Big Data
The Game Changer

Insights & Challenges

23 March 2018

Building a better working world

Outline

Background

Introduction to Big Data

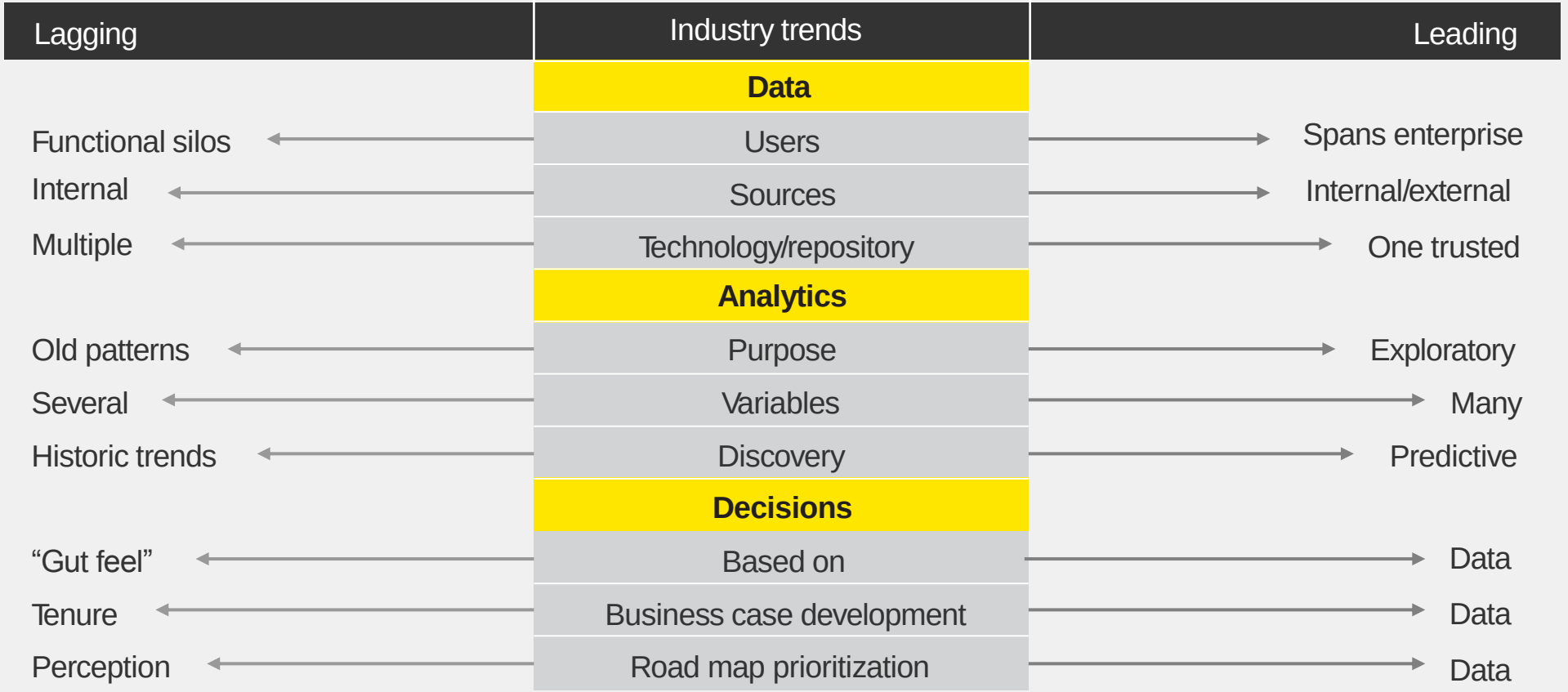
Big Data Technology & Solution

Big Data Architecture

Use Cases

Background

Industries are moving towards more data-driven decision making



Market Trends in Data..

- Neil Armstrong lands on the moon with 32KB of data (1969)
- Google processes 24PB of data everyday (2010) ~ 240K 100GB hard drives
- Twitter ~ 8TB per day (in 2010)
- Facebook ~ 500TB "processed" per day (2012)
- 2.7ZB estimated to exist in 2012, projected to grow to 8ZB in 2015 (IDC, 2012)

So What is a PetaByte?

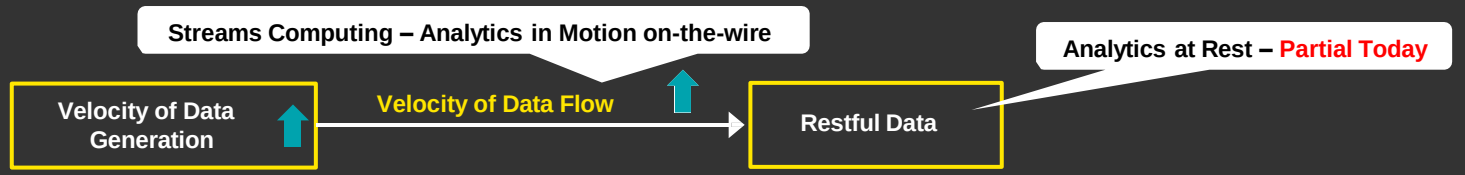
- 1 PB = 1 000 000 000 000 000 bytes = 1 million gigabytes = 1 thousand terabytes = .000001 zetabytes (ZB)
- Large Hadron Collider produces ~ 15 PB per year
- The movie Avatar took ~ 1 PB of storage for rendering 3D CGI graphics

Spending 80% of your life with 20% of the data!!!

- 80% of new world data generated is unstructured and semi-structured (this includes images, music, and video) → will grow to 90% by 2015
- Relational structures | strict schemas | canonical encodings will persist → But will progressively constitute less and less of the market data share
- Analysis requirements therefore are shifting towards unstructured and semi-structured data
- Traditional analysis platforms cannot scale to meet the new data paradigm

How Fast is Fast?

- Velocity of data generation is increasing exponentially → Proliferation of automatic sensors in software and hardware
- Necessitates proportional increase in the velocity of data analysis



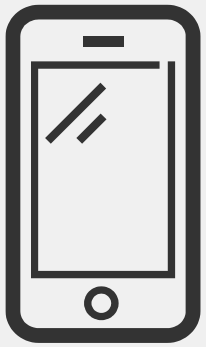
- **WHY need Velocity of Analytics? → Finding arbitrage opportunities in capital markets before asset prices balance**

Sources: IDC Predictions 2012: Competing for 2020, Twitter by the numbers (2010) Facebook processes… (CNET, 2012)

New data keeps flowing in every day

90% of the world's data today has been created in the last **2 years** alone

6 Billion smartphone subscription = **87%** of world's population

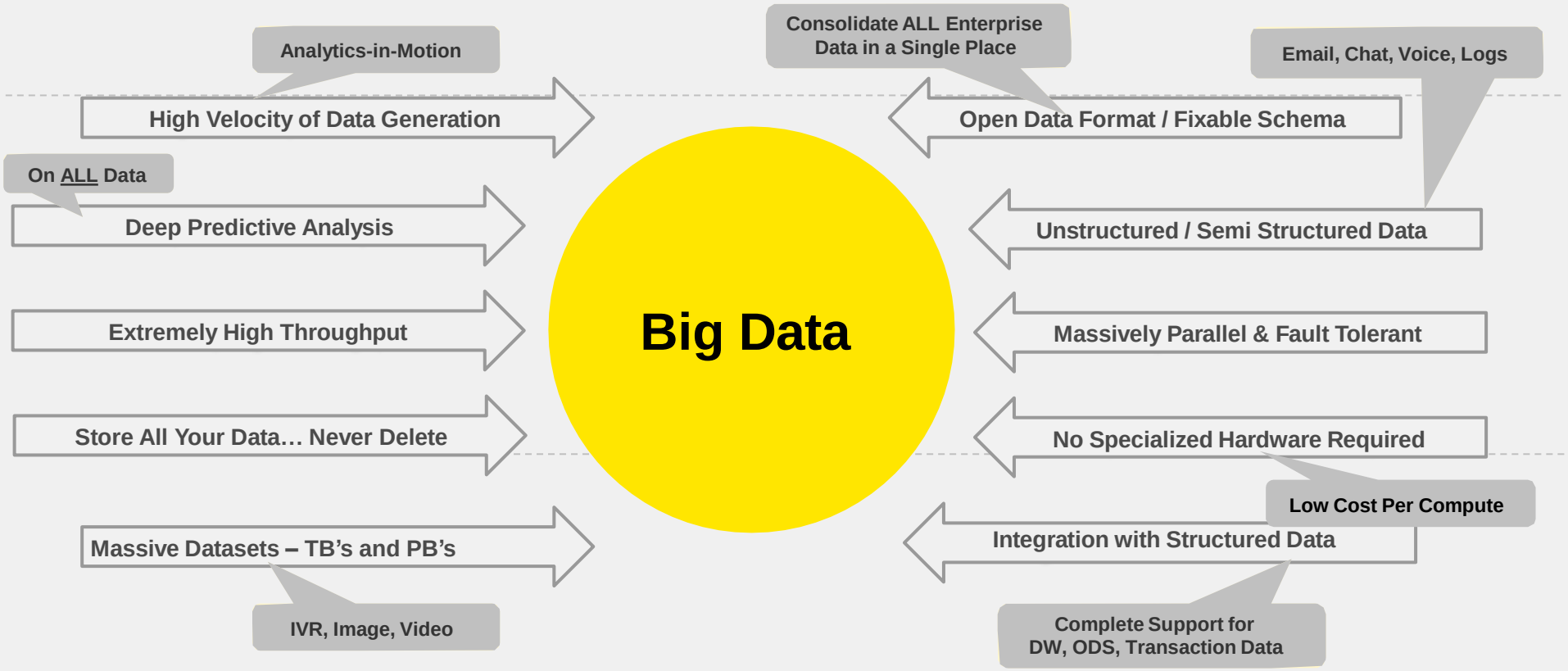


Global IP traffic has reached **1.1 zettabytes**, and by 2020 will grow up to 44 zettabytes (or **44 trillion gigabytes**)

Enormous amount of data are generated every day

~80% of new data generated is either completely unstructured (e.g. images, videos, and music) or semi-structured (e.g. text-based web content)

- Much of this data is being generated at high velocity
- Presents tremendous challenges for storage, search, retrieval and analysis
- Traditional data platforms and analysis capabilities are unable to meet these evolving demands! [17]



Source: Big Data: Big Hype? (Barrenechea, 2013)

What makes big data possible?

Commodity hardware utilization..

Distributed computing lends itself to exploiting affordable commodity hardware. The following table compares expensive, specialized supercomputing infrastructure with commodity hardware typically found in Big Data environments.

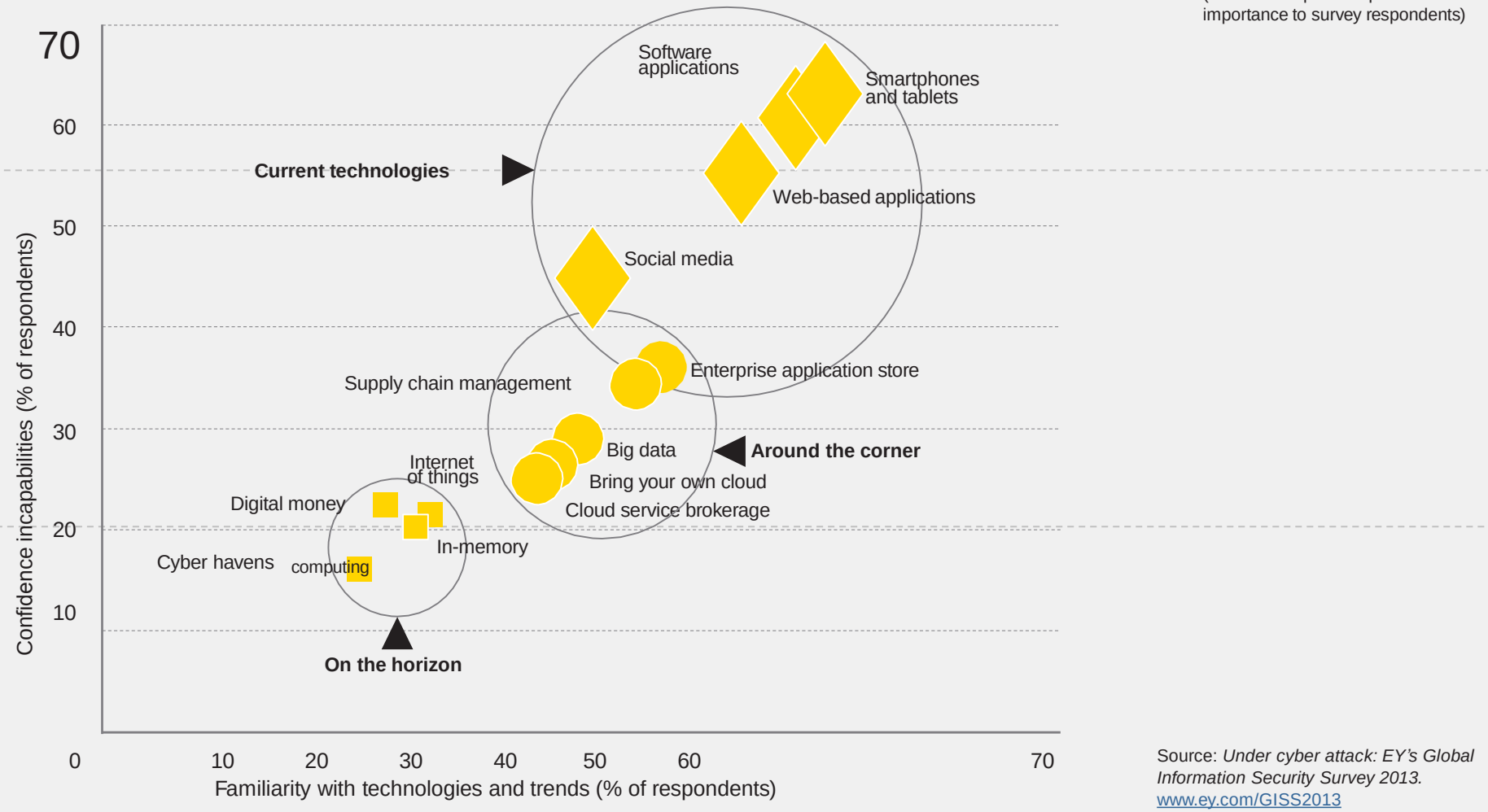
Dimension	Commodity Hardware	Supercomputers
Hardware cost	Less Expensive	Much more expensive
Sales channel	“Off the shelf”	Made to order
Hardware focus	General purpose	Specialized (function optimized)
OS Support	Commodity (e.g. Linux for enterprise)	Specialized operating systems
Plug-and-play support	Modular design	Specialized interfaces
Failure rate	Higher on individual machines	Lower on individual machines
Number of machines	Hundreds or thousands	< 100
Redundancy	At the application layer	Supported through application or hardware (e.g. RAID, backup NAS)
Computing power	Average in terms of individual machines	Leading edge
RAM resources	Average	Pushing the boundaries

- Keep in mind that yesterday’s supercomputers might be today’s commodity hardware!
- Much of the software out there in Big Data manages resources and data in distributed computing environments backed by commodity hardware, letting you focus on better supporting value-add work!

Big Data technologies are just around the corner

And starting to generate attention among businesses

Emerging technologies and trends



Source: *Under cyber attack: EY's Global Information Security Survey 2013.* www.ey.com/GISS2013

Introduction to Big Data

What is Big data?

- A generic term used to describe extremely large amounts of structured and unstructured data
- Represents the capture, storage, processing, sharing and reporting of data that is collected which is beyond the ability of traditionally used software tools & hardware infrastructure
- The applications of Big Data in the financial services industry ranges across effective customer segmentation, accurate regulatory reporting

What is Big data?

- Large volume of data
- Existing tools were not designed to handle such a huge data

Gigabyte	•10^9=1,000,000,000
Terabyte	•10^12=1,000,000,000,000
Petabyte	•10^15=1,000,000,000,000,000
Exabyte	•10^18=1,000,000,000,000,000,000
Zetabyte	•10^21=1,000,000,000,000,000,000,000

Note : All the above exponents are in bytes

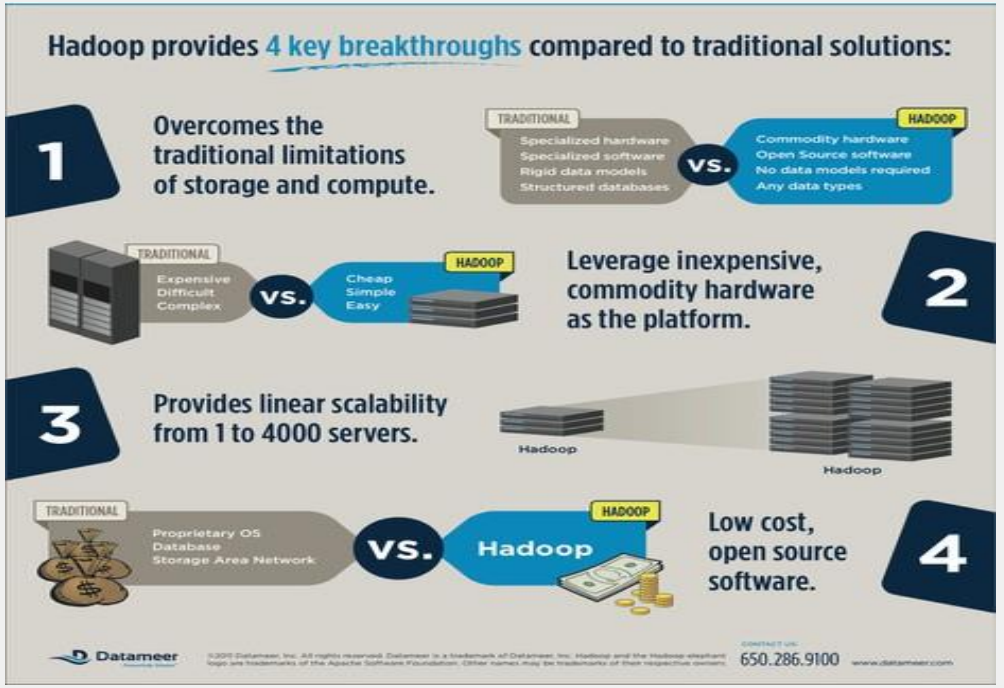
What is Big data?

Organizations these days are swimming in an increasing sea of data that is either too voluminous or too unstructured to be managed and analyzed through traditional means. '**Big data**' is the term people have begun using as a saying for smarter and more insightful data analysis.

Hadoop is one of the technology to enable big data was developed by "Doug Cutting". It started out as a subproject of Nutch by Doug Cutting. Hadoop was greatly boosted by Nutch's scalability. It was enhanced by Yahoo! And became Apache top level project. Hadoop is named after a "toy elephant" that Doug Cutting's kid used to play with.

Big Data – Journey

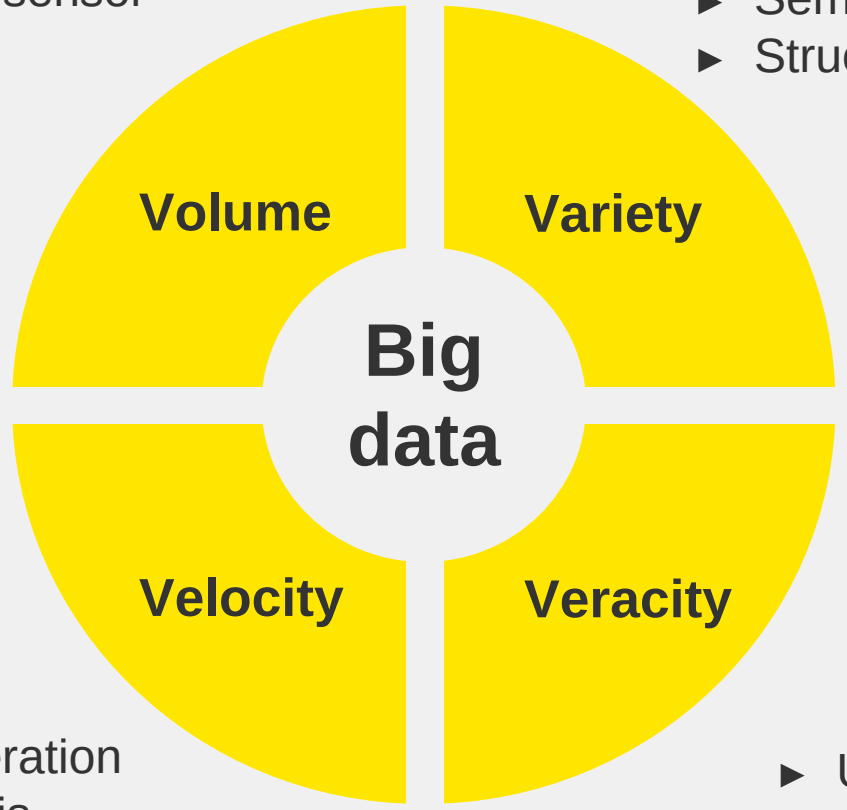
1990	2010	Hadoop
► Store 1,400MB ► Transfer speed of 4.5 MB/s ► Read the entire drive in ~ 5 minutes	► Store 1 TB ► Transfer speed of 100MB/s ► Read the entire drive in ~ 3 Hours	► 100 drives working at the same time can read 1 TB of data in 2 minutes



Big Data is typically characterized by the four “V’s”

- Click stream
- Active/passive sensor
- Log
- Event
- Printed *corpus*
- Speech
- Social media
- Traditional

- Unstructured
- Semi-structured
- Structured



- Speed of generation
- Rate of analysis

- Untrusted
- Uncleansed

Volume

The volume of persistent usable data in the analytics system at any point in time.

Variety

The form and content of data – structured (RDBMS), semi structured (social media) or unstructured (text/documents)

Velocity

How quickly the analytics system process the data to create insights

Veracity

The degree to which data is accurate, precise and trusted

State of the Big Data Market

As the big data market continues to expand (reaching more than $11 billion at the end of 2012), there are a number of factors that drive and limit its growth.

Growth drivers

- Increased awareness of benefits
- The maturation of “big data” software
- Sophistication of professional services
- Increased investment from government and web properties

Adoption barriers

- Lack of analytics specialists
- Inability to organize big data staff
- Resistance to replace gut instinct decision making
- Lack of best practices

Information Management Framework (IMF) for Big Data

Information Management Framework (IMF)

Data governance

- Organizational model
- Enablement
- Standards and policies
- Processes and procedures

Data quality

- Profiling/analysis
- Cleansing
- Controls
- Enrichment/ enhancement

Data usage

- Management reporting
- Analytics (OLAP)
- Data mining
- Quantitative analysis
- Scorecard/dashboards
- Alerts/notifications

Data management

- System of records
- Operational data stores
- Data warehouse/ data marts
- Data movement (ETL/EAI/EII)
- Data protection
- Metadata management
- Reference data management
- Master data management

Architecture

- Conceptual
- Logical
- Physical /technical
- Design patterns
- Services
- Standards

Impact

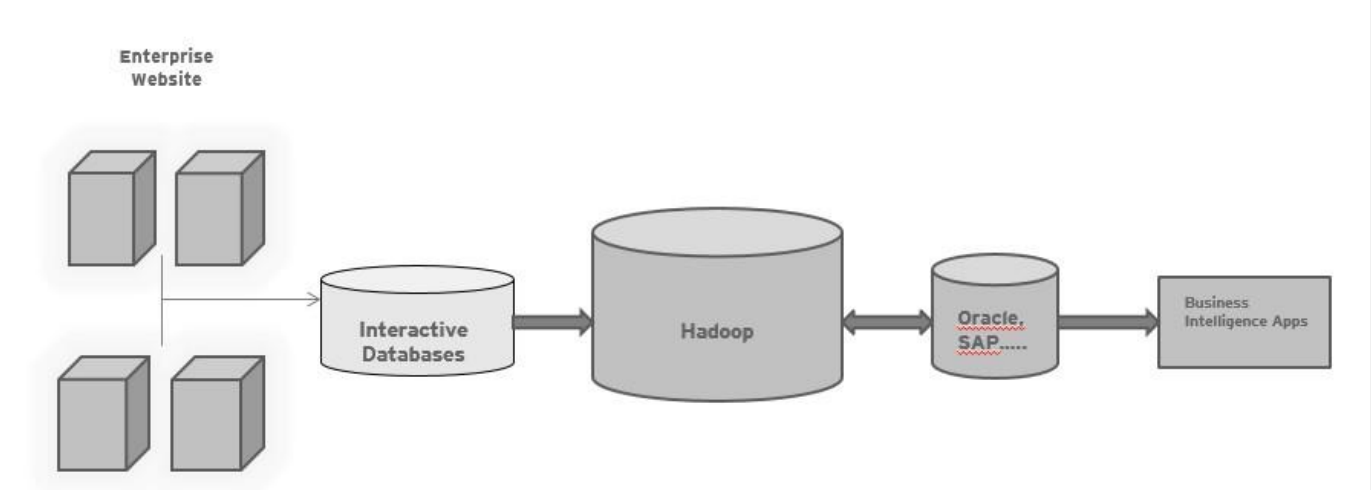
- Organizational model may expand to include more people as the variety of data increases. Policies and procedures may need to be updated to handle new data types.
- New processes may need to be put in place to profile, cleanse, and enrich new data types
- New ways to mine data and generate reports will need to be developed and implemented in consultation with business.
- Volume, Variety, Velocity and Veracity of data will change how data is stored and exchanged.
- Standards and services will need to be updated along with how logical and physical architecture is viewed.

Technology & Solution

Big Data processing platform

Hadoop

- Hadoop is a massively scalable storage and batch processing system
- Hadoop augments them by offloading the particularly difficult problem of simultaneously ingesting, processing and delivering/exporting large volumes of data
- The results can be delivered to any existing enterprise system for further use independent of Hadoop



Flexible enough to be able to work with multiple data sources

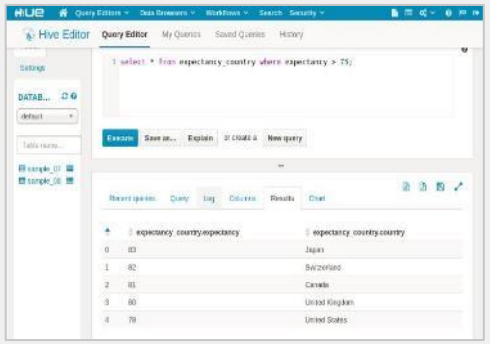
Reading data from a database in order to run processor-intensive machine learning jobs

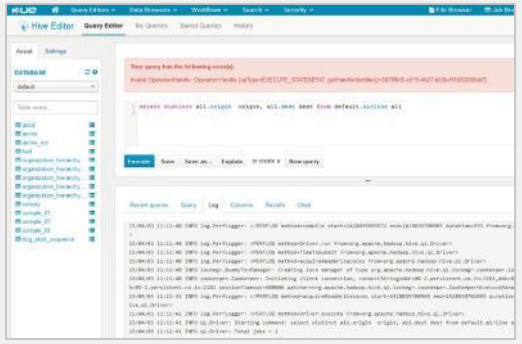
Able to handle large volumes of constantly changing data, such as location-based data from weather, web-based or social media data.

Analytics can be done efficiently & integrated within the platform

Hive

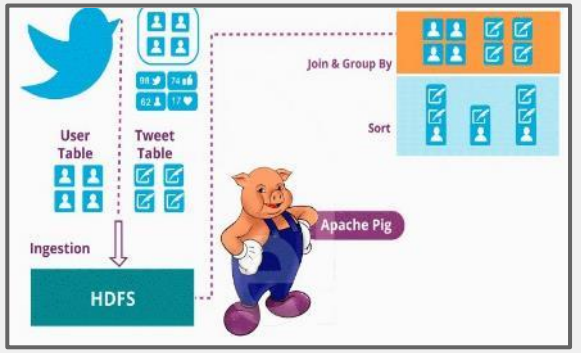
Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It amplifies the reach of Hadoop, making it more familiar for BI users.

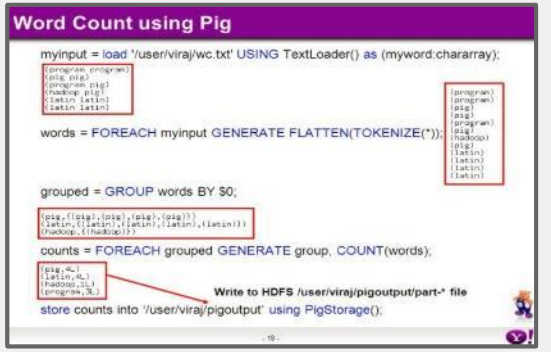




PIG

PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language.



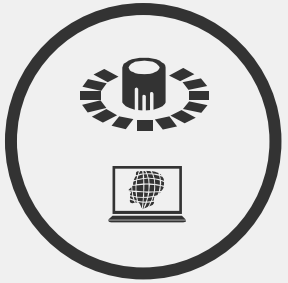


Big Data Architecture

Coexisting Data Profiles:
Data warehouse IT investment is not lost

Data Warehouse

- Cleansed, enriched, matched data
- Structured data analysis
- Analytics-at-rest
- Produces insight with known and stable measurements
- Defined based on pre-determined corpus of questions
- Inflexibility in structure due to rigid data structure design
- Rigorous data quality controls
- Performance envelope constrained due to functional limits
- High cost-per-compute
- High value-per-byte
- Data retained based on perceived business value



Enterprise Hadoop

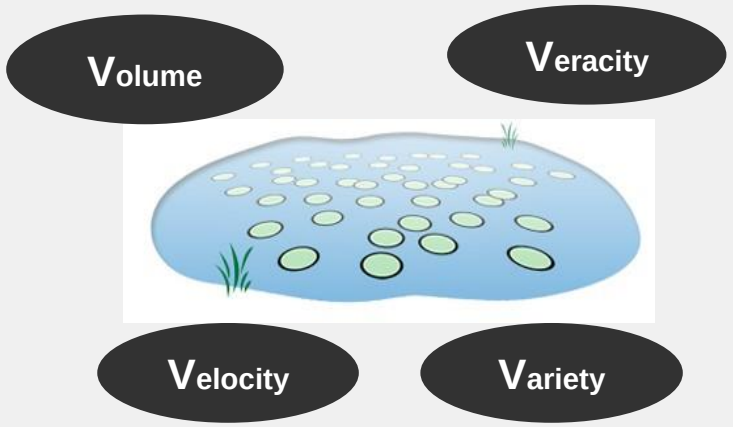
- Analyze all data (Structured, unstructured, semi-structured)
- Inherent data discovery and data value analysis
- Analytics-at-rest & analytics on-the-wire
- Multiple disparate data sources
- Store all data (retain fidelity of transactions, logs, posts etc.)
- Store data in native object format
- Flexible or no data transport encoding
- Low cost-per-compute
- Minimal performance concerns due to massive parallelism

Data lake technology

Provides a capability for the business to access and use the data across the multiple platforms

Data Lake technologies allow:

- Large data volumes to be ‘brought together’ in a flexible format from across many global data sources
- Analysts with the *right to access* can query data that is connected to the global information network
- Data to be synthesised or derived from other data in a controlled and audited manner using analyst-defined business and data rules



“Data Lake”

Hadoop ecosystem

→ Big Data Analytics

Strategically, these would bring significant changes to the operation:

- A single analyst ‘work station’ to examine, query, adjust, explore and report on finance and risk data
- Reduced cycle time for changes to: accessing new and existing data items, applying data quality ‘fix’ rules, deriving missing data from other sources
- Iterative exploration / what if scenarios become possible on global data sets for finance and risk analysts

Use Cases for Big Data

Investment Bank:
Active archive to meet BCBS239 Regulation

Regulation

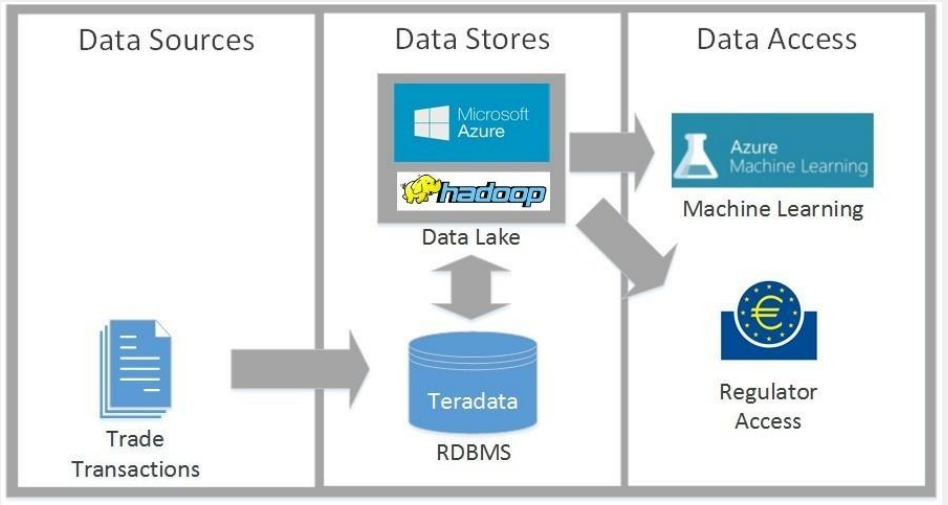
Overview

BCBS239 stipulates that historical data required to meet regulations can no longer be stored on a tape achieve and needs to be hosted on an active archive

10TB of Fixed Income trades currently reside on tape archive which are within the last 7 years. Loading the trades back into the Fixed Income Warehouse would require ~ £3 million in additional storage

Leveraging Azure Cloud storage and importing the data into a data lake is potentially 10x cheaper. An additional benefit would be that the data lake could act as a playpen to deploy technologies such as Azure Machine Learning to perform analytics

Complications currently being worked through include understanding security considerations of deploying to a public cloud



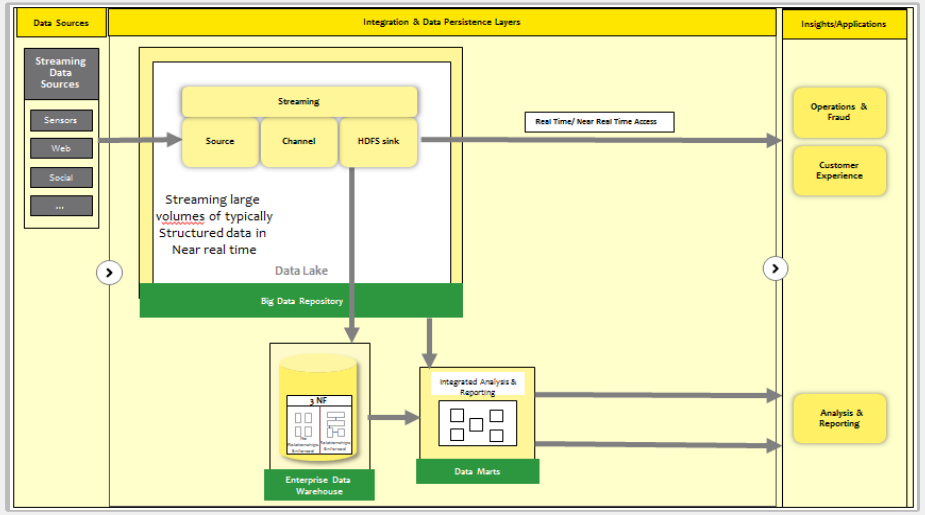
Architecture

- Hadoop hosted on Microsoft Azure Cloud Servers
- Azure Machine Learning to analyse the trade data

EY

Insurer - Telematics:
Investigating Big Data technology capability through POC

EDW Optimisation



Overview
Insurer has realised the growth in Insurance Telematic products (8,000 Telematic policies per month –Aug 2015, total current figure of 70,000 active vehicles – 8% Group sales, 14% group premium).

Challenges include Insurer moving towards a 1 second sensor data availability capacity to match competitors – reduce data latency from third party Telematics providers to end business user, reduce cost of exponentially growing volumes of data, reduce human error and support new customer channels.

Decided to investigate the power of Big Data through a Hortonworks end-to-end Big Data PoC. Moving towards near real-time data access will open Insurer to opportunities such as mobile apps, event notifications, up-selling, GEO fencing and so on.

Business Case	The business case was built around the client remaining competitive in the market, reducing overall storage cost going forward and striving towards advance analytics for pricing and product offering. Cost saving analysis underway. Initial investment will be driven by a Hortonworks PoC. **Total PoC Investment: £59k** Pre-sales/License fee for 12 months = £0.00 4 Weeks Implementation = £33k 12 Months Support – 8 Nodes/2 Clusters = £25k 2 Months Cloud Service = £1k
Current Status	Hortonworks pre-sales/design – Oct 2015 PoC - Mid Nov 2015 Start– Jan 2016 Completion Feb 2016 – PoC Review process begins Mar 2016 – Playback to Technical Governance
Architecture	• Hortonworks Solution – Kafka, HDFS, Hbase, Hive, Spark, Azure Cloud, Tableau and SAS • Data Sources – Spatial Lookup Source and raw GPS location data
Key Stakeholders	Head Architect – Adam Morton (Overseer) CTO – Charlotte
People	Hortonworks providing technical support and learning to client staff through subscription and online learning.
Governance	Technical governance structure and roles for Telematics and Cloud services are being put in place.

Predictive Analytics for Claims

Overview

Claims repudiation ratio is an important factor in measuring Insurer's position these days. Claims processing says a lot about companies performance and the maturity of its value chain with respect to the Industry. Claims scoring has been the latest technique used by the firms to describe a particular claim as soon as first loss of report is filed. Existing Predictive analytic systems are isolated and are "after the fact" analytic systems and many a times single a analytic model cannot describe all the facets of a claim. For example one particular model might be very good in answering "Will the employee join back" after the incident and there is another that is good at predicting the intensity of the claim and the related loss.

Benefits

Effective resource allocation to a claim
Timely information updates to customer
Accurate assessment of duration of claim
Effective estimation of loss and its impact on reserves

Solution

An event driven claims processing can tap into the intelligence of the multiple predictive analytic engines in real time as various events from first loss of report through closing of claim occur within the life cycle of the claim. This could be effectively achieved using modern day technologies like Sybase /Open source streaming engines and powerful capabilities of big data frameworks like Hadoop, Cloudera etc., which could enable the models to run fast over the historical data to provide timely description based on the events happening through the life cycle of the claim

As soon as a claim posting event is triggered a predictive analytic engine gets triggered to analyze the trends based on historical data which helps in identifying severity of the claim, duration of the claim etc. This enables effective management of resources internally to address the claim thus leading to timely information to the customer.

EY

Being A Data-Driven Organization

Adopting big-data technology doesn't make you a data-driven organization

Having lots of reports does not make you data-driven.

Having lots of dashboards does not make you data-driven.

Having lots of alerts does not make you data-driven.

Having a hadoop cluster does not make you data-driven.

Being a data driven organization

Must first understand the definition..

> A data-driven organization is one in which critical business data/insight automatically drives the decisions and actions of your business

> **The current typical process**
>
> Executive make a decision then find data to support it with heavily reliance on “gut feel”

> **Data-driven approach**
>
> **Data tells you not only that a decision needs to be made, but also often tells you what that decision should be**

Being a data-driven organization should start by realizing the importance and criticality of it..

- Use data to make better decisions in your organization
- Collect appropriate data and analyze it in a meaningful fashion
- Improve quality of your services
- Get the data into the hands of the people who need it
- Increase efficiency and effectiveness

Being a data-driven organization, one should see it as an asset instead of liabilities

We largely accept that data is one of the most powerful assets within an organization as it drives decisions and revenue. A best practice is establishing organizational principles to guide the treatment of this enterprise asset.

Data is an Asset	Data is a valuable corporate resource; it has real, measurable value.
Data is Managed	Data is the foundation of our decision-making, so we must also carefully manage data to ensure that we know where it is, can rely upon its accuracy, and can obtain it when and where we need it.
Data is Shared	It is less costly to maintain timely, accurate data in a single book of record, and then share it, than it is to maintain duplicative data in multiple applications.
Data is Accessible	Wide access to data leads to efficiency and effectiveness in decision-making, and affords timely response to information requests and service delivery.
Data is Protected	Data is protected from unauthorized use and disclosure.
Data is Defined	Data is defined consistently throughout the enterprise, and the definitions are understandable and available to all users.
Data is Governed	Each data element has a trustee accountable for data quality and its use within the enterprise.
Data has a Lifecycle	Data is regulated requiring procedures governing what data to keep; what data to discard and critically, how to control what's left.

Organization should start seeing data as the 4th pillar of business to win in Data Driven Organization

Challenges and Opportunities

- Although there is a mixed picture of the level of maturity at financial services companies, all businesses are on the same journey, working towards a greater emphasis on decision-making based on actionable insights generated from their data
- Part of the challenge is a willingness to assess current levels of maturity in order to determine the degree of focus and investment required. The size of the prize makes these difficulties worth confronting.

Data as the 4th Pillar of Business

- For many firms, data is now their fourth strategic "pillar," alongside people, process and technology
- Data insights can be potentially transformative – a crucial source of competitive advantage over industry rivals
- Overwhelming number of firms agree that data is their most valuable strategic asset

Holistic Approach to Data Crucial

- High growth financial services firms are noticeably more aggressive in their investment in data
- Investment in tools and technology is important but investing in people, leadership and change is equally crucial
- Technology by itself is not sufficient to create a data-centric business culture

Invest in data for the upside potential

- Firms focused on fast growth are leading the way on exploiting the upside potential of data
- Fast growth firms key focus in data analytics is in revenue sharing areas, such as marketing and sales
- Lower growth companies are more often focused primarily on compliance-related data management initiatives

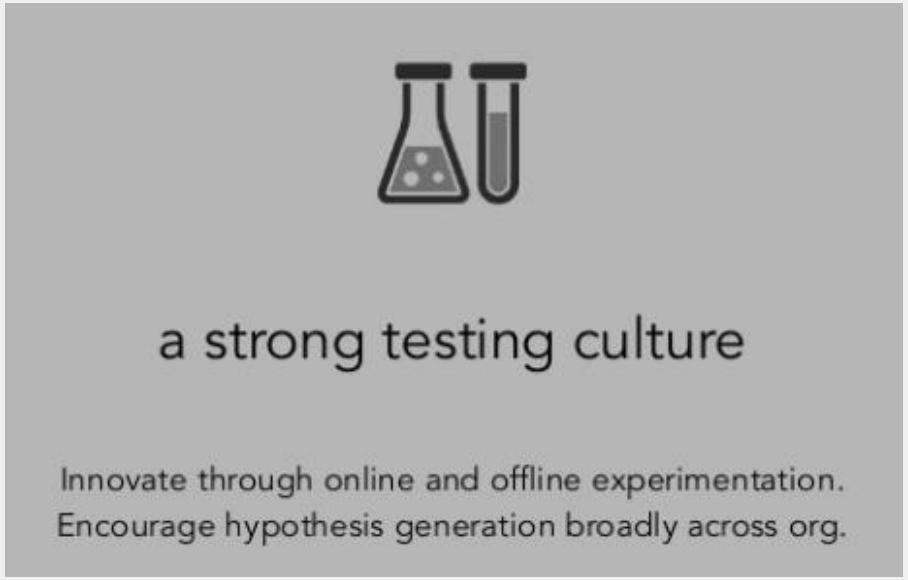
Build Analytics Capabilities

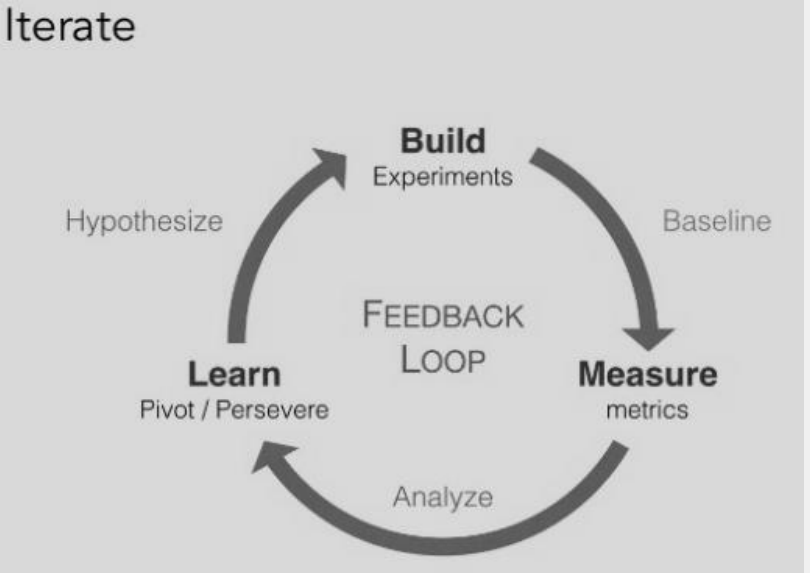
- Data leaders in financial services use cutting-edge analytics to make more key decisions on the basis of data
- Given the importance of extracting value from data there is a clear need to appoint a dedicated leader on data
- A clear vision, leadership on data, investment in people and tools and strong governance are all required in the transformation to a analytics driven, data-centric business

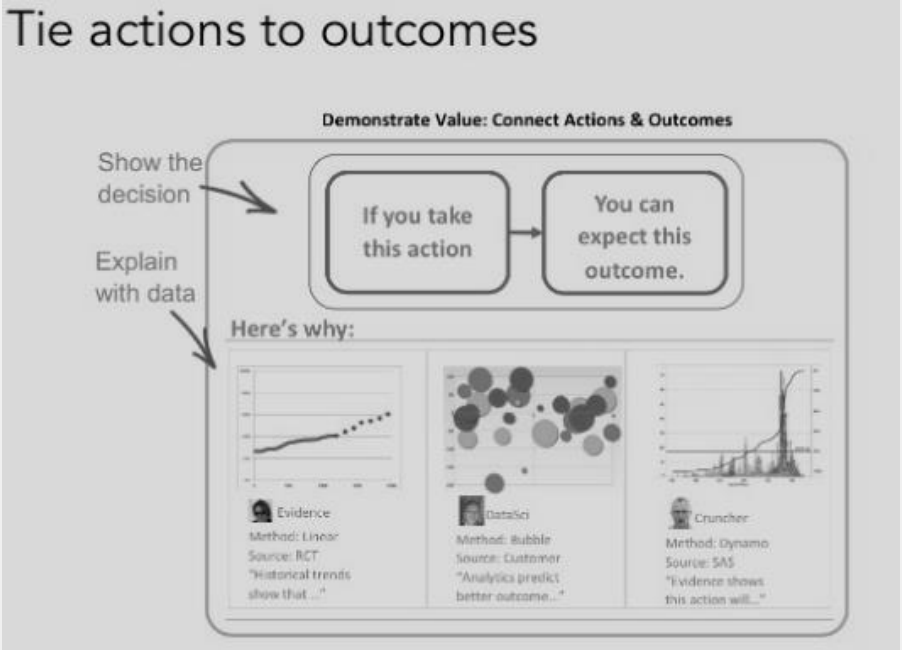
Becoming data & analytics driven requires significant cultural & organizational change

- **78%** agree big data & analytics is changing the nature of competitive advantage
- **12%** describe their analytics maturity as leading
- **66%** are investing $5Mio+ in analytics
- **89%** agree that change management is a barrier to realizing value

Becoming data & analytics driven also means..







Lastly, you don't want to be like this…

Girls Crash into Lake following Bad GPS directions

Thank You

EY | Assurance | Tax | Transactions | Advisory

About EY

EY is a global leader in assurance, tax, transaction and advisory services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for our communities.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. For more information about our organization, please visit ey.com.

APAC No. 00000308

ey.com/id