

Navigating the 'Big Data' Movement

New Value For Hard Problems

Blair Deaver – Software Product Manager
July 16, 2015



GeoEngineers

Earth Science and Technology Experts



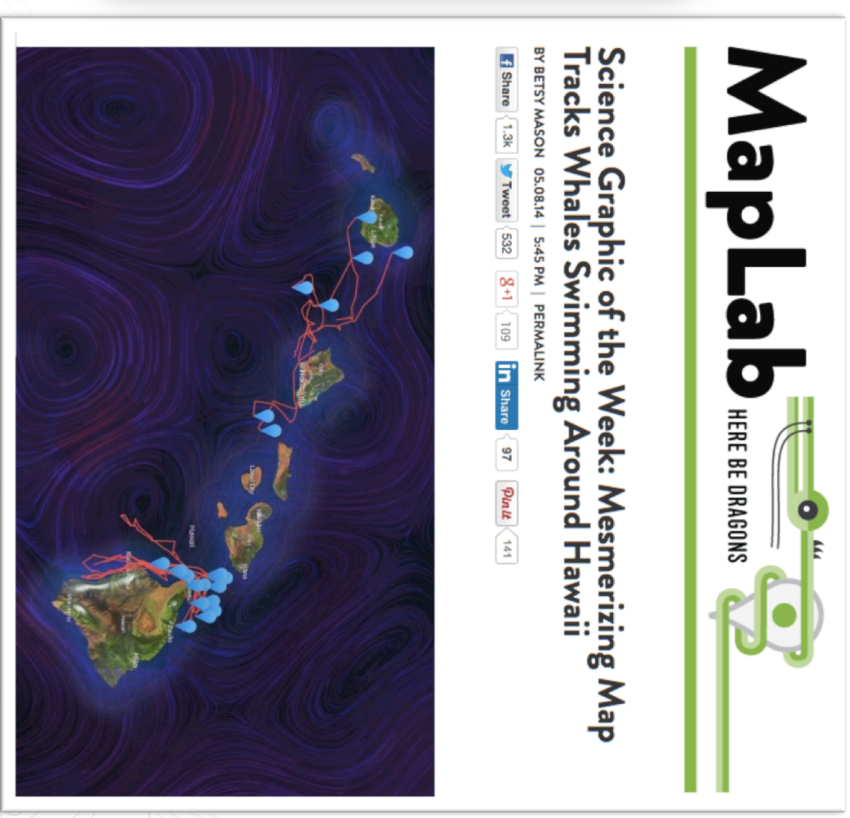
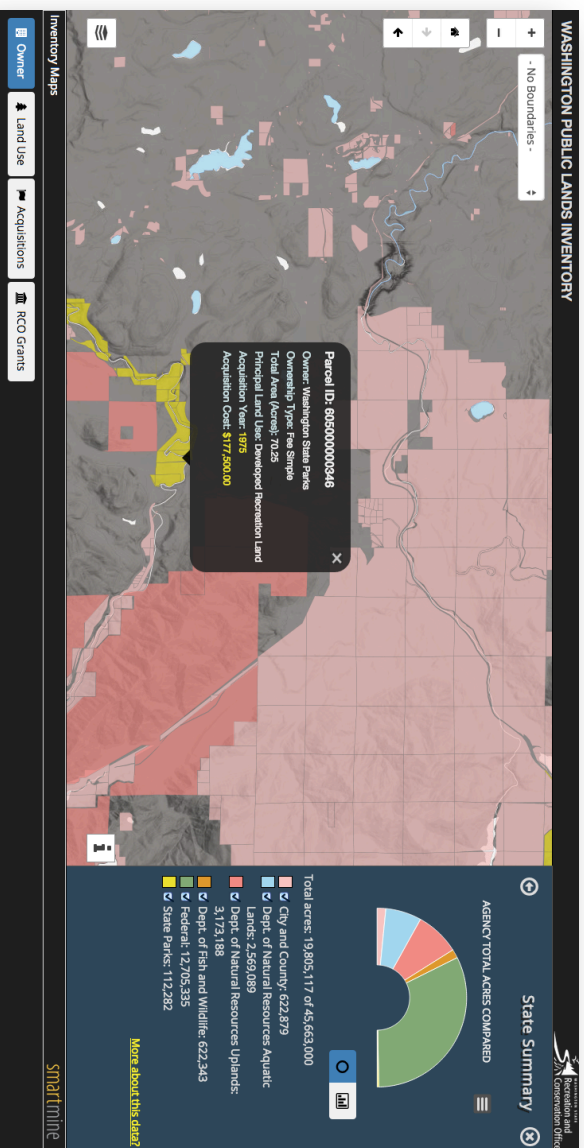
*Trusted advisors to clients who manage natural resources
and the built environment.*



smartmine

The SmartMine team at GeoEngineers works closely with our earth science experts to develop software products that help customers manage and visualize environmental data.

Our Work



Our Platform: Earth Analytics

GeoEngineers' fusion of consulting expertise...



...with industry-leading data-acquisition and visualization services

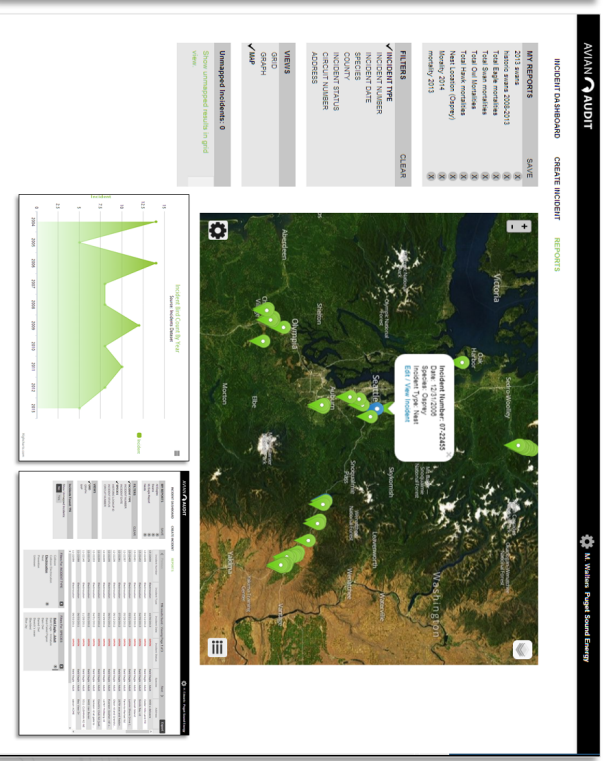


EARTH ANALYTICS

Example Solutions



Animated Temporal Modeling & Comparison Maps

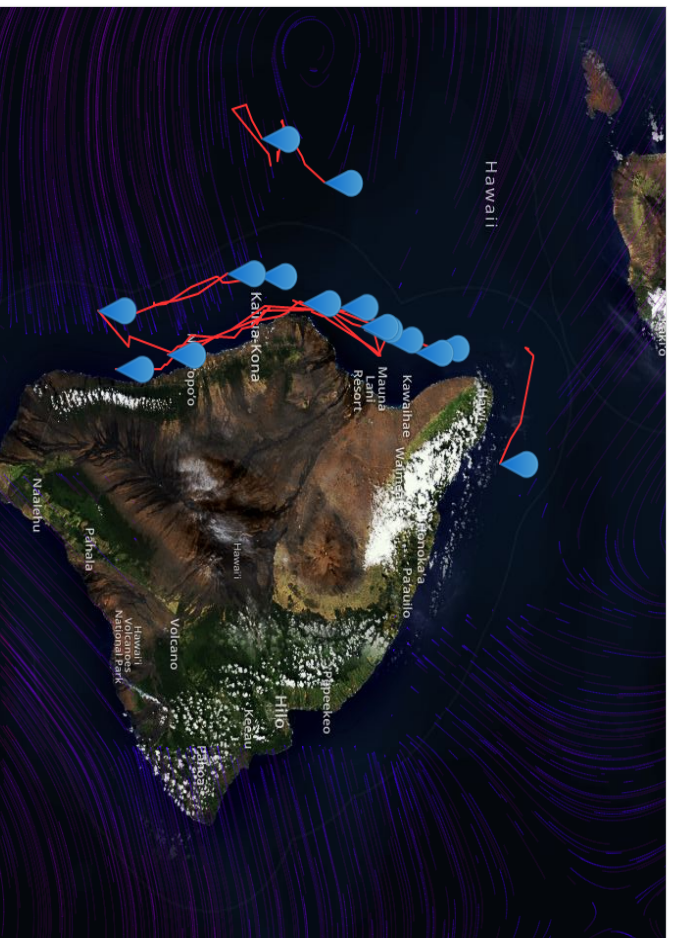


Incident Management Dashboard

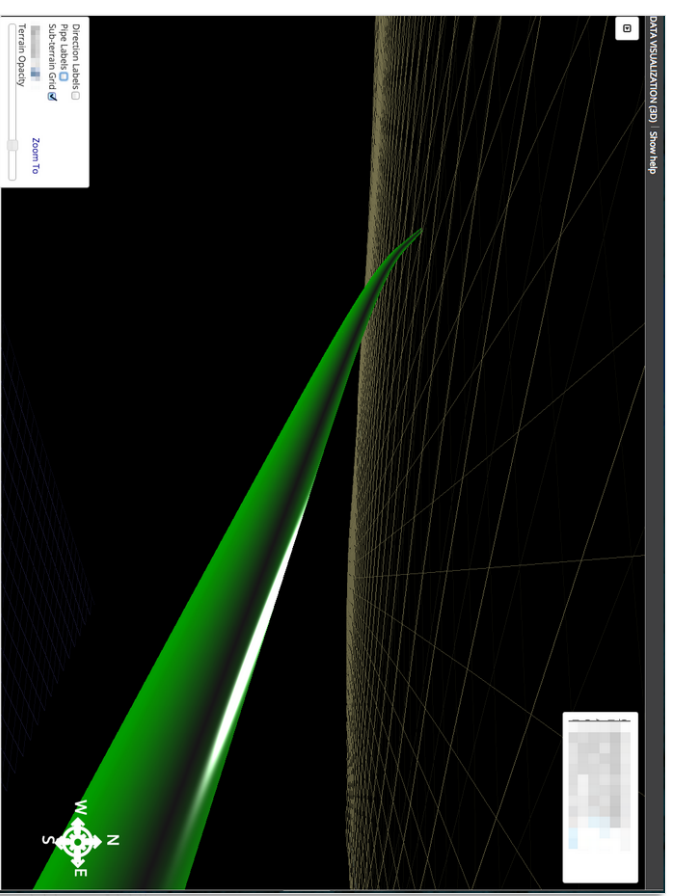


EARTH ANALYTICS

Example Solutions



GPS Tracking & Monitoring



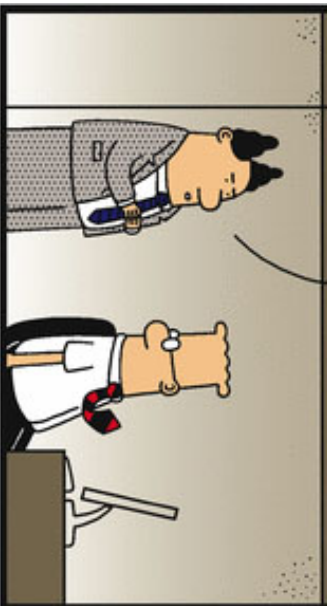
3D Visualization & Monitoring



Discussion Topics

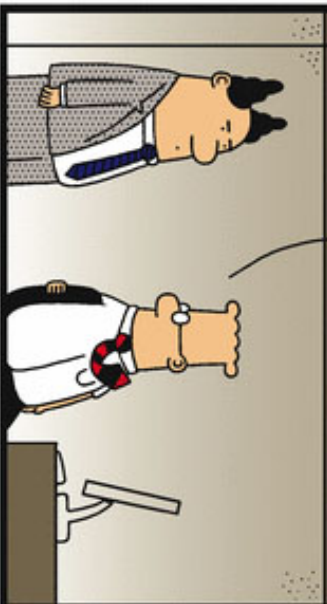
- What is Big Data
- What is a Data Scientist
- Big Data Tools
- Why Should I Care?
- Where to Start
- Questions/Answers

DO WE HAVE ANY
ACTIONABLE ANALYTICS
FROM OUR BIG DATA
IN THE CLOUD?



Dilbert.com DilbertCartoonist@gmail.com

YES, THE DATA SHOWS
THAT MY PRODUCTIVITY
PLUNGES WHENEVER YOU
LEARN NEW JARGON.



1-9-13 ©2013 Scott Adams, Inc., /Dist. by Universal Uclick

MAYBE IN-MEMORY
COMPUTING WILL
ACCELERATE YOUR
APPLICATIONS.



<http://dilbert.com/strips/2013-01-09>

[New Posts](#) ⁴⁶
[Most Popular](#)
[Lists](#)
[Video](#)
[10 Stocks to Buy Now](#)

[P](#)
[Privacy](#)
[Terms](#)
[Adchoices](#)
[Help](#)

12 Big Data Definitions: What's Yours?

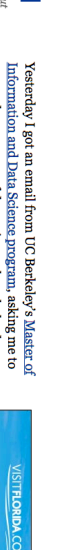
Yesterday, I got an email from UC Berkeley's *Master of Information and Data Science* program, asking me to

GI Press Contributor + Follow Comments

FOLLOW

entrepreneurs and technology.

I wrote about "What is big data?" last week. I was especially delighted to be regarded as a "thought leader" by Berkeley's School of Information, whose previous dean, Hal Varian (now chief economist at Google), answered my challenge fourteen years ago and produced the first study to estimate the amount of new information created in the world annually; a study I



visitFLORIDA.COM

WANT TO ENJOY THE BEACHES?

So many beaches.
So little time.

HOTELS TO EXPAND

2 **COMMENTS**
2 FOLLOW COMMENTS

The Berkeley researchers estimated that the world had produced about 1.5 billion gigabytes of information in 1999 and in a 2002 *publication of the study* found out that amount to have doubled in 3 years. Data was already getting bigger and bigger and around that time, in 2001, industry analyst Doug Laney described the “3Vs”—volume, variety, and velocity—as the key “data management challenges” for enterprises, the same “3Vs” that have been used in the last four years by just about anyone attempting to define or describe big data.



Scientist

on

n statistics

es and write

alysis

Data Scientist

together we are building the future

Walmart.com

Home > eCommerce Jobs > Buy Area eCommerce Jobs

Search Jobs > Career Areas > About Us > Our People

Minimum Qualifications

- Preferably PhD in computer science or similar field or MS with atleast 2-5 years of experience in machine learning, information retrieval, data mining, etc.
- Deep knowledge of machine learning, information retrieval, data mining, etc.
- Good functional coding skills in C++ or Java(Java is highly preferred) – the production code is either C++/Java/JavaScript/Python
- Expert level knowledge of one of the scripting languages such as Python
- Superior ability to analyze and interpret the results of product experimentation

Position Summary

Driving the largest product searches in international commerce, Walmart.com has 100million global user base. We are a group comprised of the brightest technical talent with the opportunity to work on the most challenging opportunities that this next generation of Commerce will bring to billions of users better.




Search processes billions of queries for millions of products on Walmart.com. We are looking for Data Scientists to work on the most challenging product categories on the web site, phone or the iPad, our service goes to social web, transactions, query logs, etc. at an unprecedented scale. We've got machine learning, and ranking to deliver a high-availability, low-latency search experience.

Joining the search relevance team shall ensure your contributions help build multi-million revenue impact!!!

Position Description

Help @Walmart Labs invent the next generation of e-commerce, integrated with the glue. Apply your strong expertise in Java, machine-learning, data generation of @Walmart Labs semantic-based engines and services

Be challenged to create applications that will impact over 1.5 billion customers. We are looking for Data Scientists to work on the most challenging language interfaces that apply semantic technology to match user intent at position! The role requires a passion for consumer experience, a willingness to take on challenges, and a desire to work on the most challenging challenges with your superb coding skills. Be excited about making an impact on the world!

- # Scientist
- ## on
- ### n statistics
- ### es and write
- ### alysis
- ## Data Scientist
- ### together we are building the future
- ### Walmart.com
- Home > eCommerce Jobs > Buy Area eCommerce Jobs
- Walmart   Careers
- Search Jobs ▶ Career Areas ▶ About Us ▶ Our People
- 
- Minimum Qualifications
- Preferably PhD in computer science or similar field or MS with atleast 2-5 years of experience in machine learning, information retrieval, data mining, etc.
 - Good functional coding skills in C++ or Java(Java is highly preferred) – the production code in either C++/Java/JavaScript/Python
 - Expert level knowledge of one of the scripting languages such as Python
 - Superior ability to analyze and interpret the results of product experimentation
- Position Summary**
- Driving the largest product searches in international commerce, Walmart.com has a 100million global user base. We are a group comprised of the brightest technical talent with the opportunity to work on the most challenging opportunities that this next generation of Commerce will bring to billions of users better.
- Search processes billions of queries for millions of products on Walmart.com. We are looking for people who can help us improve our search results for product categories on the web site, phone or the iPad, our service goes to social web, transactions, query logs, etc. at an unprecedented scale. We've got machine learning, machine learning, and ranking to deliver a high-availability, low-latency search experience.
- Joining the search relevance team shall ensure your contributions help build a multi-million revenue impact!!!
- Position Description**
- Help @Walmart Labs invent the next generation of e-commerce, integrated with the glue. Apply your strong expertise in Java, machine-learning, data generation of @Walmart Labs semantic-based engines and services
- Be challenged to create applications that will impact over 1.5 billion customers. We are looking for people who can help us improve our search results for product categories on the web site, phone or the iPad, our service goes to social web, transactions, query logs, etc. at an unprecedented scale. We've got machine learning, machine learning, and ranking to deliver a high-availability, low-latency search experience.
- The role requires a passion for consumer experience, a willingness to take on challenges with your superb coding skills. Be excited about making an impact on the world.

Scientist

on

n statistics

es and write

alysis

Data Scientist

together we are building the future of retail

Walmart.com

Home > eCommerce Jobs > Buy Area eCommerce Jobs

Search Jobs > Career Areas > About Us > Our People

Position Summary

Driving the largest product searches in international commerce, Walmart.com has a 100million global user base. We are a group comprised of the brightest technical talent with the opportunity to work on the most interesting and challenging opportunities that this next generation of Commerce will bring to billions of users better.

Search processes billions of queries for millions of products on Walmart.com. We are looking for Data Scientists to build and optimize our search algorithms for product categories on the web site, phone or the iPad, our service goes to social web, transactions, query logs, etc. at an unprecedented scale. We've got machine learning, machine learning, and ranking to deliver a high-availability, low-latency search experience.

Joining the search relevance team shall ensure your contributions help build a multi-million revenue impact!!!

Position Description

Help @Walmart Labs invent the next generation of e-commerce, integrated with the glue. Apply your strong expertise in Java, machine-learning, data generation of @Walmart Labs semantic-based engines and services

Be challenged to create applications that will impact over 1.5 billion customers. You will be working with a team of experts in machine learning, natural language interfaces that apply semantic technology to match user intent at position! The role requires a passion for consumer experience, a willingness to take on challenges, and a desire to work in a fast-paced environment. You will be working with your superb coding skills. Be excited about making an impact!

Minimum Qualifications

- Preferably PhD in computer science or similar field or MS with atleast 2-5 years of experience in machine learning, information retrieval, data mining, etc.
- Good functional coding skills in C++ or Java(Java is highly preferred) - the production code is either C++/Java/JavaScript/Python
- Expert level knowledge of one of the scripting languages such as Python
- Superior ability to analyze and interpret the results of product experimentation

Data Analyst vs. Data Scientist

Data Science

Perspective	Looking Forward
Data	Distributed, Near Real-time
Question	What will happen?
Tools	R, Hadoop, Python, JavaScript, D3.js, NoSQL, AWS, and more

A collection of various tools, including wrenches, sockets, and grinding wheels, arranged on a perforated metal surface. The text "New Tools" is overlaid on the left side.



Cassandra



amazon
DynamoDB



mongoDB



hadoop



Lightning-fast cluster computing



GEOENGINEERS



What makes these tools different?

- Engineered to handle larger datasets with less overhead
{designed for throughput}
- Scale horizontally
- Built for the cloud
- Distributed
{clusters}

Case Study: NoSQL

- New database designed to support scale (horizontal scaling)
- Built for the cloud
- Supports more unstructured data
- Works well for near real-time data and analytics

Database

Samples

TABLE

```
{
  Id = 101
  SampleNumber= "MW00010"
  SampleType=
  "Pentachlorophenol"
  SampleValue= 0.52
  SampleDate=
  "1392129408000"
  SampleComments= ""
  SampleLocation= "-123.3, 42.3"
}
```

Item 101

```
{
  Id = 102
  SampleNumber= "MW00011"
  SampleType= "Benzo[a]pyrene"
  SampleStartDepth = 0
  SampleStartDepthValue = 0.67
  SampleEndDepth = 4
  SampleEndDepthValue = 0.71
  SampleDate = "1392125408000"
  SampleComments= ""
  SampleLocation= "-122.3, 41.1"
}
```

Item 102

What is the big deal?

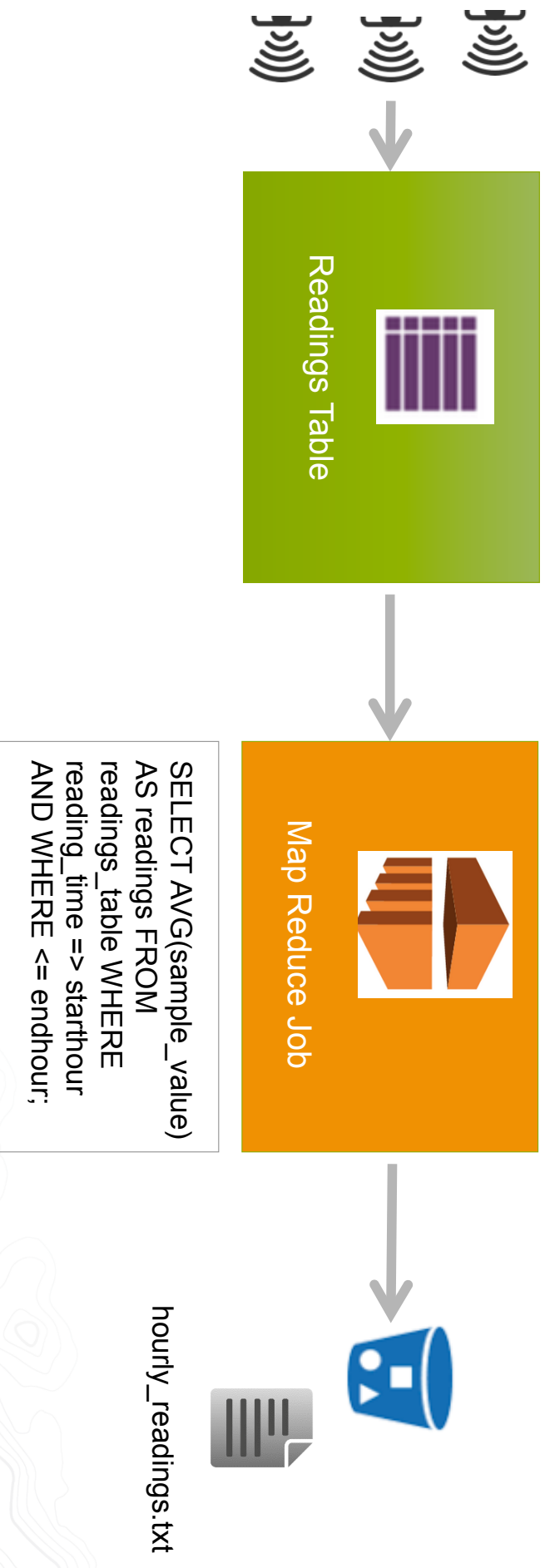
- No schema
- Built for the cloud
- Supports unstructured data
- Works well for near real-time data and analytics

Use Case: Reading Sensor Data {IoT}



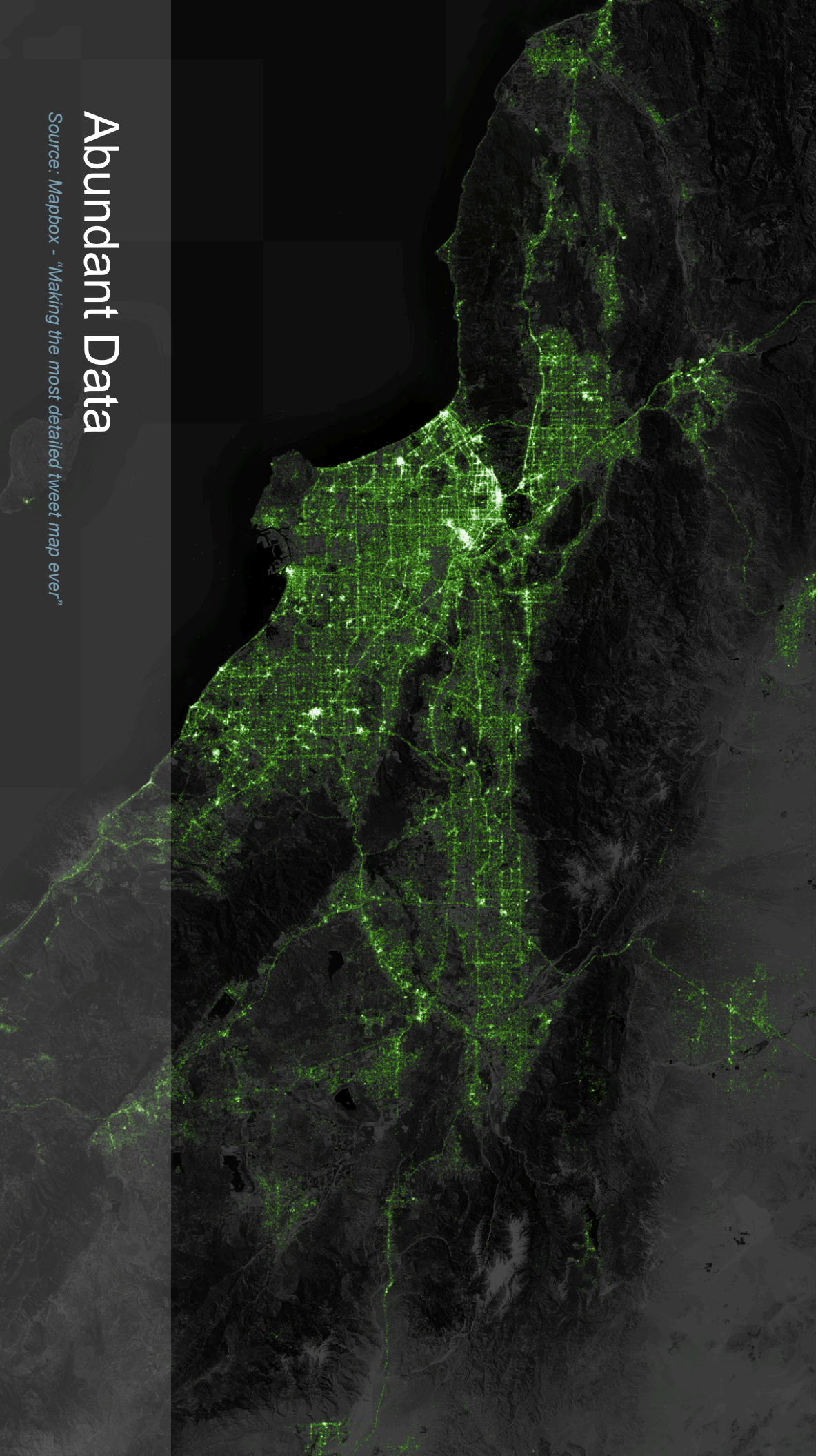
- New wireless sensor platform
- Variable support for a number of different data readings
- Customer wants readings every minute.
- Aggregate data hourly for reporting.

Using Map Reduce To Analyze Data Fire hose





Why Change




Abundant Data

Source: Mapbox - "Making the most detailed tweet map ever"

Congratulations! We are data rich





[AWS re/invent](#)
[Products](#)
[Solutions](#)
[Pricing](#)
[More](#)
[English](#)
[My Account](#)
[Sign in to the Console](#)

NASA NEX

NASA NEX is a collaboration and analytical platform that combines state-of-the-art supercomputing, Earth system modeling, workflow management and NASA remote-sensing data. Through NEX, users can explore and analyze large Earth science data sets, run and share modeling algorithms, collaborate on new or existing projects and exchange workflows and results within and among other science communities.

Three NASA NEX data sets are now available to all via Amazon S3. One data set, the NEX downscaled climate simulations, provides high-resolution climate change projections for the 48 contiguous U.S. states. The second data set, provided by the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on NASA's Terra and Aqua satellites, offers a global view of Earth's surface every 1 to 2 days. Finally, the Landsat data record from the U.S. Geological Survey provides the longest existing continuous space-based record of Earth's land.

Accessing NASA NEX Data

AWS is making the NASA NEX data available to the community free of charge. There are a variety of ways to access the data:

- Simple HTTP requests
- AWS Command Line Tools and SDKs (Ruby, Java, Python, .NET, PHP, etc.)
- [Amazon EC2](#) (Resizable compute capacity)
- [Amazon Elastic MapReduce](#) ([Amazon EMR](#) (Managed Hadoop service))

The data is hosted for free by AWS as part of the AWS Public Data Sets program.

Available NASA NEX Data Sets

Downscaled Climate Projections (NEX-DCP30)

The NASA Earth Exchange (NEX) Downscaled Climate Projections (NEX-DCP30) dataset is comprised of downscaled climate scenarios for the contiguous United States that are derived from the General Circulation Model (GCM) runs conducted under the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al. 2012) and across the four greenhouse gas emissions scenarios known as Representative Concentration Pathways (RCPs) (Mearns et al. 2011) developed for the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5). The dataset includes downscaled projections from 53 models, as well as ensemble statistics calculated for each RCP from all model runs available. The purpose of these datasets is to provide a set of high resolution, bias-corrected climate change projections that can be used to evaluate climate change impacts on processes that are sensitive to finer-scale climate gradients and the effects of local topography on climate conditions.

Each of the climate projections includes monthly averaged maximum temperature, minimum temperature, and precipitation for the periods from 1950 through 2005 (Retrospective Run) and from 2006 to 2099 (Prospective Run).

Available at <s3://nasanex/NEX-DCP30>

Global Daily Downscaled Projections (NEX-GDDP)

The NASA Earth Exchange (NEX) Global Daily Downscaled Projections (NEX-GDDP) dataset is comprised of downscaled climate scenarios that are derived from the General Circulation Model (GCM) runs conducted under the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al. 2012) and across the two of the four greenhouse gas emissions scenarios known as Representative

- NASA Earth Exchange (NEX) Downscaled Climate Projections (NEX-DCP30)
- Monthly averaged maximum temperature, minimum temperature, and precipitation from 1950 through 2005 (Retrospective Run) and from 2006 to 2099 (Prospective Run).
- **Total Dataset Size: 17 TB**
- Individual file size: 2 GB

How do I organize 17 TB of data?




Familiar Process


1. Find space to store data
2. Determine compute requirements (one massive server)
3. Run Processing Jobs

Dell Storage MD1400

Starting Price \$14,094.20
Instant Savings **\$4,800.10**

Subtotal \$9,294.10

 Discount Details
 Ships in 5 - 7 Business Days
 Print Summary

 Continue

Cloud

1. Store data on Amazon S3
2. Amazon EMR – Map/Reduce Platform
3. Run Processing Jobs

Storage Pricing

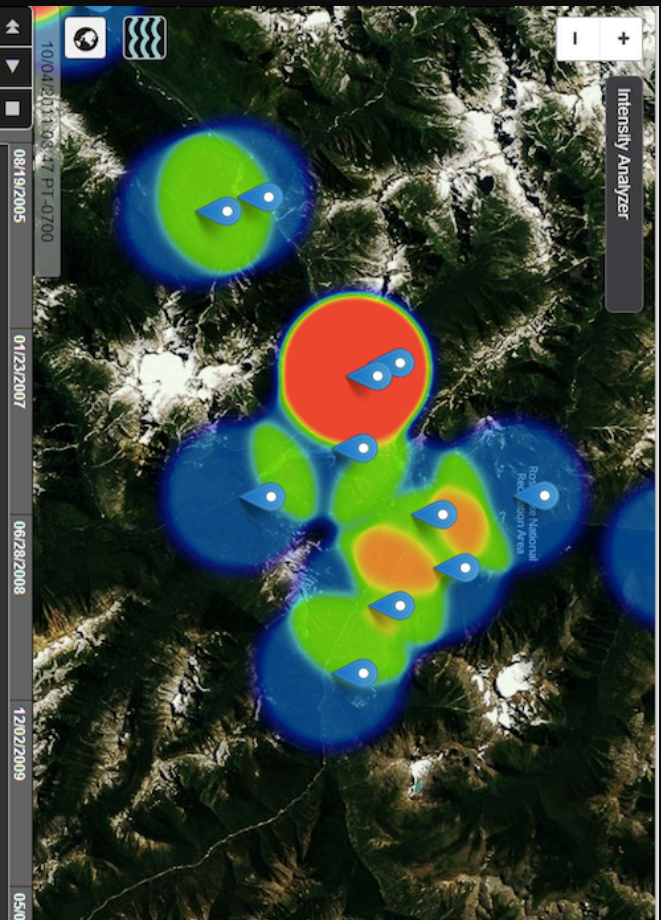
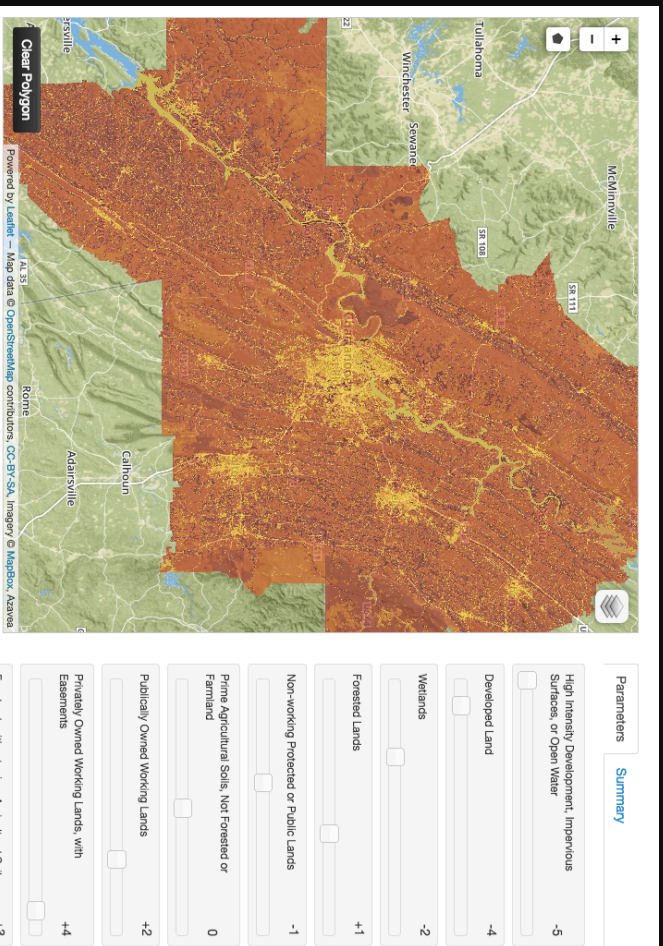
Region:

	Standard Storage	Reduced Redundancy Storage	Glacier Storage
First 1 TB / month	\$0.0300 per GB	\$0.0240 per GB	\$0.0100 per GB
Next 49 TB / month	\$0.0295 per GB	\$0.0236 per GB	\$0.0100 per GB
Next 450 TB / month	\$0.0290 per GB	\$0.0232 per GB	\$0.0100 per GB
Next 500 TB / month	\$0.0285 per GB	\$0.0228 per GB	\$0.0100 per GB
Next 4000 TB / month	\$0.0280 per GB	\$0.0224 per GB	\$0.0100 per GB
Over 5000 TB / month	\$0.0275 per GB	\$0.0220 per GB	\$0.0100 per GB



New Sources of Data

Beyond the spreadsheet. Asynchronous and synchronous data from new sources.



Analytics and Visualization



Where to Start

New Tools | Familiar Process

Core Data Management Principles Still Apply

- What analysis do your customers want to perform?
- What type of data do you need to manage?
- How often will you need to collect data?
- How will data be collected?
- How do you secure your data?
- What is your disaster recovery plan?

Don't be Afraid to Work Backward

User Experience (UX) matters!

- Creating rich and **fast** user experiences requires work
- Don't be afraid to think outside the box

Be Cognizant of Data Storage

- Storage and CPU changes can add up quickly in the Big Data world
- Have a plan for how long you will retain data and backups
- Understand upfront what Map Reduce Queries will be necessary

Contact Info

- Email: bdeaver@geoengineers.com

Questions & Answers