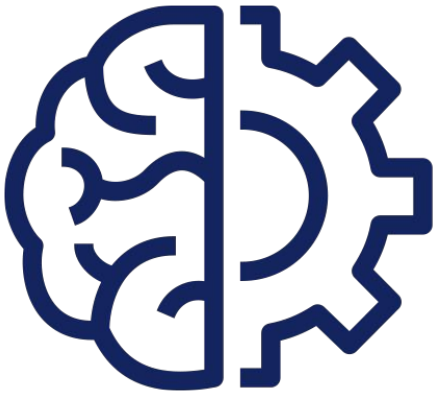


VIENNA DATA SCIENCE TOOLS

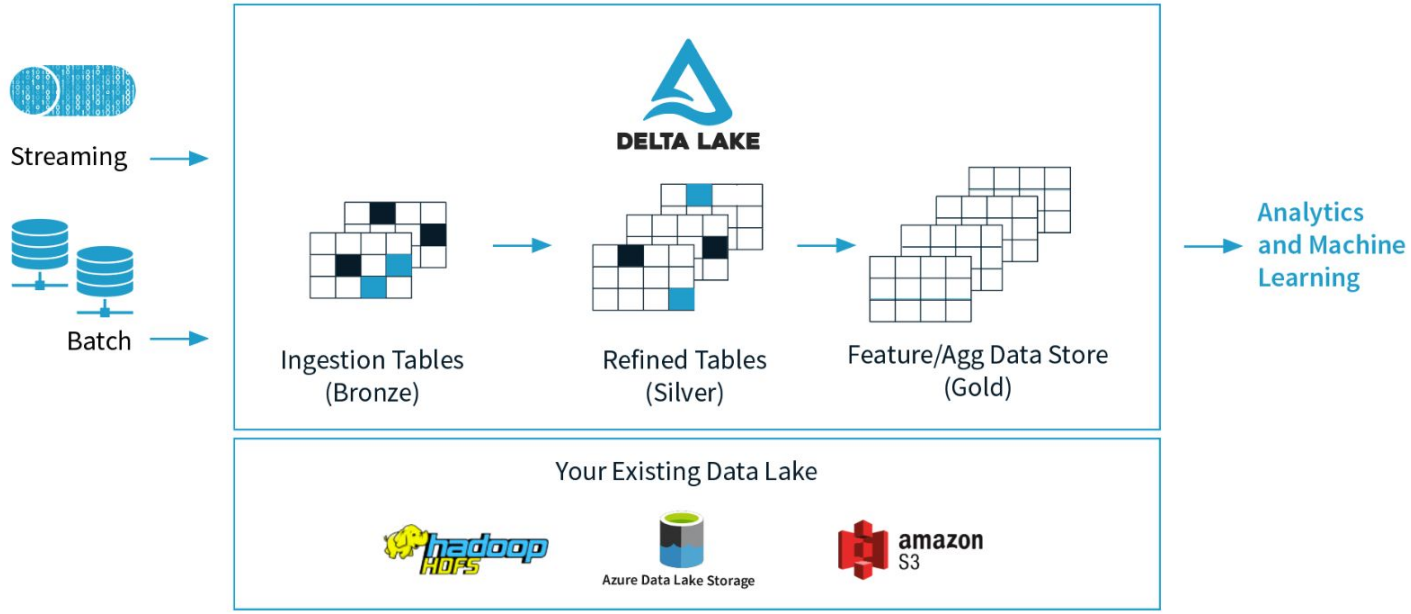
greentube
NOVOMATIC INTERACTIVE

NOVOMATIC

https://delta.io/

In a Nutshell

- Open source DWH-like solution for Big Data
- Computation: Apache Spark
- Storage: Apache Parquet



- ACID Transactions
- Schema Enforcement & Schema Evolution
- Data Versioning: Time Travel & Audit History
- Unified Batch and Streaming processing

Python Examples

```
df = spark.read.json("/data/events/")
df.write.format("delta").save("/delta/events")

df = DeltaTable.forPath(spark, "/delta/events").toDF()

deltaTable.delete("date < '2017-01-01'")

deltaTable.update(condition = "eventType = 'clck'", set = { "eventType": "'click'" })

deltaTable.alias("events").merge(
    source = updatesDF.alias("updates"),
    condition = "events.eventId = updates.eventId"
  ).whenMatchedUpdate(set =
    {
      "data": "updates.data"
    }
  ).whenNotMatchedInsert(values =
    {
      "date": "updates.date",
      "eventId": "updates.eventId",
      "data": "updates.data"
    }
  ).execute()
```

SQL Examples

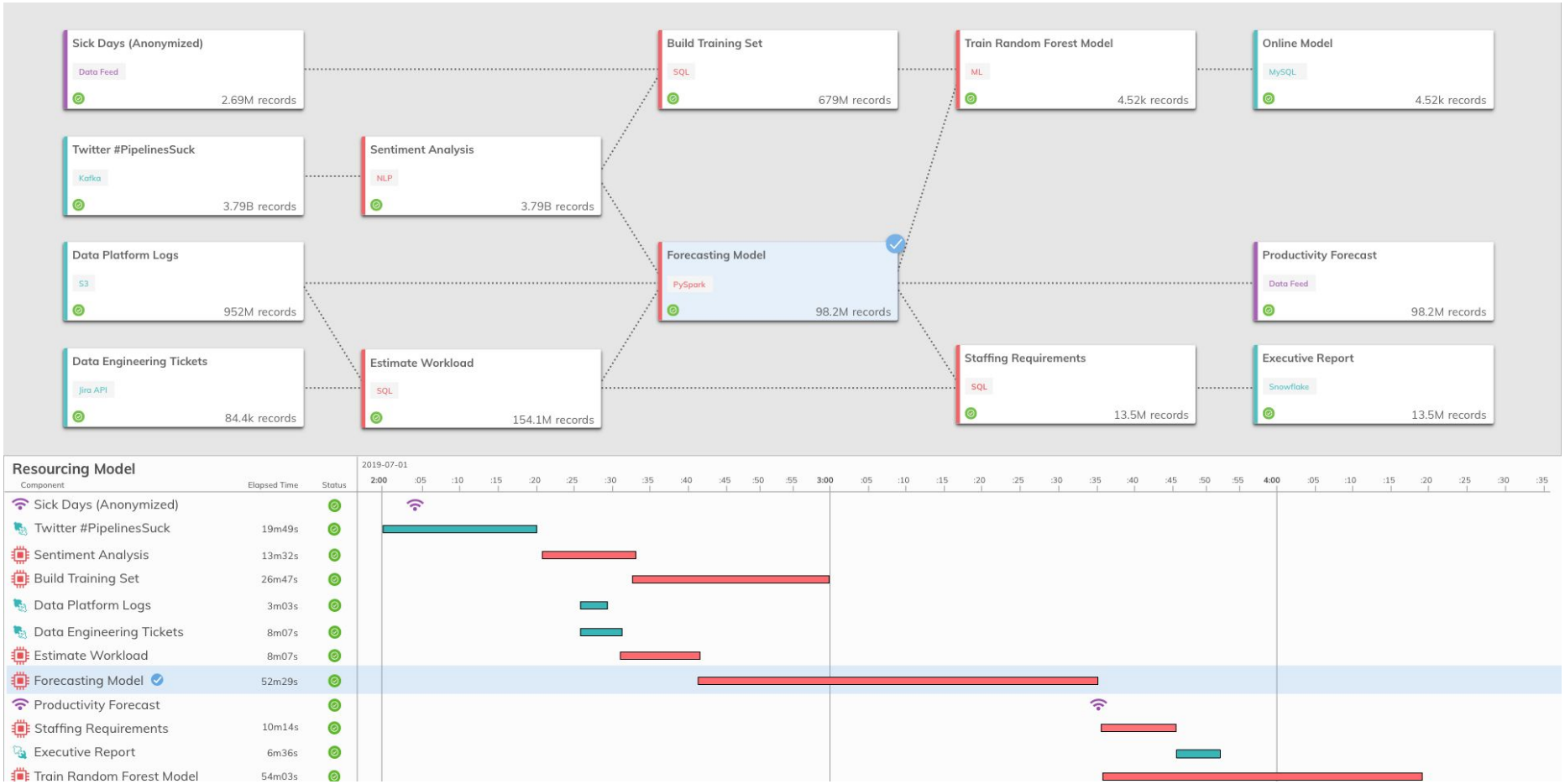
```
CREATE TABLE events
USING delta
AS SELECT *
FROM json.`/data/events/`

DELETE FROM events WHERE date < '2017-01-01'

UPDATE events SET eventType = 'click' WHERE eventType = 'clck'

MERGE INTO events
USING updates
ON events.eventId = updates.eventId
WHEN MATCHED THEN
  UPDATE SET
    events.data = updates.data
WHEN NOT MATCHED
  THEN INSERT (date, eventId, data) VALUES (date, eventId, data)
```

https://www.ascend.io/

In a Nutshell

- SaaS for "Autonomous Data Pipelines"
- Declarative Pipelines
  - Describe pipelines with simple, compact definitions in SQL, YAML or Pyspark.
- Data Plane - Managed Spark
  - Fully managed, containerized Spark on Kubernetes.
- Control Plane - Autonomous
  - Combines **declarative configurations** and **intelligent automation** to manage the underlying cloud infrastructure, dynamically construct pipelines and eliminate maintenance across the entire data lifecycle.
  - Sean Knapp: "Query planner/optimizer for Big Data workflows"
- Structured Data Lake
  - Automatically ensures **data integrity**, tracks **data lineage**, **de-duplicates data**, and **optimizes performance**.
- Multi-Cloud
  - Deploy and run seamlessly across AWS, Azure and Google Cloud Platform.

Data Pipeline Example



Announcements

We're hiring

- Software Architect (f/m)
  https://recruiting.novomatic.com/Vacancies/2071/Description/2
- Data Engineer (f/m)
  https://bit.ly/2MOzdW1
- Junior Data Engineer (f/m)
  https://bit.ly/2BFbdQ0

greentube
NOVOMATIC INTERACTIVE

NOVOMATIC

THE CONGRESS FOR DEVELOPERS IN AI, CLOUD, BLOCKCHAIN & IOT

WEAREDEVELOPERS CONGRESS VIENNA

HOFBURG VIENNA

28–29 NOVEMBER

SPEAKERS

Developers, Tech Pioneers & Heroes of Innovation

TANMAY BAKSHI
AI expert at age 15
IBM Cloud Advisor

CASSIE KOZYRKOV
Chief Decision Scientist
Google, Inc.

SIDDHA GANJU
Solution Architect
Nvidia Corporation

LEO SHIWEI LI
President
Tencent Cloud Europe

ConVienna_20_DataScienceTools

O'REILLY®

PROPOSALS SPONSORS ABOUT

Artificial Intelligence Conference

London, UK 20-23 April 2020

Call for speakers

Call closes 23:59 — 21 November 2019 GMT.

SUBMIT A PROPOSAL >

Do you have a great idea to share?

Thank you!

bpirvu@novomatic.com
jwilms@novomatic.com
anemeth@greentube.com

NOVOMATIC