



Modern Data Management and Governance

Benjamin Pecheux

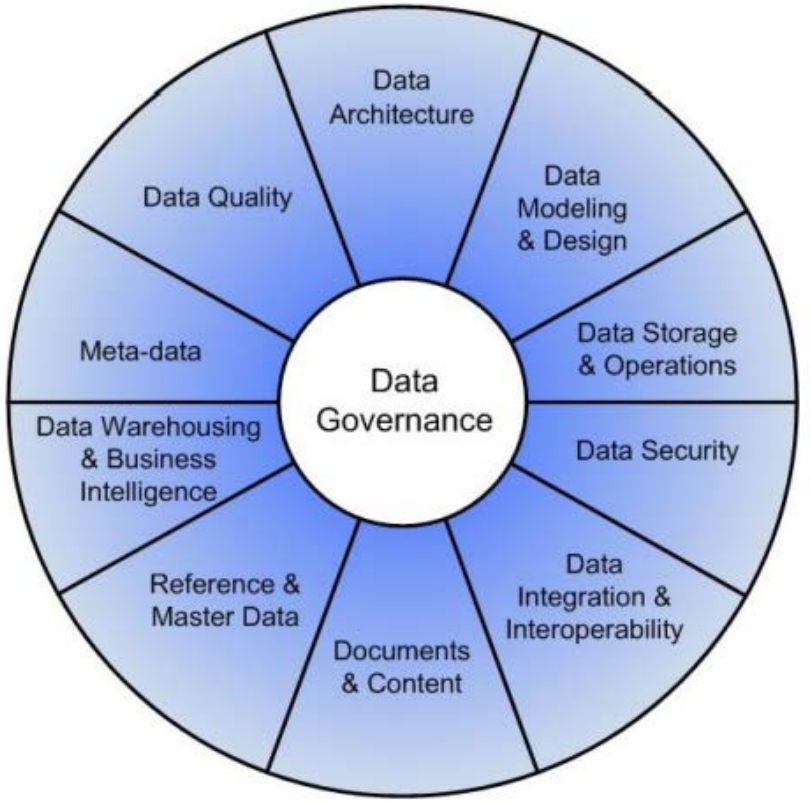
Data Management and Governance for Better Crowdsourced Data Applications
Adventures in Crowdsourcing Webinar Series
FHWA EDC-5
January 28, 2020

Agenda

- *What is data management?*
- *Traditional vs Modern*
- *Data lifecycle phases*
- *Create*
- *Store*
- *Use*
- *Share*

What is Data Management

- Data management – *the development and execution of architectures, policies, practices, and procedures that properly manage the full data lifecycle needs of an enterprise (DAMA).*
  - 11 data management knowledge areas
- Traditional data lifecycle



Data Management Knowledge Areas (DAMA Data Management Book of Knowledge – DMBOK)

Velocity of Obsolescence

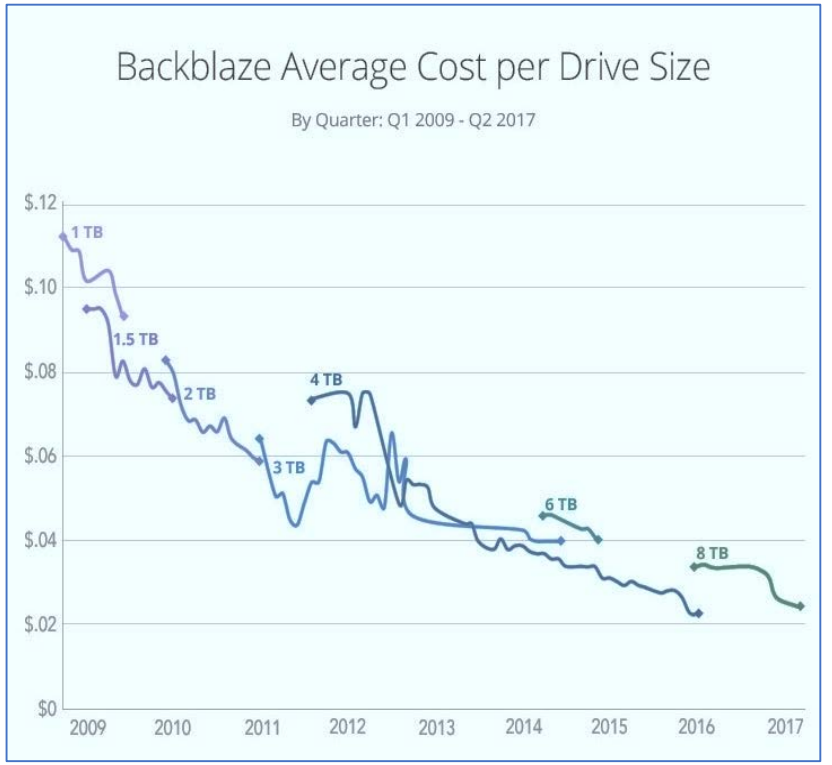
Obsolescence is defined by the time when a technical product or service is no longer needed or wanted even though it could still be in working order: Hardware, Software, and Skills

Cost of obsolescence

- Legal and regulatory compliance risks
- Security vulnerabilities
- Lower IT-flexibility
- Data silos
- Lack of skills and support

How to cope with obsolescence

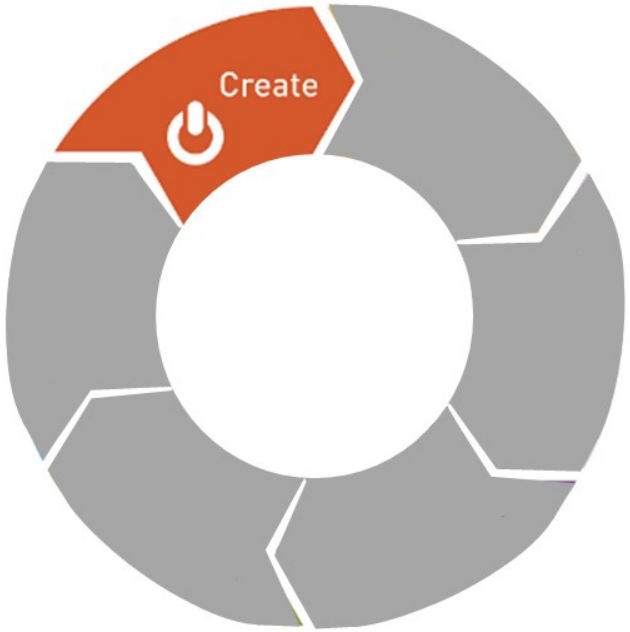
- Open standard, open source software, and cloud services



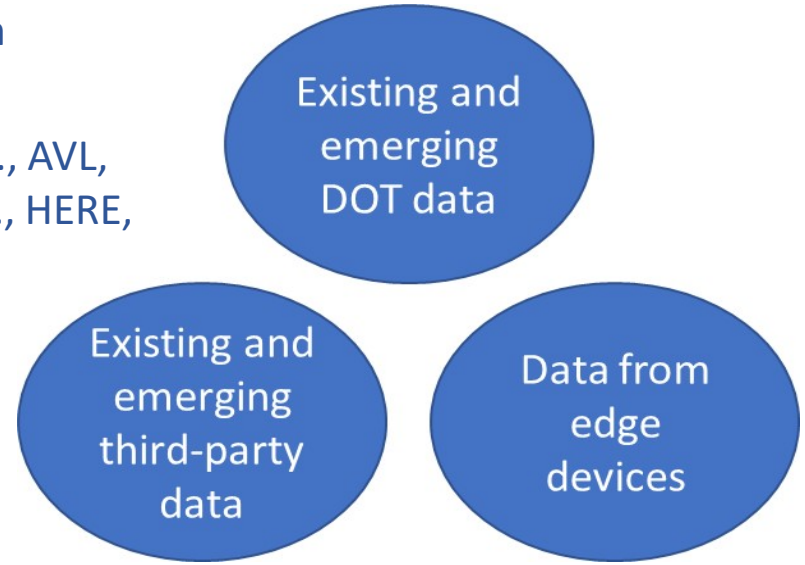
Data Lifecycle Management Framework - Overview

- Big data lifecycle
  - Create
  - Store
  - Use
  - Share
- Augmented
  and restructured to handle what's coming now and in the future without to have to redesign everything

Create

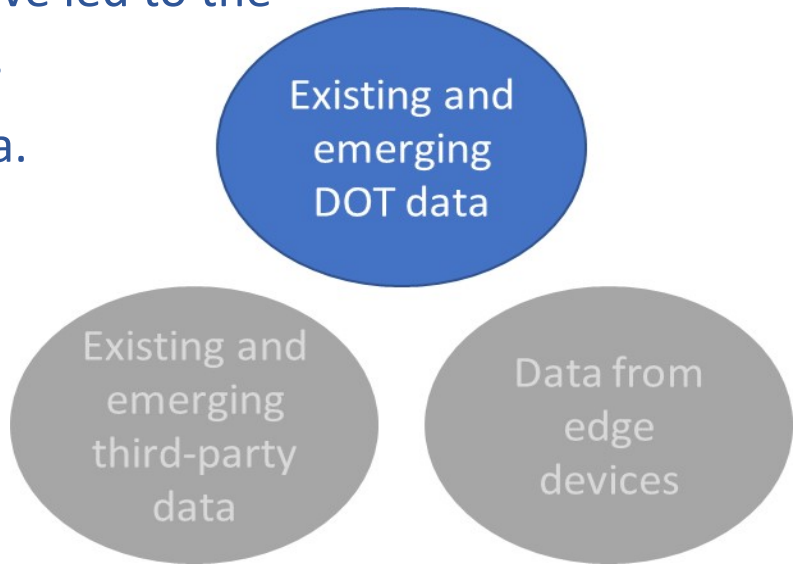


- Entails the gathering, collection, or otherwise creation of new data.
- Could include information generated from existing sensors, the discovery of a new internal dataset, access to a new external partner dataset, or the purchase of a new dataset from a third-party provider.
- Most common data source types used by transportation agencies:
    - Raw data collected and controlled by the agency – includes both existing/traditional data (e.g., ITS devices, crash, asset) and data from emerging technologies such as connected vehicles and smart cities.
    - Data obtained from third-parties – includes data from vendors (e.g., AVL, ATMS), partnership agreements (e.g., Waze CCP), crowdsourcing (e.g., HERE, INRIX), and social media platforms (e.g., Twitter, Facebook).
    - Data processed at the edge – could be DOT raw data, but it is being processed at the edge instead of being sent to storage.



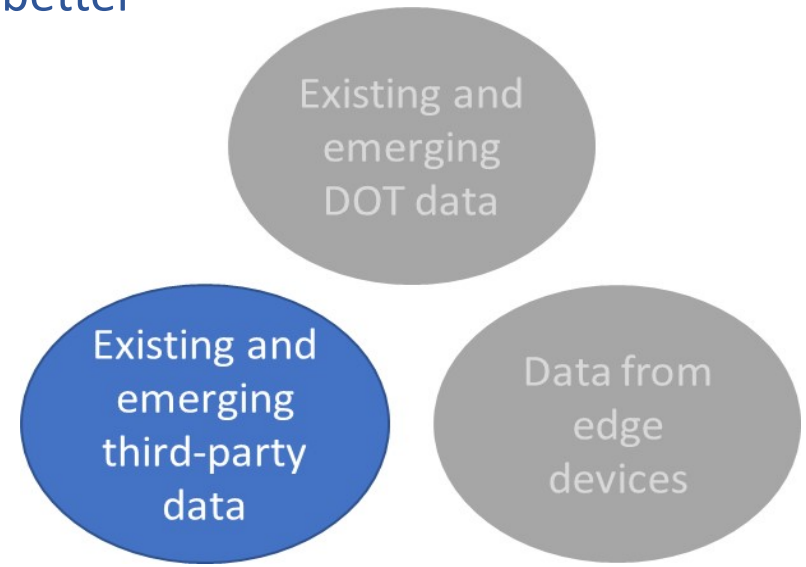
Modern Data Practices – *Create*

- Data is collected as it is generated without being modified or aggregated.
- The quality of the data is assessed, tagged, and monitored as it is being collected.
- Collected data is both technically and legally open. Potential infrastructure software and hardware vendor lock restricting data usage is avoided or resolved.
- Collection of data is not limited to known or familiar data. Each business unit is aware of what data is available outside of the unit, and investigations have led to the knowledge of if and how this data could support decision making.
- Accurate data lineage is maintained for all pieces of collected data.
- Collected data is not segregated (i.e., siloed). The same collection approach is applied to all incoming data using the same platform or system.



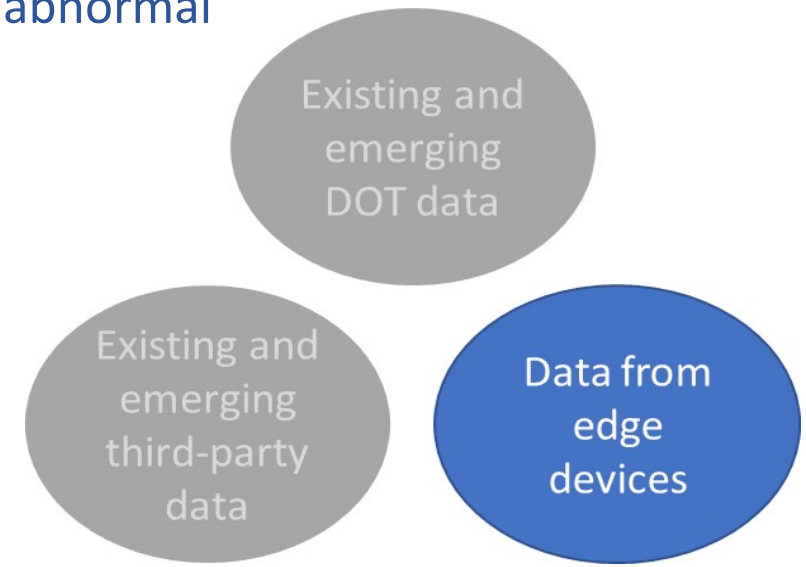
Modern Data Practices – *Create*

- A clear understanding of the purpose, lineage, value, and limitations of the data product(s) being sold or provided is established.
- Data quality rules and metrics for third-party data are established rather than solely relying on the quality metrics provided by the data provider (if provided at all).
- Third-party data products are augmented or customized to allow better understanding of their quality and establish contract clauses or communication channels with providers to fix potential issues.



Modern Data Practices – *Create*

- Data coming from the edge devices is not the sole source of data for any particular purpose or application. A sliding history of the last few minutes of raw data ingested by the edge device is collected to help diagnose variations/abnormal behavior and improve edge device algorithms.
- Edge device performance assessments are conducted using the collected raw data, and edge devices and edge device data are audited regularly to measure the performance of the edge devices.
- Edge device data is monitored in real-time to detect slow drift or abnormal behavior rapidly.
- An edge device maintenance approach based on disposability is adopted to quickly replace devices as soon as they start to drift or act abnormally.

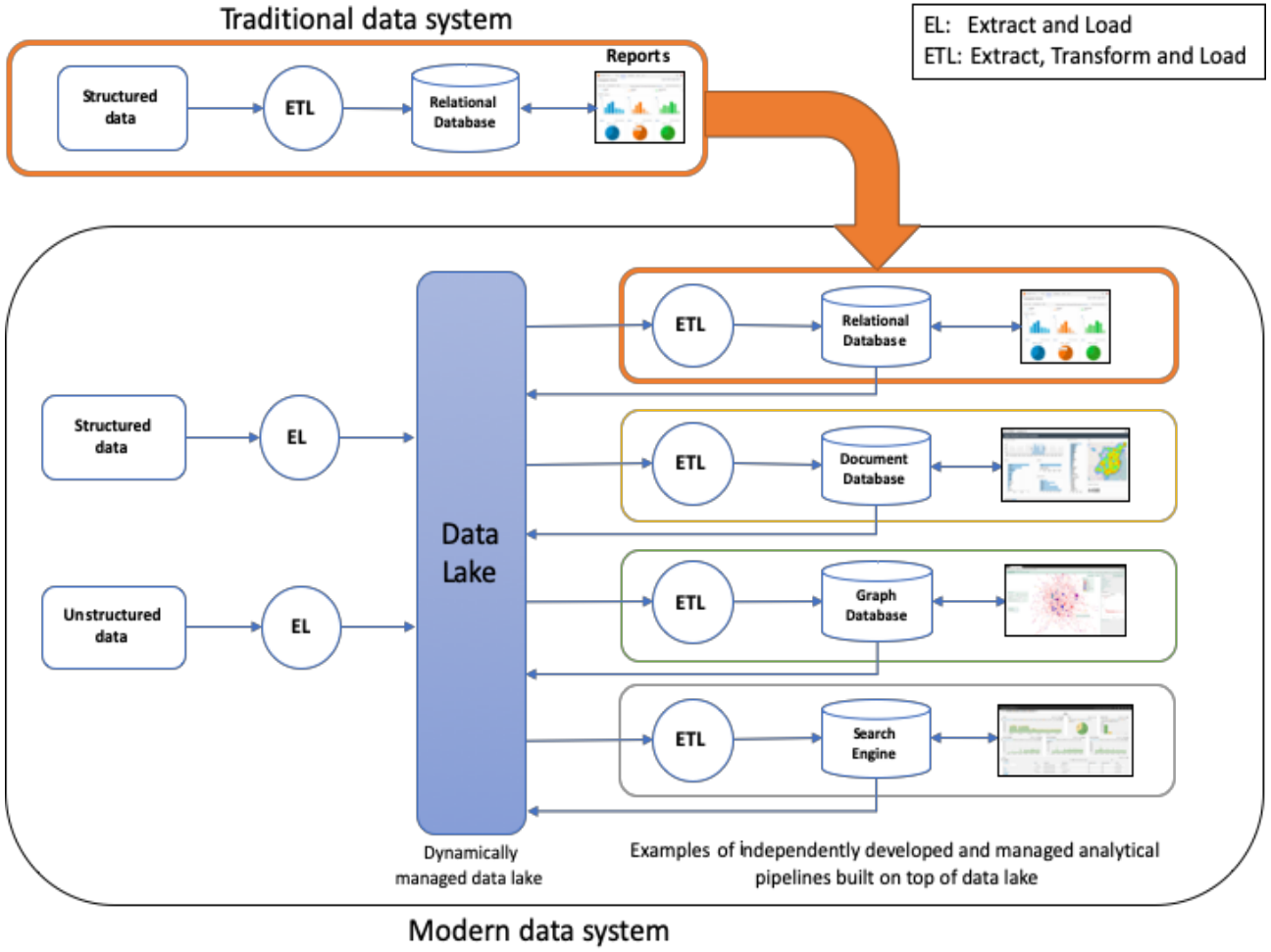


Overview of *Store*

- Encompasses the management and use of data storage architecture to store existing and newly acquired datasets.
- All data management and configuration that is performed on collected data to prepare it for future use.
- Properly managed data is securely stored in an architecture built to support its individual format and use cases while remaining scalable, resilient, and efficient.

Ideal Modern / Big Data Practices – Store

- New architectural patterns need to be adopted to cope with the wide variety of fast changing data.
- Flexible and distributed data architecture capable of applying many analytical technologies to stored data.
- Data is stored in a "data lake."
- "Schema on read" / "schema last."



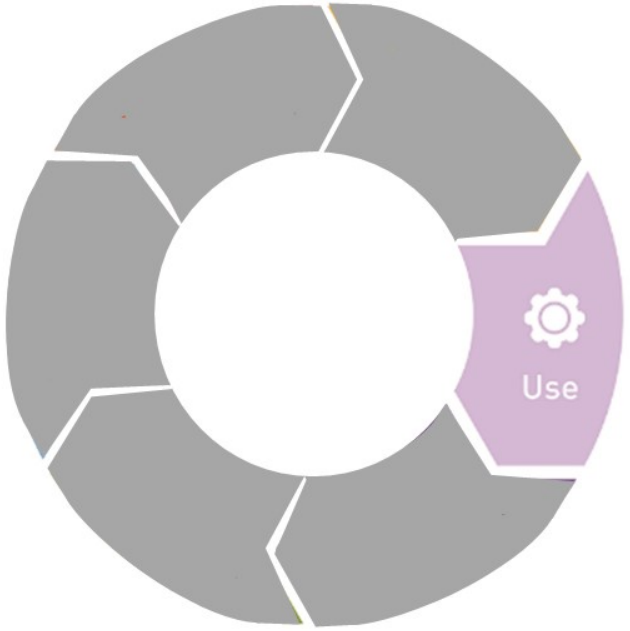
Ideal Modern / Big Data Practices – Store

- Data Storage:
  - A cloud-based, object, storage solution, also called “data lake,” is used to store all data.
  - All data is stored, both structured and unstructured data.
  - No filtering or transformation is imposed on the data prior to storing it; each end user defines and performs their own filtering/transformations.
  - Inexpensive cloud-storage solutions are used for inactive data rather than performing traditional back-ups.
  - Isolated cloud storage solutions are used if strong security requirements are needed.

Ideal Modern / Big Data Practices – Store

- Data Management:
  - Data is organized using the “regular file system” like structure offered by cloud-based object storage.
  - Raw data is augmented/enriched by adding metadata to each record to help end users understand and use the data.
  - Folder structures, datasets, and access policies are managed to accommodate end-users’ needs while maintaining the security and quality of the data.
  - Accessibility of the raw data is maximized by using open file formats and standards.
  - Data discoverability is maximized by maintaining a searchable metadata repository.
  - End users’ data access and use is monitored and controlled in real-time.
  - Open file compression standards are used to limit storage space used.

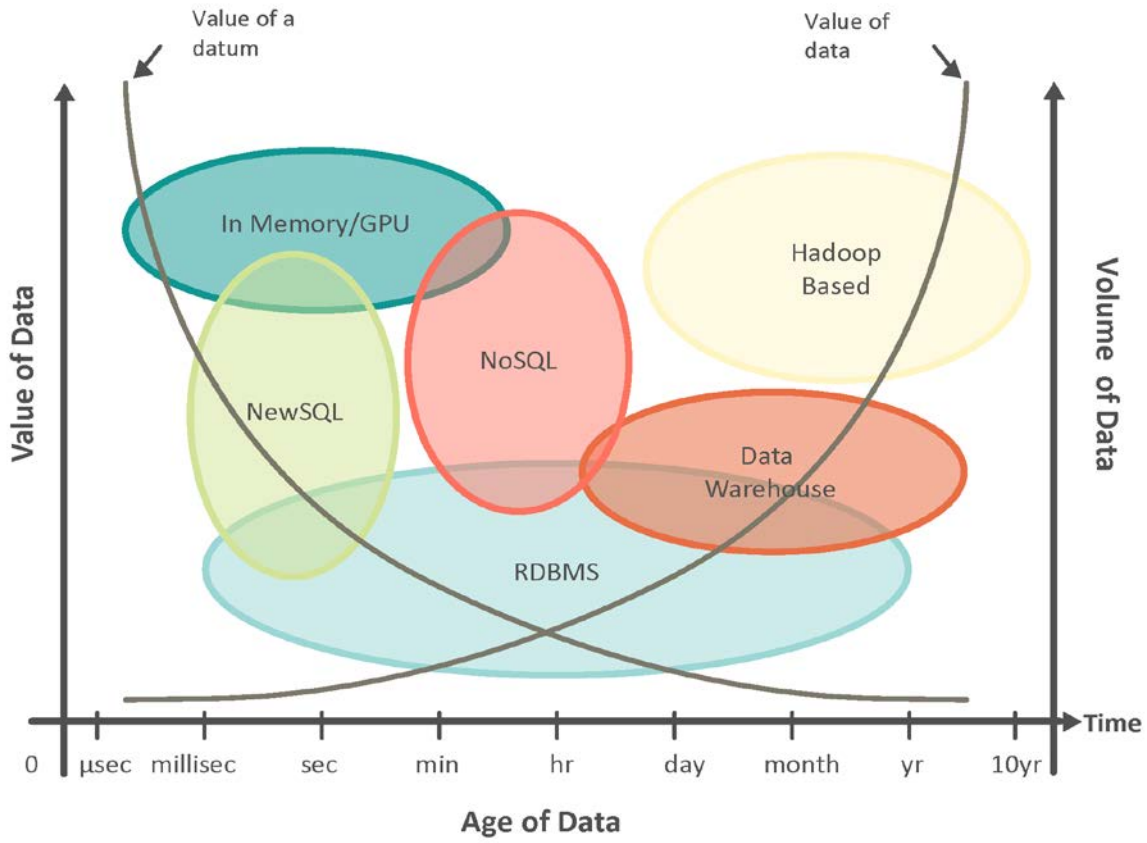
Overview of *Use*



- Includes the actual analyses performed on the data and the development of other data products such as tools, reports, dashboards, visualizations, and software.
- Includes all interactions with the data by end users, analysts, or software programs made to gain some insight or drive some business process.
- Proper management of this process includes educating end users on how best to derive decisions from the data, using effective software development cycles to create new data products, and supporting architecture that allows data to be effectively analyzed where it is stored without unnecessary computational overhead.

Ideal Modern / Big Data Practices – Use

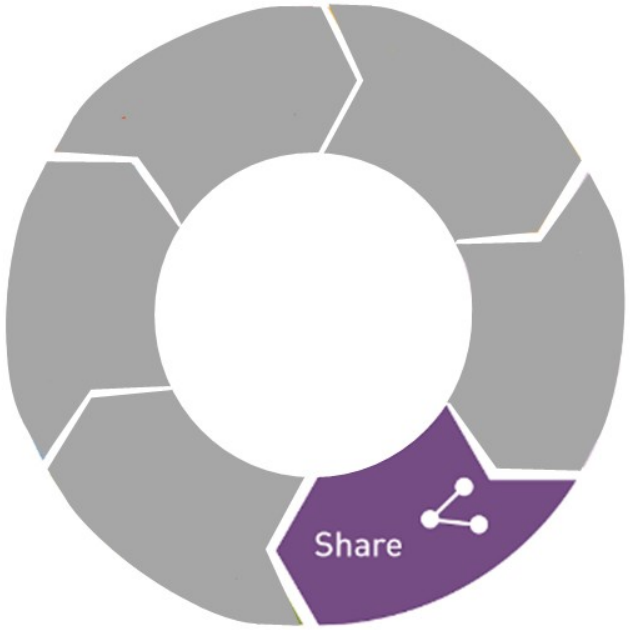
- Traditional data systems are often proprietary, and data analysts are dependent on vendors changes to meet the increasing data and analytics needs.
- Data warehouses (combine/coordinate multiple traditional data systems) were created to cope with the increasing size and complexity of the data.
- RDBMS and data warehouses were able to handle real-time analytics to a point before becoming too costly to operate and too rigid to maintain.
- Hadoop, was the first big data solution designed to run on a large group of servers on which it distributed large-scale historical data analytics.
- Hadoop has been the base model for new data analytics tools capable of handling an ever-increasing amount of rapidly changing data more efficiently and at a lesser cost.



Modern Data Practices– Use

- **Data Analysis:**
  - Data analytics are not performed by one or two tools; many, varied tools are used to meet the needs of individual business areas.
  - Data accuracy and quality of the analytics processes and products are the responsibility of the business area that developed them.
  - Each and every data analysis performs its own custom ETL.
  - Data tools are moved to where the data resides, because data is now too large to be moved around to specialized data processing environments.
  - Distributed algorithms are required to perform data analysis.
  - The nature and limitations of modern data analysis algorithms is well understood.
  - Containerization and microservices are utilized to develop custom data analysis.
  - The ephemeral nature of modern data analysis is understood.
  - Proprietary software is rarely used to develop modern data analysis; cloud provider services or open source solutions are the preferred choice.

Overview of *Share*



- Involves disseminating data to all appropriate internal and external users.
- includes creating an open data policy where appropriate, maintaining updated documentation and other content support, and providing some means by which authorized users may easily access relevant data products.
- Efficient management of this component balances the desire to provide the most use out of the data as possible with concerns over safeguarding privacy, ensuring security, and limiting liability

Modern Data Practices– Share

- Favors an open approach to data sharing with the understanding that “many eyes” are needed to extract value and intelligence from large and complex datasets.
- Access to data products (dashboards, reports) is often not under the same usage restrictions as with traditional systems, which limit the number of users out of necessity.
- The low cost of data compute and the parallel processing capabilities of the cloud, the modern data systems can support many more users (e.g., tens of thousands+).
- Data products can be available as downloads directly from cloud storage or through an API running on the cloud infrastructure.
- As data storage and compute are two distinct services in a cloud environment, external users can be given access to large datasets and process them on the same cloud infrastructure using the tools they choose at their own expense.
- The analysis desired by external users does not compete for resources with the production services.
- These users must be managed, which is not a trivial task.

Modern Data Practices – Share

- Modern data systems do not use restrictive standards (e.g., SOAP, XML); instead they use more modern protocols and formats such as REST and JSON.
- There is little fear associated with sharing large amount of data with external users.
- Corruption of data products is not a concern.
- Data being shared, especially to the public, is devoid of any sensitive data
- Encryption methods used to obfuscate shared sensitive data are updated frequently to give the best protection.
- Different versions of datasets are created based on who they need to be shared with.
- Beyond sharing data, data analysis process code is also shared.
- In addition to sharing historical data, live data streams are established.
- The responsibility of providing data services and products to external users is shared between the central governance and the business areas who own the products.
- External users allowed to access the data are clearly identified and tracked.

Traditional vs. Modern Management

Traditional Approach	Modern Approach
Systems are designed and built for a pre-defined purpose; all requirements must be pre-determined before development and deployment.	Systems are designed and built for many purposes; constant adjustments are made to the system following deployment
System features at the hardware level; hardware and software tightly coupled.	System features at the software level; hardware and software decoupled.
System designed as “set and forget” – designed once to last for many years.	System designed to expect changes (ephemeral); sees changes and adjusts automatically.
Systems are rigid and not easily modified.	Systems are flexible and can easily adapt to changes.
As technology evolves, hardware becomes outdated quickly; system can’t keep pace.	As technology evolves, hardware is disposable, system changes to keep pace.
Schema on write (“schema first”) – data is written to a storage location according to the pre-defined schema; requires extensive data modeling upfront to try and cover all datasets important to everyone using the data.	Schema on read (“schema last”) – data is loaded as-is and applied to a schema as it is pulled out of a stored location (i.e., read), rather than as it goes in/is written; requires no upfront modeling exercise, and users are not stuck with a one-size-fits-all schema.
Data and analyses are centralized (servers)	Data and analyses are distributed (cloud)
Data is heavily tied to IT system & pre-defined schema, software, & hardware.	Data is dissociated with IT and is not tied to a pre-defined schema, software, & hardware.
80% system design and maintenance, 20% data analysis	20% system design and maintenance, 80% data analysis
Dollars are spent on hardware.	Dollars spent on data and analyses.
Data governance is partly handled by system design (doesn’t have to be controlled – system won’t even allow data to come in); IT controls who sees / analyzes data (heavy in policy-setting).	More complicated data governance; open data to a lot of people; data manager needs to monitor users in real-time, what data they read, how many resources they consume, what data products are being produced, how good they are – all this across many, many users.
Uses a tight data model and strict access rules aimed at preserving the processed data and avoiding its corruption and deletion.	Consider processed data as disposable and easy to recreate from the raw data. Focus instead is on preserving unaltered raw data; making it available to all users as read-only; and using data processing methods that can be saved, rebuilt, and redeployed on the fly so that any lost or corrupted data can be recreated directly from the raw data.
Small number of people with access to data; limits use of data for insights and decision-making to a “chosen few.”	Many people can access the data; applies the concept of “many eyes” to allow insights and decision-making at all levels of an organization.