WHAMTECH®

Stop your Data Lake from turning into a Data Swamp and turn it into a well-governed Virtual Data Reservoir, aka Virtual Data Warehouse*

*With options for near real-time, distributed and federated data access to other data sources

January 2021

WHAMTECH®

A well-governed Virtual Data Reservoir could be considered a Virtual Data Warehouse (VDW) and a replacement for an Enterprise Data Warehouse (EDW)

Ten questions to ask before designing an analytics* architecture

1. Can all data be copied to a single location, e.g., a cloud database?
   - Usually, requires strong C-level support and commitment to enforce
   - What about on-soil data retention regulations or third-party data ownership?
2. If question 1 is not possible, can some data be copied to a single location and some remain on-premise, in data centers, multiple clouds and/or third-party systems, e.g., SaaS?
   - Hybrid Cloud
3. Do the original data, source schemas and formats need to be retained?
   - Regulatory compliance, trust, audit, etc.
4. Does data need to be kept up-to-date, e.g., in near real-time?
5. Do data updates need to be monitored and events processed, e.g., operational dashboards, workflows, alerts, etc.?
6. Besides analytics*, are there other uses for the data, e.g., operations or feedback to operations?
7. Are data management, including security and governance, and master data management, addressed, assuming that it is needed?
8. Is self-service by business users needed?
9. Are access control and data security in-place?
10. Do CCPA/CCPR, GDPR and other data compliance need to know where all personal data is and how it is used?

*Includes reporting, BI and all types of analytics

Most Big Data and Data Lake projects fail

From Gartner(Ref. 1) :

- 85% of Big Data projects failed to go beyond piloting and experimentation, and are abandoned
- Main reason for failure is that most companies do not know how to move Big Data programs from pilot to business cases that turn data into business value

From IDG CIO Review(Ref. 2):

- 75% of business leaders feel that future success will be driven by organizations that make most out of their information assets, whereas in reality, only 4% are set up for such success

Elsewhere:

- After over 10 years of Hadoop being publicly available, the two largest Hadoop support vendors, Cloudera and Hortonworks, were losing money and merged, and MapR has only a relatively small number of large marquee customers, which implies that the platform is also failing to deliver success
- On the positive side, cloud vendors, such as Azure, AWS and Google, are offering their own commercial versions of Hadoop and other Big Data systems, which allow easy and rapid provisioning for development, testing and, eventually, production

⇒ The question is how to monetize the value of Big Data and Data Lakes, by integrating them into the organization's business operations?

Reasons why Big Data and Data Lakes fail

Data Lakes are more like Data Swamps instead of Data Reservoirs, as they are missing data …

- Discovery/profiles
- Classification
- Security
- Metadata
- Quality
- Transformation
- Standardization
- Standard view(s)
- Relationships/links to other data
- Master data
- Governance
- Access control

Data warehouses, populated through ETL, were developed decades ago to address these data-related essentials

WHAMTECH®

Pros and cons of a Data Lake and a data warehouse

Feature	Data Lake	Data warehouse
Data type and format	All – structured, semi-structured and unstructured	Structured only
Volume	Everything of possible interest – large volume	Usually limited to defined application use – smaller volume
Schema-related processing	Schema-on-read	Schema-on-write
Storage	Medium performance commodity – low cost	High performance RAID – high cost
Flexibility	High - configuration up to application/user	Limited – usually have to copy data to data marts
Data pre-processing *Not necessarily MDM	None – copied "as is" from sources – need additional processes	ETL addresses most data aspects, including data quality and deduplication*
Standards	Limited or none – need additional layers	Standard data model, drivers and SQL
Security and privacy	Limited or none – need additional layers	High – access control and data anonymization
Users	Data scientists/analysts	Business users

Drivers for Data Lakes

- Run analytics on as much data as possible to gain insight into the business – not knowing what data is of value
- Increasingly large volume and number of data types available
- Data warehousing too complex, too expensive, inflexible and cannot deal well with a variety of data types, in particular, unstructured data
- Availability of indefinitely scalable, low-cost commodity hardware and simple cloud provisioning
- Quick and easy to instantiate and add/remove data sources and components
- Highly flexible range of uses – from discovery to analytics
- Overcome poor access to original data sources
- Support for high query performance
- Support for large number of concurrent users
- Use for archive, backup and/or redundancy

EOS

WHAMTECH®

Hurdles to overcome for Data Lakes

- No built-in data processing, data management or master data management
- Multiple copies of data subsequently made from a Data Lake to eventually land curated and useful data in an analytics environment
- CIO-level or similar high-level imposition needed to obtain ALL data, especially across organizational borders and from third-parties
- Data ownership concerns
- On-soil/on-premise data retention requirements
- Security and privacy concerns
- Replicated data/storage concerns - costs can quickly add up, especially in the cloud
- Internally perceived competition to data warehouse, therefore, usually sold as something different and not core to the organization's business, which leads to failure
- Difficult to integrate Big Data and Data Lakes with organizations' business operations

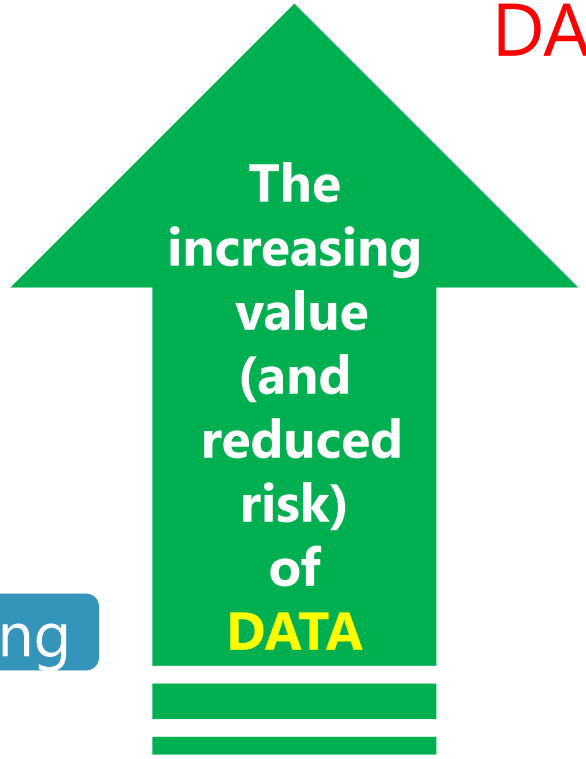
EOS

Data management fundamental truths

- Each data source is designed and created for a specific application(s), which is fit-for-purpose, and both data source and application(s) are generally siloed
- Each data source may have its own version of data governance and standards, but unlikely to be organization-wide
- Significant data management and master data management processes are needed to make data readily useable by a business user, regardless of being in a:
  - Data warehouse
  - Big Data store/Data Lake
  - Conventional data virtualization/federation system
- Master data is the glue that enables integration of data in a single data source and multiple disparate data sources
  - Each data source is its own self-standing ontology/semantic data model
  - Master data resolves and unifies entities within and across data sources using data from these sources
  - Logical relationships enable entities and other data to be integrated, and are based on physical relationships existing within and across data sources
  - Ultimately, master data enables all data sources to be viewed as a single ontology/semantic data model

Processes used to increase the value of DATA

DATA analytics
DATA integration
Master DATA management
DATA relationships
DATA management
DATA security
DATA classification
DATA discovery and profiling
Raw DATA



DATA protection regulations

DATA access control

DATA governance

Source: #WhamTech SmartData Power of Indexes

Most Data Lakes are actually Data Swamps

THE IDEAL DATA LAKE IS A WELL-GOVERNED DATA RESERVOIR, AKA DATA WAREHOUSE

From Data Swamp to clean, well-governed Data Reservoir

Options to do so include:

- Data scientist/analyst spend up to 90% of their time sorting through and making sense of the data
- Read and copy data from the Data Swamp through multiple processes to eventually land in a curated data store – recent 6x copy example with each copy involving multiple backups, i.e., up to 18 copies
- Conventional ETL – similar to a data warehouse, with all its associated hurdles
- Or…

⇒ Leave data in Data Lake "as is", but superimpose a data fabric with data virtualization and federation on top that overcomes all data source, data, access control and data security, and compute deployments issues

Option 1: Proposed basic Data Lake to Data Reservoir

A layer on top of a Data Lake with advanced SQL processing, virtual data management, master data management, event processing, graph database, link analysis and highly curated data provisioning

- Standard drivers (e.g., ODBC and JDBC) and web services (e.g., REST APIs)
- Standard applications, e.g., reporting, BI, analytics, CRM and BPM

⇒ Avoid a Data Swamp and turn a Data Lake into a well-governed Data Reservoir

Option 1: Data Lake + Virtual Data Management = Virtual Data Reservoir

+

VIRTUAL

Data Discovery
Data Indexing
Data Identification
Data Classification
Data Security
Data Quality
Data Standardization
Data Linking/Relationships Mapping
Master Data Management
Data Governance
Graph Database
Link Analysis

=

VIRTUAL DATA RESERVOIR

Option 2: Proposed near real-time operational updates to Data Lake

Maintain copies of operational data in near real-time (NRT) in Data Lake - minimize latency

- Retain original data "as is" for accountability, auditability and regulatory compliance
- As updates occur, monitor and analyze data sources, process events and interact with operational systems

⇒ Close the gap between operational systems and conventionally detached reporting, BI, analytics CRM and BPM systems

Option 3: Proposed Distributed Data Lake to Distributed Data Reservoir

A Distributed Data Lake (DDL) with distributed virtual data management, master data management, etc. layer on top = Distributed Data Reservoir (DDR)

- Maintain copies of source data in multiple Data Lakes
- Enable an integrated single view of data, including single person and other entity views
- Allow independent local/regional operations and applications
- Locally/regionally scale to support performance for data volume and velocity
- Combine with near real-time updates for NRT-DDR

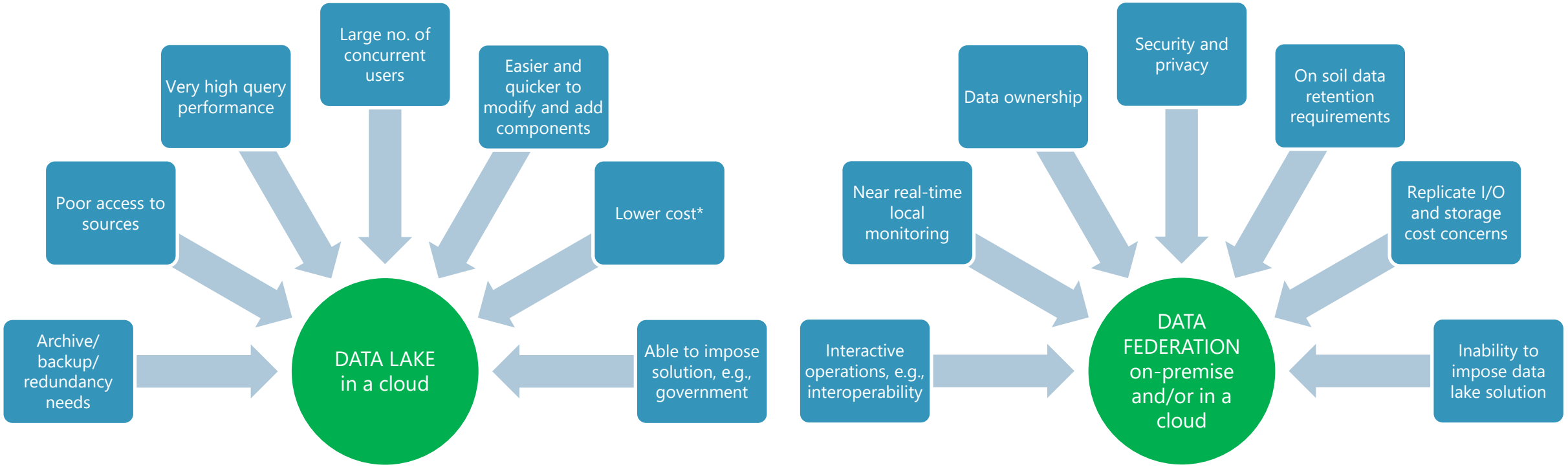
⇒ Allow regional and even local Data Lakes of perhaps different data types to be combined and viewed as a single large Data Reservoir

Option 4: Proposed Distributed Data Lake to Distributed Data Reservoir + federated data

A hybrid near real-time Distributed Data Reservoir (NRT-DDR) + federated data access

⇒ Enable the inclusion of other data sources within the cloud, multi-cloud, on-premise, in data centers and third-parties, e.g., Salesforce

Reasons for Data Lake vs. Data Federation



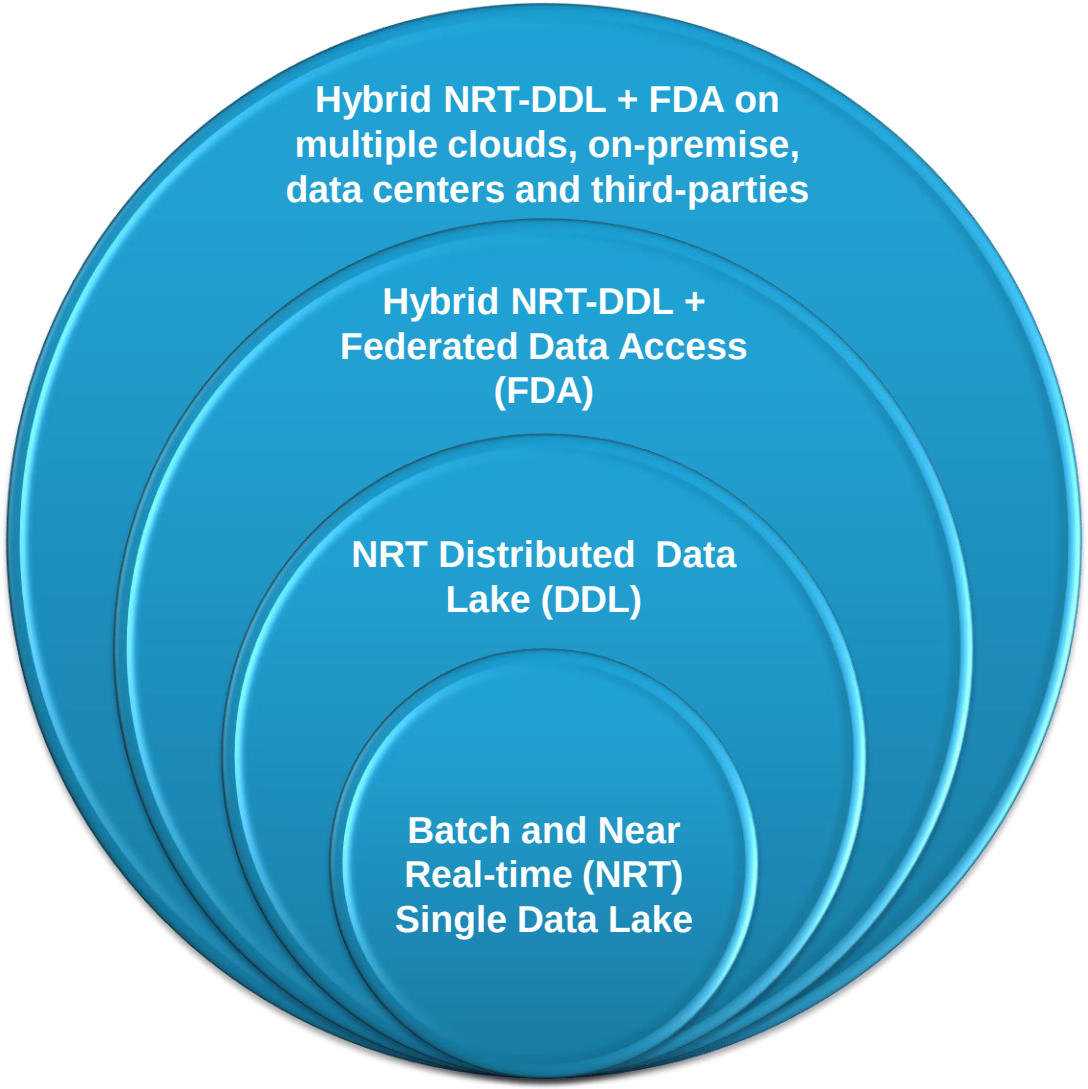
*Lower cost, as less data access work, fewer adapters, less admin, maintenance and support, cloud resources provisioning, less component tie-in work, etc.

Option 5: Proposed multiple cloud and other sources Data Reservoir

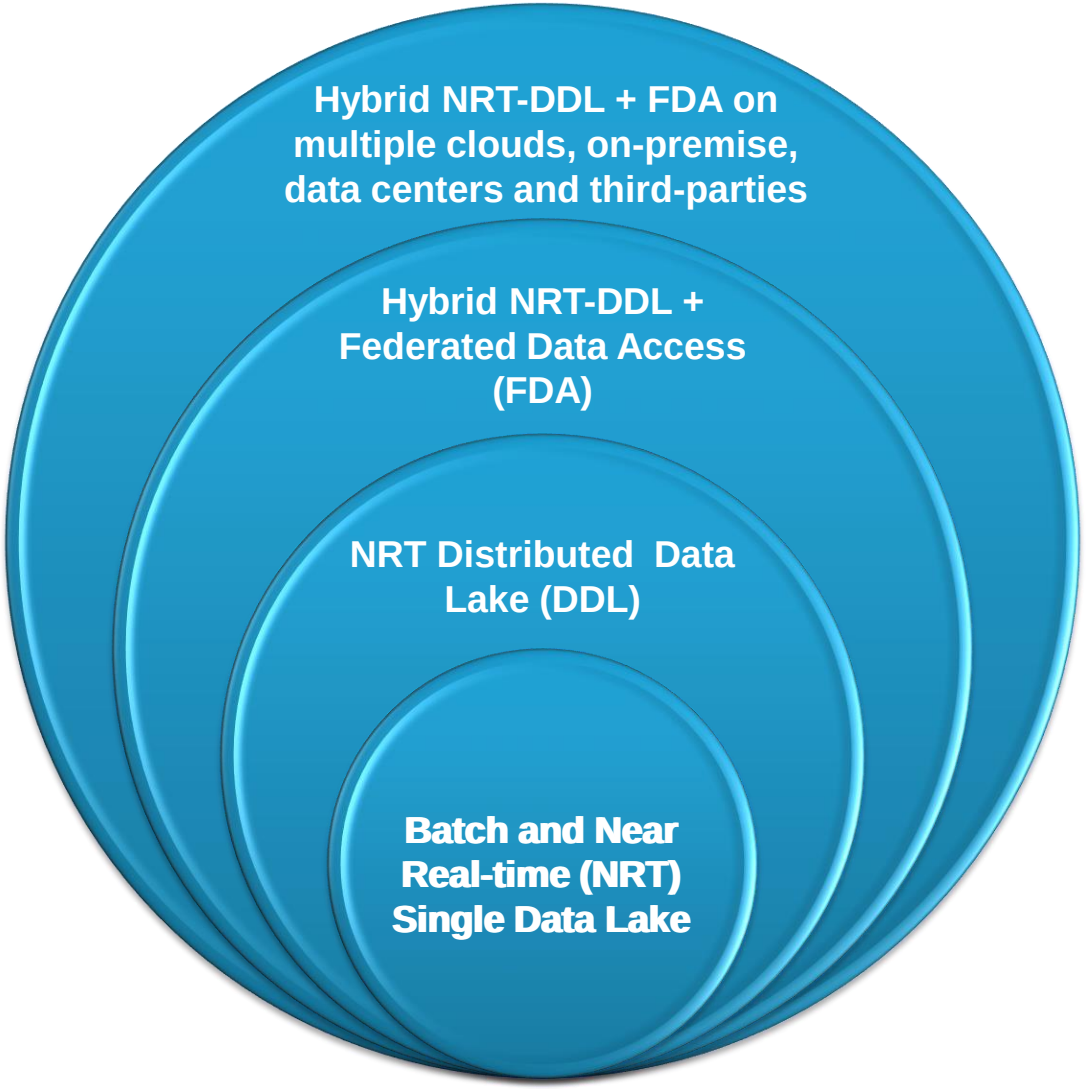
A hybrid near real-time Distributed Data Reservoir (NRT-DDR) + federated data access across multiple clouds, multiple Data Lakes and multiple other data sources

⇒ Acknowledge that customers have data in many locations – not just a single Data Lake in a single cloud

All options: Increasing range of data sources for Distributed Data Lake



All options: Distributed Data Lake + Virtual Data Management = Virtual Distributed Data Reservoir



\+

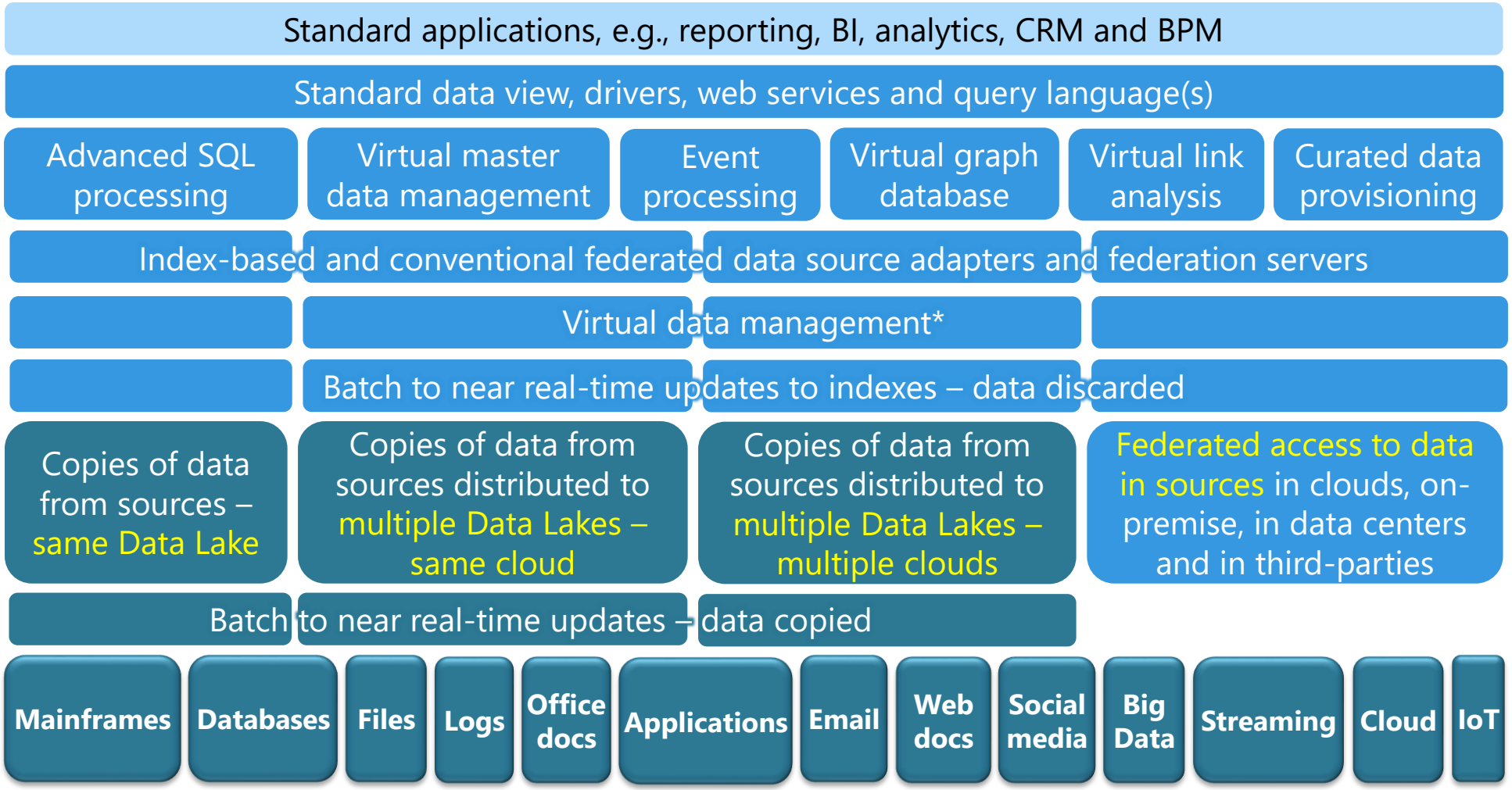
VIRTUAL

Data Discovery
Data Indexing
Data Identification
Data Classification
Data Security
Data Quality
Data Standardization
Data Linking/Relationships Mapping
Master Data Management
Data Governance
Graph Database
Link Analysis

=

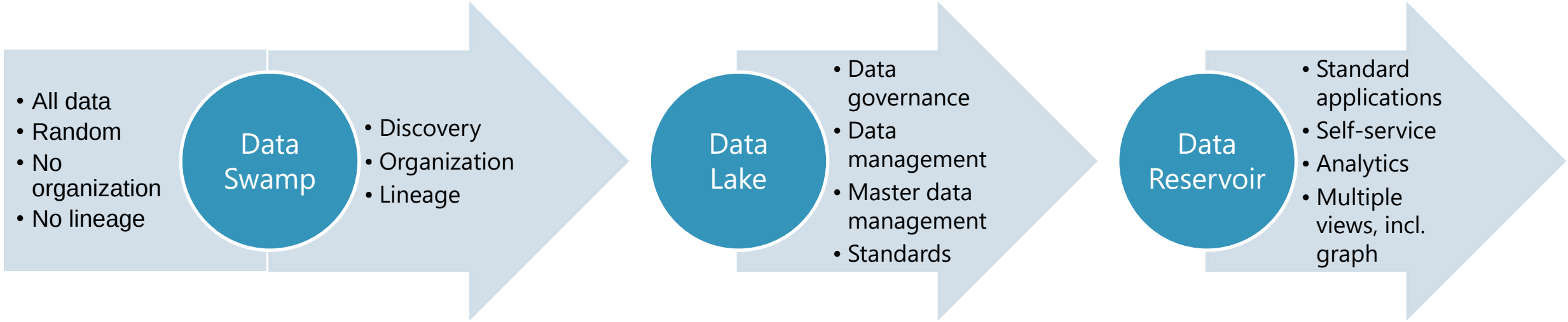
VIRTUAL DISTRIBUTED DATA RESERVOIR

Multiple options for Data Lake to Data Reservoir



***Data management = data discovery, indexing, identification, classification, security, cleansing, transformation, standardization, mapping to standard data view and linking/relationship mapping**

Proposed options: Data Value



Proposed options: Data Volume

- As much data as possible in one cloud instance
- Any data not in cloud excluded
- Needs strong cooperation and enforcement

Single and Centralized

- Multiple Data Lakes/ Reservoirs
- On soil data retention/ region-specific
- Security and privacy
- Regional data ownership
- "Cloud Hub"

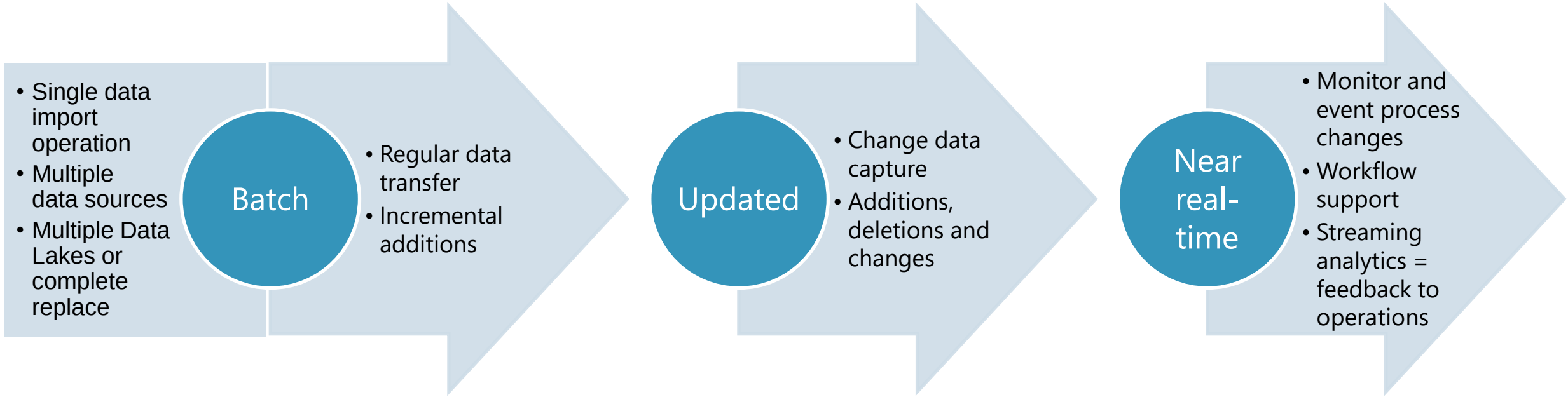
Distributed within and across Multiple clouds

- On-premise data
- Include other large data stores, e.g., EDWs
- Advanced security and privacy
- Local data ownership

Distributed + Federated Data + Third-party

- Complete customer, patient and other master data views
- Allows inter-operability with data source systems
- Simpler to monitor and event process changes

Proposed options: Data Velocity



WHAMTECH®

Hadoop as a Data Lake

- Hadoop is an ideal Data Lake as:
  - Magic sauce is Hadoop Distributed File System (HDFS) that is indefinitely scalable, e.g., Azure Data Lake
  - Can accommodate almost any data type, file, document/object, etc.
  - Supports multiple databases and types on top, e.g., Big Table/NoSQL, document/object store and SQL/RDBMS
  - Supports multiple tools to import or stream data into Hadoop, e.g., Flume, HDF/NiFi, SiriusIQ and Sqoop
  - Unique IDs to reference records, files and documents/objects in Hadoop
  - Good at simple (non-SQL) queries
- Hadoop, however, is NOT:
  - A database, per se
  - Able to address data management, master data management, data governance, etc.

WHAMTECH®

Answers to initial questions on running analytics*

Question No.	Question	Centralized Data Lake	Centralized Data Warehouse	Distributed Data Lake + Indexes, incl. Data Federation
1	Can all data copied to a single location, e.g., cloud database?	✓	✓	✓ but not necessary
2	Can some data be copied to a single location and some remain on-premise, in data centers and/or on third-party systems, e.g., SaaS?	✘ [✓ ZDP]	✘	✓
3	Does data exist and remain on multiple clouds (maybe as well as elsewhere)?	✘ [(✓) ZDP]	✘	✓
4	Do the original data, source schemas and formats need to be retained?	✓ [✘ ZDP]	✘	✓
5	Does data need to be updated in near real-time?	(✓) with near real-time updates	(✘) maybe, with near real-time updates	✓
6	Do data changes need to be monitored and events processed, e.g., operational dashboards, workflows, alerts, etc.?	✘	✘	✓
7	Besides analytics*, are there other uses for the data, e.g., operations?	✘	✘	✓
8	Are data management, incl. security and governance, and master data management addressed?	✘ [(✓) ZDP]	✓ ETL into	✓
9	Are automation and self-service by business users needed?	✘ [(✓) ZDP]	✓	✓
10	How important are access control and data security? What about GDPR?	✘ [✓ ZDP]	✓	✓

*Assumes includes reporting, business intelligence and all types of analytics
ZDP = Zaloni Data Platform exception to conventional Data Lake as they prep data before landing

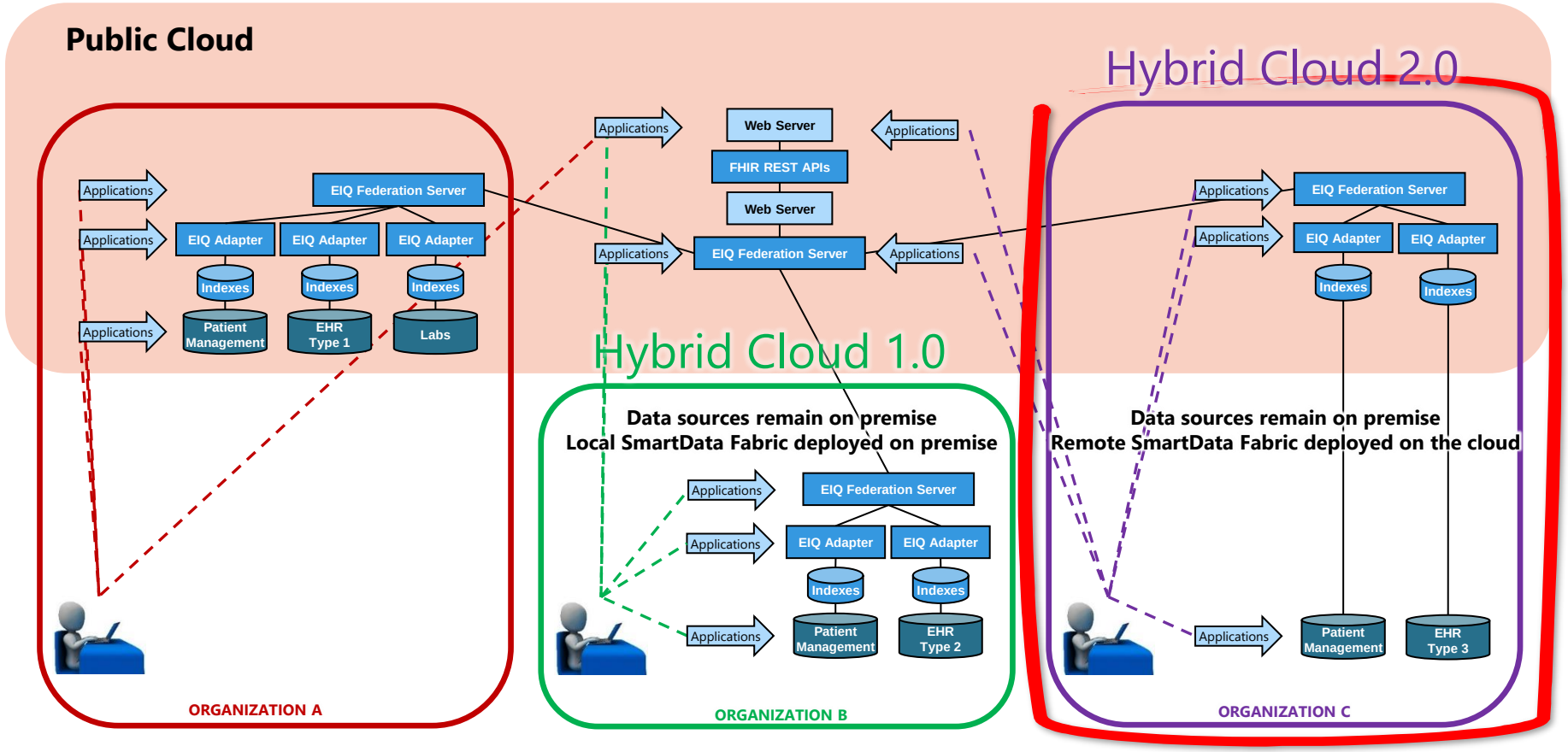
WHAMTECH®

Detailed comparison among data access approaches

Requirement Addressed	Data Warehouse	Centralized Data Lake	Index-based Distributed Data Lake (DDL)	Conventional Data Federation	Indexed-based Data Federation	Index-based DDL + Data Federation
Archive, backup and/or redundancy	(✖) Not original data	✓	✓	✖	(✓) Index inversion	(✓)
Poor access to data sources	✓	✓	✓	✖	(✓) Index inversion	(✓)
Query performance	Very High	Very High	High	Poor	Good	Good to High
Large number of concurrent users	✓	✓	✓	✖	✓	✓
Built-in data processing and management	(✖) Needs ETL	✖	✓	(✓)	✓	✓
Easy to instantiate and add/remove/modify components	✖	✓	(✓)	✖	(✓)	(✓)
Low cost	✖	✓	(✓)	✖	(✓)	(✓)
Less need for CEO-level, government or similar imposition to implement	✖	✖	(✖)	✓	✓	✓
Near real-time local data source monitoring	✖	✖	(✓) Near real-time updates	(✓) With polling	✓	(✓)
Data ownership concerns	✖	✖	(✖)	✓	✓	(✓)
On-soil/premise data retention	✖	✖	(✖)	✓	✓	(✓)
Interaction with data sources, e.g., interoperability	✖	✖	✖	✓	✓	(✓)
Security and privacy concerns	(✖)	✖	✖	✓	✓	(✓)
Replicated data/storage concerns	✖	✖	✖	✓	✓	(✓)

Example of a Hybrid Cloud Deployment with a twist

- **Single Patient 360 View across all data sources, regardless of their type, location and data – seamless and automatic integration through FHIR REST APIs**
- **Hybrid Cloud solution - data sources remain where they are – most on premise, some in the cloud**
- **FHIR standard data views and APIs through EIQ Adapters and EIQ Federation Servers**
- **Multiple application access levels – (a) direct local, as before, (b) through individual EIQ Adapters, (c) organization EIQ Federation Servers, (d) cross-organization EIQ Federation Server and (e) cross-organization FHIR REST APIs as data services**
- **All SmartData Fabric EIQ Products managed as services through the cloud or remotely for Organization B**



References

1. https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/
2. https://www.cio.com/article/3003538/big-data/study-reveals-that-most-companies-are-failing-at-big-data.html

The End