Large Scale Mapreduce Data Processing at Quantcast

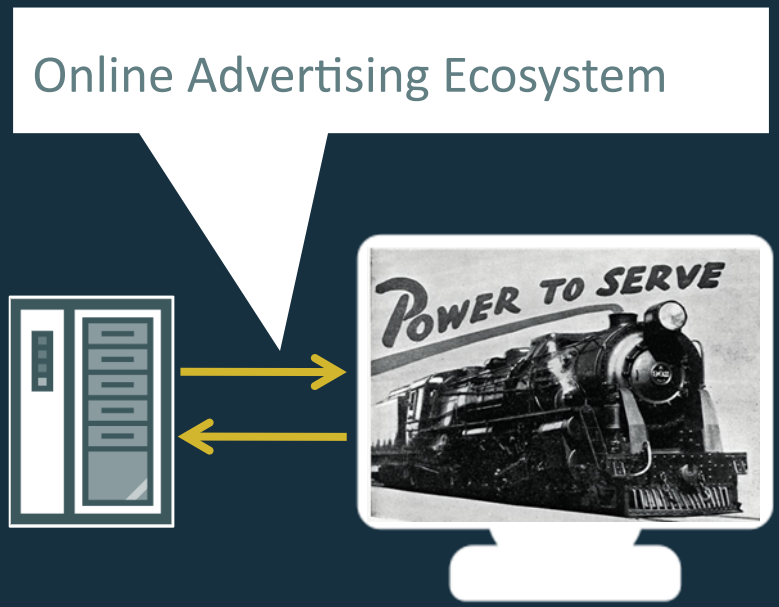
Ron Bodkin, Think Big Analytics
rbodkin@thinkbiganalytics.com

Outline

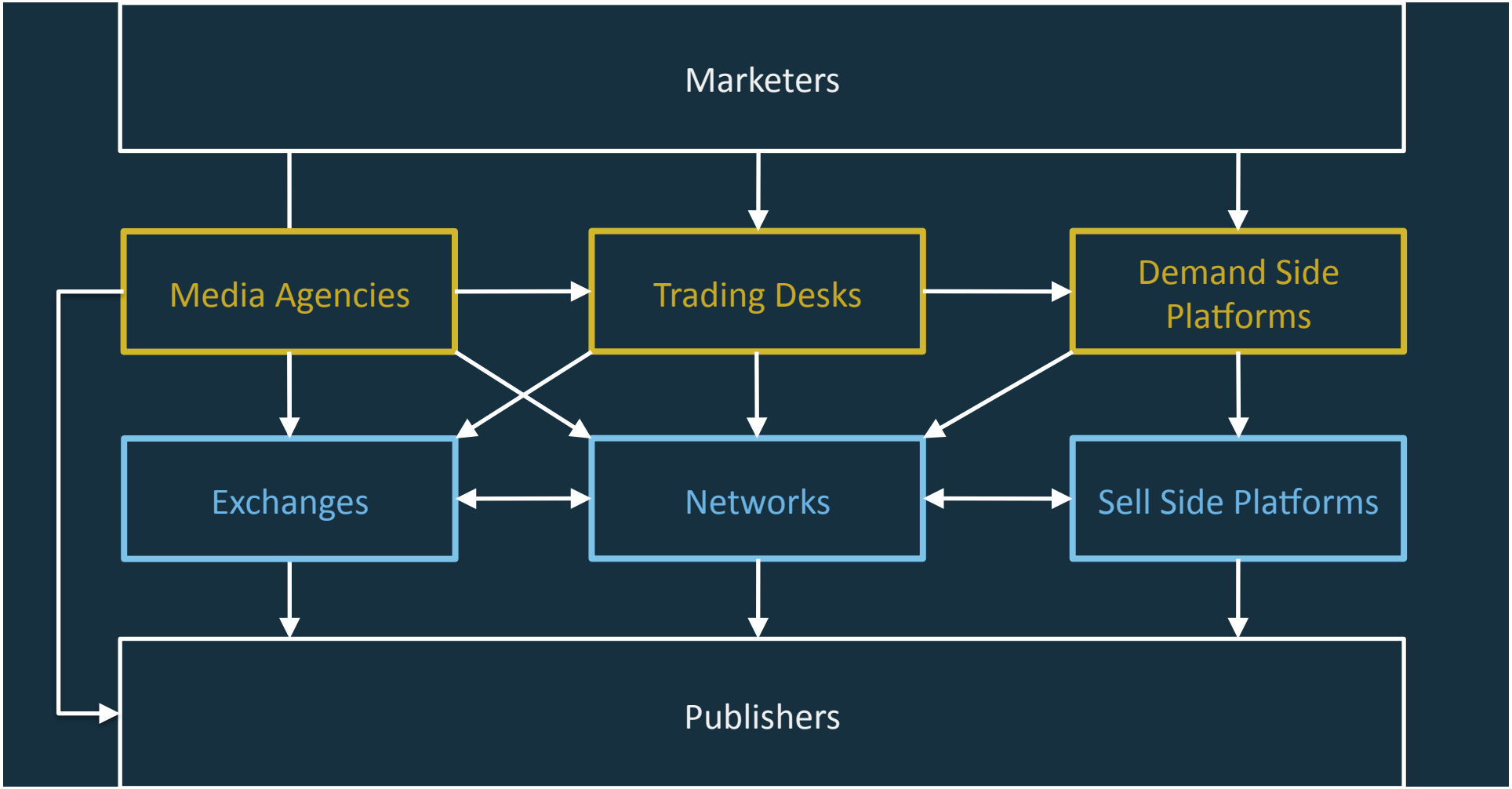
- An Internet-scale business
- Applications
- Data
- Infrastructure

The Personalized Media Economy

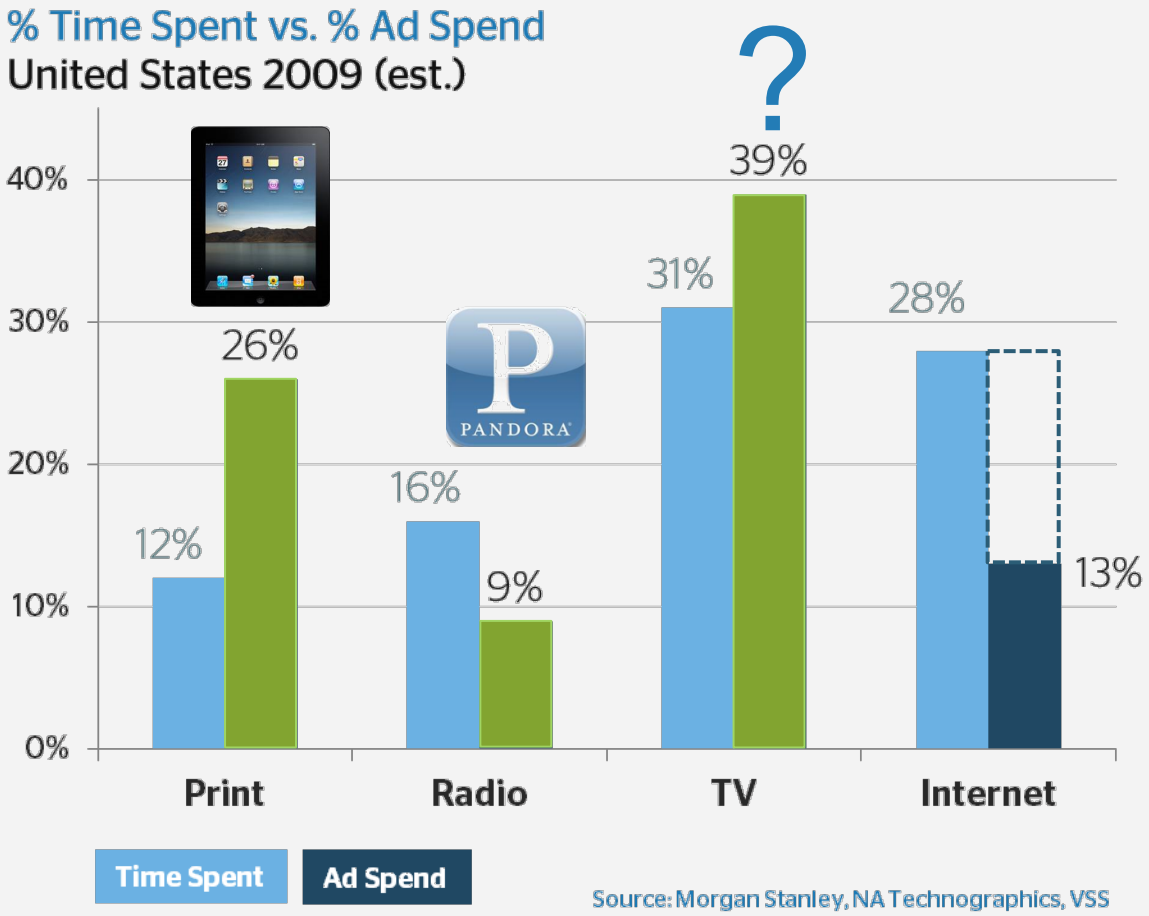
Media is transitioning from a “one size fits all” broadcast model to dynamic real-time choice



Online Advertising Ecosystem



Money Follows Media Consumption



Globally, hundreds of billions of dollars of ad spend will shift

$30B

opportunity

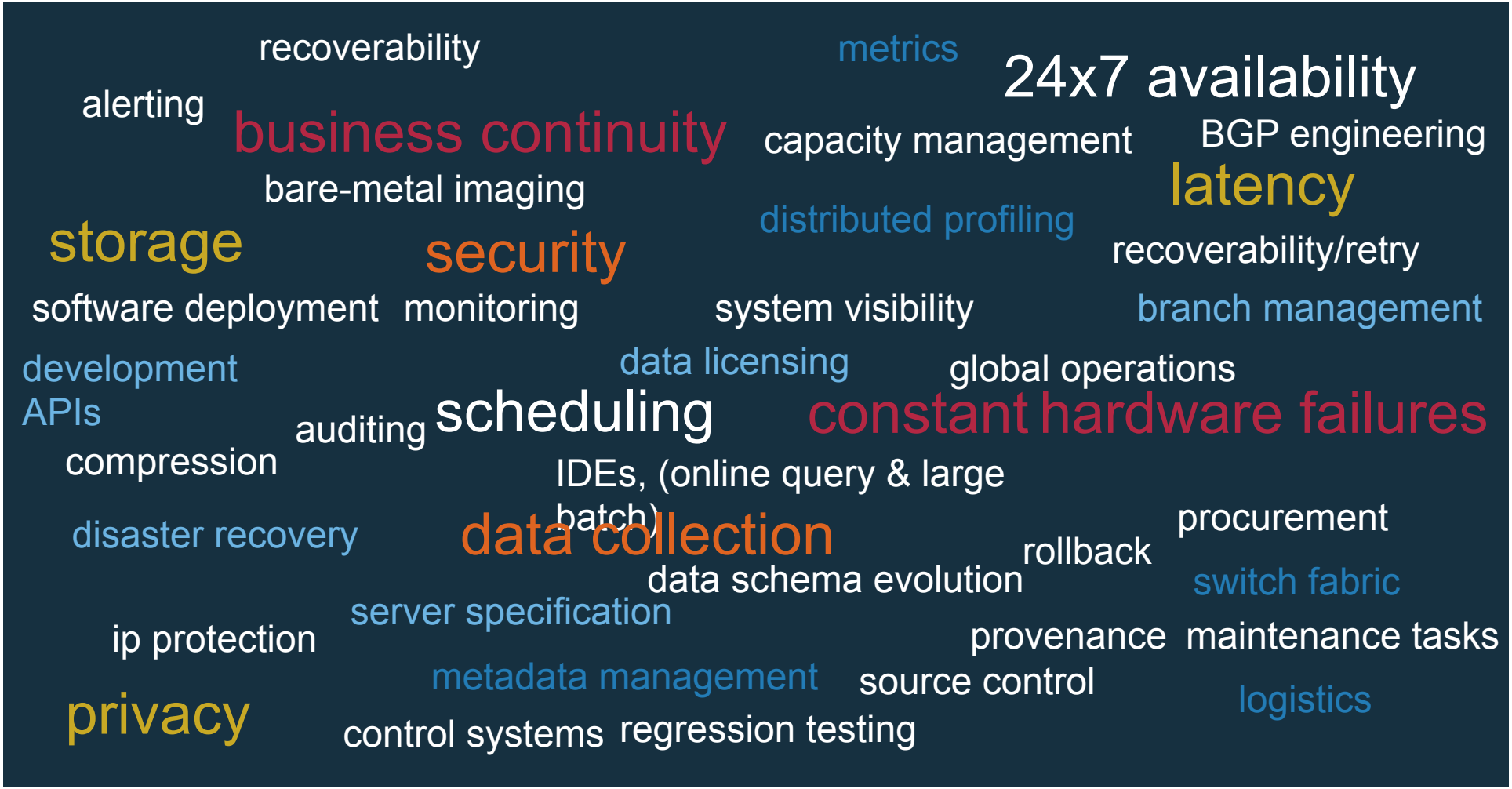
Enter Quantcast

- Launched September 2006 to enable addressable advertising at scale
- First we had to fix audience measurement
- Launched a free service based on *direct measurement* of media consumption
- Use machine learning to infer audience characteristics

Industry Adoption of Quantcast

World's Favorite Audience Measurement Service

Numerous Challenges



Data Rich Environment

300+ Billion / month
Media consumption events

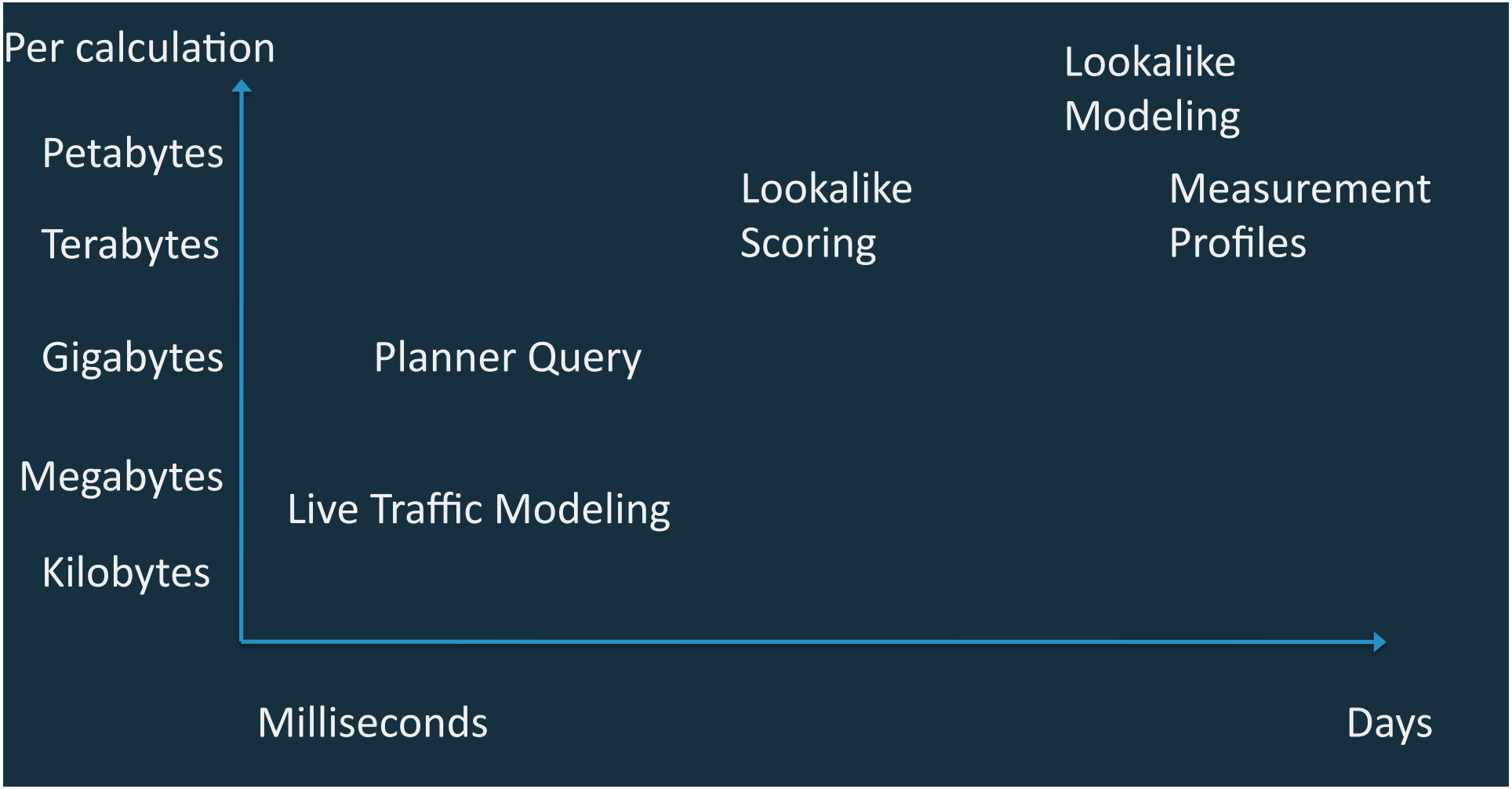
1.2 Billion
Internet users / month

Millions of Sites
Continually measured

300,000
Transactions / second

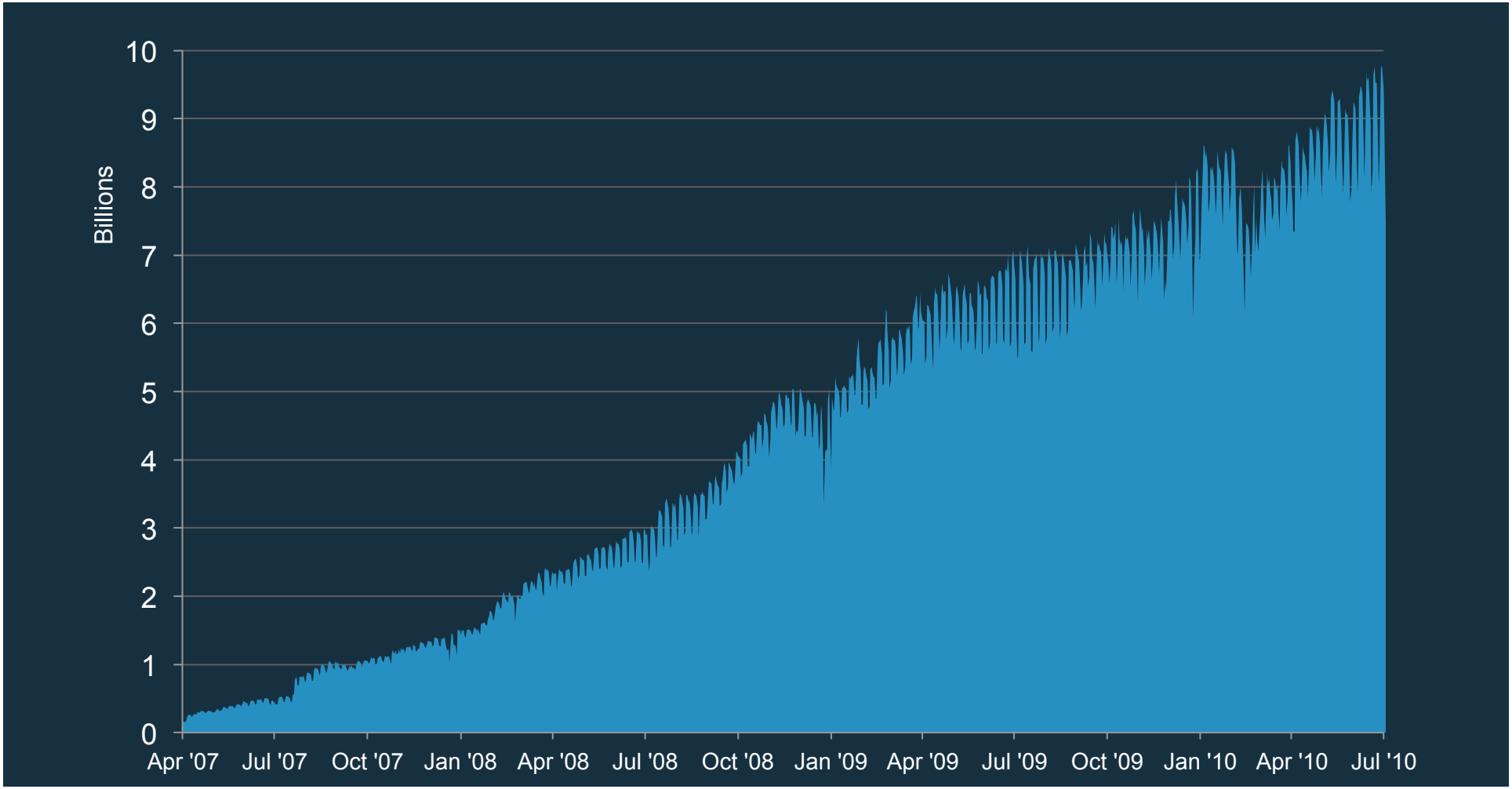
1.5 PB
Every Day
Avg. Processing Load

Response Times



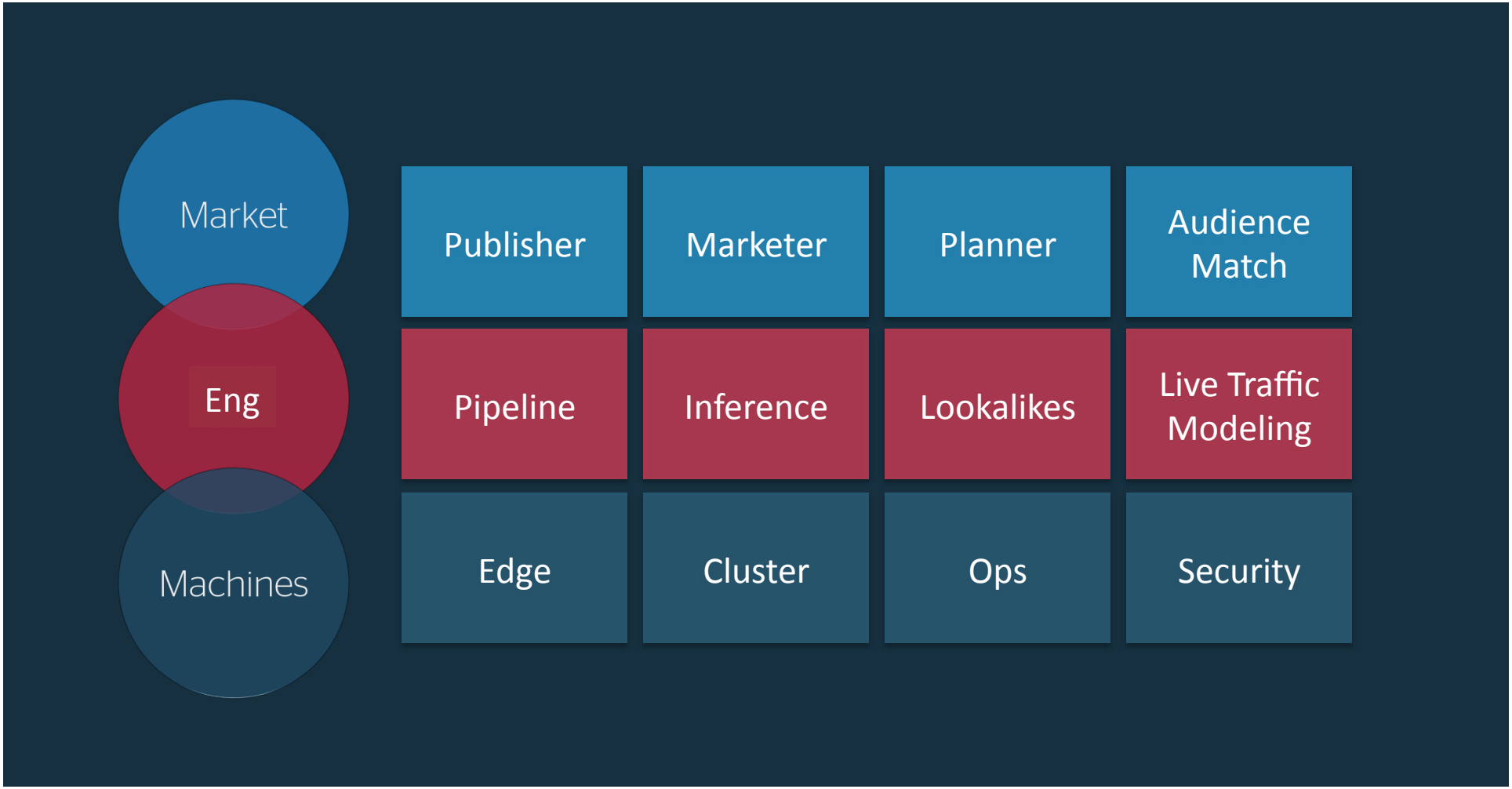
Organic Data Growth

Daily Data Volumes

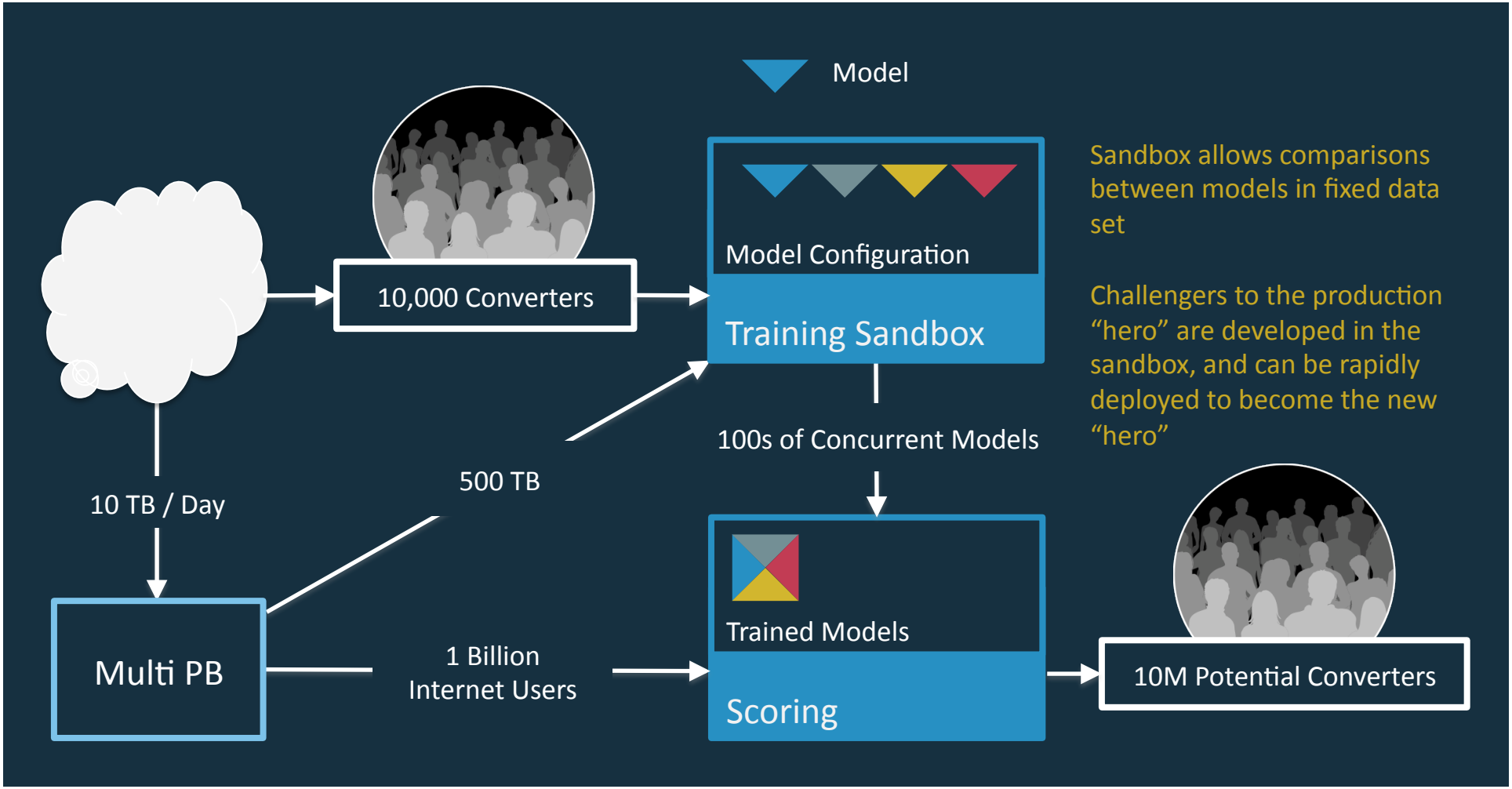


Divide and Conquer

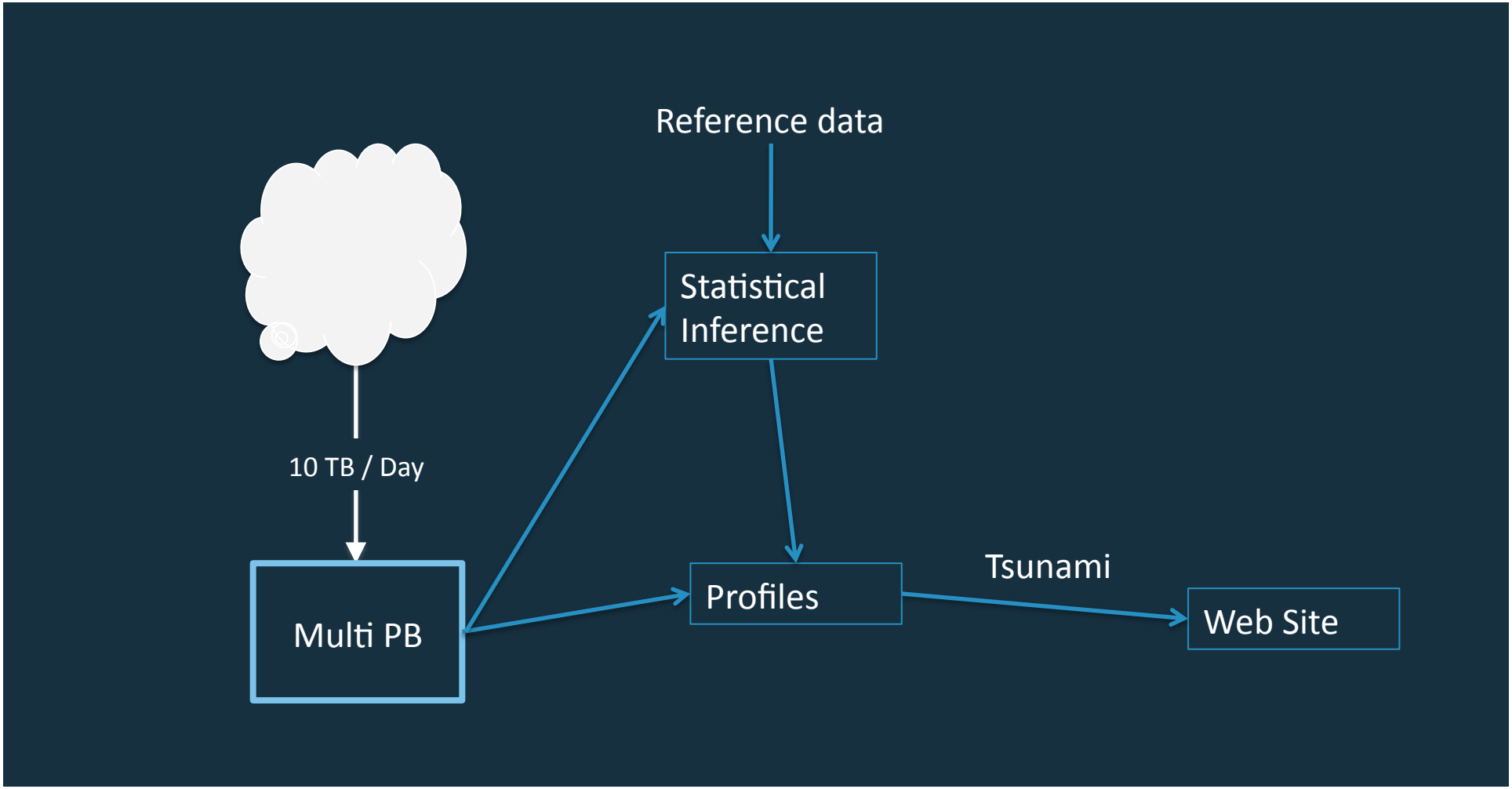
Quantcast Functional Teams



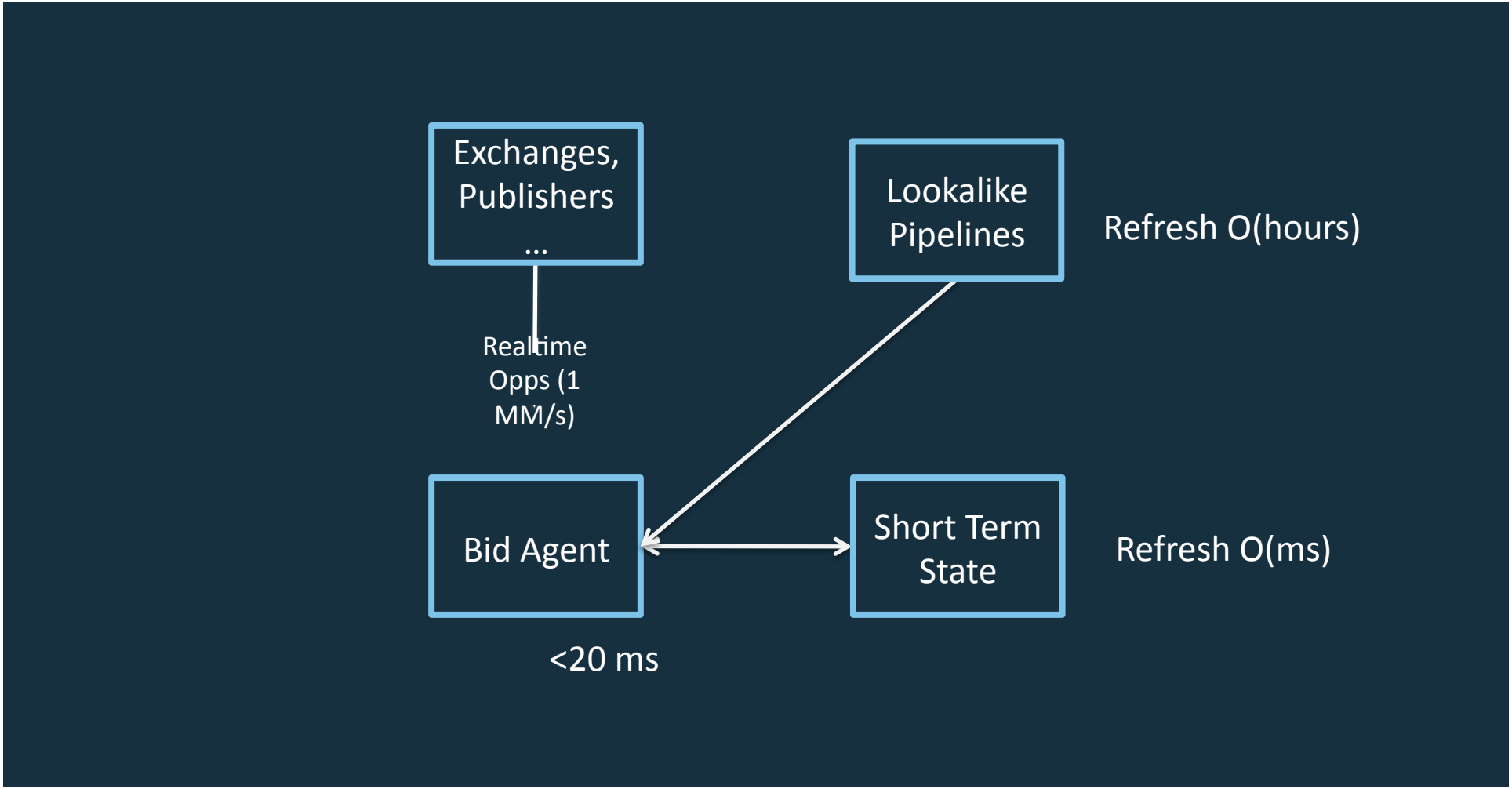
Lookalike Pipeline



Measurement Pipelines



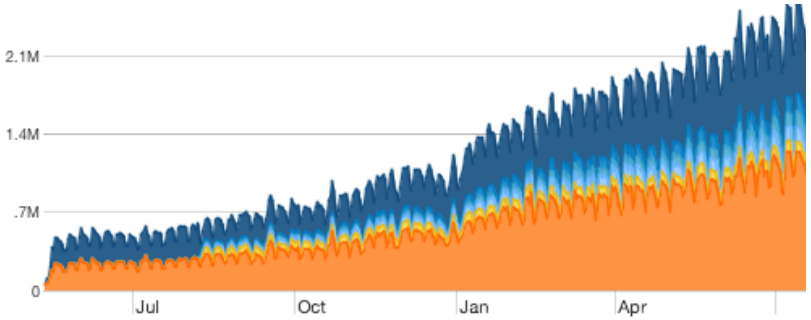
Live Traffic Modeling

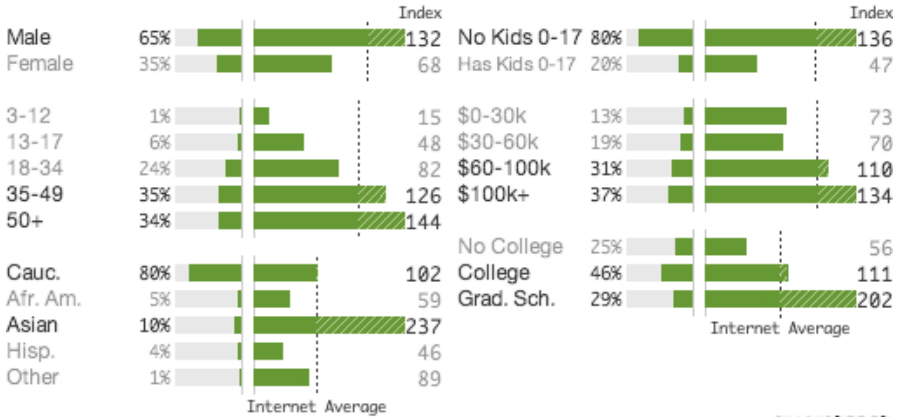


Technology Stack

Quantcast Reporting

Billion users on millions of sites





Learning ∝ experimentation

6 Hours — To process 100TB with new Map-Reduce job

2 Days — New model development

Mins — New model in production

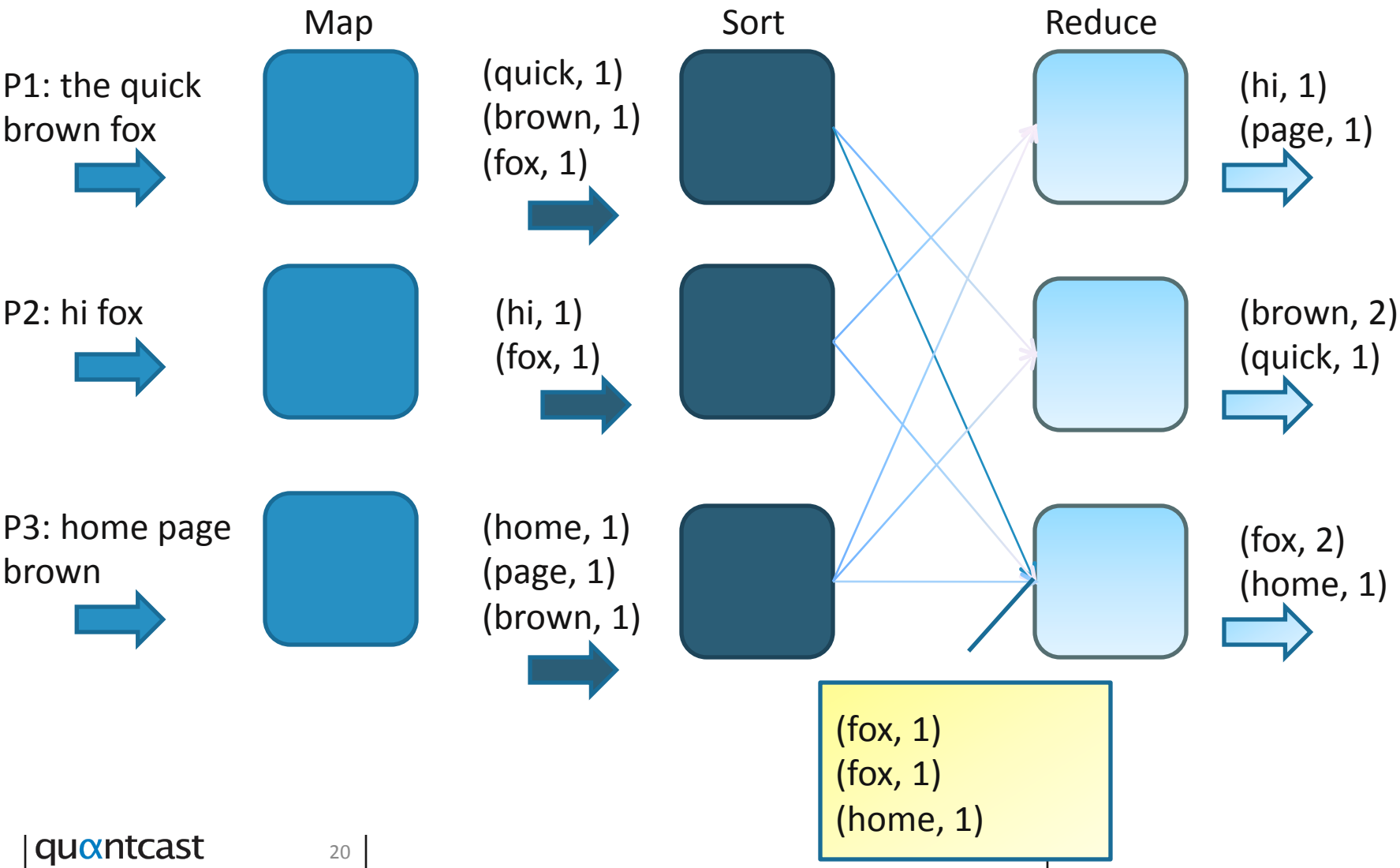
Hours — Live performance assessment

2 Weeks — To influence billions of real-time decisions every day and millions of dollars of advertising spend

Analyst Tools

- Greenplum SQL Database used for analysis
  - Valuable secondary use for quick iterative exploration
  - Allows parallel import for reasonable speed (this has been an achilles heel for traditional SQL databases)
  - Used to manipulate tables up to 1-2 TB
- Often hard to load/unload data quickly at TB scale
- When evaluating BI/DB
  - make sure you test requests that exceed memory cache
  - and run parallel workload
- Evaluating Datameer, Hive, PIG, Goto Metrics

Map-Reduce 101



…Distributed Sort Execution

- Map-Reduce provides a programming model
- The heart is sorting data and grouping:
  "the dash in map-reduce"
- Map-Reduce allows for efficient sequential disk access in processing
- Not as abstract as SQL
- More flexible than custom data flows

Hadoop

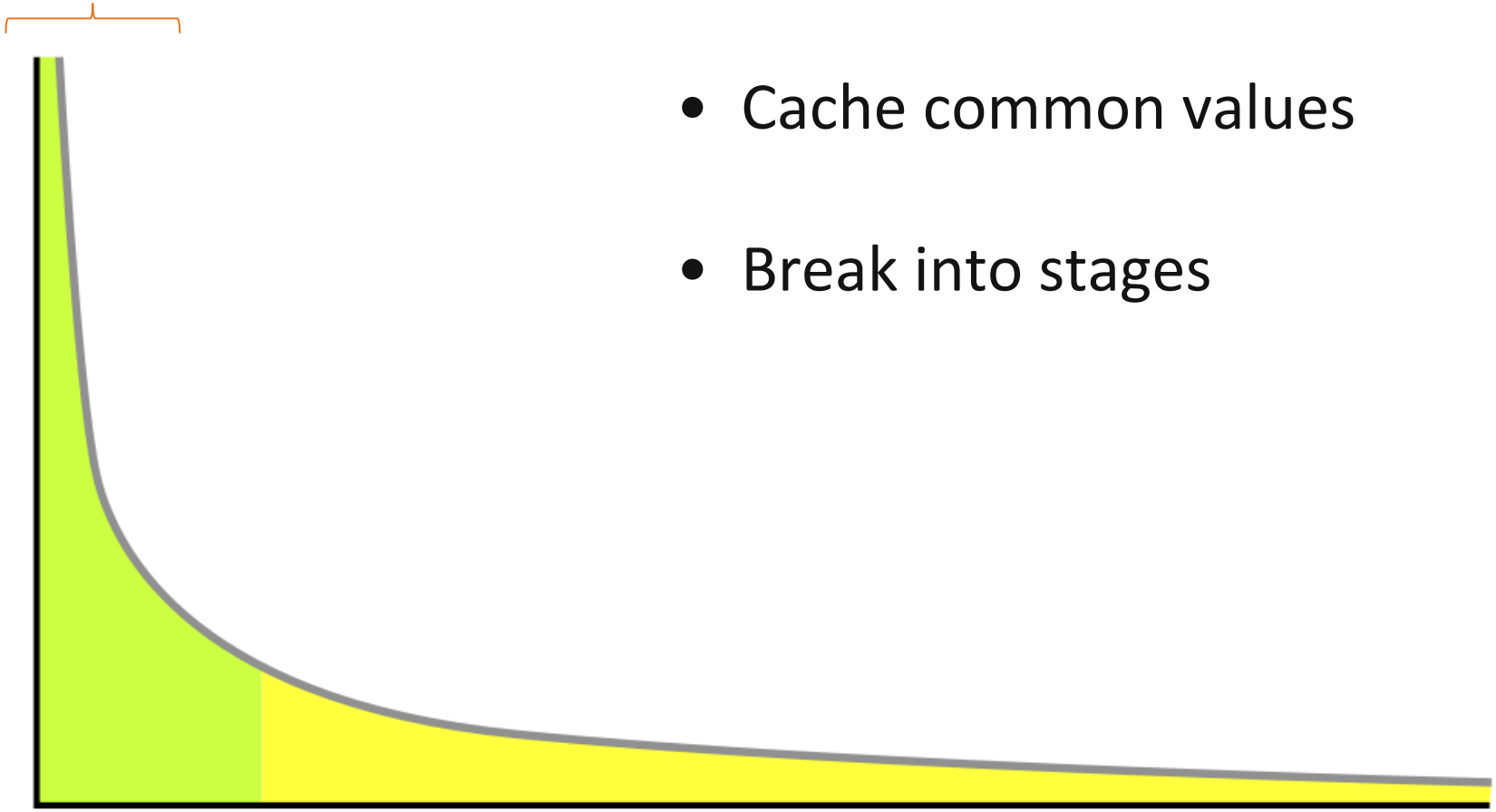
- Open source Java-based map-reduce
  - Scheduling
  - Execution
  - Monitoring
  - Recovery
- Core includes a distributed file system
- Came from Nutch project (open source crawler)

Custom Evolution at Quantcast

- Relational Map/Reduce (RMR) and MRFlow: provide productive API
- Replaced the Hadoop data path, including sort
- Need to get more processing out of hardware
- Hadoop project focused on scaling to 10,000 machines first, then addressed performance
  - E.g., Yahoo uses 10,000 cores for web map (450 TB sort)
- Push common patterns down into the platform

Power Law Distributions

Naïve jobs will take forever to process these values



- Throw out “stop words”
- Cache common values
- Break into stages

Other Compute Lessons

- Scale up gradually
  - starting locally
  - it's easy to waste massive resources in a big cluster
- Job chains get complex
  - Need for good workflow, dependency management, scheduling
- Managing data and tracing creation
  - Always a problem in any data system
  - With a more universal storage cloud, visibility improves
  - We keep metadata and keep refining APIs to support
- Anomaly detection
  - controls

Sidebar: Colossal Pipe

- 3rd generation *open source* map/reduce framework
- Built by ThinkBigAnalytics
- Builds on API lessons learned from RMR and MRFlow
- Dependency analysis
- POJO-style and fluid programming
- Avro and JSON object bindings
- https://github.com/ThinkBigAnalytics/colossal-pipe

Data Storage

- You can't back up 100 TB of cluster data "offline"
- Data corruption from distributed file system bugs is a major risk
- Erroneous jobs and operators are the next biggest risks

Distributed Filesystems

- Quantcast runs two different filesystem implementations (on the same machines)
- KFS – open source high performance distributed FS
- Hadoop DFS – open source component in Hadoop
- Both support data replication to *n* machines
- Both are simple: reliability > complex "features"
- Evaluating commercial offerings from MapR, Goto Metrics, Appistry…

Data Access Patterns

- Map-reduce leverages the efficiency of sequential access (you can shard to speed it up)
- Random-access is appropriate for near real-time access
- Indexing typically required for random reads, can be used to speed up sequential reads too
- Flash (SSD) can be a great enabler for random *reads*
- Join access patterns vary depending on strategy (merge, hash, nested loop)

High Performance Platform

Multiple Global Datacenters

Ultra-high availability with advanced traffic management

225 Thousand / Second

Real-time events

2PB / Day

Analytical throughput

Compute as a Utility

PixelTone

Availability & latency of real-time architecture

Ensure self-healing, rapid failover of our multi-master design (no single point of failure)

LogTone

Management of data consolidation – Freshness

Have to 'close the books'

ClusterTone

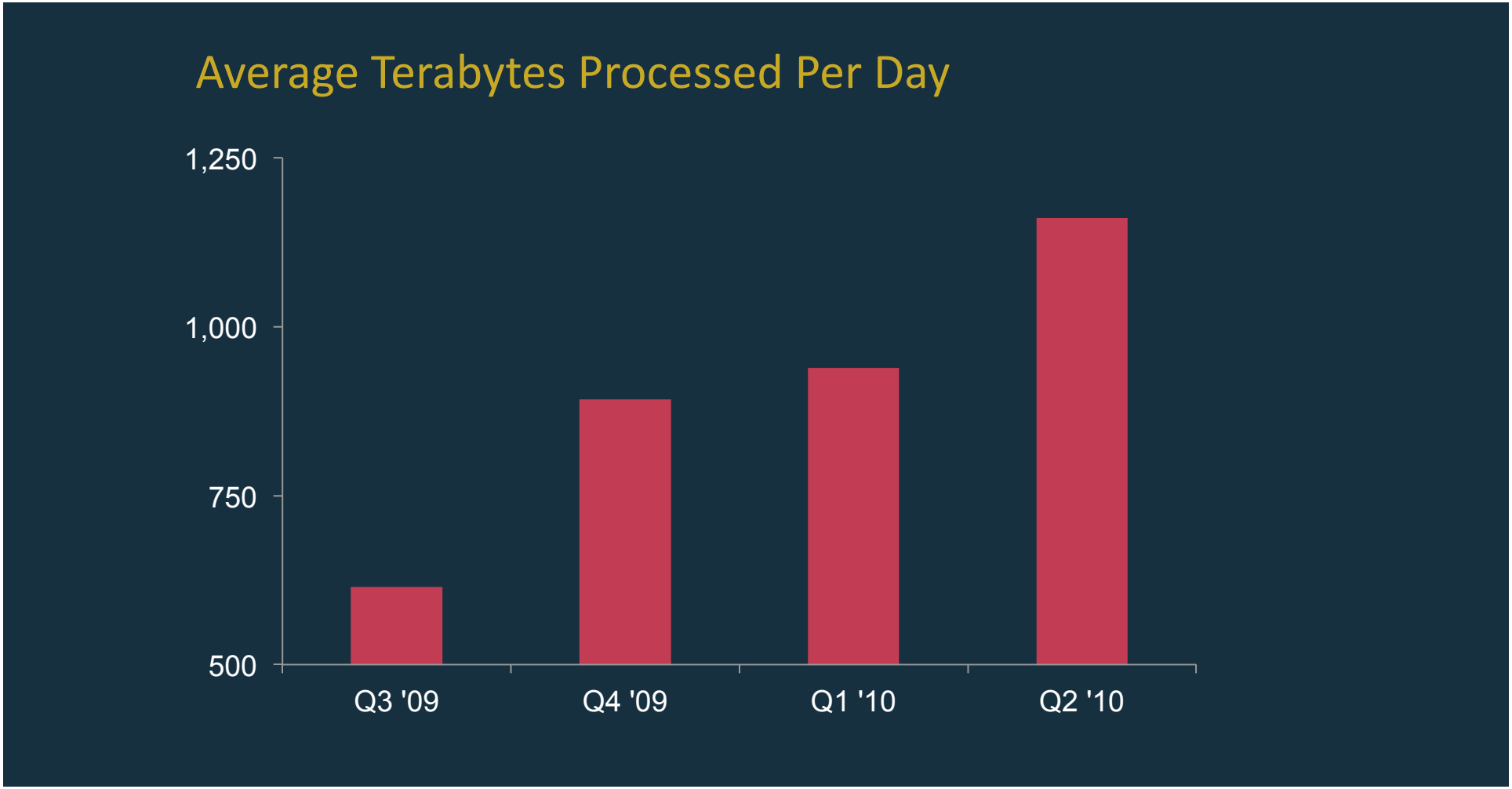
Utilization, performance, reliability and efficiency of our massively parallel distributed processing platform

AppTone

Performance, latency & user experience of all customer facing applications

Constantly Scaling Infrastructure

Whatever you have, its never enough



Quantcast Compute Environments

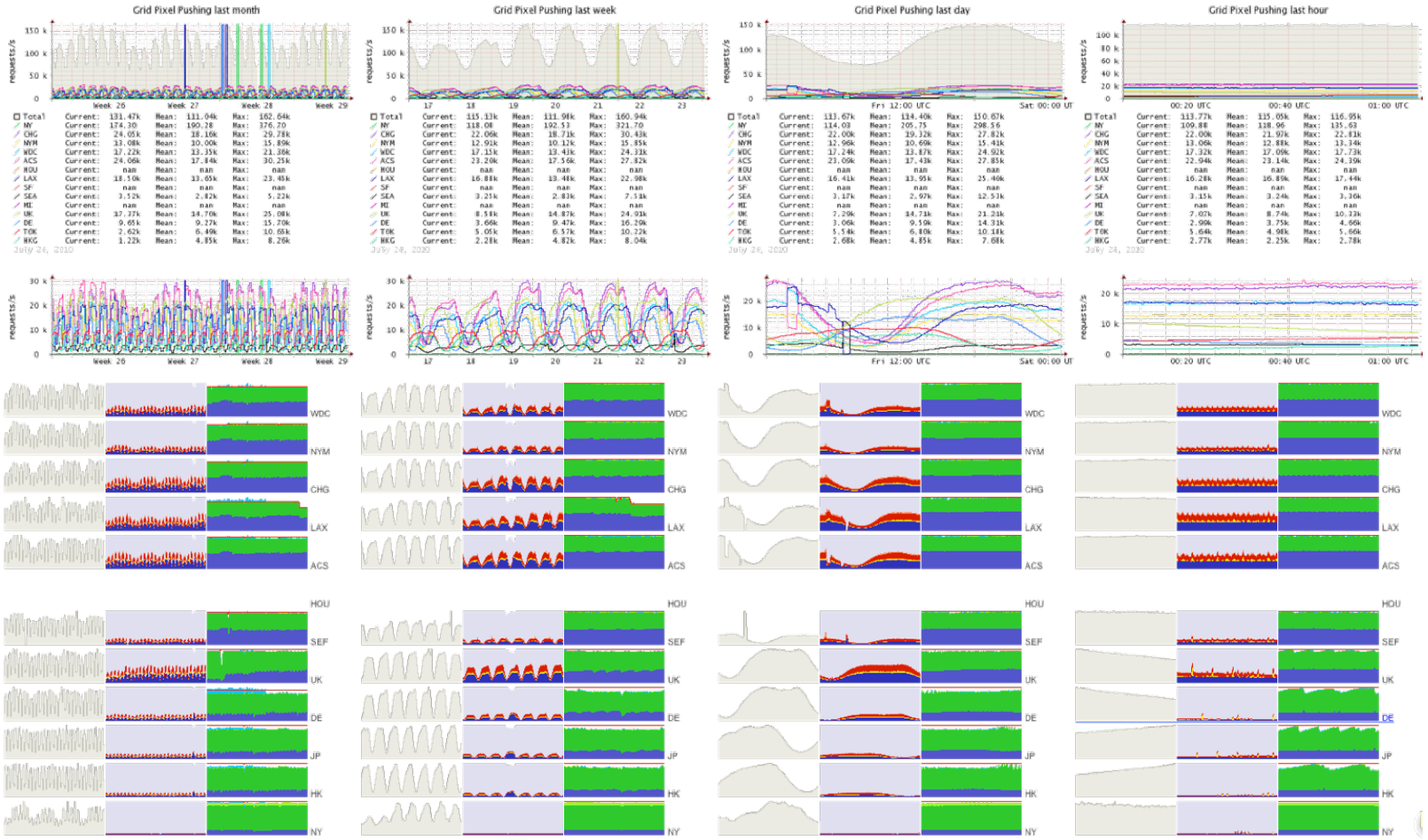
- Thousands of nodes in 13 data centers globally
- Eight people to manage and develop code
- Compute cluster: 3,000 cores, 3PB storage
- Volume commodity components and use smart software to provide reliability
- Costs of people, machines, and hosting equal
  - Power costs: $.06/kwH in Seattle vs. $.15/kwH in SF

Cloud Use: Amazon EC2

- Running RMR + Hadoop, about 50% of throughput of hosted cluster
- Provides disaster recovery capability (back up copy of data)
- Available for surge capacity
- Matured a lot from 2008 tests with network problems

Comprehensive Monitoring

350K metrics updated every 15 seconds



Management Challenges

- Rare events happen *all the time*
  - Undetected bit errors in SATA disks (10^{-14})
  - Buggy network controllers (how bad?)
  - Bit errors in network transfer
  - Memory errors
  - Must checksum all data
    (e.g., turn on memory error detection)
- Need to detect and avoid slow components…
  - Slow for an instant? (<1 s)
  - Intermittently (minutes)
  - Chronically (hours to forever)

Summary

Quantcast has built a business applying big data and analytics to change online advertising, leveraging:

- Large scale map-reduce processing for large scale
- Near-time response for responsiveness
- Metrics-driven data science
- A variety of data sets