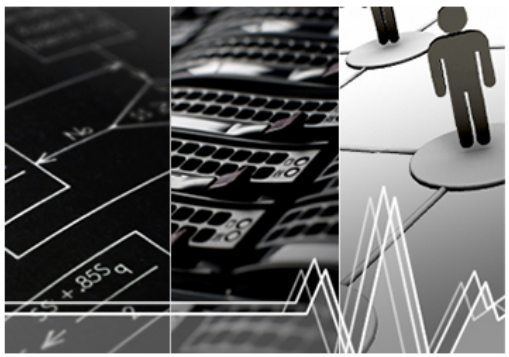
amplab
UC BERKELEY

Making Sense at Scale with Algorithms, Machines & People

PI: Michael Franklin
University of California, Berkeley

Expeditions in Computing PI Meeting
May 15, 2013



The Berkeley AMPLab

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

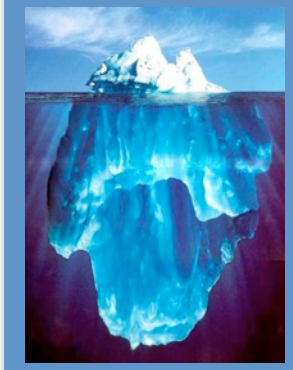
National Science Foundation: In addition to funding the Big Data solicitation, and keeping with its focus on basic research, NSF is implementing a comprehensive, long-term strategy that includes new methods to derive knowledge from data; infrastructure to manage, curate, and serve data to communities; and new approaches to education and workforce development. Specifically, NSF is:

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers;
- Funding a $10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing;

amplab

Sources Driving Big Data

It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Friend Request
Transaction
Network message
Fault
…

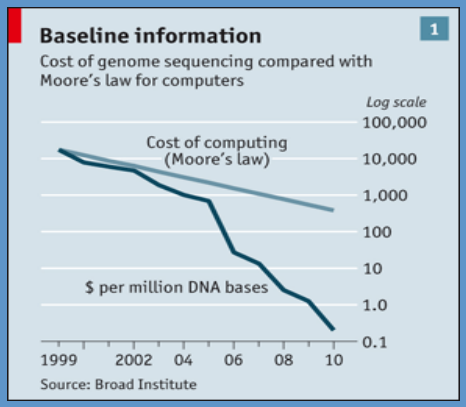
User Generated (Web & Mobile)



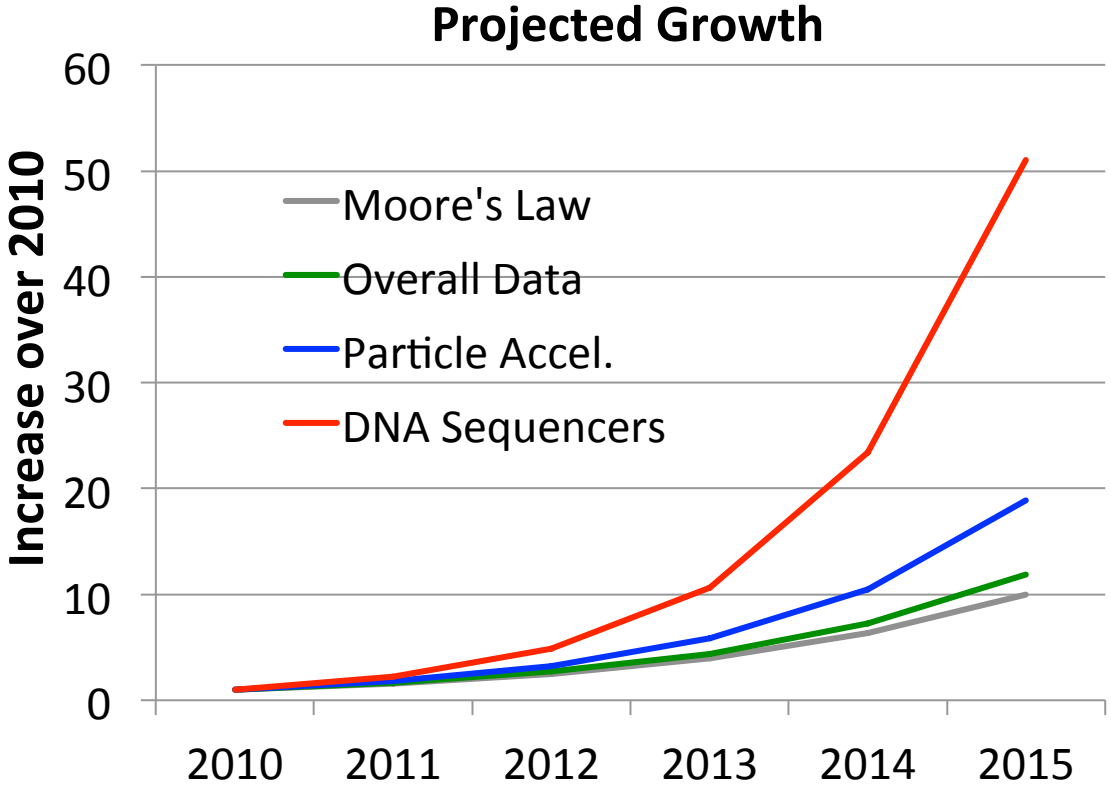

. . .
..

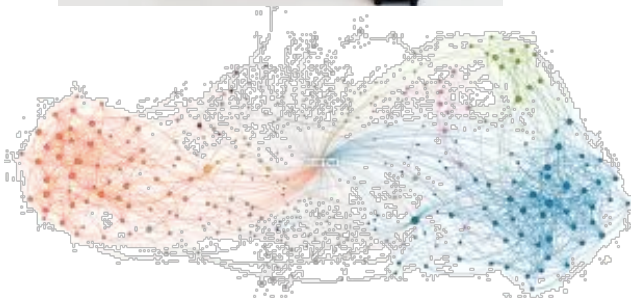

Internet of Things / M2M

Scientific Computing



Challenge 1: Data is Big





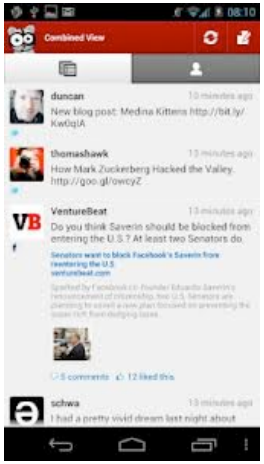
Data Grows faster than Moore's Law

[IDC report, Kathy Yelick, LBNL]

amplab

Challenge 2: Data is Dirty

- Variety of diverse sources
- Uncurated
- No schema
- Inconsistent syntax and semantics

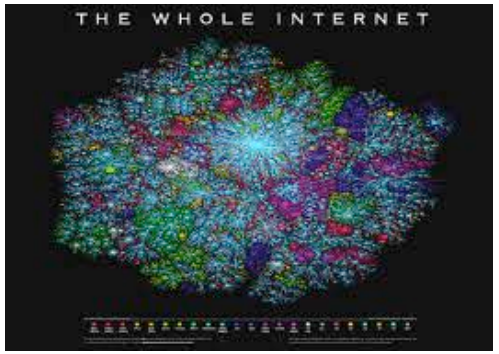


Dirty Data worse than Big Data

amplab

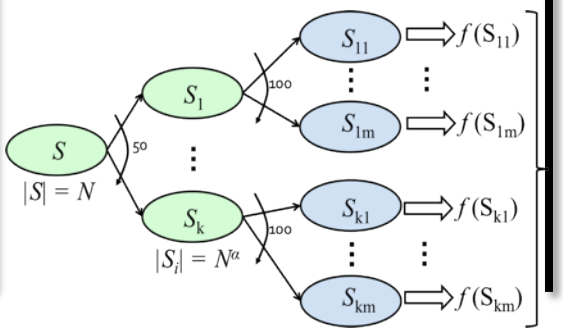
Challenge 3: Complex Questions

- **Hard** questions
  - What is the impact on traffic and home prices of building a new on-ramp?
- Detect **real-time** events
  - Is there a cyber attack going on?
- **Open-ended** questions
  - How many supernovae happened last year?



amplab

Our Vision: A Necessary Synergy

	Algorithms	Machines	People
Challenge 1: Data is Big	✔	✔	
Challenge 2: Data is Dirty	✔	✔	✔
Challenge 3: Questions are complex	✔	✔	✔
			

The AMPLab Big Bets

- Traditional intellectual borders hinder "Big Data" stacks
  - Need Machine Learning/Systems/Database Co-Design
  - Requires Cohabitation and Real Collaboration

- Now is a unique opportunity to rethink fundamental design points:
  - Changing Latency Demands
  - Changing Consistency Requirements
  - Cloud-based Elastic Resources
  - Huge Desire for New Solutions in the Marketplace
  - Open Source is the key to Tech Transfer in Big Data

- Need to consider role of people *throughout the entire* analytics lifecycle

AMPLab: Collaborative Research

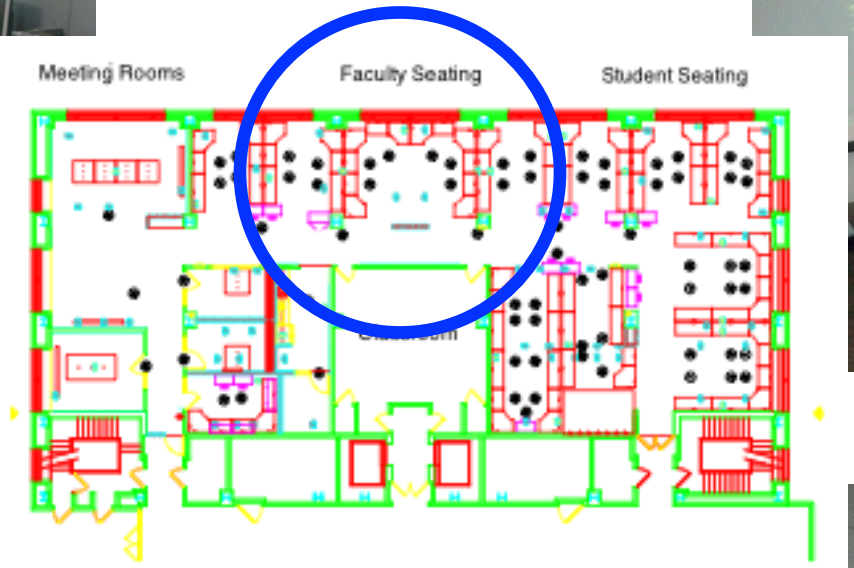
An integration of Faculty Interests (*Directors):

Alex Bayen (Mobile Sensing)	Anthony Joseph (Sec./ Privacy)
Ken Goldberg (Crowdsourcing)	Randy Katz (Systems)
*Michael Franklin (Databases)	Dave Patterson (Systems)
Armando Fox (Systems)	*Ion Stoica (Systems)
*Mike Jordan (Machine Learning)	Scott Shenker (Networking)

50+ amazing grad students, post-docs, undergrads, developers, staff & visitors

Twice-Yearly Research Retreats (industry & sponsors):

Co-Located for Collaboration



Collaboration: Industry + Government

AMPLab Launched January 2011 (5 yr plan)

Founding Sponsors:

Sponsors and Affiliates:

Federal Grants and Contracts:

Expeditions in Computing

XData Program

amplab

Collaboration: Applications

Participatory Sensing

Mobile Millenium - Traffic

Collective Discovery

Opinion Space - Opinions

Carat – Smartphone energy

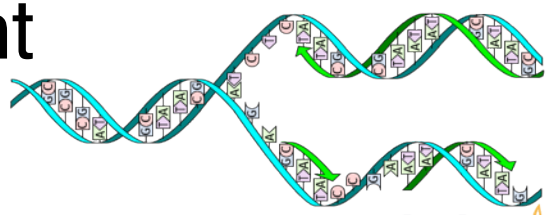
Urban Planning and Simulation

UrbanSim – data integration

Cancer Genomics/Personalized Medicine (w/ UCSF and UCSC)

SNAP: Fast Sequence Alignment

Genome Data Warehouse



amplab

Shared Deliverable: Berkeley Data Analytics Stack (BDAS)

GIGAOM Events Research Jobs paidContent

Home Apple Cleantech Cloud Data Europe Mobile Video

big data / hadoop / open source

Welcome to Berkeley: Where Hadoop isn't nearly fast enough

by Derrick Harris APR. 17, 2013 - 4:19 PM PDT

5 Comments +1

SUMMARY: *Hadoop not fast enough for you? Then you might want to get to know AMPLab, a University of California, Berkeley team*

DBMS2

December 13, 2012

Introduction to Spark, Shark, BDAS and AMPLab

UC Berkeley's AMPLab is working on a software stack that:

amazon web services

Sign Up My Account / Console

AWS Products & Solutions Articles & Tutorials Developers

Browse By Category

AWS Services

- Amazon CloudFront

Run Spark and Shark on Amazon Elastic MapReduce

Articles & Tutorials > Run Spark and Shark on Amazon Elastic MapReduce

Learn how to run Spark (in-memory MapReduce) and Shark (Hive on Spark) on Amazon EMR.

O'REILLY

Strata

Making Data Work

Shark: Real-time queries and analytics for big data

Shark is 100X faster than Hive for SQL, and 100X faster than Hadoop for machine-learning

by Ben Lorica | @bigdata | Comment | November 27, 2012

December 01, 2012

The Week in Big Data Research

Datanami Staff

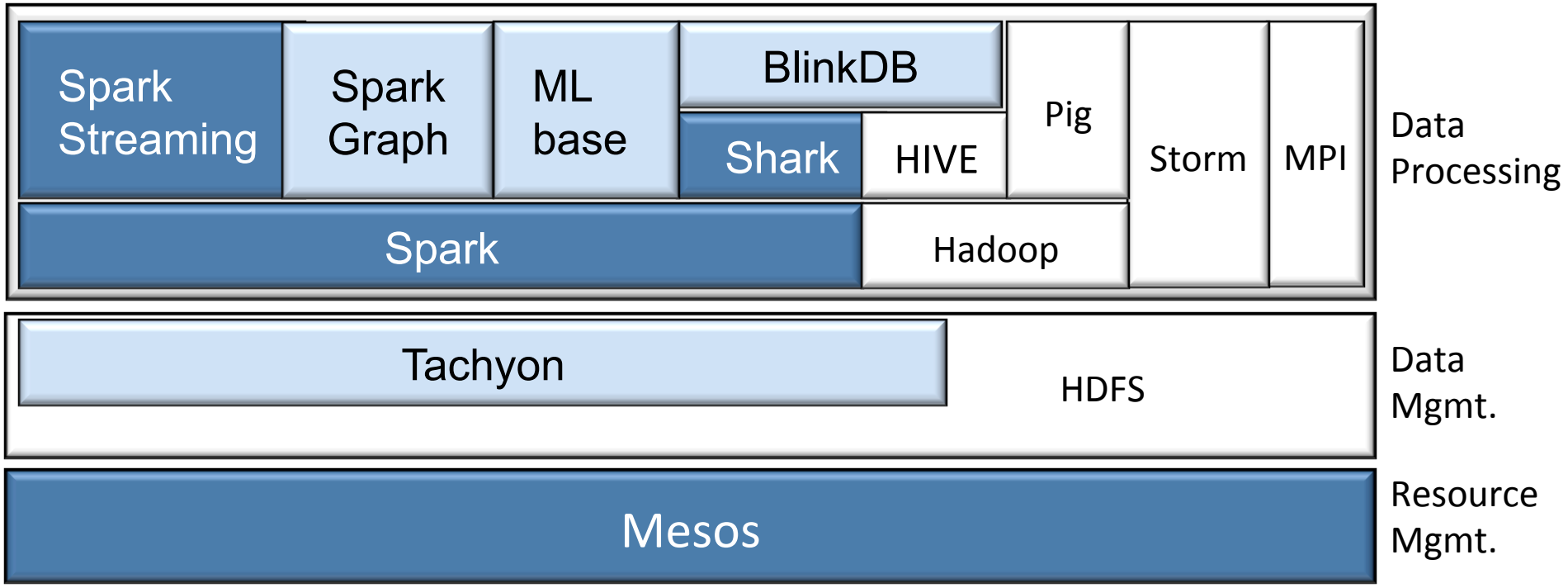
datanami

BIG DATA • BIG ANALYTICS • BIG INSIGHTS

Shark Attack on SQL and Analytics

A research team from the AMPLab at EECS on the UC Berkeley c

BDAS: Current Snapshot

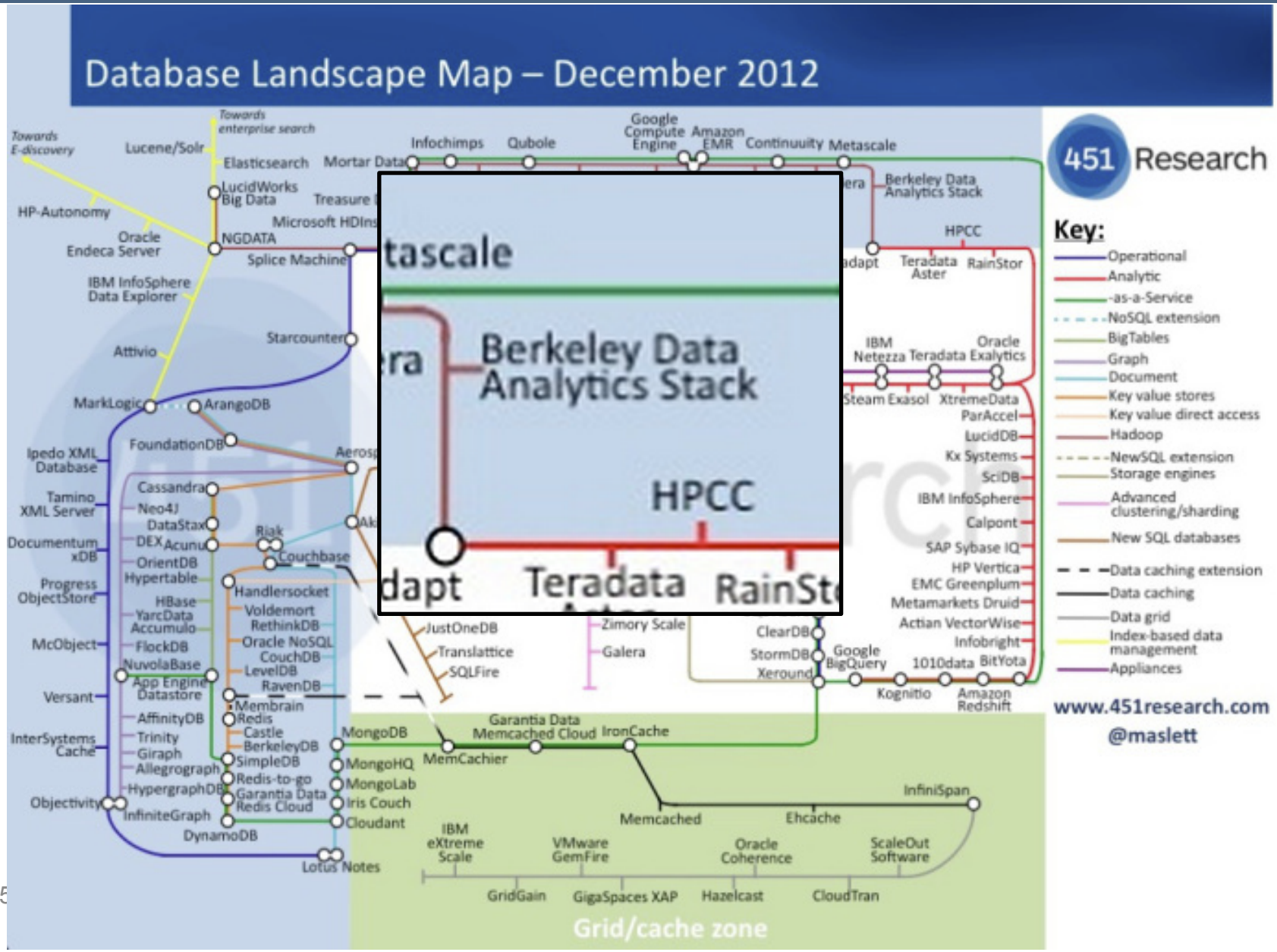


Released (BDAS) In development (BDAS) Existing open source stack

BDAS Components being released under BSD or Apache Open Source License

amplab

Big Data Landscape – Our Corner

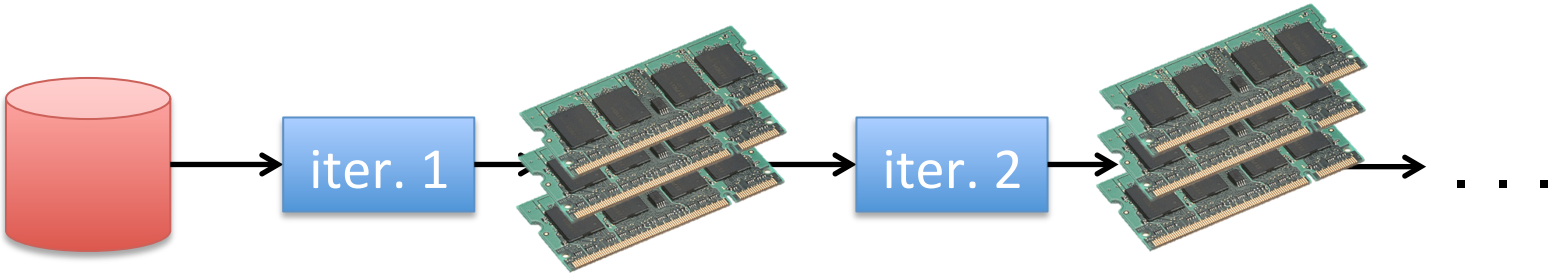


Impact (so far)

- Open Source Release of BDAS components:
  - Mesos: Cluster Virtualization
    - Business critical services on 6000+ servers at Twitter
    - see “How Twitter Rebuilt Google’s Secret Weapon” *Wired 3/13*
  - Spark: In-memory Computation Framework &
    Shark: Hive-Compatible SQL Query Engine on Spark
    - in use at large companies, start ups, and govt. agencies
    - 100x Performance Improvement over Hadoop/Apache Hive
    - available on Amazon Elastic Map Reduce
    - 700+ member Meetup group
- Best Paper Awards: Eurosys 13, ICDE 13, NSDI 12, SIGCOMM 12 and Best Demo Award: SIGMOD 12
- Students in high-demand in academia and industry

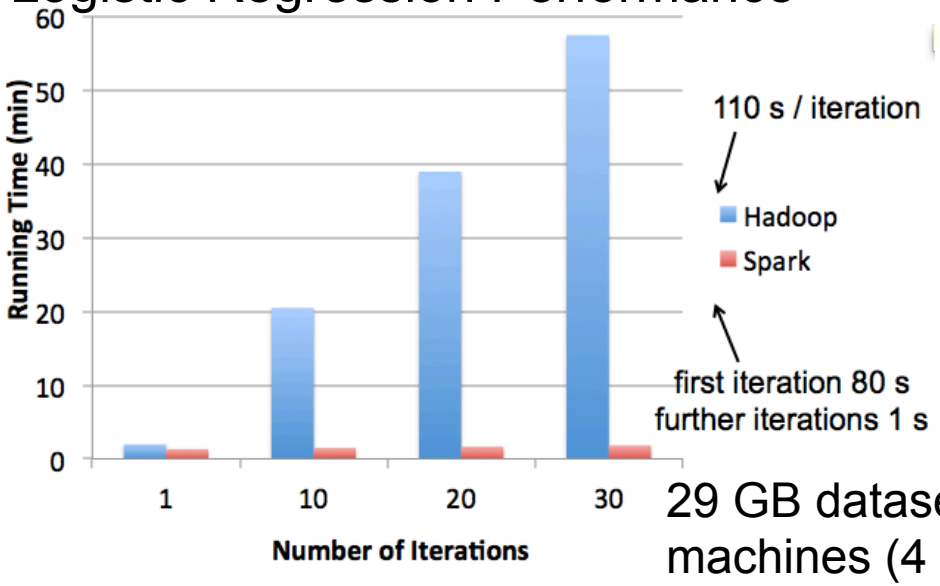
Spark: Sys/ML Collaboration at Work

Technical Challenge: disk-oriented Hadoop Map Reduce inefficient for iterative Machine Learning



Research Challenge Addressed: How to design a distributed memory abstraction that is both **fault-tolerant** and **efficient**?

Logistic Regression Performance

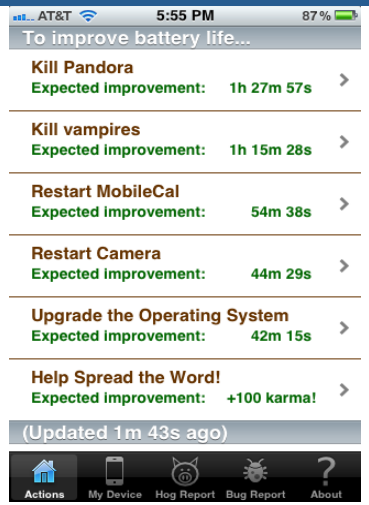


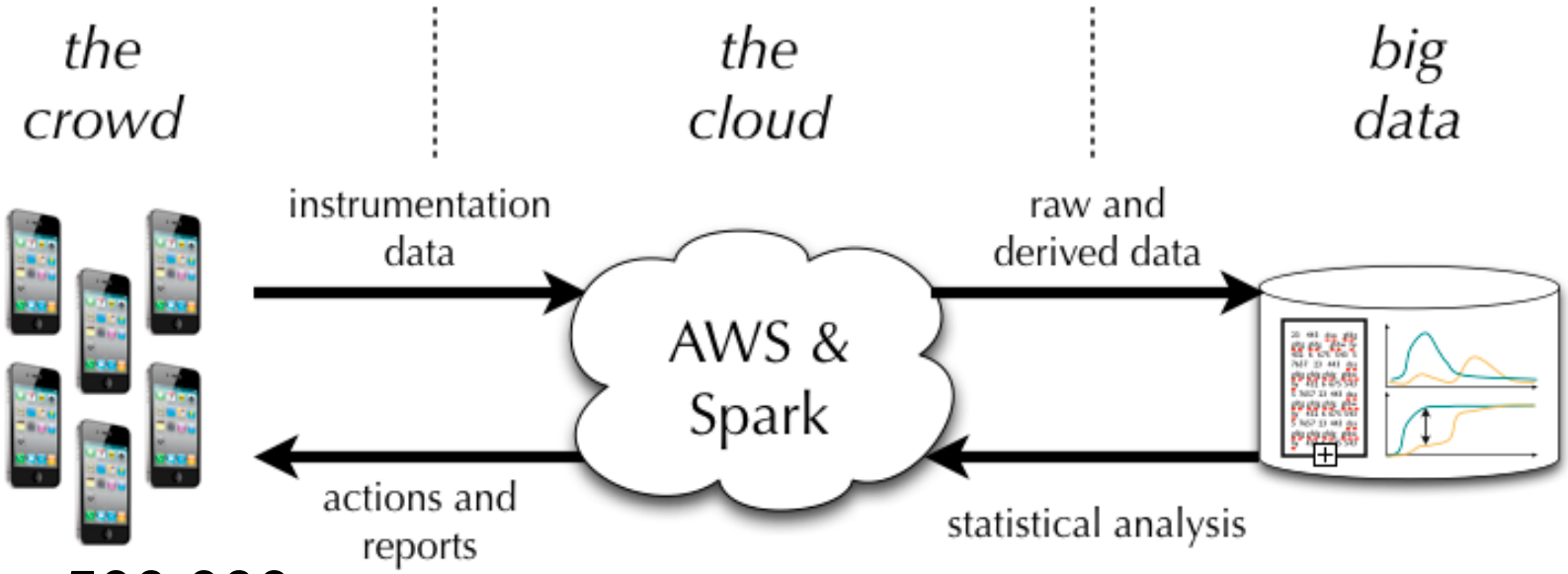
Solution: Resilient Distributed Datasets (RDDs)

Spark
Lightning-Fast Cluster Computing

29 GB dataset on 20 EC2 m1.xlarge machines (4 cores each)

Impact: Carat Smartphone App





Over 500,000 downloads

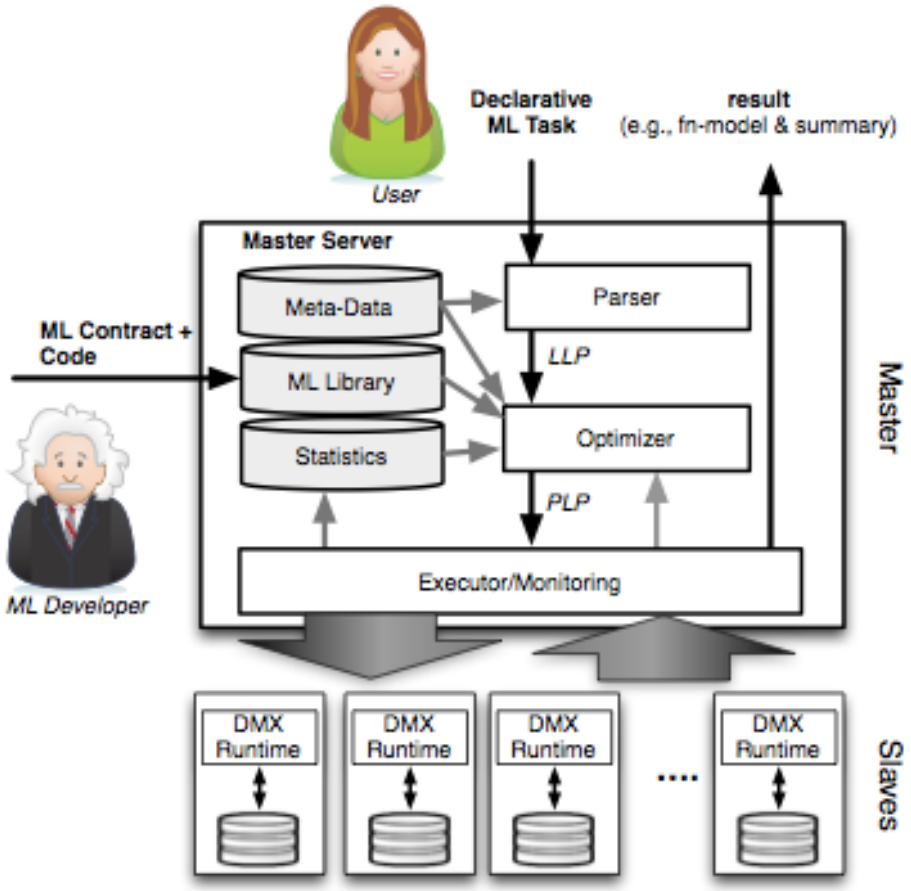
amplab

MLBase – Declarative ML

Vision:
Make Machine Learning usable by "mere mortals"

Allow high-level (declarative) specification of ML tasks

Use Database-style "query optimization to generate efficient execution strategy



amplab

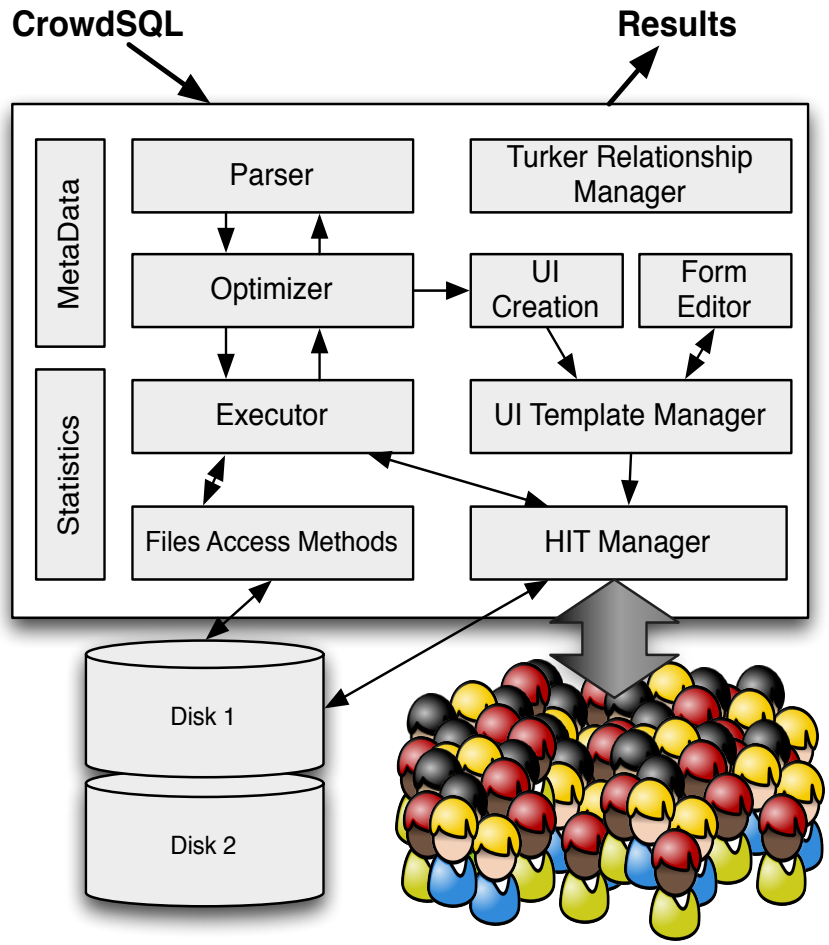
Hybrid Human/Machine Systems

Use machines for bulk data processing

Leverage human activity for data collection and event detection

Leverage human knowledge, reasoning and perception for:

- subjective entity comparisons
- complex predicates
- finding missing data
- disambiguating questions



e.g., CrowdDB Architecture

amplab

Outreach

AMPCamp I @ Berkeley, August 2012
AMPCamp II @ Strata Conf., Feb 2013
AMPCamp III @ Berkeley, August 2013
AMPCamp Online: ampcamp.berkeley.edu

ampcamp
Big Data Bootcamp

amplab

What do we get from Expeditions?

Simply put – the ability to

"swing for the fences"

For More Information

amplab.cs.berkeley.edu

- Papers and Project Pages
- News updates and Blogs

Twitter: @amplab

Github and Apache

http://spark.meetup.com

franklin@cs.berkeley.edu

amplab