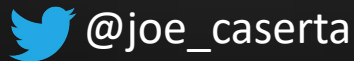
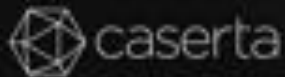
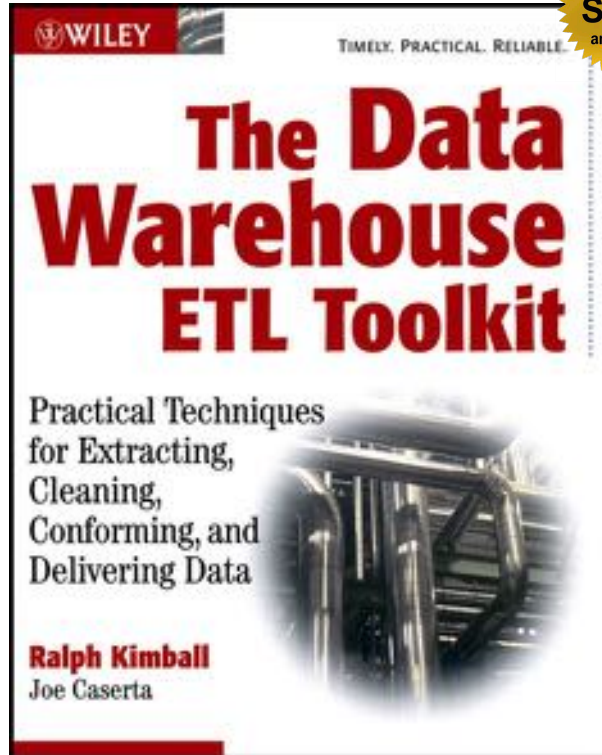




Innovation for Data and Analytics Trends



July 24, 2019



Best Practices and Standards for:

- Collecting and Understanding Requirements
- Planning and Data Design Standards
- Building the Logical Data Map
- Integrating Heterogeneous Data Sources
- Cleaning and Conforming Data
- Data Quality Screens and Their Measurements
- Delivering Dimensions and Facts
- Handling Late Arriving Data
- Metadata Standards and Practices
- Managing, Planning and Leadership
- Implementation and Operations
- Handling Streaming Data



WE PUT YOUR DATA TO WORK

Caserta solves our clients' toughest data and analytics challenges through unrivaled talent and innovation.





Transformative Strategic Consulting

Management
Consulting

Digital
Transformation

Change
Management





Advanced Technical Architecture



Cloud
Engineering

Data
Engineering

Process
Engineering



Implement, Build, Deploy

Business
Intelligence

Data
Intelligence

Artificial
Intelligence



Diverse Domain Expertise

Government

Insurance

Media

Ad-Tech

Finance

Retail

e-Commerce

Healthcare

Energy

Higher-Education

Our Clients

Finance, Healthcare & Insurance



Digital Media/AdTech Education & Services



Retail/eCommerce & Manufacturing



Our Partners



Google Cloud Platform



Microsoft Azure



Why is Data So Important?



Printing Press
1500s



Penny Post
1840s



Telegraph
1850s



Rural Free Post
1850s



Telephone
1890s



Radio
1900s



TV
1950s



PCs
1970s

Every 60 Seconds



Internet
1980s



Web
1990s



98,000+ Tweets



695,000 Status Updates



11 Million instant messages



698,445 Google Searches



168 million+ emails sent



1,829 TB of data created



217 new mobile web users

Social Media, Mobile, Big Data, Cloud
2000s

Chief Data Officer and Chief Digital Officer

Chief Data Officer

- 🔗 Evangelize a data vision for the organization
- 🔗 Provide accountability for data
- 🔗 Innovate ways to use existing data
- 🔗 Enrich and augment data
- 🔗 Support & enforce data governance & security
- 🔗 Monitor and enforce data quality
- 🔗 Set standards for analytical reporting and generate data insights







Chief Digital Officer


- 🔗 Digital strategy and Innovation
- 🔗 Challenge and cannibalize core business
- 🔗 Digital marketing and customer engagement
- 🔗 Digital user experience design
- 🔗 Customer-centric service innovation
- 🔗 All mobility solutions, data management and analytics
- 🔗 Sets standards, own data and analytics platform

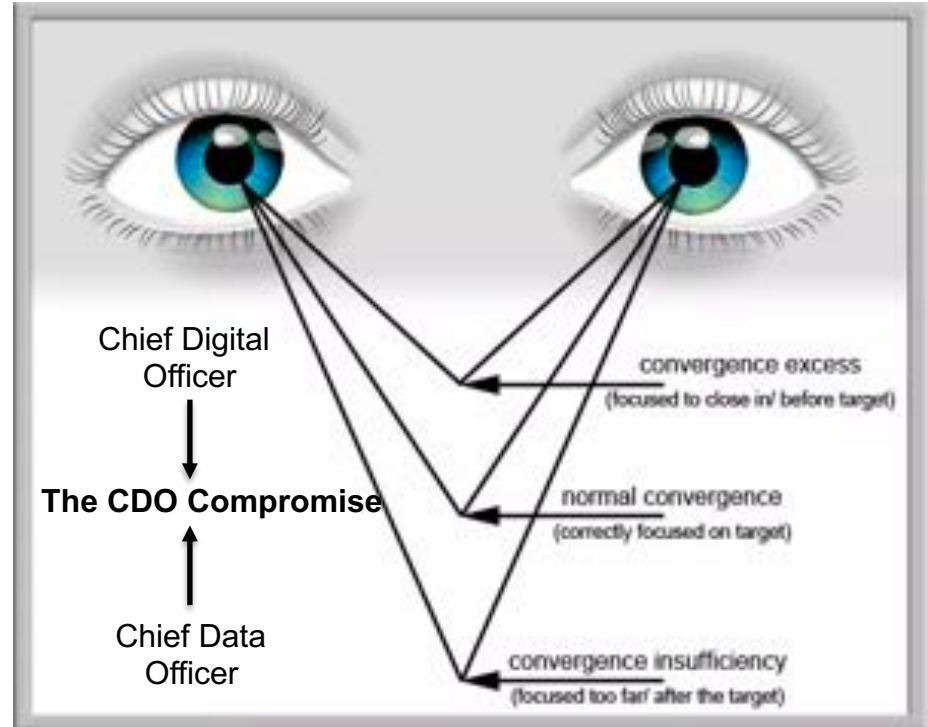
The CDO/CDO Compromise

Digital Focus

-  Time to Market
-  Company Relevance
-  Customer Advocate
-  Generate Revenue

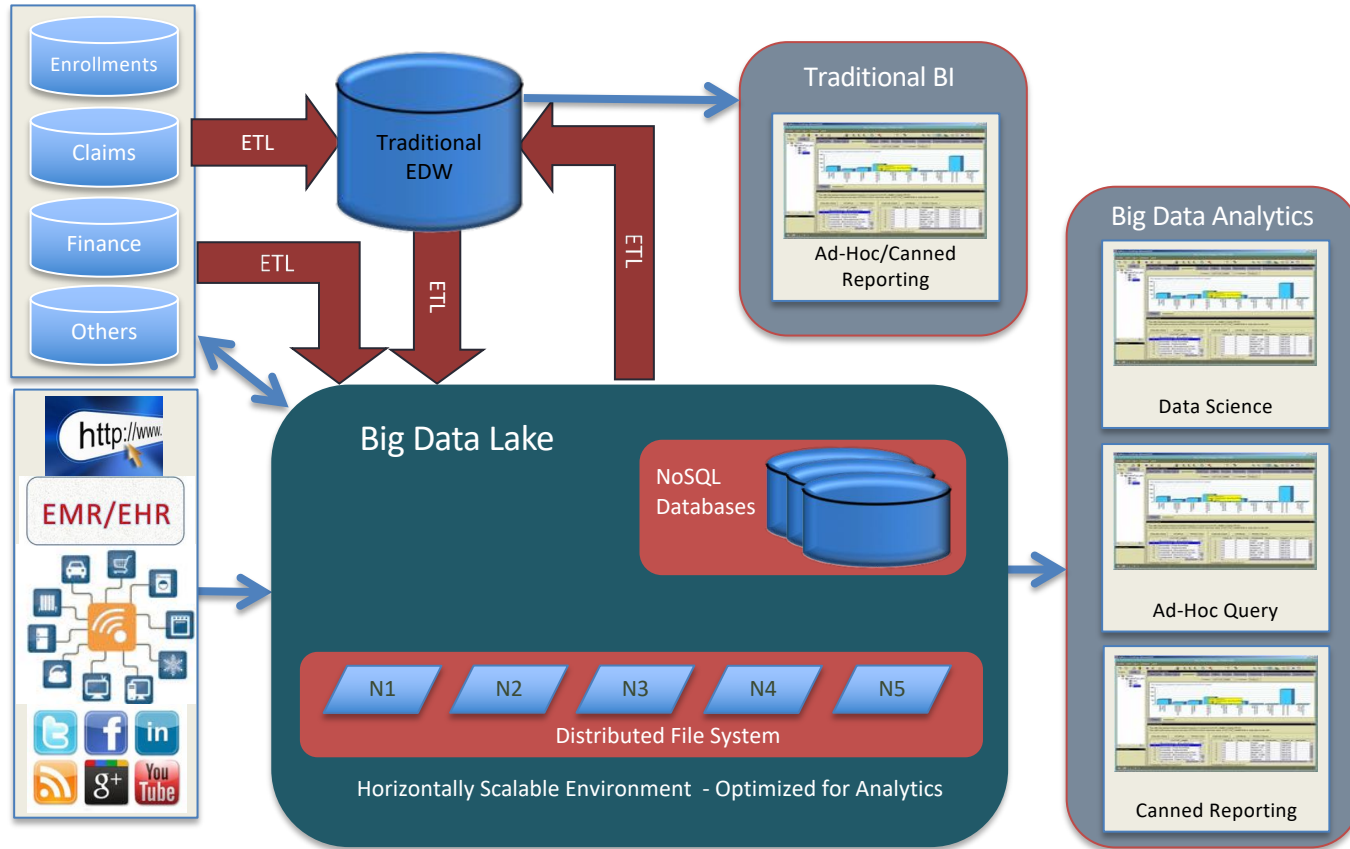
Data Focus

-  Standards
-  Governance
-  Stability
-  Reusability
-  Generate Revenue








The Evolution of Modern Data Engineering



What is This New Data?

-  Alternative data sets are information about a particular company that is published by sources outside of the company
-  An alternative data set can be compiled from various sources such as sensors, mobile devices, satellites, public records, and the internet.
-  In addition to public websites, companies are collecting and crunching data generated by credit card transactions, images of parking lots, customers reviews, etc.



24 Categories of Alternative Data







Why We Care...

...”because instead of supplying a human trader with tips about breaking news, technology sweeps up data from 300 million websites, 150 million Twitter feeds, as well as analyst presentations and FactSet reports for traders—either humans or algorithms—to analyze. It uses natural-language processing to find keywords like company names, and measures when a story is rising up the media food chain, such as from blogs to newswires, to indicate that it may be important enough to act on.”

- John Detrixhe, Future of Finance Reporter, Quartz



Alternative Data Facts

-  Analytics can save traditional money managers time by sifting through news and data on their behalf
-  Getting as much data as possible into machines running algorithms is becoming common practice
-  News and data companies like Bloomberg and Thomson Reuters now include alternative data in their offerings
-  About 75% of financial companies already use social media and social-driven news feeds to inform investing decisions

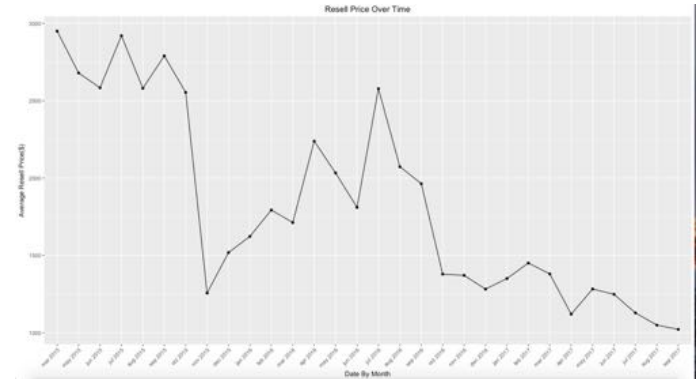


Use Case #1 – Jet Tracking

- ❖ Monitor where portfolio company's private jets are flying and meeting other jets
- ❖ Monitor the change in flying patterns between different companies and/or airports
- ❖ Send alerts to the analysts when something different or unusual is discovered
 - ❖ Cisco flights to Carlsbad, CA, home of semiconductor company Luxtera
 - ❖ Oct 2018 Cisco announced plans to buy Luxtera for \$660 million
 - ❖ Executives from HCA Health, the largest hospital operator in the U.S. tracked flying to Mission HQ in Asheville, NC eight times since May.
 - ❖ August: HCA Health, is buying non-profit company Mission Health for \$1.5B.

Use Case #2 – Price Scraping

- Hidden data in plain sight. Scrape prices of products of portfolio companies to detect trends in pricing and predict sales.
- Create web scraper with BeautifulSoup library for Python. Selenium for automated testing
- Find retail stores with products of interest. For example, sneakers are sold at Footlocker, Finishline, Adidas Stores, etc
- Plot average price of products. Here is a trend of Adidas Yeezy Sneakers
- Analyst determines if Yeezy line is leveling off and the value of Yeezy sneakers are depreciating



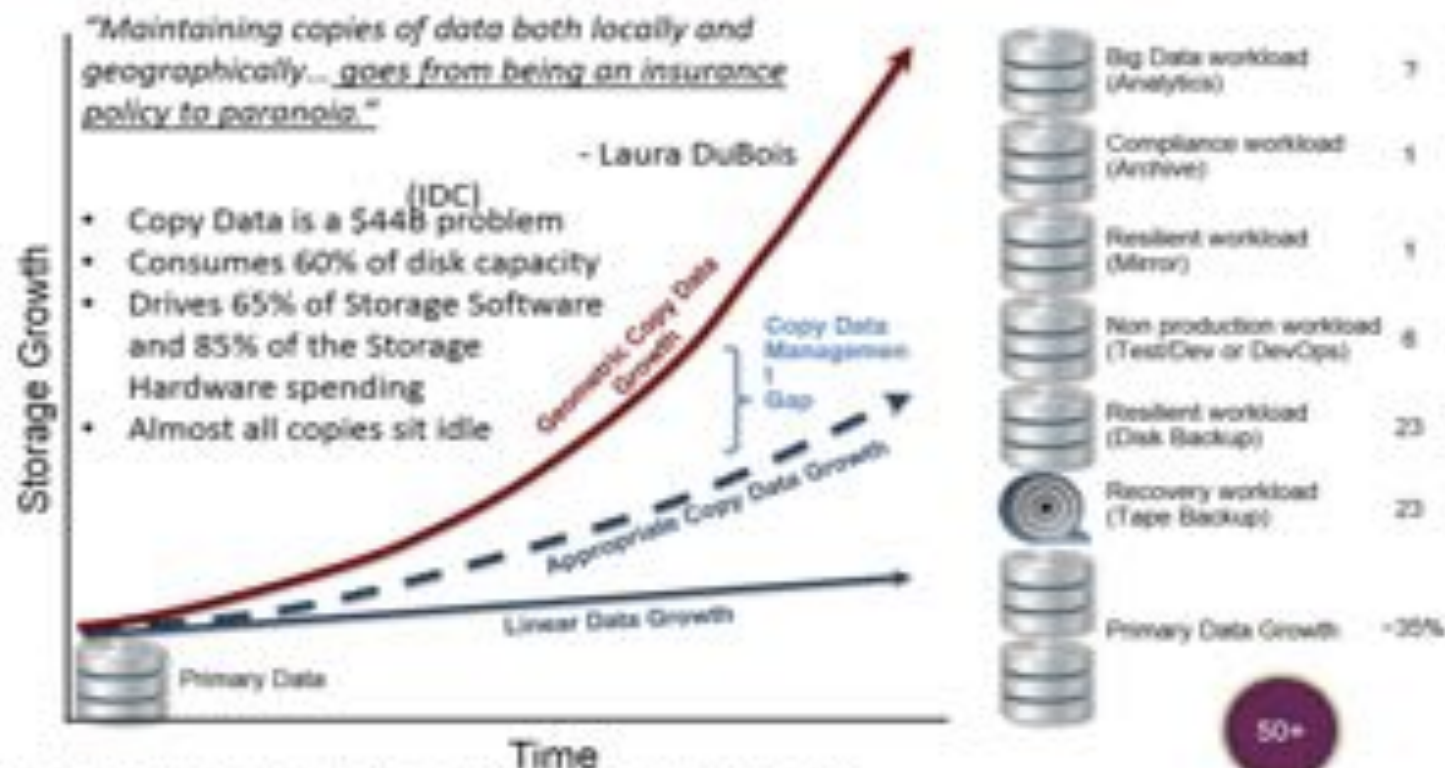
Alternative Data Providers

Source: alternativedata.org



Today's IT Challenge: Drowning in a Deluge of Copy Data

Result: Increased cost and complexity and no additional business value



The Data Sprawl Issue

- There is one application for every 5-10 employees generating copies of the same files leading to massive amounts of duplicate idle data strewn all across the enterprise.
 - Michael Vizard, ITBusinessEdge.com
- Employees spend 35% of their work time searching for information... finding what they seek 50% of the time or less.
 - "The High Cost of Not Finding Information," IDC



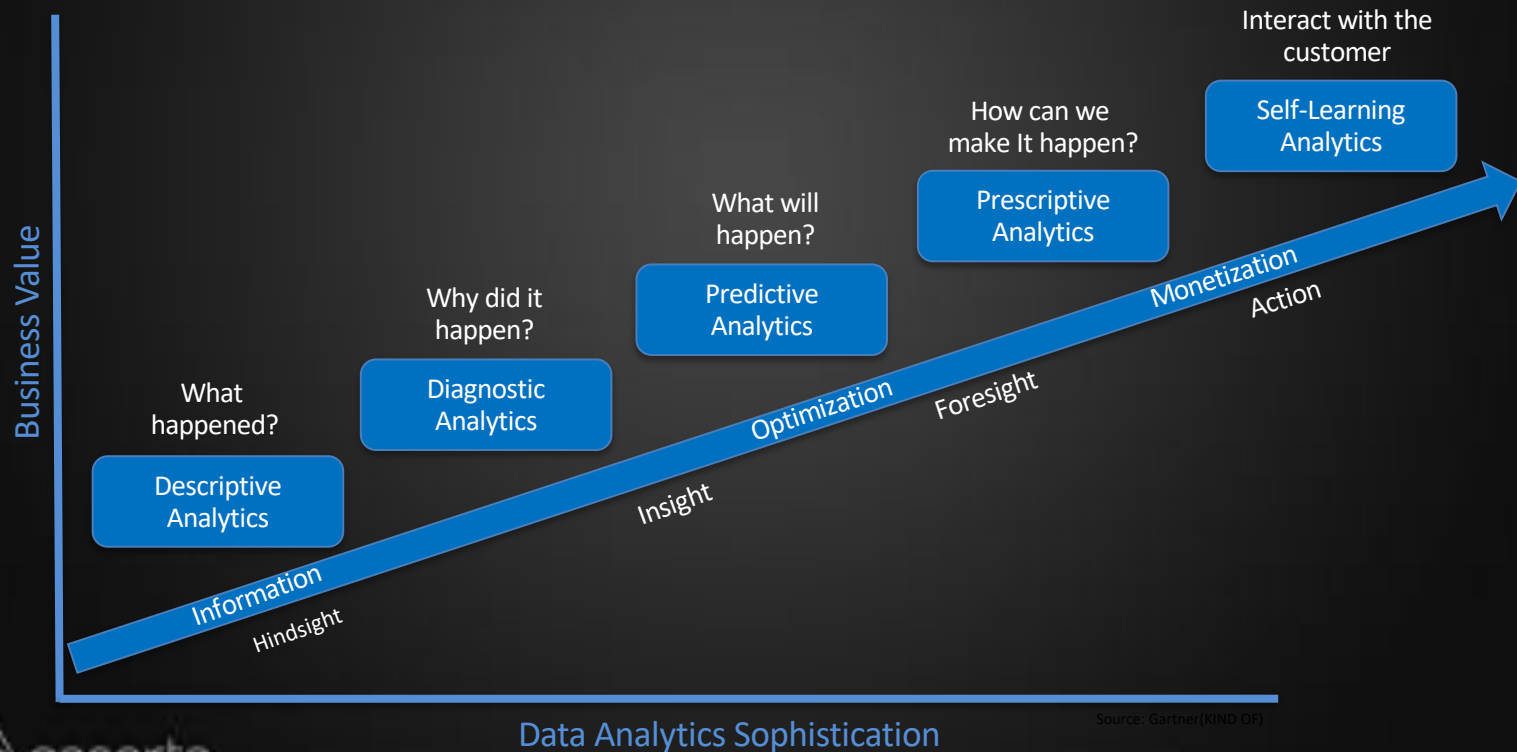
Dealing with the explosion of data sources and data types is an organizational priority.

The **need to simplify information** is driving significant change in organizations: **43 percent** of all organizations currently are making changes to how they design and deploy information and another **37 percent** are planning to make changes.



Evolution of Data Analytics

Reports → Correlations → Predictions → Recommendations → Artificial Intelligence



Source: Gartner(KIND OF)

4 steps of the Industrial Revolution



18th Century: Introduction of Water and Steam powered mechanical manufacturing facilities



19th Century: Introduction of electric powered mass production. First assembly lines in 1870



1970s: Introduction of electronics and IT to advance automation of manufacturing; for programmable controller 1969



Today: Introduction of AI based Cyber-Physical systems



How We've Built Data Warehouses

Design – Top Down, Bottom Up

- Customer Interviews and requirements gathering
- Data Profiling

Create Data Models

- Facts and Dimensions

Extract Transform Load (ETL)

- Copy data from sources to data warehouse

Data Governance

- Stewardship, business rules, data quality

Put a BI Tool on Top

- Design semantic layer

Develop reports



Cracks in the DW Armor - Onboarding New Data

Business: “I need to analyze some new data”

- ✓ IT collects requirements
- ✓ Creates normalized and/or dimensional data models
- ✓ Profiles and conforms and the data
- ✓ Sophisticated ETL programs and quality standards
- ✓ Loads it into data models
- ✓ Builds a BI semantic layer
- ✓ Creates dashboards and reports

IT: “You’ll have your data in 3-6 months to see if it has value!”

- Onboarding new data is difficult!
- Rigid Structures and Data Governance
- Disconnected/removed from business



The New Conversation

- Do we need a Data Warehouse at all?
- If we do, does it need to be relational?
- Should we leverage NoSQL, the Cloud?
- Which platform and language are we going to code in?
- Which bleeding edge Apache Project should we put in production!



The Paradigm Shift

BIG DATA IS NOT THE PROBLEM
IT'S THE CHANGE AGENT.

OLD WAY:

- Structure → Ingest → Analyze
- Fixed Capacity
- Monolithic

NEW WAY:

- Ingest → Analyze → Structure
- Dynamic Capacity
- Ecosystem

RECIPE:

- ✓ Cloud
- ✓ Data Lake
- ✓ Polyglot Data Ecosystem



The Promise of the Data Lake

Technology:

- Scalable distributed storage → HDFS, S3, GCS
- Pluggable fit-for-purpose processing → Spark

Functional Capabilities:

- Remove barriers from data ingestion and analysis
- Storage and processing for all data
- Tunable Governance

Governing Big Data

- Before Data Governance

- Users trying to produce reports from raw source data
- No Data Conformance
- No Master Data Management
- No Data Quality processes
- No Trust: Two analysts were almost guaranteed to come up with two different sets of numbers!



- Before Big Data Governance

- We can put “anything” in the Data Lake
 - We can analyze anything
 - We’re scientists, we don’t need IT, we make the rules
-
- Rule #1: Dumping data into a Data Lake with no repeatable process, procedure, or governance will create a mess
 - Rule #2: Information harvested from an ungoverned systems will take us back to the old days: **No Trust = Not Actionable**



THE DATA SWAMP: CHOOSE YOUR OWN ADVENTURE



Data Governance for Big Data

Organization	<ul style="list-style-type: none">• Add Big Data to overall framework and assign responsibility• Add data scientists to the Stewardship program• Assign stewards to new data sets (twitter, call center logs, etc.)
Metadata	<ul style="list-style-type: none">• Larger scale• New datatypes• Integrate with Hive Metastore,, home grown tables
Privacy/Security	<ul style="list-style-type: none">• Data detection and masking on unstructured data upon ingest
Data Quality and Monitoring	<ul style="list-style-type: none">• Data Quality and Monitoring (probably home grown, drools)• Quality checks not only SQL: machine learning, artificial intelligence• Acting on large dataset quality checks may require distribution
Business Process Integration	<ul style="list-style-type: none">• Near-zero latency, DevOps, Core component of business operations
Master Data Management	<ul style="list-style-type: none">• Graph databases are more flexible than relational• Lower latency service required• Distributed data quality and matching algorithms
Information Lifecycle Management (ILM)	<ul style="list-style-type: none">• Secure and mask multiple data types (not just tabular)• Deletes are more uncommon (unless there is regulatory requirement)• Take advantage of compression and archiving (like AWS Glacier)

The Corporate Data Pyramid

Usage Pattern

Arbitrary/Ad-hoc Queries and Reporting

Munging, Blending
Machine Learning

Organize, Define, Complete

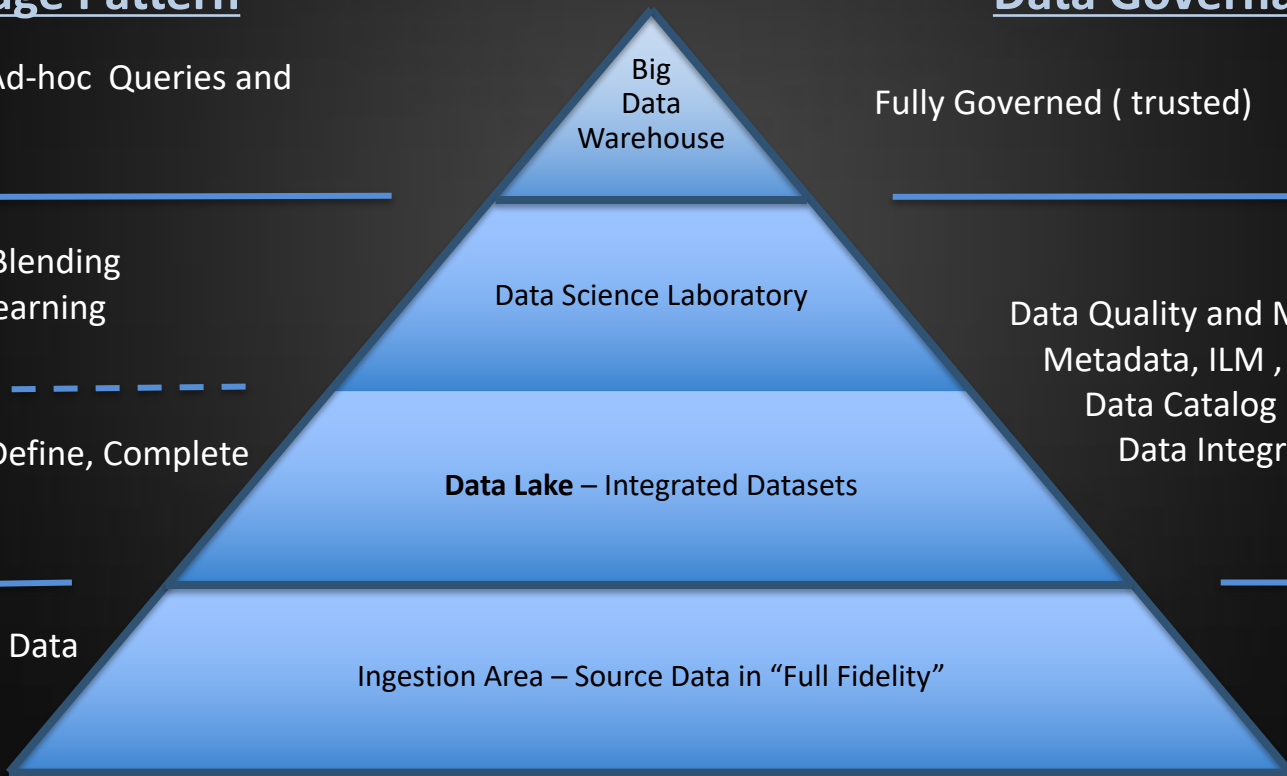
Ingest Raw Data

Data Governance

Fully Governed (trusted)

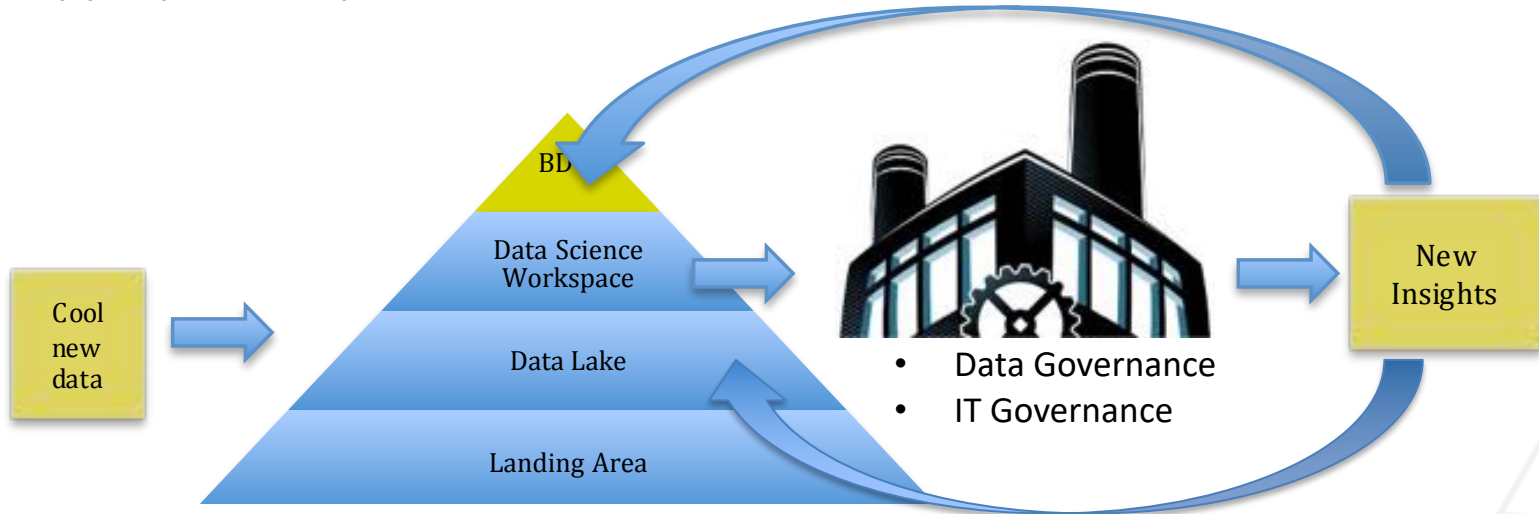
Data Quality and Monitoring
Metadata, ILM , Security
Data Catalog
Data Integration

Metadata, ILM,
Security

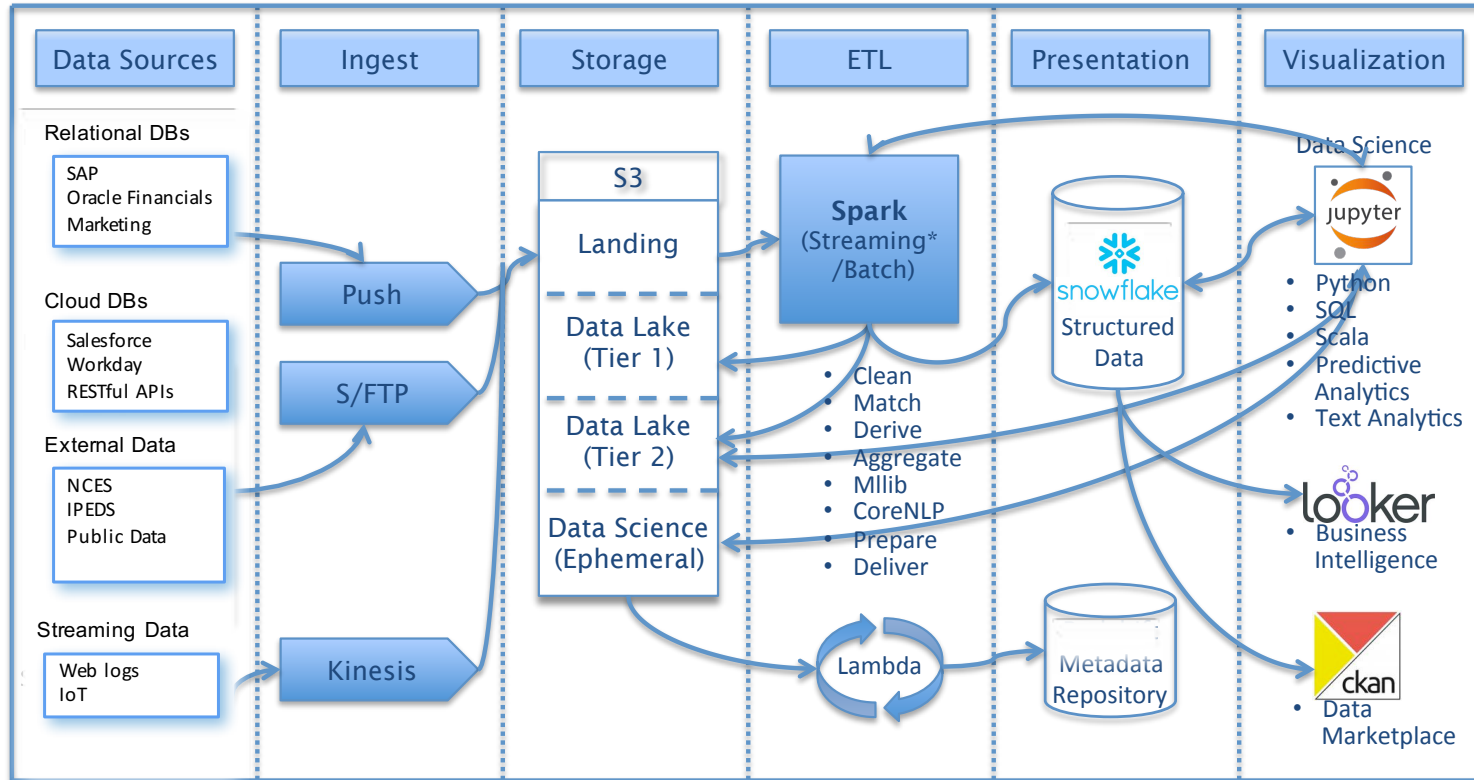


The Data Refinery

- The feedback loop between Data Science and Data Warehouse is critical
- Successful work products of science must **Graduate** into the appropriate layers of the Data Lake



Data Ecosystem Reference Architecture (AWS)

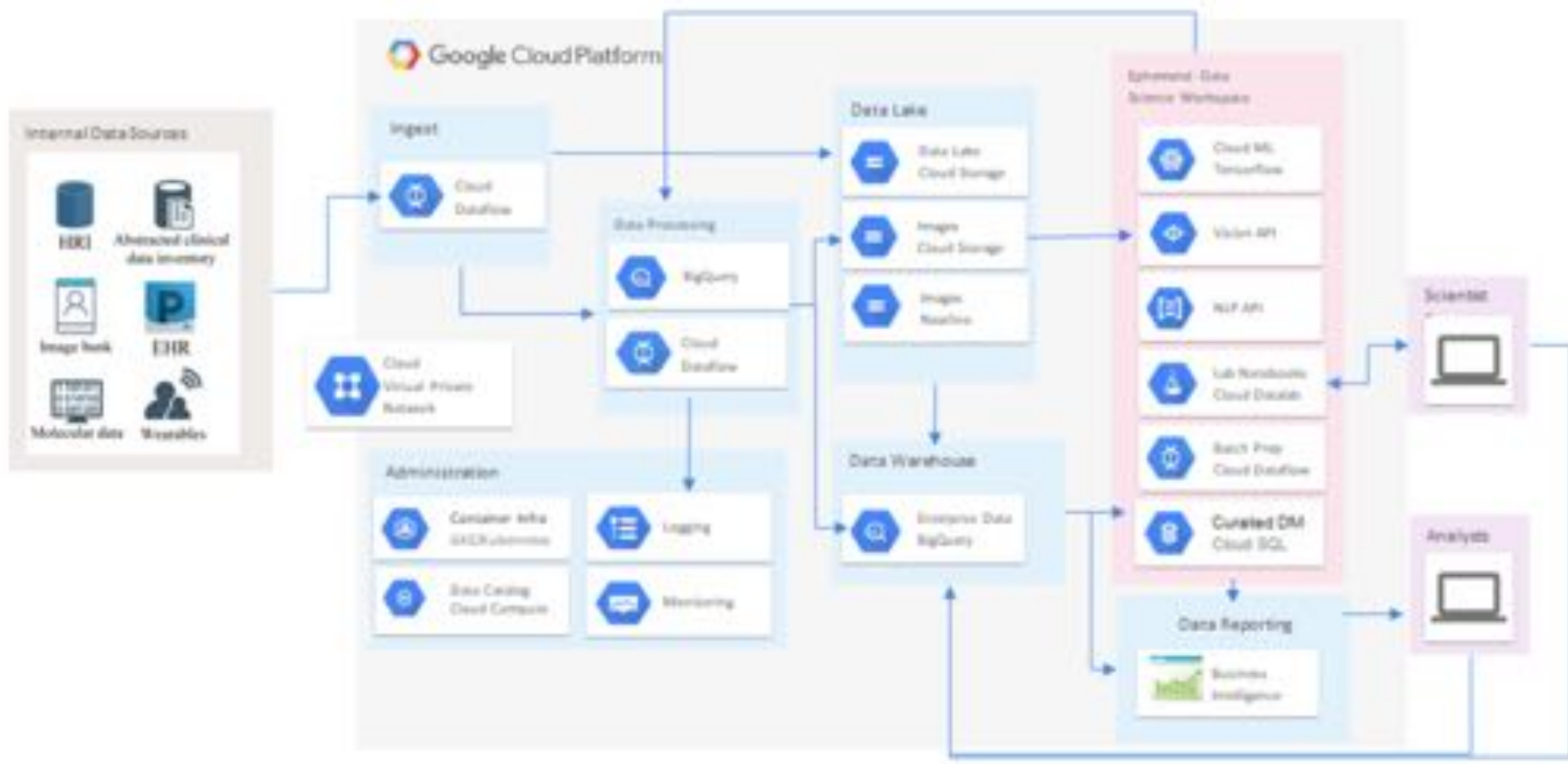


The Data Ecosystem on the Cloud

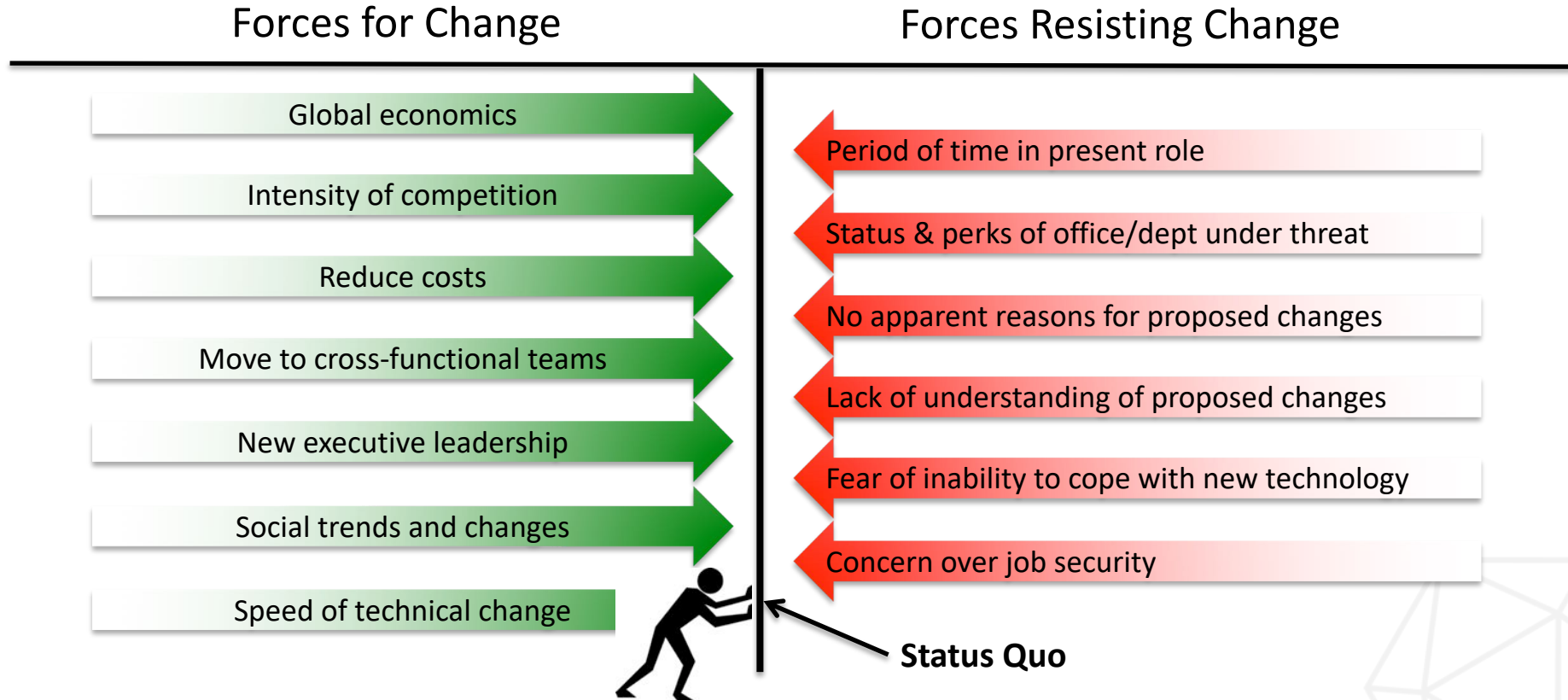
Cloud Component	AWS	Google	Microsoft
Scalable distributed storage	S3	GCS	Azure Storage
Pluggable fit-for-purpose processing	EMR	DataProc	HDInsight
Compute Services	EC2	GCE	VMs
Consistent extensible framework	Spark	Spark	Spark
Dimensional MPP Data Warehouse	Redshift/ Snowflake	BigQuery	Azure SQL Data Warehouse
Data Streaming	Kinesis	PubSub	Azure Stream
Common Interface	Jupyter	DataLab	Azure Notebook
Machine Learning	SageMaker	TensorFlow	ML Studio



Data Ecosystem Reference Architecture (GCP)



Moving the Status Quo – Change Management




It takes a Village!

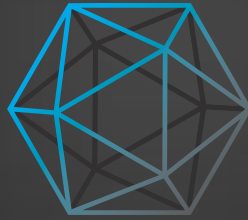


Lessons Learned from the Field

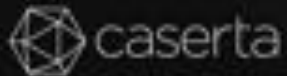
- 🔼 Data is becoming 80% exploration and 20% production. A lot of effort will be thrown away. This is normal. Embrace it.
- 🔼 Because it is mostly exploration, things WILL change and change OFTEN. Architect for rapid change.
- 🔼 Business rules usually get complicated and evolve, your ETL processes must be dynamic and metadata driven.
- 🔼 Deliver the smallest release you can and deliver, often; keep everyone focused on what is currently important. No amount of pre-development meetings will deliver as much knowledge as having your users starting to use the data. Be Agile!



Joe Caserta
President, Caserta
joe@caserta.com
 @joe_caserta



caserta



Practical Use Cases



The New York Times

How completely reengineering a legacy data warehouse to cloud-based technology unleashed analytics to **drive growth**.

The New York Times

CHALLENGE

The New York Times was on a tight deadline to migrate all systems from legacy on-premises data ecosystems to the cloud. The team tasked with the daunting project was lacking the necessary resources and expertise to get the job done right and on time.

SOLUTION

Caserta's expert consultants interfaced with the team at the NYT. We completely reimagined a cloud architecture that leverages 100% cloud-based systems. Caserta reskilled the team at the NYT to ensure a smooth transition.



The New York Times


BUSINESS OUTCOME


Caserta successfully helped The New York Times reengineer all legacy systems to the cloud by the mandated tight deadline.


Total cost of ownership for analytics was reduced. Reports to stakeholders are now delivered faster and easier. This new architecture empowers the business with speed-of-thought analysis in order to gain new insights and fuel growth.




The New York Times

 This is not a model, this is the Director, Analytic Systems shutting down the servers in the data center for good!


 Entire Data Ecosystem re-engineered and replaced with Google Cloud Platform components

 The Data Center shutdown was an interdisciplinary concerted effort across the entire enterprise


 Primary reasons were to eliminate efforts towards:


-  Infrastructure engineering

-  Permits

-  Power systems


-  Generators

-  Conduit and cabling

-  Lighting protection

-  HVAC

-  Fire suppression

-  Managed Services

-  Real Estate



“Caserta was able to integrate with our processes, project management and build out the architecture that Caserta developed. Caserta spent a lot of time making sure that the hand off between their team and ours was successful when the project was over.”

Matt Digan
Executive Director of Data Engineering
The New York Times





AARP

How architecting an analytics-driven marketing strategy helped to **boost retention** and to **acquire new members**.



CHALLENGE

AARP was struggling to collect and analyze the large amounts of data available to them on member profile, transaction, and behavior information that would give them insight into member behavior. The organization does not have in-house technical teams capable of building the necessary analytics infrastructure.

SOLUTION

AARP partnered with Caserta to architect and build a cloud-based data lake that combines data collected from all available channels and fed to an analytics platform.

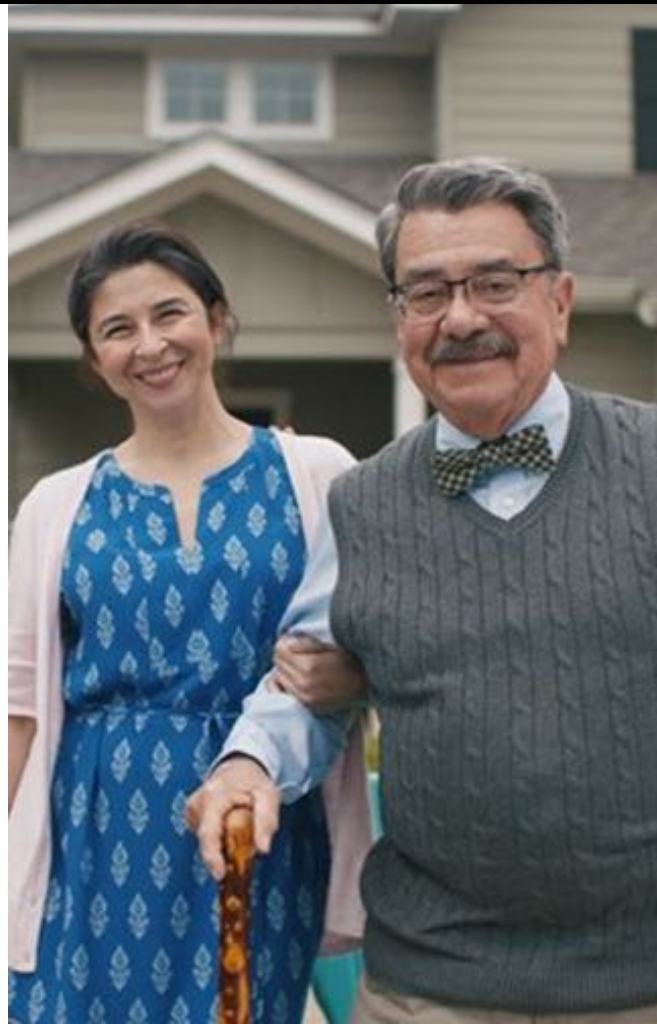




BUSINESS OUTCOME

The new unified data lake enabled business stakeholders to gain a high-resolution look into member behavior. The new data analytics platform allows unrestrained data analysis, which enables the business to create more personalized experiences and boost member retention.

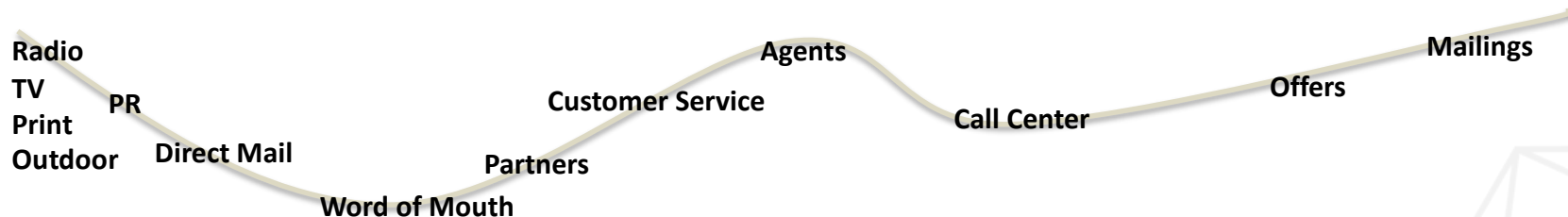
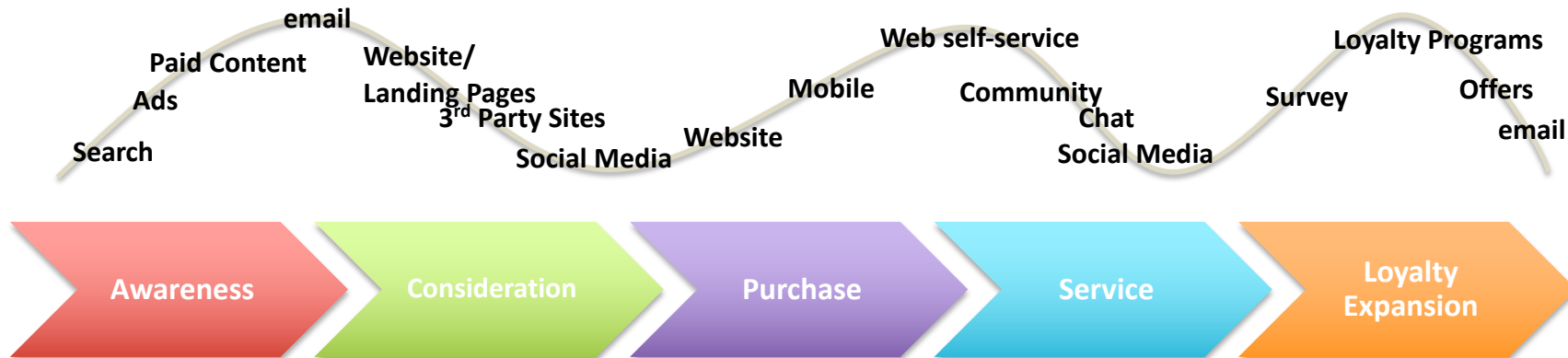
The new analytics platform is helping AARP on their goal to reaching 50M members by 2020.





CUSTOMER JOURNEY

Digital Touchpoints



Physical Touchpoints

Modern Attribution Modeling

What Works?

Isolated

Analyze effectiveness of single touch point type



100%

Rules-Based

Assess correlations between interactions based on domain expertise



33%

33%

33%

Statistically Driven

Detect interactions based full customer journey, determine success path with data-driven model



27%

49%

24%

How do we know?

- Dimensional data warehouse
- Ignores bulk of customer journey
- Undervalues other interactions and influencers

- Subjective
- Assigns arbitrary value to each interaction
- Lacks analytics rigor to determine weights

- ✓ Looks at full behavior patterns
- ✓ Consider all touch points
- ✓ Can apply different models for best results
- ✓ Use data to find correlations between touch points (winning combinations)

“Our unified data lake and cloud-based analytics platform that Caserta built have enabled us to gain unprecedented insight into member behavior thus enabling us to improve the general member experience. This has helped us retain existing members as well as increase new memberships.”

Bill Gale
Vice President, Information Strategy and Infrastructure
AARP



VÄRDE

How harmoniously integrating many data sources into a single cloud data warehouse delivers **high-value investment insights that fuel growth.**

VÄRDE

CHALLENGE

As alternative sources of data come in many types and formats, data was challenging to analyze in a unified view. Värde's analysts were struggling to gain a complete picture of the data. New sources could not be added, as each additional source of data would need to be manually added and cleansed, thus stifling scalability.

SOLUTION

Caserta helped Värde go from an onsite platform that was architected for individualized reporting to a cloud-based data warehouse solution that is both extensible and flexible to support the internal and external demands of the business.



VÄRDE

BUSINESS OUTCOME

Värde's new cost-effective cloud-based unified data warehouse enabled analysts to get a better picture of data and make better investing decisions and fuel growth.

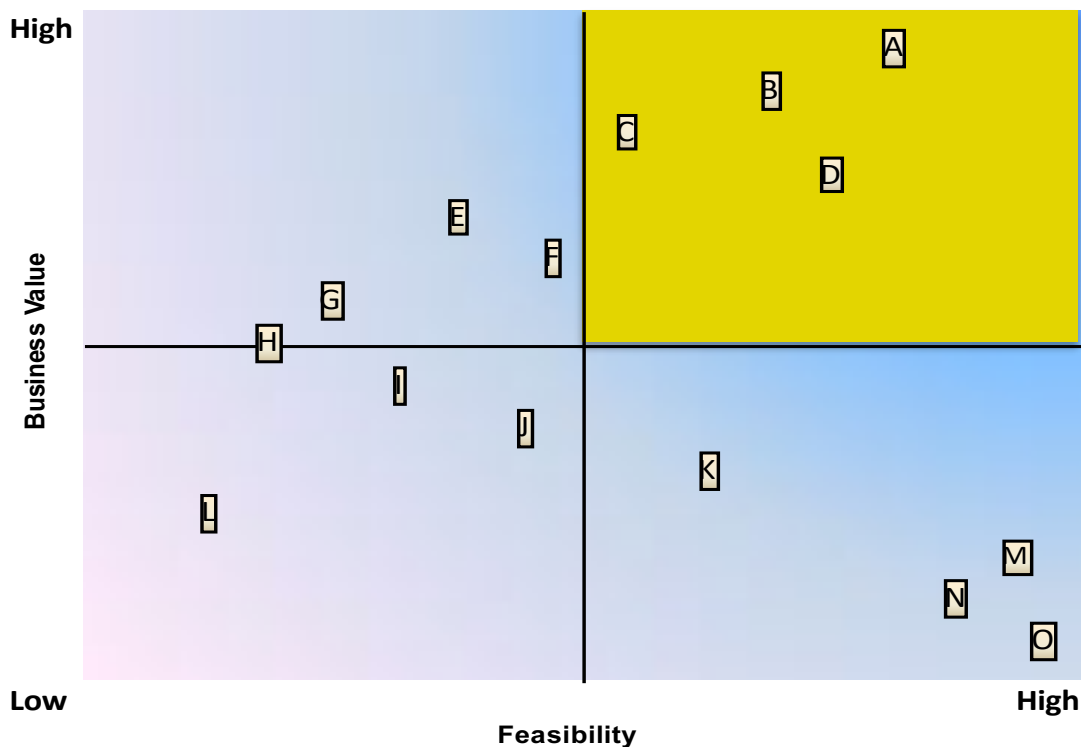
The architecture of the data warehouse enabled the integration of new and evolving data sources to address current needs and provides a foundation platform for future growth.





Proven Methods & Techniques - Feasibility Matrix

The feasibility matrix illustrates relative business value versus technical feasibility of each strategic opportunity identified. The upper right quadrant represents areas with the highest business value and highest technical feasibility.



Strategic Opportunity	
A	Reporting and Analytics Re-platform
B	Relationship Intelligence - Internal and Factset
C	Corporate Alignment & Data Governance
D	Search Platform / Deal Screen
E	Relationship Management - Other Sources
F	Financial Data Automation
G	Investor Reporting Automation/Portal
H	Customized Opportunity News
I	Collaboration Platform
J	Portfolio Management
K	eFront Financial Data Trend analysis
L	Portfolio Company Vendor Operational Data
M	Employee Reporting
N	Portfolio Company Legal Capitalization Data
O	Compliance 11 Automated update from SF

Business Value: Determined by the number of requests, rank of requestors, and evaluation of business priorities during approximately 30 interviews.

Technical Feasibility: Determined by technical complexity, data availability and collaborative The Company / Caserta Concepts assessment.



Sample Analytics Platform Roadmap

P1 – P4 includes:

- Business requirements
- Data Catalog
- Business Intelligence
- Data Integration
 - Source-to-Stage
 - Stage-to-Lake
 - Lake-to-Warehouse
- Full production roll-out of all components

Jan 2018

Recommendation for conceptual design and technologies for:

- Cloud Analytics Platform
- Analytics Platform
- Data Laboratory
- Data Pipeline
- Data Catalog

P1 May 2018

Cloud-based Analytics Platform & Data Warehouse

Re-engineer Solution Architecture for Ad-hoc reporting regarding:

- Analytics
- Reporting
- ODS (as needed)

Build Data Science Laboratory

Collect requirements for Data Science Laboratory POC

Alation Data Catalog Installed and configured

P2 Aug 2018

Fully Functional Solutions:

Security Master
Reference Master

Salesforce
Implement Data Laboratory POC

Hydrate Data Lake with Alternative Data

- Bloomberg
- Eagle Alpha

P3 Oct 2018

Fully Functional Solutions:

Middle Office
Treasury DB

Source System and Data Gap Analysis

Mortgage DB
Real Estate

P4 2018+

Fully Functional Productionalized Solution

Plan refactoring of Source-to-Stage data Feeds to read new sources as needed

Plan On-Prem Sunset strategy including migration of all data integration, reporting and dashboards to new Data Platform

“Caserta’s considerable experience manifests itself in quick and effective approaches. Their analyses are focused and yield helpful recommendations in a very short amount of time. Second, the team’s technical and engineering skills are wide-ranging.”

Steve Stryker
CTO
Vardë Partners





Media & Entertainment



Media & Entertainment

CHALLENGE

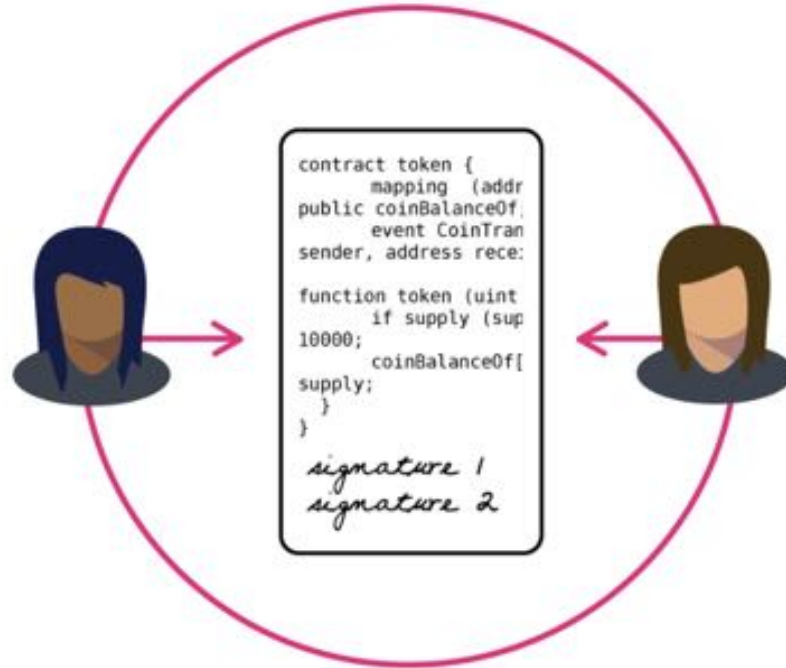
- Content is the heart of everything
- Sales Team looks at rights and availability
- The 7 Factors of a Deal Point....
- Jaguar is the source for creating all contracts
- Jaguar feeds the Rights Data Mart
- Embarking on a project to move Rights Data Mart to Snowflake

SOLUTION

Caserta helped design and architect a a cloud-based solution that combines Blockchain and Artificial Intelligence to modernize the Rights Management ecosystem.



Why Blockchain

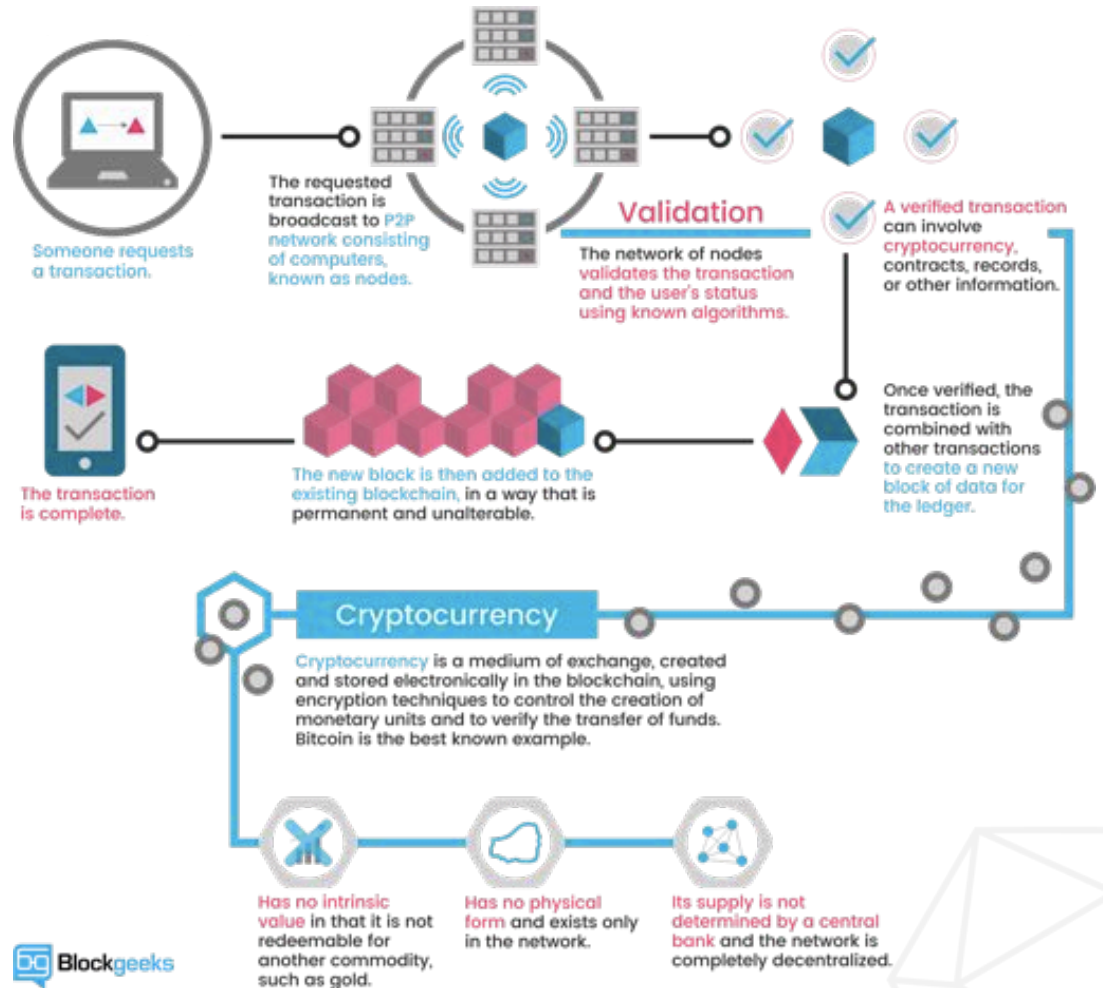


Blockchain is a Distributed Database

- Information held on a blockchain exists as a shared — and continually reconciled — database.
- The blockchain database isn't stored in any single location, meaning the records it keeps are truly public and easily verifiable.
- No centralized version of this information exists for a hacker to corrupt.
- Hosted by millions of computers simultaneously, its data is accessible to anyone on the internet.
- 'Miners' collect all of the transactions made during a set period into a list, called a block. It's the miners' job to confirm those transactions, and write them into a general ledger.

How it Works

- A self-auditing ecosystem. The network reconciles every transaction that happens in ten-second intervals.
- Each group of these transactions is referred to as a “block”.
- Every Node connected to the blockchain network uses a client that performs the task of validating and relaying transactions.
- Each Node has an incentive for participating in the network.
- In fact, each Node is competing to win Bitcoins or Ether (or other exchangeable value tokens) by solving computational puzzles.



Smart Contracts

- Distributed ledgers enable the coding of simple contracts that will execute when specified conditions are met.
- Ethereum is an open source blockchain project that was built specifically to realize this possibility.
- Ether is the incentive ensuring that developers write quality applications (wasteful code costs more)
- Every few seconds a new block is added to the blockchain with the latest transactions processed by the network and the computer that generated this block will be awarded 5 ether. Due to the nature of the algorithm for block generation, this process (generating a proof of work) is guaranteed to be random and rewards are given in proportion to the computational power of each machine.

Coding a Smart Contract

Ethereum accounts:

- external: like wallet addresses.
- Internal: contract addresses
(a class with a bunch of methods)

OpCode. Lowest level code

Python and Node.js have libraries

- Solidity – programming language
Like javascript
- Serpent – python based.
Not used as much
- Web3 – distributed apps on
Ethereum
- “Gas” is the internal pricing for
running a transaction or contract

The screenshot displays a Solidity IDE interface. On the left, a code editor shows a smart contract named `HelloWorld` with the following code:

```
1 // HelloWorld.sol
2
3 contract HelloWorld {
4
5     function displayMessage() constant returns (string) {
6         return "Hello World, from a smart contract!";
7     }
8 }
9
10
```

On the right, the IDE shows the compilation and deployment details. The Solidity version is 0.4.9+commit.3644da25. The contract is compiled to 451 bytes. The deployment details show the contract address and the deployment transaction hash.

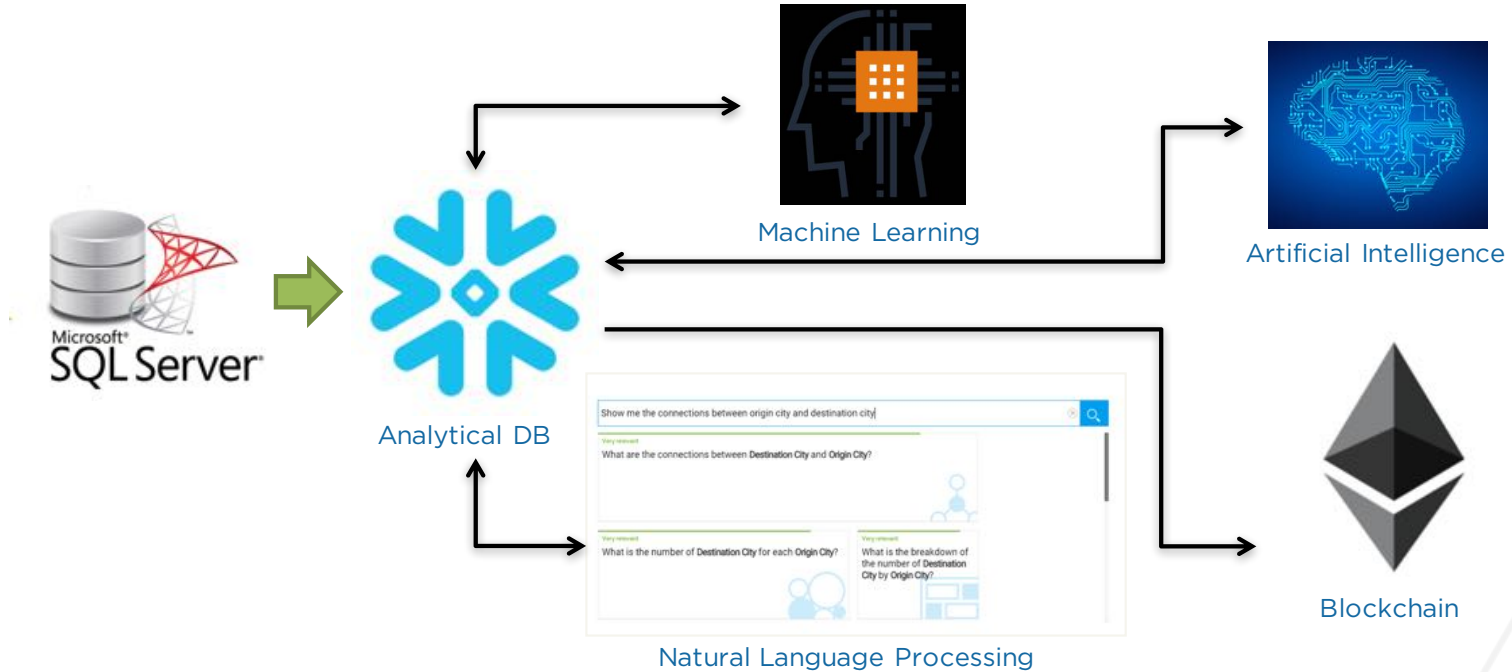
Deployment details:

- Contract address: `0x00`
- Deployment transaction hash: `0x00`

A warning message at the bottom states: "Warning: Source file does not specify required compiler version (see compiler option -v). Compiler will use latest supported version (0.4.9+commit.3644da25)."

BLOCKHEAD Architecture

Use Case: Rights Management for the Entertainment Industry



Data is Food for the Future



Imagine the Possibilities



Forge Ahead – Change the World

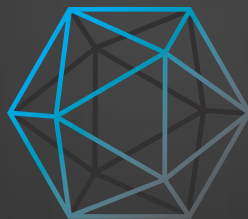
Life is like riding a bicycle, to keep your balance, you must keep moving.

- Albert Einstein



Thank You / Q&A

Joe Caserta
President, Caserta
joe@caserta.com
[@joe_caserta](https://twitter.com/joe_caserta)



caserta

