MacroBase: Prioritizing Attention in Fast Data

Peter Bailis
DAWN Project, Stanford InfoLab
Stanford DAWN (Data Analytics for What’s Next) Project

Peter Bailis
Streaming & Databases

Chris Ré
MacArthur Genius
Databases + ML

Kunle Olukotun
Father of Multicore
Domain Specific Languages

Matei Zaharia
Co-Creator of Spark and Mesos
It’s the golden era of data

Incredible advances in image recognition, natural language processing, planning, info retrieval

Society-scale impact: autonomous vehicles, personalized medicine, human trafficking

No end in sight for advances in ML

*for the best-funded, best-trained engineering teams*
The DAWN Question

What if anyone with domain expertise could build their own production-quality ML products?

- Without a PhD in machine learning
- Without being an expert in DB + systems
- Without understanding the latest hardware

What’s needed: end-to-end systems that cover the full user workflow
The DAWN Stack
Understanding Customer Interactions

Given smart customer data streams (e.g., retail telemetry, purchasing decisions, customer service interactions), how can we fuse, filter, and aggregate data to deliver actionable customer, business insights?

Our research: scalable ML-powered tools that prioritize analyst attention in large data streams
Customer **Scale**: Opportunity and Challenge

Click impressions (10K-MMs / day)
Retail interactions (10s-1000s video feeds)
Purchasing history (years of data warehousing)
External data sources (demographic, social media, 3rd party)

**Opportunity**: combining large, fast data sources boosts quality

**Challenge**: efficient, cost- and time-effective analytics
After “Big Data”, Data Continues to Grow!

Big Data systems (e.g., HDFS, S3, Kafka), cloud reduced storage costs

Further, instrumentation of complex applications via sensors, processes, production telemetry has led to exploding data volumes

e.g., today, Facebook, Twitter, LinkedIn collect 12M+ events/sec
Challenge: Limited human attention

Human attention is scarce! Infeasible to manually inspect large volumes.

In practice: data only accessed for post-hoc root cause analyses.

*top SV orgs say:* < 6% data read

MEMS & sensors: manufacturing, monitoring, "IoT"
Example: Cambridge Mobile Telematics

Product: collect, analyze telemetry to improve driver behavior

Question: do users enjoy the application on every platform?
over 24K different Android device types; 2x since 2013
Example: Cambridge Mobile Telematics

Product: collect, analyze telemetry to improve driver behavior

Question: is the application behaving well on every platform?

Challenge: 24K different Android devices, 25 Major API releases spending even 1 second per combination requires 7 days

“IOS 9.0 beta 1–5 (but not 9.0.1) had a buggy BLE stack that prevented iOS devices from connecting to devices.”
Challenge: Limited human attention

Human attention is scarce! Infeasible to manually inspect large volumes.

In practice: data only accessed for post-hoc root cause analyses.

*top SV orgs say: < 6% data read*

Needed: systems that automatically prioritize attention.
MacroBase Architecture and Topics

**CLASSIFY**
- Identify data in tails

**EXPLAIN**
- Find disproportionately correlated attributes
- Outliers: {iPhone6, Canada}, {iPhone6, USA}, {iPhone5, Canada}
- Inliers: {iPhone6, USA}

**TRANSFORM**
- Extract domain-specific signals
- E.g., identify data in tails

**Key research question:**
- How can building end-to-end systems improve scalability and result quality?
MacroBase Architecture and Topics

**Key research question:**
how can building end-to-end systems improve scalability and result quality?

- **Classify**
  - e.g., identify data in tails
  - find disproportionately correlated attributes

- **Transform**
  - extract domain-specific signals

- **Explain**
  - Outliers: {iPhone6, Canada}, {iPhone6, USA}, {iPhone5, Canada}
  - Inliers: {iPhone6, USA}, {iPhone6, USA}, {iPhone5, USA}

Note: 100x more inliers...
MacroBase System: Prioritize Attention

Execute cascades of operators that transform, filter, aggregate the stream

**Example Output:** readings from Android Galaxy S5 devices running app version 52 are 30 times more likely than others to have abnormally high frequency

[Bailis et al., CIDR 2017]
Early Production Usage

automotive
monitoring fleet QoS

online services & datacenters (DevOps / monitoring)
identifying slow containers, exception telemetry

industrial manufacturing
key sources of process variance in product

mobile applications
diagnosis of misconfigured platforms
Early Production Usage

“MacroBase discovered a rare issue with the CMT application and a device-specific battery problem. Consultation and investigation with the CMT team confirmed these issues as previously unknown...”

[Bailis et al., SIGMOD 2017]
MacroBase Architecture and Topics

**Transform**
- Extract domain-specific signals
  - e.g., identify data in tails

**Classify**
- Find disproportionately correlated attributes
  - Outliers: {iPhone6, Canada}, {iPhone6, USA}, {iPhone5, Canada}
  - Inliers: {iPhone6, USA}, {iPhone6, USA}, {iPhone5, USA}

**Explain**
NoScope: 1000x Faster Video Extraction

What if we want to extract higher-level features from complex streams, like video?

Neural networks offer promise

Problem: state-of-the-art neural networks run ~30fps on $1K GPU

[Kang et al., arXiv:1703.02529]
NoScope: 1000x Faster Video Ext

Idea: use ideas from query optimization to speed NN evaluation

1. A **difference detector** to see if video has changed

2. Models that are **specialized** just in time to operate on a given feed

   **End result:**
   Process up to 1000x more video streams for same processing cost

   [Kang et al., arXiv:1703.02529]
MacroBase and Customer Interactions

Given smart customer data streams (e.g., retail telemetry, purchasing decisions, customer service interactions), how can we fuse, filter, and aggregate data to deliver actionable customer insights?

Our research: scalable ML-powered systems that prioritize analyst attention in large data streams
The DAWN Stack
Conclusions

Increasing data volumes demand new infrastructure for model training, fast inference, prioritizing human attention

MacroBase: combine feature extraction, classification, explanation

Major systems opportunity: efficient implementation at each stage, spanning query optimization, cardinality estimation, specialization

Stanford DAWN Project: A new stack for next-gen analytics

http://dawn.cs.stanford.edu/