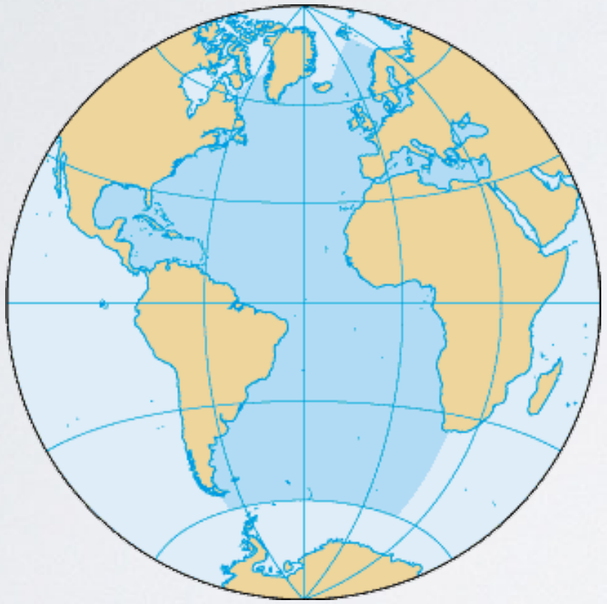
Data Intelligence for All

Christopher Nguyen, PhD
Co-Founder & CEO

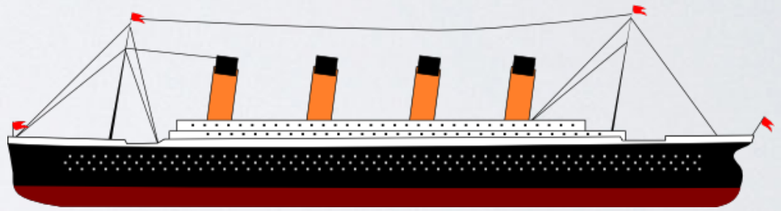
Presented on December 2, 2013

What do you get when you cross …



Atlantic

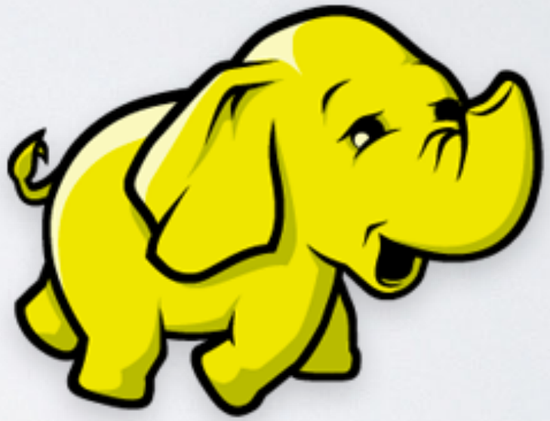
×



Titanic


About
Halfway!!
Chart of the
ATLANTIC OCEAN

What do you get when you cross …



What do users need?

Let’s take a look!

… and that's what we're working on at Adatao

0 0 8 9 1

In the beginning,
there was **darkness**...

1990

Then came Business Intelligence

You could see what happened with your business

BIG DATA
PROBLEMS

Huge Volume

High Velocity

Great Variety

2010

HADOOP
HIVE
MAPREDUCE

helped solve big data problems

but businesses still
LOOKING BACKWARD

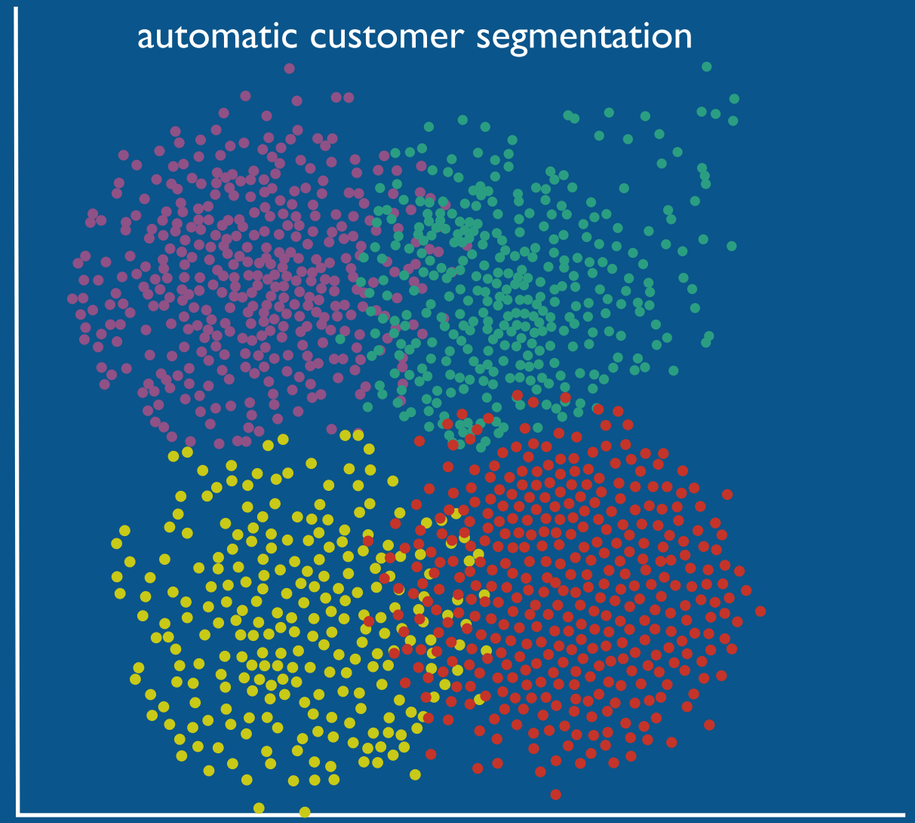
BIG DATA + BIG COMPUTE = OPPORTUNITIES

~~BIG DATA= PROBLEMS~~

Machine Learning

Predictive Analytics

Natural Language



2012

Business Users | Data Scientists | Data Engineers

PINSIGHT

BIG INSIGHTS

- Visually Beautiful
- Interactive Data Exploration
- Narrative Web App

PANALYTICS

BIG COMPUTE

- Powerful In-Memory Data Mining
- Machine Learning Big Analytics Platform

ADATAO

BIG DATA

(Hadoop HDFS, Cassandra, SQL DMBS, Streaming Data)

cassandra

Deep engineering & business experience from Google, Yahoo et al. PhD's in DM & ML from UIUC, Georgia Tech, Stanford, Berkeley, ...

Big-Data Compute Engines, Google Apps Engineering Director, Google Founders' Award, HKUST Prof, 2 successful enterprise exits, Stanford PhD

Hadoop distributed/streaming analytics, Yahoo Hadoop Eng, UIUC PhD

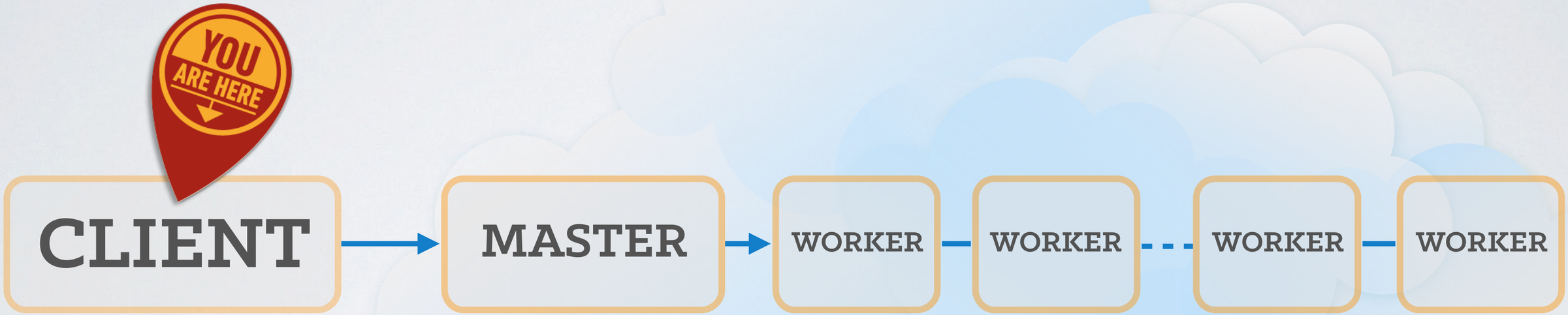
Machine learning & machine vision, US Army Research Lab, Johns Hopkins PhD

Adatao *pInsight* demo

pINSIGHT

Demo Deployment

Diagram



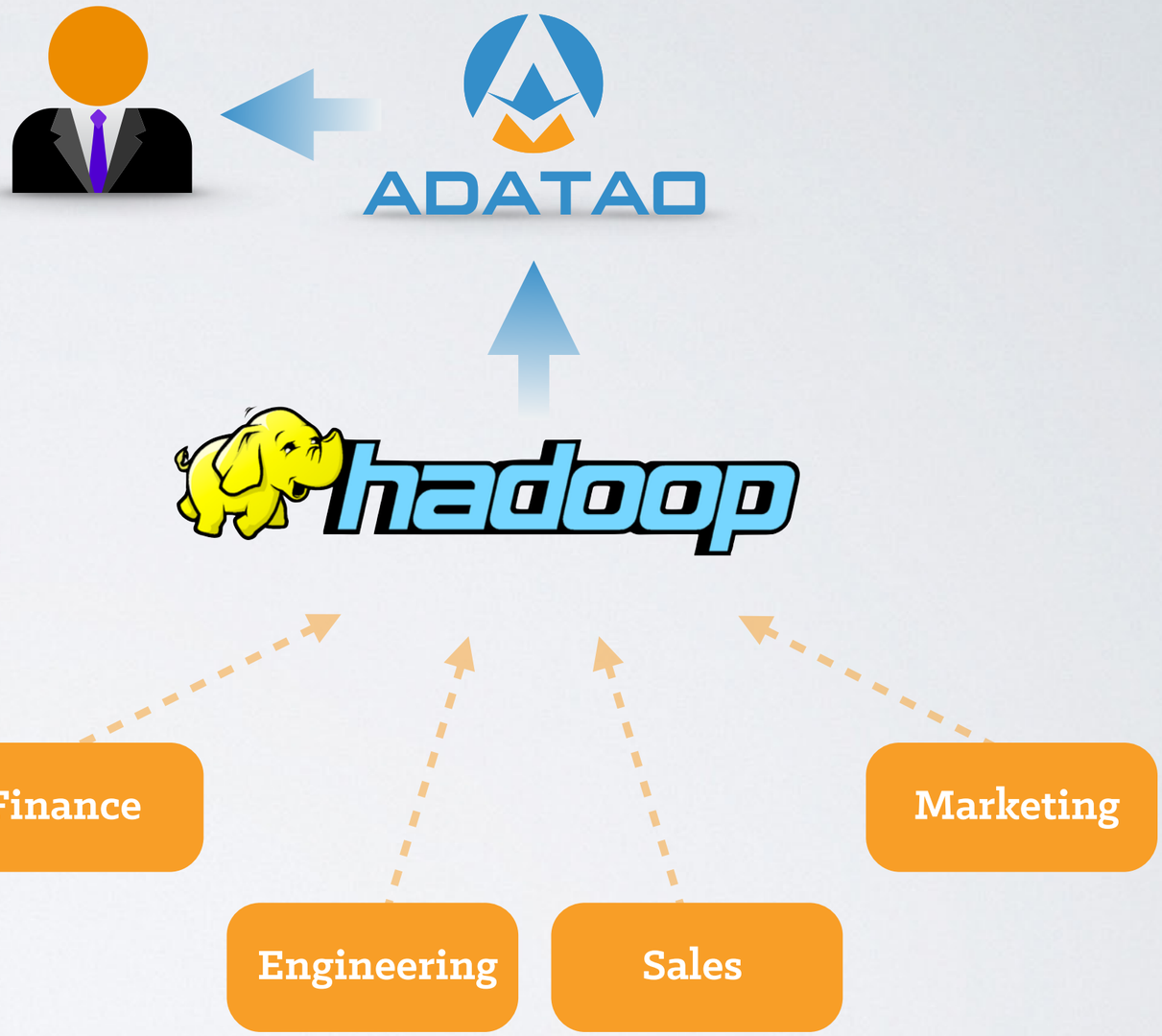
Adatao *pAnalytics* demo

PANALYTICS

Bonus Demo

Use Cases

Internet Service Provider



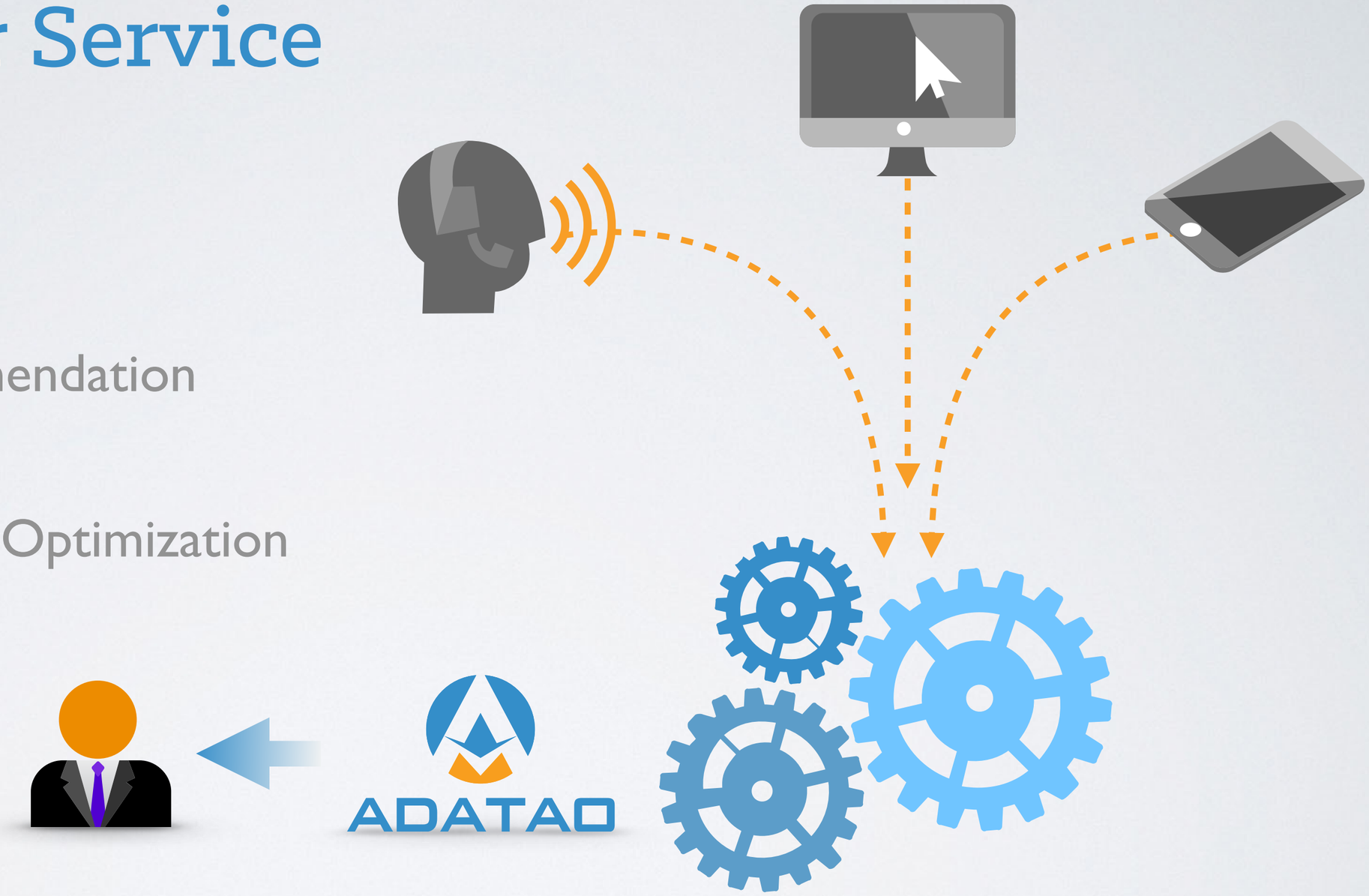
Interactive,
Ad Hoc Business Query

Insight Discovery on
Aggregated
Operational Data

Customer Service
Provider

Product Recommendation

Cross-channel
User Experience Optimization



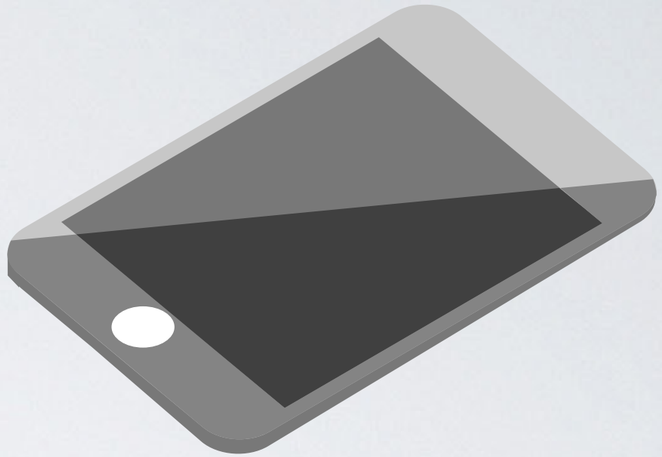
Heavy Equipment Manufacturer

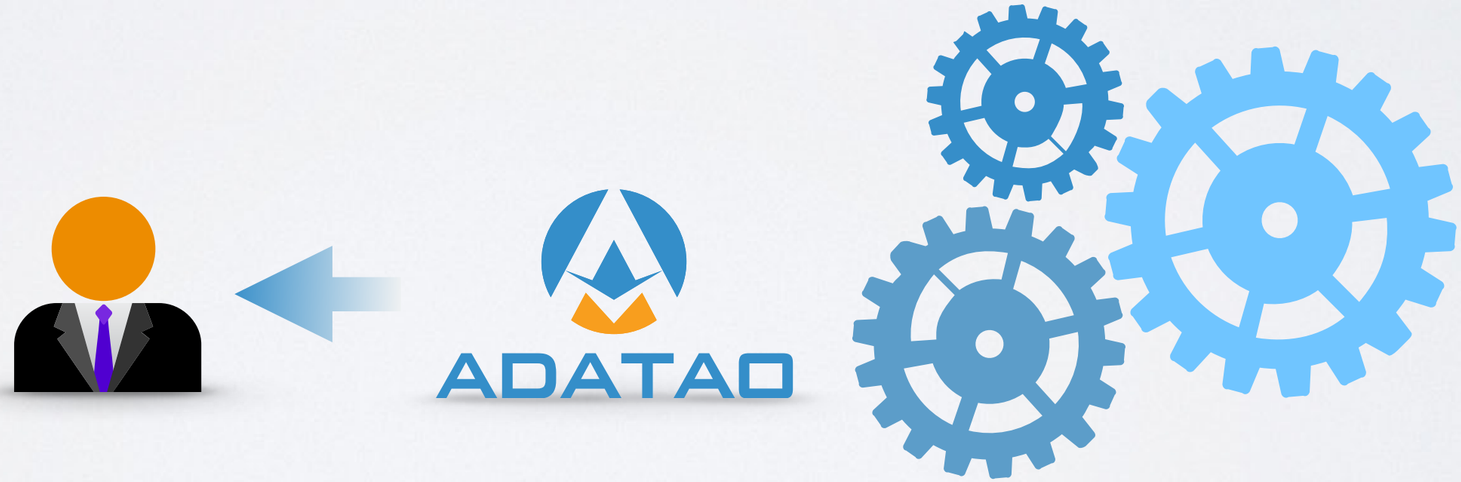
Sensor Network Analytics for Predictive Maintenance

Mobile Ad Platform

Ad Targeting

CTR Prediction





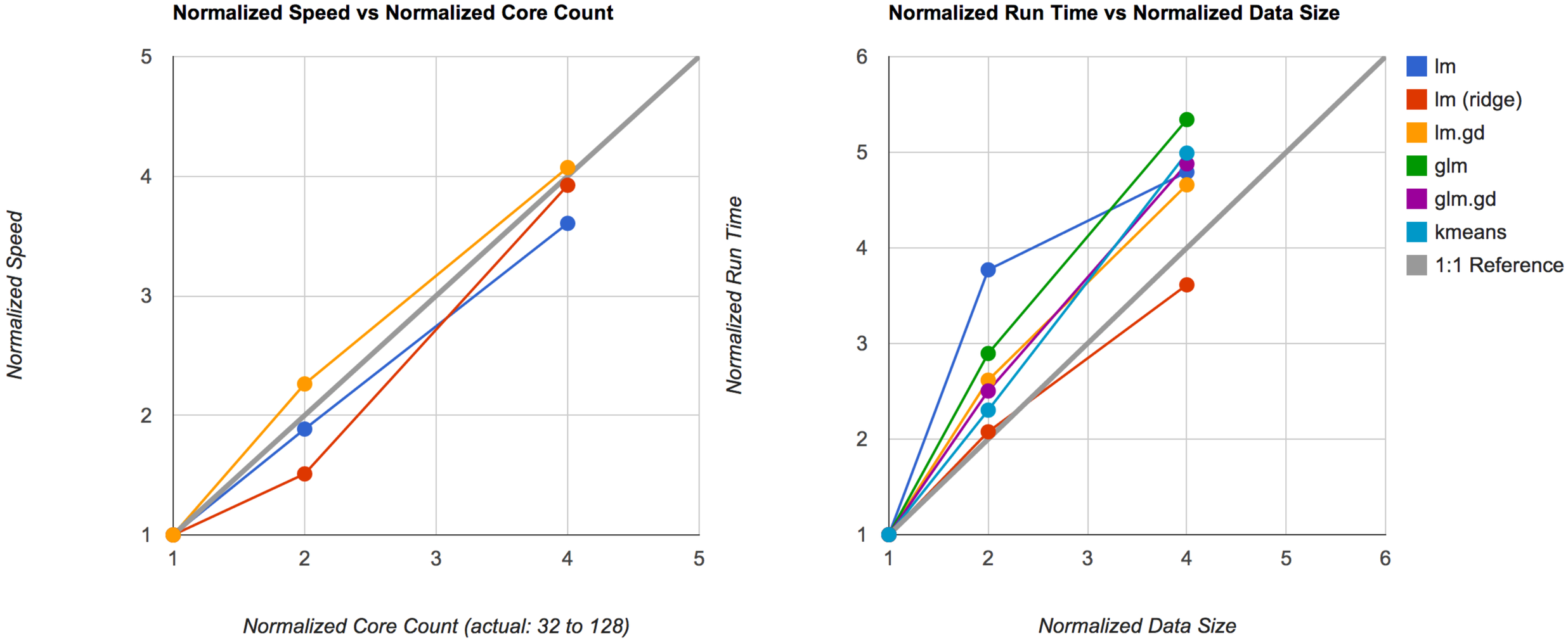
Scaling Performance

Algorithm	Run time (sec) for 50GB (800M rows)	Per-Core Throughput (MB/sec-core)	Per-Machine Throughput (MB/s)
bigr.lm (ridge)	3.04	130	1,040
bigr.lm	4.05	102	816
bigr.lm.gd	12.2	32	256
bigr.glm.gd	24.5	16	128
bigr.glm	36.1	11	88
bigr.kmeans	335	1.2	9.6

pAnalytics performance on building machine learning models with cluster Adata016 (m3.2xlarge) on a 50GB data set of 5 features and 800 million rows. (Gradient descent algorithms are over 5 iterations)

Algorithm	Run time (sec) for 1.1 TB dataset (1.6B rows)	Per-Core Throughput (MB/sec-core)	Per-Machine Throughput (MB/s)
bigr.lm (ridge)	70.9	130	1,040
bigr.lm	74.9	123	984
bigr.lm.gd	127	72.8	582
bigr.glm.gd	145	63.6	509

pAnalytics performance on building machine learning models with cluster Adata040 (m3.2xlarge) on a 1.1 TB data set of 40 features and 1.6 billion rows. (Gradient descent algorithms are over 5 iterations)


Normalized Speed vs Normalized Core Count
Normalized Speed
Normalized Core Count (actual: 32 to 128)
1
2
3
4
5
Normalized Run Time vs Normalized Data Size
Normalized Run Time
Normalized Data Size
1
2
3
4
5
6
lm
lm (ridge)
lm.gd
glm
glm.gd
kmeans
1:1 Reference

Data Intelligence
for All

ADATAO

Business Users

PINSIGHT

Fast & Easy
Business Analytics

Natural Language

Beautiful Web UI

Data Scientists
& Engineers

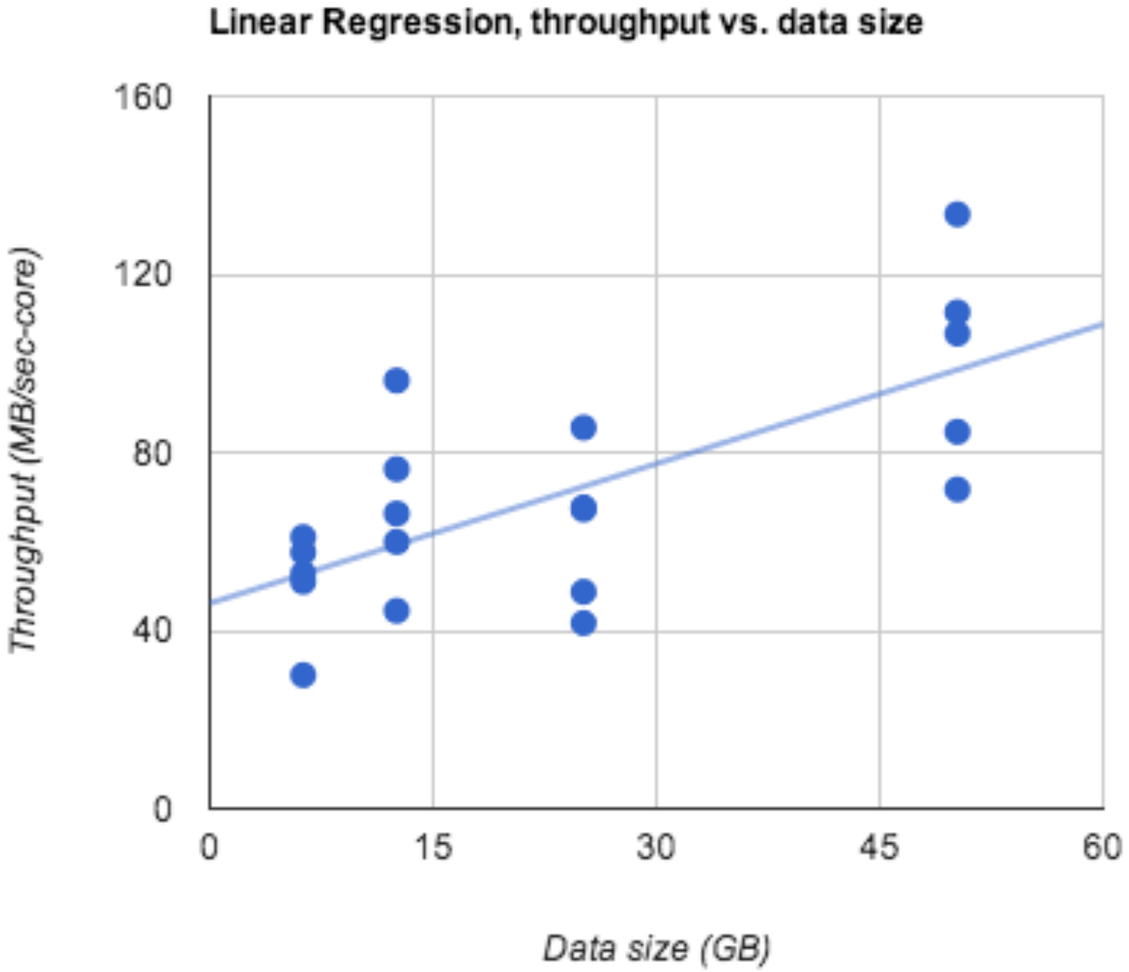
PANALYTICS

Big & Fast
Data Science

R, Python, REST API

Data Mining & ML

Thanks for contributing to the Spark Community!


Linear Regression, throughput vs. data size
160
120
80
40
0
0
15
30
45
60
Throughput (MB/sec-core)
Data size (GB)