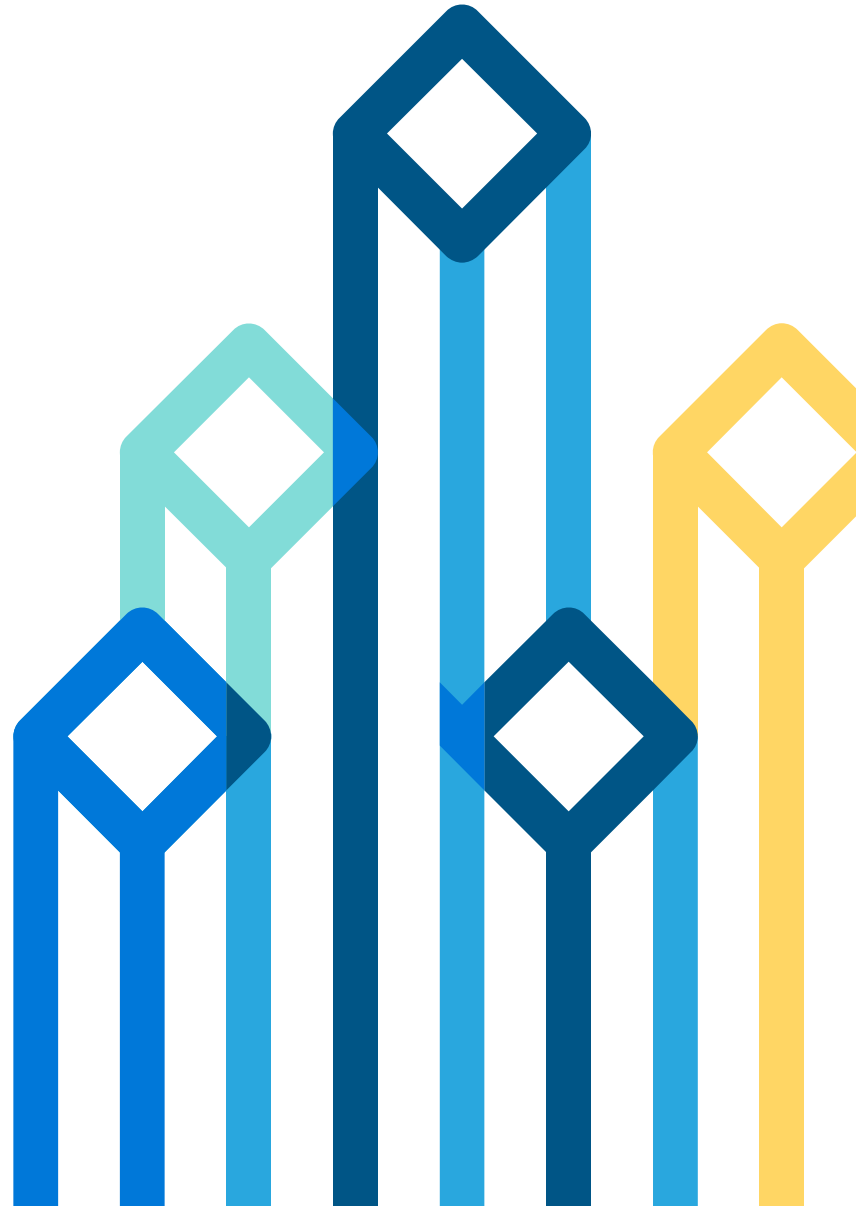
cloudera®

Five Challenges for Energy Efficient Computing Research

Yanpei Chen, Software Engineer, Performance Team

NSF Workshop on Sustainable Data Centers 2015



Our mission:

Cloudera helps organizations profit from all their data

cloudera

Cloudera company snapshot

Founded	2008, by former employees of 
Funding	$670M cumulative investment
Employees Today	800+ worldwide
Mission Critical	Production deployments in run-the-business applications worldwide – Financial Services, Retail, Telecom, Media, Health Care, Energy, Government
Diverse Customers	Majority of Fortune 100 companies are Cloudera customers
Cloudera University	Over 40,000 big data professionals trained
Open Source Leaders	Cloudera employees are leading developers & contributors to the complete Apache Hadoop ecosystem of projects

Protecting consumers from fraud

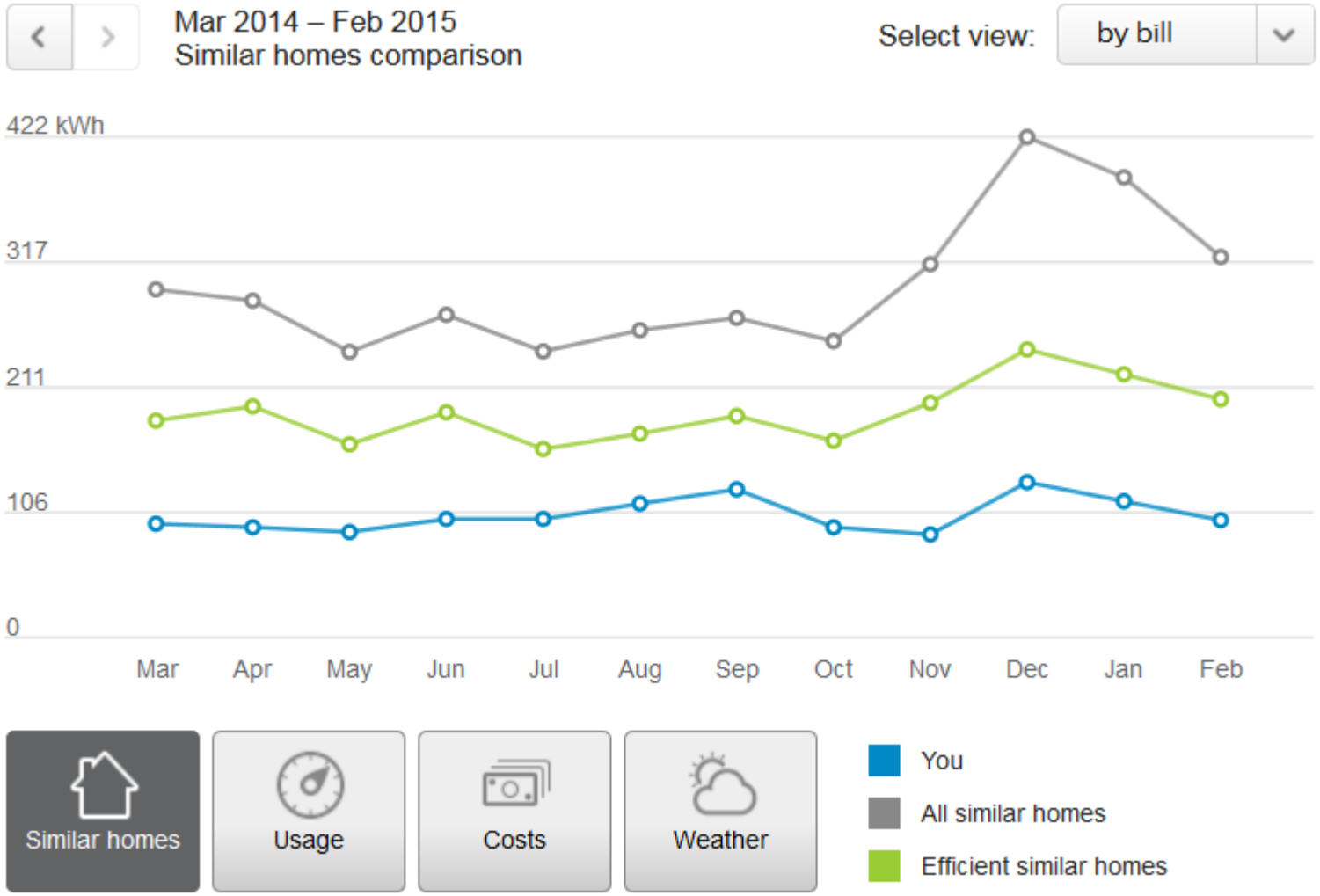
Credit card companies use Cloudera Data Hub to analyze timing, location, $ amount of transactions to distinguish normal and fraudulent behavior *for each customer*. Caught largest fraud case in a provider's history.

cloudera

Improving neonatal care

The Children's Healthcare of Atlanta uses Cloudera Data Hub to monitor 24/7 the light, noise, patient vital signs in their neonatal wing. Improved care by adjusting environmental factors. Found ways to improve pain management.

Reducing electricity use



Pubic utilities use Cloudera Data Hub to make visible residential electricity use at a per-hour granularity.

Led to behavior change that saved 2 terawatt-hrs globally in 2013, average 1-3% reduction.

cloudera

Three of Fortune 100 are consumer Internet companies: Google, Amazon, Apple. Cloudera and big data trace our technical heritage there, and consumer Internet is an important use case. There is also a much bigger world of “big data” beyond these companies!

Diverse Customers

Majority of Fortune 100 companies are Cloudera customers

What Cloudera sees re energy efficiency

- Energy efficiency important to our largest (~5%) of customers
- All customers rapidly expanding clusters, sometimes 2 expansions per yr
- Hence expect growing interest in energy efficiency

How Cloudera can contribute

- Provide insights/workloads on customer user cases
- Influence how energy efficiency is measured in industry benchmarks
- Channel energy efficiency improvements into open source

My past work

- MapReduce energy efficiency (2009)
- Realized we can't run/measure stuff for real, nor at scale
- MapReduce workload capture & replay (2011) – 5x Cloudera customer workloads
  - Validated Hadoop fair scheduler (2011)
  - MapReduce energy efficiency (2012)
  - TCP incast fix validated on MapReduce (2012)

- Performance engineering at Cloudera - "make things go fast"
  - Work on SQL-on-Hadoop, Search, Resource Mgmt, MapReduce, HBase

Challenge 2:

How can energy efficiency measurement methods capture realistic behavior?

Many hard design challenges arise from the dynamics of the serviced workload over time. Efficiency and energy efficiency measurement methods can distort technical merit when they measure an unrealistic good case or corner case.

Challenge 3:

How do we design for common workloads given that we do not know what is common?

Companies such as Google, Facebook, Twitter have visibility into their own workload as a single case study. Vendors such as Cloudera see many customer workloads but do not and should not have access to their proprietary information. How do we progress and avoid a multitude of point solutions that do not generalize? This is a concern beyond energy efficiency.

Challenge 4:

How do we design for energy efficiency without distorting cluster operations in a way that vastly decreases business value?

Many energy efficiency techniques place constraints on cluster operations that vastly decreases business value. Such constraints relate to how data is placed, when do the jobs/queries get executed, and the performance of the jobs/queries. Energy efficiency must be achieved while still serving the customers' businesses needs.

Challenge 5:

How can we design for energy efficiency at large scale?

Many design challenges are challenging only at scale. Customers ask for proof-points on clusters of 100s of nodes. The demand for scale and the cost of proof-of-concept at scale increases continuously. How do we proceed? This is a concern beyond energy efficiency.

Challenge 1:

How is energy efficiency different from regular computational efficiency?

Even without considering energy, the more efficient system in the traditional sense will lead to the same hardware being able to serve a larger workload. Hence for the same workload increase, slower capacity increase, hence less energy spent. PUE approaches 1 means computational efficiency and data center efficiency converge. So do we need to be concerned with energy efficiency at all, or is it already fully incorporated within traditional measures of computational efficiency?