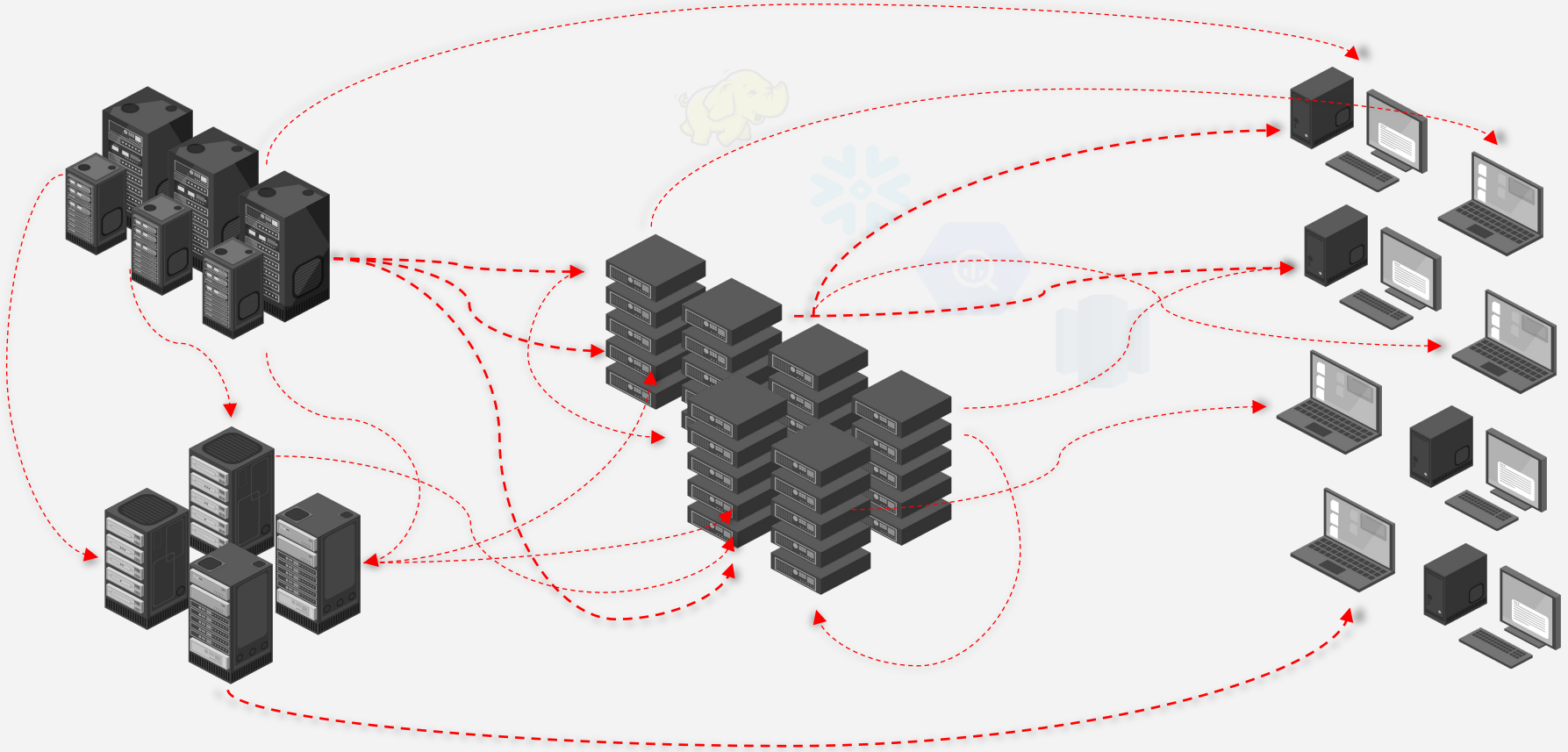
Cloud Pak for Data

Binu Midhun

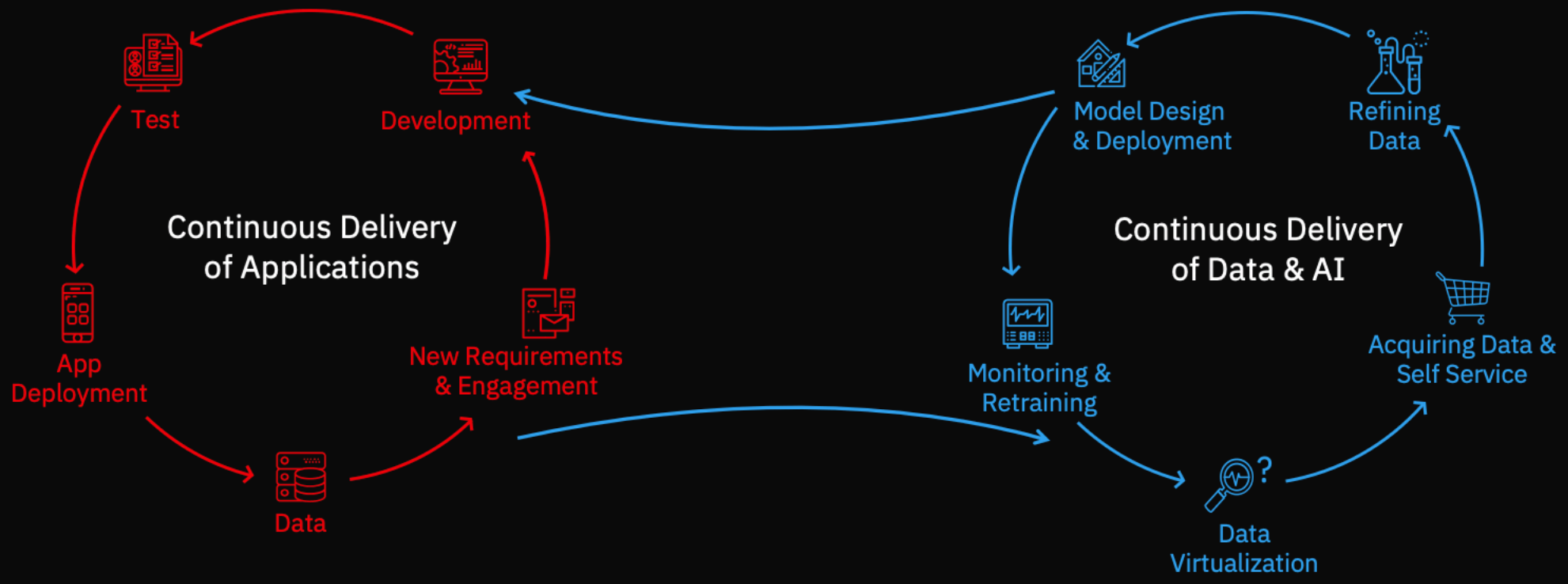
IBM Developer Advocate

New challenges with data centric delivery…

...are associated with inefficient **information architecture**



Transforming the Delivery of Data & AI to the Business



> Hybrid Multicloud

> Security

> Unified User Experience

What is Cloud Pak for Data

Cloud Pak for Data

Everything you need for enterprise data and AI, on any cloud

Infuse

Analyze

Organize

Collect

Business Users and Applications

Comprehensive Information Architecture Design
Unified experience for all your capabilities

Any Cloud
Infrastructure / Data Lake / Data Sources

aws On-Prem

IA (Information Architecture) for AI and Cloud - The AI Ladder

A prescriptive, proven approach to accelerating the journey to AI

IBM Ladder to AI

Infuse – Deploy trusted AI-driven business processes

Analyze - Scale insights with AI everywhere

Organize – Create a trusted analytics foundation

Collect – Make data simple and accessible

Requires a strong foundation that is built on a modern cloud-native architecture
with underlying infrastructure that is highly performant, secure, reliable and cloud-ready

IBM Cloud Pak for Data

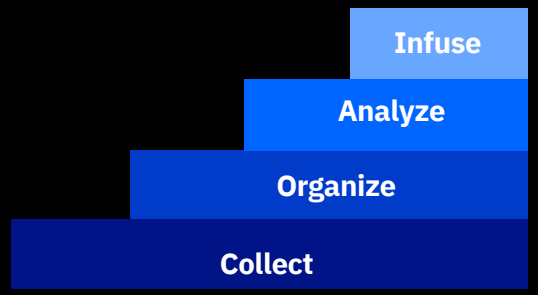
Open, Extensible Platform **Personalized, Collaborative Platform**

App Developers and SREs| Business Partners | Data Engineers | Data Stewards | Data Scientists | Business Users

APIs | Integrated User Experience

Extensible: partner ecosystem, accelerators, and solutions

The AI Ladder



Collect and connect

- Data virtualization
- Provision SQL and NOSQL Databases
- Warehouses and Marts
- Event Ingestion and Streaming Analytics
- Distributed compute – Apache Spark

Organize and integrate

- Discovery and search
- Data transformation
- Data cataloging and Classification
- Business glossary
- Policies, rules and privacy

Analyze and infuse

- Data Science and Visualization
- Dashboards and reporting
- AutoAI, ML deployments and operations
- AI Trust and Transparency - Explainability and Bias detection
- AI services – Chat, NLU

Core services

- Logging
- Monitoring
- Metering
- Persistent Storage
- Kubernetes
- Security
- Identity Access Mgmt.
- Docker Registry / Helm

Red Hat OpenShift

IBM Cloud | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

The Cloud Pak for Data Platform

1. Services Ecosystem

With a click, access and deploy an ecosystem of 45+ analytics services and templates from IBM and third parties.

2. Data Virtualization

Quickly and easily query across multiple data sources without moving your data.

3. Platform Interface

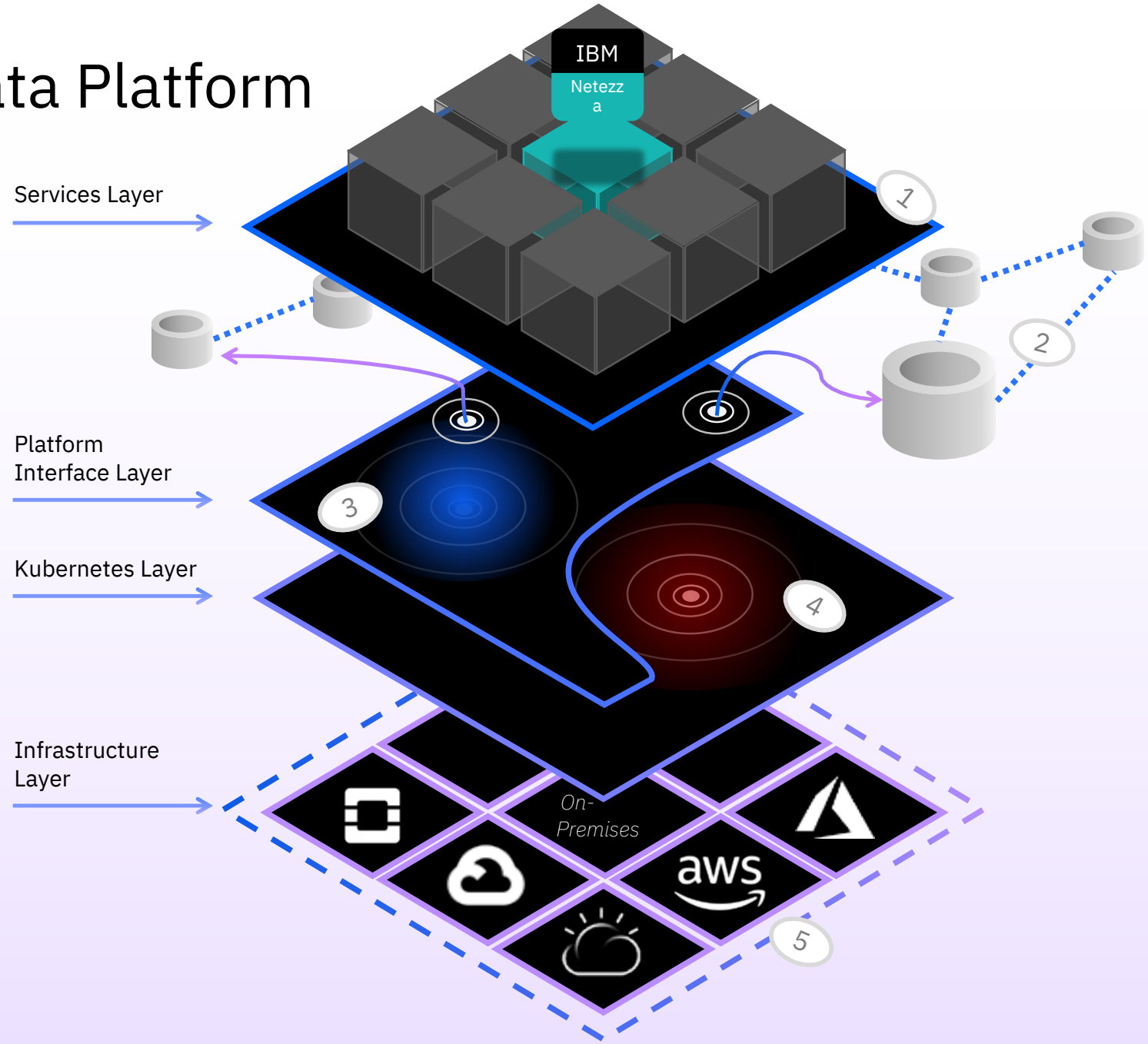
Speed time-to-value with a single user experience that integrates data management, data governance and
analysis for greater efficiency and improved use of resources.

4. Red Hat **OPEN**SHIFT®

Leverage the leading hybrid cloud, enterprise container platform for an innovative and fast deployment strategy

5. Any Cloud

Avoid lock-in and leverage all cloud infrastructures with our multi-cloud approach.



SAME

public cloud

hybrid cloud

redhat

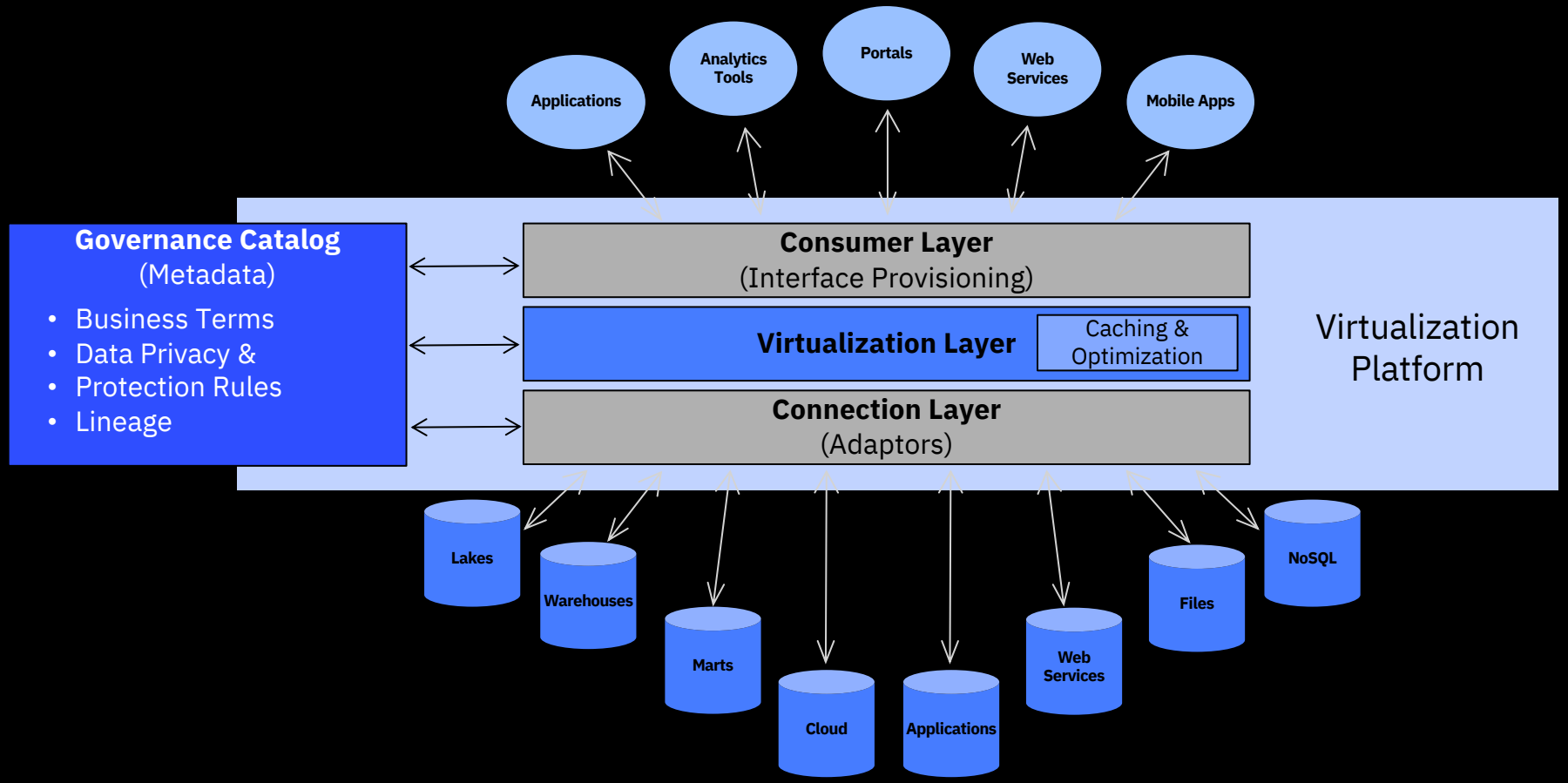
private cloud

on premises

SAME

Data Virtualization in Cloud Pak for Data

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



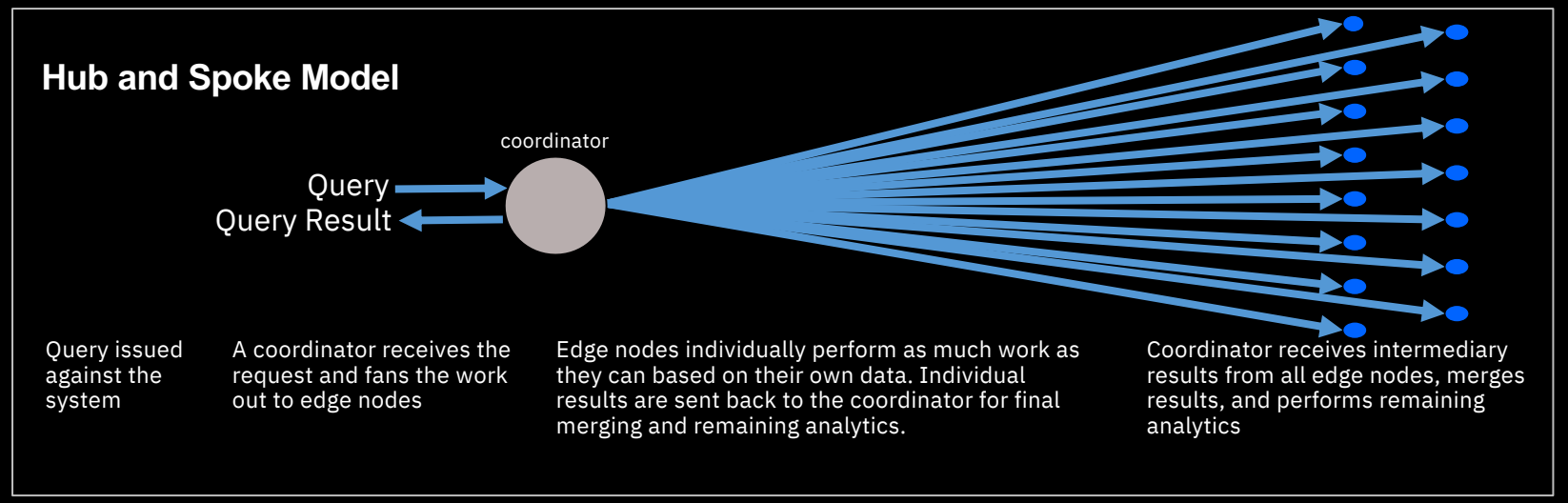
Key Architectural Differentiation

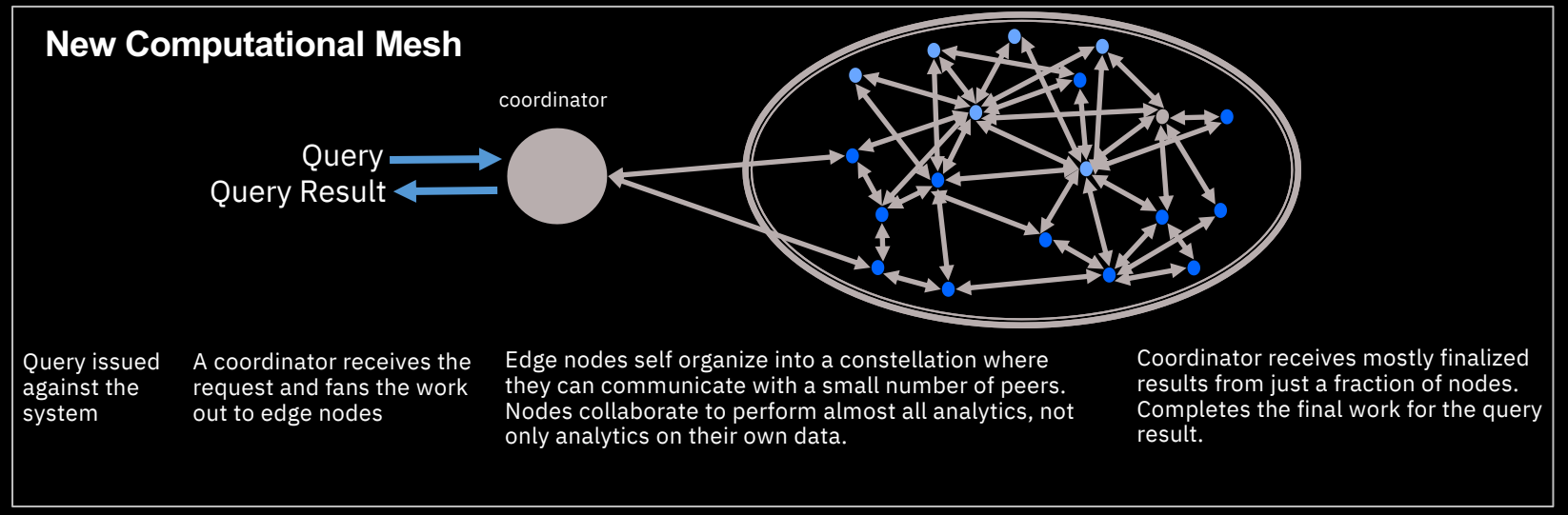
Hub and spoke execution models:

- Lacks scalability
- Performance constrained
- Basis for Federation and our competitors

IBM is first to market with a parallel processing model:

- Theoretically unlimited scalability
- Ease of addition/removal of sources
- Execution pushed down into the constellation mesh



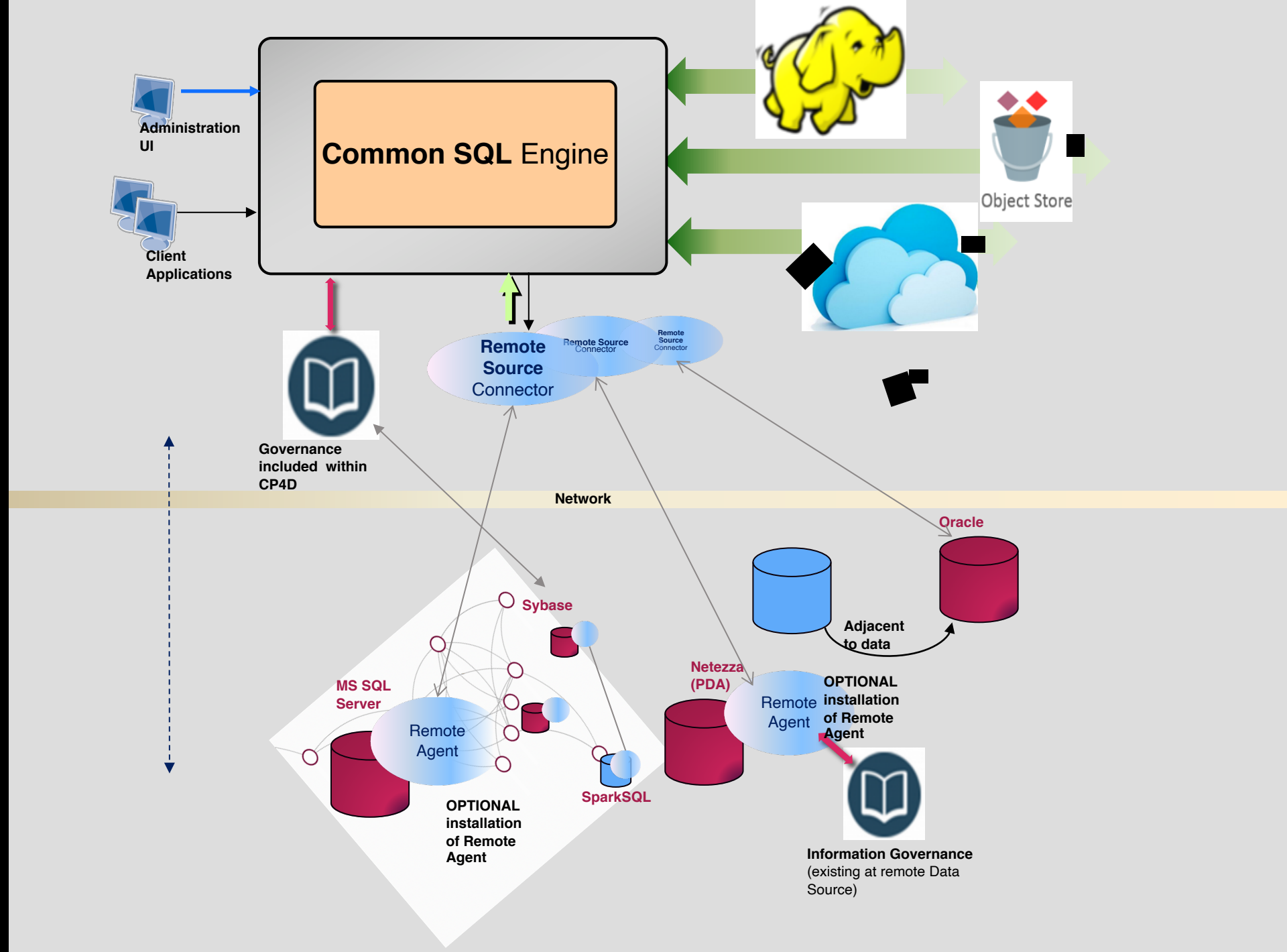


DEMO – Cloud Pak for Data

https://www.ibm.com/cloud/paks/experiences/cloud-pak-for-data

https://developer.ibm.com/components/cloud-pak-for-data/series/

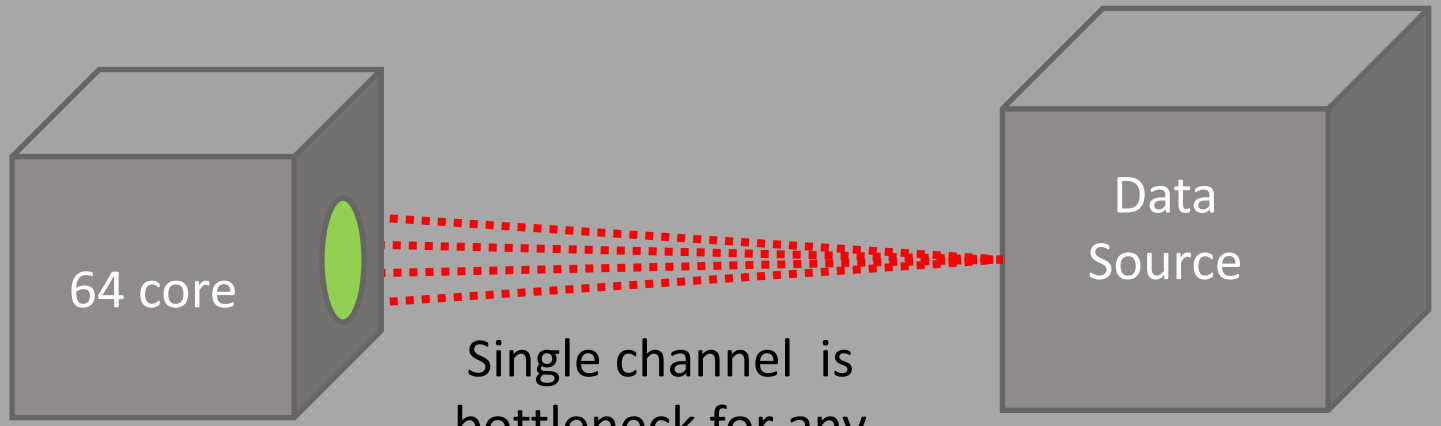
Architecture



Bringing parallelism to each data source.

Traditional processing queries the data source (on right) then processes large results on a single thread (executor) on left.

Solution: Data Virtualization queries data source (on right). Merges results in many **parallel threads** on left by leveraging parallel read streams.

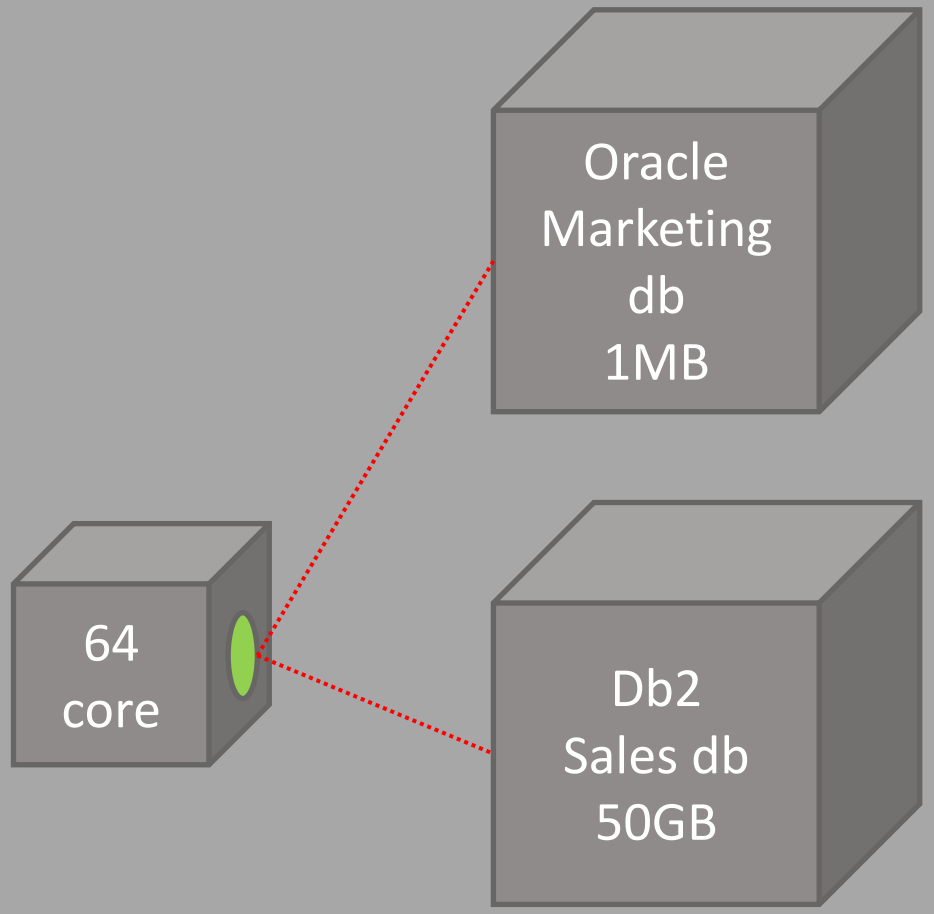


Single channel is bottleneck for any large result set. Data Virtualization resolves this through parallel read from sources.

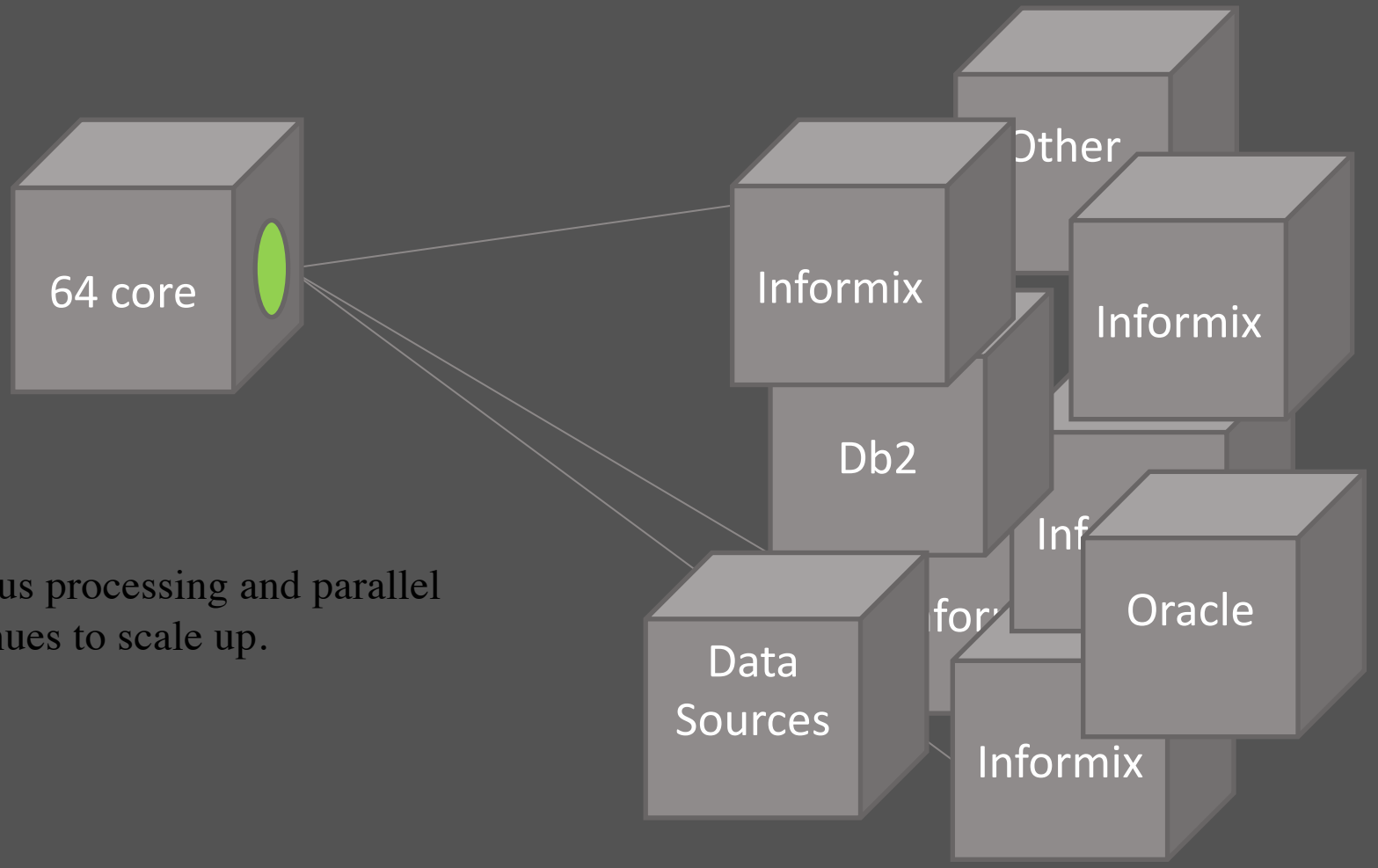
Efficient Cross Source Joins

Problem: Joining data between multiple databases is a common pattern, but brutally slow. With traditional processing, data from both tables is shipped over the network to the server, where the join is processed.

Solution: Data Virtualization uses the **early filtering** techniques to dramatically reduce transmission costs. Data from the smaller table (inner) is used to build a small filter query that is applied to the larger (outer) table before data is transmitted. In many cases this is 85% effective at filtering data that will be removed by the join before the join is processed, leading commonly to a ~10x acceleration.



20 databases? Traditional processing gets worse as the number of databases increases



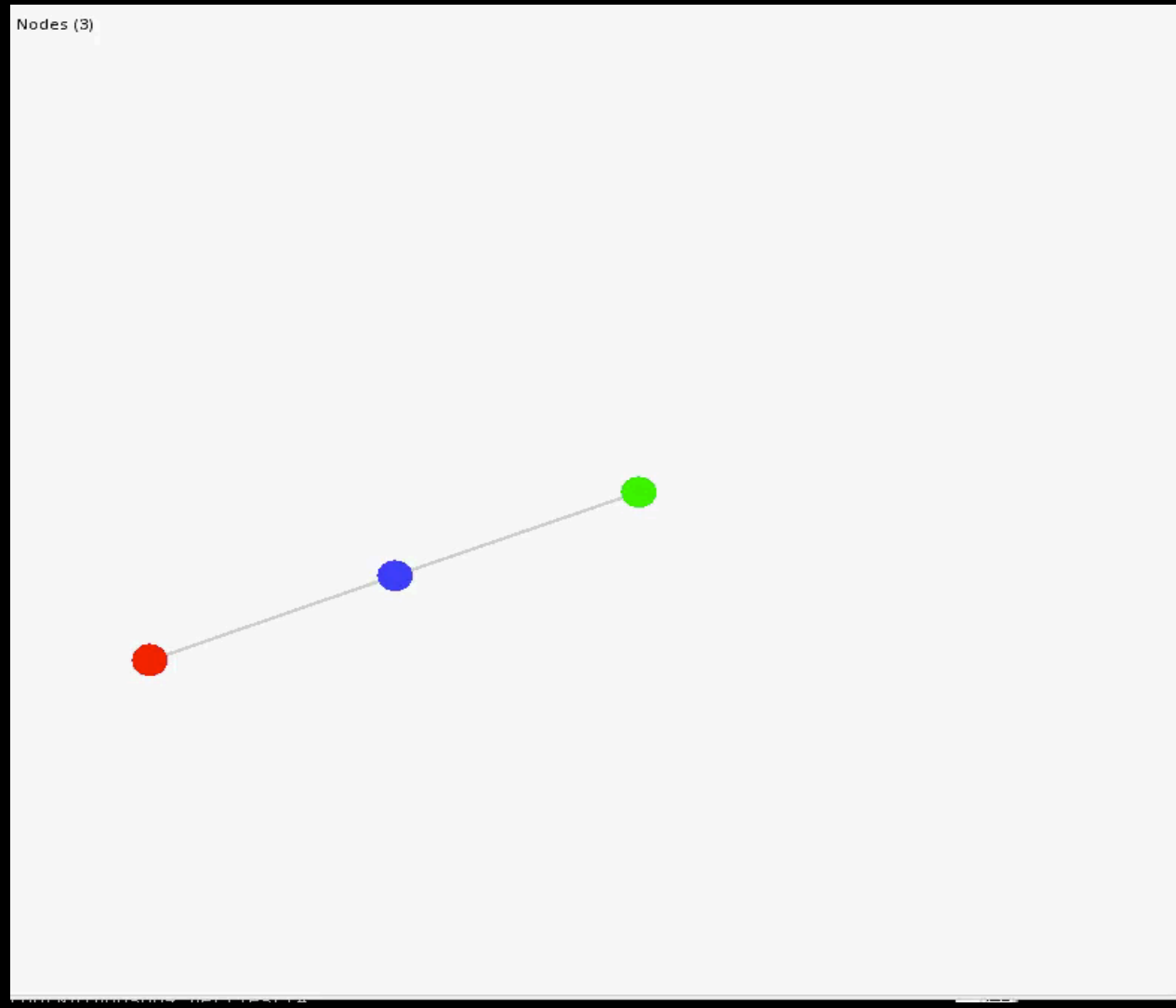
Simultaneous processing and parallel fetch continues to scale up.

Real System test
Growing a Constellation

- Video of constellation growing to 349 Nodes.
  - Network stays compact.
  - 2 and 10 links between nodes
  - No manual configuration.

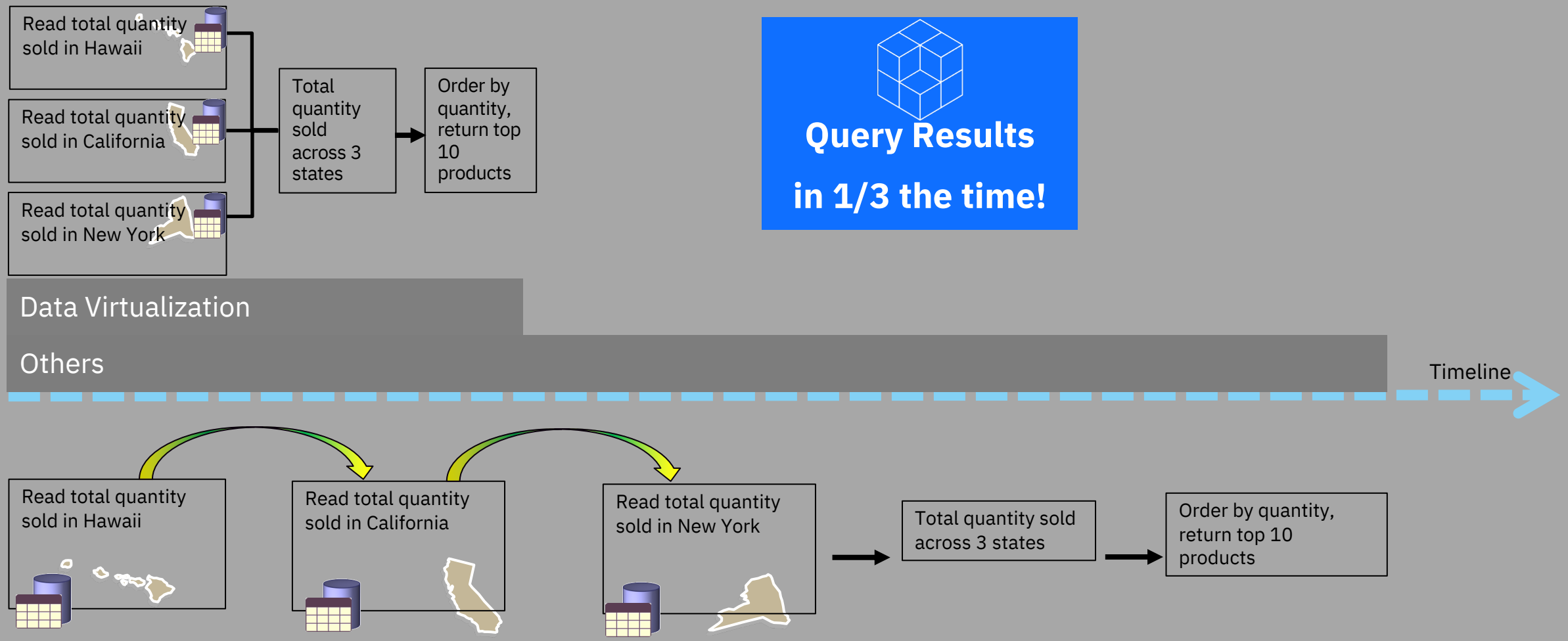
- Latency aware connection between nodes
  - Which nodes connect to which others?
  - Fastest reply strategy

- Diameter of the constellation (i.e. the number of hops between the two furthest nodes) grows logarithmically. Small diameter is ideal for communications.



Smart Query Processing

IBM Data Virtualization reduces query time by using Parallel Processing, Pushdown Optimization and Connection Pooling



DV Engine Generated SQL for Distribute Aggregation

Original

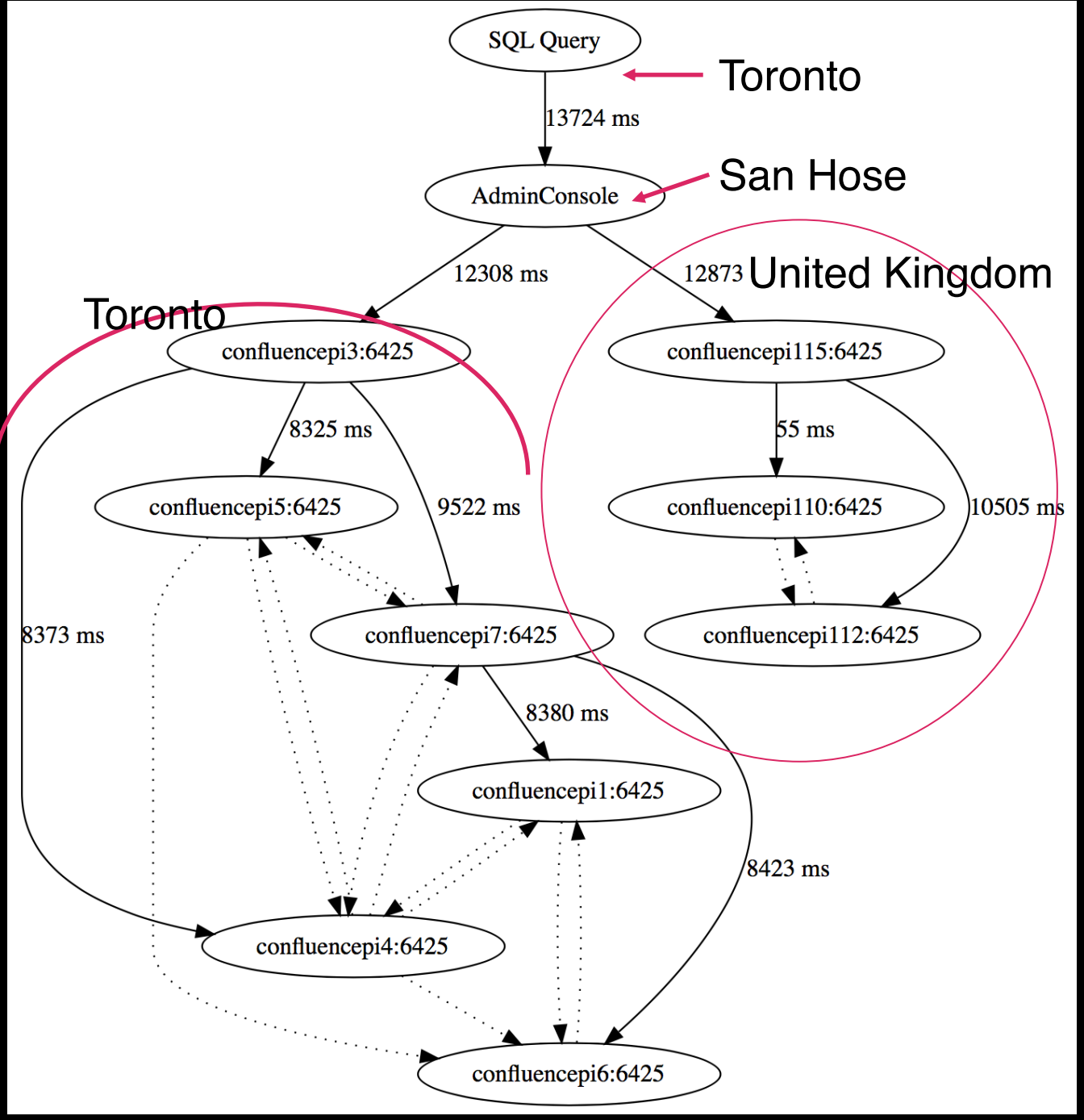
```
SELECT COUNT(PROMOVALUE2) FROM PROMOTION
```

Remote SQL

```
SELECT SUM( A0.C0)
  FROM (
    SELECT A1.C0 C0
      FROM new com.ibm.db2j.GaianQuery(
            'SELECT COUNT( A2."PROMOVALUE2") C0
               FROM new com.ibm.db2j.GaianTable(
                      ''PROMOTION'',
                      ''SOURCELIST=(MYSQL10000:"POPS_node1", MYSQL10001:"POPS_node2",
                                    MYSQL10002:"POPS_node3", MYSQL10003:"POPS_node4",
                                    MYSQL10004:"POPS_node5")'',
                      '''"PROMOKEY" INTEGER, "PROMOTYPE" INTEGER, "PROMODESC" CHAR(30),
                        "PROMOVALUE" DECIMAL(5, 2), "PROMOVALUE2" DECIMAL(5, 2),"PROMO_COST" DECIMAL(9, 2)''
             ) A2',
            'GAIAN_EXTENDER=[pAgg]','',
            'C0 DECIMAL(5, 2)') A1
       ) A0
```

Query Processing in the Constellation

- Fixed execution within the constellation is impossible because of the highly dynamic nature of the network.
- Each node instead simultaneously sends the relevant portions of the query to both the connected data source to it's peers in the network.
- Combines and process the results as they are received.
- Duplicate results are avoided by a given node only returning results to the first peer that requested them.
- Implicitly results in balanced processing of the query through the constellation.



Language Translation in Data Virtualization

Broad set of data sources supported by Data Virtualization each with unique syntax variations.

Constellation is not limited only a single data source type. A logical schema is created across all connected sources.

Multiple levels of translation as we move from the applications through the constellation down to the data source.