splunk> .conf2017

Productizing ML For Behavior Modeling and Security

Janet He | Chief Solution Architect at *SAIC, Inc.*
Marios Iliofotou | Principal Data Science Engineer UBA

September 27, 2017 | Washington, DC

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

splunk> .conf2017

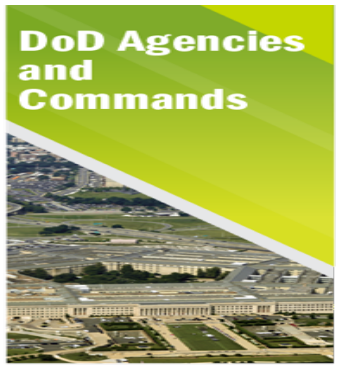
Agenda

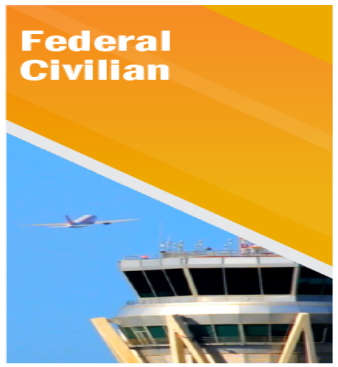
- SAIC Overview
- The Challenges and Opportunities
- SAIC Approach
- Solution Examples

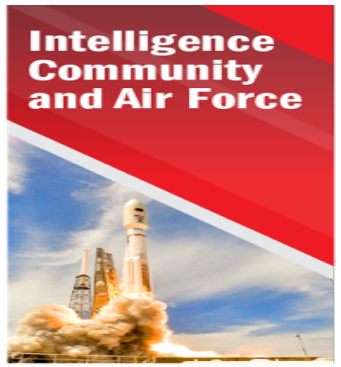
splunk> .conf2017

SAIC Overview

- Leading technology integrator specializing in technical, engineering, intelligence, and enterprise IT services to the U.S. government
- 47-year history of mission service delivery and customer relationships
- Significant scale of about $4.3 billion with diversified contract base
- Highly skilled workforce of about 15,000 employees, with presence in all 50 states, and focuses on
  - low-cost and low-risk enterprise project management,
  - visualization and data analysis, modeling and simulation, data publication and distribution, and big data analysis,
  - technology integration, IT security, application services and tools
  - user services, networking and communications, facility operations







Challenges and Opportunities

Effectively understanding and exploiting the rich content of machine-generated data is both an increasing challenge and huge opportunity for our customer organizations

- Lack of Operational Situational Awareness/Understanding trending/Ability to act
- Lack of visibility for Security threats and fraud detection across enterprise

- Network, System, Applications, User and Entity Activity Monitoring
- Using machine data to improve customer mission execution
- Streamlining and automating enterprise wide effort to gain the efficiency

splunk> .conf2017

SAIC Approach

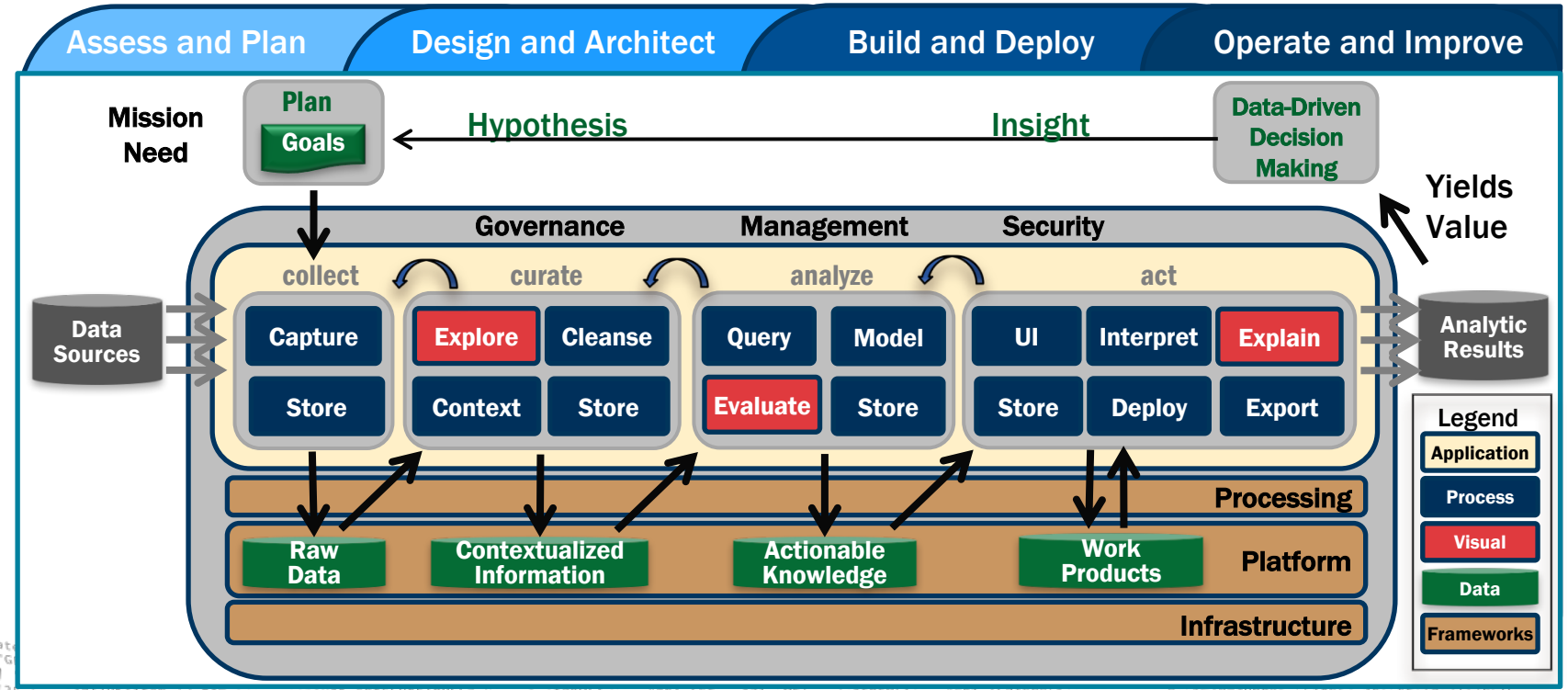
Technology Platform – Splunk

SAIC Approach

Data Management Methodology – SAIC DSE™

SAIC's Data Management Model enables innovation of analysis

- Data Science Edge™ (DSE) is SAIC's proprietary data lifecycle model geared toward the efficient planning and execution of enterprise data planning and analytics
- Model includes four phases of execution; Assess, Design, Build, and Improve. DSE Improve focuses on the performance and optimization of existing data and analytic systems
- SAIC has successfully used this process model to design a big data lake for our clients, and perform real-world testing of airport check-in biometrics devices



splunk> .conf2017

SAIC Approach

Data Protection – SAIC CSE™

CyberSecurity Edge Three Phase Methodology

1 — **Discover** offers highly trained objective experts to identify real-world security risk and validate the implementation and effectiveness of an organization's existing security controls against industry recognized best practices and adversarial threats.

2 — **Mitigation** is a highly tailored offering designed to help a customer design, plan, and implement solutions to meet specific goals and improve overall cybersecurity.

3 — **Manage** provides cost efficient, low risk options for ongoing and continuous monitoring support by certified cybersecurity experts.

- Three options include managed, staff, and hybrid.

Advantages of SAIC's Approach

Verified | Recognized

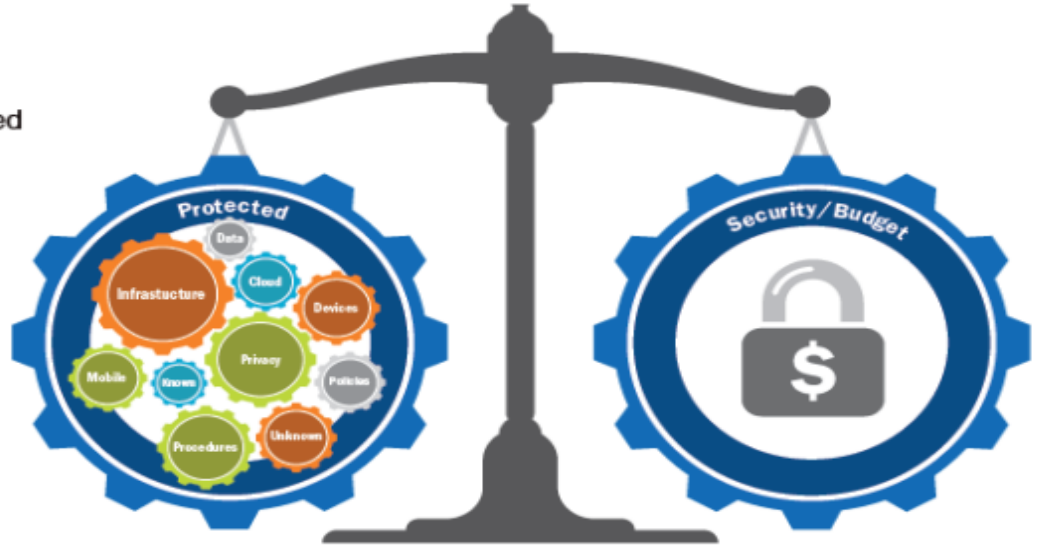
- Proven methodologies that have been developed and refined over countless engagements.

Automated | Optimized | Balanced | Tailored

- Offers customer-tailored solutions without the customization price tag.
- Optimizes current customer toolset.
- Fills gaps to strengthen ecosystem.
- Automates information assurance tasks.
- Balances tools, risk tolerance, and budget.

Packaged | Defined

- Clearly defines scope across all three phases with a fixed-priced model.

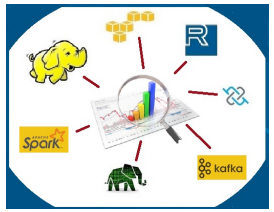


SAIC Approach

SAIC Big Data Analytics Solutions

Repeatable Solutions

- Big Data Assessment and Roadmap – templates and processes to assess an organization's big data maturity and devise a roadmap.
- Big Data Platform Accelerator – reference architecture, blueprints, conops and security guidance to accelerate development and deployment of a big data platform.
- Big Data Analytics Sandbox – an SAIC cloud-based platform enabling client organizations to "play" with big data tools and technologies and develop advanced analytic products. Augmented for Deep Learning tools.
- Big Data as a Service – a scalable "as a service" offering allowing streaming analysis, batch analytics and data exploration in a secure fashion. Augmented for logical data analytics solution to handle the *Variety* problem of big data.



splunk> .conf2017

Solution Examples

SAIC Internal Splunk UBA Implementation

- SAIC Splunk UBA implementation based on machine data and our existing SIEM infrastructure.
- Objectives
  - Detect hidden security threats
  - Monitor networking, system, application, user and device anomalous behavior
  - Provide threat visualization
  - Increase SOC response to threats efficiently and effectively

Integrating with currently Splunk infrastructure include ES, ITSI, and SAIC capabilities in DSE, CSE and big data analytics services.

Aaron Bishop, @SAIC*inc*

A CISO's Perspective on User Behavior Analytics: Setting the Right Expectations for All Stakeholders

splunk> .conf2017

Solution Example

Machine Learning – Threat Analysis

Need:

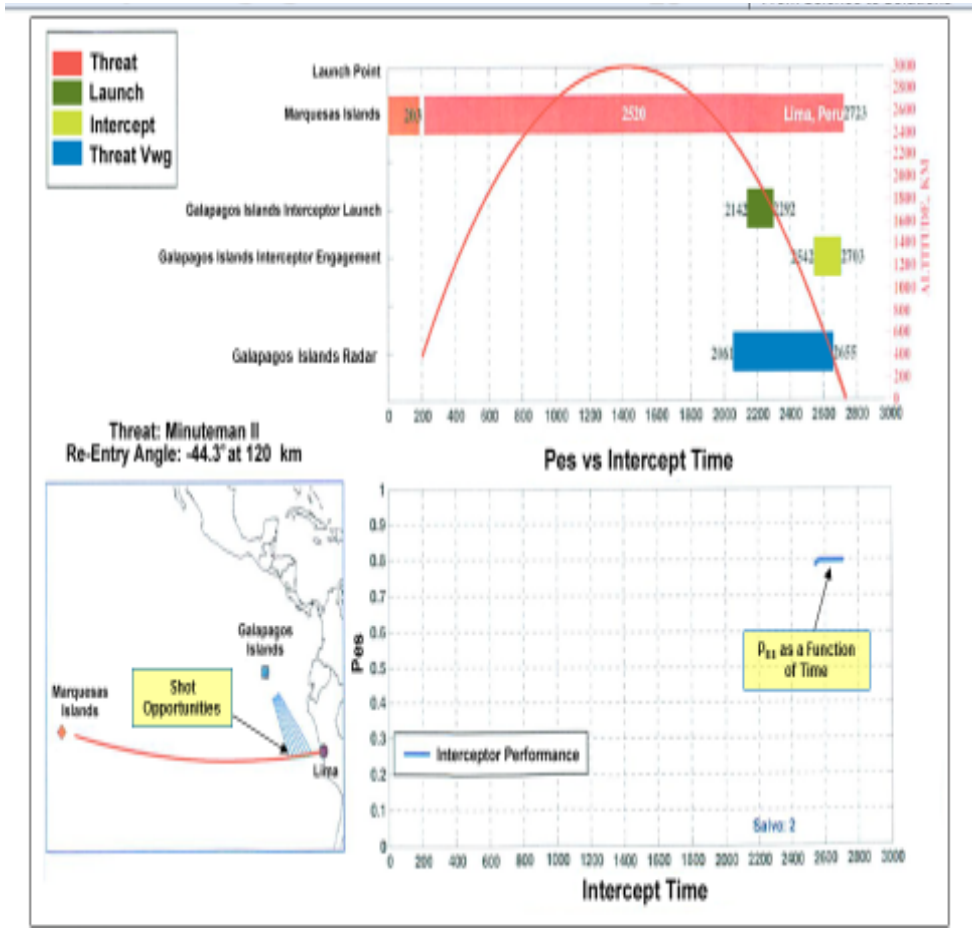
- Support studies on emerging threats and impacts. Manage and coordinate over 30 TB of raw data and processed products between multiple sites

Solution:

- Use of SQL databases as well as NoSQL (MongoDB)
- Developed and modeled advanced threat discrimination algorithms using Neural Networks, and Bayesians classifiers
- Automated tools to run simulations, generate KPIs, and create briefings
- Variety of tools used for visual displays including GIS and 3-D plots

Benefits:

- Eliminates laborious manual effort on part of analysts
- Provides frequent insights to leadership



splunk> .conf2017

Contact

- For additional information please contact us
- Splunk .conf 2017 SAIC booth M36

Sanjay Sardar
VP | Advanced Analytics
Advanced Analytics, Simulation and Training
Cell: 703.861.5620 | Desk: 703.676.5028
email: sanjay.sardar@saic.com | @SAICinc

Janet He
Chief Solution Architect | Advanced Analytics
Advanced Analytics, Simulation and Training
Cell: 301.366.2078 | Desk: 703.676.2378
email: janet.he@saic.com | @SAICinc

splunk> .conf2017

Overview

- ~~Introduction~~
- **Challenges**
- Platform
- Programmability (SDK)
- Conclusions Q/A

splunk> .conf2017

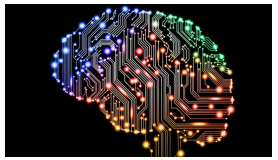
Why use Machine Learning (ML)?

- You are probably trying to solve one of these problems:
  - Insider threats
  - Malware/hackers
  - Fraud

What they all have in common?
Simplistic solutions don't work!

Complete Solutions are More than Just ML

- Machine learning capabilities
  - Supervised, Unsupervised



- Ability to quickly analyze lots of data (Big Data)
  - Need to use a cluster of machines

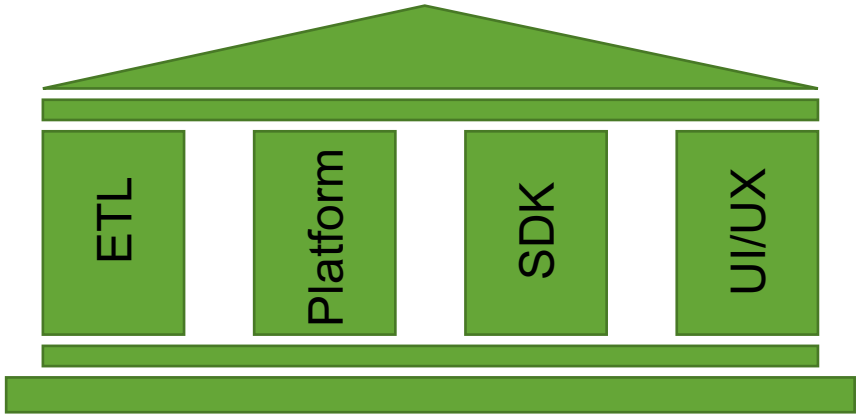
- Quick response to new events (automation)
  - Cannot have a Data Scientist look and analyze the data manually (takes too long!)

Productizing(Big data + ML + Automation) = HARD

splunk> .conf2017

Why Productizing a Solution is Hard?

1. ETL – Parsing, normalizing, cleaning
2. Platform – Scalability, performance, monitoring, orchestration
3. Programmability – Change/add new logic, test, develop (SDK)
4. Presentation – UI/UX, exploration/investigation



Any of the four pillars being weak the solution will fail!

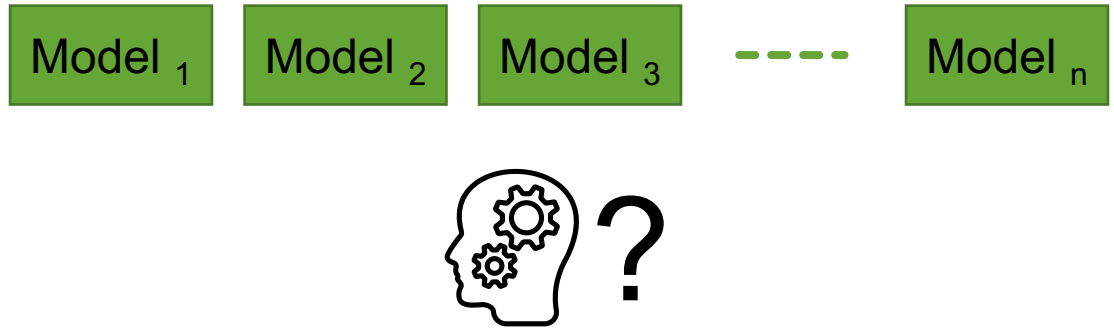
Overview

- ~~Introduction~~
- ~~Challenges~~
- **Platform**
- Programmability (SDK)
- Conclusions Q/A

splunk> .conf2017

Goal: Manage Multiple ML Models

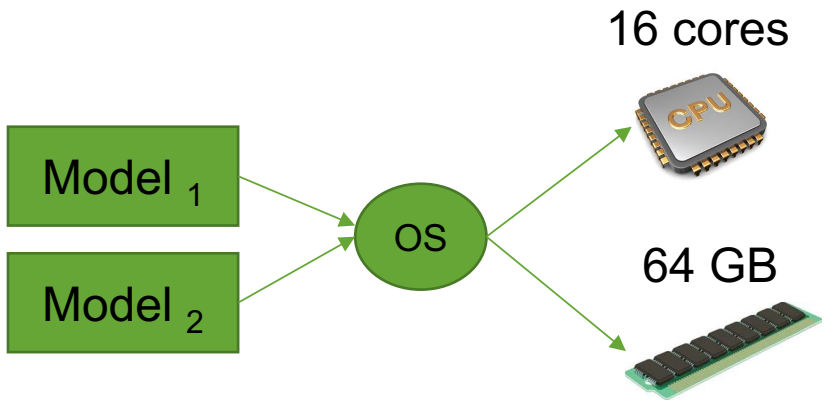
Model = “Training and scoring of ML models plus utility tasks”



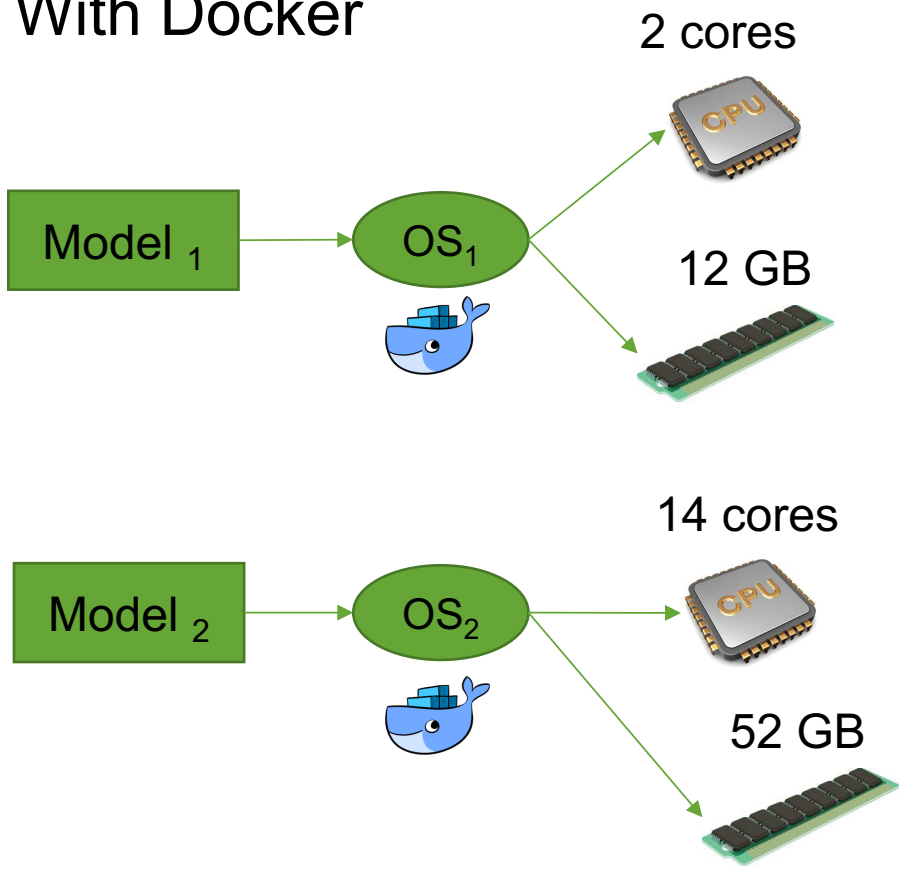
- Isolate models (processes)
  - Out of memory
  - Out of disk space
  - High CPU usage

Model Isolation With Docker

Without Docker



With Docker



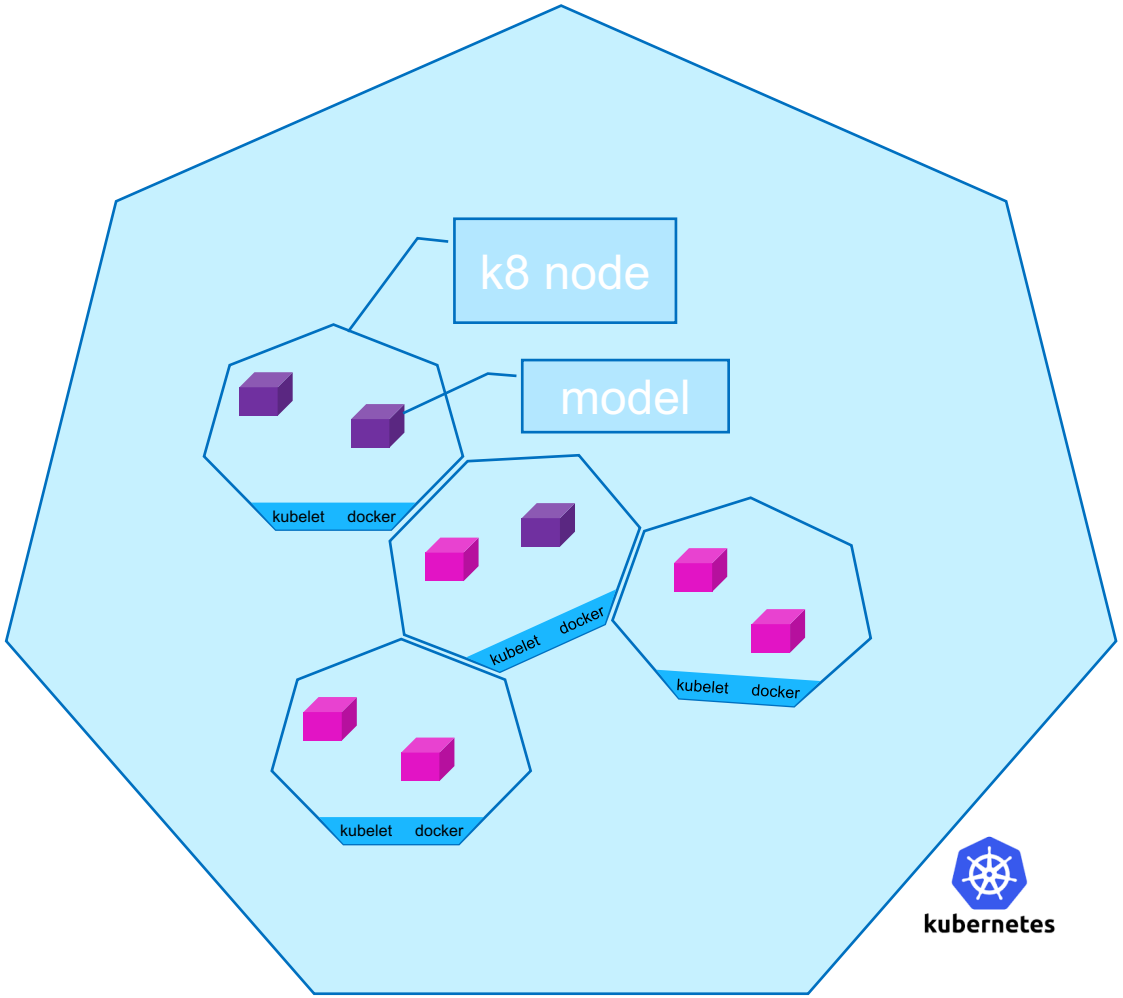
If “Model 2” takes all the resource, “Model 1” is not affected

Elastic Resource Allocation for Models

- Deploy models over multiple nodes
- A critical model runs too slow
  - Give it more resources
- A model holds resources without utilizing them
  - Give the resources to the ones that need them
- New models are loaded into the system

Multi Node Containers with Kubernetes

- Scenarios
  - A model is struggling
    - Spin more instances of the model and balance the load
  - New model is added
    - Create new containers and assign them to a node
  - When cluster gets overloaded
    - Add extra nodes



Kubernetes Cluster

Overview

- ~~Introduction~~
- ~~Challenges~~
- ~~Platform~~
- **Programmability (SDK)**
- Conclusions Q/A

splunk> .conf2017

Programmability: From Ideas to Production

Research

Explore initial idea, visualize data, understand the problem, choose the best algorithm

Develop

Program solution, evaluate tradeoffs

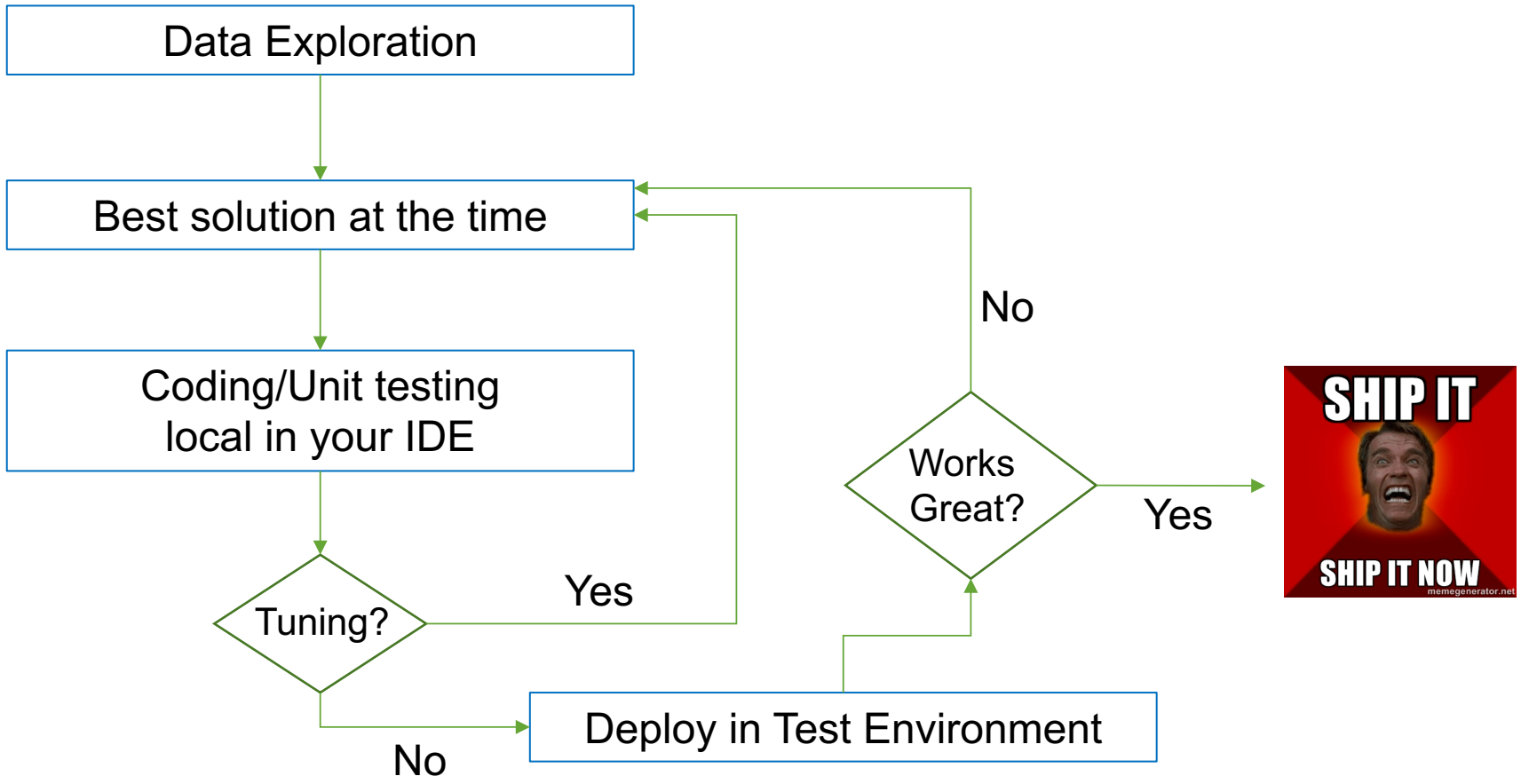
Test

Test that the solution works well in terms of performance and accuracy

Deploy

Run the model on the production cluster

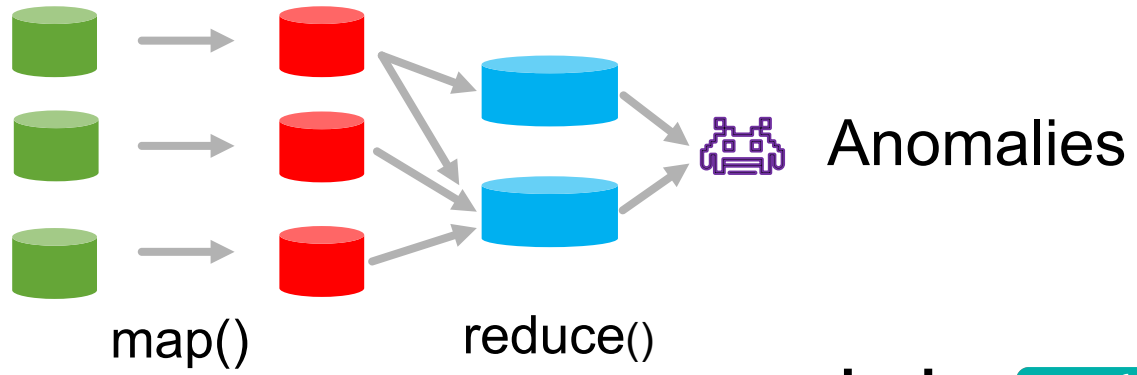
Realistic Model Development Life-Cycle



SDK should support all of these steps!

Different Use Cases Require Different Model Types

- Streaming
  - Single pass over the data
  - Quick response to events
  - Run continuously
- Batch
  - Multiple passes over the data
  - Can run expensive correlations (joins)
  - Run at scheduled intervals (think Linux Cron jobs)



splunk> UBA: Streaming Model APIs

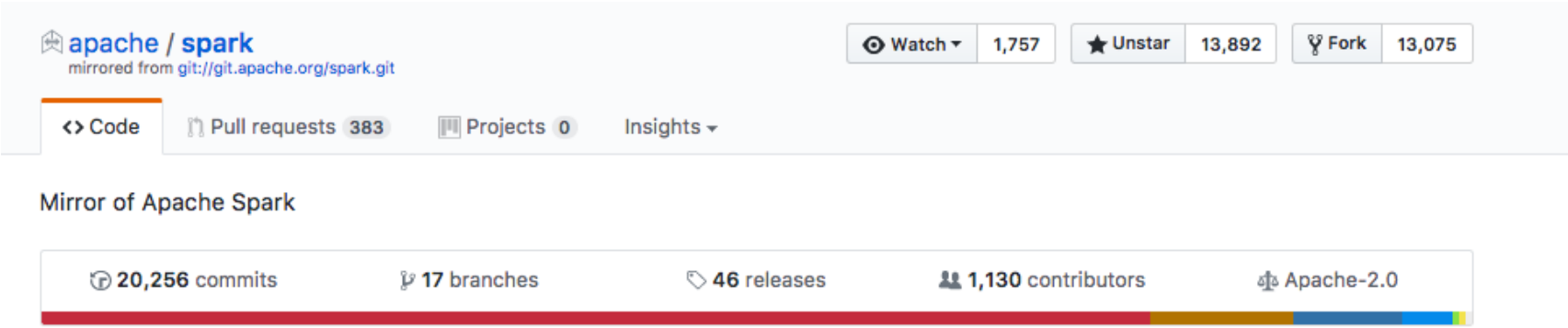
```
Option[Anomaly] analyzeData(DataEvent currentEvent)
```

- State is checkpointed internally
  - Serialization (Protocol Buffers, Kryo)

- Streaming models choose
  a) Pivot (e.g., per user or device)
  b) Input types (e.g., HTTP traffic data)

splunk> UBA: Batch Model APIs

- Apache Spark



- Apache 2.0.x full set of APIs are supported
  - RDD, Dataset, DataFrames, Spark SQL

```
httpData.groupBy('userId).agg(sum('bytesOut), unique('dstIP))
```

Overview

- ~~Introduction~~
- ~~Challenges~~
- ~~Platform~~
- ~~Programmability (SDK)~~
- **Conclusions Q/A**

splunk> .conf2017

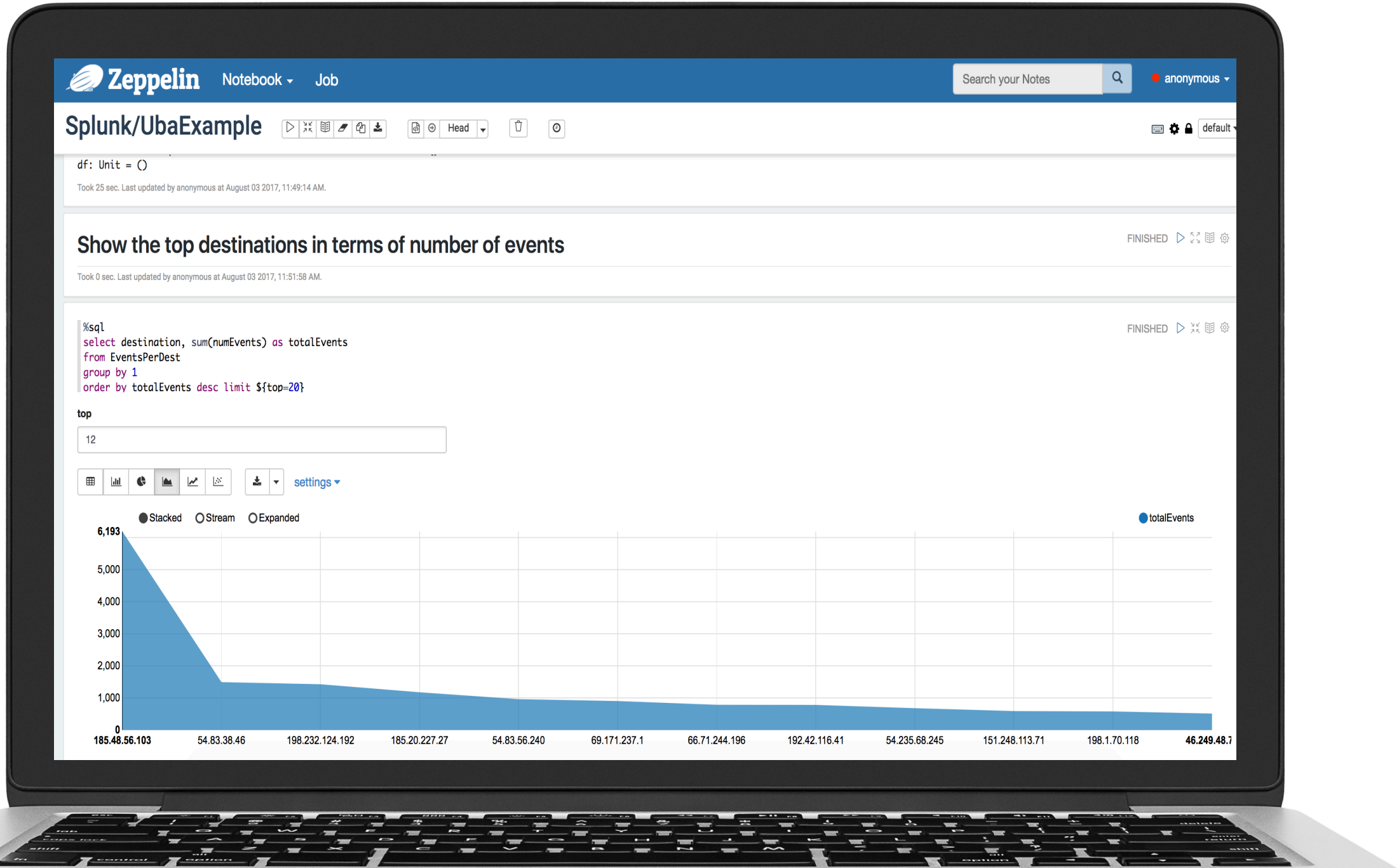
Conclusions

- Productizing ML solutions has many challenges
  - Orchestration (fault tolerance, isolation, elasticity)
  - Friendly developer environment

- splunk> UBA addresses all of these challenges
  - ETL: Normalizes all data and attributes events to actual users/devices
  - Platform: Scaling, monitoring, orchestration
  - SDK: UBA 4.0 comes with SDK support
  - UI: Integrated experience

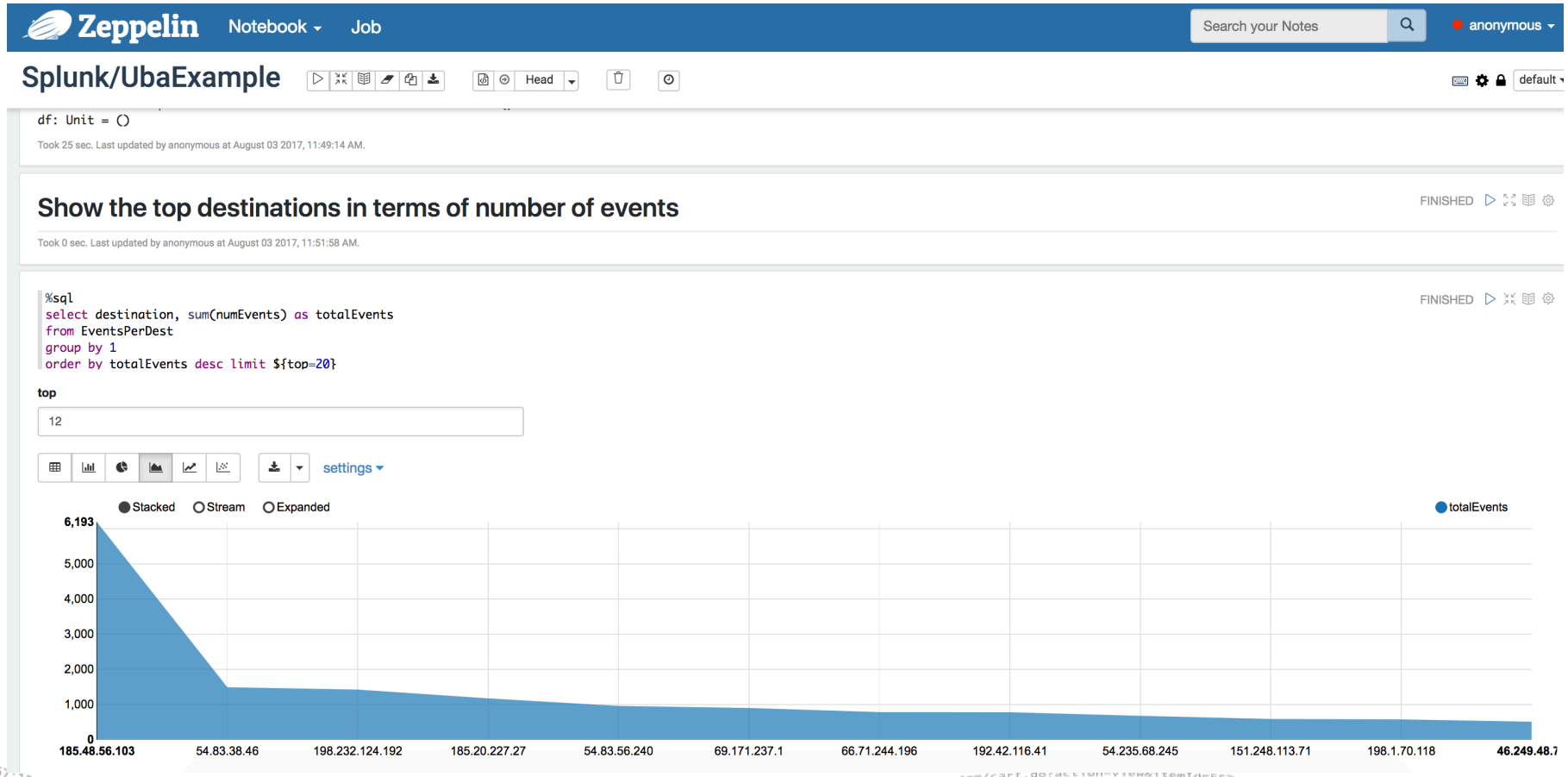
Thank You

Don't forget to rate this session in the .conf2017 mobile app



Notebook Driven Development

- Zeppelin notebook example



splunk> .conf2017

Different Model Types and Challenges

- Streaming: Quick response to events (Kafka)
  - State explosion
  - Slow EPS
- Batch: Stronger correlations (Apache Spark)
  - Job execution time (timeout)
- Common challenges
  - Scaling up/down
  - CPU/Memory/IO fairness (one task interfering with another)