Data Lake 4.0

Powered By AI, Data Ops & Data Virtualization

Kitman Cheung
CTO, IBM Data and AI – Asia Pacific
cheungjk@sg.ibm.com

© 2019 IBM Corporation

Rorschach Inkblot

MIT Experiment – Norman AI

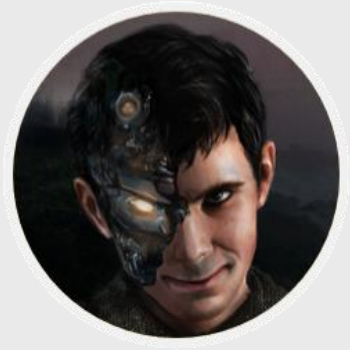
http://norman-ai.mit.edu/

WHAT DOES AI SEE?

We trained Norman on Reddit, and compared captions with standard image captioning neural network.
Here is what both AIs see on Rorschach's inkblot tests.

CAPTIONS BY

NORMAN AI



CAPTIONS BY

STANDARD AI

Why Data Lake?

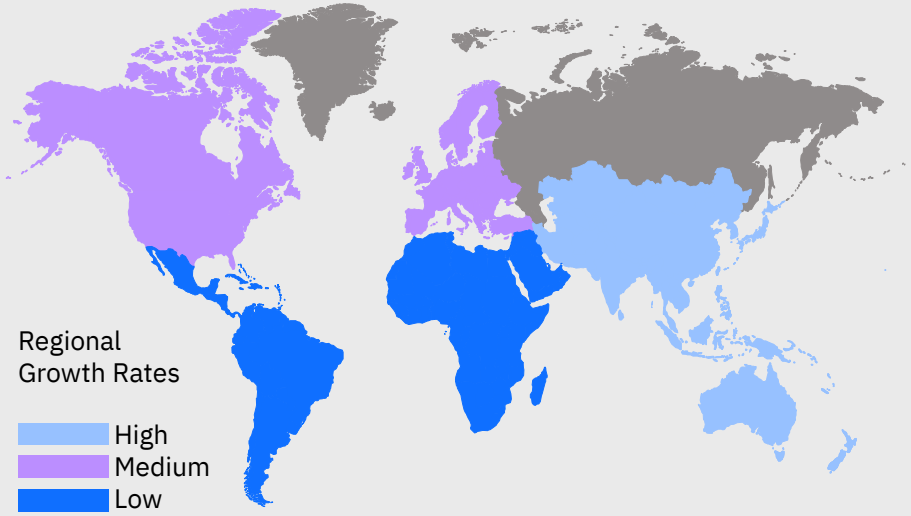
- Rapid technology advancements have led to a dramatic increase in information traffic and as cloud adoption grows streaming data is rapidly becoming more commonplace.
- Vast volume of data has introduced new challenges in data capturing, storage, analysis, search, sharing, transfer, visualization, querying, updating, and information privacy.
- Inevitably, these challenges required completely new architecture design and new technologies, which help us to store, analyse, and gain insights from these large and complex data sets.

Data Lake supports agile data acquisition, natural storage model for complex multi-structured data, support for efficient non-relational computation, and provision for cost-effective storage of large data-sets

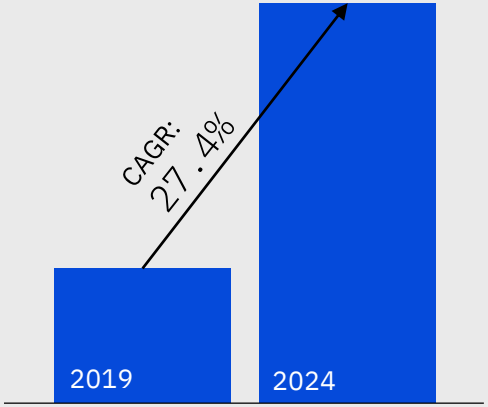
© 2019 IBM Corporation

Data Lake – Growth

Data Lake Market – Growth Rate by region 2019-2024



Data Lake Market



Study period:	2018-2019
Base year:	2018
Fastest growing market:	Asia-Pacific
Largest market:	North-America
CAGR:	27.4%

The data lakes market is expected to witness growth at a CAGR of 27.4% over the forecast period 2019-2024. Data lakes have become an economical option for many companies. The cost of maintaining a data lake is lower owing to the number of operations and space involved in building the database for warehouses.

© 2019 IBM Corporation

Source: https://www.mordorintelligence.com/industry-reports/data-lakes-market

The Hadoop conundrum

With the advent of cloud, organizations are providing cheaper ,scalable and faster solutions than on-premise HDFS. One of the primary factors in the adoption of HADOOP was its ability to scale for exponential data volume growth with cheaper hardware. But , being an on-premise platform and the maintenance overhead loomed over organizations. Cloud providers are able to do much more by providing the same scalability along with no maintenance overhead and cheaper costs. A new architecture is gaining momentum on the cloud which allows the separation of computer and storage layers, thus enabling scaling on both fronts.

© 2019 IBM Corporation

Experience from pure Hadoop based data lake deployments

- Multi-tenancy with a centralized data lake being a myth , most of the customer ended up having multiple data lake deployment.
- With adoption of horizontal scaling and elastic infrastructures (cloud model) Hadoop cluster struggled to get the same level of automation and elasticity as other services (devops).
- Most of the Hadoop workload shifting to Spark.

© 2019 IBM Corporation

Evolving to address the challenges

Data Lake 1.0: Data Warehouse

Proliferation of data silos and need for enterprise Insight , led to data warehouse.
Design principle: focus on KPI , known matrix.

Data Lake 2.0: Hadoop

Growing digital footprint and web scale analytics , demanded a low cost , scalable infrastructure.
Design principle: focus on ad-hoc analysis and batch processing for large scale data.

Data Lake 3.0: Cloud Object Storage + Spark

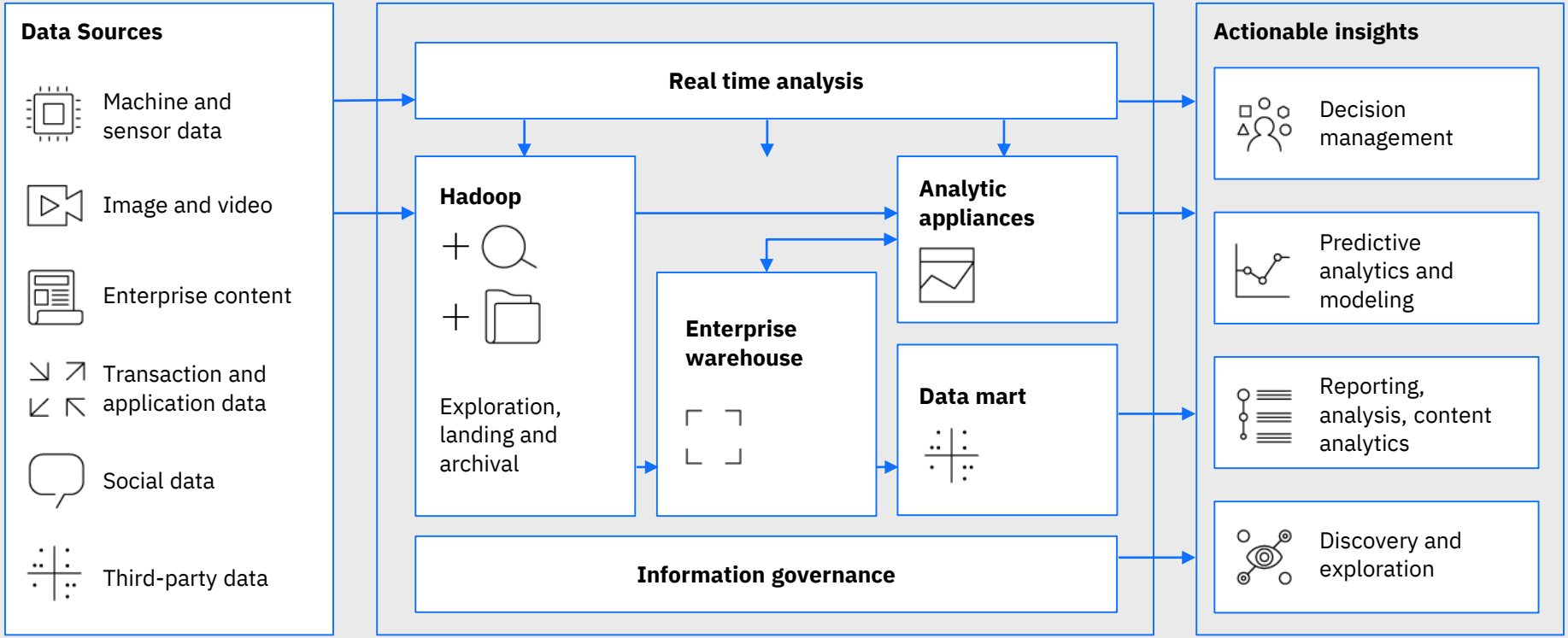
Growth of cloud and AI experimentation led requirement for elastic compute environment.
Design Principle: isolation of compute and storage.

Data Lake 4.0

Need for automation (dataops), focus on ML/AI workload, hybrid cloud
Design Principle: Microservices, AI enabled, business outcome.

© 2019 IBM Corporation

Current architecture for most customer based on Industry reference architecture commercialized 4-5 years back

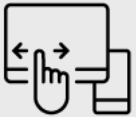


© 2019 IBM Corporation

Major Architecture and technology shifts is happening

Cloud Experience

Easy to use, self-service, on-demand, elastic, consumption



Compute and storage

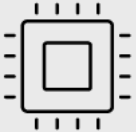
Separation in public and private clouds for increased performance

Kubernets and containers

Adoption as standard operating environment for flexibility and agility

Streaming and ML/AI

Multi-function analytics for the data-driven enterprise

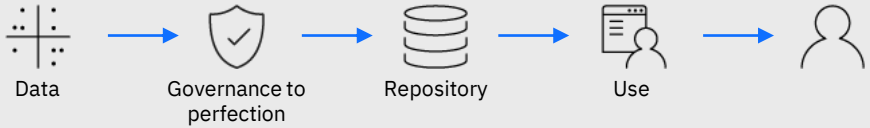


© 2019 IBM Corporation

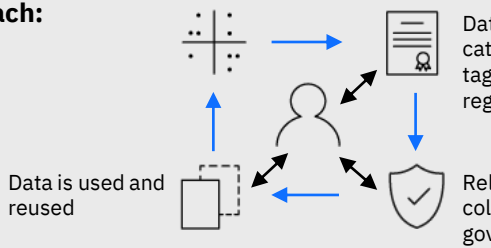
Key drivers for Data Lake Initiatives

- An analytics **sandbox for exploring** data to gain insight.
- An **enterprise-wide catalogue** to find data across the enterprise and to link from business term to technical metadata.
- An environment where users can **access** vast amounts raw data at low cost.
- Tools and technologies for processing large volume of data.

Traditional approach:



Data Lake approach:



Key drivers for Data Lake Transformation 4.0

Data Ops	Data as a Service	ML/AI Driven
Focus on Business	Focus on Publish/Subscribe Model	Focus on Data Monetization

Business Value

© 2019 IBM Corporation

"...a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and consumers across an organization. The goal of DataOps is to create predictable delivery and change management of data, data models and related artifacts. DataOps uses technology to automate data delivery with the appropriate levels of security, quality and metadata to improve the use and value of data in a dynamic environment."

Gartner 2018

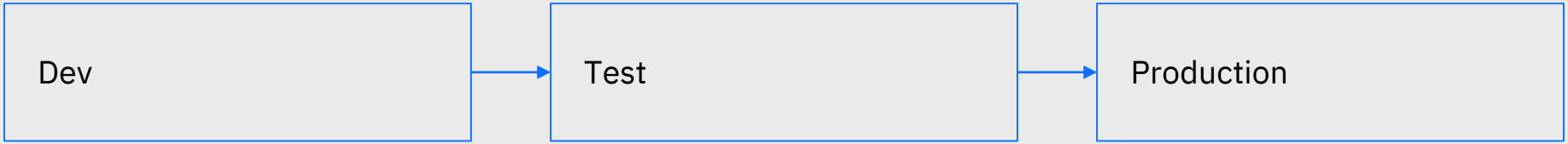
"DataOps incorporates the Agile methodology to shorten the cycle time of analytics development in alignment with business goals"

Wikipedia

Using automation to help operationalize data

© 2019 IBM Corporation

DataOps



Business Process Owner

Designs business processes, leads a team.

Ops for AI

Ensures ongoing health and transparency of AI in production environments.

Continuous improvement and health (including Bias)

Evolve and improve applications

Understands business problem and domain

Build/test models and AI functions

Runtime explain-ablility for trust

Trace AI outcomes to business results

Prepare for fairness and robustness

Orchestration into application

Data Scientists

Builds and trains models.

App Developer

Builds and supports applications that leverage AI.

© 2019 IBM Corporation

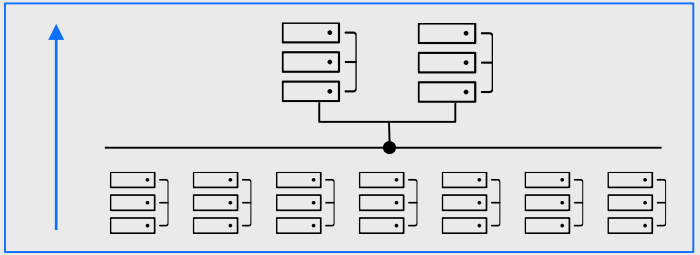
IBM Technology for Data Lake 4.0

© 2019 IBM Corporation

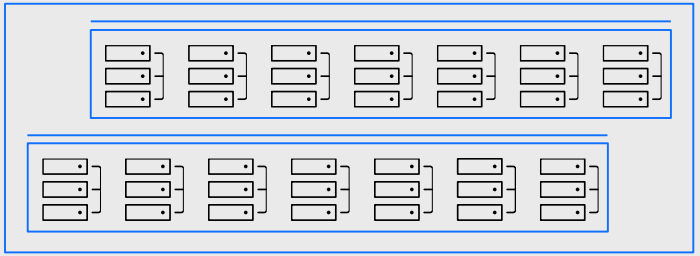
IBM Cloud Object Storage creates a transformational approach for large scale data storage

- **Next generation web accessible storage**
  - REST API accessible through standard S3 interface with no proprietary API calls
  - Parallel access and scale out architecture
  - Access & share your data from multiple applications
- **Storage data protection**
  - In line erasure coding - single copy of data
  - No RAID or replication
- **Suitable for long term data retention**
  - Design for automatic migration of data during HW refreshes
  - Designed for zero down time with automated maintenance
- **Security**
  - In-built or customer provided encryption keys
  - FIPS certified & lockable WORM Storage
- **#1 Ranked Object Storage (Gartner 2018)**

Move from traditional approach



To a transformational scalable architecture built for the cloud era



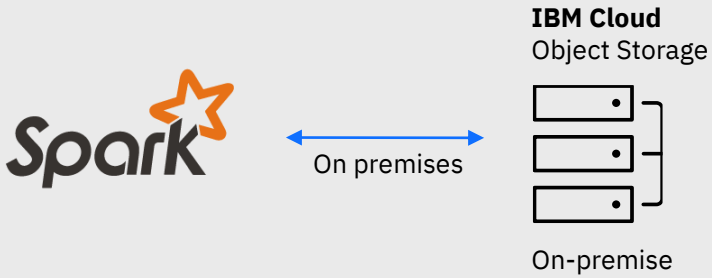
No centralized controller(s)
Access from anywhere

© 2019 IBM Corporation

Spark workloads can be run directly against Cloud Object Storage

- Interactive querying of very large data sets (e.g. BI)
- Running large data processing batch jobs (e.g. nightly ETL from production systems, primary Hadoop use case)
- Building and deploying rich analytics models (e.g. risk metrics)
- Implementing near-real-time stream event processing (e.g. fraud / security detection)
- Complex analytics and data mining across various types of data (e.g. data science)

Use case with object storage

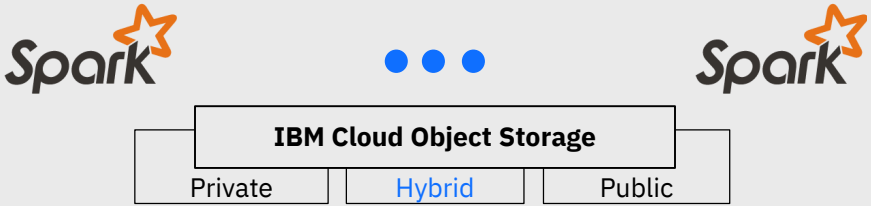


© 2019 IBM Corporation

Spark is Ideal for Cloud Object Storage - addresses challenges of traditional HDFS deployments

Traditional deployment

Cloud Object Storage deployment



Before	After
Data not well protected; requires multiple copies • Need to copy from persistent store	Data is well protected by single copy in COS
Need to scale storage with compute • Expensive and poor match for explosive data growth and scale	Ability to scale storage independently • Directly use data for other analysis/cloud services
Up to six copies required for a local fully redundant HDFS Spark environment	One Repository geo-dispersed with a single copy of data • Built in redundancy compared to traditional Spark Architecture • Enables accessibility to all Spark workflows
Data is local to HDFS cluster	Data available to other applications in addition to Spark
Standard data protection via Hadoop architecture	Increased data protection via SecureSlice technology

© 2019 IBM Corporation

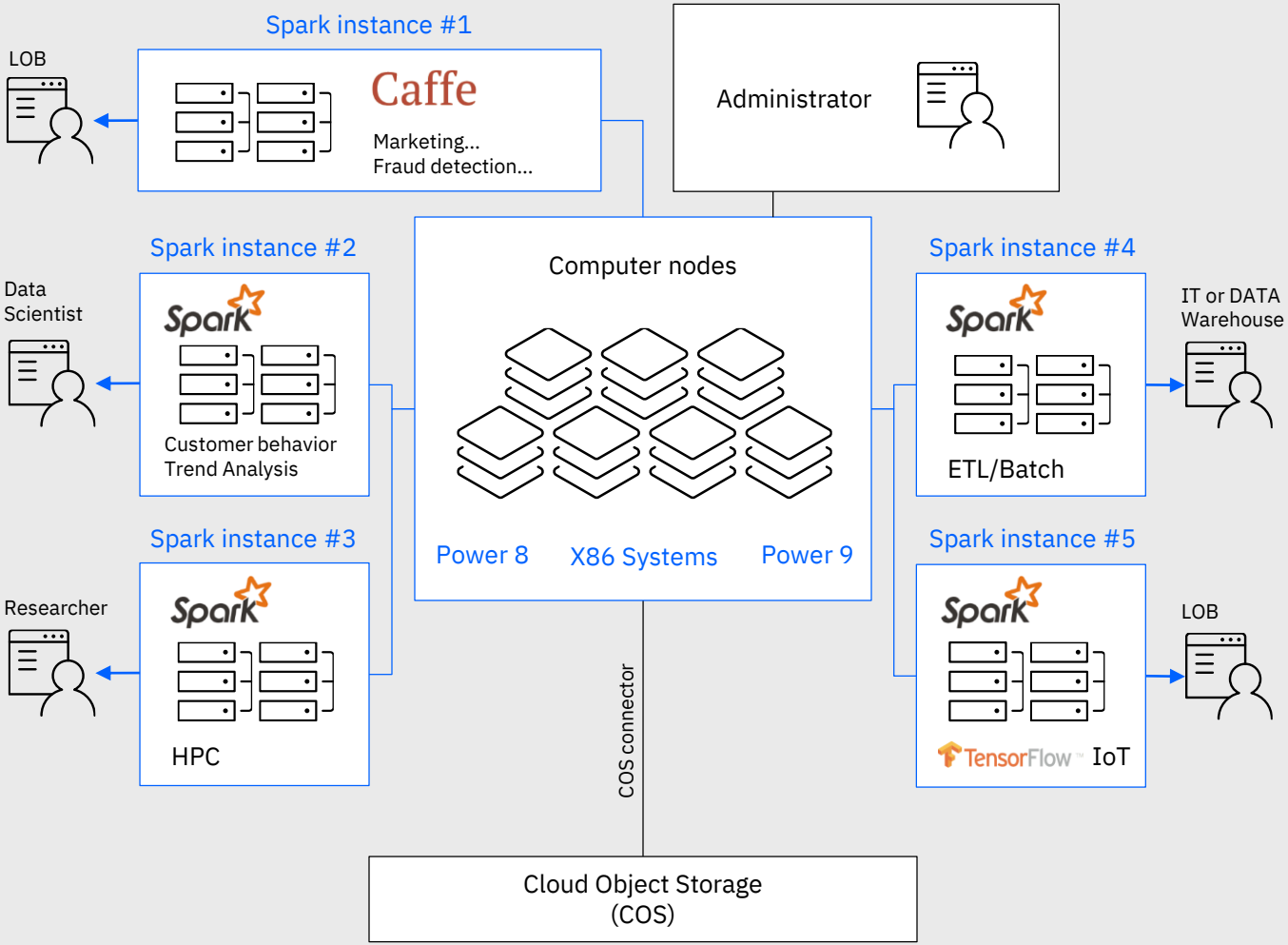
Multitenant Shared Service Use Case

Physical view

Single cluster with SPARK Executer POD on each Server

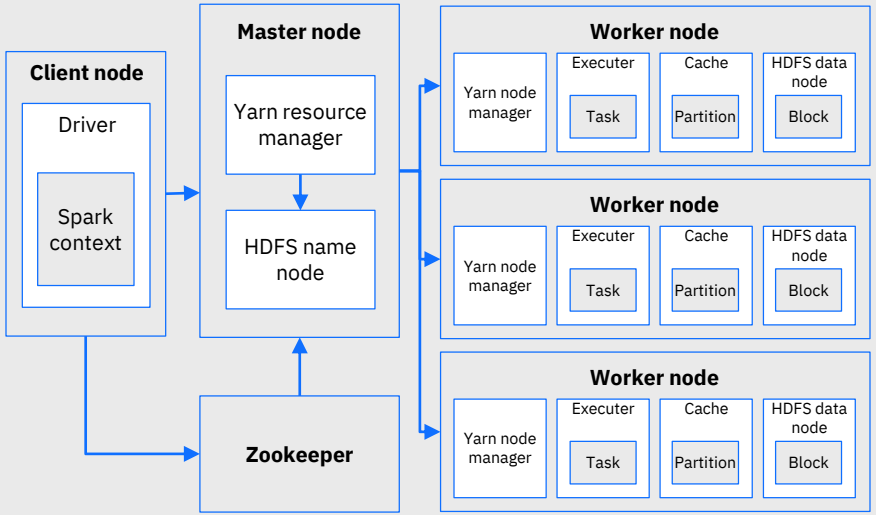
Logical view

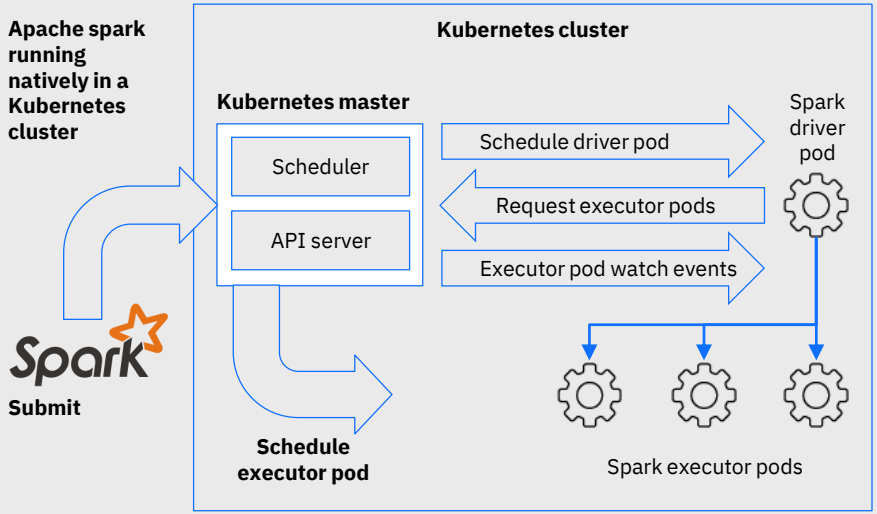
Each group has their own Spark cluster - Isolated, Protected, Secured & SLA Managed resource allocation



© 2019 IBM Corporation

Cloud Data Lake – Hadoop Spark Vs Dynamic Spark Cluster



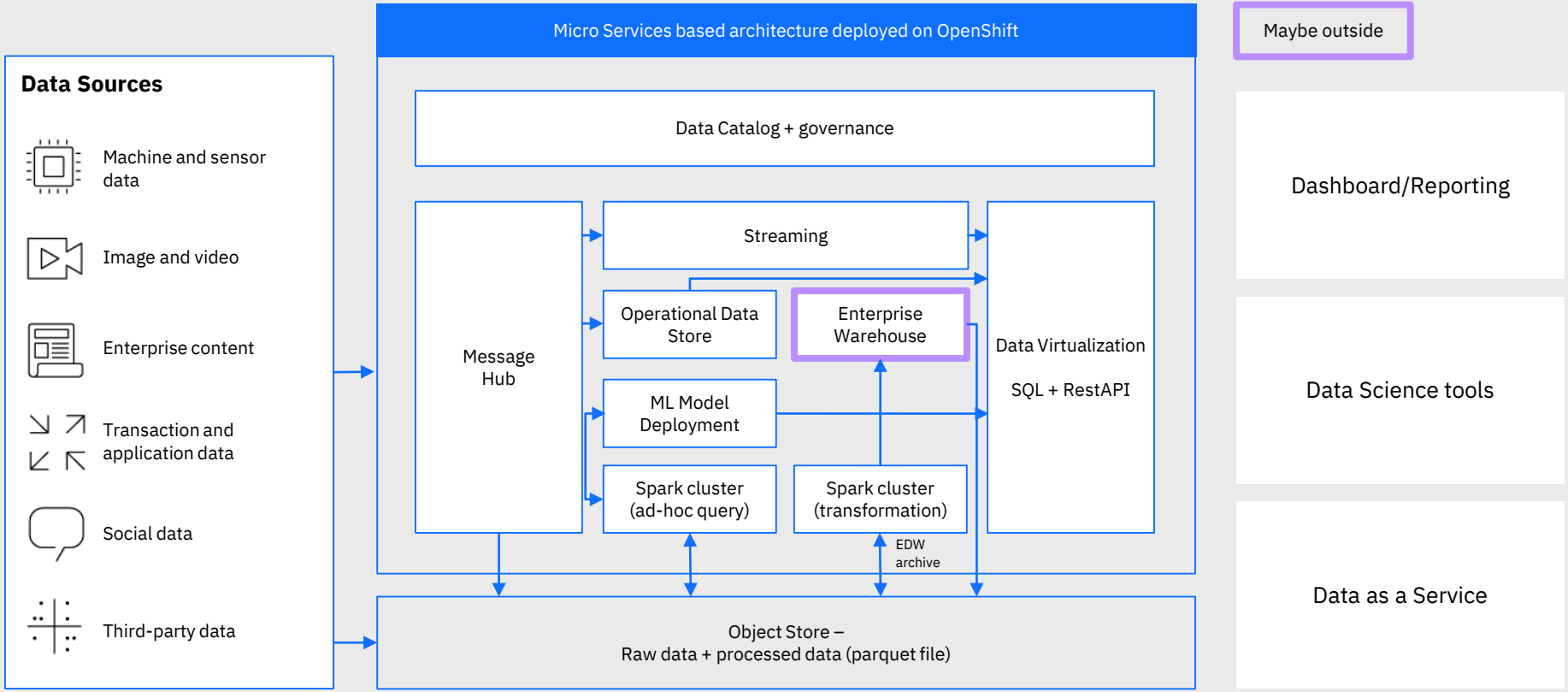


- persistent spark cluster.
- Yarn manages multiple job submission and workload management.
- Suitable for ETL/continuous workload.
- Bound hive repository.

- Ephemeral spark cluster.
- Multiple cluster, each cluster is single tenant.
- Suitable for data science and ad-hoc jobs processing.
- Independent metadata(hive) repository.

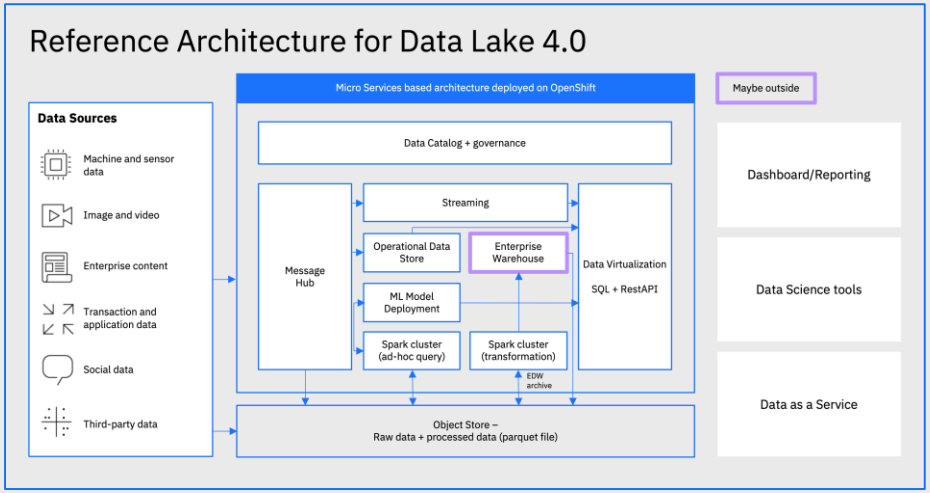
© 2019 IBM Corporation

Reference Architecture for Data Lake 4.0



© 2019 IBM Corporation

Reference Architecture for Data Lake 4.0



IBM Cloud Pak for Data

Data & AI Platform for Data Lake 4.0

1 Isolation of compute and storage provide seamless scalability and elasticity.
*60% reduction in compute infrastructure.

2 Microservices based job deployment allow dataops automation.
**80% quicker onboarding of new data.

3 Multi-tenant spark as a service, allow experiment and production workload to coexist without any conflict.

4 Easily onboard new tools and services.

5 Hybrid deployment architecture, support multi-cloud deployment.

© 2019 IBM Corporation

Thank you

© 2019 IBM Corporation

© 2019 IBM Corporation