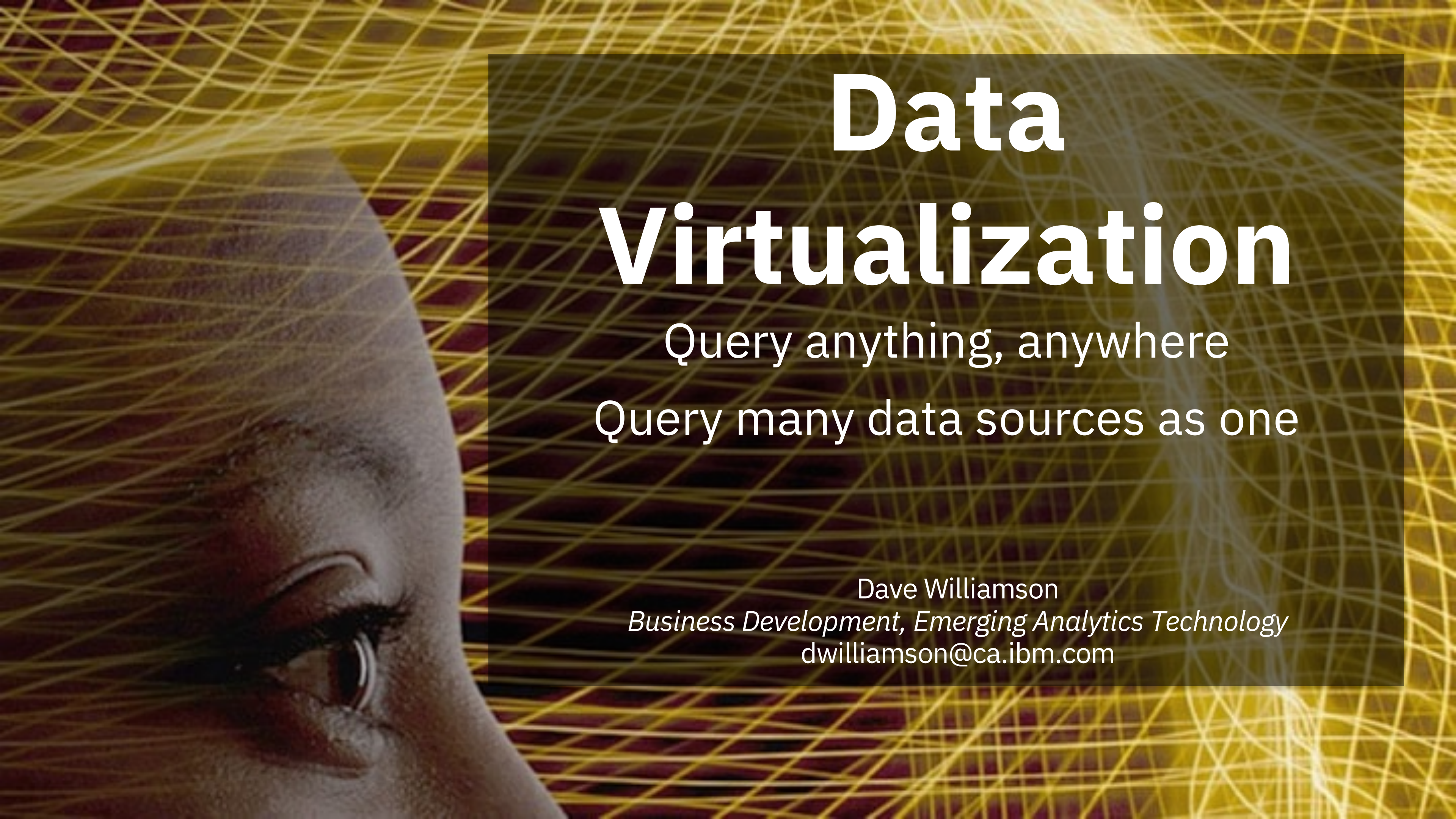


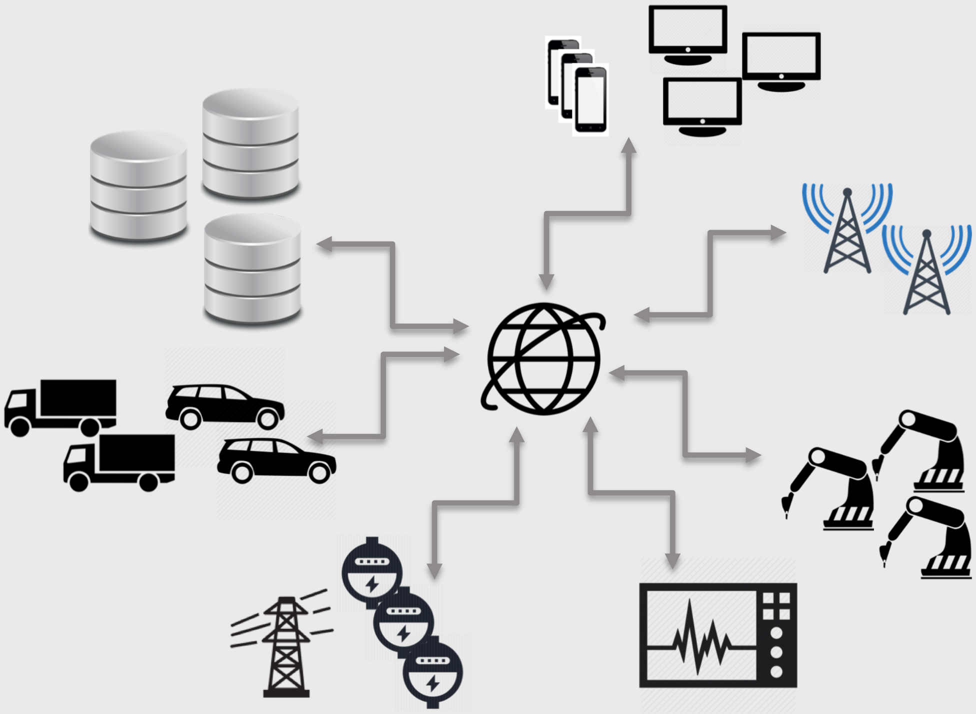
Data Virtualization

Query anything, anywhere

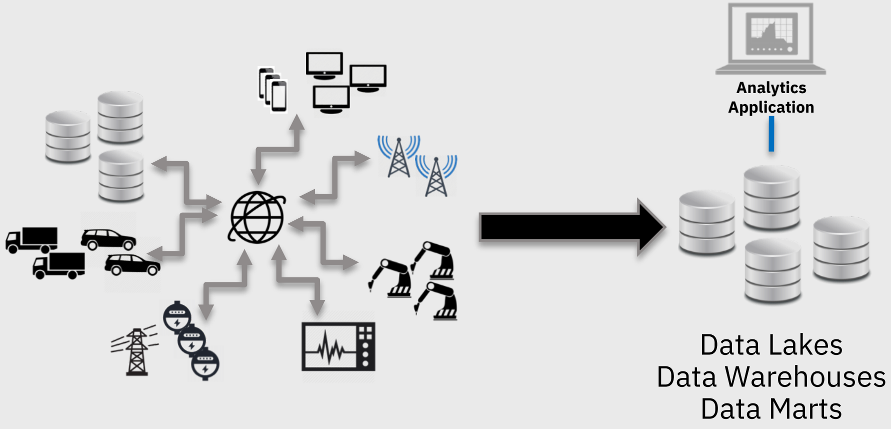
Query many data sources as one

Dave Williamson
Business Development, Emerging Analytics Technology
dwilliamson@ca.ibm.com

Data is Everywhere and increasingly heterogeneous



Performing Analytics Today



Costly and Complex

High Latency from source to use

Does not meet Business need

Compute resources at source not utilized

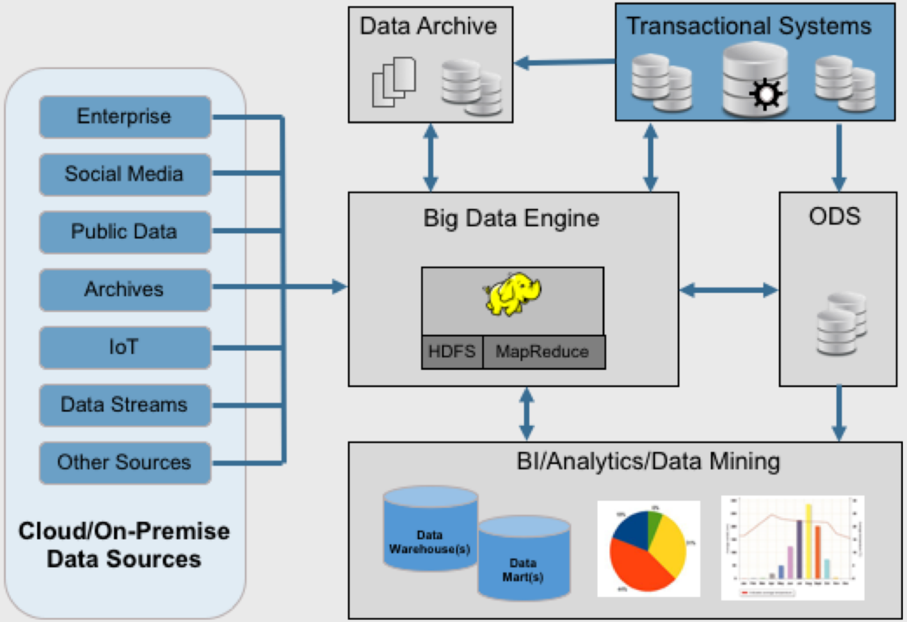
Error prone, data integrity challenges

Applications expect homogeneity

Not scalable

Not all data needs to be moved or copied

Resulting Data Architectures



Numerous ETLs

Unnecessary duplication, replication

Data governance issues accelerating

What is Data Virtualization?

The ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location

A new approach to Data Virtualization

Now in beta trial. Coming to ICP for Data in November

1

Query anything, anywhere.

Query **many heterogenous data sources as one** across cloud, on-premise and mobile with advanced analytics using the most popular languages and tools

2

Simplicity and scalability.

Automatically discover, and connect **few to many devices and data stores** into a single self balancing constellation. Avoid the complexity of centralized copies. Data only persists at the source.

3

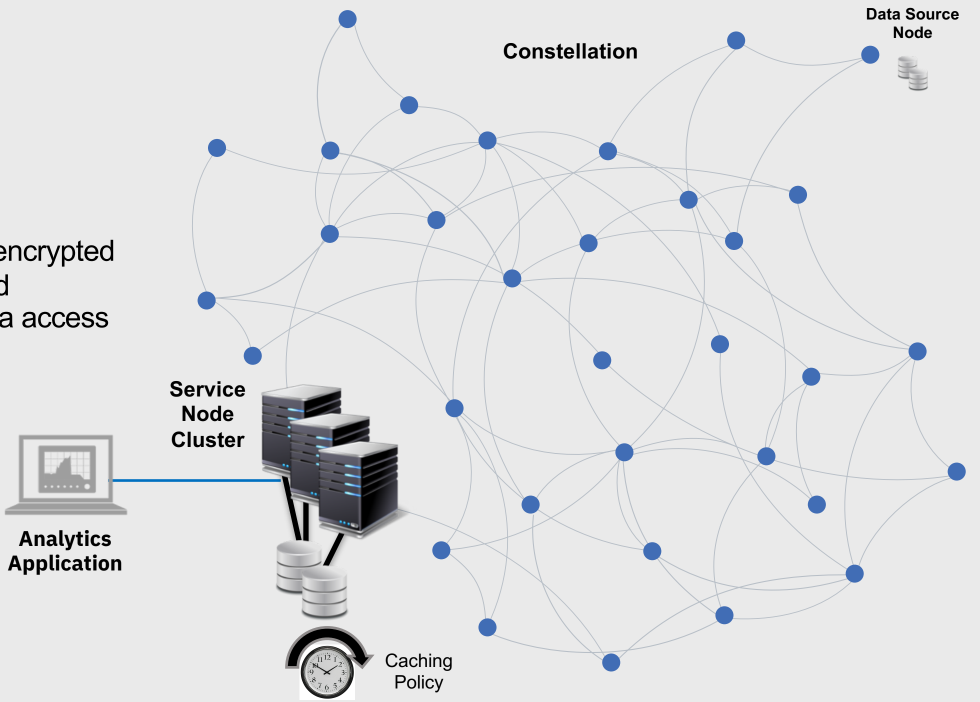
Execution speedup.

Many times acceleration using the power of every device to compute and aggregate results.

4

Security.

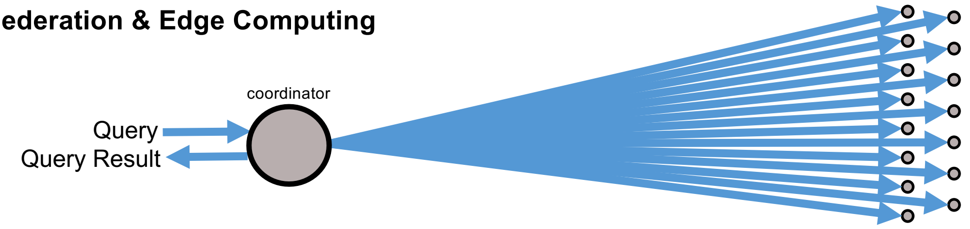
Fully secure and encrypted communication and preservation of data access rights at source.



IBM Cloud / © 2018 IBM Corporation

What is fundamentally different?

Classic Federation & Edge Computing



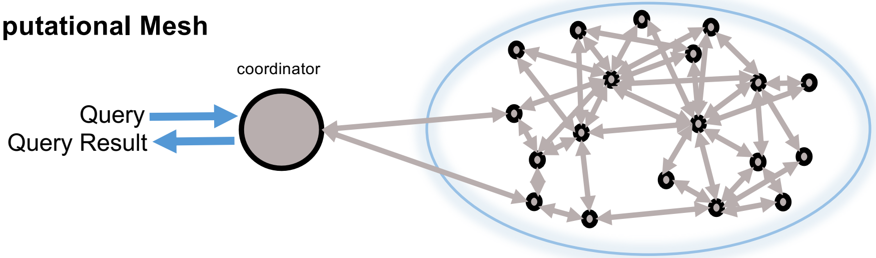
Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes individually perform as much work as they can based on their own data. Individual results are sent back to the coordinator for final merging and remaining analytics.

Coordinator receives intermediary results from all edge nodes, merges results, and performs remaining analytics

New Computational Mesh



Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes self organize into a constellation where they can communicate with a small number of peers. Nodes collaborate to perform almost all analytics, not only analytics on their own data.

Coordinator receives mostly finalized results from just a fraction of nodes. Completes the final work for the query result.

Supported Languages & Data Sources

Query Languages	
SQL (ANSI)	✓
SQL (Oracle)	✓
SQL (DB2)	✓
SQL (PostgreSQL, Netezza)	✓
Spark SQL	✓
SQL Python	✓
SQL in R & SparkR	✓
PL/SQL (stored procedures)	***Future***
SQL PL (stored procedures)	***Future***

Mix Any Combination of Data Sources			
Oracle	✓	Excel	✓
Db2 (software, appliance, cloud)	✓	CSV (delimited text)	✓
Netezza	✓	Cloudera	✓
PostgreSQL	✓	Teradata	✓
Informix	✓	Db2/Z	✓
MySQL	✓	MongoDB	***Future***
SQLServer	✓	DVM	***Future***
DerbyDB	✓	Redis	***Future***
Big SQL	✓	Cloudant	***Future***
Db2 Event Store	✓	Greenplum	***Future***

ICP for Data

Use Case: **Manage All Your Data** – regardless of what it lives

IBM Data Virtualization

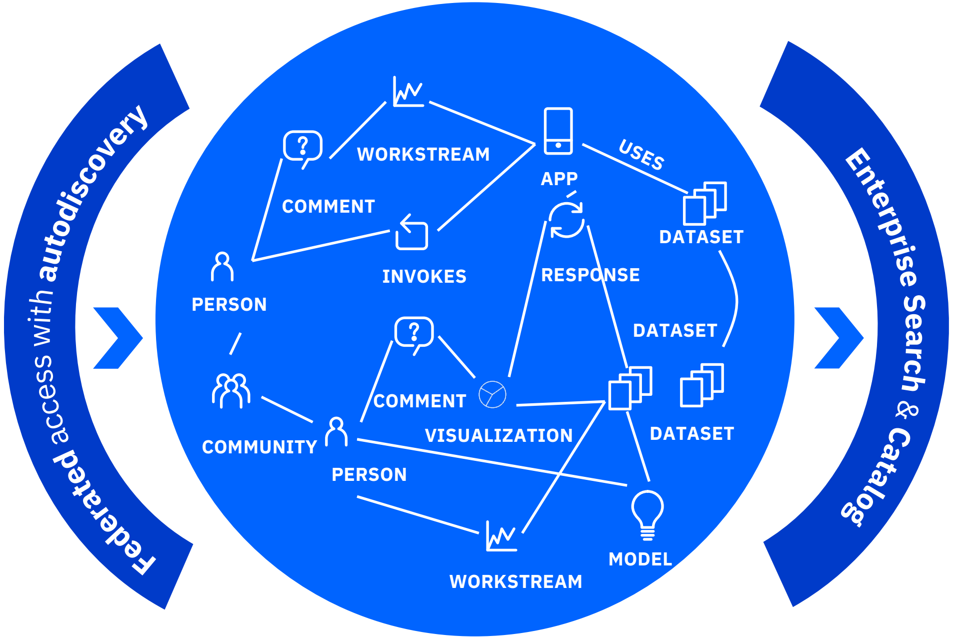
Your Data Anywhere

On-Premises

Private Cloud

Public Cloud

Structured & Unstructured Data



Governance of Consumers and Producers

With lineage, context & metrics

End Users

Data Engineer

Data Scientists

Business Analysts

App Developers

IBM

http://queryplex.com
info@queryplex.com