Les nouvelles architectures data lake dans le cloud hybride

Marceau GABIN
Cloud Platform Sales
marceau.gabin@fr.ibm.com
+33 624740334

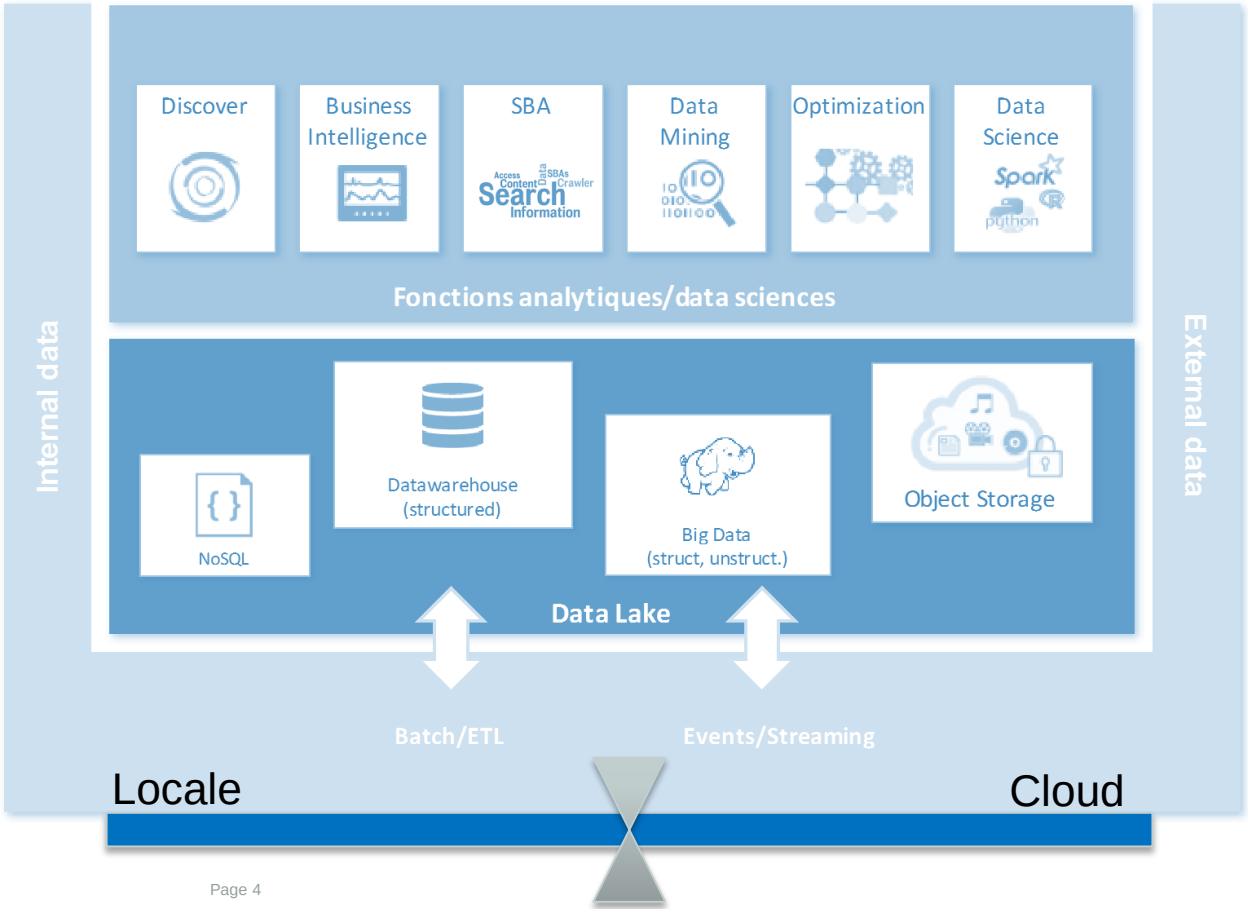
Christophe BURGAUD
Data Architect & Data Scientist
christophe.burgaud@fr.ibm.com
+33-612 360 852

Contents

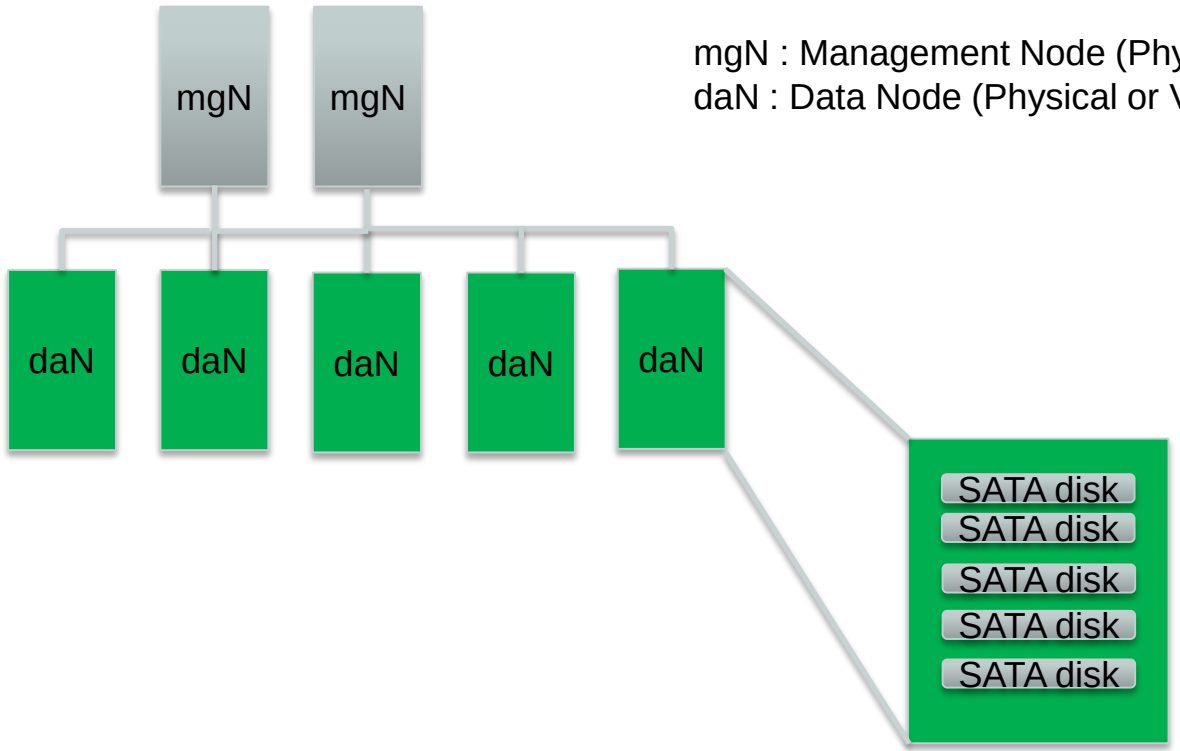
1. The different data lake architectures
2. Data lake storage is evolving
3. Cloud Object Storage ready for data & AI

Global presence of IBM in the different strategy
Customisation vs. Standardisation

Data Lake Typical Architecture

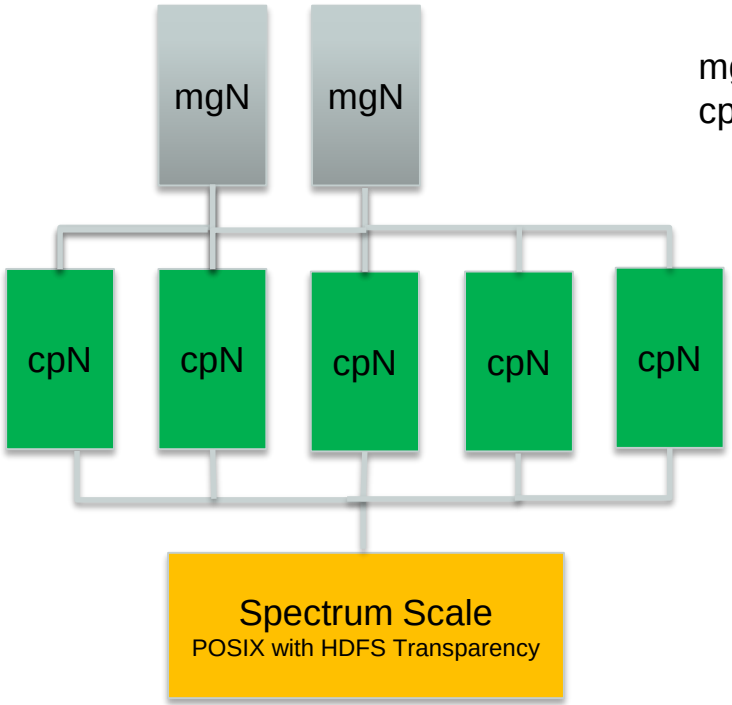


Data Lake with Hadoop and Local Storage



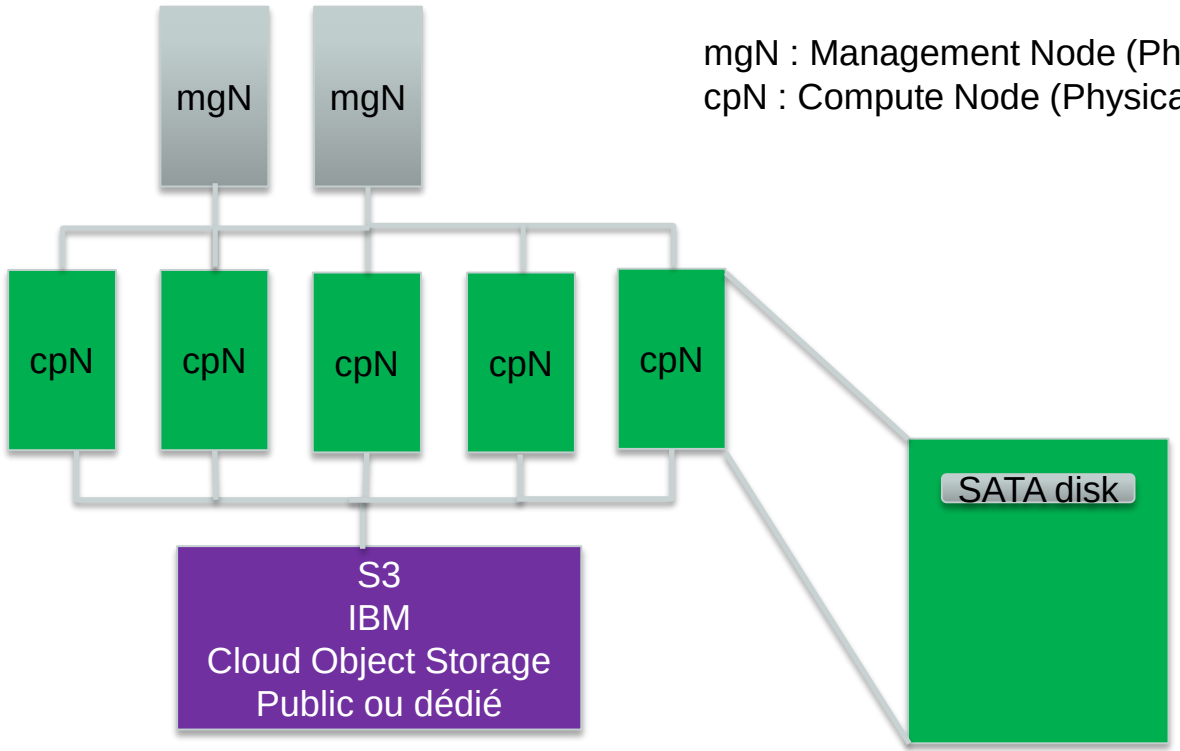
mgN : Management Node (Physical or VM)
daN : Data Node (Physical or VM)

Data Lake with Hadoop and Centralized Storage



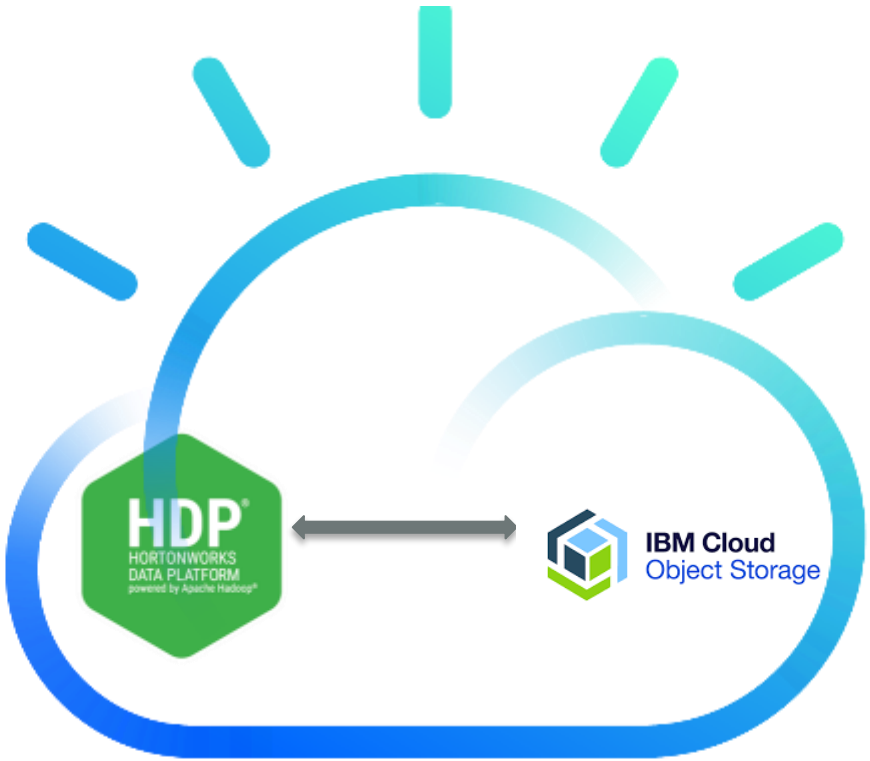
mgN : Management Node (Physical or VM)
cpN : Compute Node (Physical or VM)

Data Lake with Hadoop and Soft Storage and COS (S3)



mgN : Management Node (Physical or VM)
cpN : Compute Node (Physical or VM)

Hybrid Data Lake Architecture with Hadoop PaaS and COS (S3) on IBM Cloud



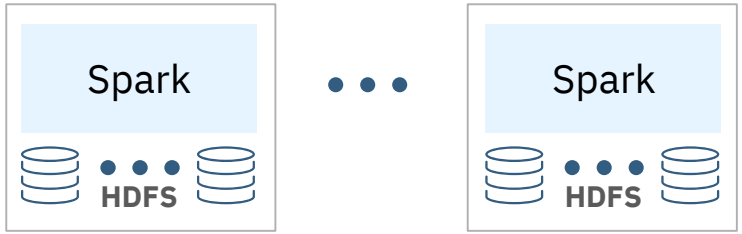
IBM Cloud Object Storage

Redefines availability, security and economics of data storage.

Always-on Availability	• Tolerates a catastrophic regional outage without down time or intervention • Continuous availability architecture • Some legacy providers place the burden of data management and cost for creating and maintaining an out of region second copy on the client
Built-in Security	• Protects against digital and physical breeches • Provide strong data-at-rest confidentiality by combining encryption and information dispersal
Better Cloud Storage Economics	• More cost-efficient than competitors • Enterprise class support, at no additional cost • We cap costs for workloads with unpredictable data access pattern
Simplicity	• Immediate consistency (v. eventual consistency) simplifies app development • Integrated with the rest of IBM Cloud, IBM Bluemix, IBM Watson, IBM Video Services • No “black boxes,” we share how our technology delivers durability, availability, security

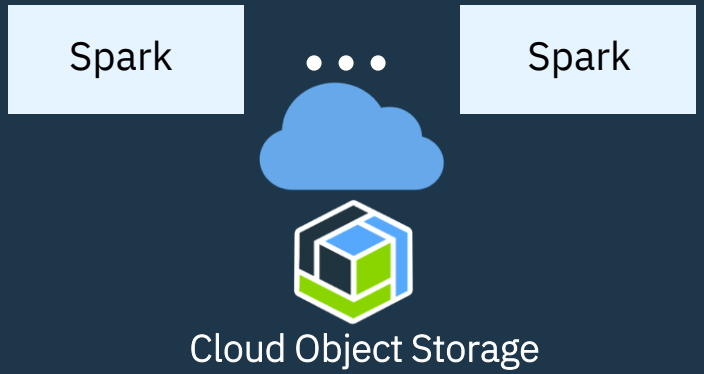
Cloud Object Storage is ideal for Spark analytics

Traditional deployment



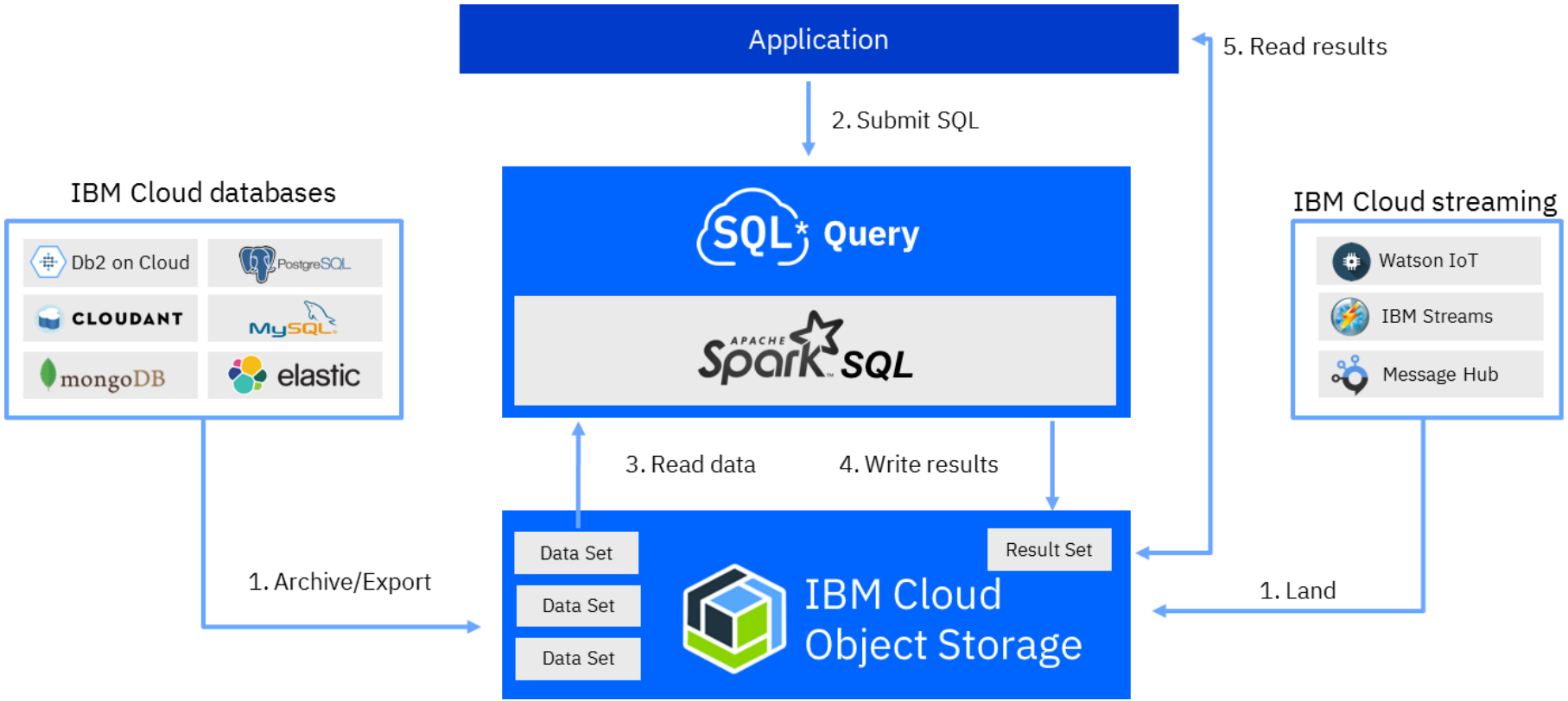
- Data not well protected in HDFS
- Need to scale and manage long-lived, tightly coupled compute and storage infrastructure

Deployment with Object Storage



- Data well protected in Cloud Object Storage
- Use infinitely scalable, cloud managed storage and ephemeral compute.

IBM SQL Query Service



AI & Big Data Analytics

Cloud Object Storage is an integral part of IBM Analytics Engine, Watson Studio, SQL Query and other IBM Cloud Services to provide self-service data analytics and business intelligence solutions that go well beyond the scalability, security, and cost efficiencies of traditional solutions.

Query data in place

Combining SQL Query with data in IBM Cloud Object Storage creates an active workspace for a range of big data analytics use cases. IBM SQL Query is a serverless, interactive querying service for analyzing data directly in IBM Cloud Object Storage.

Perform Apache Spark Analytics directly against data stored in Object Storage

IBM Cloud Object Storage offers optimized connectively to Apache Spark and can be used as a low-cost, scalable persistent storage layer for analytics.

Store data for AI training models

IBM Cloud Object Storage is integrated with IBM Watson Studio to accelerate machine and deep learning workflows required to infuse AI into your business. Build and train AI models, and prepare and analyze data, in a single, integrated environment.

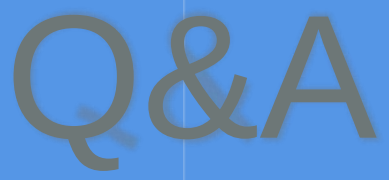
Move data from HDFS clusters to Cloud Object Storage

Free up space on expensive Hadoop cluster by using IBM Big Replicate to efficiently move data between Hadoop data clusters. You can also use IBM COS Distributed Copy (DistCp), an open source tool for migrating large amounts data from Hadoop to Cloud Object Storage.

Build and Analyze IoT Pipelines

IBM Cloud Object Storage is perfectly suited to storing massive amounts of IoT data at low cost and allows analytics frameworks to access the data directly. Data pipelines can be easily set up and managed to generate analytics-ready data, which can be analyzed directly by Watson using Spark as a Service.

Thank you



Disclaimer

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETETENSS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCALIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICUALR PURPOSE. THE SOWFTWARE VERSION OR SOFTLAYER'S OFFERING MAY HAVE CHANGED OR DISAPPERAED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

THE PRICE IS FOR PLANNING PURPOSES ONLY AND IS NOT A FINAL COMMITMENT BY IBM. THIS PRICE IS SUBJECT TO CHANGE FOR REASONS SUCH AS, BUT NOT LIMITED TO, REFINEMENT OF SCOPE AND ASSUMPTIONS AND DEPENDENCIES, AND NEGOTIATION OF TERMS AND CONDITIONS. A FIRM PRICE PROPOSAL CAN BE PREPARED AT THE CUSTOMER'S REQUEST.