

How to Go From Big Data to Big Insights

Stanford Engineering “Big Data
for Energy” Lecture Series

Tuesday, May 14, 2013



Presenters

Drew Hylbert

VP, Technology and
Infrastructure



Jeff Kolesky

Chief Software Architect

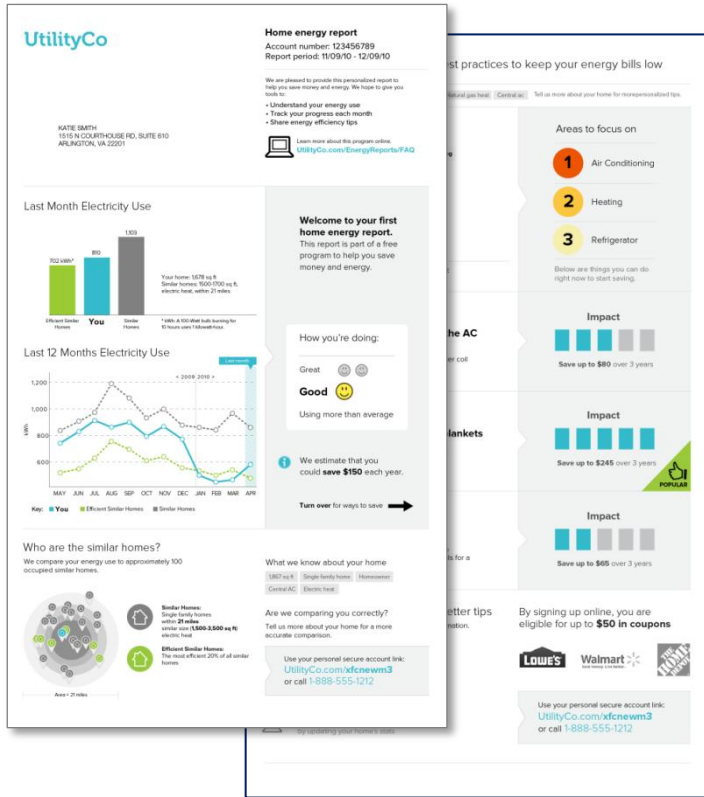


COMPANY OVERVIEW

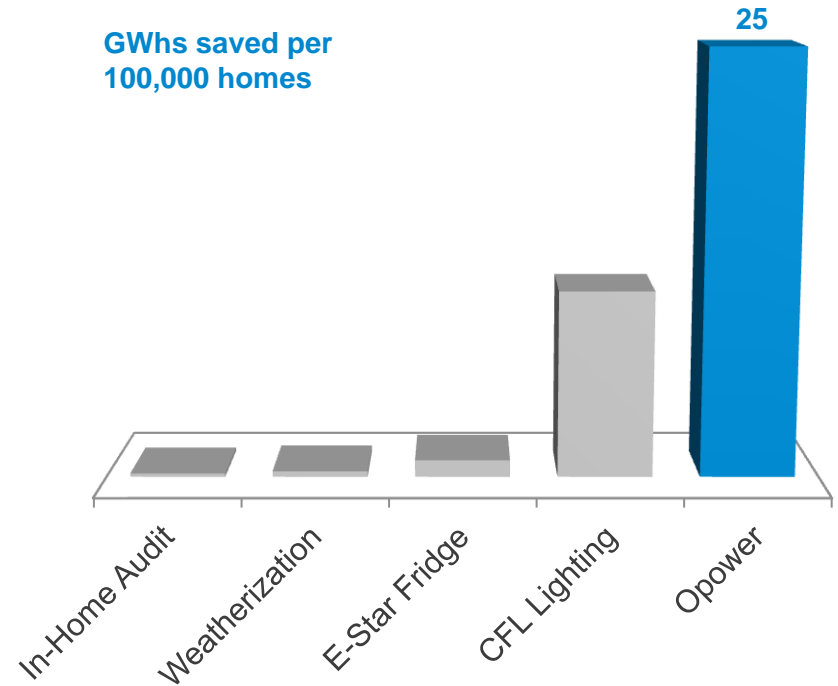
How we started: behavior change at scale

Pioneered Home Energy Reporting....

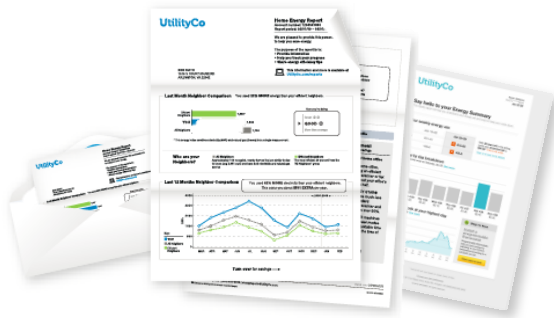
...And a New Type of Energy Efficiency



GWs saved per 100,000 homes



We've since added more points of interaction



Energy reporting



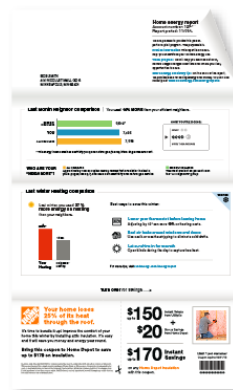
Web, mobile and alerts



Facebook



Call center



Retail marketing



Home Energy Management Systems

Opower today

The world's leading Customer Engagement Platform for utilities

The Company

- Serving leading utilities in **6 countries**
- Forbes **#10** of 100 **Most Promising Companies**
- **300** people in Washington, San Francisco, London, Singapore



Our DNA

- Behavioral science software
- Data analytics
- Consumer marketing
- User-centric design

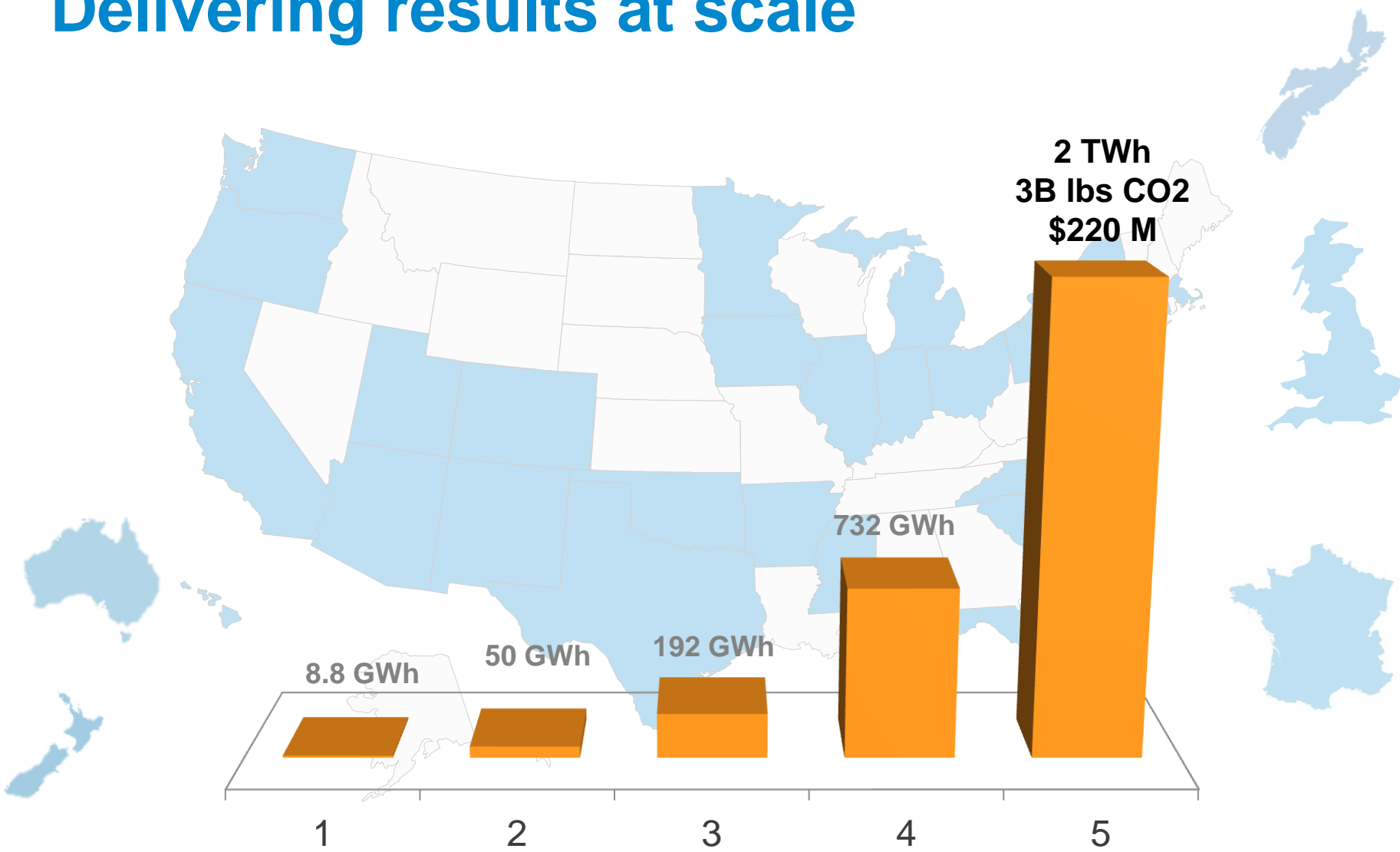
Technology Investment

- \$25M R&D investment annually
- World-class partners: Facebook, Honeywell, Home Depot, Best Buy

Our Global Footprint: 82 utilities, ~50M homes

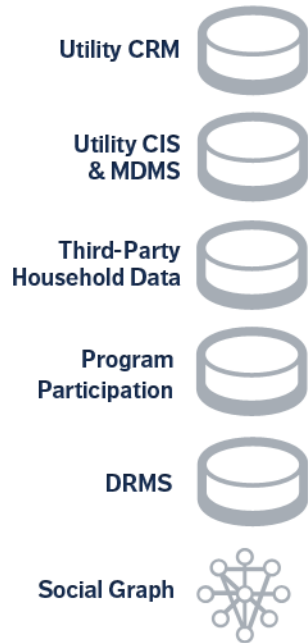


Delivering results at scale

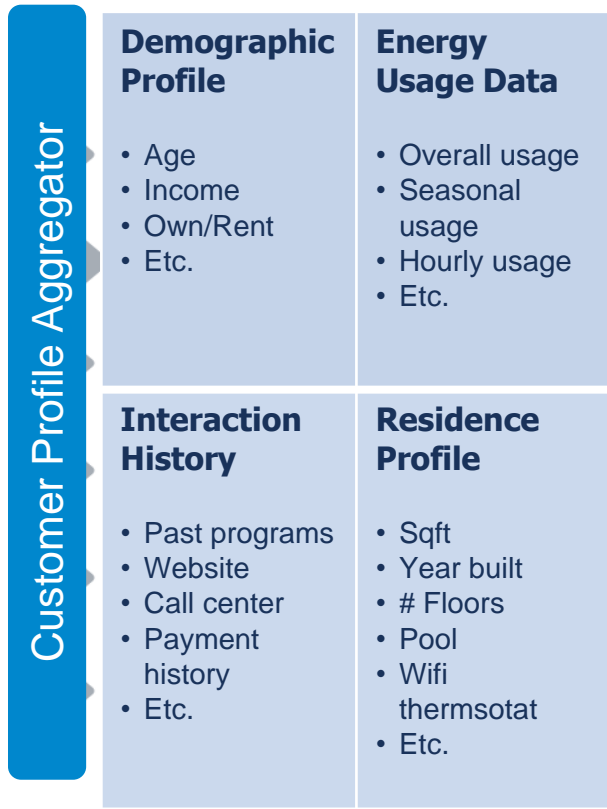


Deep analytics make all the difference

Utility & Third-Party Data



360 Degree Customer View



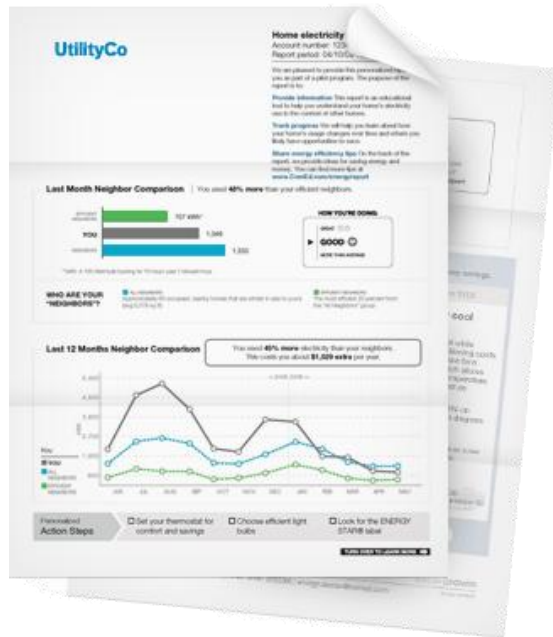
Energy Insight Engine

Actionable Customer Insights

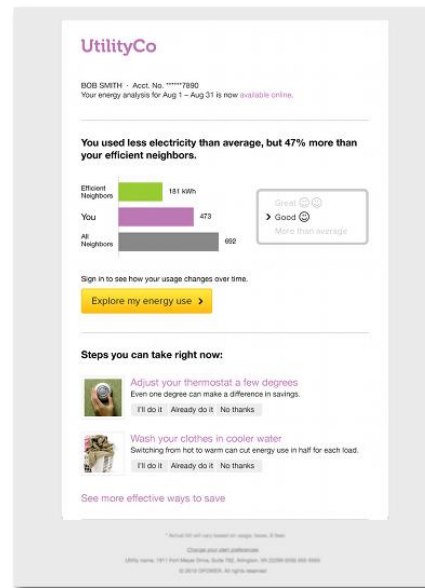


Export to Utility CRM

Push Insights, enabled by Big Data



Home Energy Reports

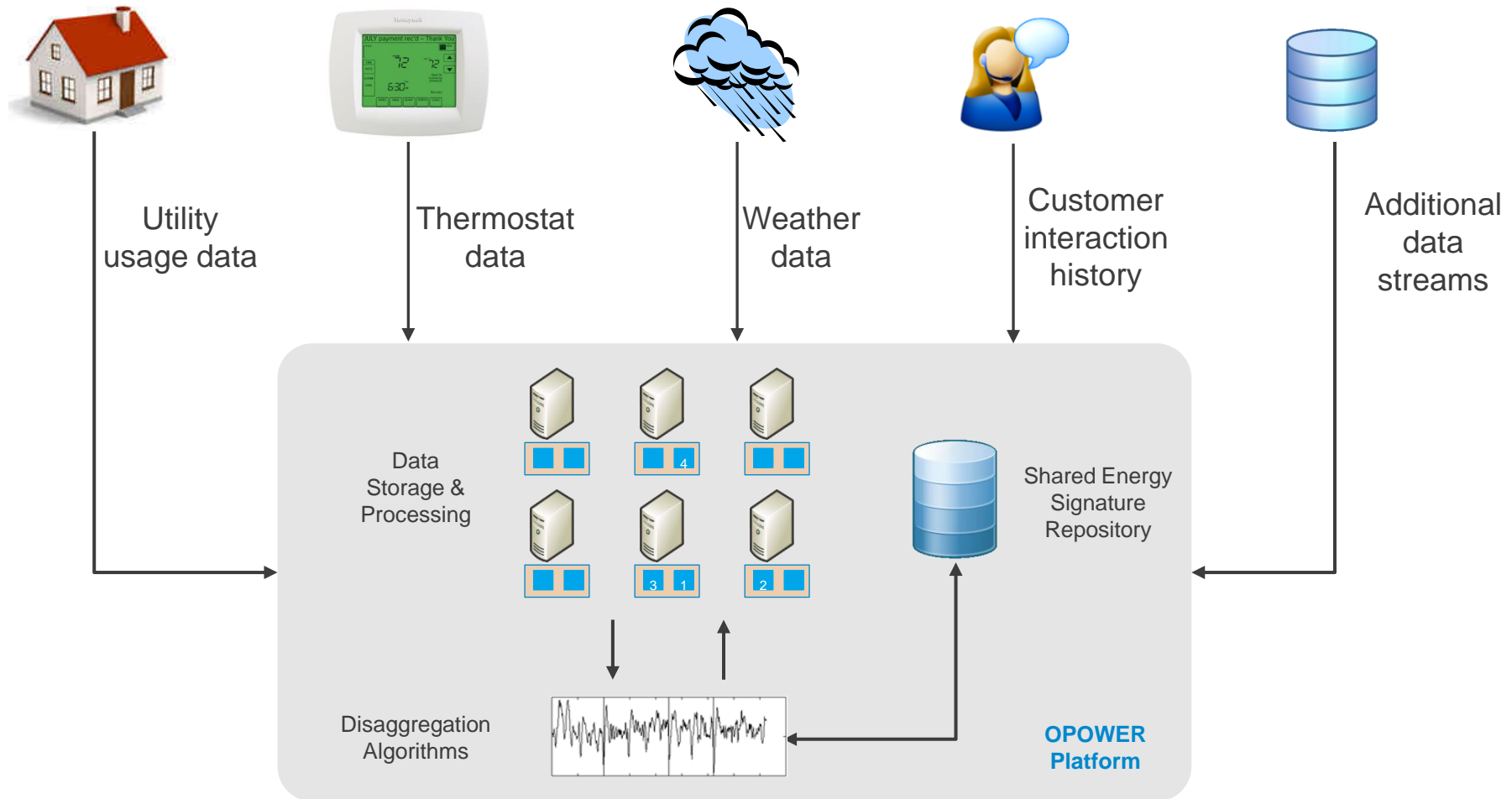


Monthly emails



Usage Alerts

Our analysis relies on data from a variety of sources



Opower Data Infrastructure

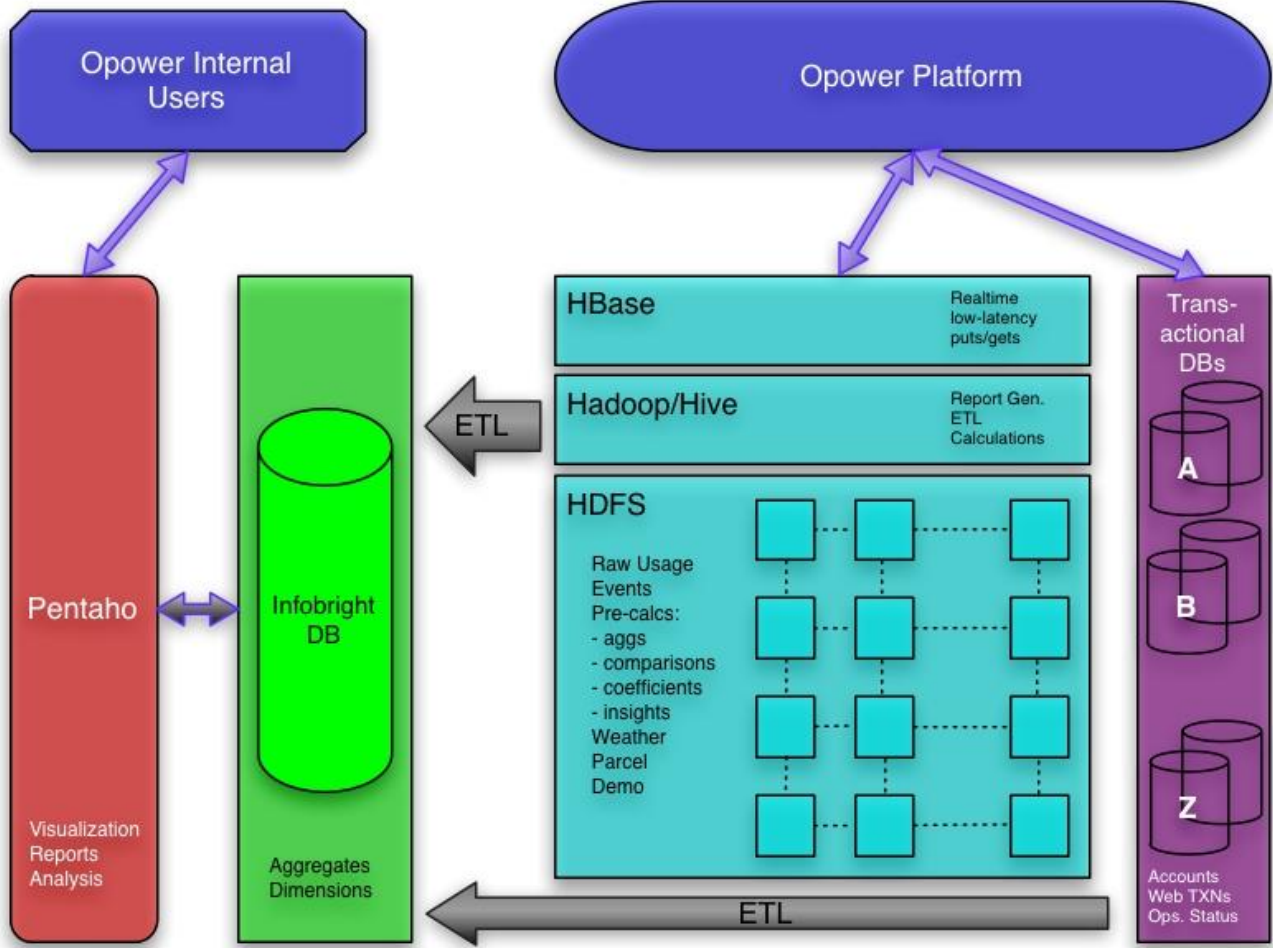


Patterns for Dataset Requirements

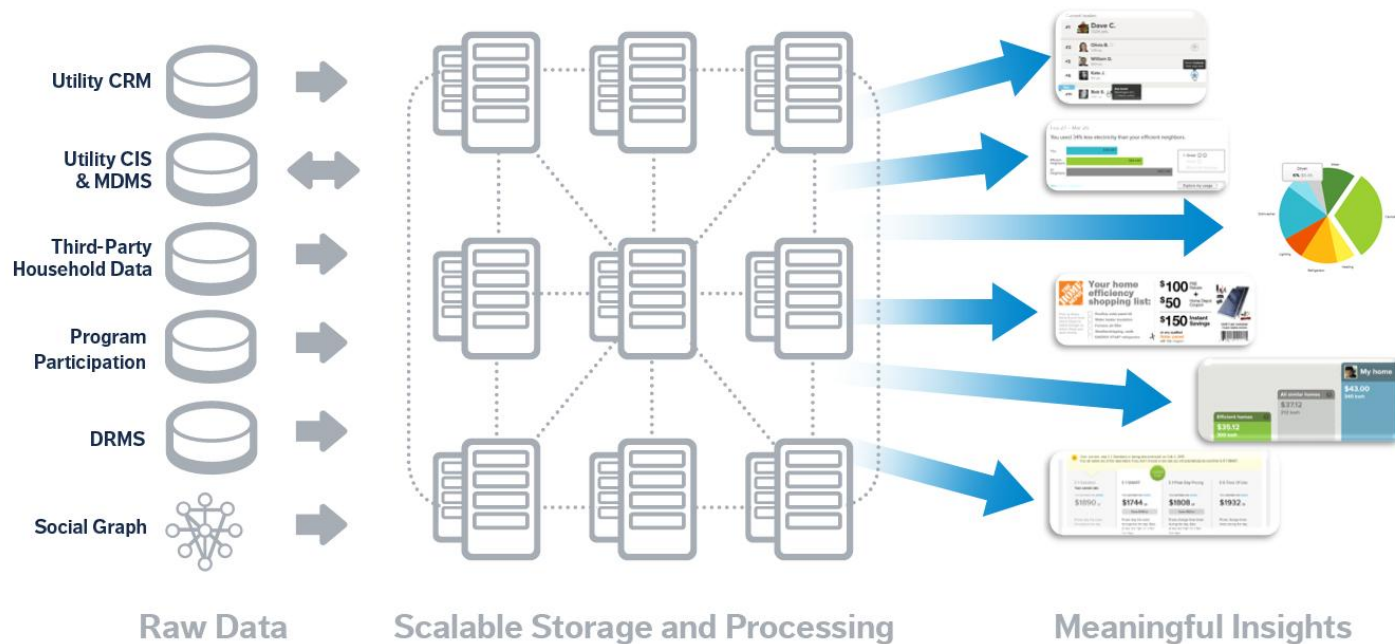
- » Access Patterns
- » Dataset Size
- » Atomicity
- » Resiliency
- » Budget

- » Opower Requirements
 - Transactional Dimension Datasets
 - Immutable Time Series Fact Datasets (Consumption)
 - Data Warehousing
 - Aggregates & Statistics

Opower Data Infrastructure



From Big Data to Big Insights



Our Scale:

- 50M Households, 15M with AMI
- 30TB of Usage Data
- 100k events per day per t-stat
- High Throughput Requirements
 - ~10M Bill Forecasts in 12 hours
- High Sequential IO Requirements
 - 1-3 years of data for each personalized comparison
 - Comparisons may require processing data for 100s of other consumers

HDFS, Hadoop, and HBase...

» The Apache Hadoop project provides a great technology set for processing, storing, and serving time series data.

» Opower has 5 Hadoop clusters

- 60 nodes
- 600TB of raw storage

» Benefits

- Optimized for sequential IO
- Locality: Blocks are processed where they are stored
- Linearly Scalable
 - Scale compute and storage simultaneously
- Open Source
- Cohesive Product Suite
- Commodity Hardware

Why Hadoop?

Me
WANTS
THE
DATA



Choose your own adventure...

Relational Databases

RDBMS = Relational Database Management System

Most common products: Oracle, MS SQLServer, MySQL, PostgreSQL

A.C.I.D.

- Atomicity – manipulation within a transaction is “all or nothing”
- Consistency – every transaction takes the DB to another valid state
- Isolation – no transaction can be effected by another
- Durability – transaction completion results in a persisted, recoverable DB state even in the event of power loss to the system or fatal error.

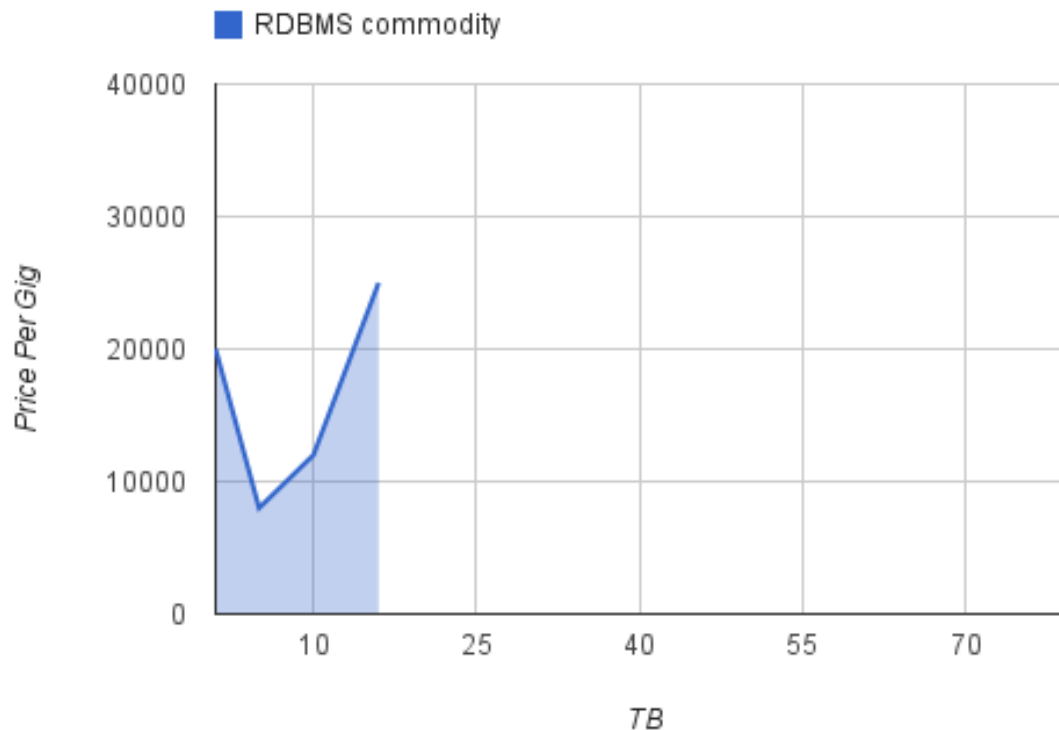
Optimized for transaction throughput

Common Installations

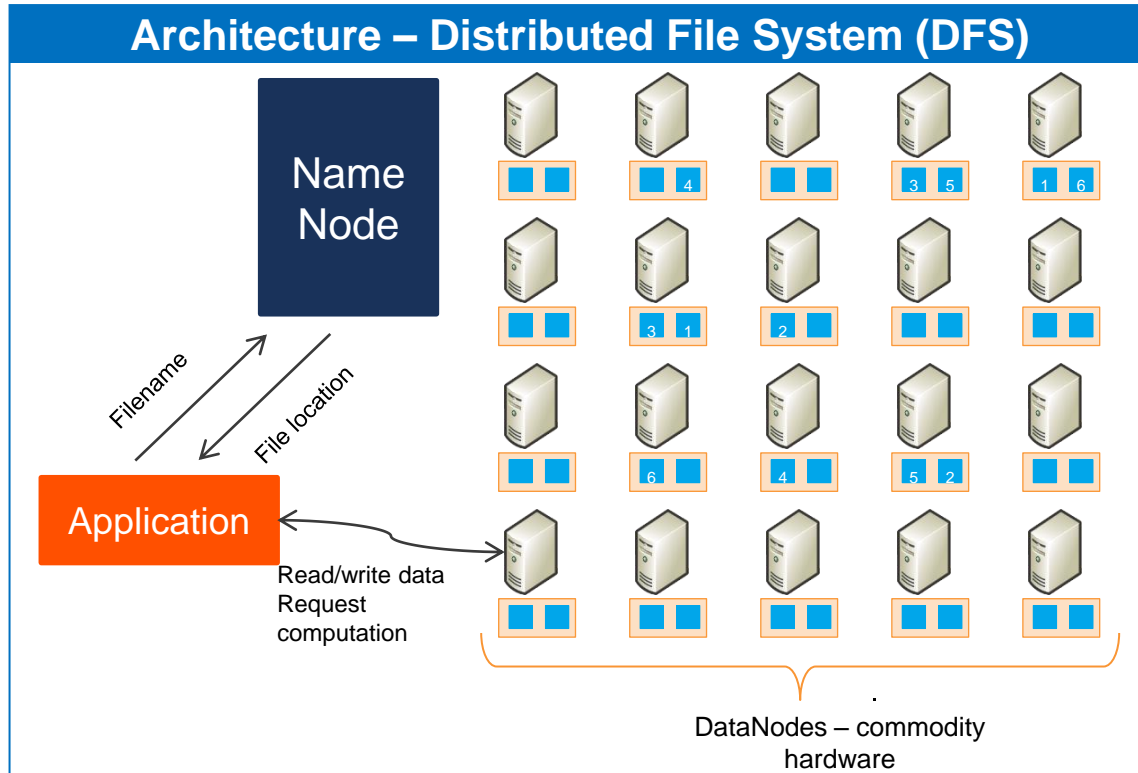
- Stand-alone commodity machine
 - Local Disk
 - Attached SAN
- Special Hardware – Sun/Oracle Rack

But traditional database technologies can only get you so far

- » Optimized for transactions and events aren't transactional
- » Handling large datasets is expensive
- » High Sequential IO is necessary and just not available

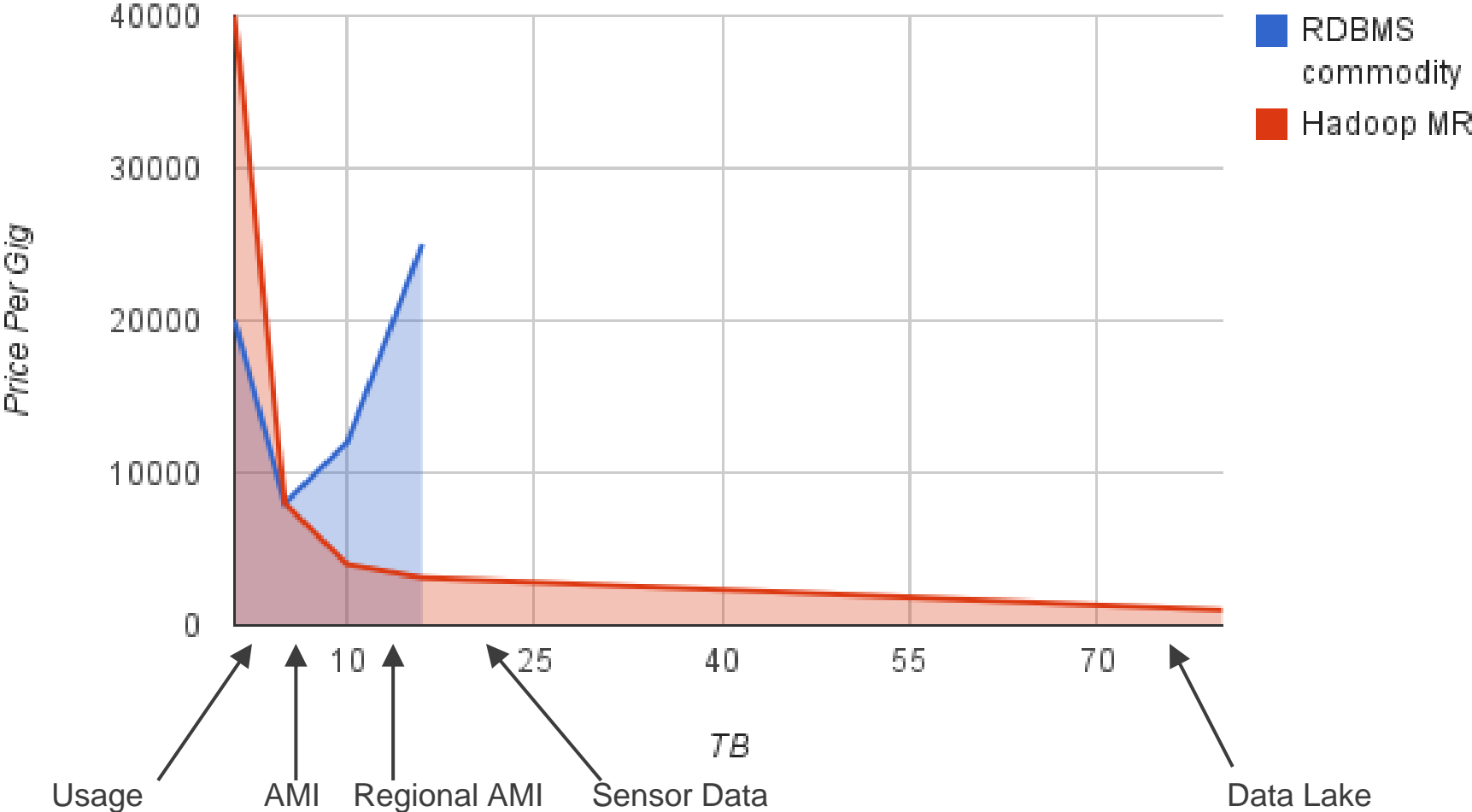


We use Hadoop and Map/Reduce



- ## Hadoop Properties
- **Open Source License:** Large user base ensures future technology innovation and leadership
 - **Scale:** Supports multiple PB of data by adding servers
 - **Low cost:** Runs on commodity hardware
 - **Fault tolerant:** Data replication
 - **Optimized for AMI data:** Write once, read many times
 - **Moves computation** to where data is located
 - **Portability** across hardware platforms: Java language

Efficiencies in performance and cost



Challenges in using Hadoop


- » Finding experienced Sysops teams
- » Dealing with Open Source tools
- » Delegating data to Hadoop vs RDMS
- » Managing security and access control
- » Fewer ETL and automation tools right now
- » **Data Quality...**

Walkthru: Unusual Usage Alerts



Unusual usage alerts

You're receiving this alert to help you keep your bills low. [Unsubscribe](#)



Acct # *****5678


Unusual electric usage




Your last 8 days
\$58
May 22 – 29
[See your use each day](#)

Your next bill could be
\$175*
Projected for May 22 – June 20

Your typical June bill 2009 – 2010: **\$106**

Based on your usage since May 22, you could be headed towards a bill that is **40% higher** than what you normally use this time of year.

 You still have time to minimize your next bill.

Steps to take	Impact
Turn off unused lights & devices	
Clean or replace air filters monthly	
Adjust your thermostat 3 – 5 °	

[See more ways to save](#)

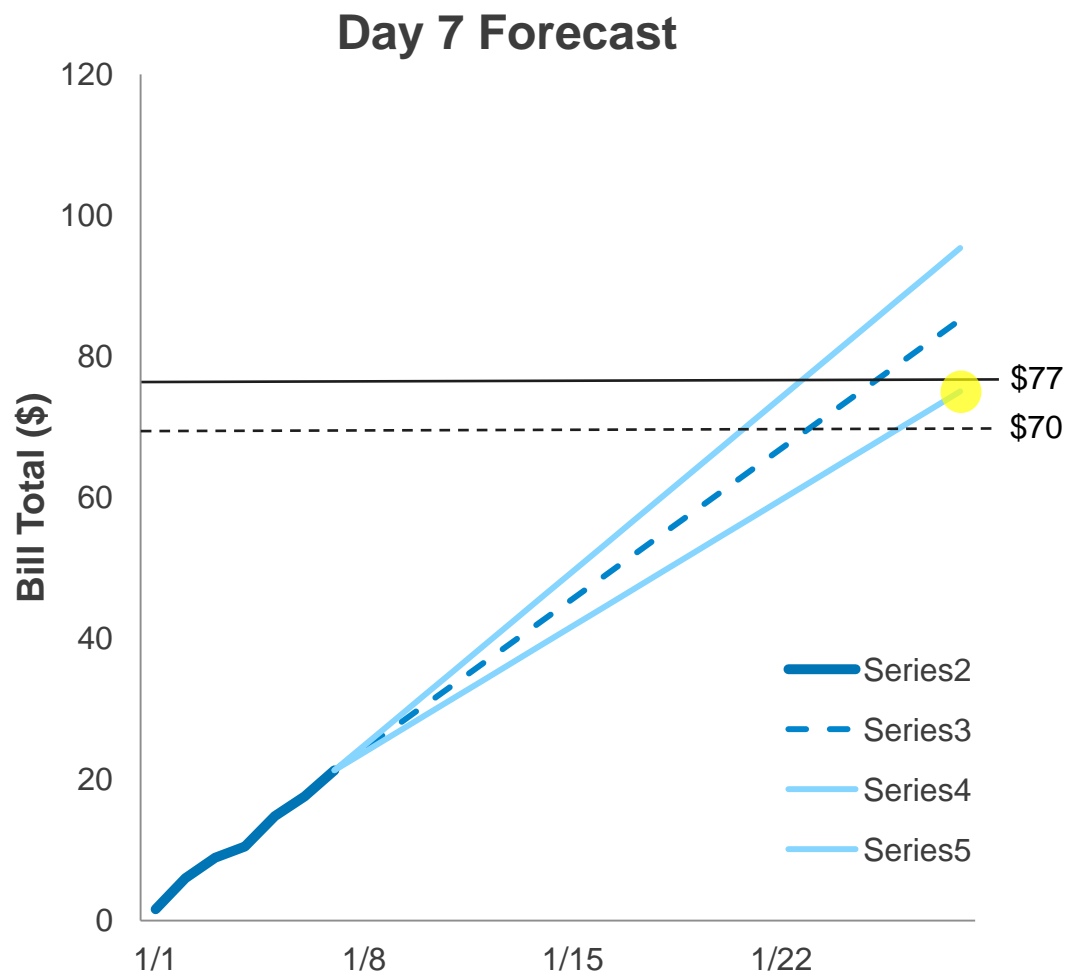
* Actual bill will vary based on usage, taxes, & fees

[Change your alert preferences](#)

Utility name, 1911 Fort Meyer Drive, Suite 702, Arlington, VA 22209 (555) 555-5555
© 2011 OPOWER. All rights reserved.

- » Empower customers and manage expectations with alerts based on energy use
- » Being leveraged for unusual usage (high bill) alerts in the US and UK

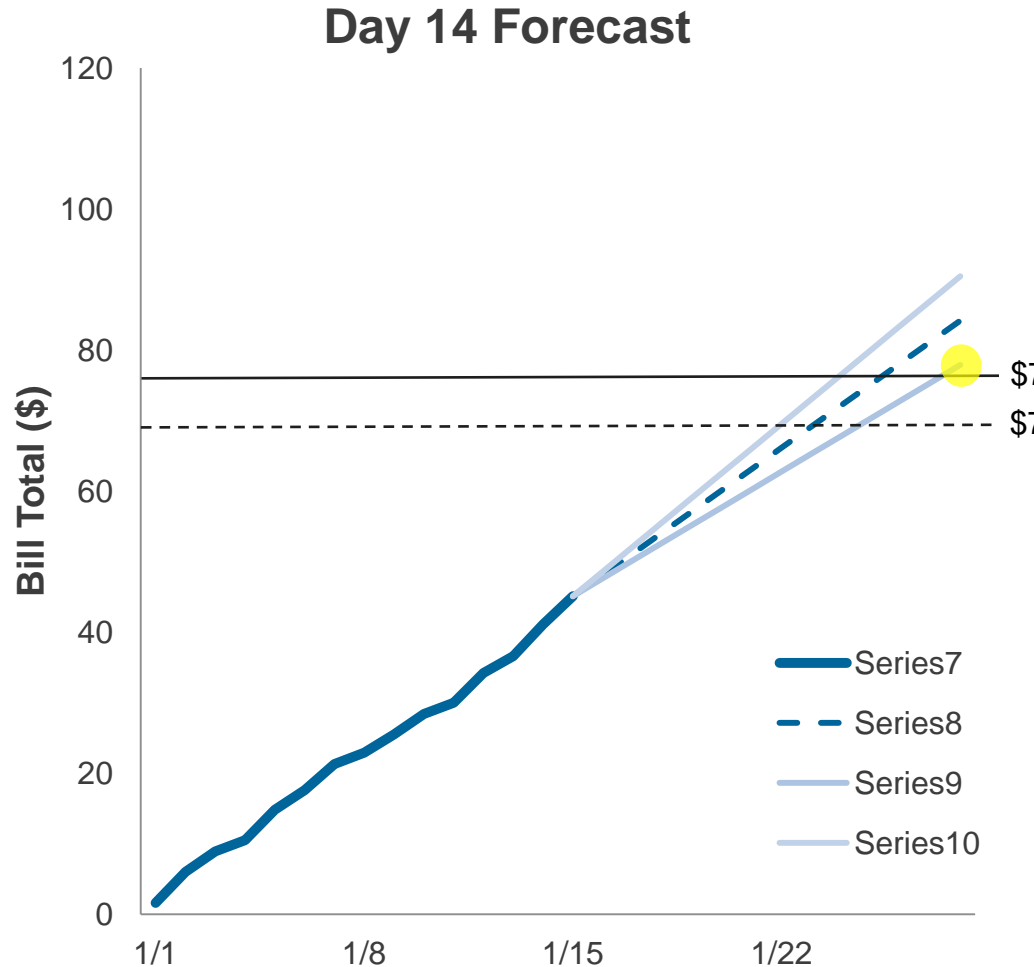
How we forecast your next bill



- Total usage-to-date
- Estimate end of bill cycle
- Project average value based on historical data
- Calculate variance (90% confidence distribution)
- Add buffer to expected bill
- Compare minimum forecast to threshold

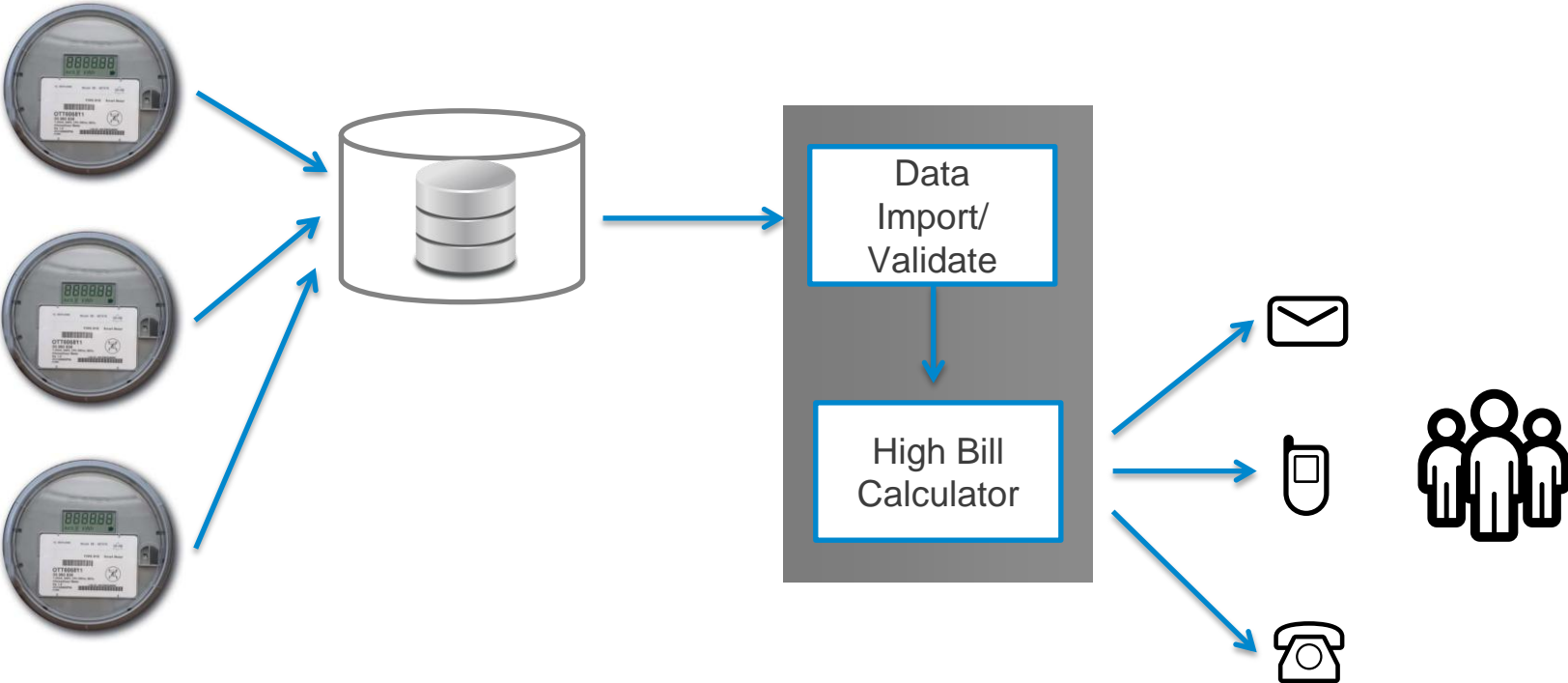
x No alert on Day 7

When we send high bill alerts



- Re-evaluate during bill period as new AMI data arrives
- ✓ Alert sent on Day 14
- No more alerts sent this bill period
- Avoid sending alerts near end of bill period

Information Flow



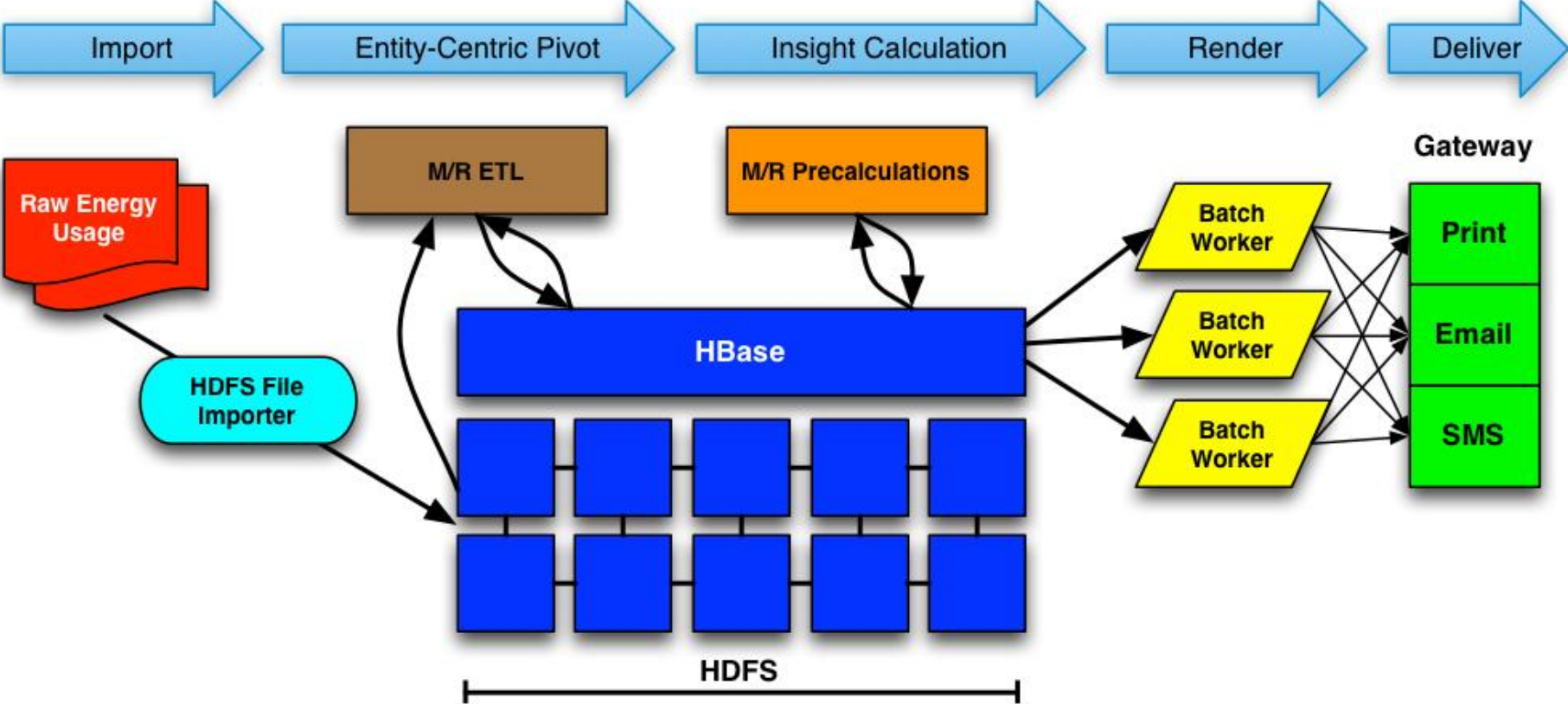
Collect usage data from customer meters

Transfer daily interval data to Opower

Opower processes latest data

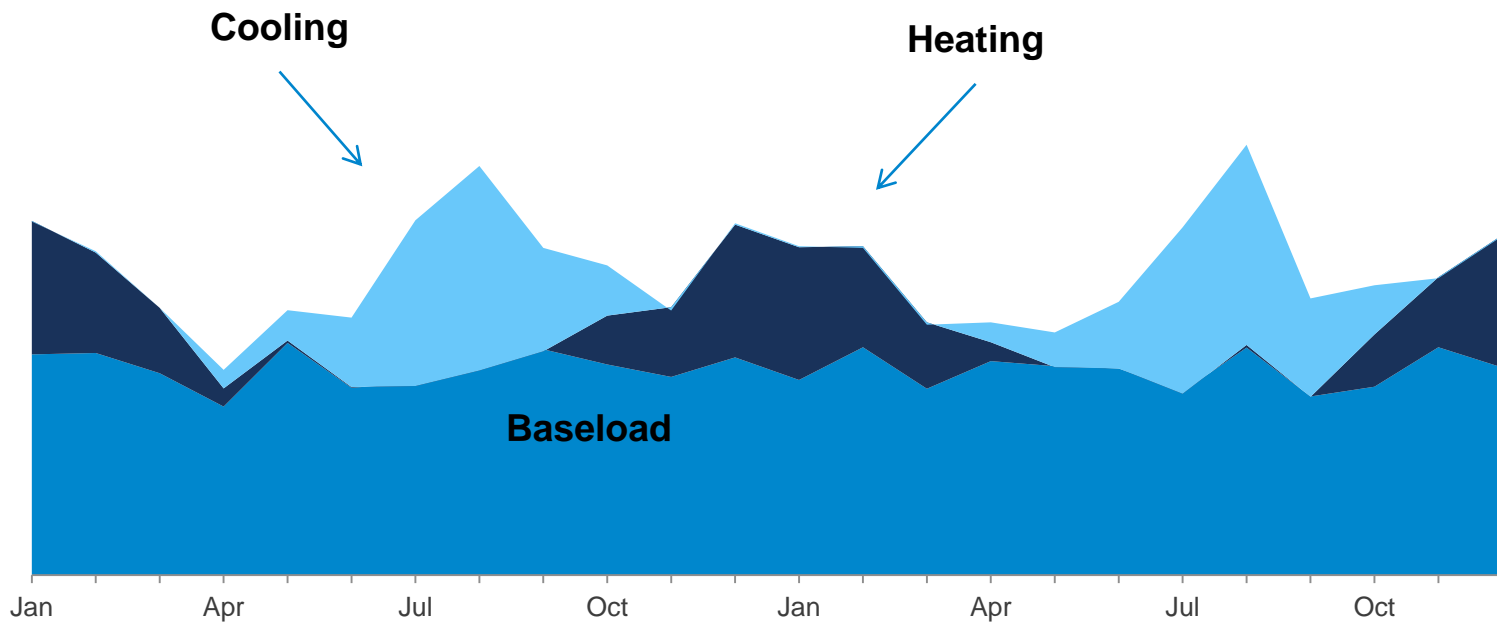
Generates and sends high bill alerts

Data Flow

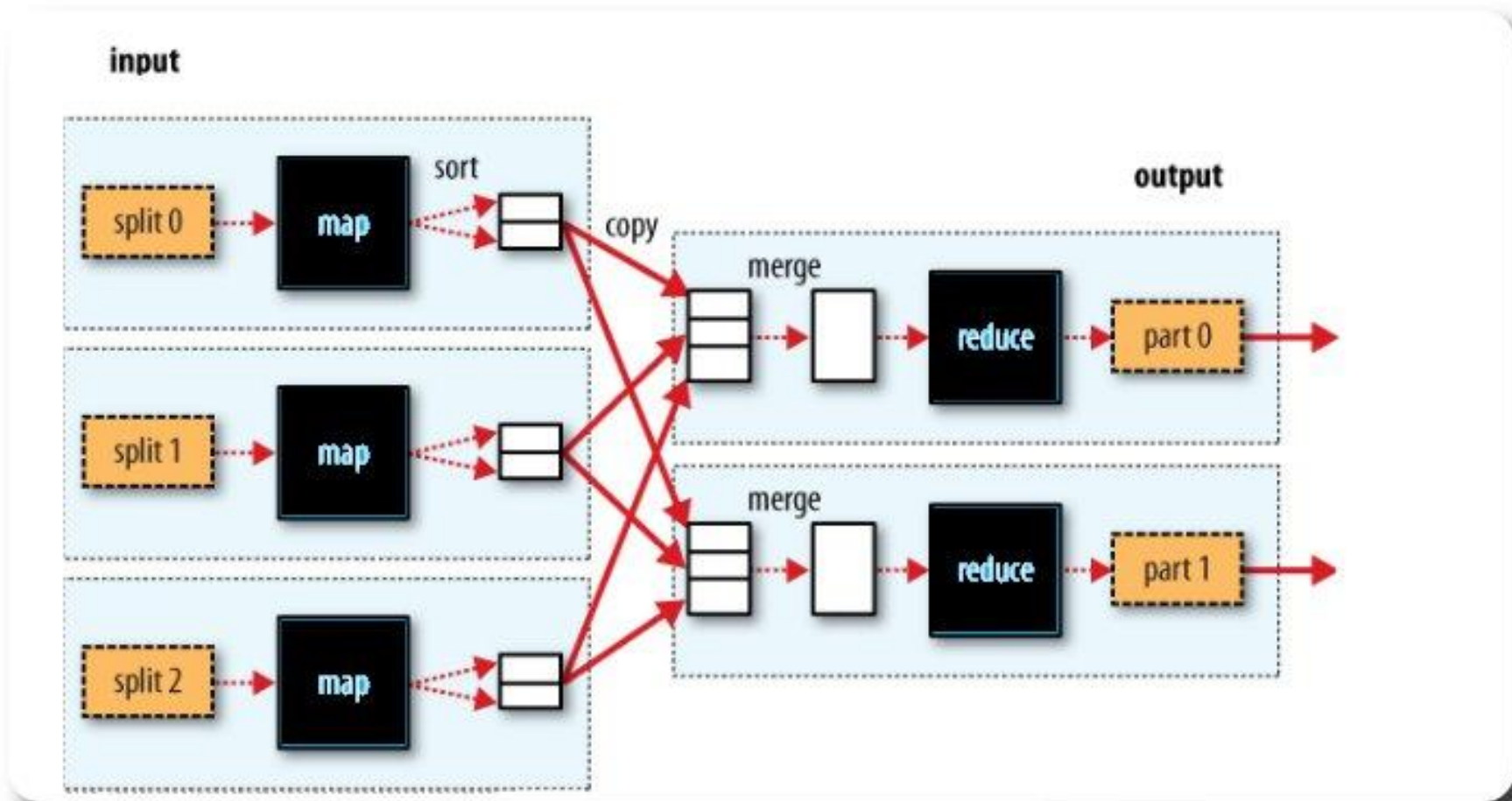


Analytical Output

- » Bill Forecasting & Unusual Usage Detection
- » Heating and Cooling Disaggregation
- » Baseload Disaggregation
- » Neighbor Comparisons and Rankings



MapReduce Data Flow



Borrowed from <http://xmlandmore.blogspot.com/2011/12/volume-rendering-using-mapreduce.html>

HBase Overview

It is a

- » Sparse
- » Distributed
- » Sorted
- » Key/value


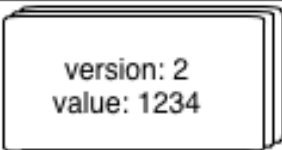
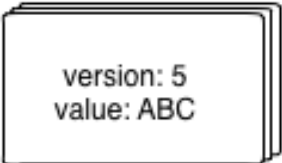

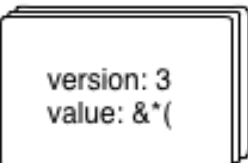


data store.

Modeled after Google's BigTable, which is a “sparse, distributed, persistent multi-dimensional sorted map.”

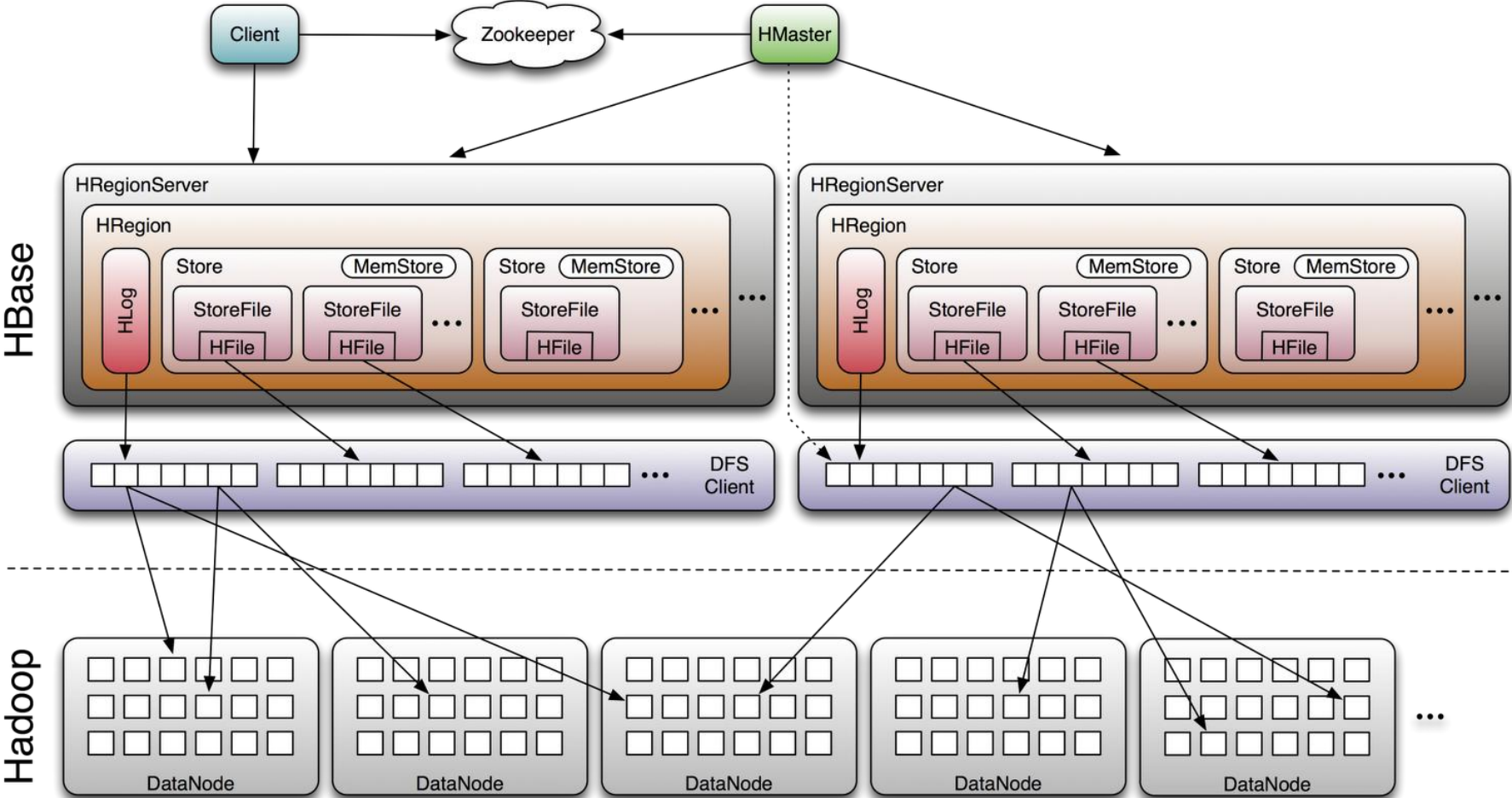
HBase Schema

Three-dimensional table.

- » Row
- » Column
- » Timestamp version

	column-1	column-2	column-3
row-00		 version: 2 value: 1234	version: 1 value: !@#
row-01	 version: 5 value: ABC	version: 1 value: 5678	version: 1 value: \$%^
row-10	version: 1 value: DEF		 version: 3 value: &*('
row-2	version: 1 value: GHI		

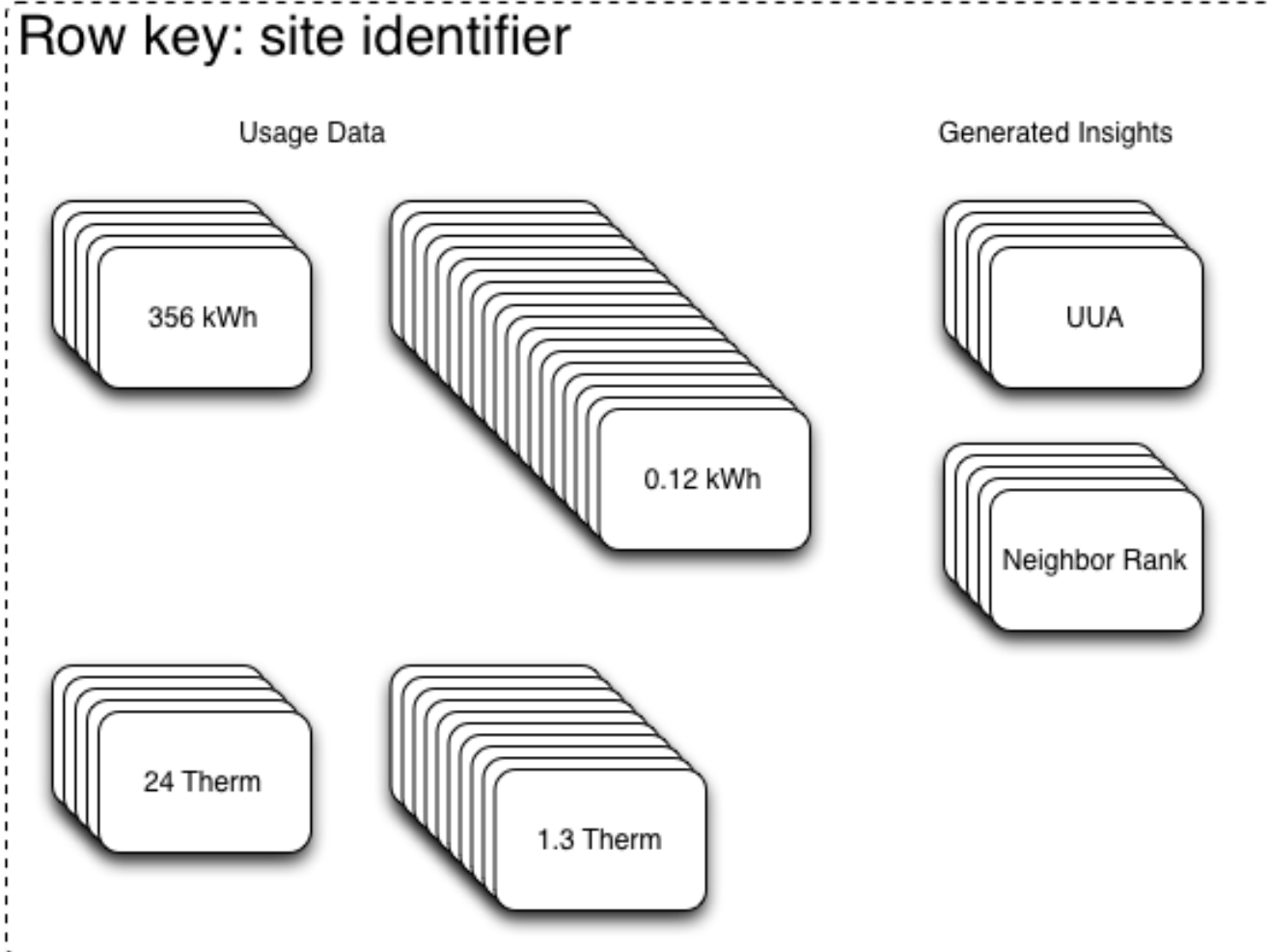
HBase Architecture Overview



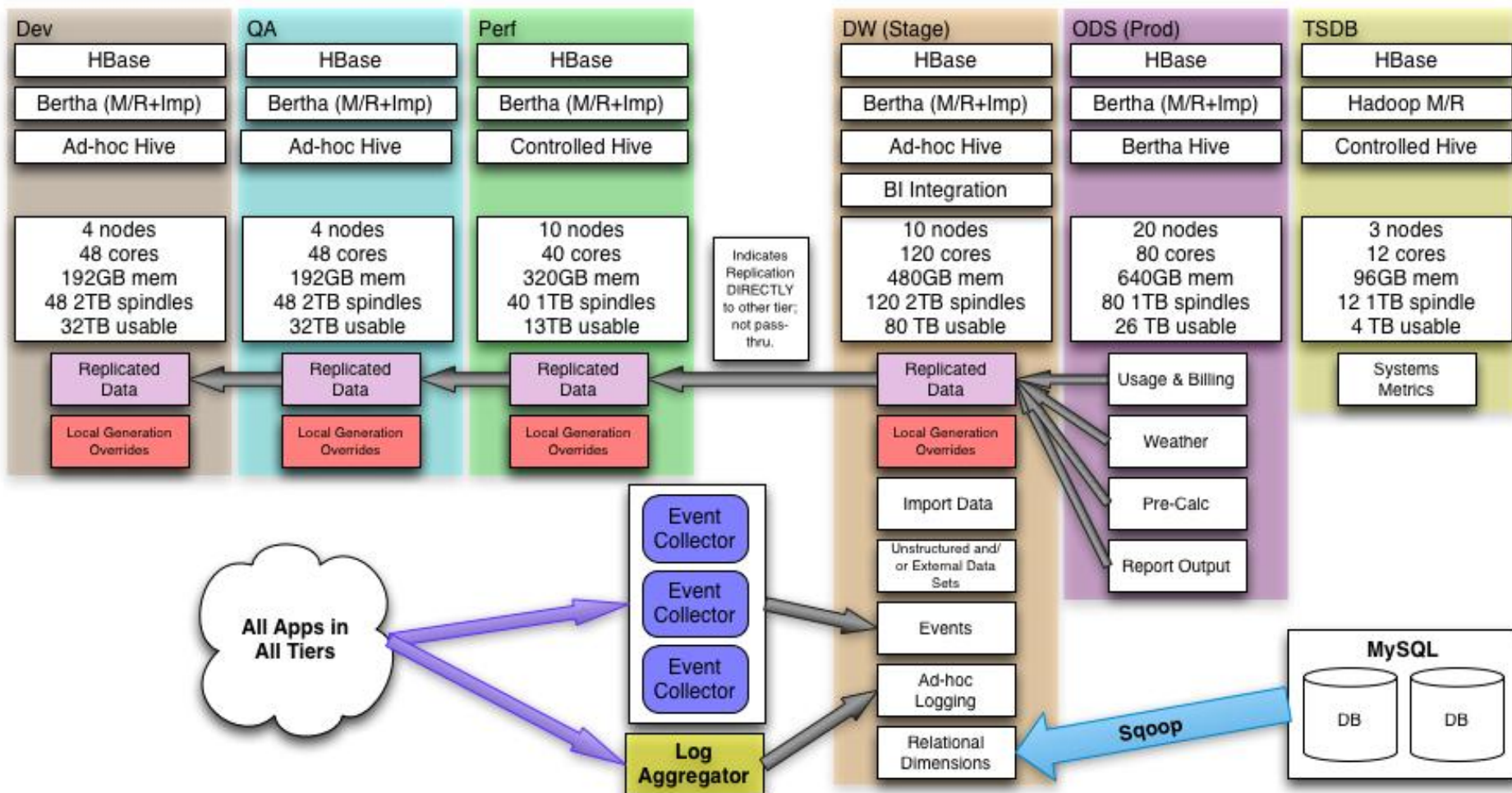
Borrowed from <http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html>

Our Data In HBase

- » Entity-centric
- » Timeseries
- » Raw & generated data



Opower Hadoop Infrastructure



Appendix



Maintaining Quality when using Hadoop





Yahoo! front page - Case Study

The screenshot shows the Yahoo! front page with several key areas annotated with blue ovals:

- Content Optimization:** Located over the top image of a woman's face.
- Search Index:** Located over the 'POPULAR SEARCHES' list.
- Machine Learned Spam filters:** Located over the Yahoo! Mail interface.
- Content Management:** Located over the news headlines.
- Ads Optimization:** Located over the Progressive Direct advertisement.

POPULAR SEARCHES

1. Paula Abdul
2. Angelina Jolie
3. David Beckham
4. Moon Landing
5. Pearl Jam

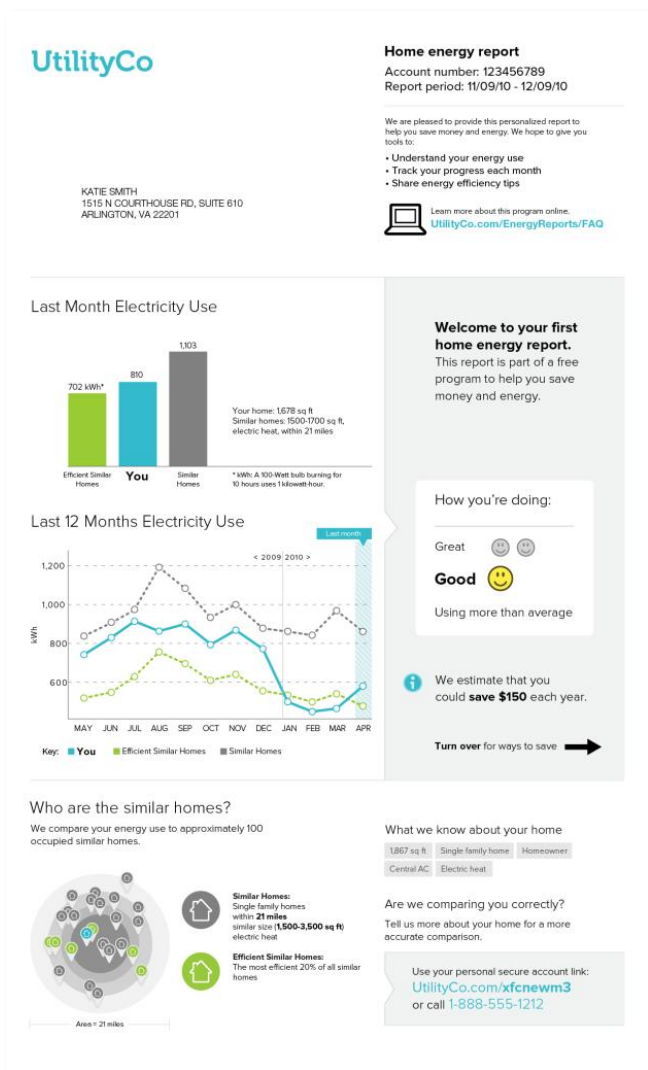
News Headlines:

- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 hijackers are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
- Company designs... mobile video editing
- Sony bids \$50...
- Defense Secretary Gates...
- Frank McCourt...
- Annual all-star football...

Markets: Dow: 8,821.70 **0.80%** Nasdaq: 1,908.95 **0.73%**

Deal Of The Day: GEICO Car Insurance. You could save over \$500 on car insurance. Get a free quote today.

Opower M/R Use Case has key differences



Low tolerance for quality issues because:

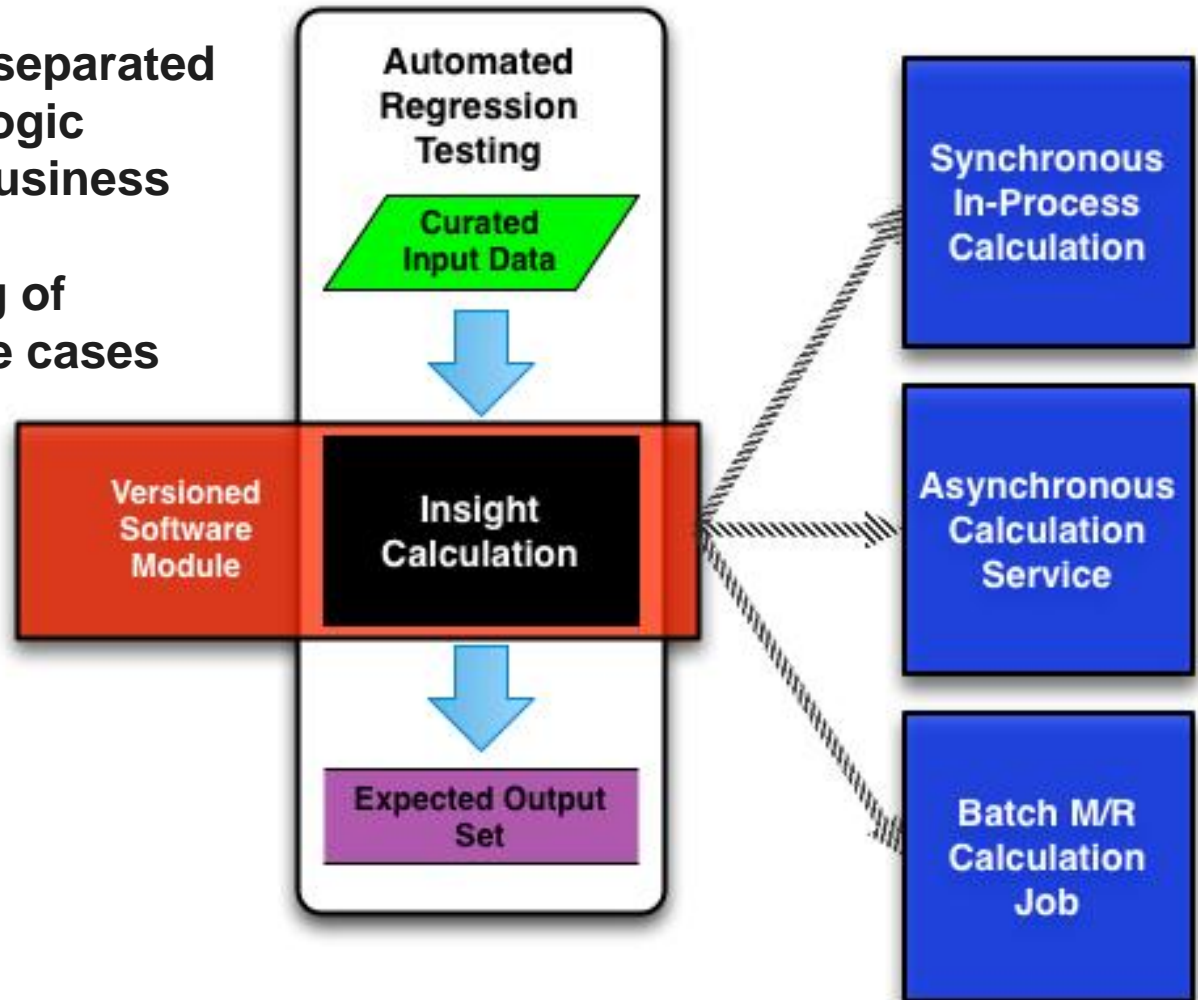
- Limited engagement opportunities; sometimes just 4 times a year
- Most insights go on to paper, which lasts indefinitely
- Must engage all users in a target sample
- Results of EE program depend greatly on the actual values produced

Maintaining Quality when using Hadoop

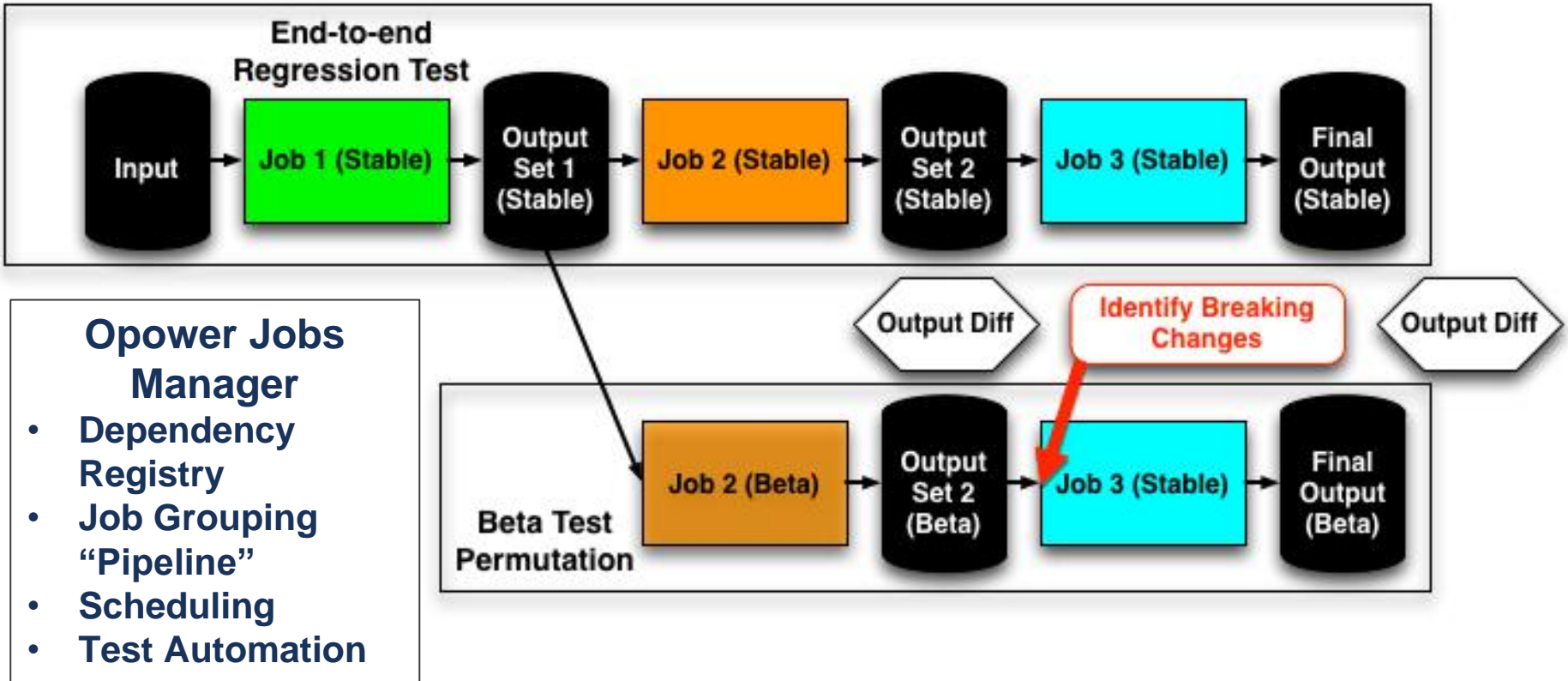
- **Business Logic Abstraction**
- **Data Pipeline Testing**
- **Multi-Cluster Strategy**

Business Logic Abstraction

- » Business logic is separated from processing logic
- » Direct testing of business logic
- » End-to-end testing of business logic use cases



Data Pipeline Testing

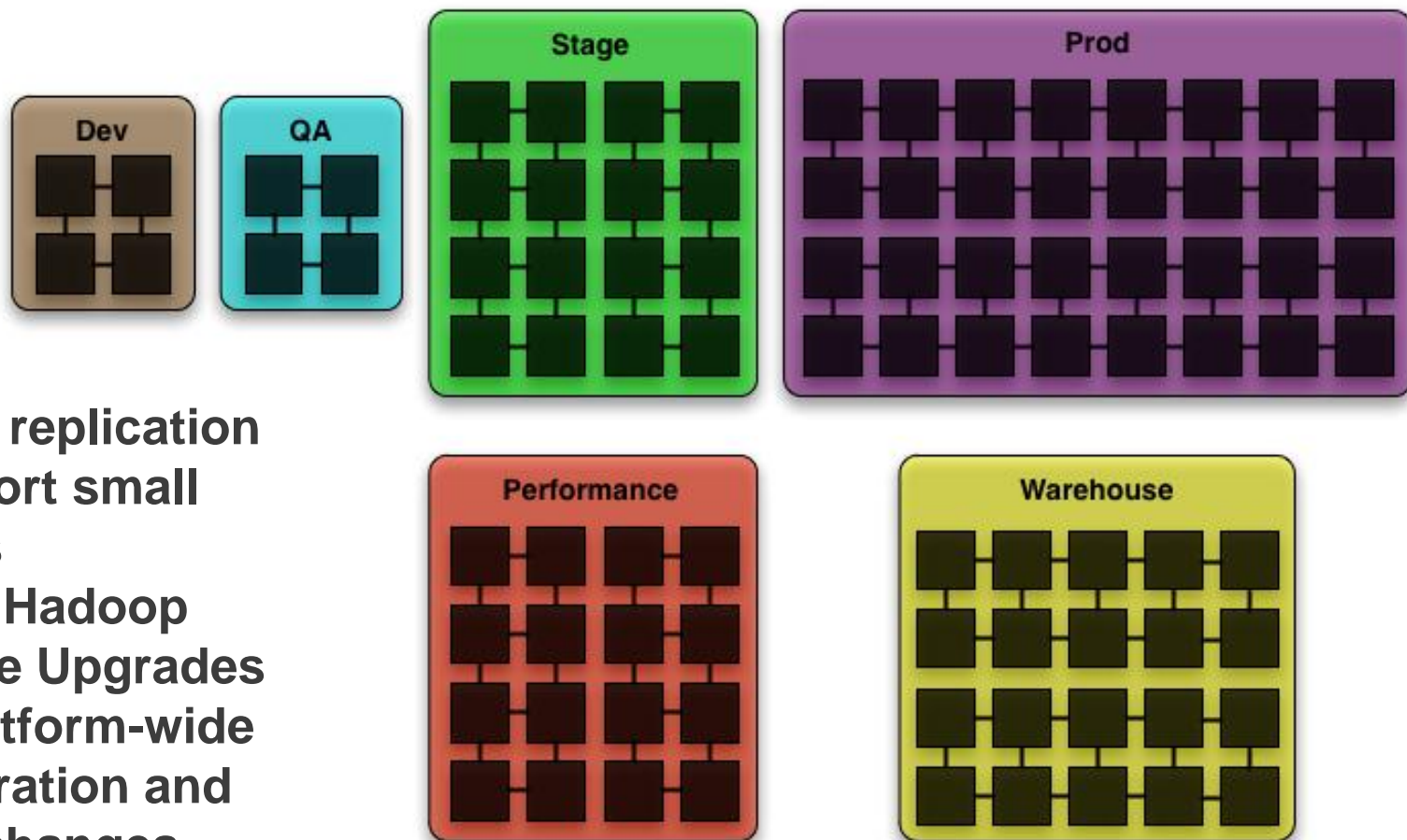


Framework Features:

- Maintain separate data access methods for verification of data on disk


- Pipeline breakage alerts
- Smart dataset pointers
- Dataset promotion

Multi-Cluster Strategy



- Change replication to support small clusters
- Burn-in Hadoop Software Upgrades
- Test platform-wide configuration and library changes
- Performance Testing

Ensuring Success with Hadoop

- » Focus on data quality
- » Hire great developers
- » Train systems teams properly
- » Get help (we use )