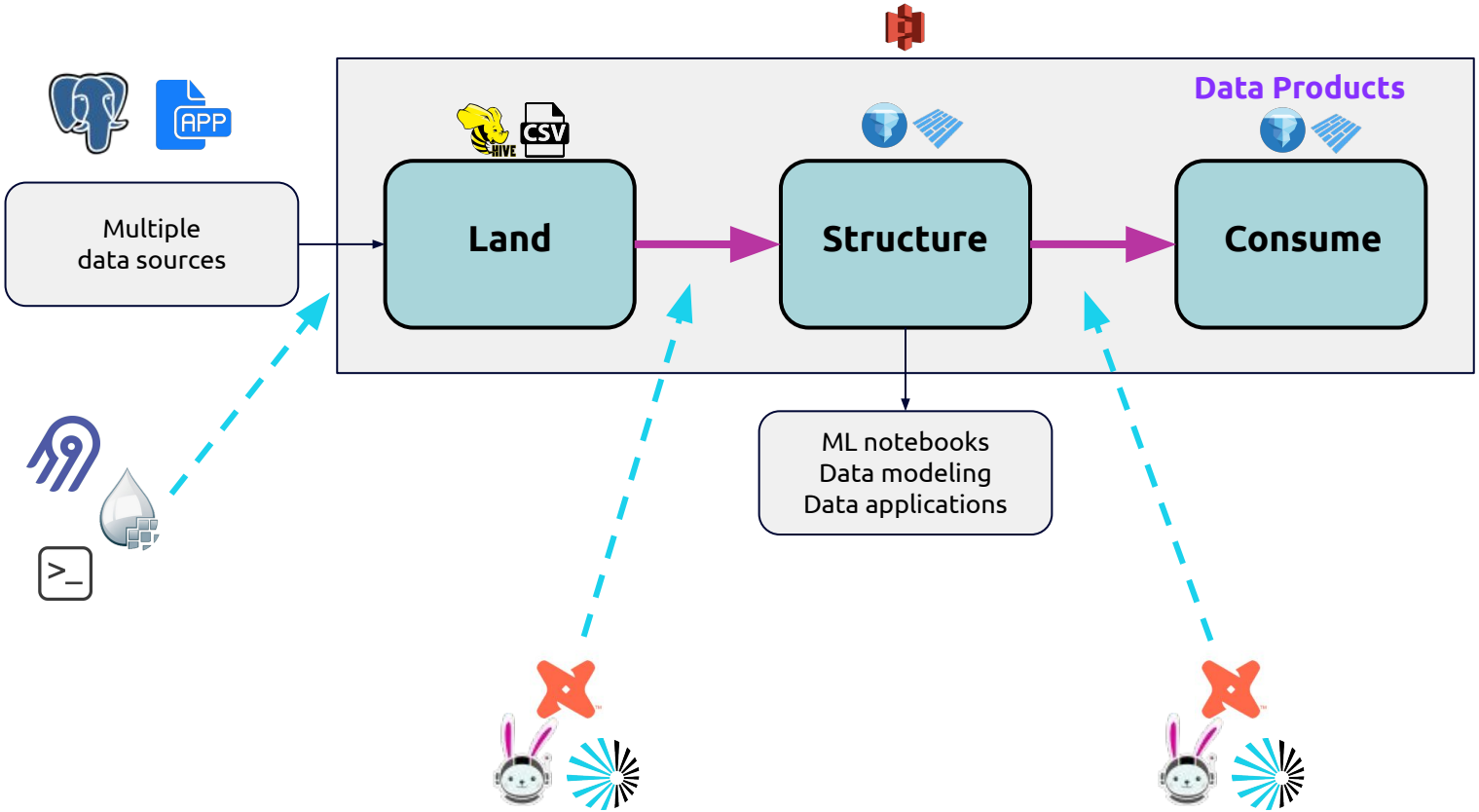


Workshop tools & technologies



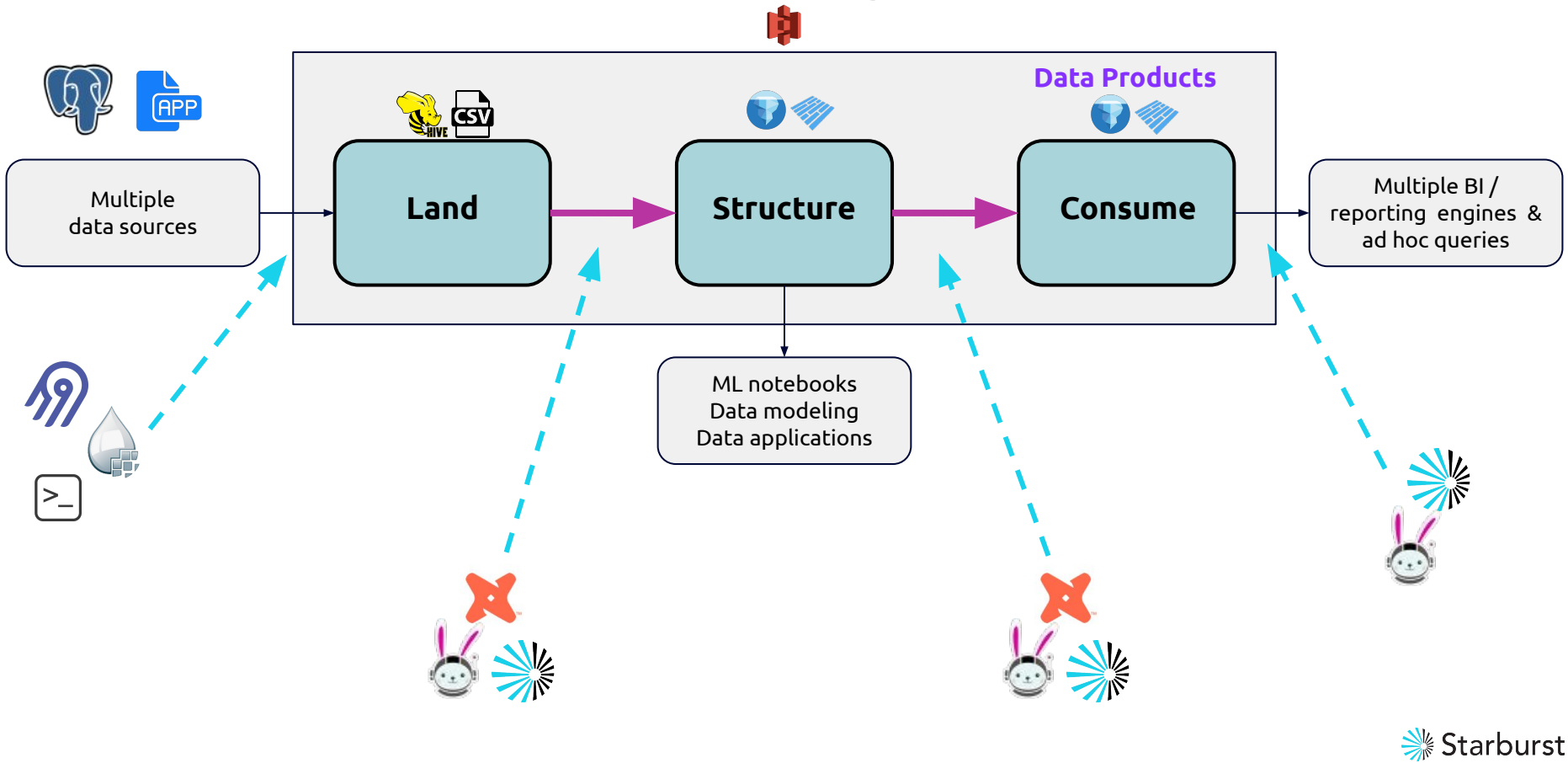
Starburst

Workshop tools & technologies



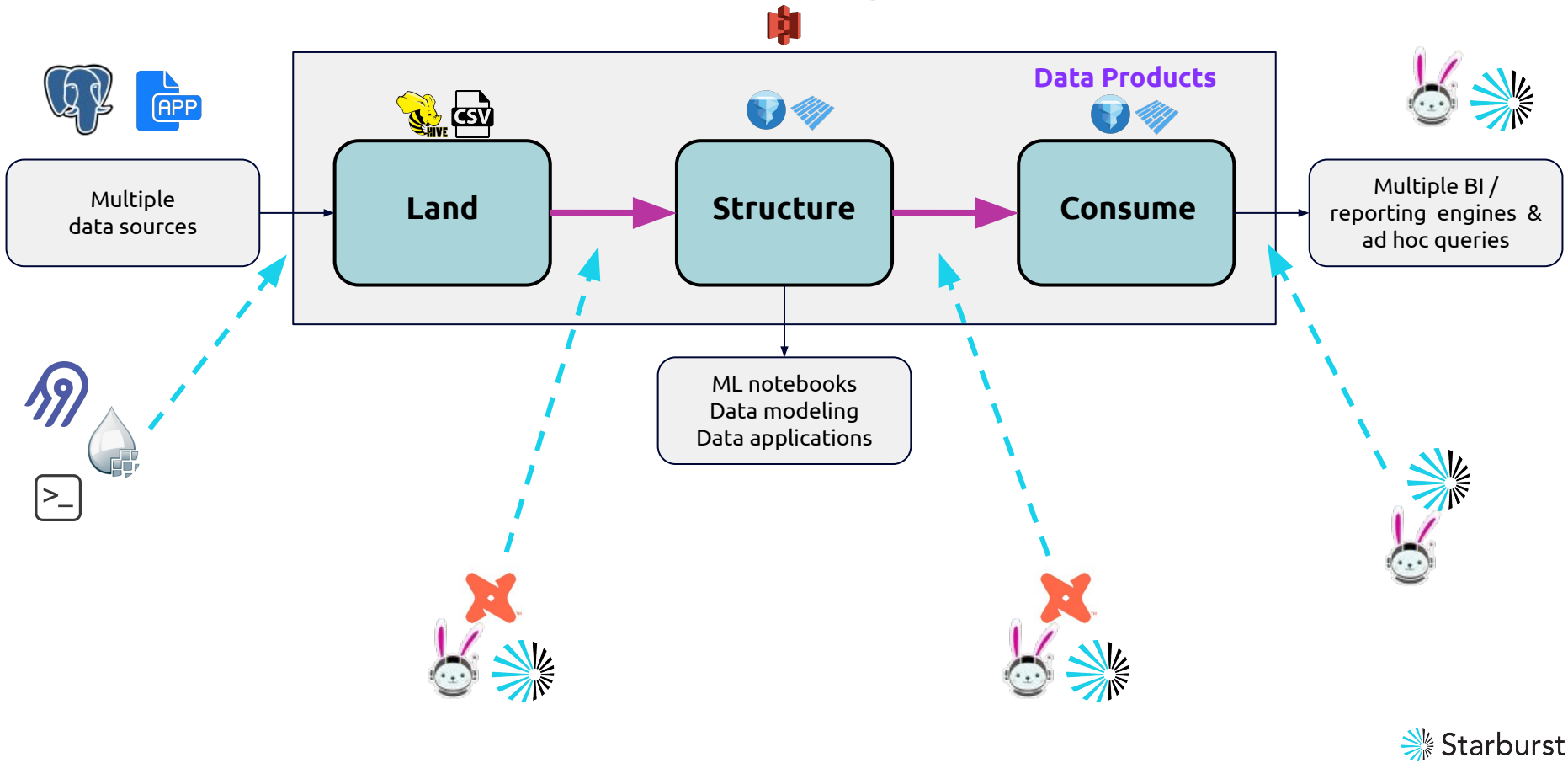
Starburst

Workshop tools & technologies



Starburst

Workshop tools & technologies



Pipeline scenario (dbt's Jaffle Shop)

- Leverage land zone datasets
  - Customers (PostgreSQL)
  - Orders (PostgreSQL)
  - Payments (Amazon S3)
- Validate, standardize & build structure zone datasets (Apache Iceberg)
- Define a consume zone model using joined/aggregated structure zone datasets *(View)*
- Verify correct land > structure > consume zone data lineage *(dbt Cloud & Starburst Galaxy)*
- Expose curated datasets as data products *(Starburst Galaxy)*

Starburst

HANDS-ON LABS!!

Starburst Academy