

Fisher CIO Leadership Program: “Big Data Analytics: Making Big Data Work”

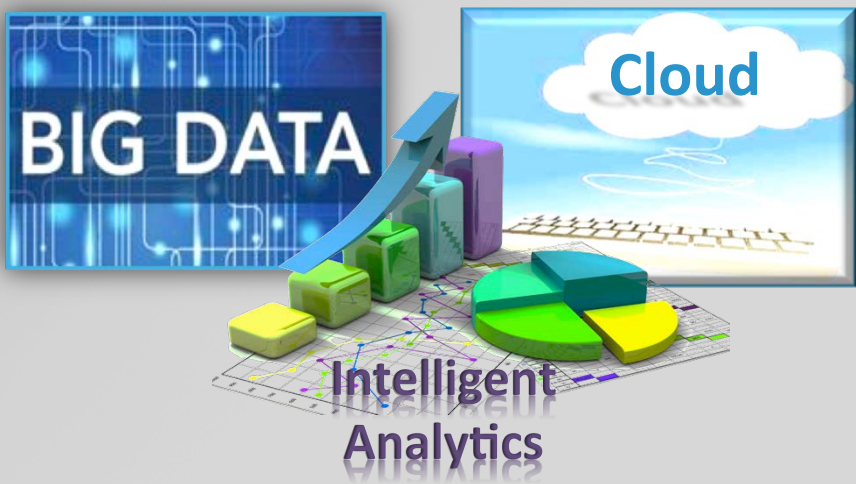
Bill Ruh
Vice President, Global Software Center, GE
November 1, 2012



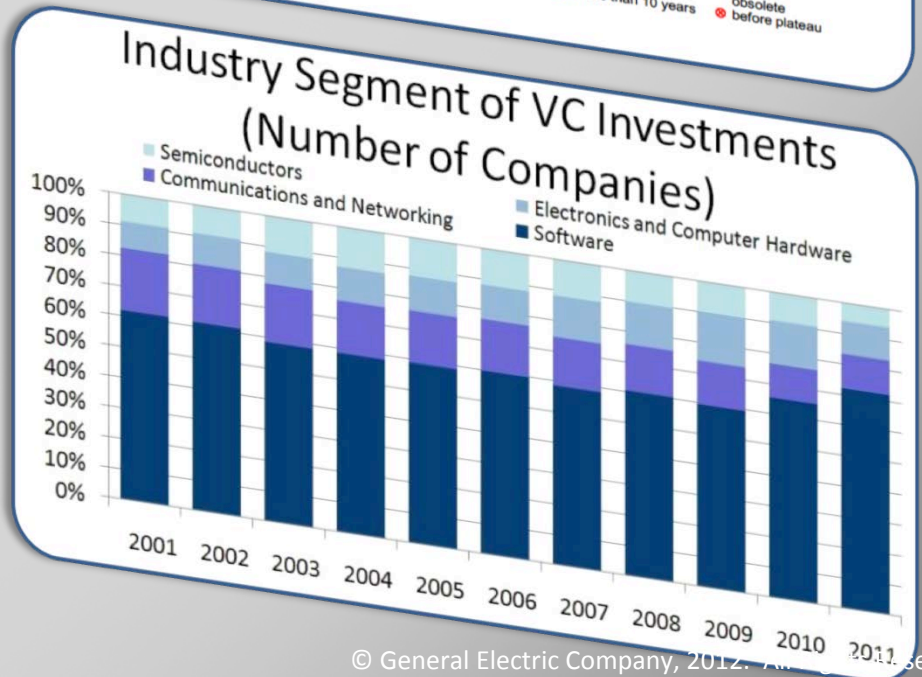
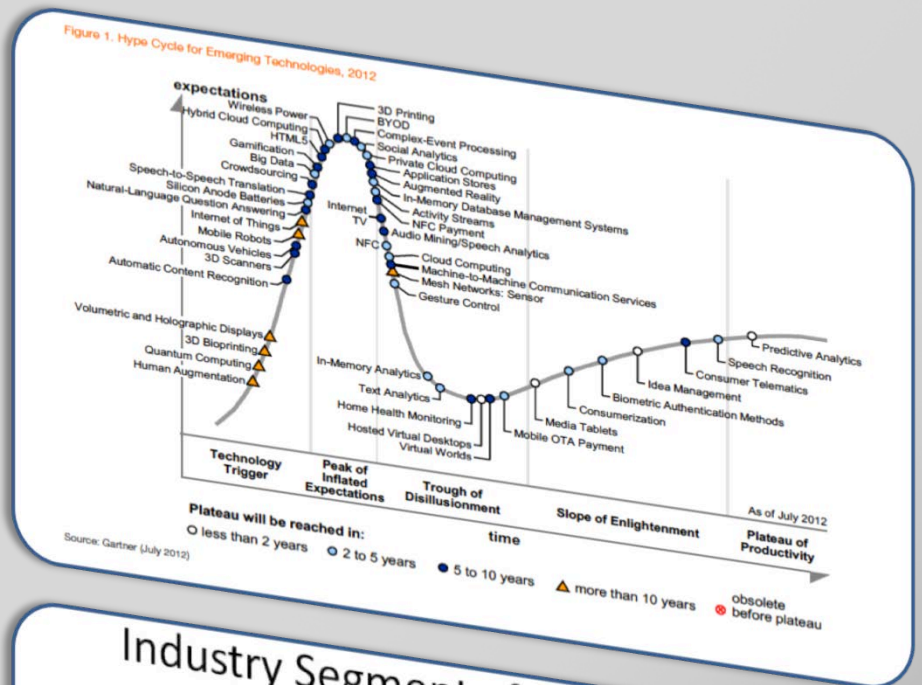
imagination at work

Software is Eating the World*

'Tipping Point' Technologies Will Unlock Long-Awaited Industry Scenarios



In 2001, 62% of VC investments were in software companies, compared to 72% today



Disruption in Industrial Market

Digital industry transformation trends

	INDUSTRY	TRANSFORMATION	ANALOG INDUSTRY	DIGITAL INDUSTRY
2000	Communications: Telco's and Cable	Data Transmission	Landline POTS & MCI	Mobile Internet
2005	Consumer: Retail Media Gaming, Advertising	Transactions & Interactions	Stores – Music, Book, DVD & Tower Records, Borders, Blockbusters	iTunes Kindle Online Media
NOW	Industrials: Energy Healthcare Aviation Mining Transportation ...	Sensing Analytics Control & collaboration	Analog products Manual processes Limited use of sensors & software	HC – telemedicine, digital health records, medical devices Aviation – integrated modular avionics Energy – smart grid, smart buildings

First movers & fast followers win

- New opportunities emerging...enabled by technology, and driven by mega trends
- Rising customer expectations in both cost & complexity reductions
- Accelerating pace of Software innovation...real-time capabilities
- New competitive threats and challenges...and new business models

Forces Shaping the Future

GE is a company that builds the machines that make the world work and has access to and deep understanding of the information that can make them work better

1. Internet

Hyper-connectivity: a living network of machines data and people

Internet of things: more devices tap into the Internet than people on Earth to use them

2. Intelligent Machines

Increasing system intelligence through embedded software

Rise of machines: networked devices overtook the global population in 2011

3. Big Data

Democratization of data

Data overload: 2.5 quintillion bytes of data created every day

4. Analytics

Generating data-driven insights

Enhancing asset performance by detecting & predicting forecasts

Algorithms on installed base

10011010
10101010
10101010
11000101



imagination at work

Scale of Industrial Internet

Social media versus electric generating power source

2012 Twitter Usage

Gas Turbine Compressor Blade Monitoring potential*

VS.



80 Gigabytes per day

enabling social connections



588 Gigabytes per day

enabling capital asset productivity

Data volume potential is 7x greater from a gas turbine than current Twitter usage



imagination at work

Value of Data & Analytics

Monitor fleet of ~25,000* engines ... 3.6MM flight records/month

B777




Prognostics

- ✓ Dispatch reliability
- ✓ Preventive maintenance
- ✓ Asset utilization

+

GE90



Asset Productivity

- ✓ Enhanced service offerings
- ✓ Airline cost structure
- ✓ Fuel performance

=

DATA

90,000 flight records analyzed

~200 parameters per flight record

~18MM parameters per month

System & Optimization

- ✓ Time & space management
- ✓ Fuel efficiency
- ✓ Airspace capacity

Drives strong alignment with customers

Creates productivity in long-term service agreements

Value-added services fuels growth

Prevent failures = customer efficiency

Streamline operations = increased airline productivity

Integrated systems = value-added services



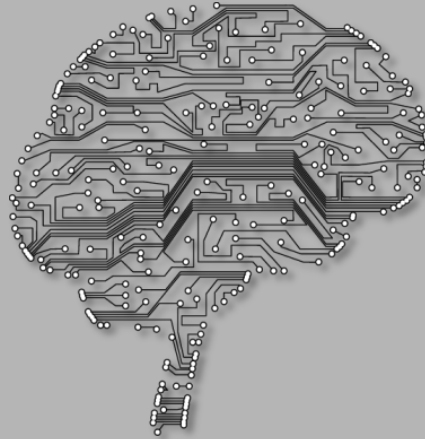
* Includes GE & joint-venture engines with CFM and Engine Alliance. CFM is 50/50 JV with SNECMA. Engine Alliance is 50/50 JV with Pratt & Whitney

The Industrial Internet

Intelligent Machines

Intelligent Information

Intelligent Collaboration

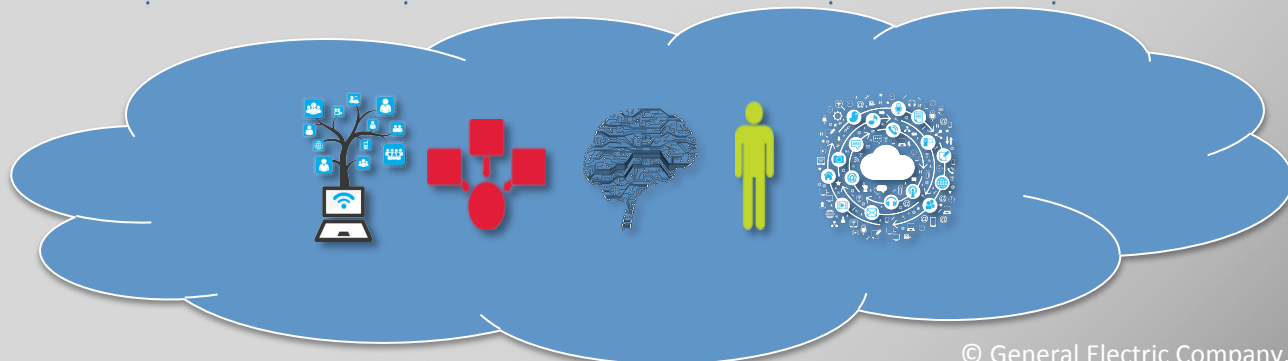


The Hyper-Connected infrastructure

Improved
responsiveness

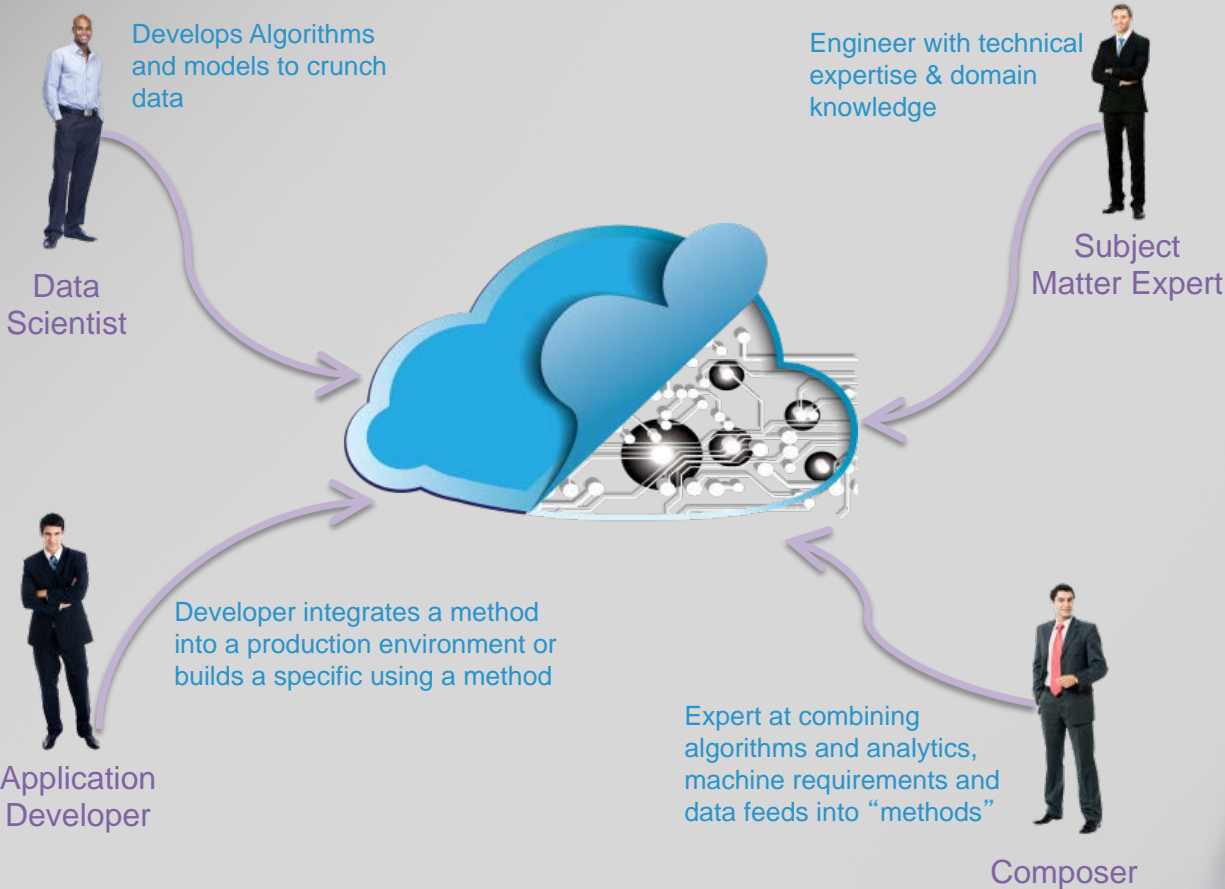
Move from “reactive”
to “**proactive**”

Tracking performance &
accuracy in real-time



GE Analytics Cloud

Ecosystem for collaborative analytics development



Make analytics scalable & repeatable

Accessible & amenable: empower different users

Speed up learning process, discover what we don't know

Deploy analytics into applications & processes

Software & Analytics Strategy

Core technologies to drive productivity...

	Real Time	Physics	Historical Data	Data Analytics
Remote monitoring & diagnostics	-	•	•••	•
Controls/sensors	•	••	•	•
Performance optimization	••	•••	••	••
Usage based	-	•••	•••	•••
<p><i>Note: dots represent level of importance & difficulty</i></p> <p>When to inspect, when to repair, how to operate.</p> <p>GE expertise</p>				

Build solid foundations in every business

GE's advantage...

Customer productivity & operational flexibility

Big Challenges Are Our Future

5 multi-disciplinary R&D centers
40k engineers worldwide
3k research team
8k software developers

500 Industrial Internet experts
150 Data Scientists
Over \$5B in R&D spend

Niskayuna: New York



Over 1,200 engineers
R&D Headquarters
Software Sciences & Analytics

China Technology Park: Shanghai



Over 1150 engineers
Leading ICFC efforts
Connected to innovation centers

John F. Welch Technology Center: Bangalore



Over 4,200 engineers
First global site...1999
Growing emerging markets

Global Research Europe: Munich



Over 170 engineers
Located on tech campus (TUM)
Clean, distributed energy focus

Brazil Technology Center: Rio de Janeiro (2013)



Over 400 engineers
O&G, transportation focus
Customer & university relations

“I find out what the world needs...
Then I proceed to invent it.”

- Thomas Edison



The Next Chapter: Software COEs

Tapping the world's most important information is the science of work

Building a Silicon Valley presence

Focused on software & analytics

190 employees hired in 12 months

Targeting 1000 staff

Award winning facility

Gold LEED

Open architecture: consolidation opportunity

Shared services

GE digital architecture for industrial solutions

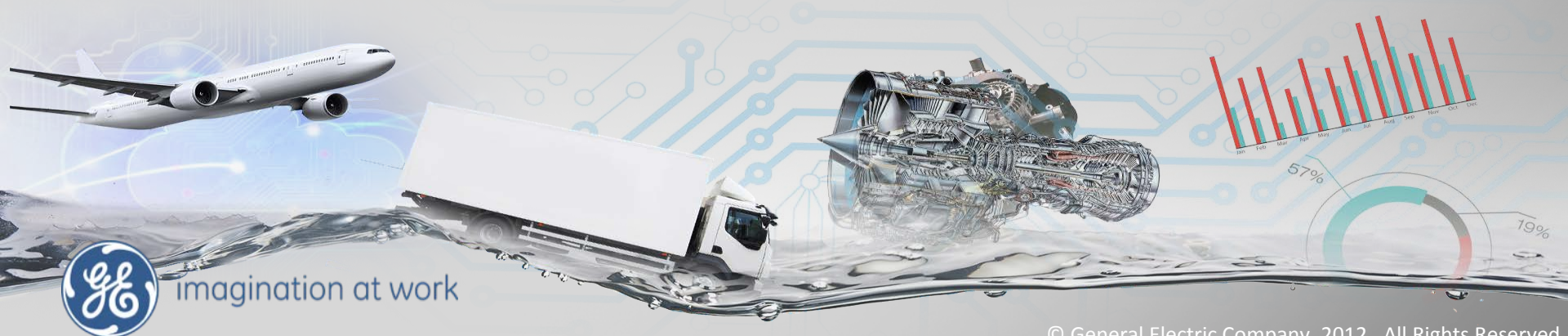
Expertise: user experience, cloud, analytics



imagination at work

What this means?

- Disruption is occurring in every industry – Analog to Digital Industries
- Software coupled with new processing architectures are the enabler for these digital industry architectures
- Future of software is in analytics that drive meaningful & real-time insight
- R&D is critical to lead the change: not just new products but new solutions, systems & industry architectures



Uncovering Hidden Business Potential Through Big Data And Analytics

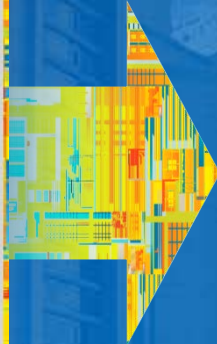
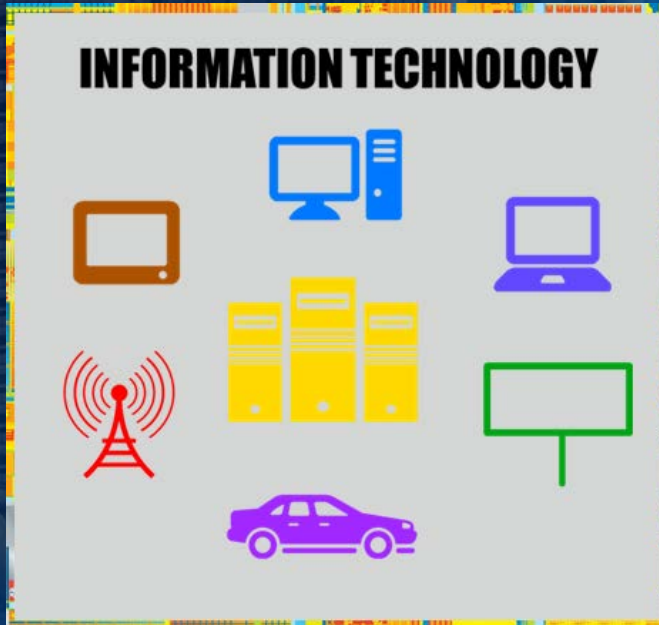
Kim Stevenson
Vice President
Chief Information Officer
Intel Corp.





IN A NEW ERA OF COMPUTING

From Era of **IT** Productivity...



...To Era of **Business** Productivity

The "I" in IT

IMPORTANT
IMPOSSIBLE
INNOVATION
IRRELEVANT
INSPIRED
INGENIOUS
IMPEDIMENT
IN THE WAY
INITIATOR
INEFFICIENT
INVENTIVE
INSIGHT
IMPLEMENTER
INFORMATION
INTEGRATOR
INSIGNIFICANT
INSULAR
INFRASTRUCTURE
INCLUSIVE
INHIBITOR
INSINCERE
IMAGINATION

The “I” in IT



The background features a city skyline at night, with the Oriental Pearl Tower on the left. The scene is overlaid with a grid of binary code (0s and 1s) in a light blue color. The text is centered in the upper half of the image.

Making Sense of Data to
UNLOCK THE POTENTIAL



THE RECIPE

for Insights

Identify The Info

Synthesize It

Ask The Right Questions

Results in a Valuable
Outcome

Sales and Marketing



Manufacturing



Supply Chain



Examples Can Be
Found Across
Multiple Functions

Channel Reseller





Intel Inside Fraud prevention



Post-silicon Validation



Spare Parts Forecasting

A man in a dark polo shirt and glasses stands in a server room, pointing at a server rack. A digital overlay of another man in a colorful shirt is visible on the left. The background is filled with rows of server racks under blue lighting.

Challenges and Opportunities

External Data
Data Visualization
Skills Development

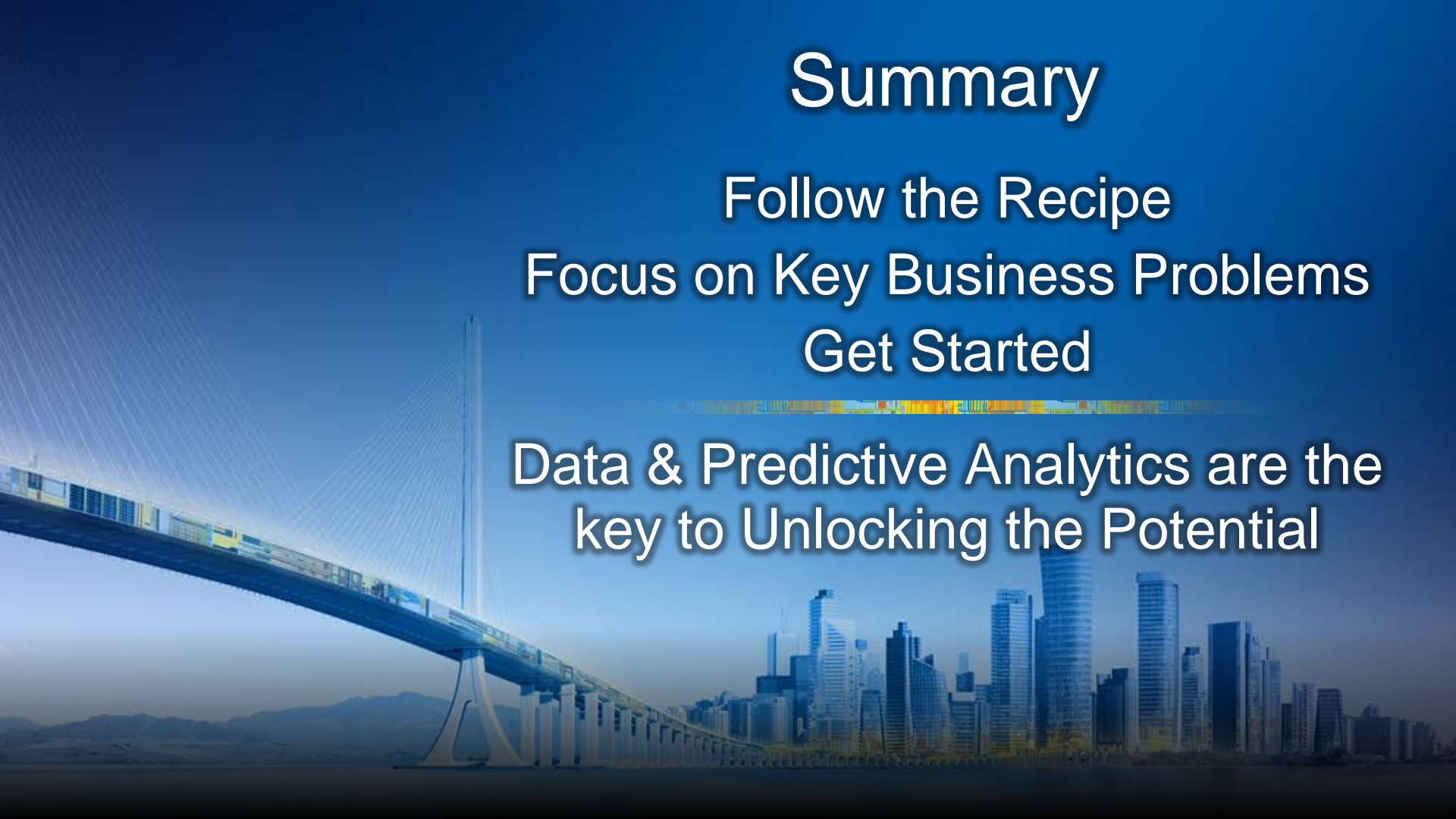
Summary

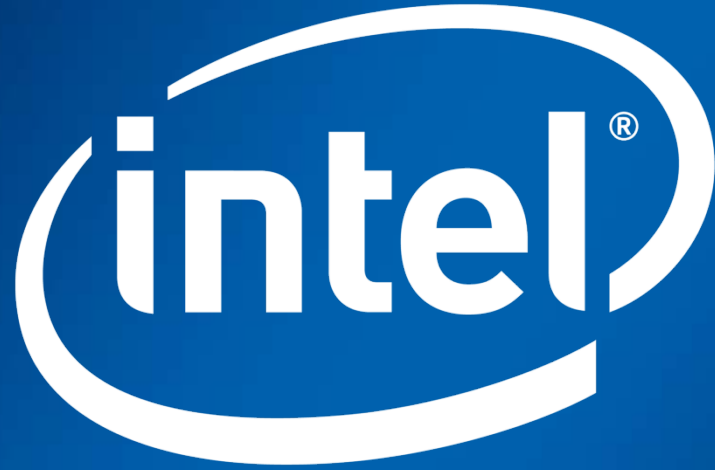
Follow the Recipe

Focus on Key Business Problems

Get Started

Data & Predictive Analytics are the
key to Unlocking the Potential





Smarter Decision Making Leverage Big Data to Gain New Actionable Insights

Anjul Bhambhri
VP, Big Data, Information Management, IBM



Where is big data coming from?

? TBs of data every day

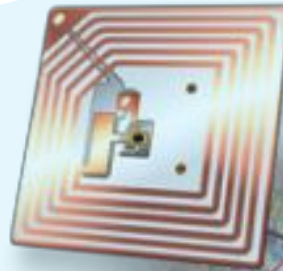
12+ TBs of tweet data every day



25+ TBs of log data every day



30 billion RFID tags today (1.3B in 2005)



76 million smart meters in 2009... 200M by 2014



4.6 billion camera phones world wide

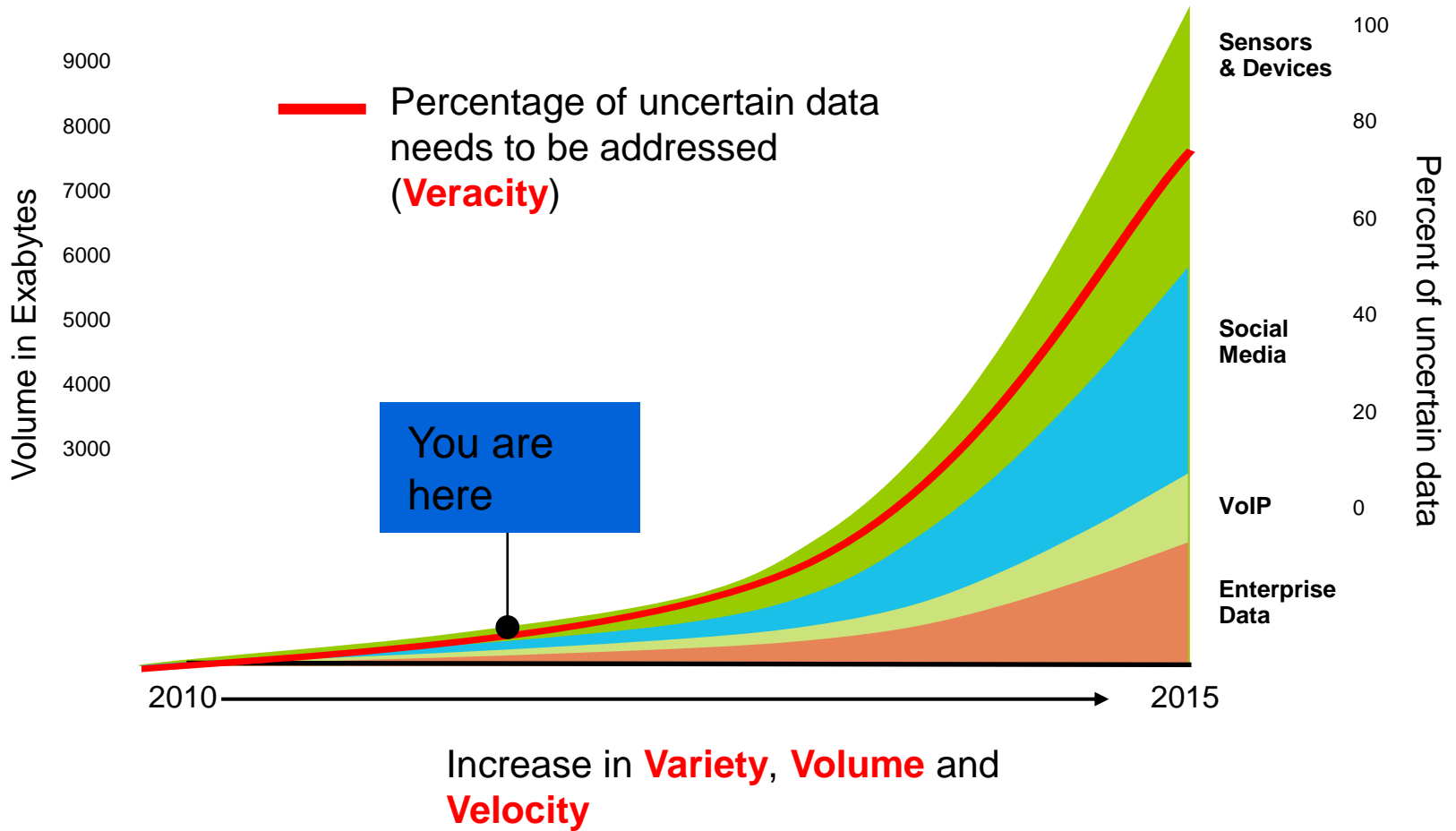


100s of millions of GPS enabled devices sold annually



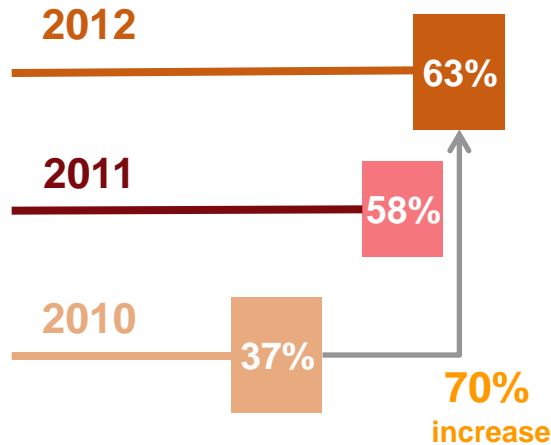
2+ billion people on the Web by end 2011

Big Data (4Vs): This is just the beginning



Benchmark of Global Big Data Activities (Oct 2012)

Realizing a competitive advantage



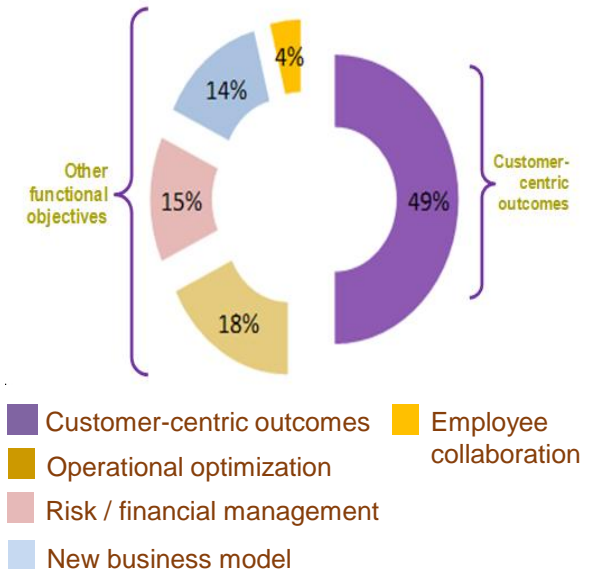
Nearly two out of three respondents reports realizing a **competitive advantage** from information and analytics

Big data activities



Three out of four organizations have **big data activities** underway; and one in four are either in **pilot or production**

Big data objectives

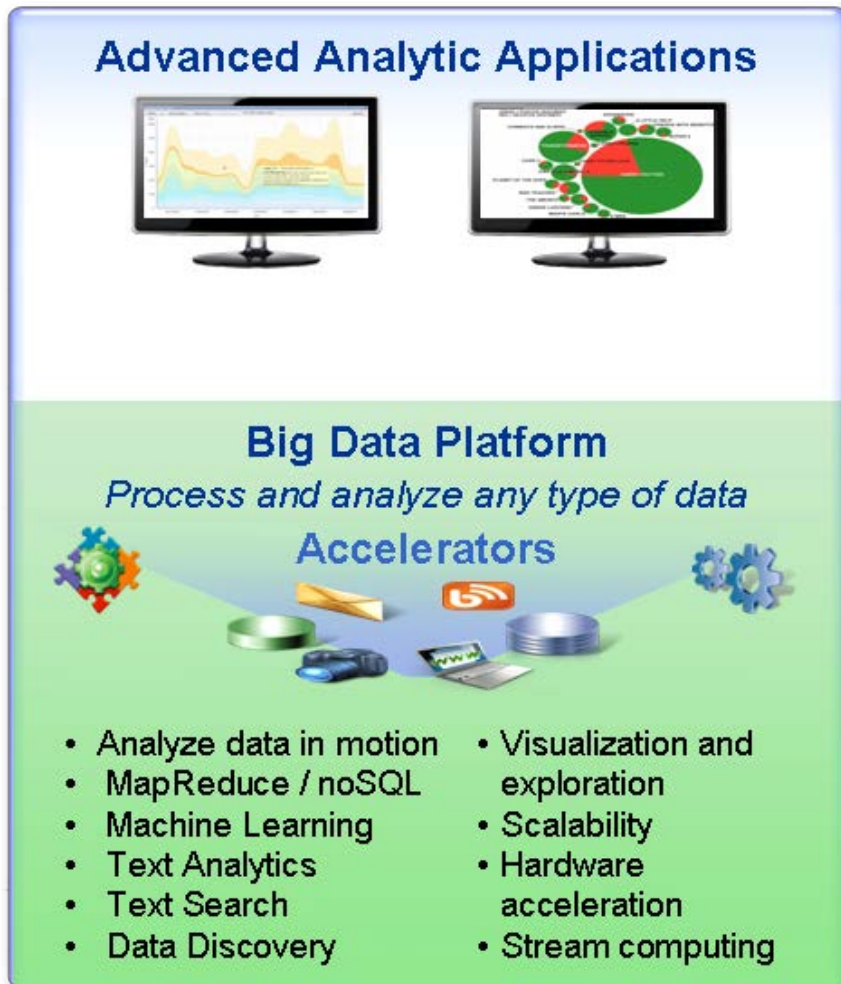


Improving **customer experience** by better understanding behaviors drives almost half of all active big data efforts followed by **Operational Optimization**

www.ibm.com/2012bigdatastudy

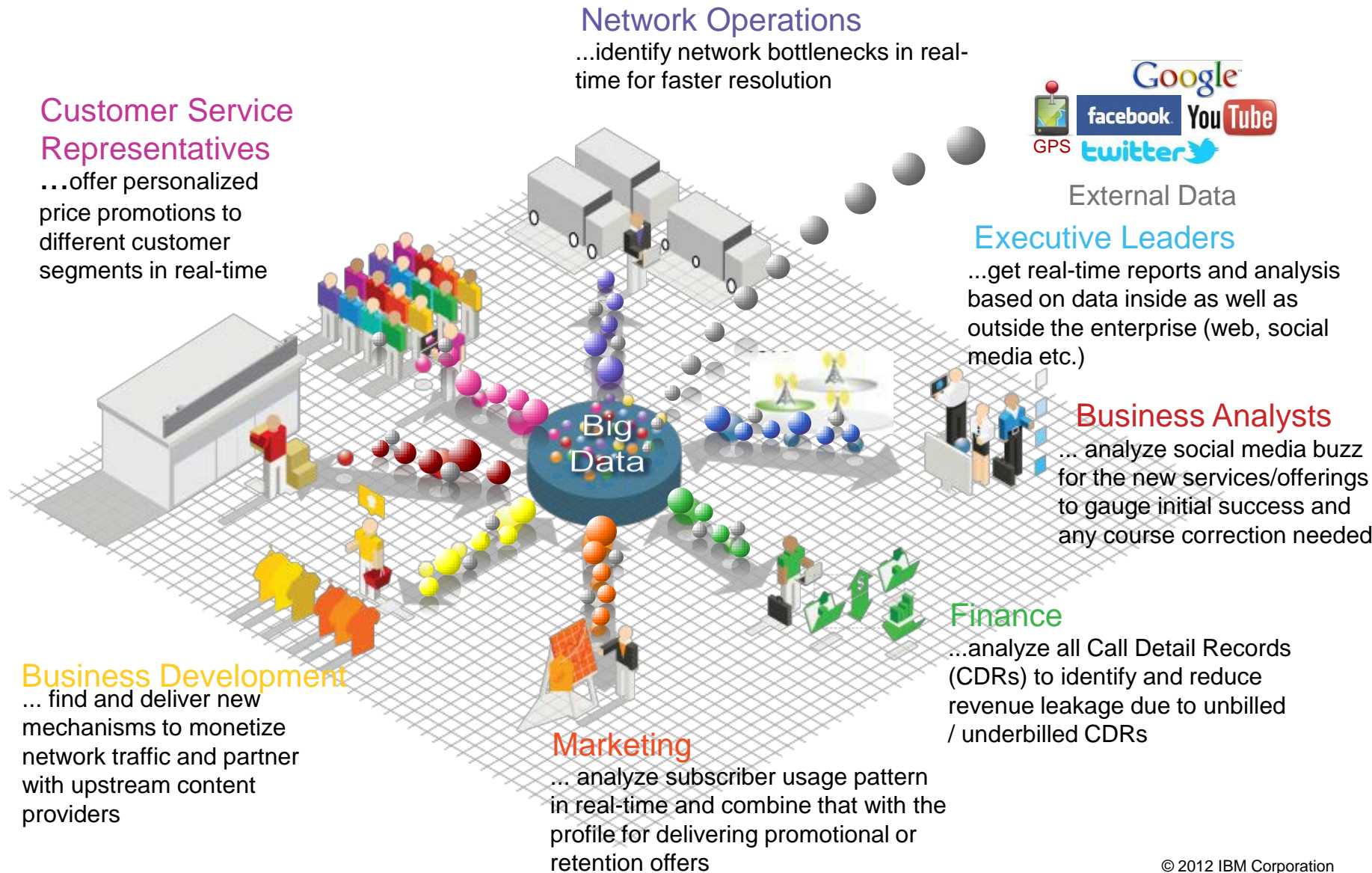
IBM Institute for Business Value and the University of Oxford Saïd Business School


More Mission-Critical Apps Ride on Big Data Platforms



- Integrate and manage the full **variety**, **velocity** and **volume** of data
- Apply **advanced analytics** to information in its **native** form
- Visualize all available data for **ad-hoc analysis and discovery**
- Development environment for **building new analytic applications**
- Integration and deploy applications with enterprise grade **availability**, **manageability**, **security**, and **performance**

The new era of analytics delivers value across the enterprise





Vestas optimizes capital investments based on **2.5 Petabytes** of information.

- Model the weather to optimize placement of turbines, maximizing power generation and longevity.
- Reduce time required to identify placement of turbine from weeks to hours.
- Incorporate 2.5 PB of structured and semi-structured information flows. Data volume expected to grow to 6 PB.

Vestas



Cisco turns to IBM big data for intelligent infrastructure management

- Optimize building energy consumption with centralized monitoring
- Automate preventive and corrective maintenance

Capabilities Utilized:

- Streaming Analytics
- Hadoop System
- Business Intelligence

Applications:

- Log Analytics
- Energy Bill Forecasting
- Energy consumption optimization
- Detection of anomalous usage
- Presence-aware energy mgt.
- Policy enforcement





Dublin City Centre Increases Bus Transportation Performance

Capabilities Utilized:

Stream Computing

- Public transportation awareness solution improves on-time performance and provides real-time bus arrival info to riders
- Continuously analyzes bus location data to infer traffic conditions and predict arrivals
- Collects, processes, and visualizes location data of all bus vehicles
- Automatically generates transportation routes and stop locations

Results:

- Monitoring 600 buses across 150 routes
- Analyzing 50 bus locations per second
- Anticipated to Increase bus ridership



Asian telco reduces billing costs and improves customer satisfaction.

Capabilities:

Stream Computing
Analytic Accelerators

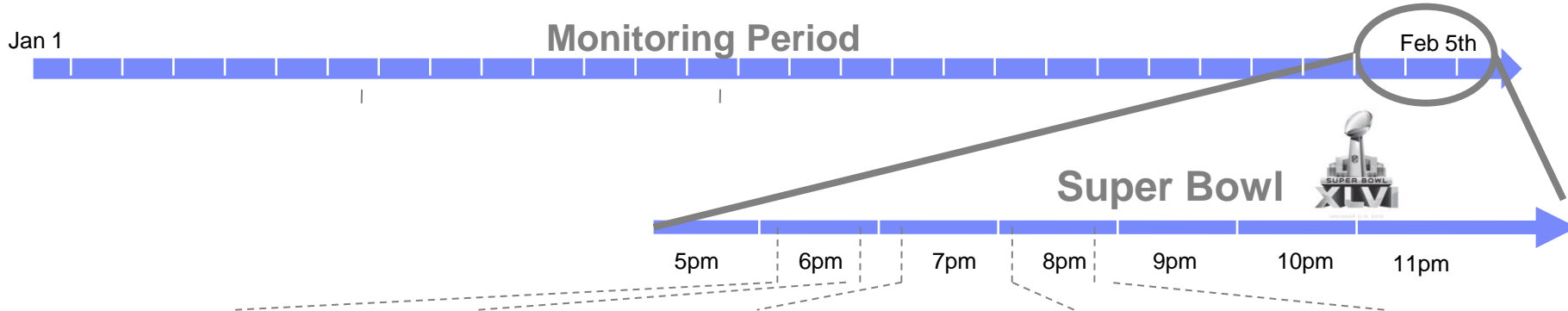
Real-time mediation and analysis of
6B CDRs per day

Data processing time reduced from
12 hrs to 1 sec

Hardware cost reduced to 1/8th

Proactively address issues
(e.g. dropped calls) impacting customer satisfaction.

To-the-minute and historical product insight

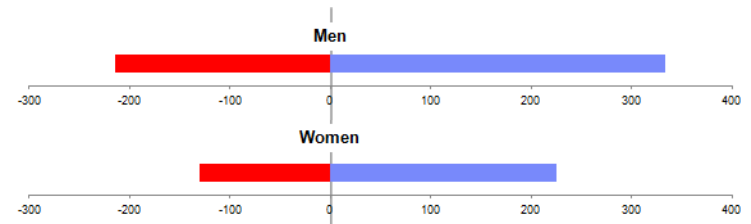
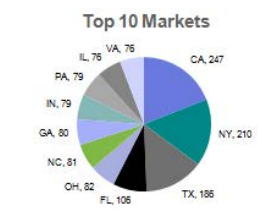
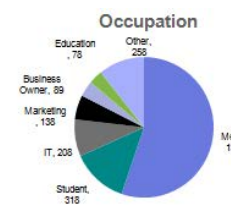
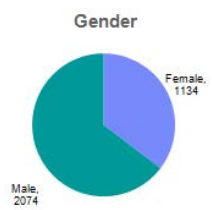


Data Set

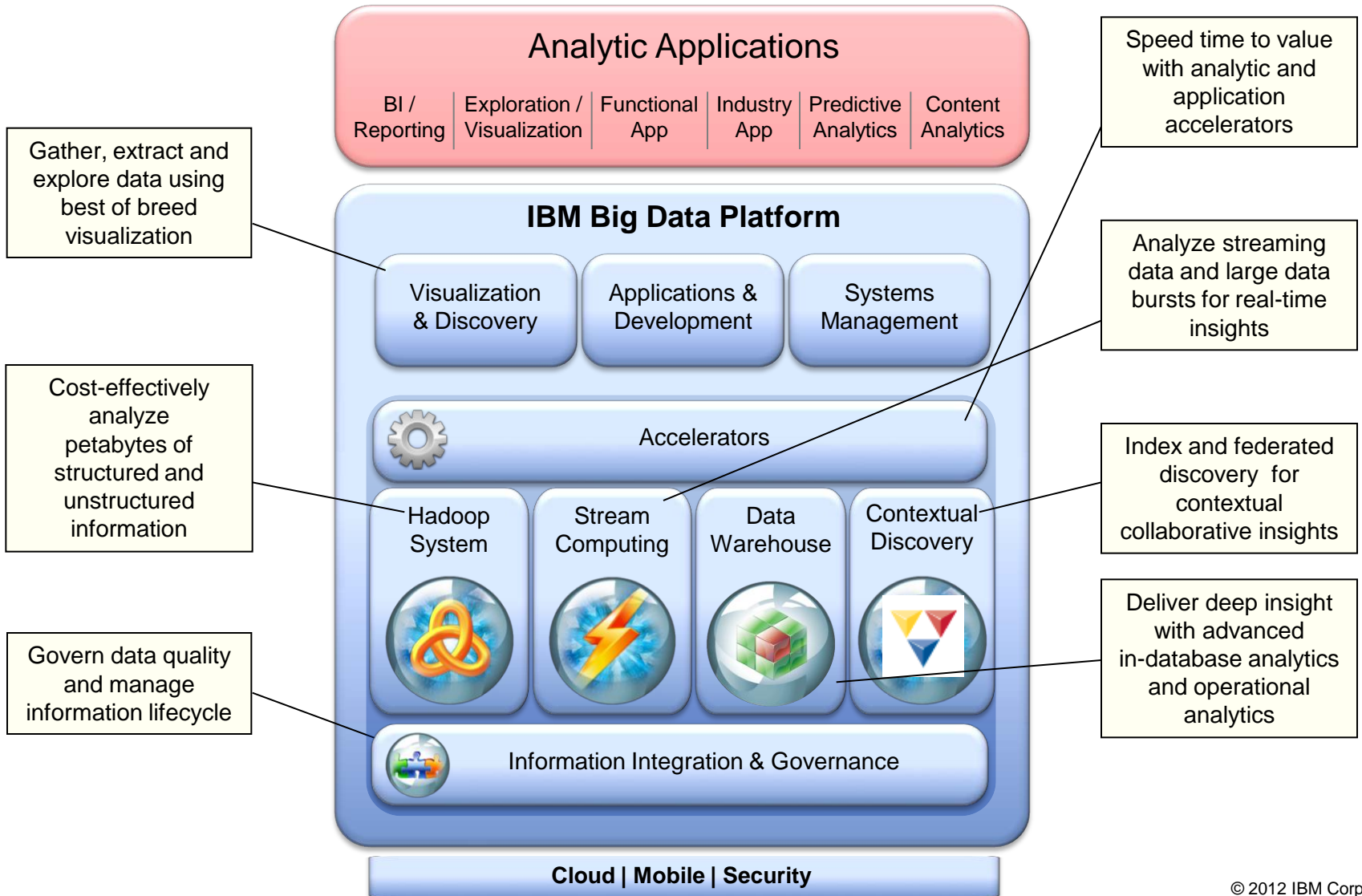
- 1.1B tweets
- 5.7M blog and forum posts
- 3.5M relevant messages
- 97K referencing Product_A
- 18K referencing Product_B

Information extracted

- Buzz and sentiment
- Gender, Location and Occupation
- Fans
- Intent to in purchase
- Specific attributes of products

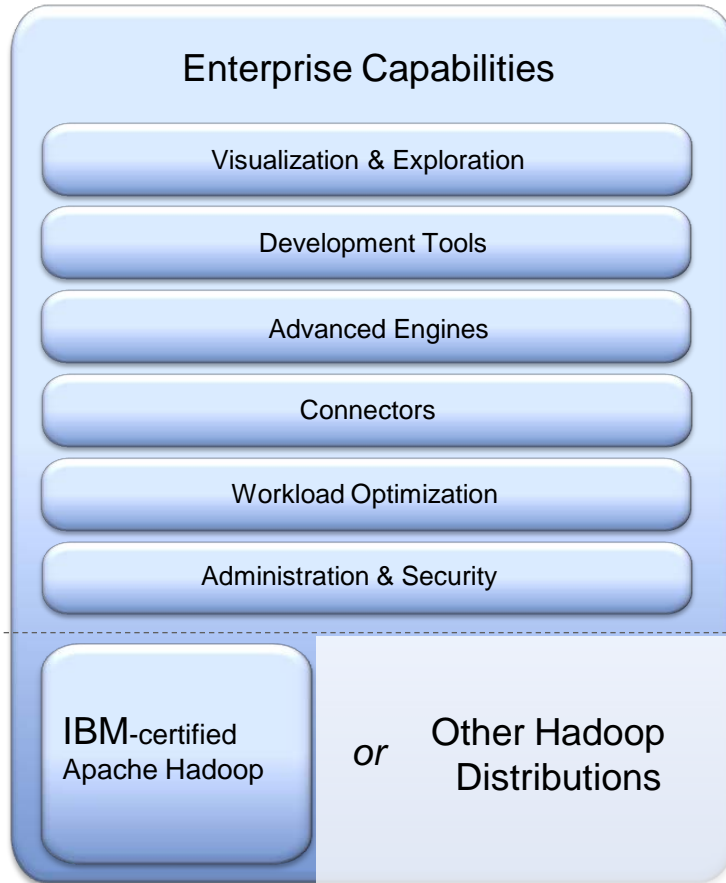


Big Data Platform and Application Framework





Big Data Platform – Internet Scale Analytics



Platform Capabilities

- Built-in analytics
 - Text analytics engine, annotators, Eclipse tooling
 - Interface to project R (statistical platform)
- Deep integration with enterprise software stack
- Analytical tool for analysts
- Ready-made business process accelerators
- Integrated installation of supported open source and other components
- Web Console for admin and application access
- Platform enrichment: additional security, performance features, . . .
- World-class support
- Full open source compatibility

Business benefits

- Quicker time-to-value due to IBM technology and support
- Reduced operational risk
- Enhanced business knowledge with flexible analytical platform
- Leverages and complements existing software





Massively Scalable Stream Analytics

Linear Scalability

- Clustered deployments – unlimited scalability

Automated Deployment

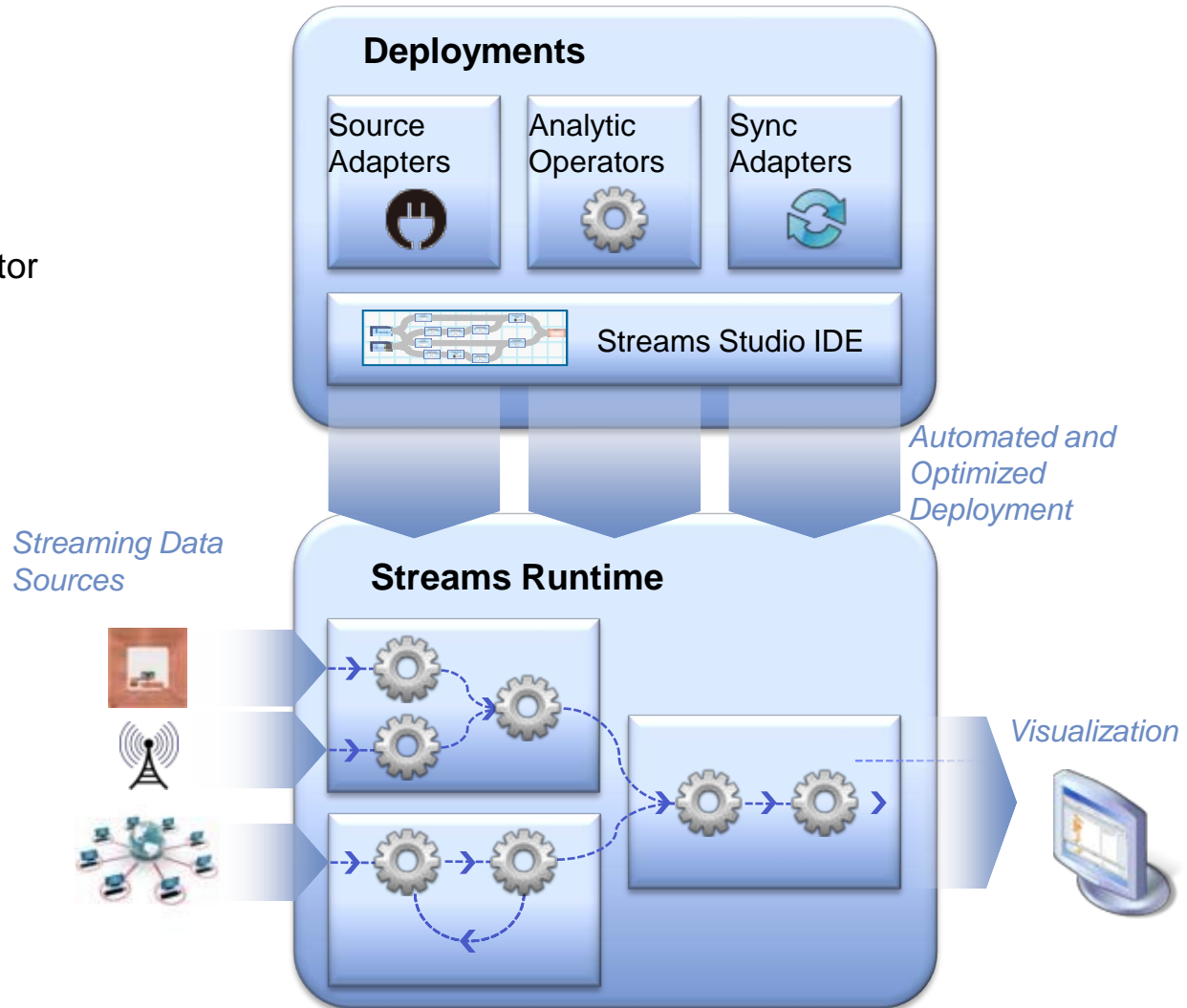
- Automatically optimize operator deployment across clusters

Performance Optimization

- JVM Sharing – minimize memory use
- Fuse operators on same cluster
- Telco client – 25 Million messages per second

Analytics on Streaming Data

- Analytic accelerators for a variety of data types
- Optimized for real-time performance





Deep Analytics Appliance – Revolutionizing Analytics

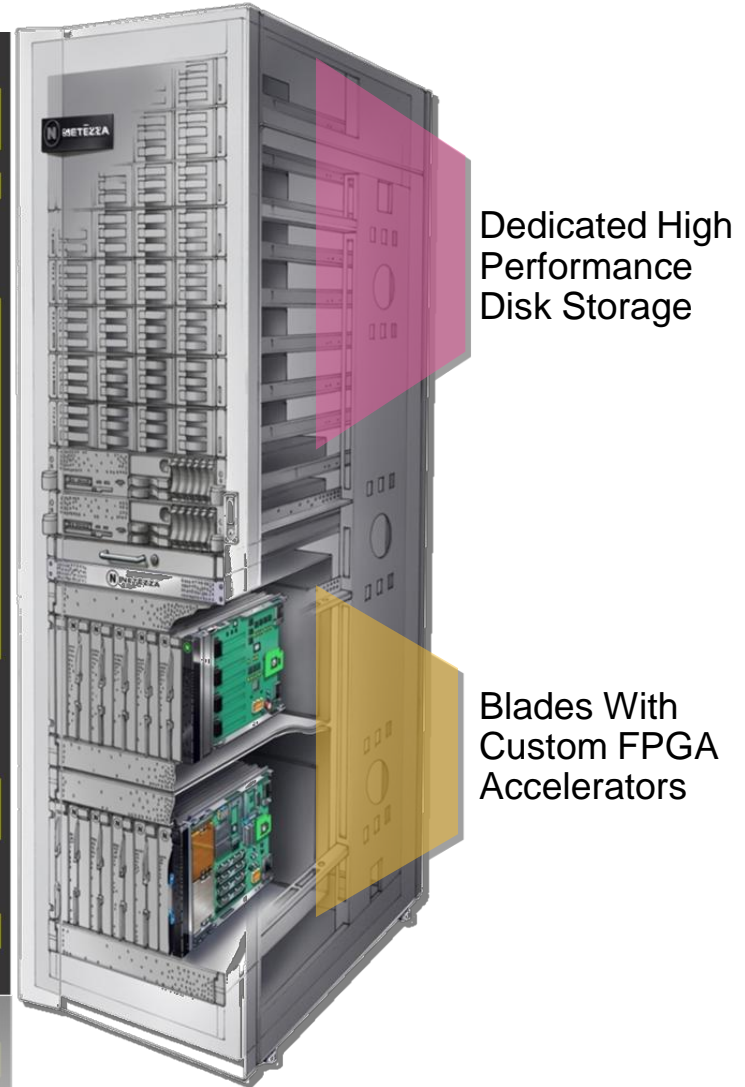
Purpose-built analytics appliance

Speed: 10-100x faster than traditional systems

Simplicity: Minimal administration and tuning

Scalability: Peta-scale user data capacity

Smart: High-performance advanced analytics



New classes of applications for end-users



Real Time Analytics

Internet Scale Analytics

In-Database Analytics

Enterprise Data Connectors

Federated Discovery

Navigation and Visualization

Application Framework

OPTIMUM INVESTMENTS

Logged in as Frank Gelato | Help

Followed High Net Wealth Clients

- Isabella Jones
- Thomas Jackson
- Theresa Mayer
- Michael Kleinfelder

Tracked Products

- 529 Plan
- 401K
- Money Market IRA
- Fixed Income & Bonds

Financial Blogs

CNBC: Warren Buffett: 'Disruptive' Debt Limit Debates Are 'Waste of...
Everything Warren Buffet

CNBC: CNBC Transcript: Warren Buffett on Russian Roulette, Tax...
Everything Warren Buffet

BLOOMBERG: Munger Treats 'Hard-Core Addicts' as Wesco Stock Exits...
Everything Warren Buffet

Investment News

Texas Gains New Billion Dollar Bank
Business Wire - 35 minutes ago

Ally Financial Reports Preliminary Second Quarter 2011 Financial Results
PR Newswire - 54 minutes ago

Pinnacle West Reports Second-Quarter Results
TheStreet.com - 55 minutes ago

Action Needed

Sentiment	Customer	Type	Format	Time	Product
Negative	Isabella Jones	Support	Tweet	05.29.12 10:30 am EST	Mutual funds
Neutral	Thomas Jackson	Support	Support ticket - email	05.29.12 10:25 am EST	401K
Positive	Isabella Jones	Sales	Tweet	05.29.12 10:05 am EST	529
Negative	Theresa Mayer	Satisfaction	Blog	05.29.12 09:30 am EST	401K

Activity Feed

Isabella Jones - @tzyJones
Wow is the stock market really this bad, my monthly mutual fund statement looks terrible! Time to call my financial advisor.
Twitter - Minutes ago

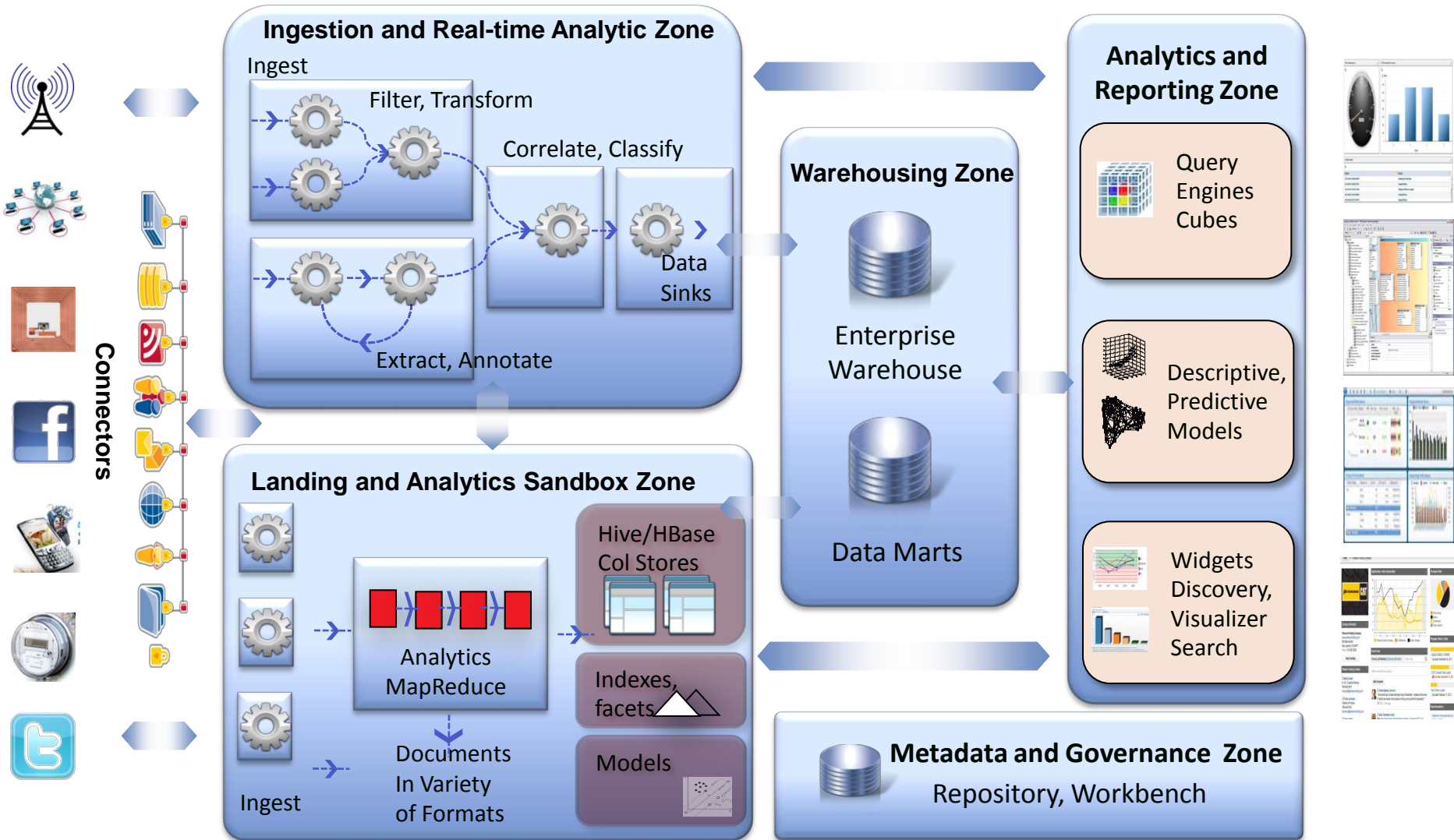
Theresa Mayer - tmayer@gmail.com
Great advice on retirement planning! Thinking about increasing my 401K contribution but are there other retirement plans I should be looking at given I hope to retire in 10-15 yrs?
<http://mextavenue.org/blog/why-women-need-embrace-retirement-planning>
Blog - Minutes ago

Thomas Jackson
How can I change my 401K contribution using your online system?
Remedy - Minutes ago

Sentiment

Sentiment By Age

Emerging Pattern of Big Data Implementation



IBM's Big Data Business Partner Ecosystem

180+
Big Data Business Partners

Thank You!



Analytics on Big and Small Data

Tom Davenport

UC Berkeley

November 1, 2012

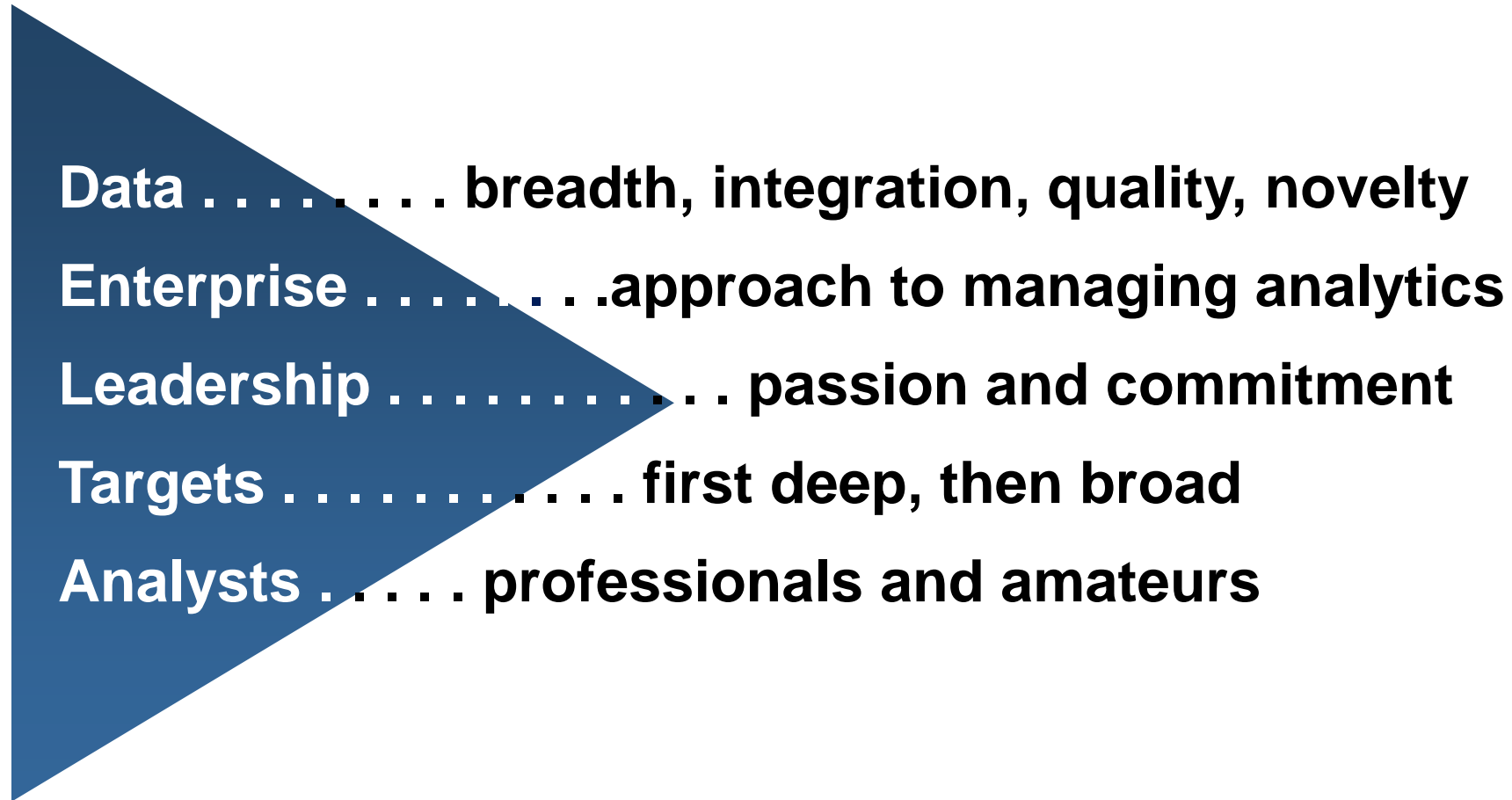


A Bright Idea – Analytics on Small and Big Data

It works for:

- Old companies (GE, P&G, Marriott, American Express)
- Middle-aged companies (Capital One, Google, Ebay, Netflix, etc.)
- New companies (Quid, Recorded Future, Kyruus, GNS Healthcare, and a host of Silicon Valley and Boston companies you don't know)
- Technology companies (SAP, HP, Teradata, EMC, IBM, etc.)

The Analytical DELTA (Small Data, but Relevant to Big)



The Rise of Big Data

What is it?

- Data that's too big (petabytes), too unstructured (not in rows and columns), or too diverse (mashups) to be stored and analyzed by conventional means (also relative)

Where does it come from?

- Internet/social media
- Genomic analysis
- Voice and video
- Sensors everywhere

What is to be done with it?

- Structure, classify, and count it
- Then analyze it (just as you would small data)



What's Different About Big Data?

The need for continuous flows of data, not stocks

- Stocks may be useful to develop models, but big data eventually requires a continuous process of analysis on moving data

Data scientists, not analysts

- IT “hacking” abilities in addition to the usual analyst attributes
- Scientific and exploration focus
- Closer to the product or process

New ways of deciding and acting on it

- It just keeps on coming, so have to establish ongoing processes to manage or decide on it



What's Different About Big Data? (cont.)

- Filtering, structuring, and classification tools – MapReduce, Hadoop, etc.
- Content analytics tools--NLP
- Data redundancy management
- Cloud analytics
- Machine learning
- Open source everything, including R (it's capable, it's free, and that's what everybody coming out of school wants to use)

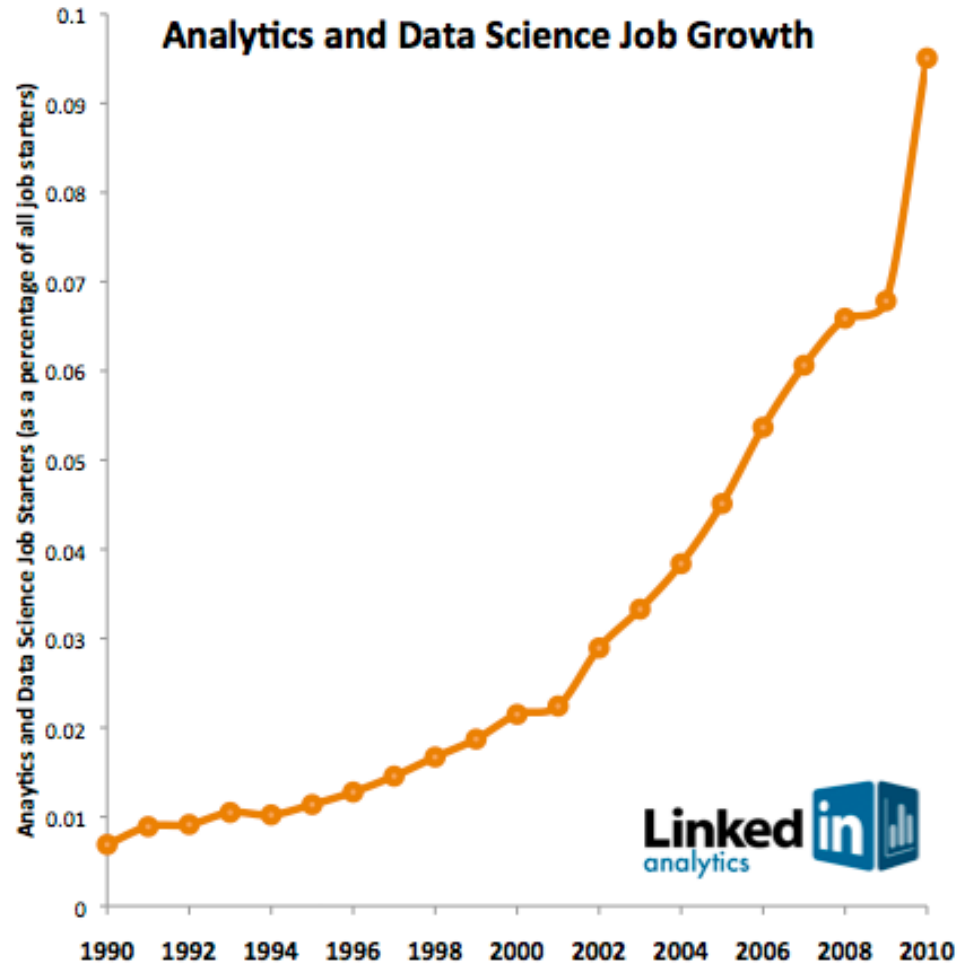
New technologies
to manage it



The Rise of the Data Scientist

Hybrids	<ul style="list-style-type: none">• Half analytical, with modeling, statistics, and experimentation skills• Half focused on data management – extraction, filtering, sampling, structuring• Lots of programming skills – Python, Ruby, Hadoop, Pig, Hive
Scientific	<ul style="list-style-type: none">• Experimental physicists• Computational biologists• Statisticians with dirty hands• Ecologists, anthropologists, psychologists, etc.
Impatient	<ul style="list-style-type: none">• Try something and iterate• Don't wait for a data person to get your data• “We're a pain in the ass”• Job tenure is short
Ground-breaking	<ul style="list-style-type: none">• “Nobody's ever done this before”• “If we wanted to deal with structured data, we'd be on Wall Street”• “Being a consultant is the dead zone – too hard to get things implemented”• “The output should be a product or a demo – not a report”

The Rise of Data Scientists and Analysts



Courtesy LinkedIn Corp.

Some Use Cases for Big Data

- Social media analytics – “People You May Know” at LinkedIn
- Voice analytics – Call center triage
- Text analytics – Voice of customer, sentiment analysis, warranty analysis
- Video analytics – Intelligence, policing, retail applications
- Genome data – what genetic profiles are associated with certain cancers?



Big Data at eBay



“Analytics platform,” with heavy focus on testing

40 petabytes of storage in Teradata EDW, with hundreds of “virtual data marts”—and much more in Hadoop clusters

50 new terabytes per day

Platform includes:

Hadoop and MapReduce for image similarity networks

R for statistical analysis

User-developed apps described in “Data Hub”

Big Data at GE



New \$2B corporate center for software and analytics

Hiring 400 data scientists—200 already on board

Includes financial and marketing applications, but with special focus on industrial uses of big data

When will this gas turbine need maintenance?

How can we optimize the performance of a locomotive?

What is the best way to make decisions about energy finance?

Big Data at EMC

EMC²

where information lives[®]

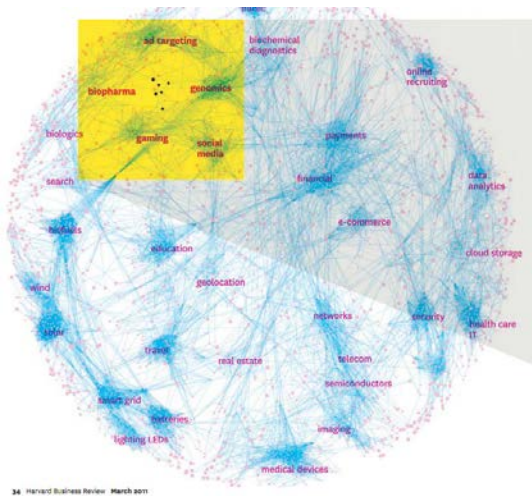
Bought Greenplum, a big data appliance vendor, in 2010

Realized that data scientist availability would be gating factor in big data capabilities

Developed a big data analytics course for employee and customer consumption

Using early graduates to examine probability that product innovation ideas will succeed

Big Data at Quid



Small startup, but working with big organizations

Works to map the structure of technology ideas, funding, and product breakthroughs using primarily Internet data

e.g., opportunities at intersection of biopharma, social media, gaming, and ad targeting

Works with major IT vendors and governments; beginning to work with strategy consulting firms

Big Data and Small Data Analytics – How Do They Compare?

Focus

- Big data is often external, small data often internal
- Big data is often part of a product or service, small data is used to manage

Relationships

- Big data and small data analysts require good relationships
- But relationships are different: product managers and customers for big data analysts; internal managers for small data analysts

Technologies

- Big data requires data management (Hadoop, Pig, Hive, Python)
 - Analysis in visual (Tableau, Spotfire), open-source (R) tools
- Small data requires less data management – SQL is sufficient
 - Analysis in BI (BO, Cognos, Qlikview) or statistical (SAS, SPSS) tools

Flies in the Big Data Ointment

- Labor intensive, and labor expensive
- “Not much abstraction going on here”
- Big data = small math
 - Step 1 is just getting the data counted
 - Step 2 is providing nice visualizations of it
 - Step 3 will be doing real analytics on it
- Lots of interfaces and integration necessary
- Technologies and people will be easier if you can wait



Elements in Common: Leadership



Gary Loveman at Caesars

- “Do we think, or do we know?”
- “Three ways to get fired”

Jeff Bezos at Amazon

- “We never throw away data”

Reid Hoffman at LinkedIn

- “Web 3.0 is about data”

“Our CEO is a real data dog”

Sara Lee
executive

Elements in Common: Getting Much Faster!

In-memory analytics

- HANA from SAP

High-performance analytics

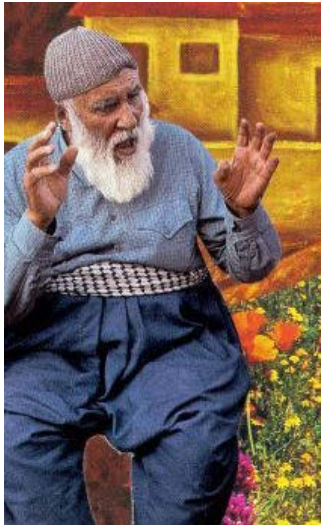
- From 19 hours to 19 minutes to optimize prices for all categories at Macy's

In-database processing

- Propensity scoring for all customers in seconds, not weeks, at Cabela's



Elements in Common: New Analyst Skills



“Tell a story with data”



“Be courageous”



“Build a rapid prototype”

Turn It On!

- Make sure your leaders are on board
- Figure out your targets
- Build, buy, or borrow the people you need
- Assess your technology
- Improve some decisions or some products and services!





BIG DATA AT EBAY

Hugh E. Williams

Vice President, Experience, Search, and Platforms

eBay Marketplaces

hugh.williams@ebay.com

@hughewilliams

Welcome to today's online marketplace...

...the market that brings buyers and sellers together in an honest and open environment...

Welcome to eBay's AuctionWeb.

Welcome to our community. I'm glad you found us. AuctionWeb is dedicated to bringing together buyers and sellers in an honest and open marketplace. Here, thanks to our [auction format](#), merchandise will always fetch its market value. And there are plenty of great deals to be found!

[Take a look at the listings.](#) There are always several hundred auctions underway, so you're bound to find something interesting.

If you don't find what you like, take a look at our **Personal Shopper**. It can help you search all the listings. Or, it can keep an eye on new items as they are posted and let you know when something you want appears. If you want to let everyone know what you want, post something on our [wanted page](#).

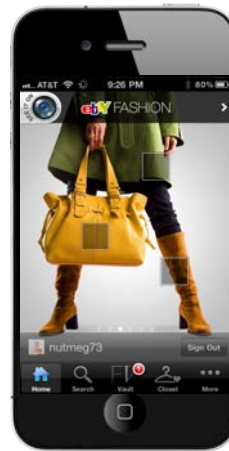
If you have something to **sell**, start your auction instantly.

Join our community. Become a registered user. Registered users receive [additional benefits](#) such as daily updates and the right to participate in our user feedback forum and the bulletin board.

every
49 minutes
a Ford Mustang
is sold



every
5 seconds
a cell phone
is sold



every
6 seconds
a pair of shoes
is sold





\$68.6 billion

in merchandise sold in 2011



108+ million

Active buyers and sellers
worldwide

250+ million

Queries every day to the eBay
search engine

350+ million

Live global listings



20+ petabytes

Of data in our Hadoop and Teradata clusters

2 billion

Page views each day

75 billion

Database calls each day



EVEN OUR QUALITATIVE DATA IS BIG

- Inline and other surveys to capture ratings and verbatims
- Touch point NPS to capture promoter and detractor effects
- Customer service data
- User experience research studies to watch and learn from customers
- Meeting, listening, and recording customer experiences

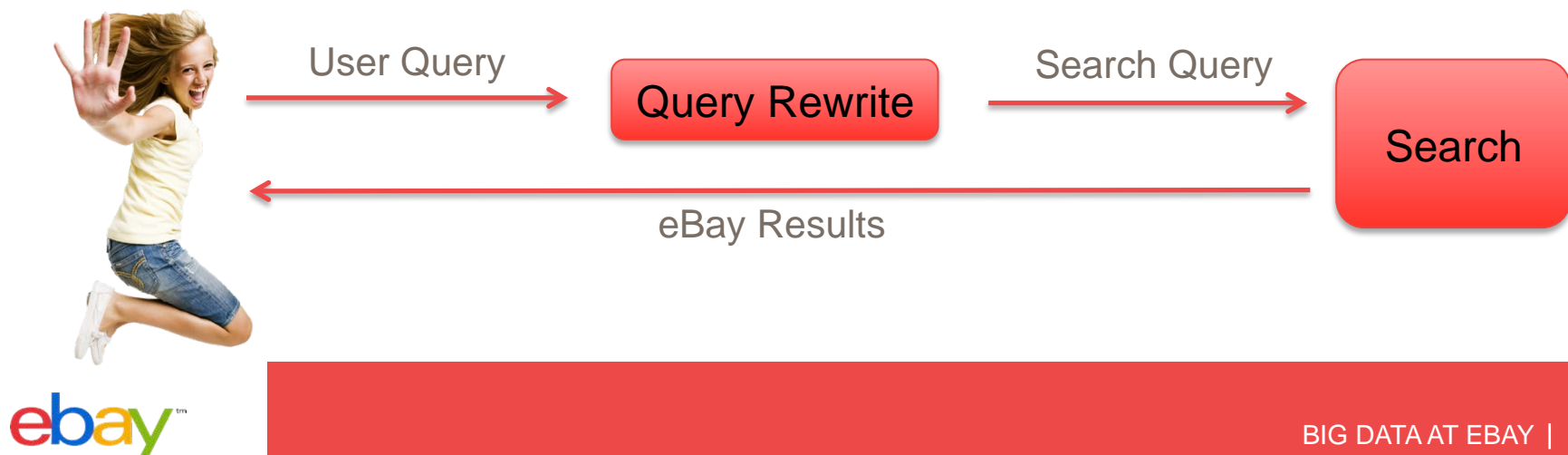
WHY IS BIG DATA TRANSFORMATIONAL?

BIG DATA IS TRANSFORMATIONAL

- Big data informs on:
 - *Patterns*
 - Anomalies and outliers
 - Generalizations
 - *Predictions*
 - *Relative performance*
 - An holistic customer picture
- Vast array of applications at eBay:
 - *Product development, A/B testing, system performance, fraud and risk detection, purchase prediction, customer support, buyer demand, seller intelligence, financial performance, ...*

PATTERNS: QUERY REWRITES

- In 2010, our search engine was very literal: it matched exactly what you typed
- We're on a journey to make it more intuitive, so it does a great job of understanding user intent and finding all of the relevant results
- Idea: Mine our extreme data, look for patterns, and use these to map words in user queries to **synonyms** and **structured data** associated with items for sale at eBay



PATTERNS: QUERY REWRITES ...



HOW DO BUYERS PURCHASE THE PILZLAMPE?

- It turns out, they try one (or more) of a few things:
 - Type **pilzlampe**, and purchase
 - Type **pilzlampe**, ... , **pilz lampe**, and purchase
 - Type **pilzlampe**, ... , **pilzlampen**, and purchase
 - Type **pilz lampen**, ... , **pilzlampe**, and purchase
 - ...

PATTERNS: QUERY REWRITES ...

- From our big data mining:
 - We automatically discover that *pilz lampe* and *pilzlampe* are the same
 - We also discover that *pilz* and *pilze* are the same, and *lampe* and *lampen* are the same
- From these patterns, we rewrite the user's query *pilzlampe* as:
pilzlampe OR "pilz lampe" OR "pilz lampen" OR pilzlampen OR "pilze lampe" OR pilzelampe OR "pilze lampen" OR pilzelampen

ARE QUERY REWRITES EASY?

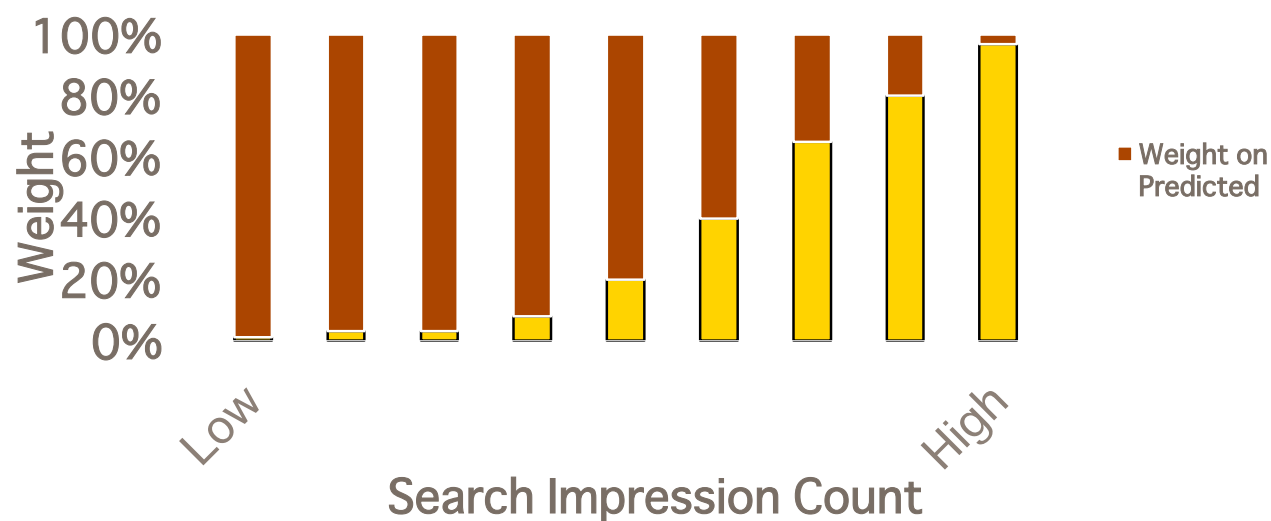
- Nothing is easy at scale
 - Incorrect strong signals:
 - CMU is not Central Michigan University
 - Colliding signals
 - Mariners is not the same as Marines
 - Context matters
 - Correcting Seattle Marines to Seattle Mariners is (generally) right
 - Jacksonville Jaguars is not Jacksonville in the Motors category

PREDICTION: ITEM QUALITY IN BEST MATCH

- Our goal in search is to show the best matching items for each user's query
- We use tens of *ranking factors* to rank
 - Factors are drawn from item text, item images, seller information, buyer information, and *behavioral big data*
- Factors are combined into a *ranking function* using a machine-learned ranking model

PREDICTION: ITEM QUALITY IN BEST MATCH

- One ranking factor we compute is *item quality*
 - The likelihood that an item will sell, and its likely selling price
 - Predictions are based on our vast data sets of item and seller performance
- At listing time, we compute *predicted* item quality
- As users interact with the item, we observe and learn its *true* quality



EVERY BIG DATA IDEA IS CHALLENGING TO BUILD AT SCALE

- We often have more than six million item updates per hour
 - Log events flow from individual machines when the events occur
 - A *listener cluster* accumulates events
 - Events are sorted by their unique identifier
 - A queue of events is created by likely impact of the changes
 - We process queues and update the Best Match factors
- Rinse, repeat frequently

TEST VS CONTROL EXPERIMENTATION

- Divide customers into populations
- One population is the *control*
- One or more populations are the *tests*
- Collect data from each population
- Compute metrics from the data, including confidence intervals
- Understand the results
- Make decisions

A FLEXIBLE APPROACH

ONE SIZE DOESN'T FIT ALL

- There isn't one solution for driving an organization with big data:
 - **Hadoop** is for:
 - Engineers, batch (asynchronous), map reduce (divide and conquer), unstructured, flexible problems
 - **HBase** is for:
 - Engineers, real-time, large data blob, unstructured, key lookup, flexible problems
 - **Teradata** (or another data warehousing solution) is for:
 - Analysts, real-time or batch, structured, flexible problems
 - **Cassandra** (or **MongoDB** or ...) is for:
 - Engineers, real-time, smaller data blob, unstructured, key lookup, flexible problems
 - Some problems warrant specialized solutions

THE BIG DATA OUTLOOK

SHOPPING IS CHANGING

We are at an inflection point!



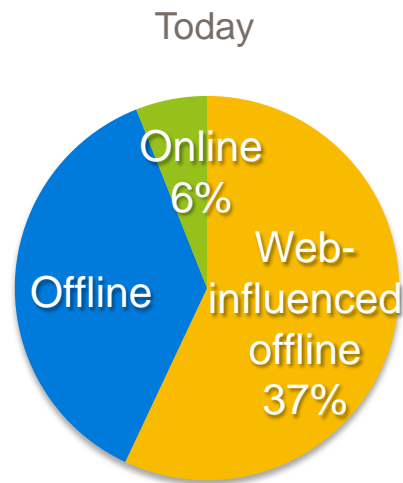
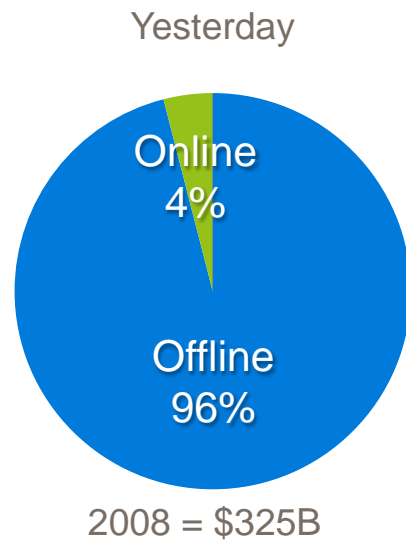
TRADITIONAL BOUNDARIES HAVE BLURRED:

Technology is fundamentally changing the way people shop



IT'S JUST COMMERCE

There's no longer online and offline

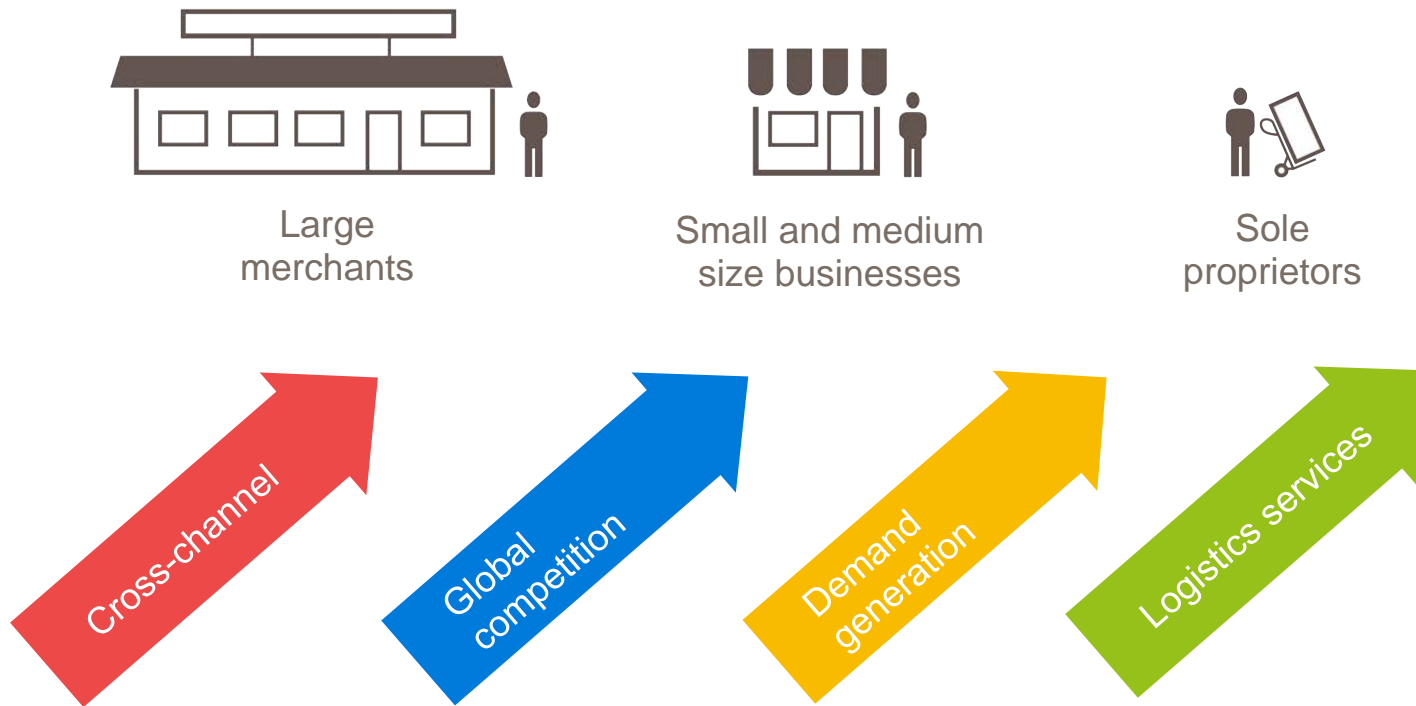


Source: Forrester, Euromonitor and Economist Intelligence Unit

Source: Forrester

Source: Economist Intelligence Unit

THE PLAYING FIELD IS BEING LEVELED



Cost of entry is lower than ever, consumers are in the driver's seat

THE BIG DATA OUTLOOK

- Vastly more data, from:
 - New customers
 - New applications
 - Noisier applications
 - A widening landscape
 - Engineers and analysts creating derivatives
- A new set of challenges:
 - Curation
 - Cleaning up
 - Documentation
 - Managing the user population
 - Stability and scalability of big data systems



Psst... We're hiring. Email: hugh.williams@ebay.com





Geoffrey Moore

AUTHOR, SPEAKER, ADVISOR

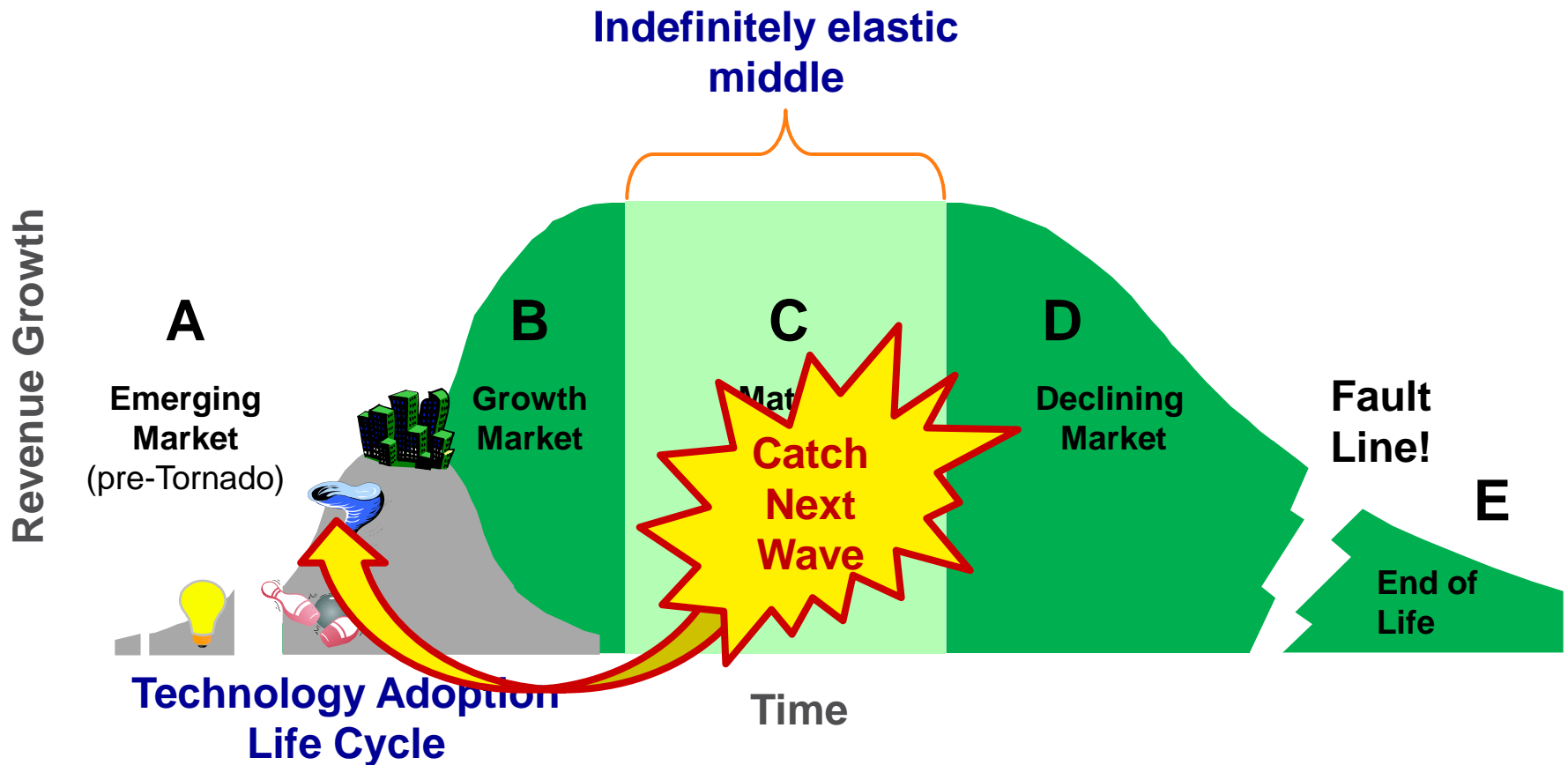
Big Data

Where Are We?

Fisher CIO Leadership Hadoop Conference
November 1, 2012

Category Maturity Life Cycle

Where is Big Data Today?

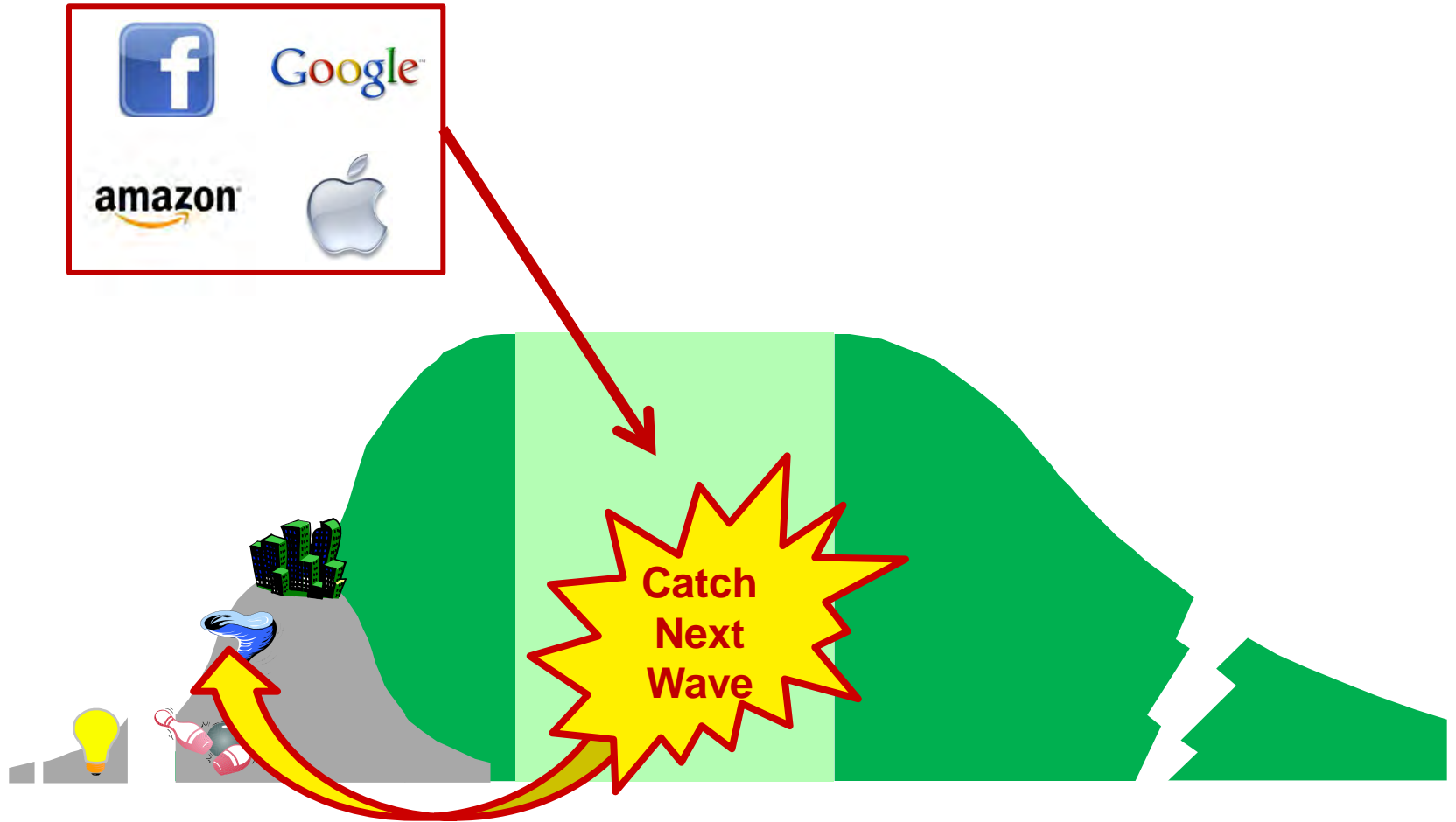


It depends!

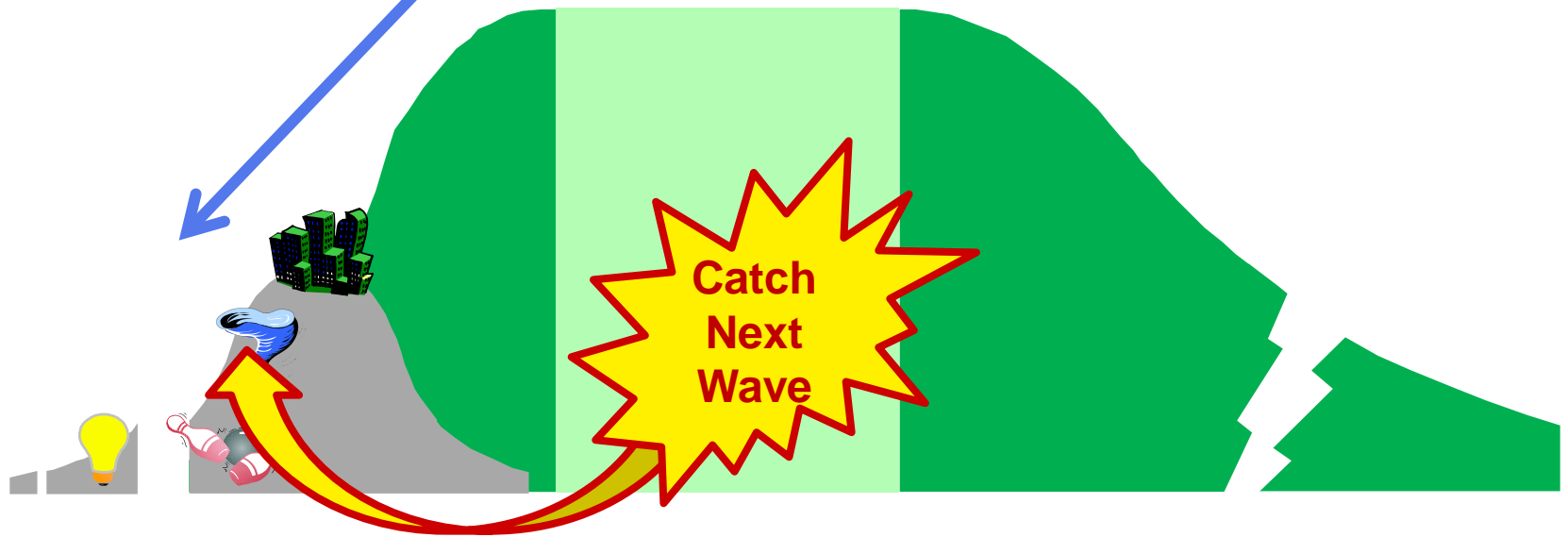
Digital Disruptors



Digital Disrupters



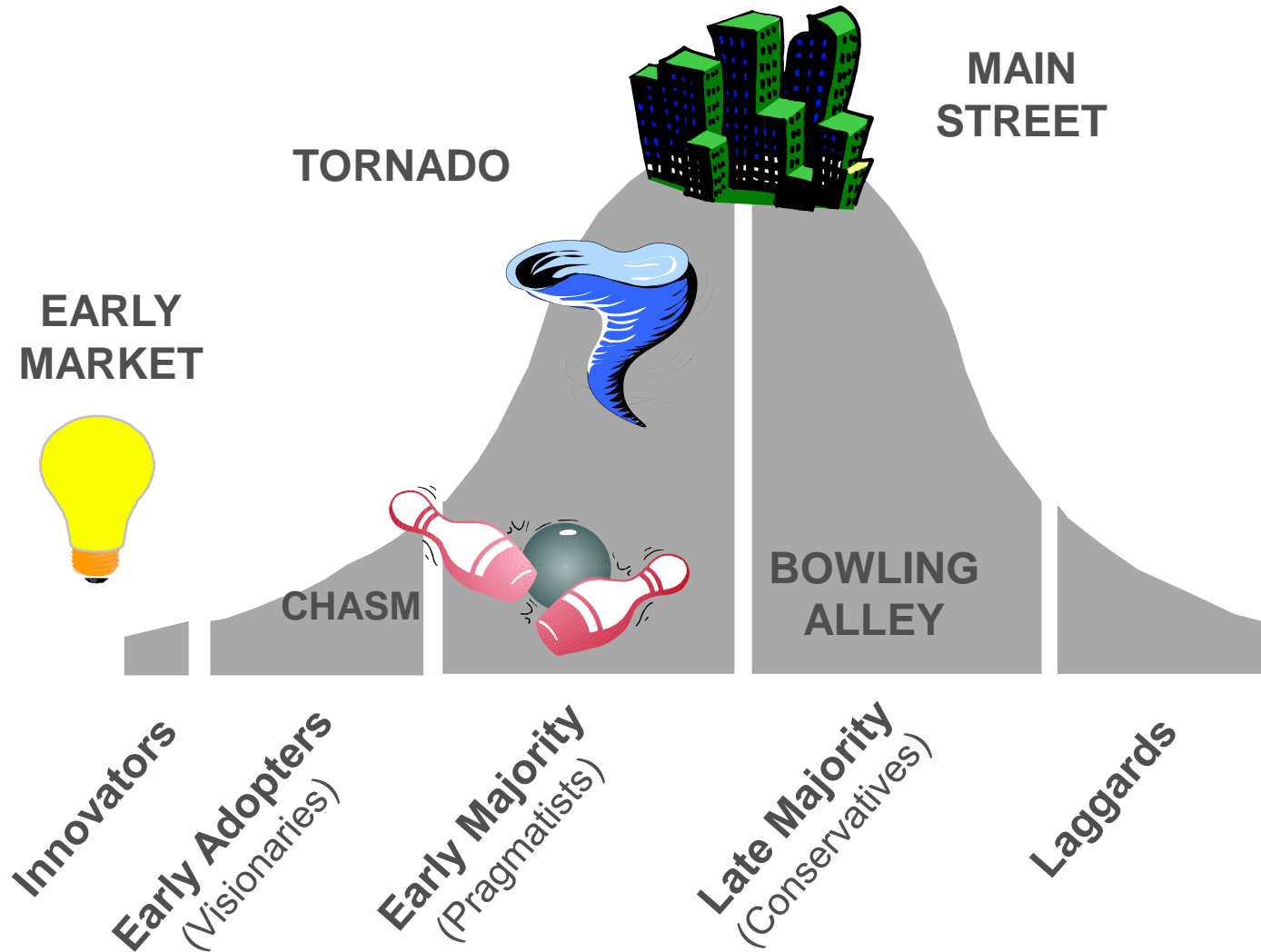
Digital Disrupters vs. Digital Disruptees



Your mileage may vary!

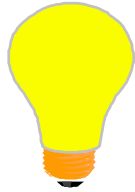
Technology Disruptions

How Technology Enters the Mainstream

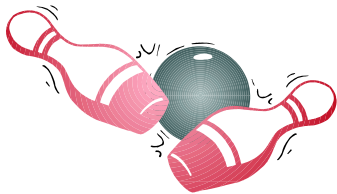


Impact of Category State on Company Power

Business Model Adapts to Life Cycle Dynamics



- Project orientation
- **Sell, Design, Build**
- **Focus on *performance***



- Solution orientation
- **Design, Sell, Build**
- **Focus on *performance/price***



- Product orientation
- **Design, Build, Sell**
- **Focus on *price/performance***

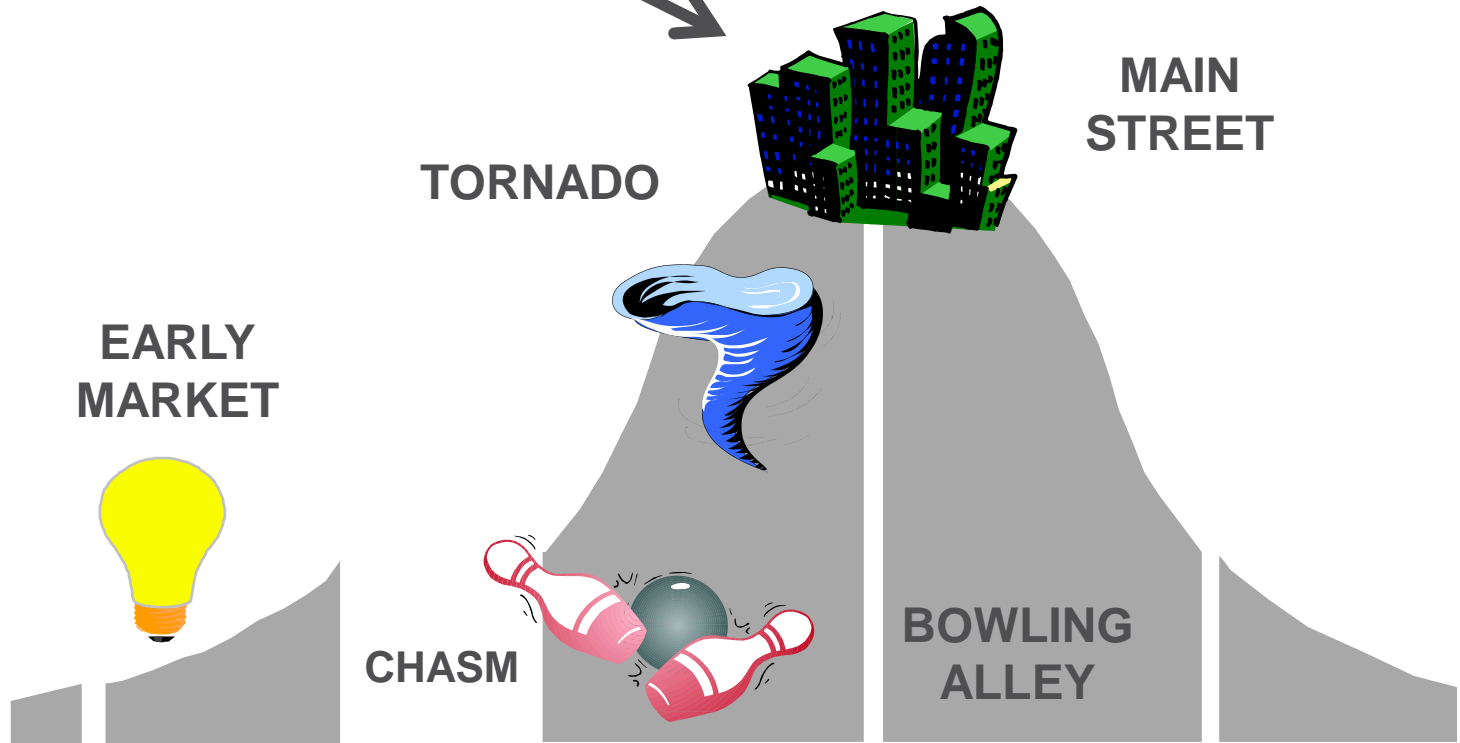


- Systems orientation
- **Build, Sell, Design**
- **Focus on *price/TCO***

Genomics



Genomics

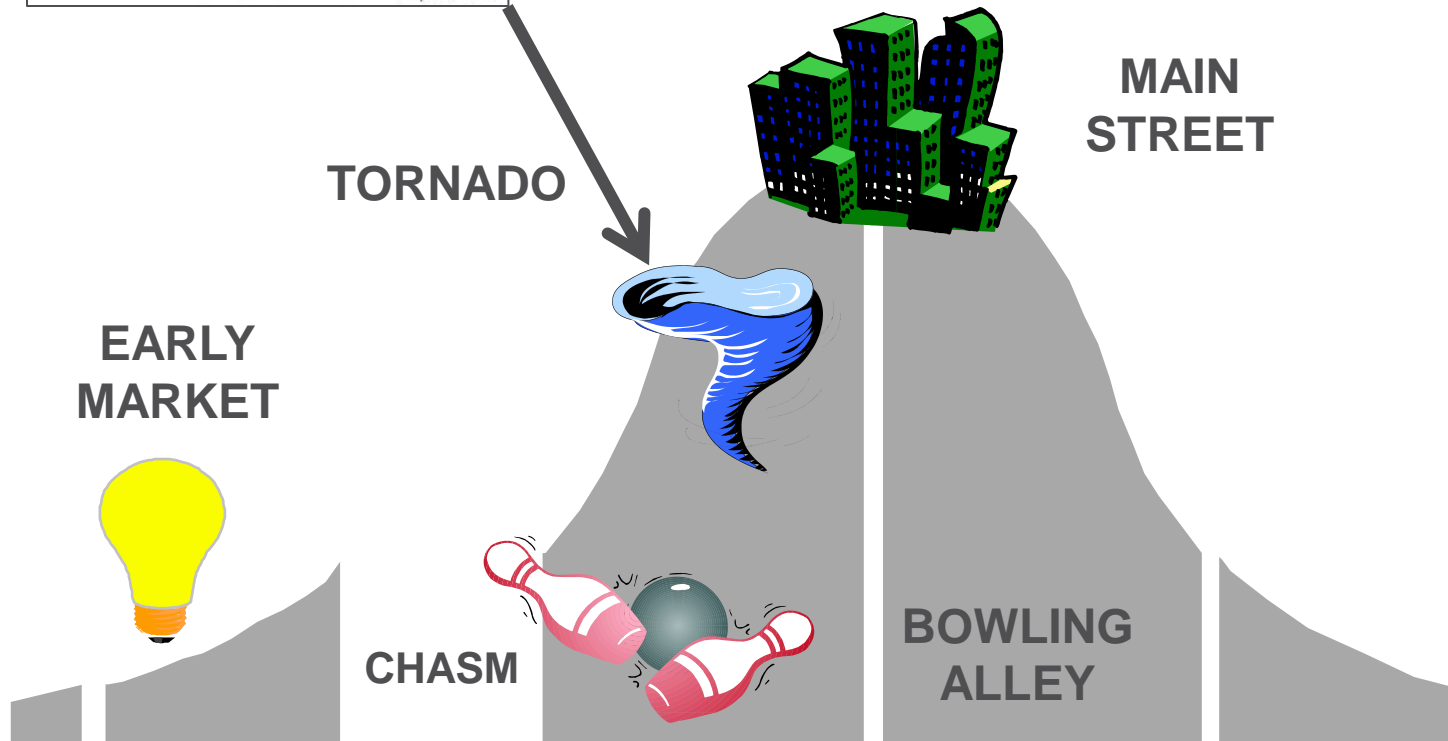


Digital Advertising: Real-Time Bidding



SpotMarket RTB™

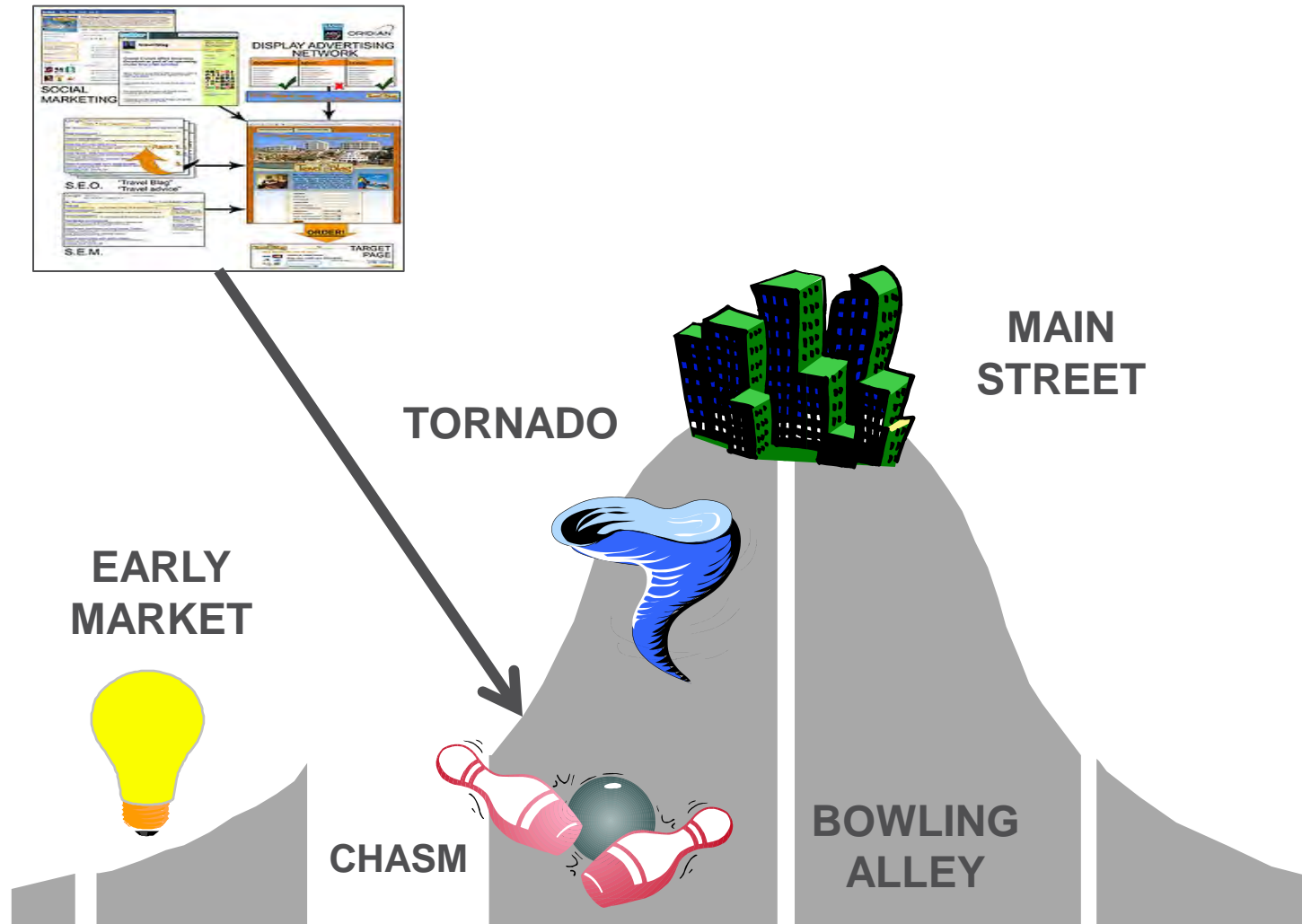
Digital Advertising: Real-Time Bidding



Digital Marketing



Digital Marketing



Algo Trading

The screenshot displays a comprehensive trading application interface with the following components:

- Lab49 WPF Equities Trading Application:** Main application window with menu (File, Launch, Feeds, Window, About), toolbar, and navigation tabs (Stocks list, Stocks graph, Stock 3D Chart, Trade entry, Trade history, Blotter). It also features a status bar with text like "vs 1.0.459 links on big prints and sizes".
- Ticking stock list:** A table showing real-time market data for various stocks.

Symbol	Open	Low	Bid	Ask	High	Change
MSFT	£18.4561	£1.2133	£16.9449	£16.9449	£22.1473	-8.19%
LAB49	£5.5421	£1.2372	£15.0433	£15.0433	£24.7402	171.43%
MS	£8.9690	£2.0632	£11.8979	£11.8979	£24.5019	32.66%
YHOO	£14.6152	£3.0375	£24.5394	£24.5394	£30.5675	67.90%
GM	£1.5193	£1.2154	£20.7949	£20.7949	£23.0061	1268.73%
F	£6.5678	£3.2161	£16.9131	£16.9131	£31.9034	157.52%
- Trade entry:** A window for executing trades for MSFT, showing price (£16.9449), high (£22.1473), low (£1.2133), and issued shares (23639000000). It includes a "Buy" button with a quantity of 100 and an "Execute" button. A confirmation message states "Bought 100 MSFT @ £12.1196".
- Stock 3D Chart:** A 3D bar chart showing price movements for various stocks, with labels for symbols like MSFT, DD, HHH, NNN, and UUU, along with their respective percentage changes.
- Stock price graph:** A 2D bar chart showing price data for MSFT, with a "Stock symbol: MSFT" label and other details like "Company name:", "Bid:", "Ask:", "Open:", "Stock day high:", "Stock day low:", and "Issued shares:".
- Short v Long:** A line chart comparing Long CCI (green line) and Short CCI (orange line) over time, with a y-axis ranging from -200 to 0.
- Coral8 Summary:** A table showing a list of stocks with their current prices, high, low, and number of shares.

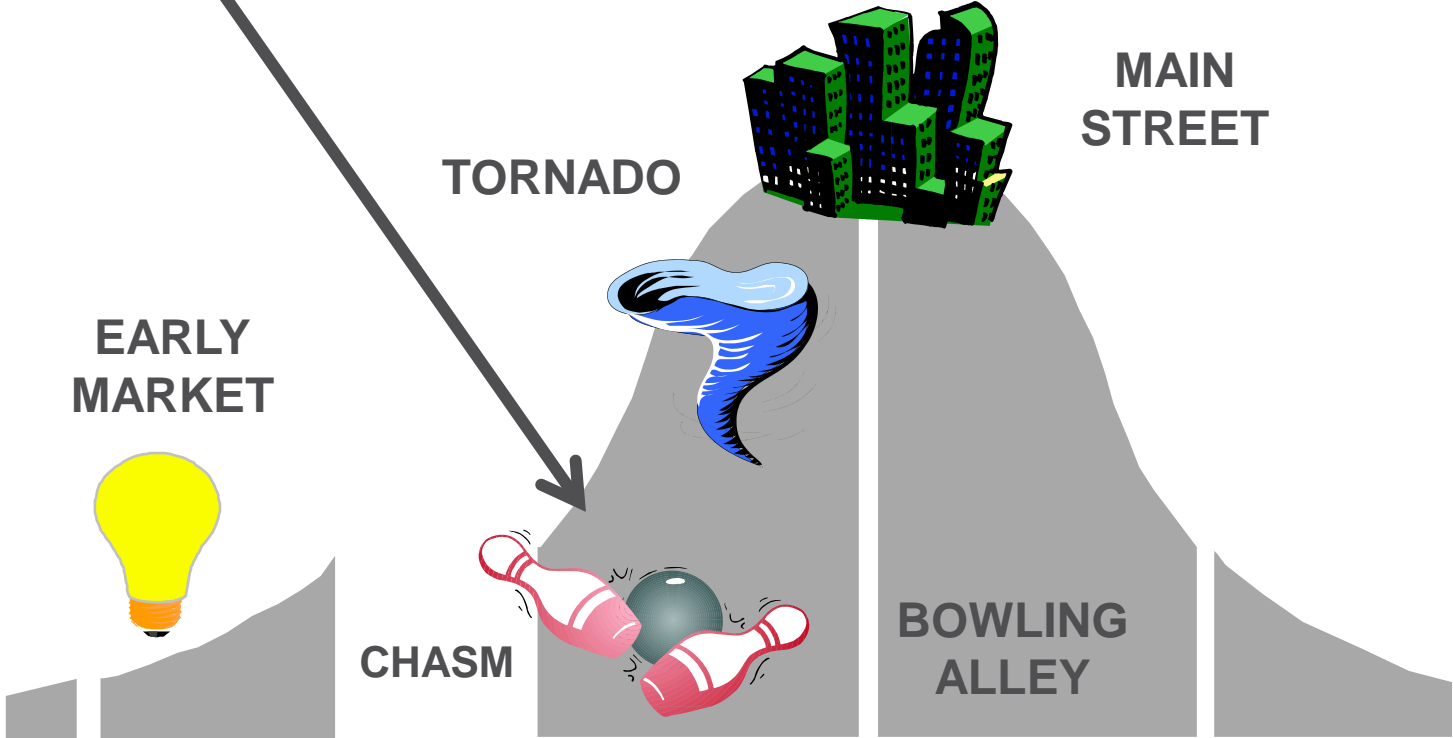
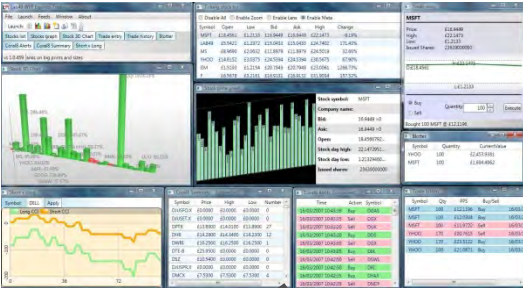
Symbol	Price	High	Low	Number
DJUSFO.X	£0.0000	£0.0000	£0.0000	0
DJUSSET.X	£0.0000	£0.0000	£0.0000	0
DPTR	£13.8900	£14.0100	£13.8900	27
DHB	£14.2900	£14.3400	£14.2300	12
DWRI	£16.2500	£16.2500	£16.2500	1
DTE-B	£25.9500	£0.0000	£0.0000	0
DSZ	£10.5400	£0.0000	£0.0000	0
DJUSPR.X	£0.0000	£0.0000	£0.0000	0
DMCX	£7.5300	£7.5300	£7.5300	4
- Coral8 Alerts:** A table showing a list of alerts with their time, action, and symbol.

Time	Action	Symbol
16/03/2007 10:43:36	Buy	DGAS
16/03/2007 10:43:35	Sell	DGX
16/03/2007 10:43:20	Sell	DUK
16/03/2007 10:43:20	Buy	DDS
16/03/2007 10:43:05	Sell	DGX
16/03/2007 10:43:05	Buy	DIA
16/03/2007 10:42:50	Sell	DSWL
16/03/2007 10:42:50	Buy	DRI
16/03/2007 10:42:35	Buy	DYAX
16/03/2007 10:42:35	Sell	DNDI
- Trade history:** A table showing a list of trades with their symbol, quantity, price per share (PPS), and buy/sell status.

Symbol	Qty	PPS	Buy/Sell
MSFT	100	£12.1196	Buy
MSFT	100	£12.0304	Buy
MSFT	100	£11.9722	Sell
YHOO	170	£20.7615	Sell
YHOO	170	£23.5122	Buy
YHOO	100	£21.0871	Buy
- Blotter:** A table showing the current holdings in the portfolio.

Symbol	Quantity	Current Value
YHOO	100	£2,453,9381
MSFT	100	£1,694,4862

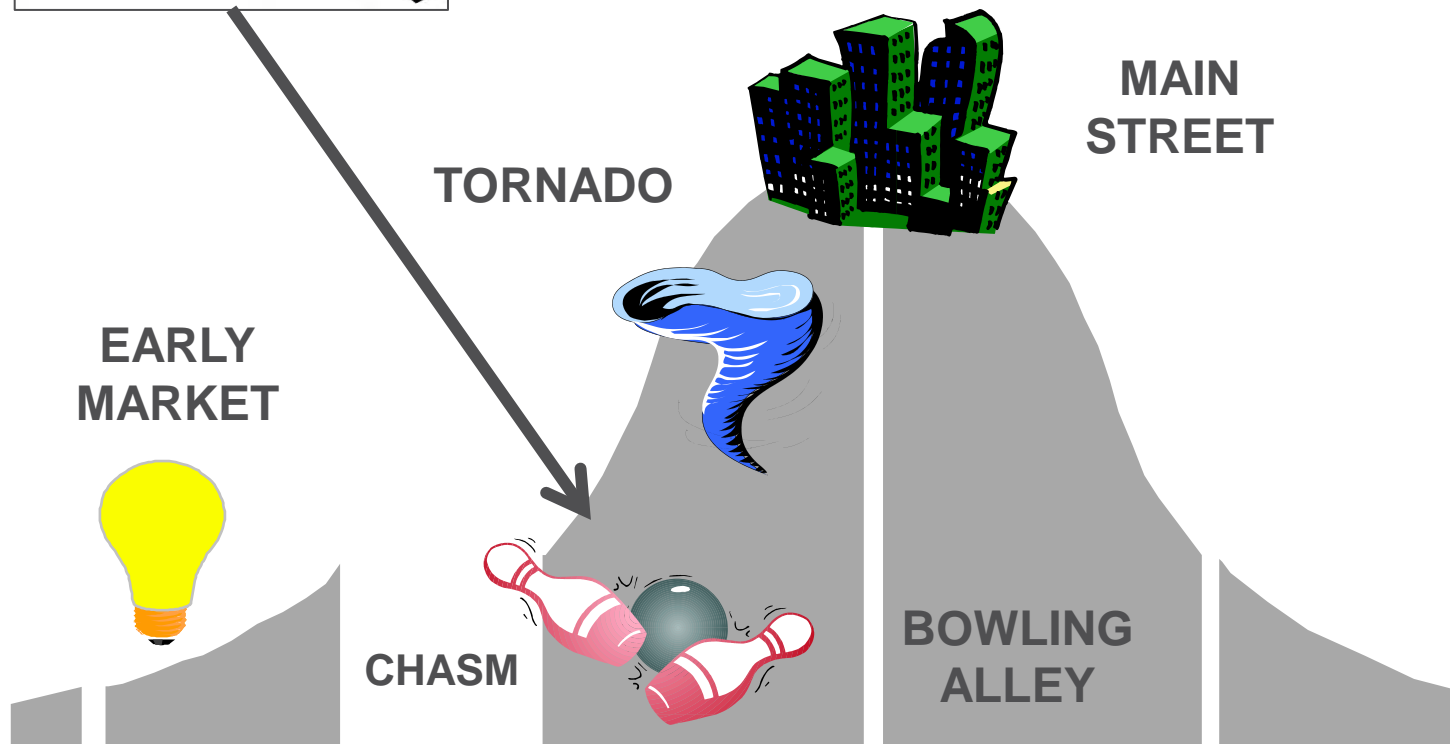
Algo Trading



Fraud Detection



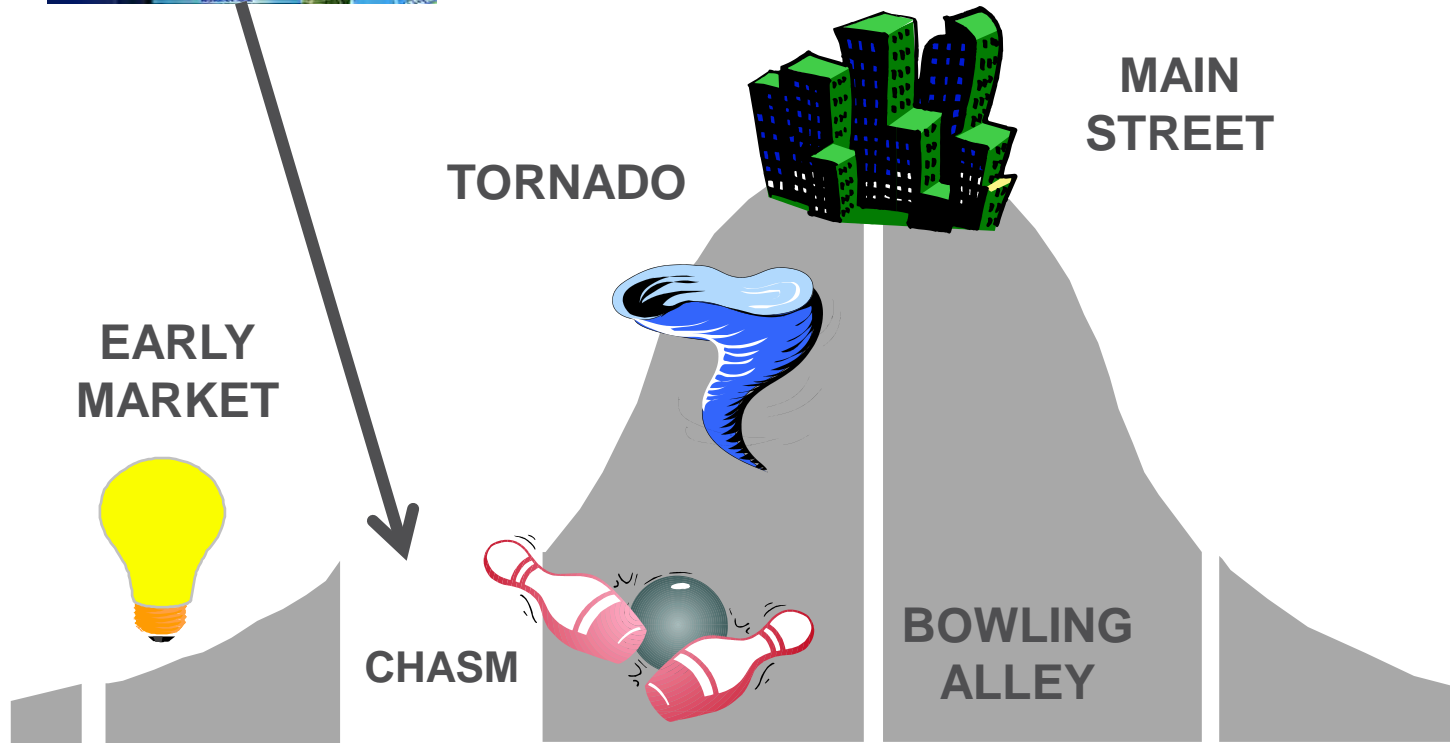
Fraud Detection



Loyalty Programs



Loyalty Programs



Brand Management



Canon

MOVADO
the art of time

**WARING
PRO**

Callaway

NINE WEST

KitchenAid®
FOR THE WAY IT'S MADE.®



TaylorMade®

SONY®



DOONEY & BOURKE

Fisher-Price®

BLACK&DECKER®

**Boston
acoustics®**



GARMIN®

Cuisinart®

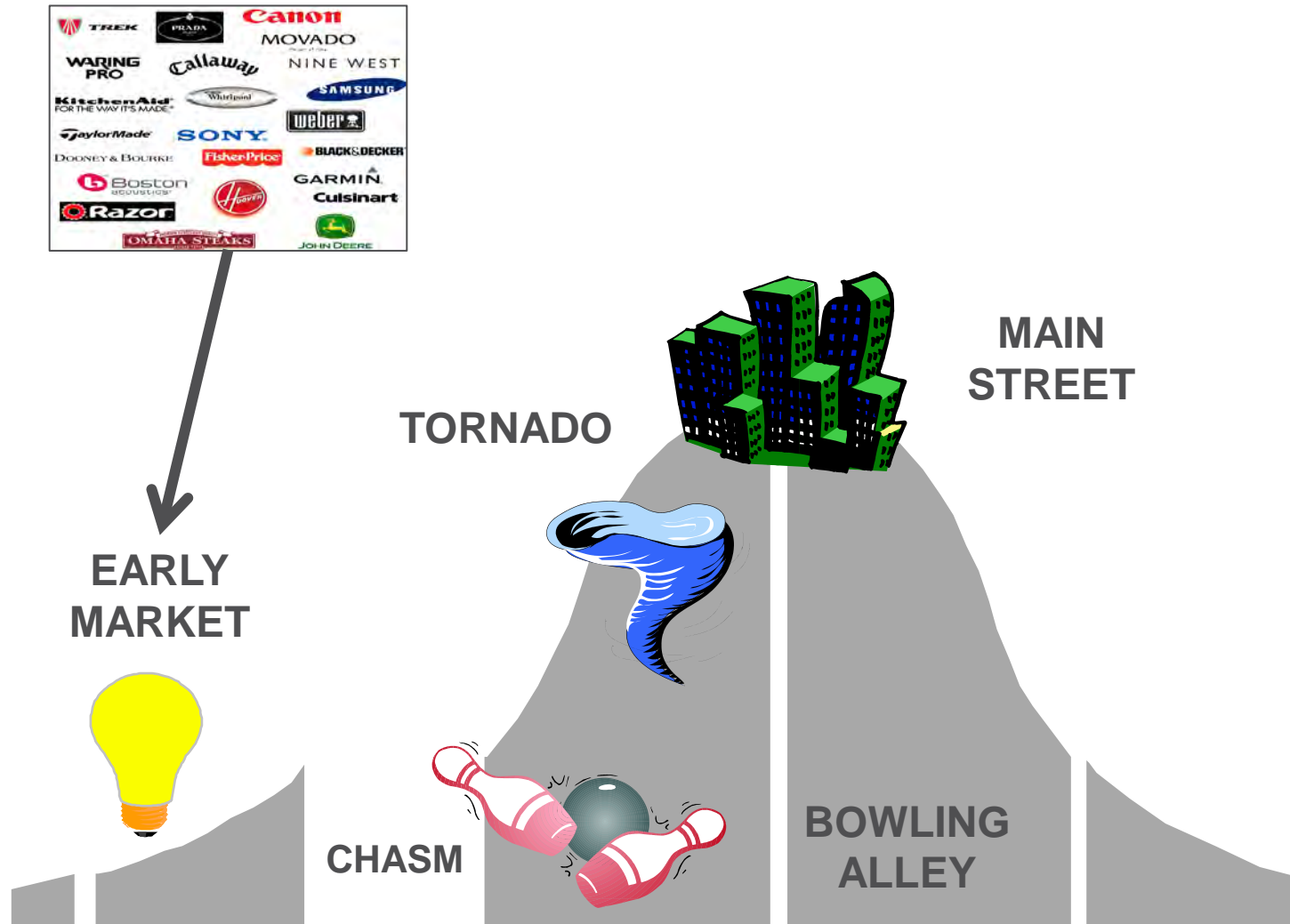
Razor®



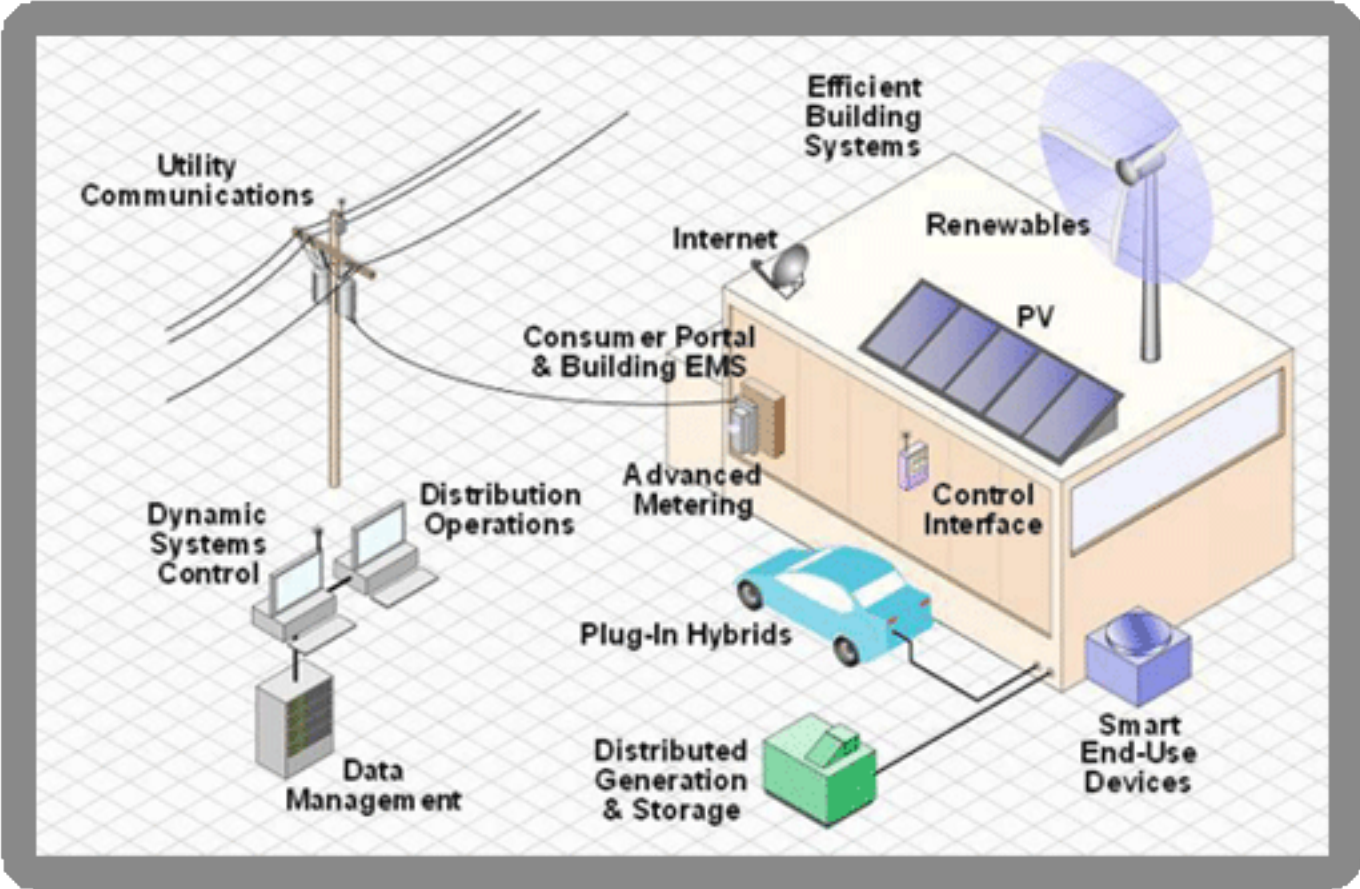
JOHN DEERE

OMAHA STEAKS®
PREMIUM HEARTLAND QUALITY
SINCE 1913

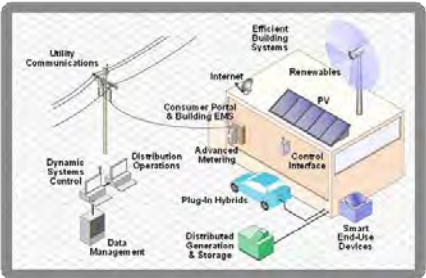
Brand Management



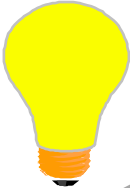
Smart Grid



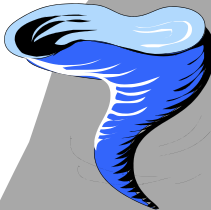
Smart Grid



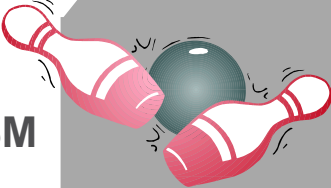
EARLY MARKET



TORNADO



CHASM



MAIN STREET

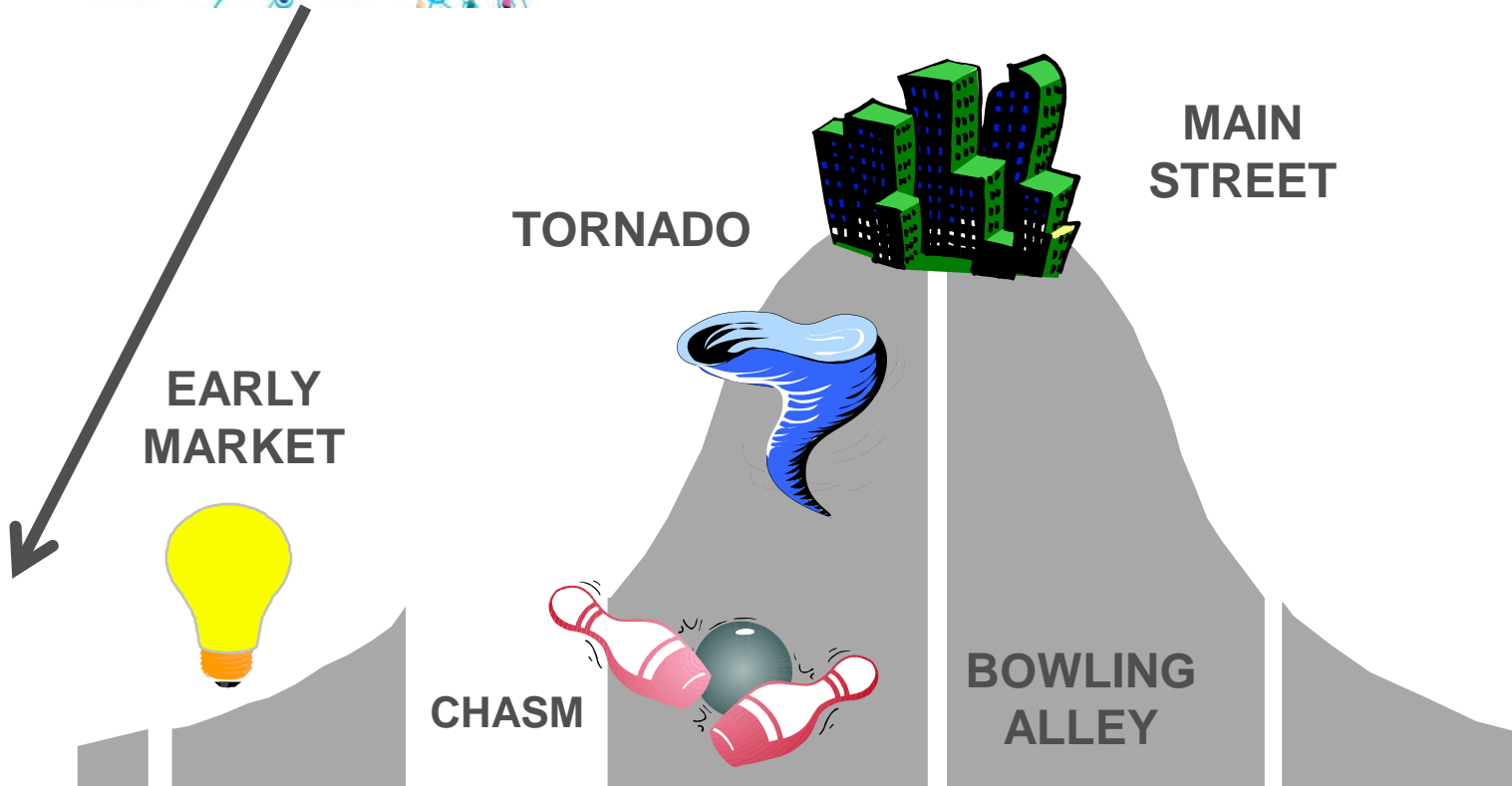
BOWLING ALLEY



Internet of Things



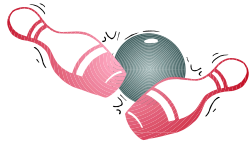
Internet of Things



Key Takeaways



- Project orientation
- **Sell, Design, Build**
- Focus on *performance*



- Solution orientation
- **Design, Sell, Build**
- Focus on *performance/price*



- Product orientation
- **Design, Build, Sell**
- Focus on *price/performance*



- Systems orientation
- **Build, Sell, Design**
- Focus on *price/TCO*

- **Four business models**
- **Each fit for purpose**
- **Match to market maturity**
- **Match to your own crown jewels**

Geoffrey Moore

AUTHOR, SPEAKER, ADVISOR

gmoore@geoffreyamoore.com

twitter.com/geoffreyamoore

charles SCHWAB

Big Data At Schwab

- Nicholas Grabowski, Charles Schwab & Co.
- Stephen Sorkin, Splunk Inc.

splunk™ >

About Us

charles SCHWAB

splunk™ >

- Founded in 1973 as a discount brokerage.
- Grown to a full service financial services company: Brokerage, Banking, Investment Advisor Services, Retirement Planning, etc.
- 13,700 employees.

- Founded 2004, first software release in 2006
- April 2012 IPO
- 4,400 customers in 80+ countries, Over half of the Fortune 100
- Major use cases: Application Management, Operations Management, Developers, Security, Business and Web Analytics

charles SCHWAB

Why Big Data at Schwab?... To better serve our clients

- Serving our clients means understanding what is happening to their transactions and assets at all times.
- Our clients trust Schwab with:
 - Over 1.8 trillion in total assets.
 - Over 9 million accounts.

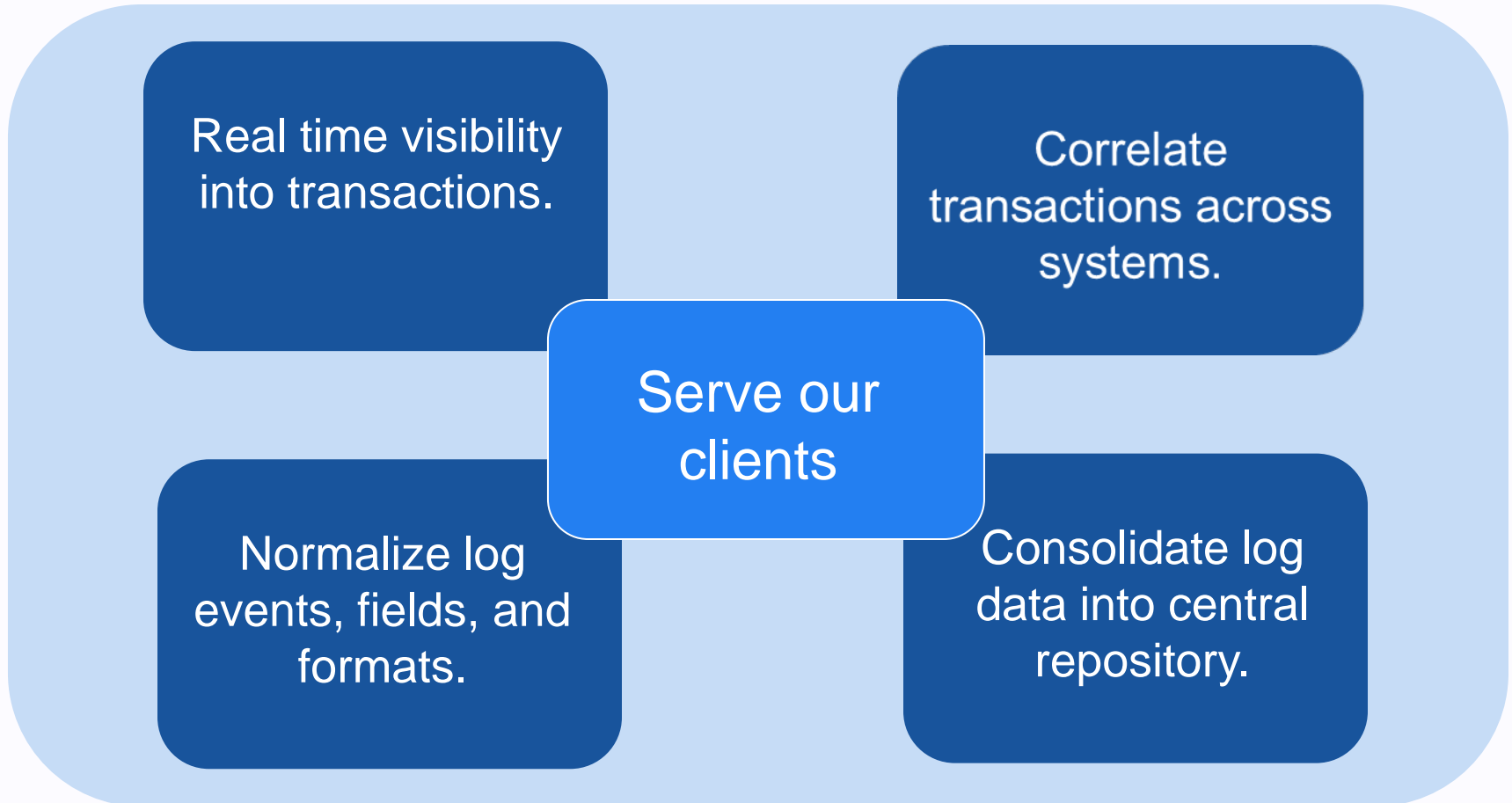
The Schwab ecosystem:

- Terabytes of log data per day.
- Log collection and storage for multiple lines of business
- Troubleshooting and cross system data mining abilities are critical.

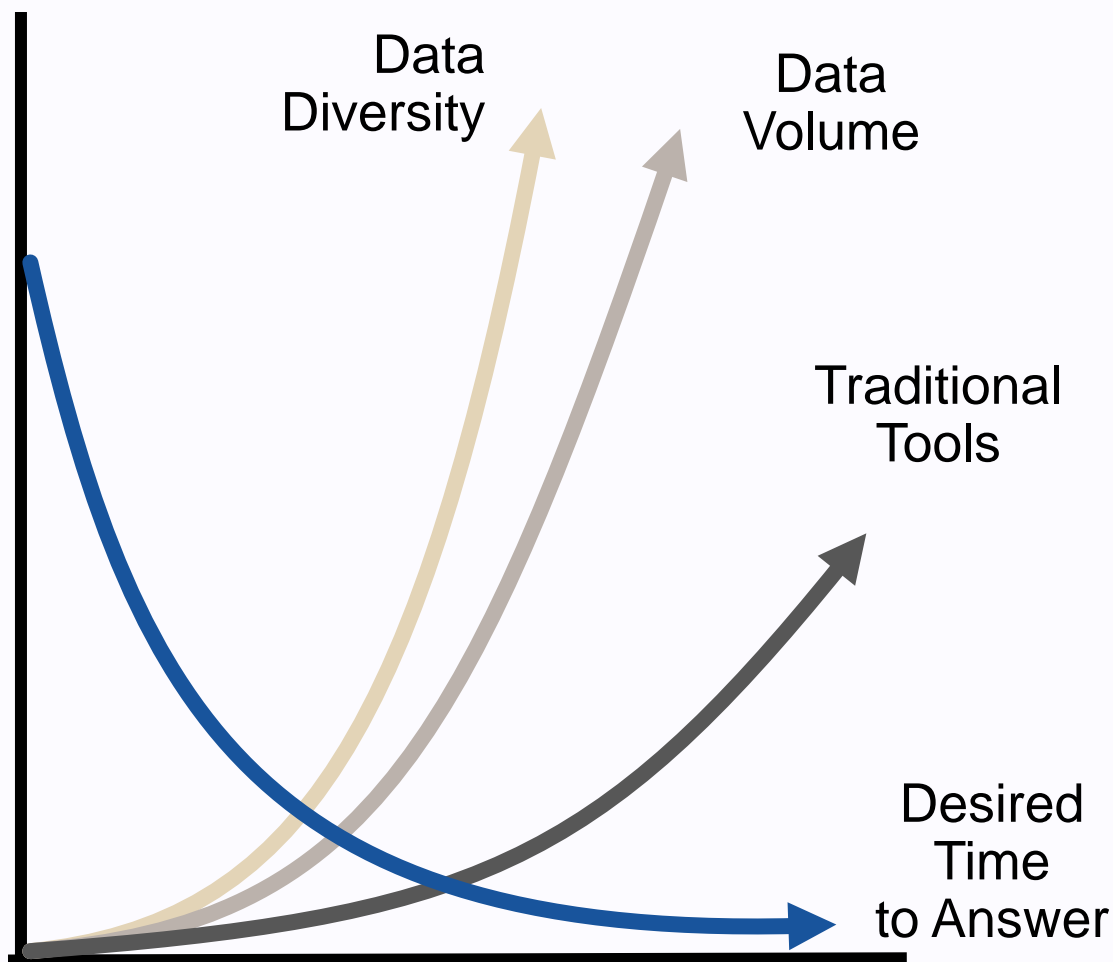


charles SCHWAB

Key Requirements



World's Digital Data Growing Exponentially



“ Big Data is the next frontier for innovation, competition, and productivity. ”

McKinsey Global Institute

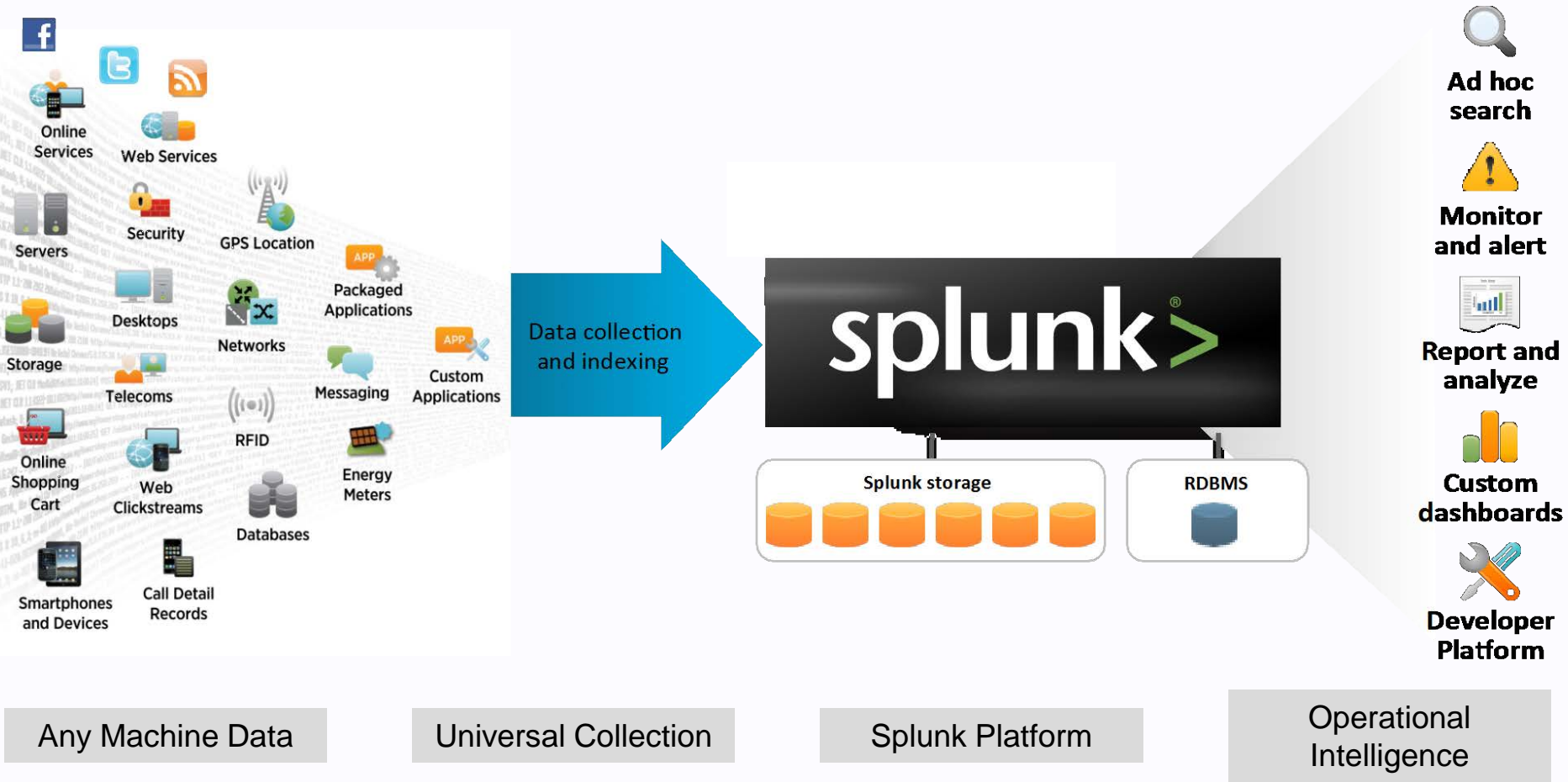
“ Unstructured data accounts for more than 90% of the digital universe ”

IDC 2011 Digital Universe Study: Extracting Value from Chaos

charles SCHWAB

Splunk Turns Machine Data into Real-time Insights

Optimized for real-time, low latency and interactivity



charles SCHWAB

New Approach to Analyzing Heterogeneous Data

Universal Indexing

- No data normalization
- Automatically handles timestamps
- Parsers not required
- Index every term and pattern “blindly”
- No attempt to “understand” up front

Late Structure Binding

- Knowledge applied at search-time
- No brittle schema to work around
- Multiple views into the same data
- Find transactions, patterns and trends

Analysis and Visualization

- Normalization as it’s needed
- Faster implementation
- Easy search language
- Multiple views into the same data

Rapid time-to-deploy: hours or days



charles SCHWAB

Making Sense of Machine Data: Inside Universal Indexing

```
Type:8 Code:0 ID:47447 Seq:4 ECHO
[**] [1:384:5] ICMP PING [**]
[Classification: Misc activity] [Priority: 3]
05/04-11:51:26.224713 10.2.1.48 -> 10.2.1.222
ICMP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:84 DF
Type:8 Code:0 ID:47447 Seq:4 ECHO
[**] [1:408:5] ICMP Echo Reply [**]
[Classification: Misc activity] [Priority: 3]
```

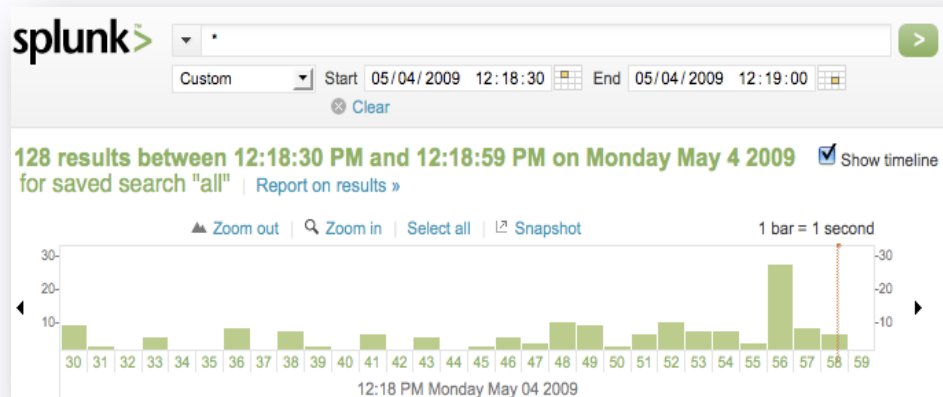
Automatic event
boundary identification

Automatic timestamp
normalization

11:51:26.224713

[Classification: Misc activity] [Priority: 3]
05/04-11:51:26.224713 10.2.1.48 -> 10.2.1.222
ICMP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:84 DF

...enable accurate searching and
trending by time across all data:



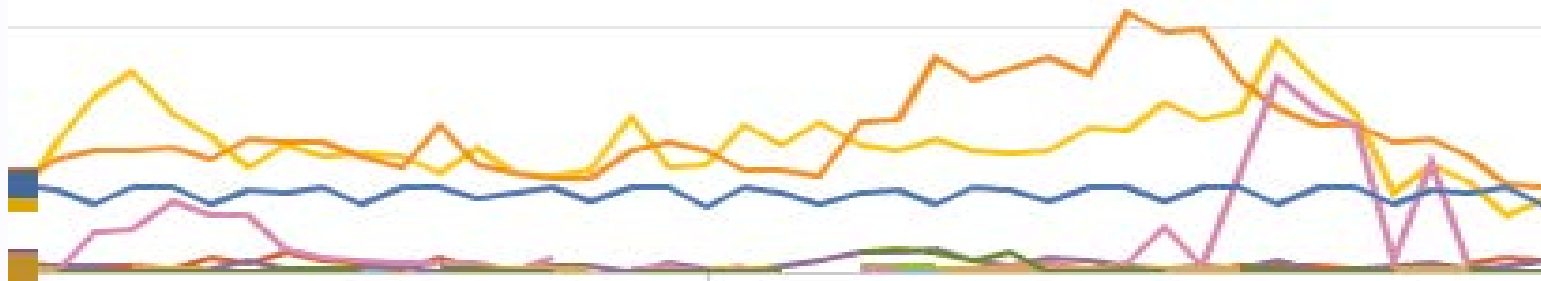
Splunk's Search: Ask Any Question

“What is the average price of Pad Thai in Berkeley over the last 6 months as a chart broken out by zip code?”

```
“pad thai” “Berkeley” | timechart avg(price) by zipcode
```

```
“pad thai” “Berkeley” sourcetype=menu | timechart avg(price) by zipcode
```

```
“pad thai” earliest=-3m | stats max(price) by restaurant
```



charles SCHWAB

Splunk's Search Processing Language

Lots of random “hypothetical examples” from our Mugs

Find **happiness** `happiness` Find **true love** `"true love"` **Down** and **dirty**, or **fast** and **furious** `(down dirty) OR (fast furious)`

Where's **Waldo**? `name="waldo" | fields latitude longitude altitude` **Friend**, friendly, friends, friendlier... `friend*`

What were you doing at the time of the **murders**? `sourcetype=actions person="you" [search action=murder | eval earliest=_time-600 | eval latest=_time+600 | fields earliest latest | format "(" "(" "" ")" "OR" ")"]`

Zombie infestation trends, daily **simple** and **exponential** moving averages `sourcetype=zombies | timechart span=1day dc(id) as z_count | trendline sma10(z_count) ema10(z_count)` Where the **streets** have **no name** ... `source=streets NOT name=*`

Hosts that have not reported in lately `| metadata type=hosts | where lastTime < now()-3600` Is **San Francisco** really **colder** in the summer? `source=weather city=sf-ca | timechart span=1d avg(temp) max(temp) min(temp)`

How much have you had to **drink** tonight, sir? `earliest=@d+17h+15m latest=now item=beer OR item=wine OR item=liquor | lookup nutritioninfo item OUTPUT alcohol_pct | stats sum(eval((alcohol_pct/100)*qty)) as oz_alcohol`

How **long** is this going to take? `source=history | stats stdev(dur) as stdev, avg(dur) as avg | eval soonest=avg-(3*stdev) | eval latest=avg+(3*stdev)` All the king's **horses** `source=hm_stables | top limit=0 horse`

splunk > How about a nice hot cup of search and analytics?

charles SCHWAB

Solution: Enterprise Logging Initiative (ELI) and Splunk

Key Requirements : Correlation, Normalization, Visibility, and Centralization

- Real time collection of data
- Centralized collection and storage
- Democratizes logs - searching all machine data via a single web interface
- Stores large volumes of data (10s of Terabytes for Schwab)
- Format agnostic

splunk® >

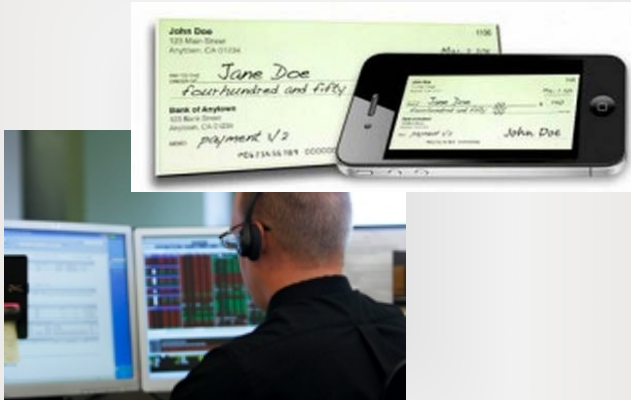
- Provides long term storage via Hadoop
- Specifies log events, format, fields, and format
- Provides Splunk to Hadoop integration

ELI

charles SCHWAB

Enterprise Logging Initiative & Splunk in Action

Mobile



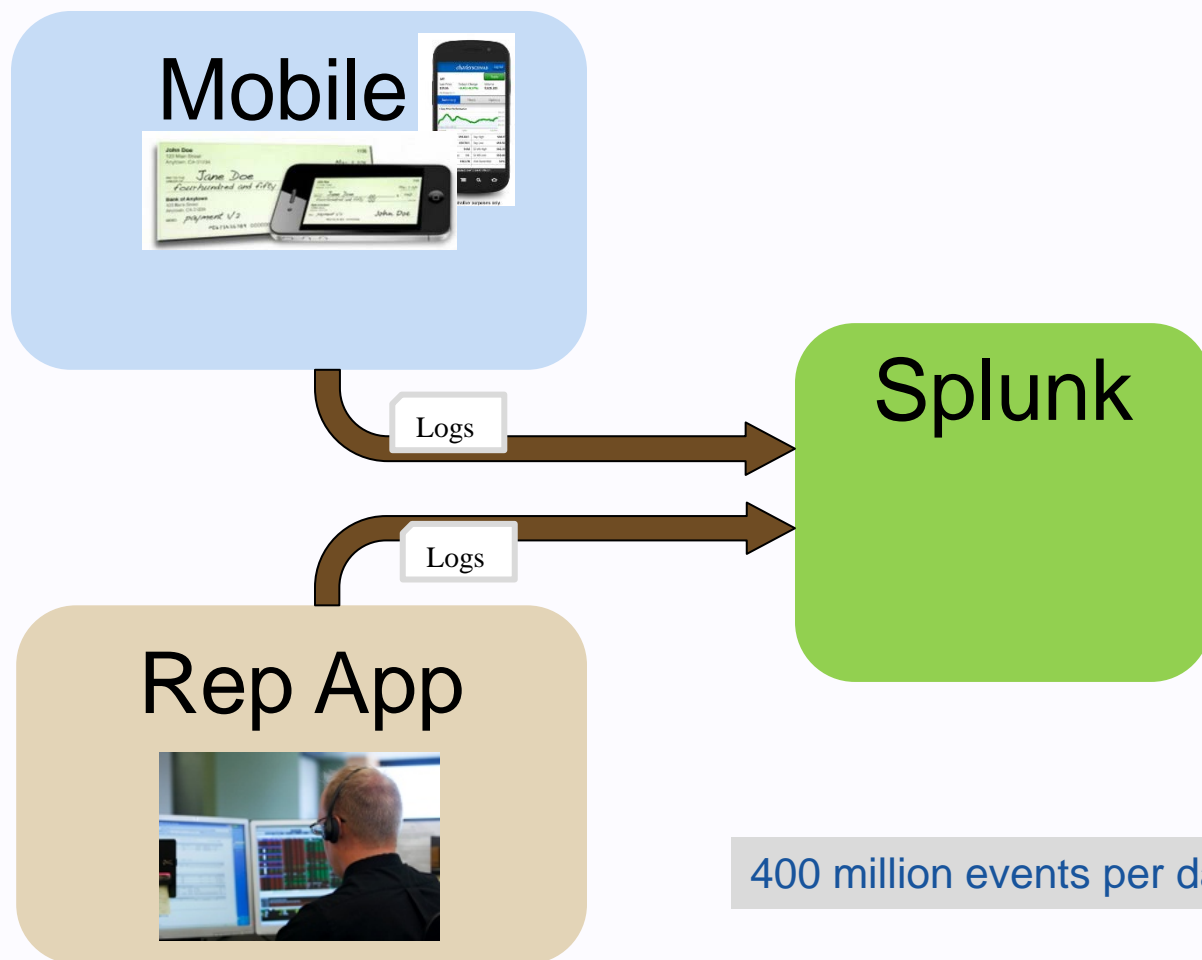
Feature Launch Analysis &
Behavior Tracking

A Big Data demo at Schwab.

Mobile Check Deposit

- Tracking success.
- Not all deposit attempts succeed.
Why not?
- How do clients learn over time?
- What can we do to help clients?

Enterprise Logging Initiative & Splunk in Action



What did we learn from Splunk?

Tracking success of the Mobile Check Deposit.

- About 85% of clients succeed on first attempt.
- Within first 10 days 98+% of clients succeed.

Not all deposit attempts succeed. Why not?

- Image reading is the most common mishap.
 - blurry image, check not in image, etc.

How do clients learn over time?

- With the help of customer service or through trial and error, they figure it out.
98+%

What can we do to help our clients?

- Reach out to them directly.
- Note their account incase they call us.

How Schwab Uses Splunk

Behavior tracking

“How do reps behave during market storm events?”

KPI reporting

“Are trades increasing or decreasing? Is site performance improving?”

Feature analysis

“How did mobile check deposit do?”

Operational monitoring

“There is a production outage!”

Application debugging

“What caused the production outage?”

Splunk at Schwab by the Numbers

2,000+

Servers

20+

Applications

500GB+

Data per Day

100s

Users per Day

90TB

Splunk Capacity

45

Splunk servers

30+

Dashboards

250GB+

Data Archived to Hadoop
per Day for Long-term Storage

charles SCHWAB

Questions



Big Data Ecosystem @ LinkedIn

Outline

- LinkedIn Overview
- Why Data is important
- Big-Data Ecosystem

- LinkedIn is the world's largest professional network at 175M members and growing, with a vision of connecting talent with opportunity at massive scale for the world's 640M professionals.
- We have a selection of data driven products.
 - Recruiter (Hiring Solutions)
 - Premium Subscriptions
 - Marketing Solutions

Some Data Stats.....

On-Line systems (Oracle)

Data Size: 150 TB

Growth month-over-month (approx): 10TB

Queries Per Sec : 150K qps

Offline (TD + Hadoop)

Teradata

Data Size approx: 300TB

Daily Data Load: 2.5 TB

Growth month-over-month (approx): 30TB

Hadoop

Total Size approx : 15 PB of raw storage, 5 PB of usable storage

Total # of (grids) clusters: 9

Total # machines : 5000

Total Jobs per day: 20K

Internal Users: 550

Dedicated Dev & Ops team

What does this translate to

- 1TB of compressed data written to local 'Kafka' clusters per day. Compression ratio is about 3x. This data is replicated to out of the local colo to create global feeds for Data Warehouse and live consumers

.....so what is actually sent over the wire is $\sim 2x$

- We ETL around 1.5 TB (similar compression ratio) into Hadoop and less than that into Teradata. This is the above Kafka data plus database dumps.

Mission: Connect the world's professionals to make them more productive and successful.

Vision: Create economic opportunity for every Professional in the world

We must leverage this mountain of data to fulfill the mission & vision of the company....

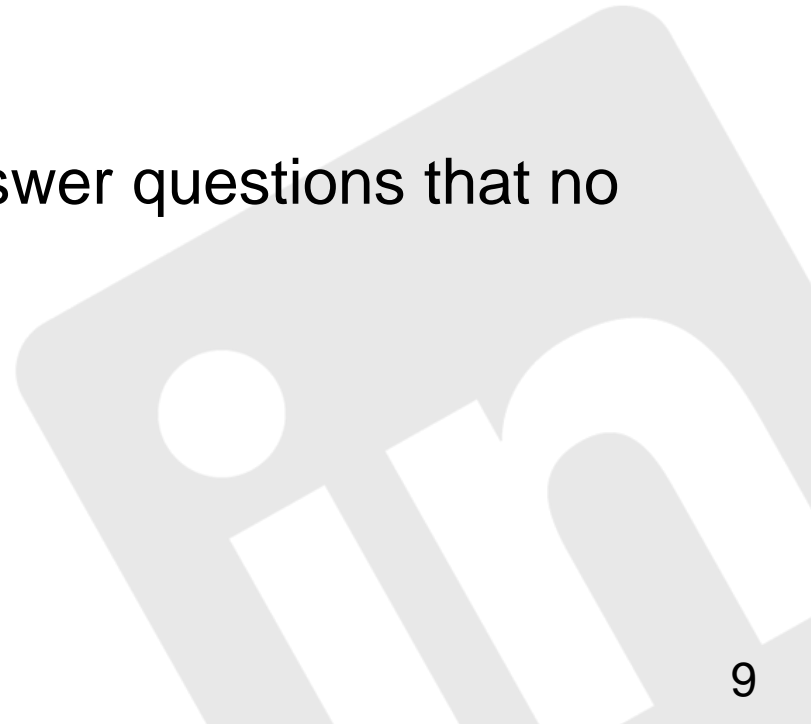
Outline

- LinkedIn Overview
- **Why Data is important**
- Big-Data Ecosystem

Data and Infrastructure Drivers

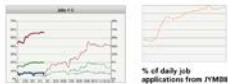
1. Building Enterprise and User-facing Products
 1. Business analytics (e.g., growth, forecasting)
 2. Sales analytics (e.g., customer segmentation, targeting)
 3. Marketing (e.g., campaigns)
 4. Talent Connect (Best candidate for job and vice versa)
 5. Data insights for Customers (e.g., Career site analytics)
2. Measuring and Iterating on Products
 1. On-line: Experimentation (we run ~1000 experiments on the site daily, based on data/analytics)
 2. Off-line: Product analytics (e.g., what is working, what is not)
3. Running the Site
 1. Engineering/Operations/Security metrics (real-time monitoring/alerting for site failures, fraud/abuse)
4. Data Discovery (given the aggregation and the ability to slice/dice in many ways, we can answer questions no one can)

- ❖ Our members come first – they are the most important asset for LinkedIn!
- ❖ They provide profile data, professional graph/connection data and activity stream data.
- ❖ With this data, we can answer questions that no one else can answer



A Sampling of Our Data Driven Products

Jobs You May Be Interested In



Companies

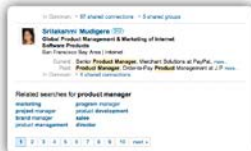
Recommendations, similar companies search, peer companies, and company browse maps, company products and services browse maps



Talent Match



Related search



Behind the Scenes



CAP



Profile browse maps



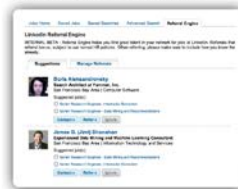
Jobs browse maps



Ad matching engine

$pCTR = f(\text{member, creative, advertiser, context, inventory, OCTR})$

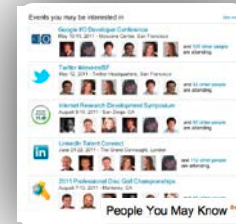
Referral Engine



Pandora Search for People



Events You May Be Interested In



Groups browse maps

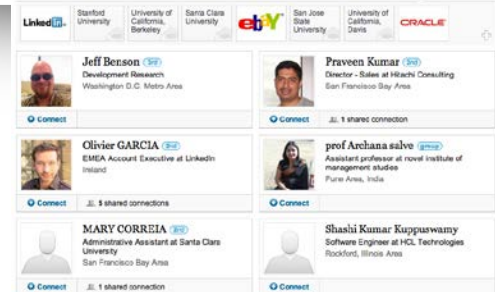


Groups

Recommendations, similar groups search



Similar jobs



Students + Colleges + Companies + LinkedIn: A win-win-win-win based on data & network

COLLEGES

Career Outcomes

Outcome data for marketing, rankings, reporting

Admissions

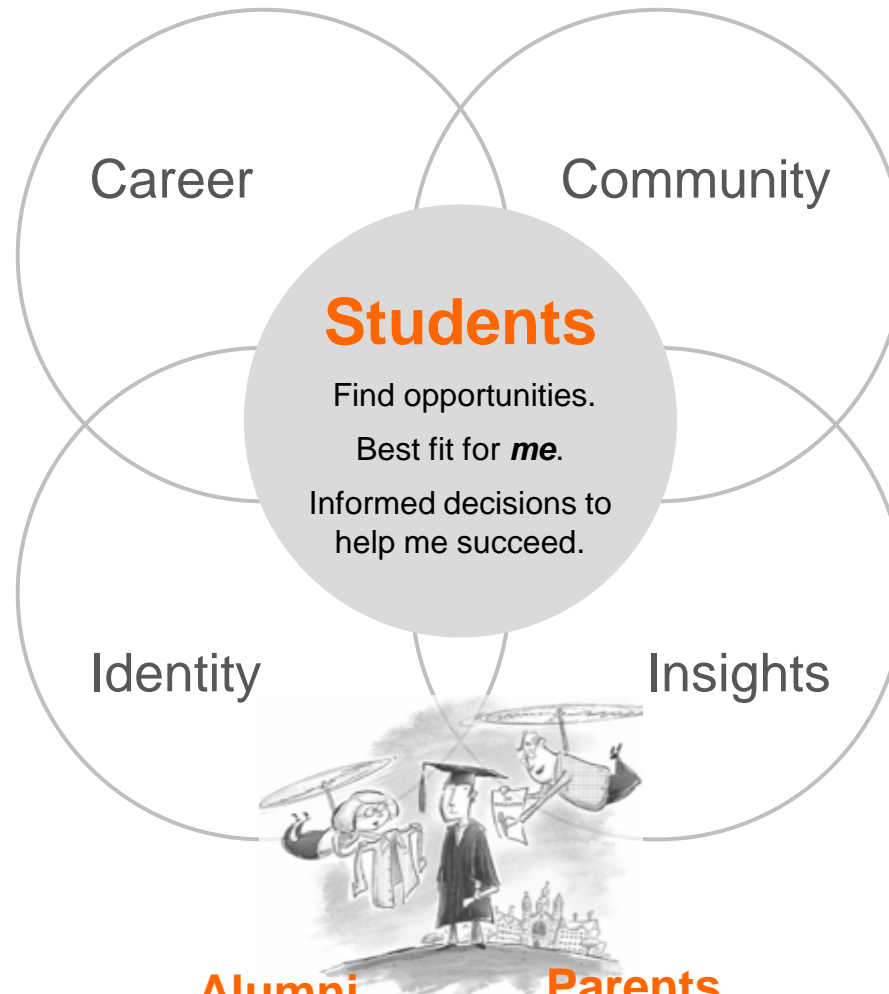
Identify and attract the best students – show success of programs

Alumni Offices

Engage alumni for gifts, relations

Career Centers

Help students land good jobs



COMPANIES

Recruiting

Identify, target and hire the best new talent – early!

Marketing

Reach students at important milestones, target by education

Partnerships

Extend value of systems in higher education

Maintain network, leverage insights, return on investment

Best fit for child. Return on investment. Leverage their connections, nudge, nudge...

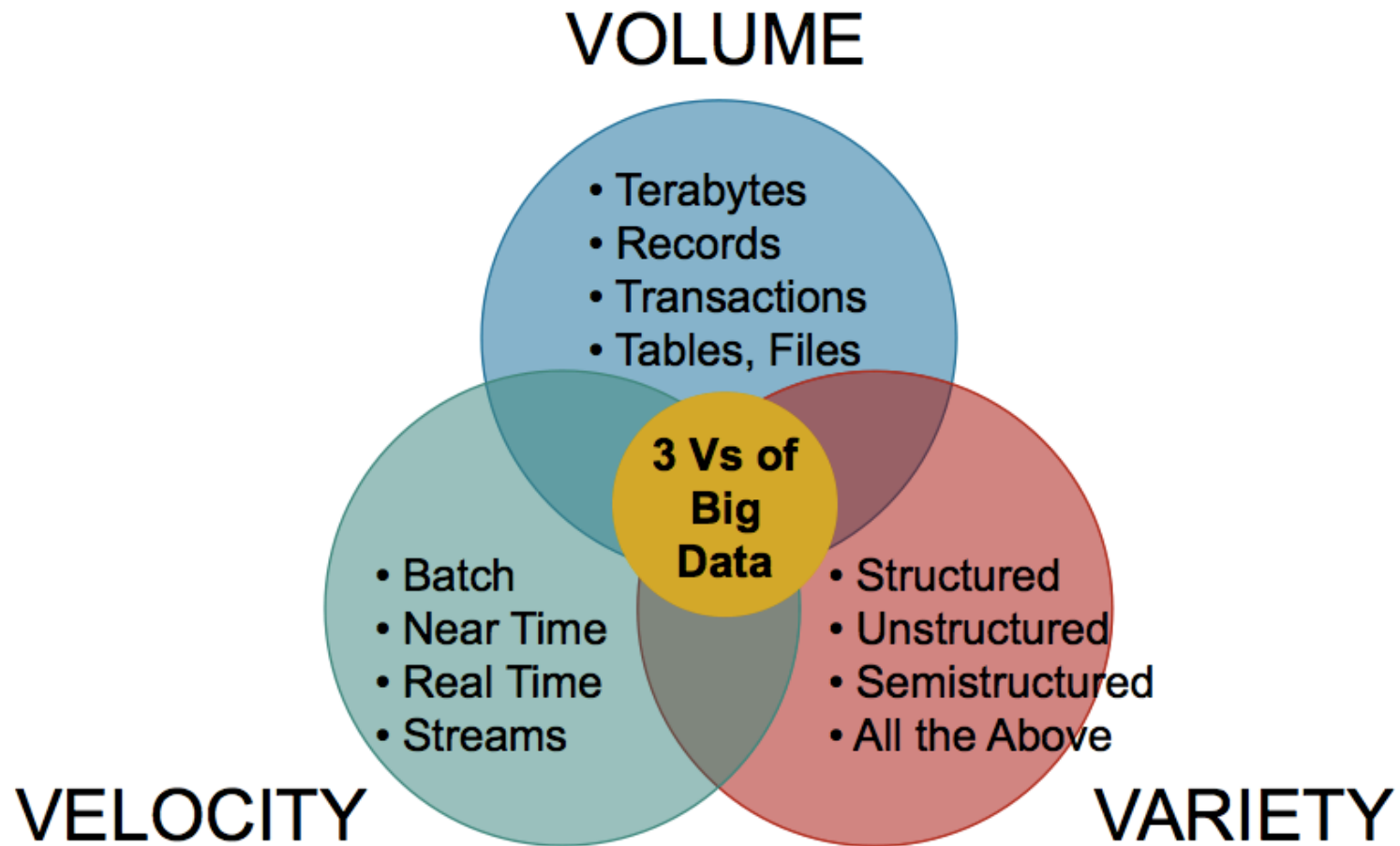
What does Big-Data mean at LinkedIn

- Platform and solutions that
 - Enable scaling with data complexity
 - Simplify the data continuum across online, near-line and offline

Outline

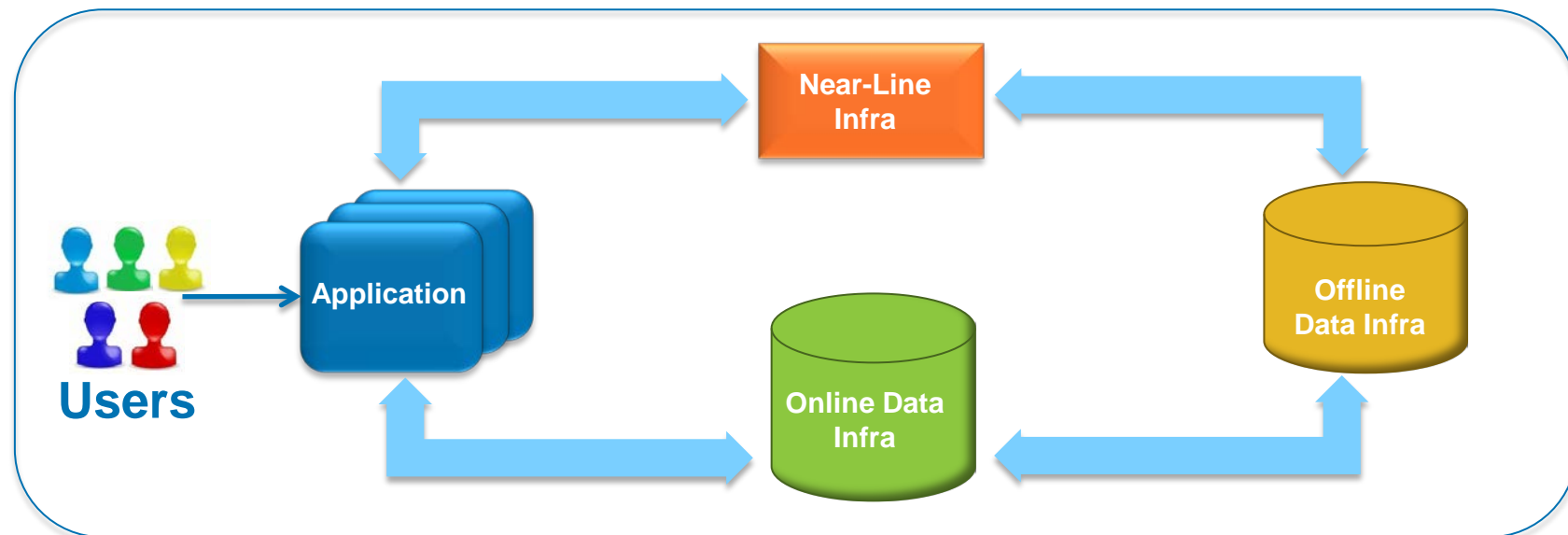
- LinkedIn Overview
- Why Data is important
- **Big-Data Ecosystem**

Big Data at LinkedIn



* Chart from Philip Russom- Research Director: TDWI

LinkedIn Data Infrastructure: Three-Phase Abstraction



Infrastructure	Latency & Freshness Requirements	Products
Online	Activity that should be reflected immediately	<ul style="list-style-type: none"> Member Profiles Company Profiles Connections Messages Endorsements Skills
Near-Line	Activity that should be reflected soon	<ul style="list-style-type: none"> Activity Streams Profile Standardization News Recommendations Search Messages
Offline	Activity that can be reflected later	<ul style="list-style-type: none"> People You May Know Connection Strength News Recommendations Next best idea...

LinkedIn Data Infrastructure: Sample Stack

ORACLE®

ESPRESSO

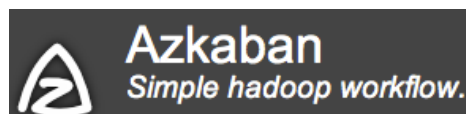


MySQL



HELIX

Kafka



Databus

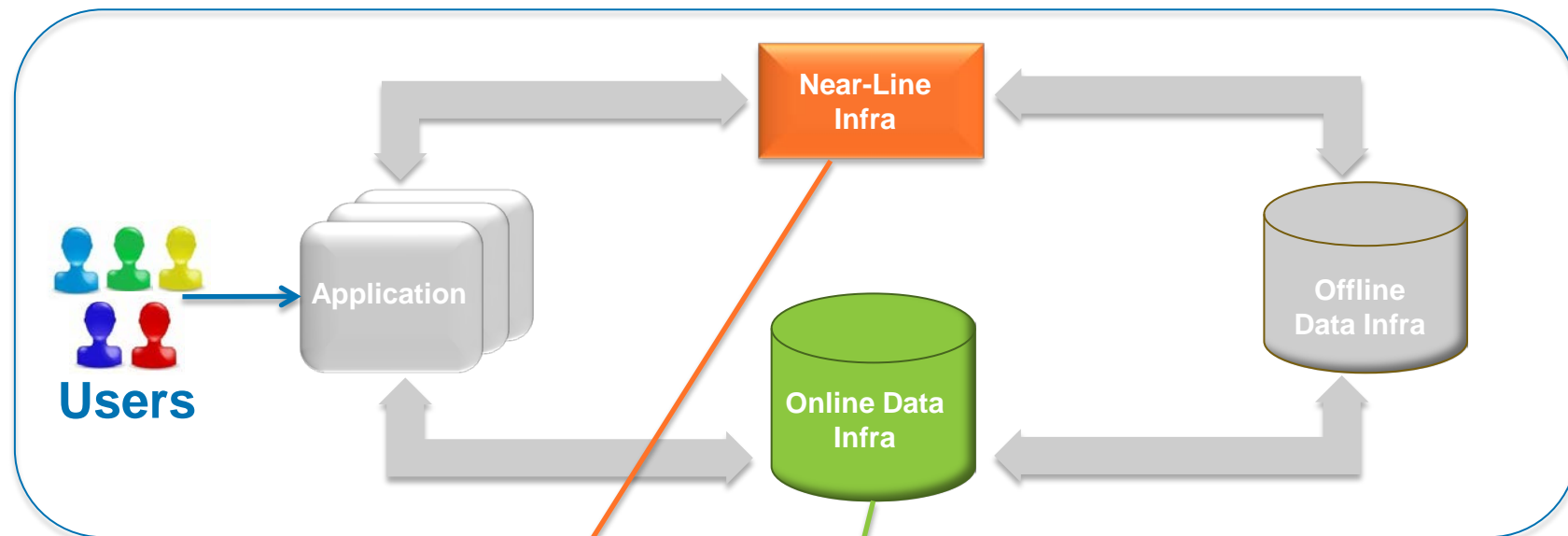
TERADATA®

Infra challenges in 3-phase ecosystem are diverse, complex and specific



Some off-the-shelf. Significant investment in home-grown, deep and interesting platforms

LinkedIn Data Infrastructure: Data Stores



Systems

ORACLE®

ESPRESSO

MySQL™

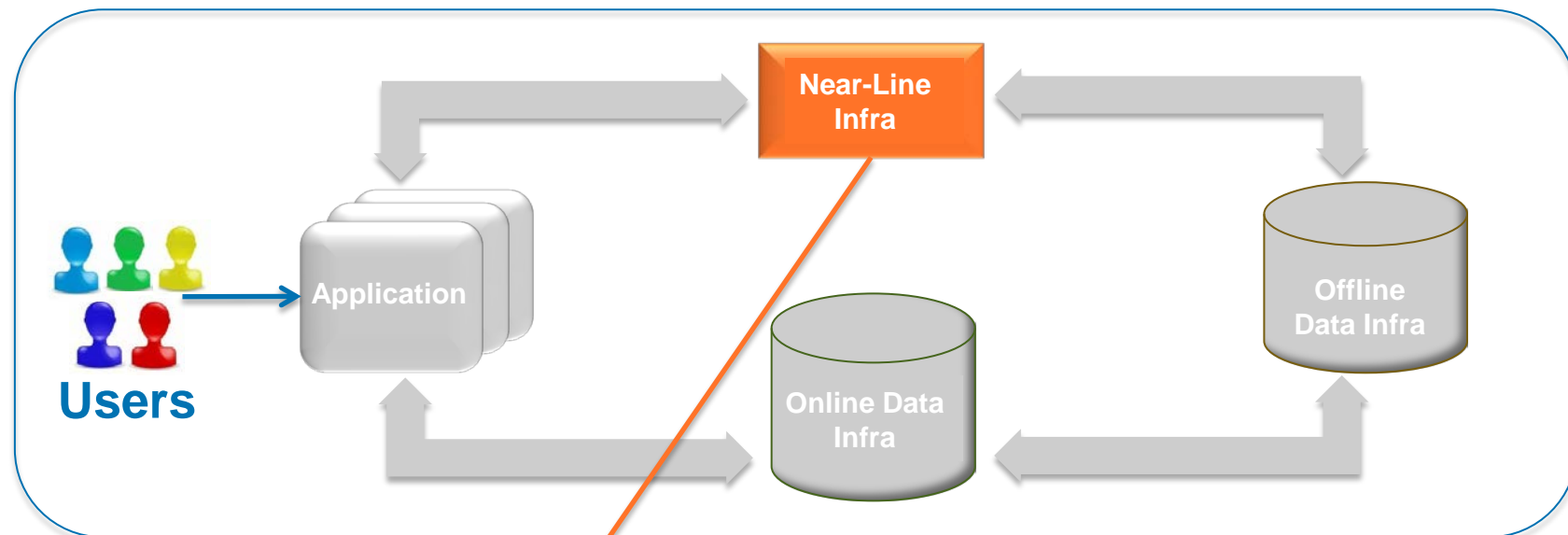


Voldemort

Capabilities

- Transactions
- Rich structures (e.g. indexes)
- Change capture capability
- Key value / document storage

LinkedIn Data Infrastructure: Specialized Indexes



Systems



Zoie



Bobo



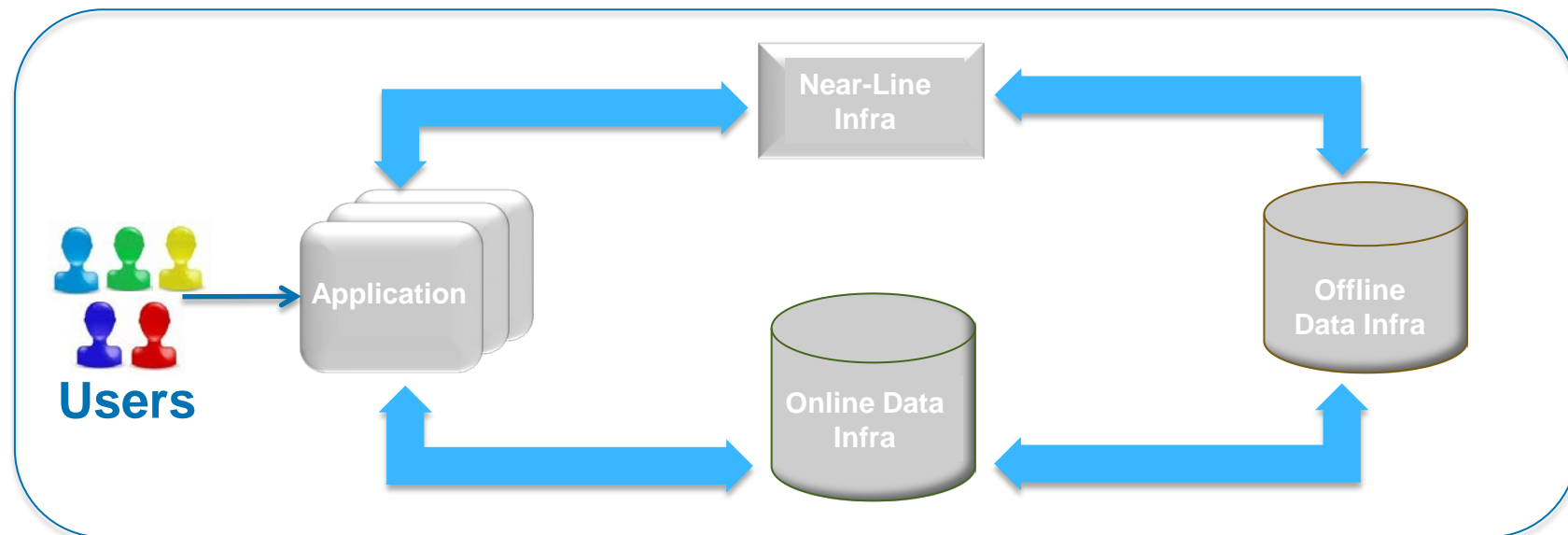
Sensei

GraphDB

Capabilities

- Search platform
- Distributed graph engine

LinkedIn Data Infrastructure: Pipelines



Systems

Kafka

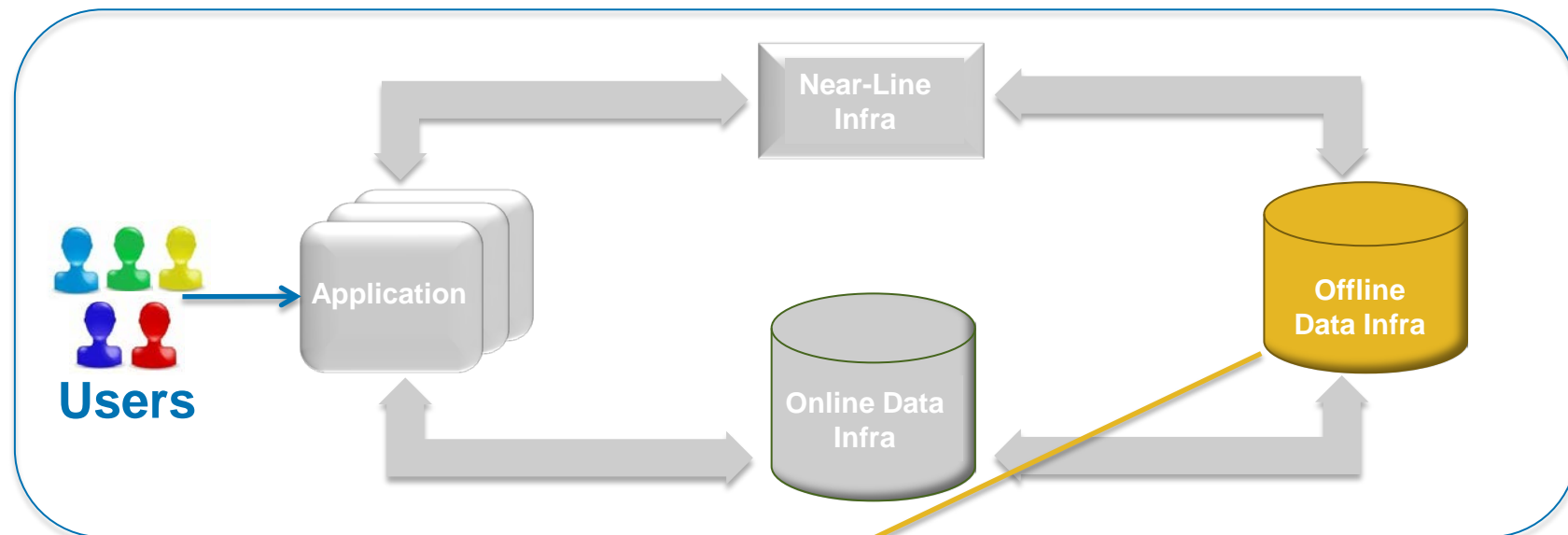
- Messaging for site events, monitoring
- High throughput

Capabilities

Databus

- Change data capture stream
- Reliable, consistent, low latency pipe

LinkedIn Data Infrastructure: Off-line Analysis



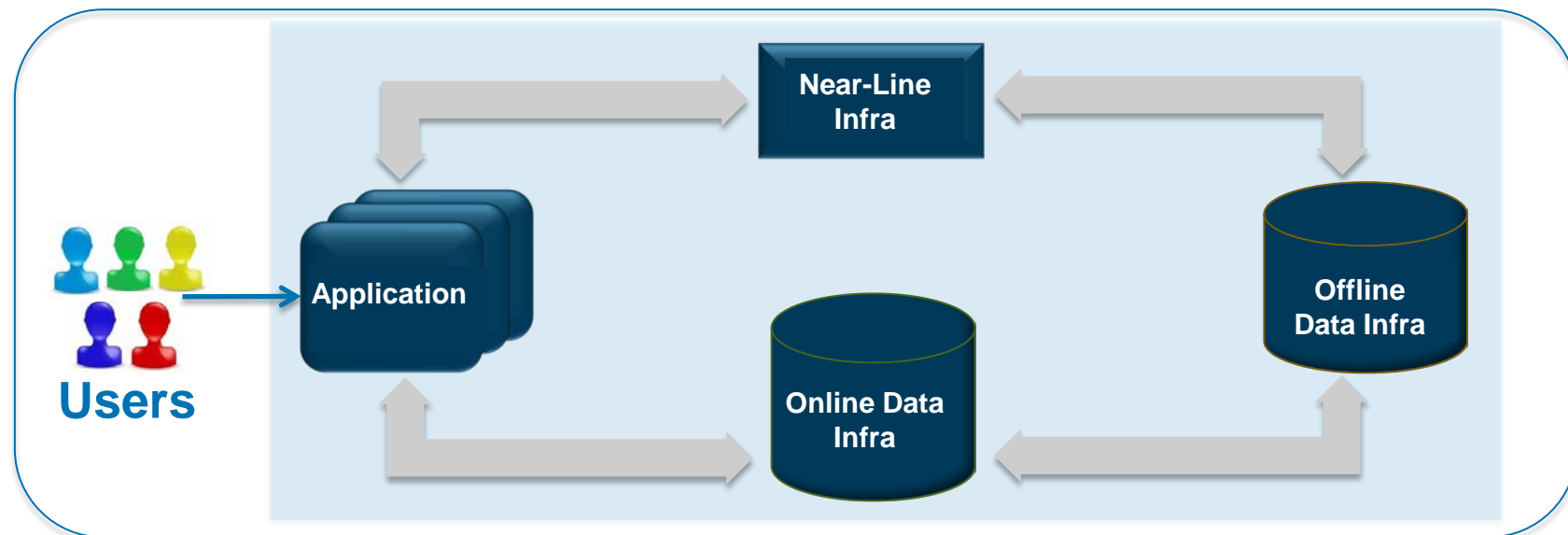
Systems



Capabilities

- ML, Ranking, Relevance
- Insights and Analytics
- ETL, Metadata and Pipes
- Business Source of Truth

LinkedIn Data Infrastructure: Cluster Management



- Generic framework for building distributed systems
- Declarative model of cluster management Primitives
- Encapsulates multiple shard-management logic (Assignment of Master-Slave, Elasticity and Rebalancing, Fault Tolerance and Recovery)
- Leverage: Used in Search, Databus, Espresso

Core tech stack



Open Source Contributions



Kamikaze
Utility package for compressed arrays



Voldemort
Distributed key-value storage system. LinkedIn created



Sensei
A distributed, elastic, real-time, searchable database. LinkedIn created



Zoie
Real-time search and indexing system built on top of Apache Lucene



Azkaban
Simple hadoop workflow. LinkedIn created



Bobo
Fast faceted search with Lucene



Kafka
Data pipeline and messaging. LinkedIn created



Helix
Generic Cluster Manager

LinkedIn

Simpleton History of Data and Computing

- First there was horizontal distribution of the data (e.g., sharding in Oracle)
- Next there was horizontal distribution of the computation (e.g., GFS, Hadoop, Map-reduce)
- What is next? Data1, Data2,Data n fed into computation engineA, computation engineB, ... Computation engine n in real-time (holy grail)
- Why is this important: the nature of the data has changed with use models on the Internet, including social/professional media, mobile computing, cloud computing
- ***More than ever we need to derive useful signals from the noise and clutter of information at speed and scale (“Hyper-cube”)***

Web 3.0 – It's all about data!!

