Deep Analytics Pipeline
A Benchmark Proposal

Milind Bhandarkar
(Milind.Bhandarkar@emc.com)
(Twitter: @techmilind)

Quest for "Typical" Workloads

- Benchmarks most relevant if representative
- Tune systems for broadly applicable workloads
- Designing optimized systems: Make common tasks fast, other tasks possible

Encouraging Early Results

- Analyzed characteristics of 1M+ real Hadoop jobs on production clusters at Yahoo, 100+ features
- Identified 8 Job types
- Verified with GridMix 3
- Characterization of Hadoop Jobs Using Unsupervised Learning, Sonali Aggarwal, Shashank Phadke & Milind Bhandarkar, in 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, Indiana, December 2010 (http://doi.ieeecomputersociety.org/10.1109/CloudCom.2010.20)

TeraSort

William Vambenepe @vambenepe 24 Oct

MapR sets new world record for Hadoop **TeraSort**. On Google Compute Engine. insights.wired.com/video/mapr-goo…

Expand

Rohit Valia @rohitvalia 24 Oct

IBM demonstrates breakthrough #hadoop 100TB **Terasort** results on private cloud. More details at my session @strataconf tinyurl.com/HiPerfHadoop

Expand

Nyyra Jvggranhre @nyyrajvggranhre 24 Oct

Stop using terasort as a benchmark! It is almost always very far from reality or measures different things than you think it does!

Expand

Andrew Purtell @akpurtell 24 Oct

Prevalence of **Terasort** benchmark as yardstick in various talks means we desperately need an industry benchmark suite for Hadoop. #strataconf

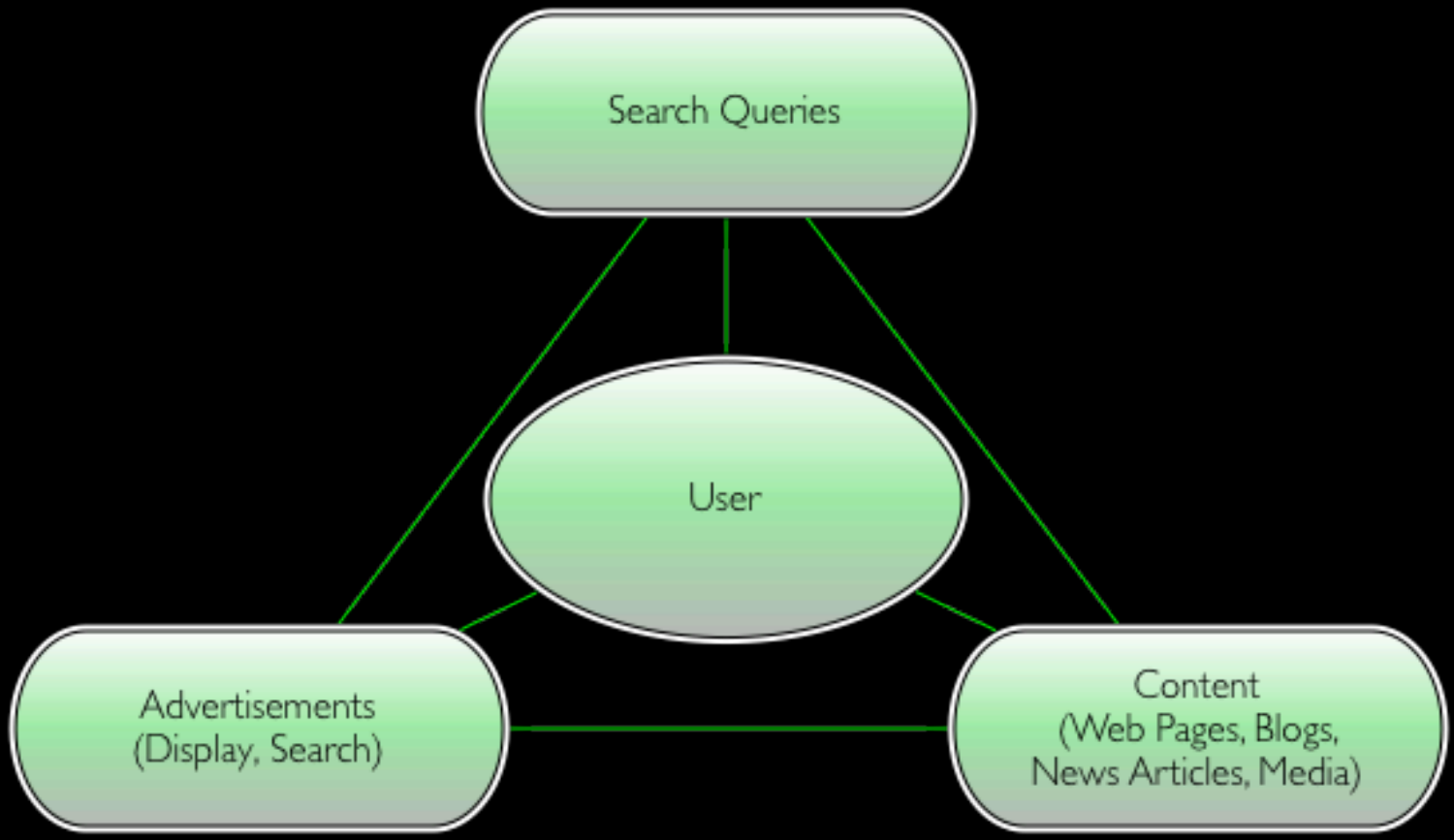
Expand

Drivers for Big Data

- Ubiquitous Connectivity
- Sensors Everywhere
- Democratization of Content

Big Data Sources

- Events
  - Direct - Human Initiated
  - Indirect - Machine Initiated
- Software Sensors (Clickstreams, Locations)
- Public Content (blogs, tweets, Status updates, images, videos)



Online: Major Data Sources

“User” Modeling

- Objective: Determine User-Interests by mining user-activities
- Large dimensionality of possible user activities
- Typical user has sparse activity vector
- Event attributes change over time

Domain: Retail

- User = Customer
- Activities
  - Online: Purchase, Ad click, FB Likes
  - Offline : Brick-and-mortar purchases, returns, coupon clipping, gift cards
- Personalized Product Recommendation

Domain: DataCenters

- "User" = HW & SW Components
- Activities
  - Log messages, Metrics, connectivity, communication events
- Goal: Proactive alerting of imminent failures

Domain: HealthCare

- User = Patient
- Activities
  - Dr Visits, Medicine refills, Medical History
  - 3G/WiFi-enabled Pillbox...
- Goal: Prevent Hospital Readmissions

Domain: Telecom

- User: Subscriber
- Activities
  - Calls made, duration, calls dropped, locations, ...
  - “social” graph, status updates
- Goal: Reduce customer churn

Domain: Ad-Supported Web

- User = User :-)
- Activities
  - Clicks on content, Likes, Repost
  - Search Queries, Comments, Participation
- Goal: Increase Engagement, Increase Clicks on revenue-generating content (ads/ premium content)

User-Modeling Pipeline

- Sessionization
- Feature and Target Generation
- Model Training
- Offline Scoring & Evaluation
- Batch Scoring & Upload to serving

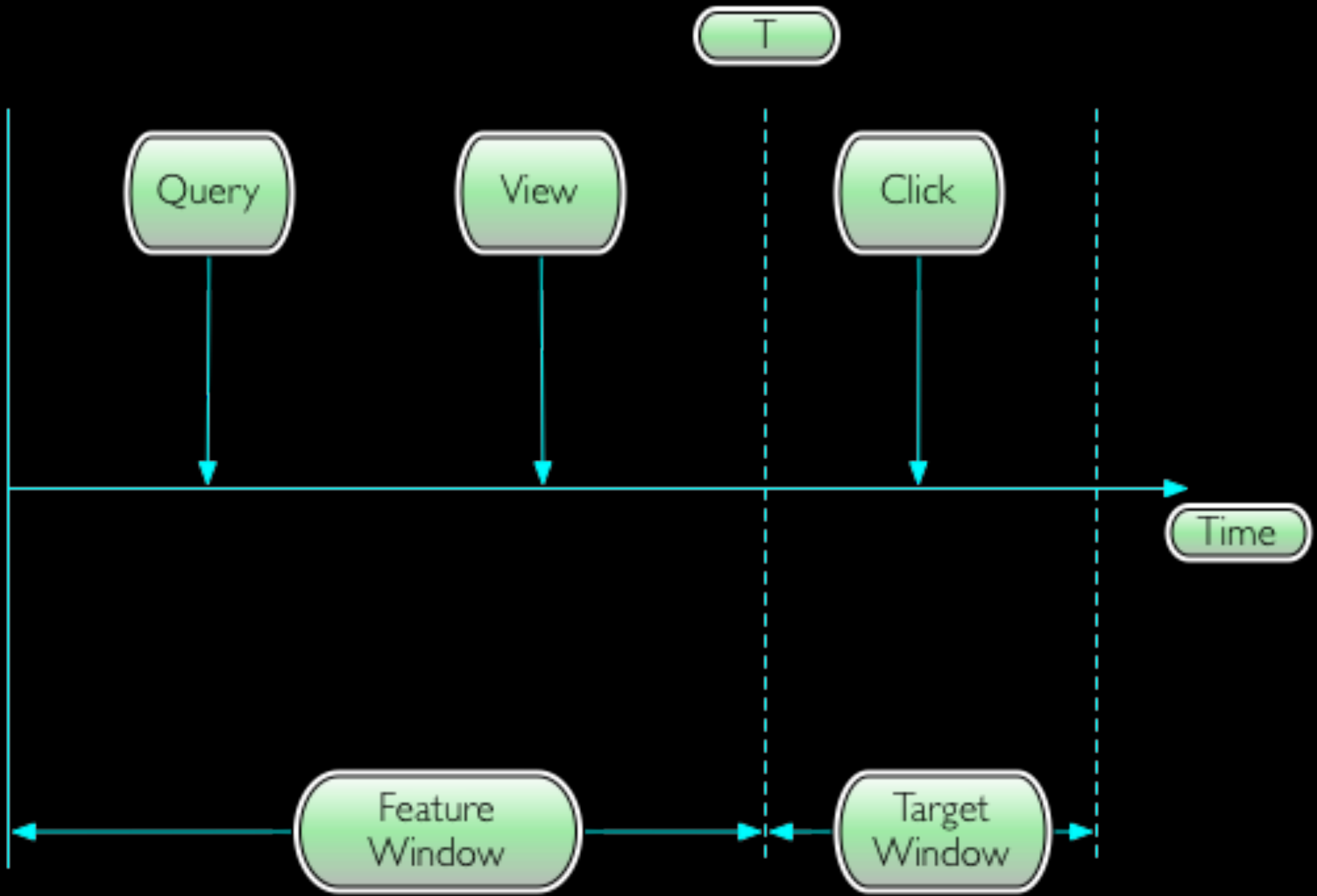
Data Acquisition

User	Time	Event	Source
U0	T0	Visited Auto website	Web Server logs
U0	T1	Searched for "Car Insurance"	Search Logs
U0	T2	Browsed stock quotes	Web Server Log
U0	T3	Saw ad for "discount brokerage", did not click	Ad Logs
U0	T4	Checked Mail	Web Server Log
U0	T5	Clicked Ad for "Auto Insurance"	Ad Logs, Click Lo

Normalization

User	Time	Event	Tag
U0	T0	View	Category: Autos, Tag: Mercedes Be
U0	T1	Query	Category: Insurance, Tag: Auto
U0	T2	View	Category: Finance, Tag: EMC
U0	T3	View-Click	Category: Finance, Tag:Brokerage
U0	T4	Browse	Irrelevant Event, Dropped
U0	T5	View+Click	Category: Insurance, Tag: Auto

Features & Targets



Targets

- User-Actions of Interest
  - Clicks on Ads & Content
  - Site & Page visits
  - Conversion Events
    - Purchases, Quote requests
    - Sign-Up for membership etc

Features

- Summary of user activities over a time-window
- Aggregates, moving averages, rates over various time-windows
- Incrementally updated

Joining Targets & Features

- Target rates very low: 0.01% ~ 1%
- First, construct targets
- Filter user activity without targets
- Join feature vector with targets

Model Training

- Regressions
- Boosted Decision Trees
- Naive Bayes
- Support Vector Machines
- Maximum Entropy modeling
- Constrained Random Fields

Model Training

- Some algorithms are difficult/inefficient to implement in Map-Reduce
  - Require fine-grain iterations
- Different models in parallel
- Model for each target response in parallel

Offline Scoring & Evaluation

- Apply model weights to features
- Pleasantly parallel
- Sort by scores and compute metrics
- Evaluate metrics

Batch Scoring

- Apply models to features from all user activity
- Upload scores to serving systems

User Modeling Pipeline

Component	Data Volume	Time
ata Acquisition	1 TB	2-3 Hours
ature & Target Generation	1 TB * Feature Window Size	4-6 Hours
lodel Training	50 - 100 GB	1-2 Hours for 1 of Models
Scoring	500 GB	1 Hour

Issues

- Different modeling techniques for different kinds of data
- Different notions of a “session”
- Widely varying number of events per entity

Proposal: 5 Classes

- Tiny (100K entities, 10 events per entity)
- Small (1M entities, 10 events per entity)
- Medium (10M entities, 100 events per entity)
- Large (100M entities, 1000 events per entity)
- Huge (1B entities, 1000 events per entity)

Proposal: Publish results for every stage

- Data pipelines constructed by mix-and-match of various stages
- Different modeling techniques per class
- Need to publish performance numbers for every stage

Questions ?