H_2O

WORLD

2017

Donald Gennetten

Data Engineer

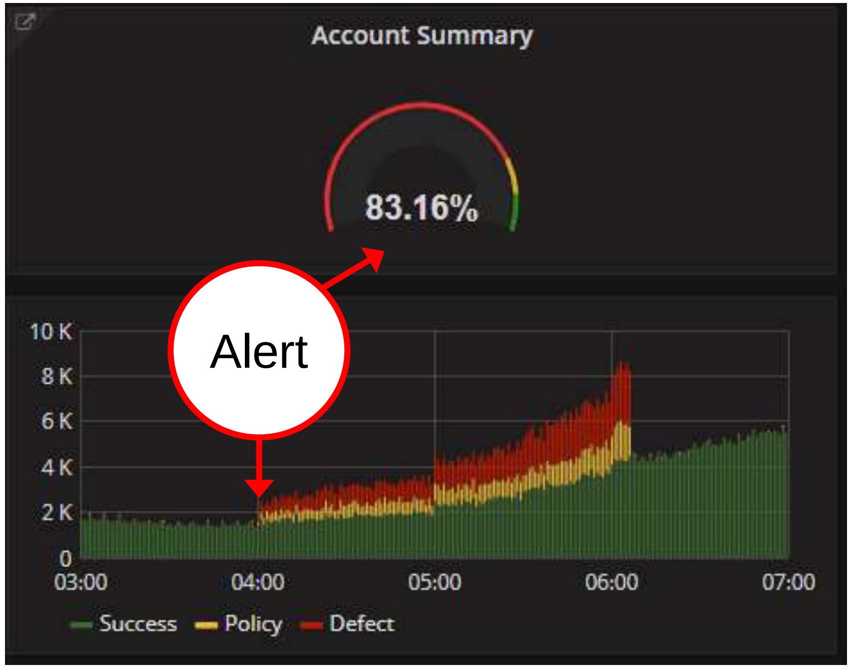
Rahul Gupta

Data Engineer

Using H2O for Mobile Transaction Forecasting & Anomaly Detection

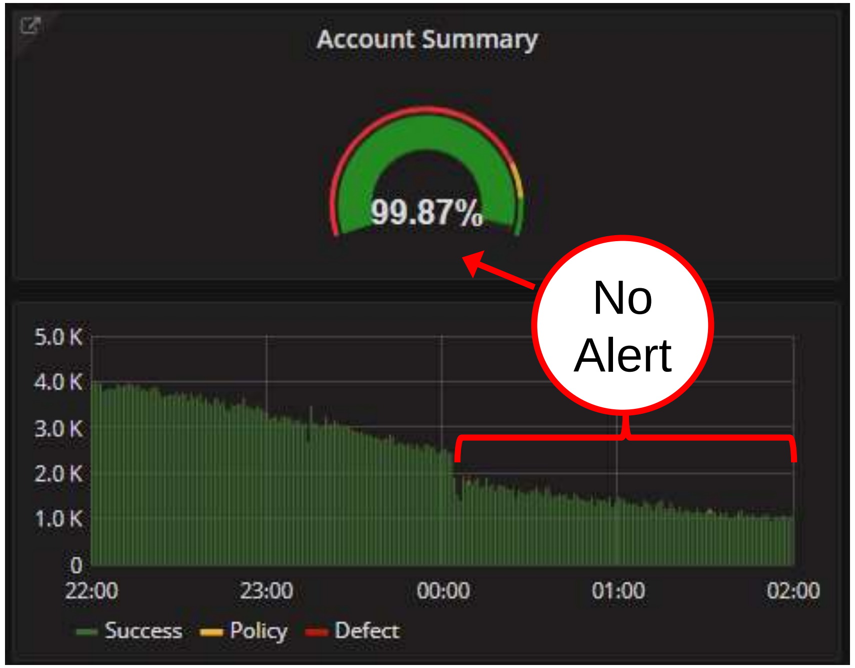
Problems are usually identifiable through elevated failures or volume anomalies

Elevated Failure Rate



Easy to detect, measure, and alert

Low Volume Anomaly



Hard to detect, measure, and alert

Why not set volume alerts?

Unlike failure alerts, volume-based thresholds vary by event type, hour, minute, day of week, week of the year, holiday, and much more.

100+ customer event types

x

24 hours/day

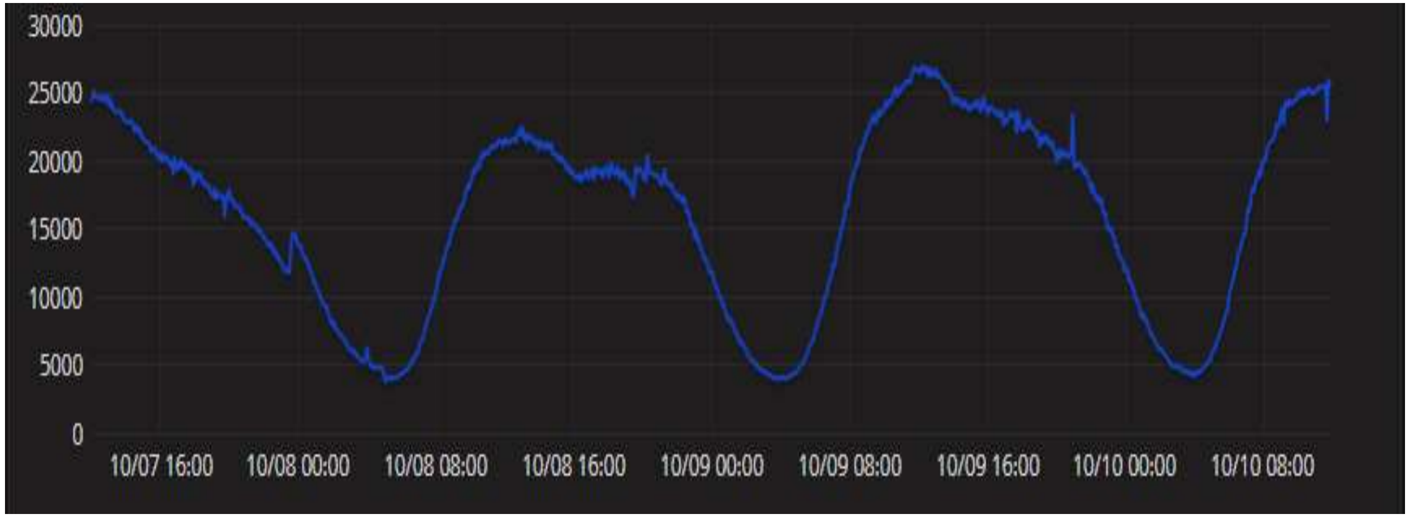
x

7 days/week

x

52 weeks/year

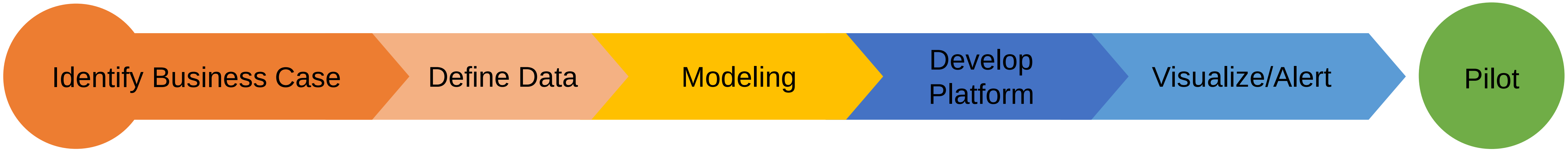
Over 873k distinct thresholds to calculate, set and maintain.



Machine Learning should be used when:

- You cannot effectively code the solution

Solving the problem required going beyond modeling

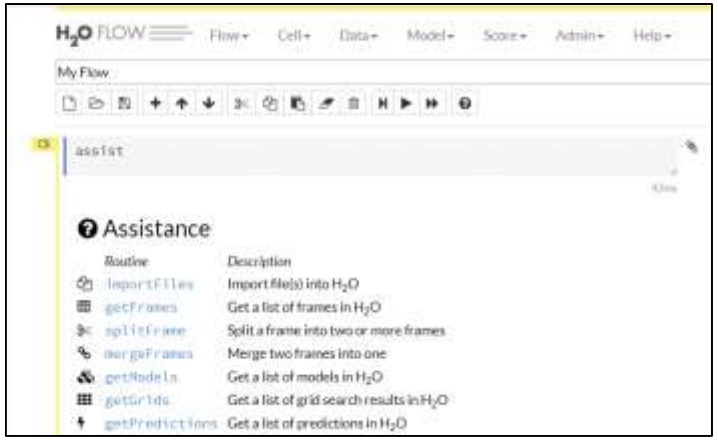


Our goal was to deliver Machine Learning for Production Monitoring that:

- Followed Governance Requirements
- Used Available Data Science and Machine Learning Resources
- Leveraged Platform Engineering and Open Source Technology
- Ensured Usability and Scalability

Sparkling Water allowed us to rapidly test and deploy machine learning

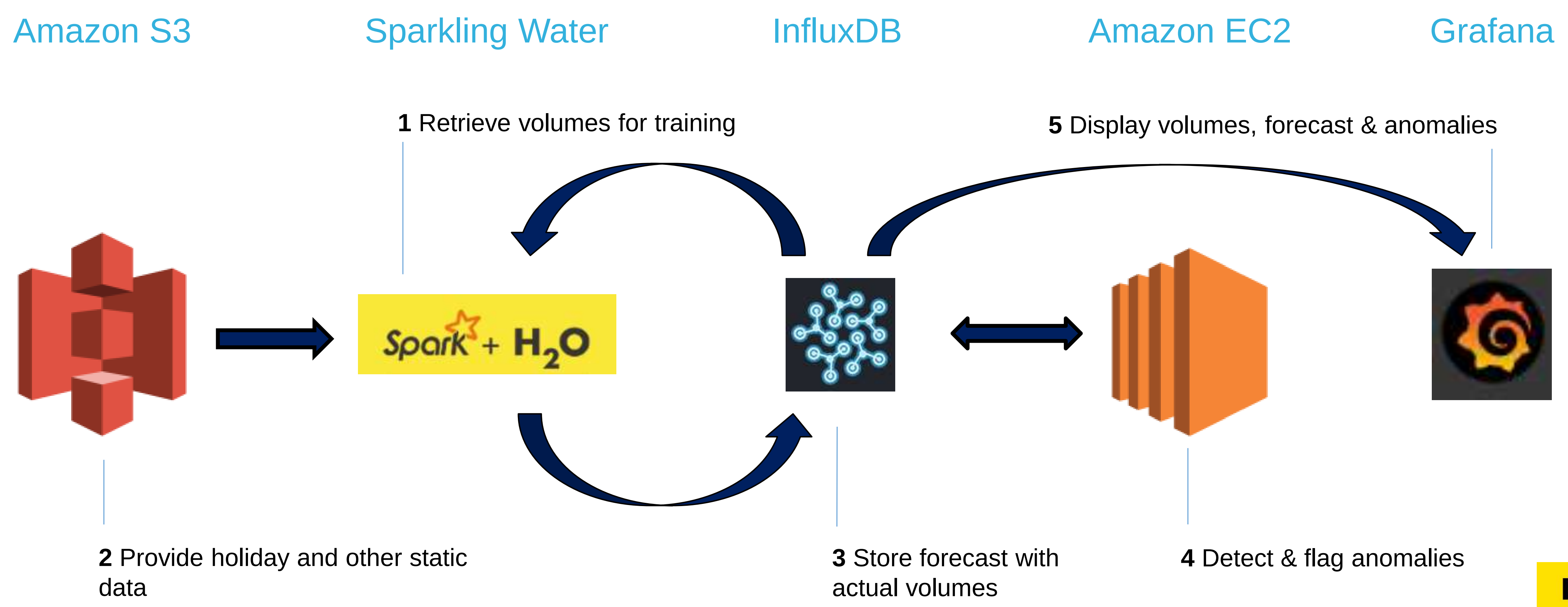
- **Sparkling Water combines the fast, scalable ML algorithms of H2O, the H2O Flow UI, Scala, and Python with the capabilities of Apache Spark**
- **In-memory processing supports big data environment needs**
- **Spark + Python + Scala enables a unified coding pipeline**
- **Grid search options allow for greater efficiency**
    - **Test models**
    - **Optimize hyperparameters**
- **H2O Flow facilitates ad-hoc experimentation**
- **REST API is easily integrated into production software**



GBM provided greater flexibility and benefits over traditional methods

- **Traditional time series techniques assume stationary data (no trends/seasonality), constant variance over time**
- **Univariate time series consists of single, sequential observations over equal time increments**
- **GBM model accepts external explanatory variables**
  - **# accounts having payment due**
  - **Incidents**
  - **Change orders**
  - **Payment due dates**
- **GBM also enables data filtering/exclusion (e.g., incident data for training set)**

We developed an open source, cloud-based platform for rapid delivery



H_2O WORLD 2017

What does it look like?

Monitoring teams are easily able to visually inspect forecasted and actual volumes in real-time

Forecasts are available for future dates to aid in capacity planning

What does anomalous volume look like?

Small changes in expected volume are easy to detect, measure, and alert

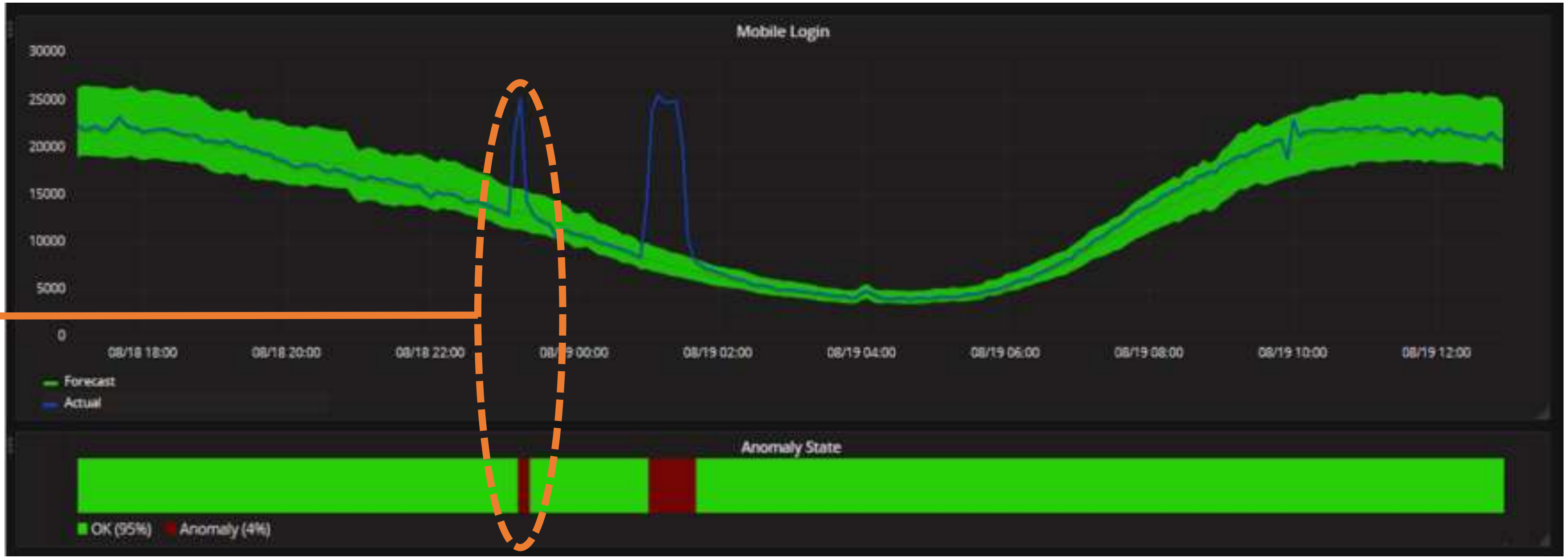
~12% of expected events were missing after a planned change to the streaming data platform

Alerts triggered due to lower than expected volume; Root cause analysis determined a platform release was casing dropped data and a code roll back was required to resolve the issue

Does it improve incident detection times?

Anomaly detection alerts are sent ahead of escalation and detection times, including when other alarms aren't triggered

Anomaly detected at 11:15 p.m. when Login volumes spiked ~20k higher than expected



Incident response teams were alerted at 11:17 p.m., more than 4 minutes before other incident alarms

Solar events as a predictor?

Variation from predicted login volume was easily quantified during the August 21st solar eclipse; Interest appears to have been lost within 15 minutes of totality

A. 12:06 p.m. EDT (9:06 a.m. PDT) the solar eclipse starts in Salem, Oregon

B. 2:41 p.m. EDT (11:41 a.m. PDT) totality begins in Columbia, South Carolina

C. 4:06 p.m. EDT (1:06 p.m. PDT) eclipse ends