BIG DATA MODERNIZATION

23 June 2020

Solomon Williams

AVP, Data Management

WHAT IS BIG DATA – WHERE DID IT COME FROM?

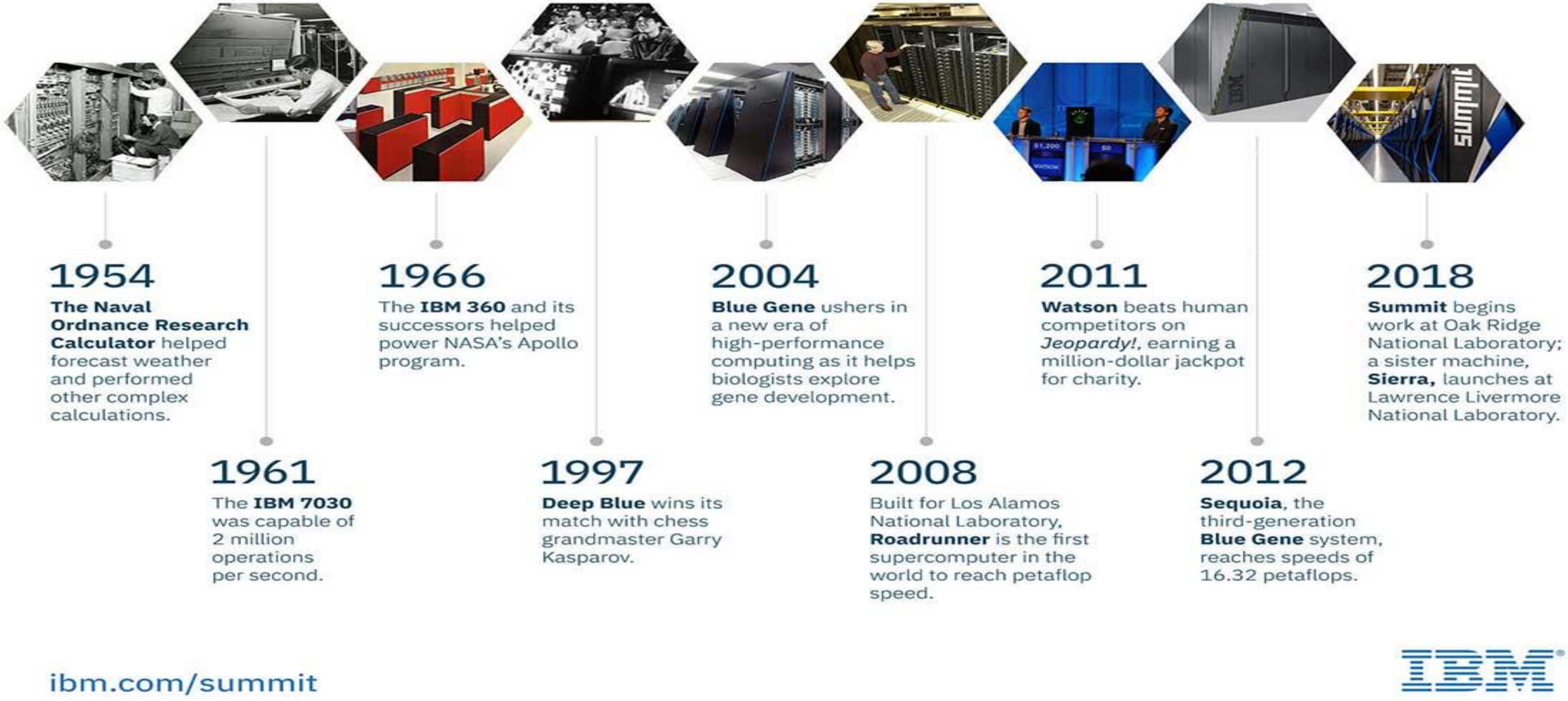
Big Data as a function of enterprise data management is not new. Large volumes of data requiring storage, integration, analysis and the ability to query it (quickly) has been a mainstay of data management for a very long time…So where did BIG DATA come from?

- Big Data as a term, was coined approximately 20 years ago. Since that time “big data” as a concept has grown and has been promoted as this functional area with its own methodology for strategy development, implementation, management and modernization and this is not as it should be...
- Why?

Simple answer… It was a marketing device to sell more technology!

- Storage, processing and access of large, complex datasets for the purpose of developing insights has been a common activity for some time. If we go back to the mid-50’s we see the introduction of what are rightly termed super-computers. These super-computers have grown in storage and compute capacity for nearly 50 years – why? - because data **volume**, **velocity**, and **variability** kept growing. Where was the term big data then?
- In our discussion today, we are going to look at big data for what it is, a framework of integrated capabilities and technologies used for the purpose of gaining deeper insights into data which continues to grow in volume and complexity
- Regarding approaches to modernizing big data, we will examine modernizing the data architecture which will allow the utilization of more capable technologies without breaking the bank, or the business each time a technology advancement comes to market

EVOLUTION OF COMPUTING…AND DATA



1954
The Naval Ordnance Research Calculator helped forecast weather and performed other complex calculations.

1961
The **IBM 7030** was capable of 2 million operations per second.

1966
The **IBM 360** and its successors helped power NASA's Apollo program.

1997
Deep Blue wins its match with chess grandmaster Garry Kasparov.

2004
Blue Gene ushers in a new era of high-performance computing as it helps biologists explore gene development.

2008
Built for Los Alamos National Laboratory, **Roadrunner** is the first supercomputer in the world to reach petaflop speed.

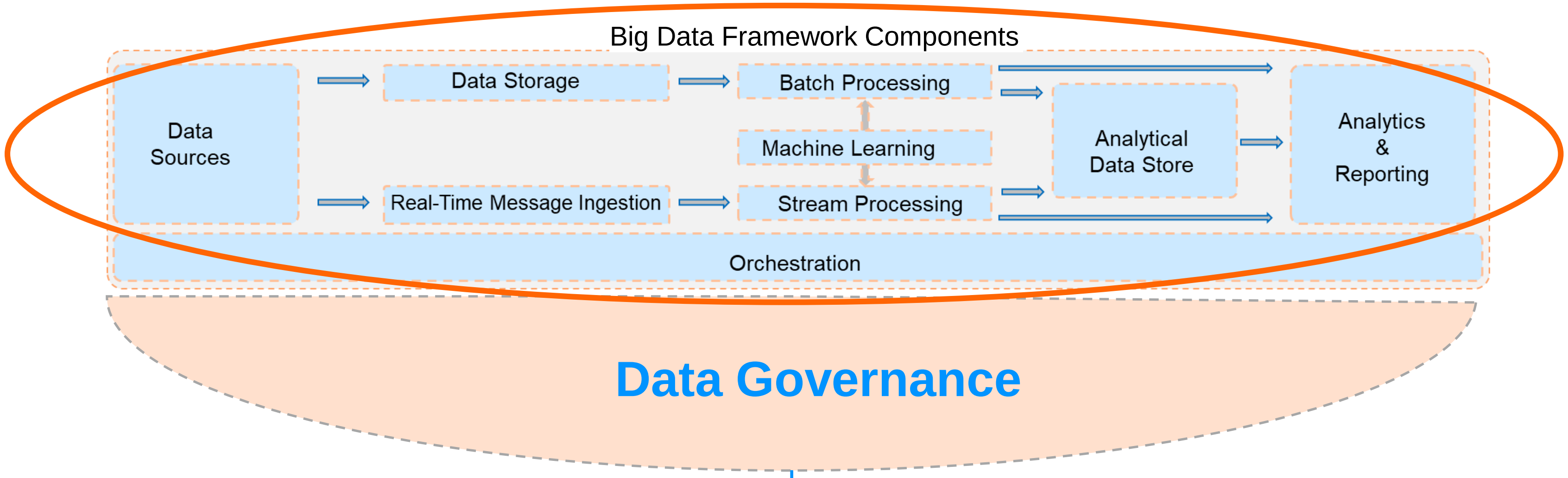
2011
Watson beats human competitors on *Jeopardy!*, earning a million-dollar jackpot for charity.

2012
Sequoia, the third-generation **Blue Gene** system, reaches speeds of 16.32 petaflops.

2018
Summit begins work at Oak Ridge National Laboratory; a sister machine, **Sierra,** launches at Lawrence Livermore National Laboratory.

ibm.com/summit

BIG DATA IS A FRAMEWORK OF MULTIPLE CAPABILITIES AND TECHNOLOGIES



Data Sources

- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT devices.

Data Storage

- Often called a data lake data for batch processing is typically stored in a distributed file system

Real-time Message ingestion

- Capture and store real-time messages for stream processing

Batch Processing

- Often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis.

Machine Learning

- Often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis.

Stream Processing

- Capability that allows the solution to filter, aggregate, and further prepare real-time messages for analysis

Analytical Data Store

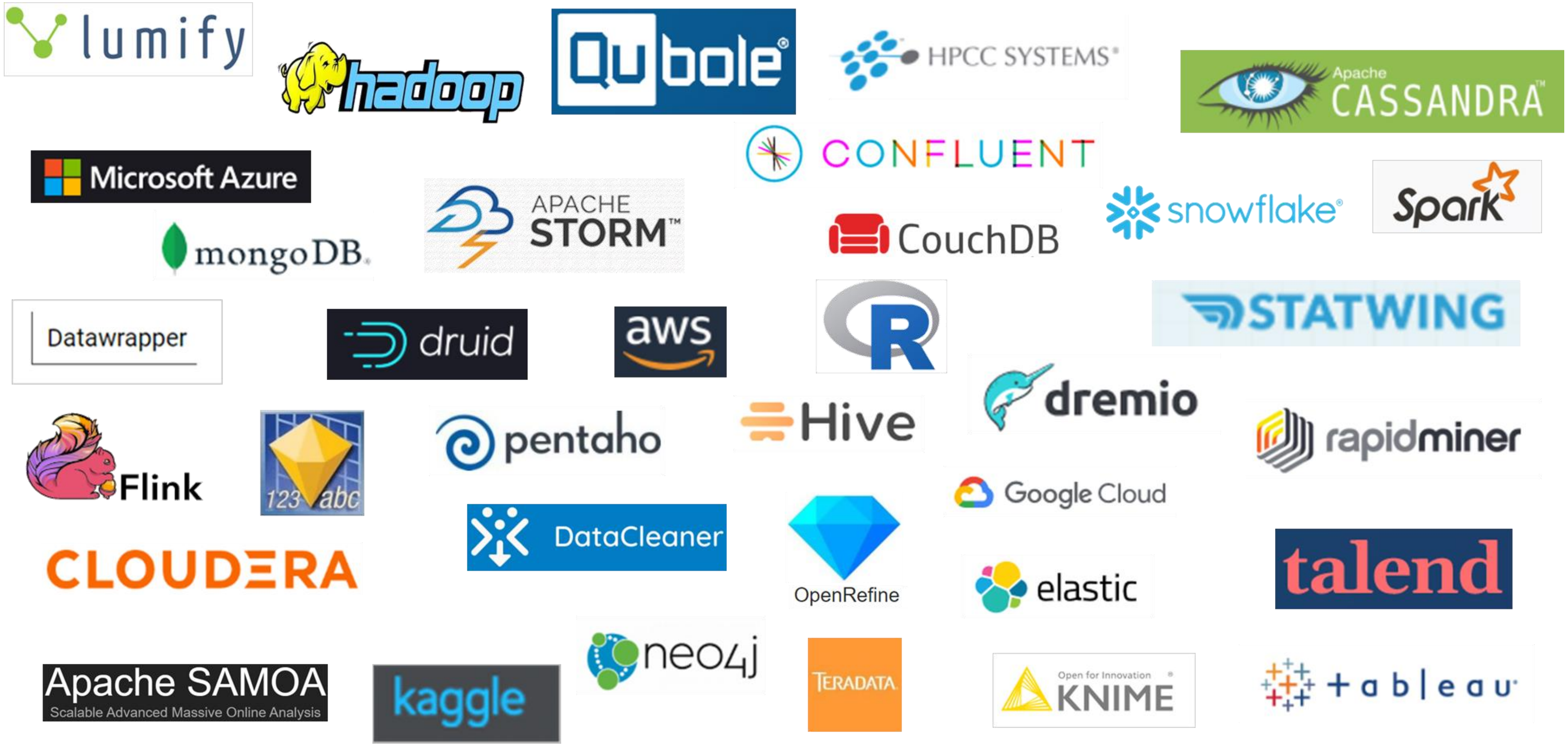
- Data stored post processing in a structured format for query analysis

Orchestration

- Capability of the solution to manage workflows, data movements and other automated functions

SMALL SAMPLE OF THE BIG DATA VENDOR LANDSCAPE

Worldwide revenues for big data and business analytics solutions are expected to grow at a compound annual growth rate of 13.2 percent over the next several years to reach $274.3 billion in 2022, according to market researcher IDC.



MODERNIZING BIG DATA OR MODERNIZING DATA ARCHITECTURE

"Big data" was a marketing term coined to sell technology. Most organizations were already doing this work, however, it you weren't doing "big data" you weren't cool…So when it comes to big data modernization, at the heart of it is data architecture modernization

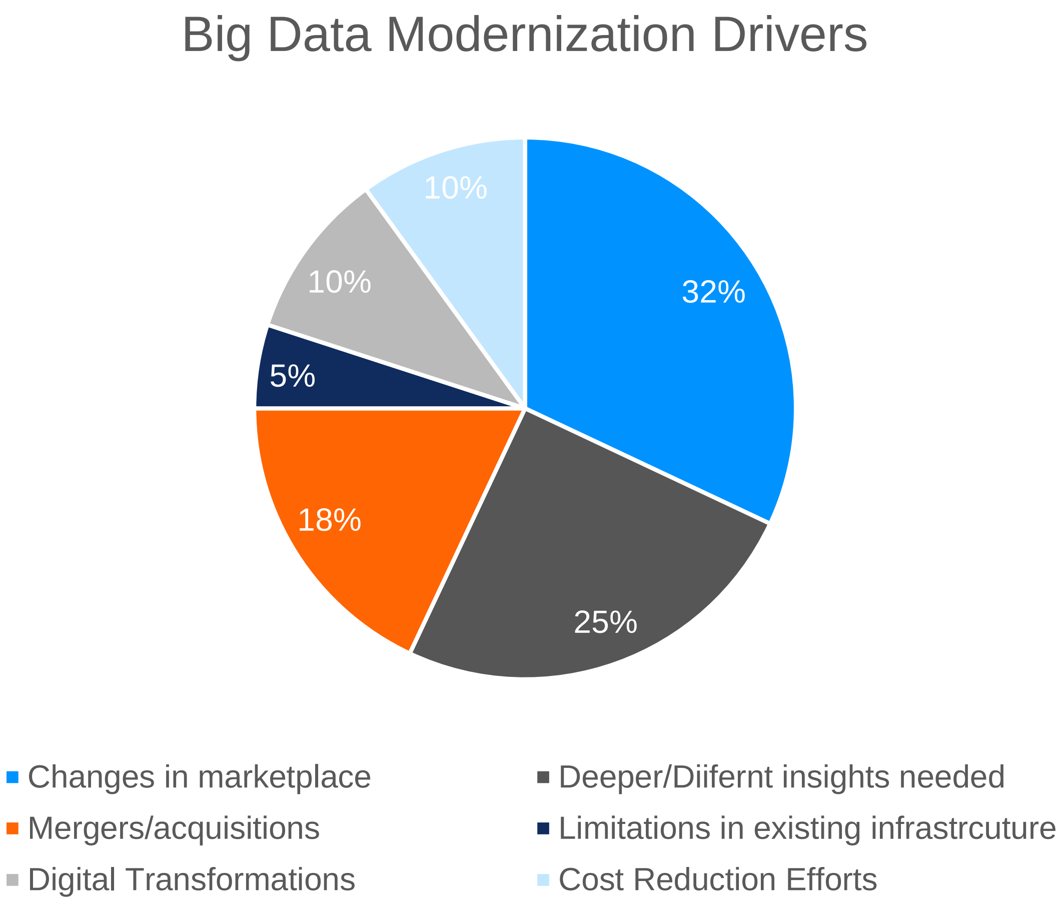
Why data architecture modernization instead of big data modernization?

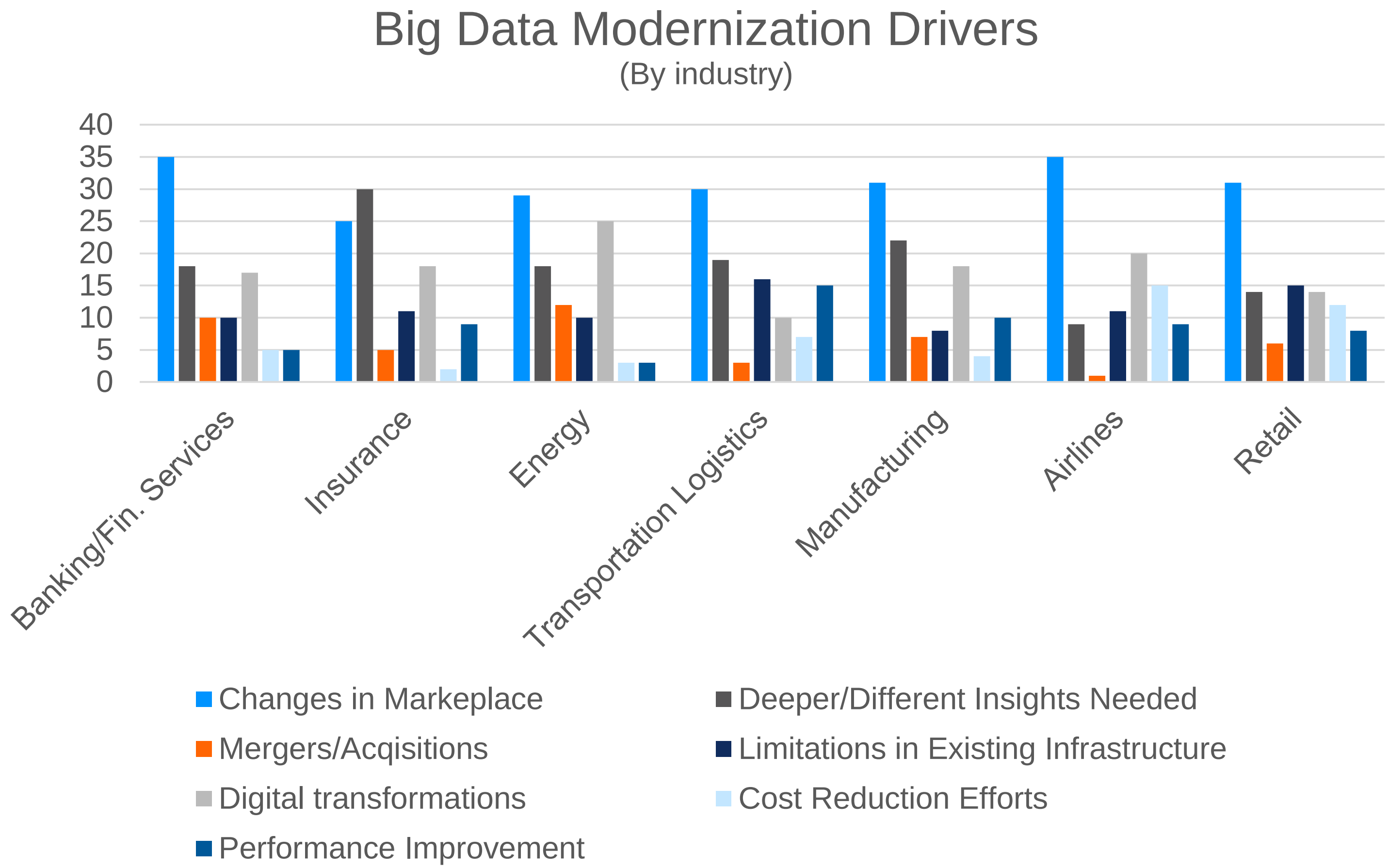
- Organizations have been struggling with the increase in volume and complexity of data for hundreds of years (yes, there was data two centuries ago…it just that it wasn't digital)
- Then, like now, faster and deeper insights were the request of business leaders
- Then, like now, it was advances in technology that was sought to help solve the latency between data ingestion and insights produced
- The biggest difference between then and now, is we have an **entire industry of technology vendors** who produce newer and more advanced capabilities nearly every month
- These new capabilities are developed against the ever-increasing volume, velocity, variability, variety and veracity of the data
- How is an organization supposed to manage this data management issue alongside the new technology availability issue?
- By creating a flexible, scalable data architecture which enables the adoption of new technology capabilities in a seamless manner without having to re-architect and re-design each time…This is the purpose of the **modern data architecture**

DATA ARCHITECTURE MODERNIZATION AND DRIVERS

DATA ARCHITECTURE MODERNIZATION DRIVERS

Based on our experience we have identified several key drivers of big data modernization efforts and those drivers across selected industries



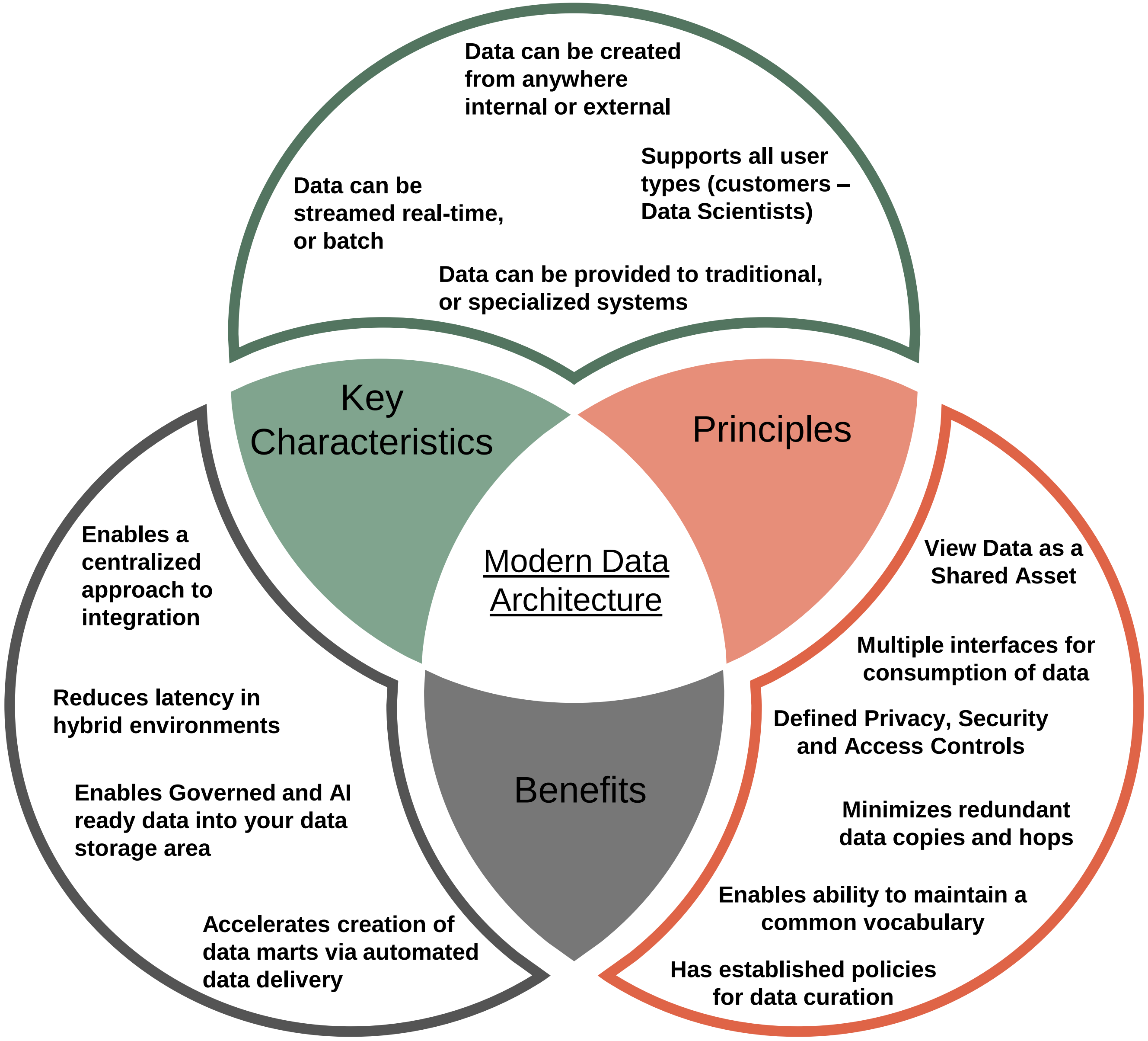


KEY CHARACTERISTICS OF A MODERN DATA ARCHITECTURE

- **Data is the foundation to delivering analytic insights for making knowledgeable and supportable decisions.**

- **The way data and data management assets are organized is the data architecture. Data architecture is a set of models, rules, data flows integration patterns and policies which illustrates and informs how data is captured, stored, processed, and synchronized throughout the organization.**

- **Many organizations that use traditional data architectures today are rethinking their data architecture. This is because existing data architectures are unable to support the speed, agility, and volume that is required by companies today.**



Data can be created from anywhere internal or external
Data can be streamed real-time, or batch
Supports all user types (customers – Data Scientists)
Data can be provided to traditional, or specialized systems
Key Characteristics
Principles
Modern Data Architecture
Benefits
Enables a centralized approach to integration
Reduces latency in hybrid environments
Enables Governed and AI ready data into your data storage area
Accelerates creation of data marts via automated data delivery
View Data as a Shared Asset
Multiple interfaces for consumption of data
Defined Privacy, Security and Access Controls
Minimizes redundant data copies and hops
Enables ability to maintain a common vocabulary
Has established policies for data curation

UNDERSTANDING – THE BUSINESS SIDE

The business side of data architecture is not just a critical input, but the most important input to a big data solution. In order to achieve any measure of success, there must be support from the business, clear understanding of objectives and goals the business desires to achieve, and a realistic measure of the organization's data literacy

Why modernize the Data Architecture…For what purpose?

- **Performance Management**
  - It involves using transactional data like customer purchase history, turnover and inventory levels to make decisions relating to store management and operational supremacy. This data is available within the organization and gives insights into subjects relating to short term decision making and long term planning. It works well with companies with large historical databases that can be leveraged without much pain. It can also help with better customer segmentation and targeting
- **Data Exploration**
  - This approach makes heavy use of data mining and research to find solutions and correlations that are not easily discoverable with in-house data. Currently, it is used by companies focusing on robust inbound marketing to generate insight on prospects behavior on the website. It helps you identify new segments of data and bring out insights regarding customer's behavior and preferences.
- **Social Analytics**
  - Social analytics measures the non-transactional data on various social mediums and review sites like Facebook, Twitter and Google+. It is based on the analysis of conversations and reviews that come up on these platforms. It brings out three primary analytics viz. awareness, engagement, and word-of-mouth. In-stream data analysis techniques like sentiment analysis prove very effective in these cases. It gives insights on the brand identity and customer's opinions on new offerings and services. The social analysis also proves effective in predicting spikes in demand for certain products.
- **Decision Science**
  - Decision science refers to the experiments and analysis on non-transactional data, such as consumer-generated content, ideas, and reviews. Decision science is more about exploring possibilities than measuring known objectives. Unlike social analysis, that is based on engagement analytics, decision science focuses on hypothesis testing and ideation process. This involves extensive use of text and sentiment analysis to understand customer's opinions about new services and schemes.

UNDERSTANDING – THE TECHNICAL SIDE

Big data is a framework of multiple components. Understanding the constituent components will better inform how your organization's data architecture needs to evolve. In this way a more informed strategy can be defined

Data Characteristics

Volume

Organizations collect data from a variety of sources (business transactions, smart (IoT) devices, industrial equipment, videos, social media and more. With big data storage tools the burden has been reduced

Velocity

With the growth in IoT, data streams inbound at an incredible speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Variety

Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

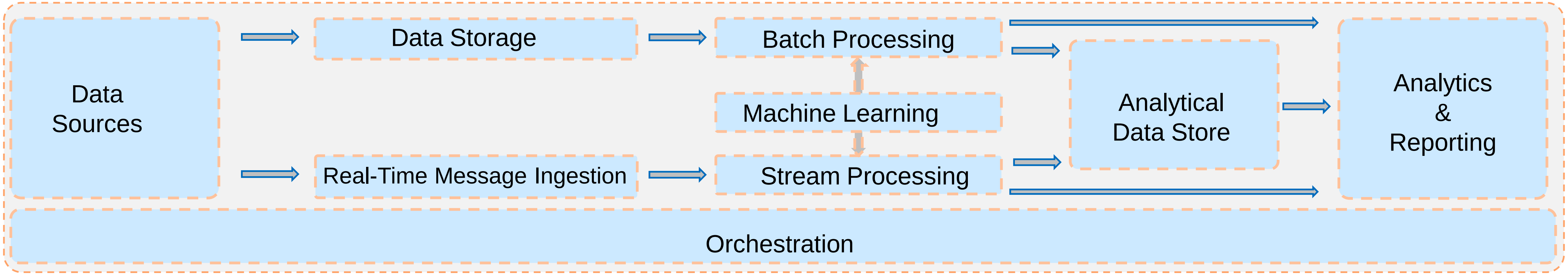
Variability

Increasing velocities and varieties of data, data flows are unpredictable – changing often and varying greatly. Businesses need to know when something is trending in social media, and how to manage multiple peak data loads.

Veracity

Veracity is the quality of data. Data comes from so many sources, it's difficult to link, match, cleanse and transform data across systems. Businesses need to connect and correlate relationships, hierarchies and multiple data linkages..

Solution Framework Components



Data Sources

Data Storage

Real-Time Message Ingestion

Batch Processing

Machine Learning

Stream Processing

Analytical Data Store

Analytics & Reporting

Orchestration

KEY COMPONENTS OF DATA ARCHITECTURE MODERNIZATION STRATEGY

A data architecture modernization program requires a comprehensive strategy

Key components of a modernization strategy are…

Clearly Defined Goals

Your end goal has the biggest impact on the shape of your overall strategy.

You need to decide whether you want to increase the efficiency of customer reps, improve operational efficiency, increase revenues, provide better customer experience or improve marketing.

The goal you have should be precise, certain and direct. Any strategy with just the sole purpose of exploring possibilities is likely to end up in confusion.

Data Availability

A key component of a big data strategy is the data.

Big data as the name states relates to large data sets. However, not all data is required for big data analysis

Part of the big data strategy is to determine the appropriate data sources for analysis:

Technical Constraints & Debt

big data solutions are not plug-and-play. Any implementation of big data tools will require changes/modifications to the existing infrastructure

If the old company data was stored in traditional formats it might not facilitate the running of complex algorithms and analysis.

different departments may need integration to collect and streamline data to put it to more usable format.

Data Literacy

Data literacy is an understanding of data sources, constructs, analytical methods and techniques and the ability to describe the use case, application and resulting value. – In short, data literacy is the ability to communicate business value in the context of data.

A big data strategy must consider the maturity data literacy of the organization to ensure usability of the solution

Data Governance

Data Governance is equally as important as data literacy to the big data strategy.

Without data governance the lineage, meaning and intended usage of the data will be lost

If there is any ambiguity of the lineage, meaning and usage of the data is lost, there will be an immediate impact to the usability and adoption of the big data solution and loss of confidence in the insights provided

Resource Availability

The right team of business and technical resources for the developing the big data strategy is essential

Business SMEs, Statisticians, Data Scientists, big data architects and developers are all required

ARCHITECTURE PRINCIPLES

Cost Optimized

- An architecture with appropriate tradeoffs: buy vs. build, proven technology vs. leading technology, open source vs. commercial, etc.

Governed

- EDM foundation services (such as quality, audit, and security) are mandatory and applied to the entire architecture

Modular

- Modular approach decouples interdependency, increases reusability, and allows simpler recoverability in the event of failure

Maintainable

- Leverage metadata driven frameworks, reusable patterns, and configurable processes in the architecture where applicable

Scalable/Reliable

- Think managed service, serverless, and containers for elasticity, scalability, overhead and maintenance reduction
- Match Supply and Demand

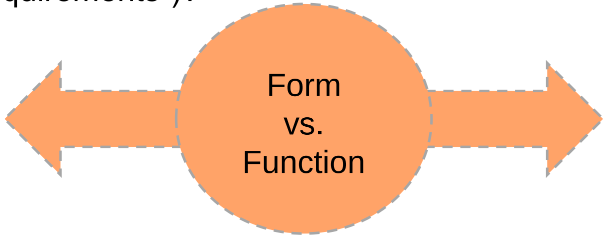
Flexible

- “Toolbox” (multi-option) approach to avoid locking into a single technology. It leaves room to take certain architecture risks!

UNDERSTANDING TRADE-OFFS

Design Principles - The fundamental core of all design is “Does the architecture serve the function (or in this case functional-al requirements”)?

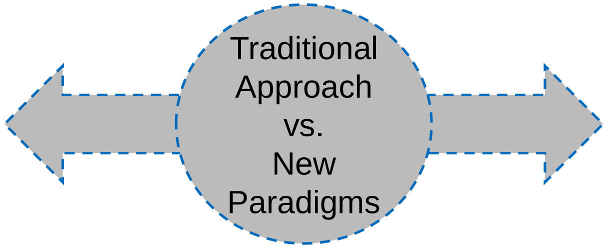
- Is a cutting edge / bleeding edge design required?
- Does a technology serve only one capability?



- Should the architecture strongly adhere to rigid “separation of concerns”?
- De-prioritize certain capabilities, if enabling makes the solution less usable?

Data Management Approach – There are tried and true methods for Data warehousing and analytics. For a packaged analytics solution, however, is a new offering and requires new thought paradigms

- Relational Databases
- Traditional Datawatehouse
- Traditional Extract, Transform and Load (ETL)



- Data Lake
- Cloud Based Object Storage
- Receive and Load Data Integration

Technology Stack - There is no one definitive technology stack. The options are myriad, as well as the configuration

- Established, more mature technologies can equate to higher costs and greater complexity of implementation

Established
vs.
Challengers

- Newer Technology vendors, who are less established may also have fewer use cases, and limited support and support communities

KEY REQUIREMENTS WHEN MODERNIZING DATA ARCHITECTURE

1. Number of sources to integrate
2. Total data volume
3. Incremental data volume
4. Latency
5. Availability
6. Data consistency
7. Volatility of underlying sources
8. Data quality
9. Variety of analytics
10. Data types
11. History (as-is vs as-was)
12. Workloads
13. Concurrency
15. Concurrency
16. Auditability
17. Security (access & obfuscation)
18. Criticality (what if you lose data)
19. Speed of delivery
20. Performance expectations
21. Storage options
22. Locality of data
23. Technology compatibility
24. Internal skill sets
25. Cost sensitivity
26. Vendor maturity
27. Software maturity

DATA ARCHITECTURE MODERNIZATION BEST PRACTICES

Decouple the Architecture - The core of the solution is not the technology, it’s the data architecture that the supports the technology

- The Data Architecture…
- Should not be limiting
- Deals with change more easily and at scale
- Does not enforce requirements and models up front
- Does not limit the format or structure of data
- Assumes the range of data latencies in and out, from streaming to one-time bulk
- Allows both reading and writing of data from outside

BIG DATA GOVERNANCE

DATA GOVERNANCE IS THE FOUNDATION OF ALL DATA MANAGEMENT

Data Governance is often overlooked when it comes to big data. This oversight is what leads many organizations into the data swamp

- Just as big data is really a component of enterprise data management, so too is big data governance a component of an organization's information governance capability
- Big data governance is the capability of decision-making regarding policy definition, privacy and monetization of data on the big data platform by coordinating the objectives and priorities of multiple business units
- Big data governance must set rules on how the data is to be used, AND how it is not to be used
- Big data governance must also manage the metadata – It must build information about the inventory of data held
- Big data governance must manage the quality of the data and set policies for data hygiene, data ingestion and data synchronization
- Big data governance must have as a core principle the mission to mature the **data literacy** of the organization

EFFECTIVENESS OF DATA GOVERNANCE

- Un-governed large volumes of data with varying velocity, variety, variability and veracity

Ineffective Data Governance

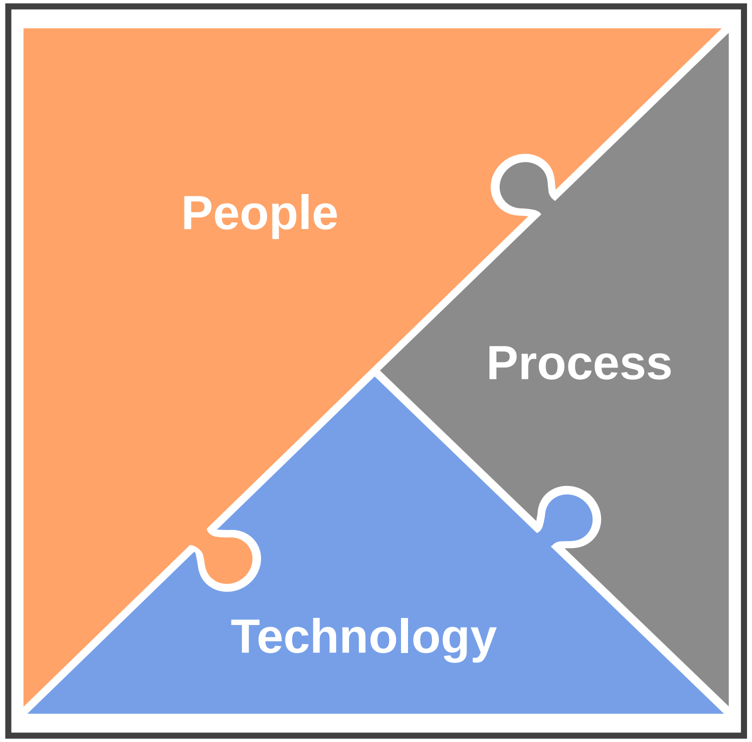
Effective Data Governance

- Governed large volumes of data with varying velocity, variety, variability and veracity

- This rather extreme example is used to drive home the point and need of data governance as the underlying foundation for big data.
- The very definition of big data is that data is coming from everywhere, in multiple formats and varying velocities, how can we expect to get any value if it is not governed effectively

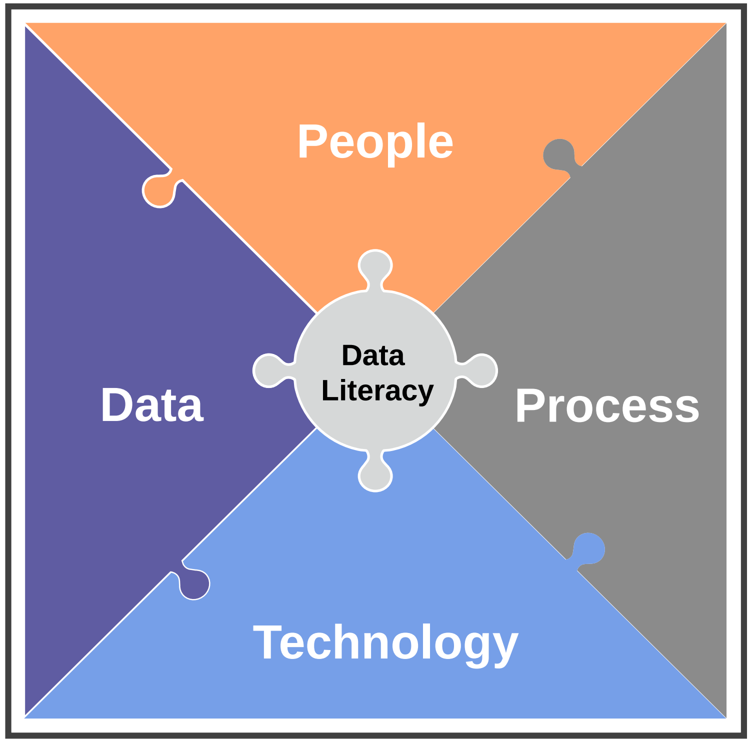
DATA LITERACY

Digital skills are critical, including an understanding of sensors, robots, digital twins, mobile, cloud and seamless collaboration. However, there is a fundamental element that flows through all of these — data. The need to understand how insight can be derived from data through analytics and artificial intelligence (AI) is foundational to how every employee engages with it and, in so doing, adds value



Traditional Thinking

- Breaking down traditional thinking about how to succeed in business with data-driven decision making is a key objective of big data governance
- Executives must realize that speaking data is a missing link that will uncover unrealized value from years of data and analytics initiatives. They also must view **data as the new core capability** driving how the business should compete, innovate and be efficient in a digital context.



Required Thinking

DATA GOVERNANCE FRAMEWORK FOR BIG DATA

The big data framework is the intersection of data, people, process and technology - which are also the key aspects of data literacy. It is this critical intersection where the value of data governance really comes into appreciation

Business

Functions

- Sales & Marketing
- Operations
- IT
- Accounting
- Finance
- Risk

Objectives

- Improved customer Satisfaction
- Growth in marketshare
- Streamlined processes
- Responsive applications
- More accurate forecasts
- Better insights into order to cash

01
02
03
SQL
CSV
DOC

Big Data Categories

- Clickstream and Social Media Data
- IoT – Technology-to-Technology Interfaces
- Digital Process Data
- Systems Logs
- Biometrics
- Images
- IVR, Email & scans

Information Governance Domains

- Identity and Access Management
- Data Literacy
- Organization Change Management
- Business Process Orchestration and Integration
- Data Privacy
- Data Hygiene and Quality
- Metadata Management
- Master & Reference Data Management

SUMMARY

Modernizing your big data platform is less about adopting the newest technology and more focused on the underlying data architecture which enables the technology.

The take-aways from this session should be….

1. Have a clear understanding and agreement with all stakeholders as to the objectives of the modernization
2. The key characteristic of big data is the data. Large volumes, highly complex, from multiple sources. Ensure the architecture required to meet the needs of the program objectives is also flexible, and scalable
3. Don’t get caught up in the technology. Technology will continue to evolve, ensure the architecture selected will be able to support technology changes without having to re-design each time
4. Decouple the architecture. Data acquisition should not be directly tied to the needs of consumption. It must operate independently of data use.
5. Leverage the cloud as much as possible in architecting your modernized architecture
6. Data governance is the foundation of success. At each component of the big data solution (data collection, data analysis, data storage and data querying), data governance should have a seta at the table

THANK YOU

Solomon Williams

AVP, Enterprise Data Management
solomon.williams@exlservice.com
M: +1 480 543 7541

EXL (NASDAQ: EXLS) is a leading operations management and analytics company that designs and enables agile, customer-centric operating models to help clients improve their revenue growth and profitability. Our delivery model provides market-leading business outcomes using EXL's proprietary Business EXLerator Framework®, cutting-edge analytics, digital transformation and domain expertise. At EXL, we look deeper to help companies improve global operations, enhance data-driven insights, increase customer satisfaction, and manage risk and compliance. EXL serves the insurance, healthcare, banking and financial services, utilities, travel, transportation and logistics industries. Headquartered in New York, New York, EXL has more than 27,000 professionals in locations throughout the United States, Europe, Asia (primarily India and Philippines), South America, Australia and South Africa.

EXLservice.com

GLOBAL HEADQUARTERS

280 Park Avenue, 38th Floor
New York, New York 10017
T +1 212.277.7100 **F** +1 212.771.7111

United States • United Kingdom • Czech Republic • Romania • Bulgaria • India • Philippines • Colombia • South Africa