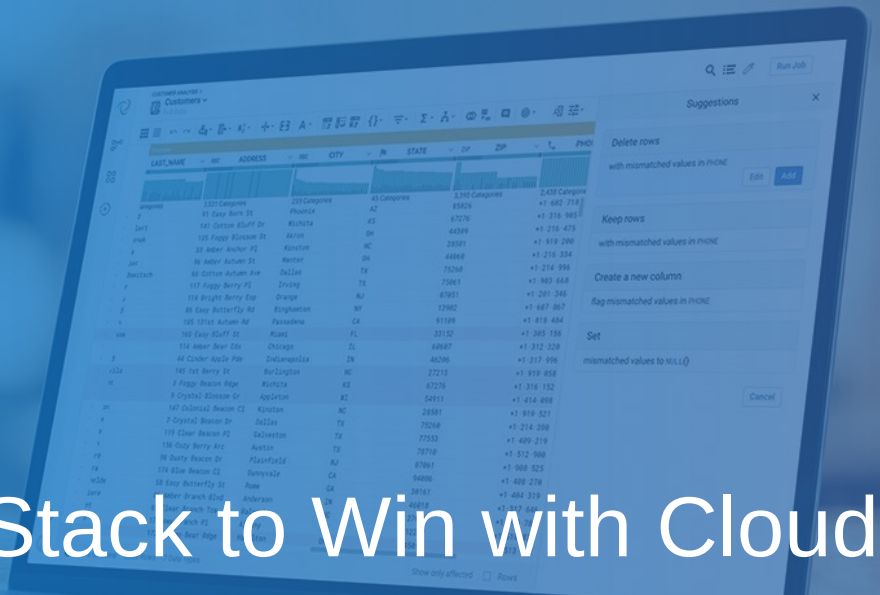
Fivetran

TRIFACTA

aws



The Modern ELT Stack to Win with Cloud Data Warehousing

January 28, 2021

Today's Speakers

Will Davis
VP of Marketing

TRIFACTA

TJ Holsman
Partner Engineer

Fivetran

Vijay Balasubramaniam
Director, Partner Solutions Architect

TRIFACTA

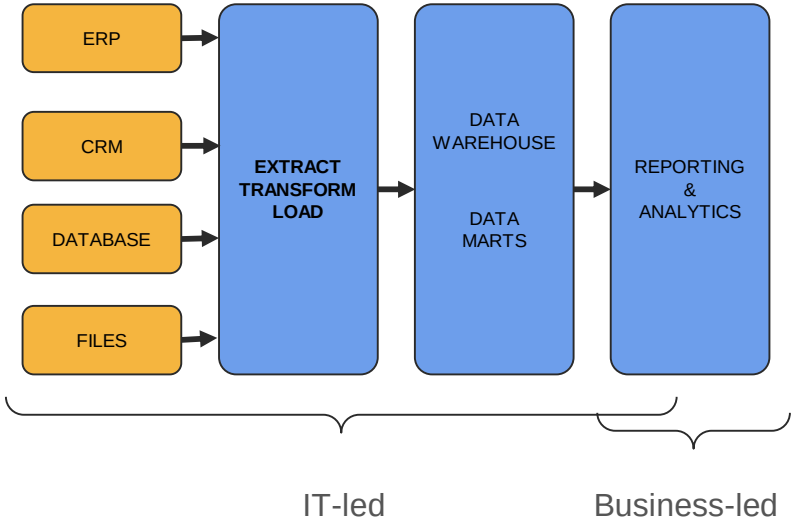
Greg Khairallah
Director of Analytics

aws

Agenda

- Rise of Cloud Data Warehousing
- Modern ELT Stack Overview
- ELT in the Wild!
- Demo - ELT for Marketing Analytics
- Q&A

Traditional Analytics Process with ETL

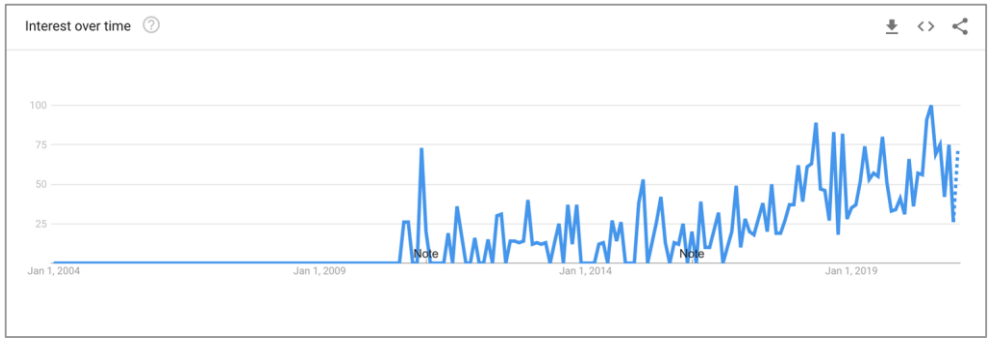


Challenges with ETL

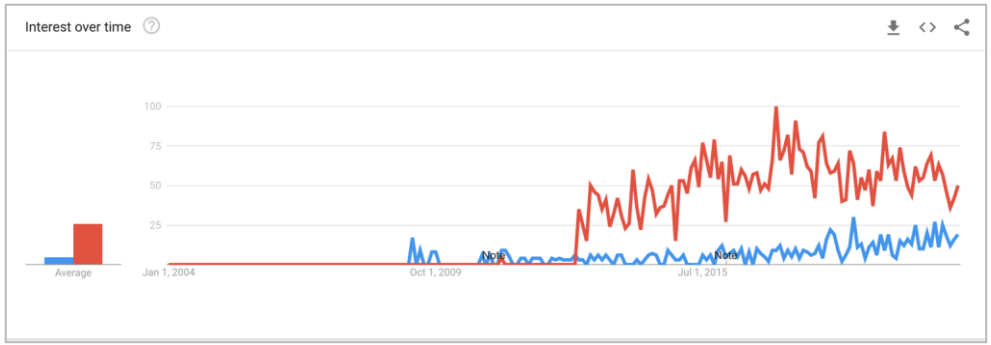
- **Rigid** - hard to adapt to changing requirements/data
- **Siloed** - typically IT-led tools...not exposed broadly
- **Technical -** not designed for data analysts & scientists

The Rise of the Cloud Data Warehouse

Search Volume - Cloud Data Warehouse (CDW)



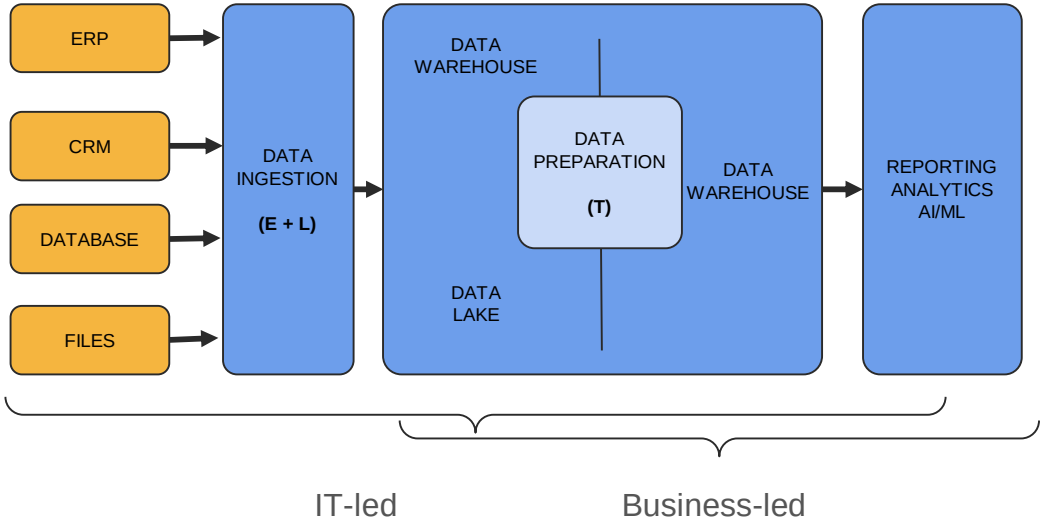
Search Volume - Amazon Redshift (red) compared to CDW (blue)



"Traditionally, data was extracted, transformed, then loaded – ETL, in short – into a data warehouse. For ETL, complex transformation pipelines were built at the data source. However, cloud data warehouses have finally made it cost-effective to store all of a company's data in a central location: we no longer need to transform data before we load it into a data warehouse. Transformation can be done when running analytics in a data warehouse."

- **Martin Casado, Andreessen Horowitz**

Modern ELT Stack for Cloud DW



Why ELT?

- **Flexible** - can transform data on-the-fly to meet requirements
- **Collaborative** - fosters collaboration between data engineers & analysts
- **No-code/low-code -** empowers a variety of users to do this work

The Modern ELT Stack

DATA INTEGRATION

Automate data integration from source to destination

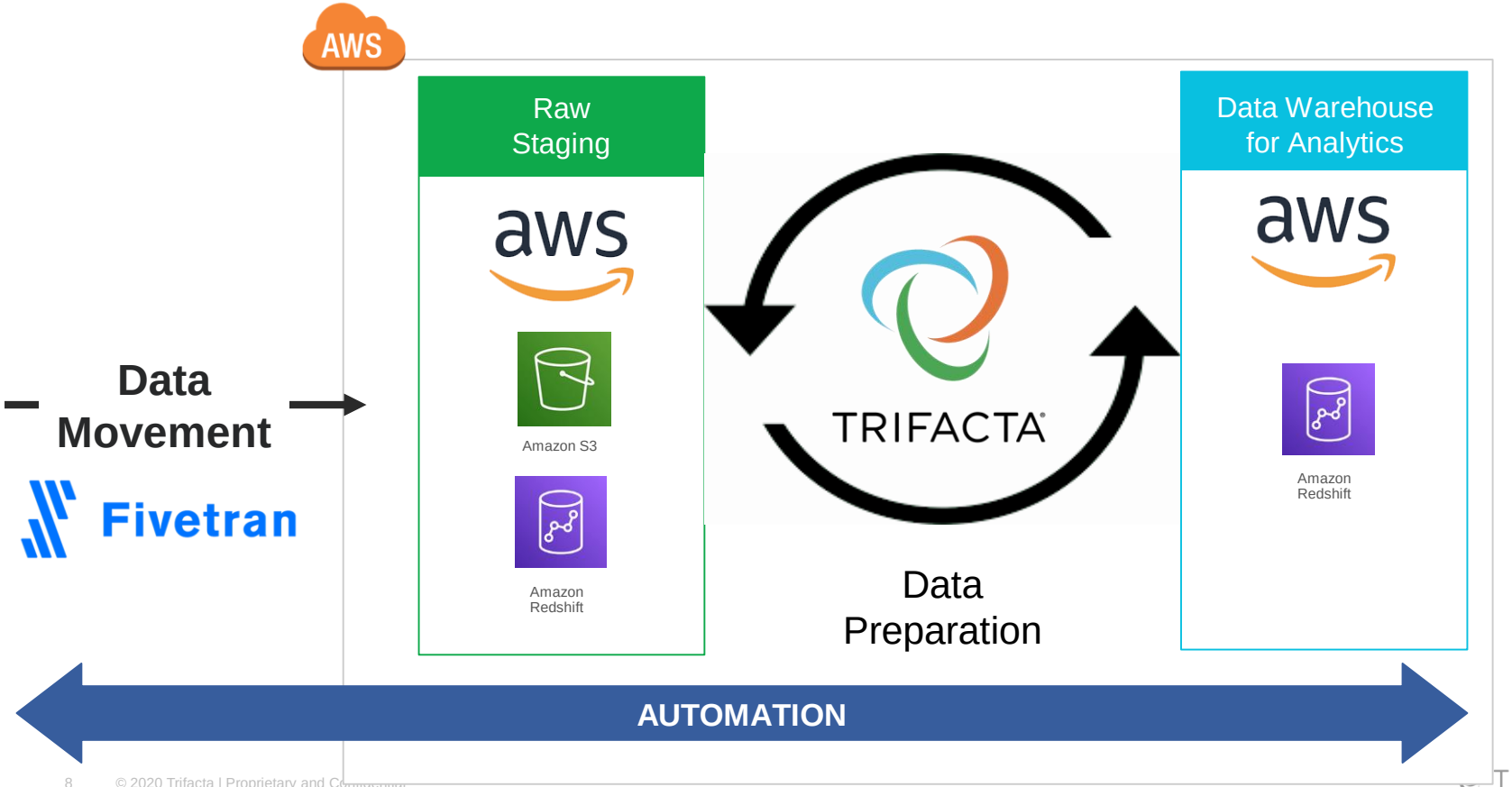
DATA PREPARATION

Explore, clean & blend data for use in analytics

DATA WAREHOUSE

Centralized data warehouse for reporting & analytics

Analytics Workflow in the Cloud

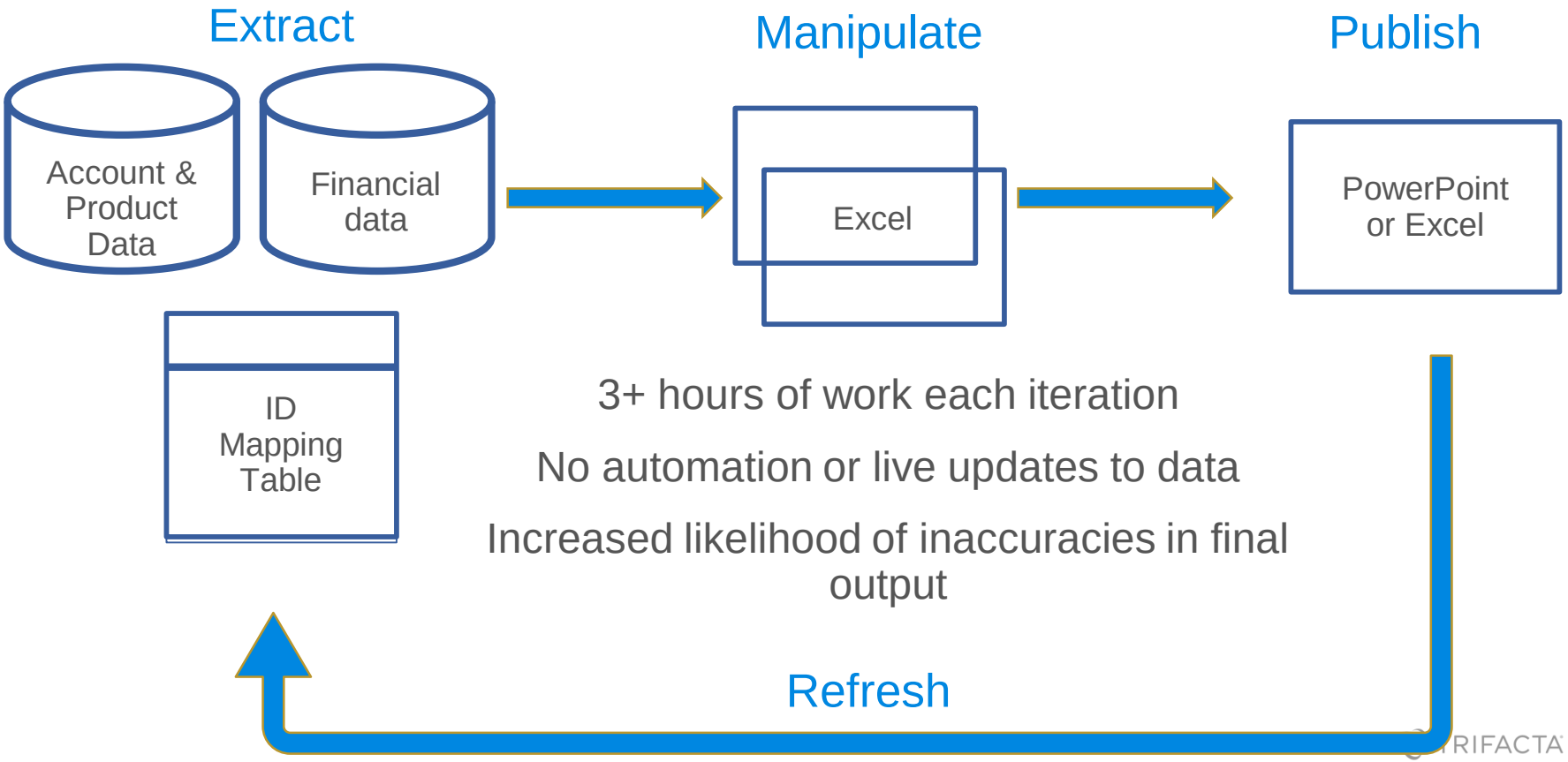


ELT in the Wild

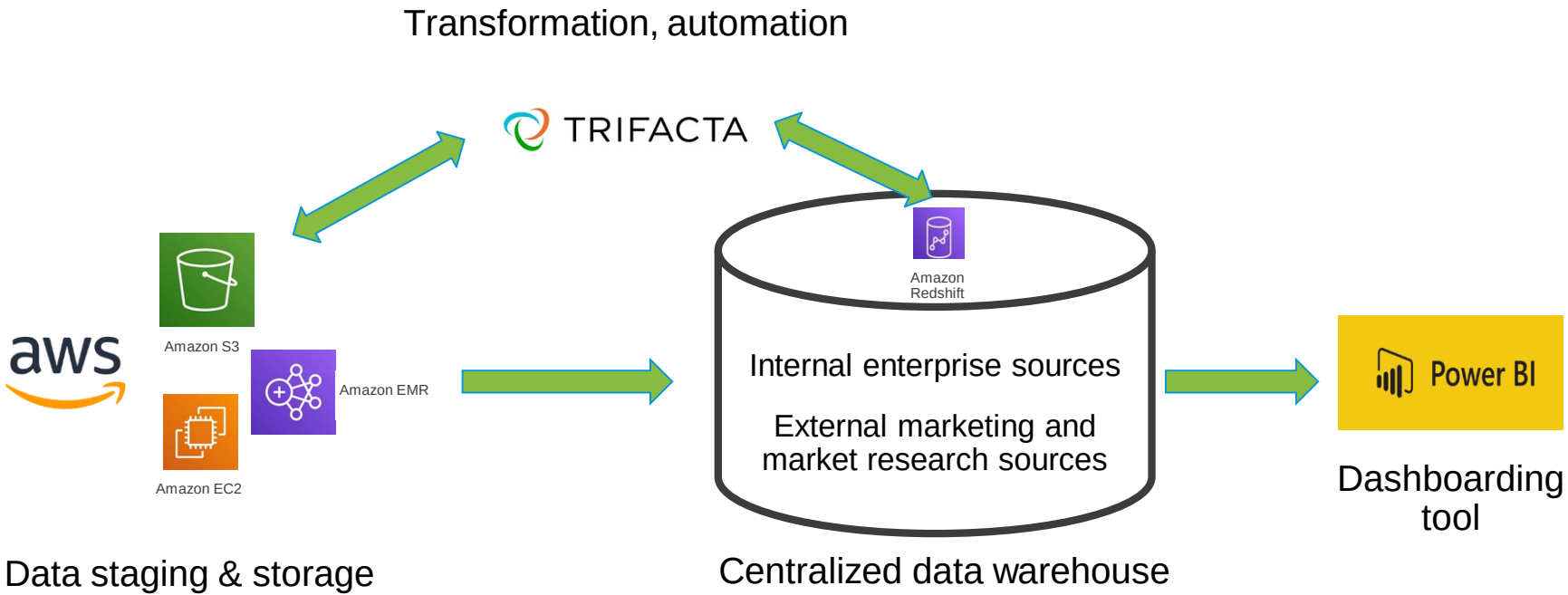
AUTODESK®

Make anything™

Data Challenges at Autodesk

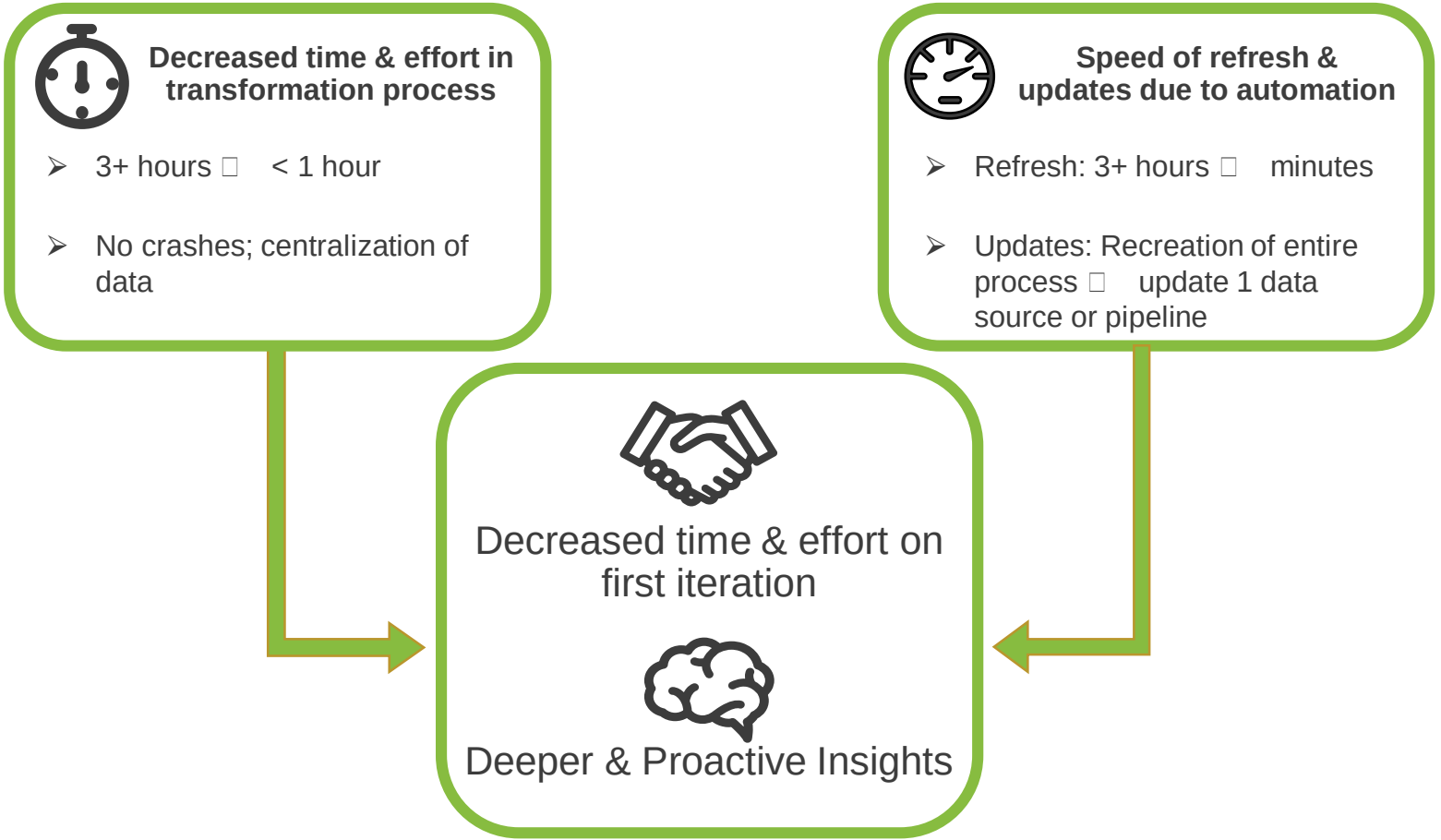


ELT at Autodesk



ROI at Autodesk

AUTODESK.




allahar
A community
of thinkers & makers

Data Challenges at Callahan

callahan

- Lots of data coming from different sources and in different formats
- Lots of dirty data that requires constant cleansing before it can be used or reported on
- A team of analysts that need to be able to ask a lot of questions of the data very quickly
- Demanding client-base who needs to be able to understand and communicate results fast
- Often tackling problems we have never run into before, and where there is no playbook to refer back to
- No database admins or data engineers on staff

ELT at Callahan

callahan

DATA SOURCES

- PoS & eCommerce
- Advertising & web data
- CRM & other databases

SAP teradata. ORACLE shopify Microsoft SQL Server LinkedIn Jira Software INTERCOM HubSpot salesforce servicenow

DATA TRANSFER

- Custom APIs
- Fivetran
- Trifacta

Fivetran

DATA STORAGE

- Cloud DW
- Cloud Storage

Cloud DW

Cloud Storage

DATA PREP

Trifacta
- SQL

TRIFACTA

VISUALIZATION

- Tableau

Tableau

DATA OVERLAYS

- NOAA Weather
- ESRI Location Science
- US Census & Labor Data
- etc.

MODELING

- Python
- R

python R

ROI at Callahan

- Fundamentally changed the way Callahan does business, in a competitively advantageous way
  - Has kept time spent on setting up and managing data pipelines to less than 30% of overall time spent on projects, allowing analysts to spend more than 70% of their time on analysis
- Brought extreme value to our clients in terms of improved business results, and cost efficiencies
  - **Media Result: 90% improvement in media impact, on a 50% reduction in media budget**
  - **CRM Result: 2x increase in customers, on a 60% reduction in leads purchased**
  - **Sales Result: 5% sales improvement during peak periods with ability to predict inventory out of stocks 3 weeks in advance**

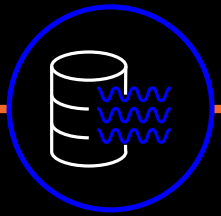
Demo

Amazon Redshift

Amazon Redshift

THE MOST WIDELY USED CLOUD DATA WAREHOUSE, WITH TENS OF THOUSANDS OF CUSTOMERS

ANALYZE ALL YOUR DATA

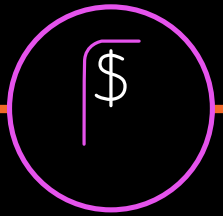


Take a **lake house approach** by analyzing all your data across your data warehouse, your Amazon S3 data lake, and operational databases with consistent security and governance policies

PERFORMANCE AT ANY SCALE

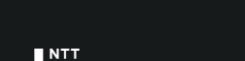
Get up to **3x better price performance** than other cloud data warehouses with a **self-tuning** system, boost queries up to **10x with AQUA,** and achieve <1s latency with materialized views

LOWER YOUR COSTS



Start small and pay only for what you use with **predictable** monthly costs; Amazon Redshift is **50% less expensive** than other cloud data warehouses

Tens of thousands of customers process exabytes of data with Amazon Redshift daily



NTT DOCOMO

Moved >10 PB of data from on-premises to cloud

FOX Corp.

Taking a lake house approach with RA3 nodes and Amazon S3

Yelp

Enabling a data-driven organization with concurrency scaling

Jack in the box

Improved ops by moving off of on-premises DW

Warner Bros. Games

Performance, scale, cost-effective

Amazon Redshift innovates to meet your needs

Analyze all your data

Lake house with AWS integration

- NEW! — Amazon Redshift ML
- NEW! — Data sharing
- NEW! — SUPER data type with JSON support
- UPDATED! — Federated Query
- NEW! — Lambda UDF
- NEW! — Partner console integration
- NEW! — Materialized Views via AWS Glue Elastic Views
- Amazon Redshift Spectrum + Lake Formation
- Data Lake Export

Performance & scale

Fast and self-tuning

- UPDATED! — RA3 nodes & managed storage
- COMING SOON! — AQUA
- NEW! — Performance tuning: automated
- UPDATED! — Materialized views with auto refresh & rewrite
- NEW! — 100K tables
- NEW! — HyperLogLog
- Concurrency scaling

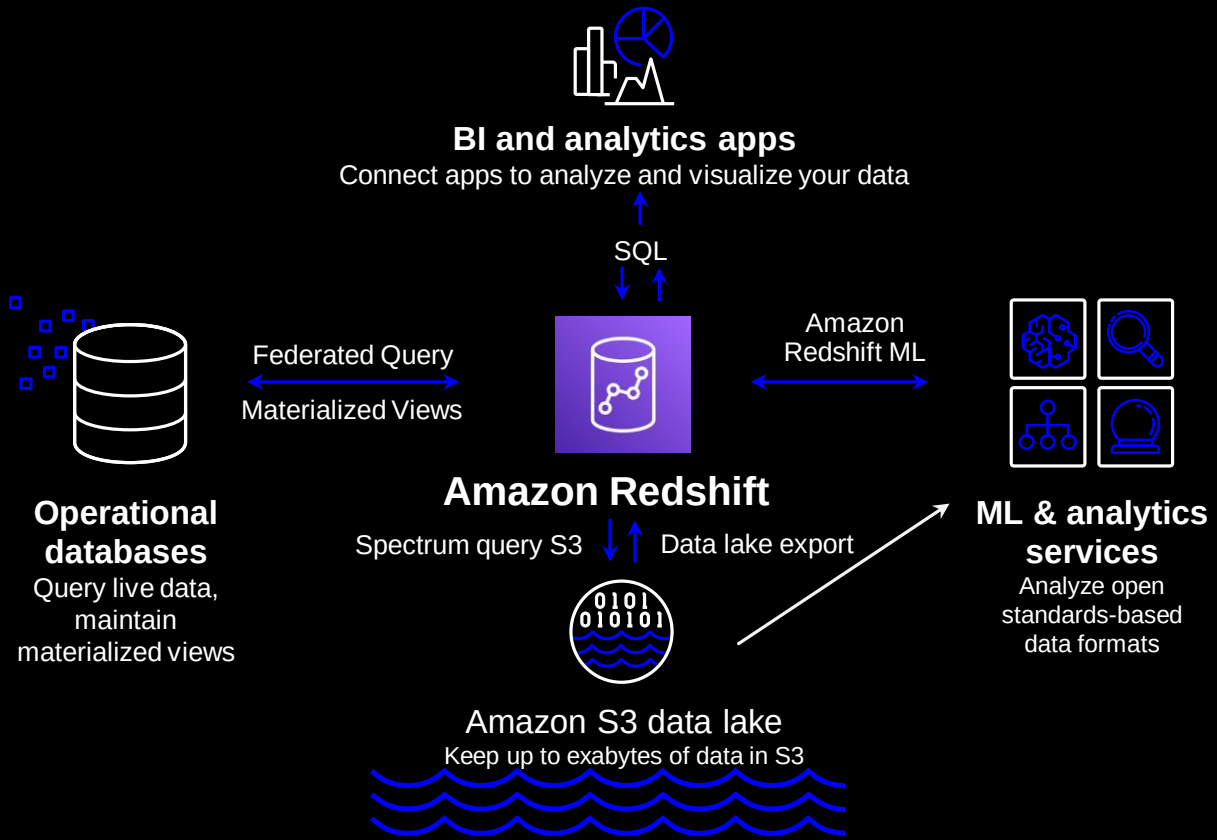
Low cost & best value

Predictable costs

- UPDATED! — Automatic workload manager
- NEW! — Cross-AZ cluster recovery
- NEW! — Data API
- On-demand and RIs
- Pause and resume
- Cost controls
- Built-in security features

Analyze all your data

WITH A LAKE HOUSE APPROACH TO ANALYTICS

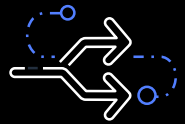


RA3 nodes with managed storage

SCALE COMPUTE AND STORAGE INDEPENDENTLY

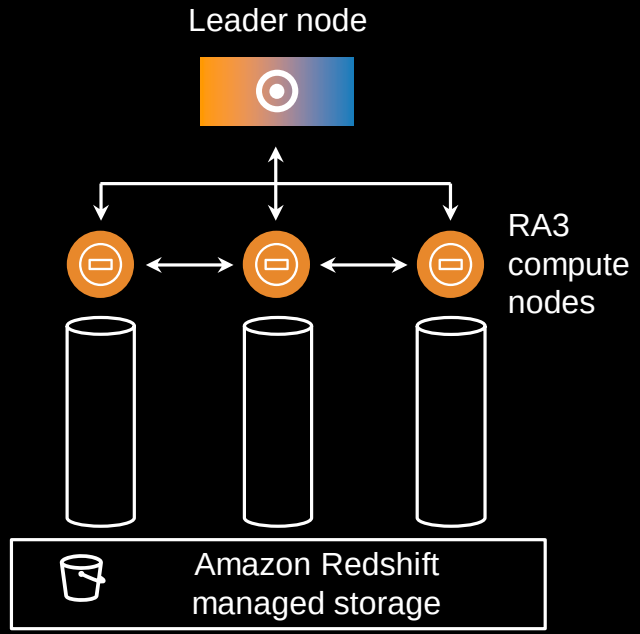
Managed storage

Large high-speed cache



High-bandwidth networking

- Size of data warehouse only based on steady state compute needs
- Scale and pay independently for compute and storage
- Automatic, no changes to any workflows, no need to manage storage



Concurrency scaling

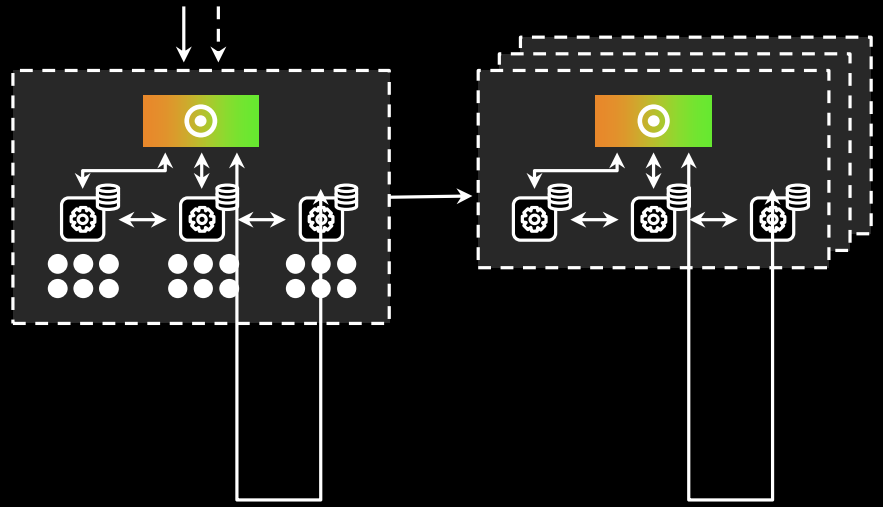
COMPUTE ELASTICITY AND SCALABILITY TO HANDLE UNPREDICTABLE USER DEMAND

Scale out to multiple Amazon Redshift clusters from a single endpoint in seconds

Support virtually unlimited concurrent users and queries while maintaining SLAs

Per-second billing for additional clusters used

Cost controls and free one-hour usage per day



Amazon Redshift automates performance tuning

NEW

ML-BASED OPTIMIZATIONS TO GET STARTED EASILY AND GET THE FASTEST PERFORMANCE QUICKLY

Automates physical data design and optimization

Optimizes for peak performance as data and workloads scale

Leverages machine learning to adapt to shifting workloads

Automated performance tuning

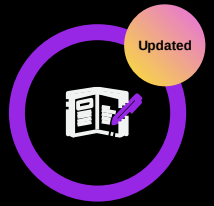
Automatic vacuum delete

Automatic distribution keys

Automatic sort keys

Auto workload manager

Automatic table sort



MV auto-refresh and rewrite

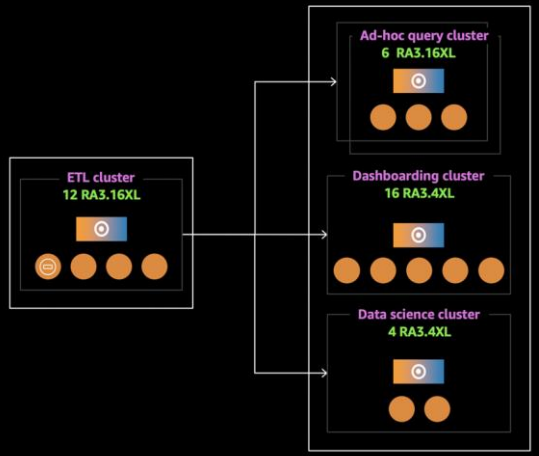
"When we tested ATO in our development environment the performance of our queries was 25% faster than our production workload not using ATO, without requiring any additional effort by our administrators."

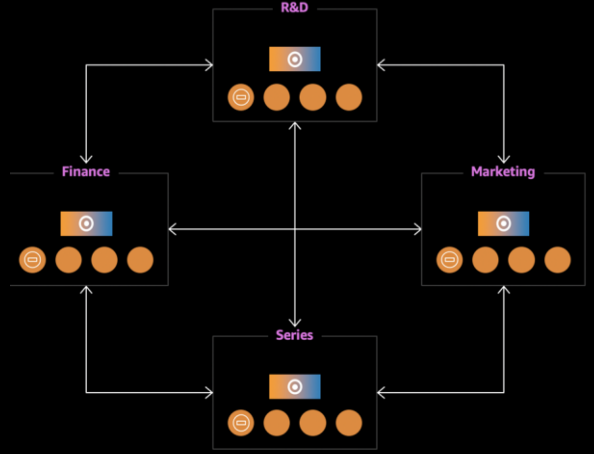
Nishesh Aggarwal, Enterprise Architecture Manager, ZS Associates

Data sharing

PREVIEW

A SECURE AND EASY WAY TO SHARE DATA ACROSS AMAZON REDSHIFT CLUSTERS





- Instant, granular, high-performance data access without data copies / movement
- Live and consistently updating views of data across all consumers
- Secure and governed collaboration within and across organizations and with external parties

- Workloads accessing shared data are isolated from each other
- Use cases: Cross-group collaboration and sharing, workload isolation and chargeability, data as a service

yelp

"Data sharing feature seamlessly allows multiple Amazon Redshift clusters to query data located in our RA3 clusters and their managed storage. This eliminates our concerns with delays in making data available for our teams, reduces the amount of data duplication and associated backfill headache. We now can concentrate even more of our time making use of our data in Amazon Redshift and enable better collaboration instead of data orchestration."

Steven Moy, Yelp

Materialized views auto refresh and query rewrite

UPDATED

SPEED UP QUERY PERFORMANCE BY ORDERS OF MAGNITUDE WITH PRECOMPUTED RESULTS

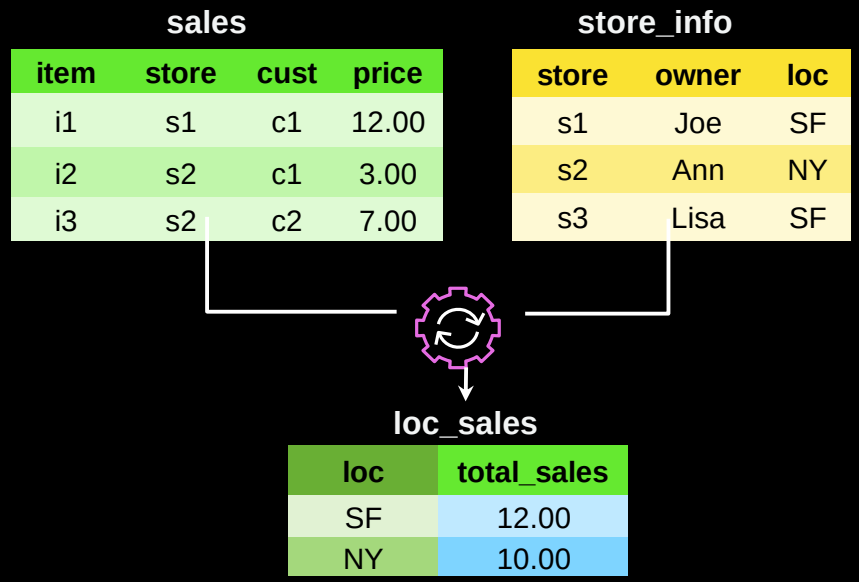
Simplify and accelerate iterative and predictable workloads, such as ETL, BI/dashboarding queries

MVs can be based on one or more Amazon Redshift tables or external tables (Spectrum, Federated)

Efficient incremental maintenance

Scheduled, automatic, or manually timed refresh

Amazon Redshift auto query rewrite optimizes queries by replacing native tables with materialized views



"The Amazon Redshift materialized view auto query rewrite feature reduced dashboard load times from 8 minutes to just 500 ms. The best part is that this is completely transparent for Tableau and the business user."

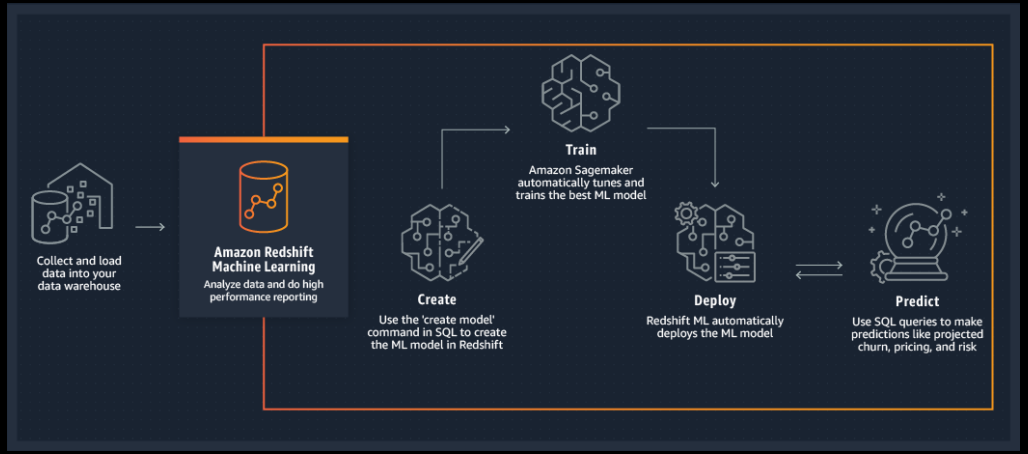
Arman Nasrollahi, Home24

Amazon Redshift ML

PREVIEW

EASILY CREATE AND TRAIN ML MODELS USING SQL QUERIES WITH AMAZON SAGEMAKER

- Use case: Product recommendations, fraud prevention, reduce customer churn
- Train and apply ML models using SQL
- From fully automated training to partially or fully guided training
- Automatic pre-processing, creation, training, deployment of your model



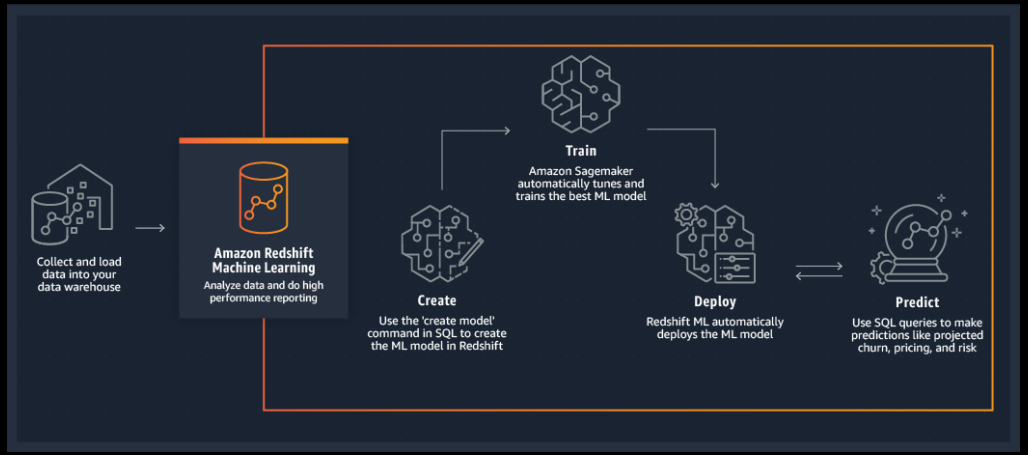

```
CREATE MODEL customer_churn
FROM (SELECT c.age, c.zip, c.monthly_spend,
c.monthly_cases, c.active FROM customer_info_table c)
TARGET c.active
FUNCTION predict_customer_churn
…;
```

Amazon Redshift ML

USE ML MODELS USING SQL QUERIES

PREVIEW

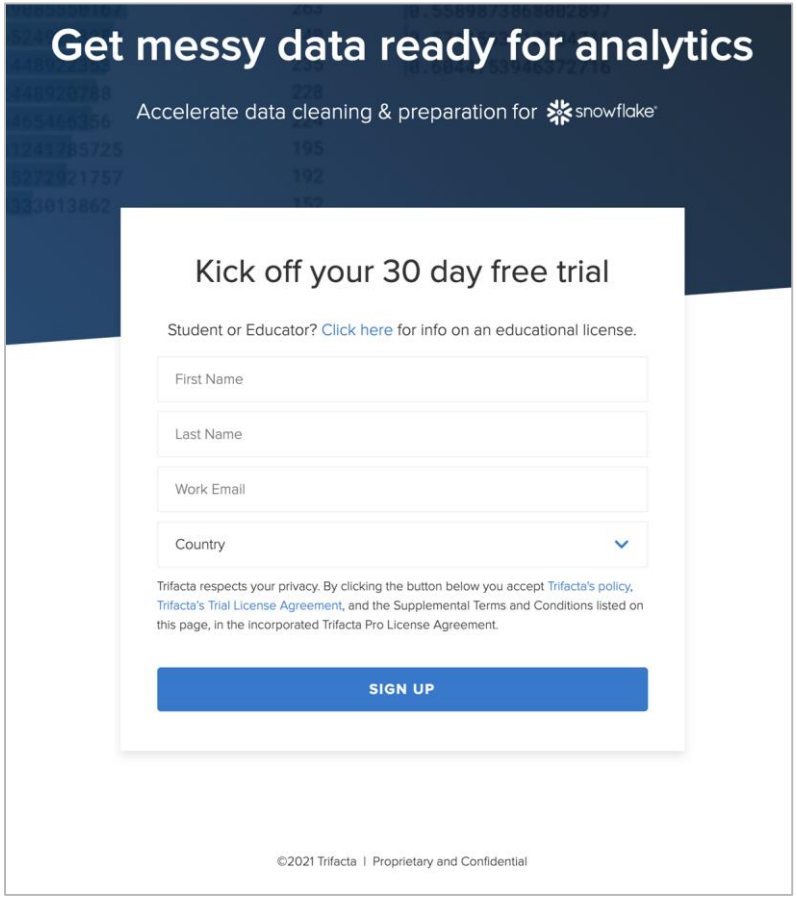
- Deploy inference models locally in Amazon Redshift
- Run an inference as invoking a user-defined function as part of SQL statements




```
SELECT n.id, n.firstName, n.lastName,
predict_customer_churn(n.age,c.zip,..)
AS activity_prediction
FROM new_customers n
WHERE n.marital_status = ‘single’
…;
```

Q & A

How to Get Started?



Start Free:
https://www.trifacta.com/start-wrangling/

Thank You

team@trifacta.com | Trifacta.com