Una Nueva Mirada al Gobierno de Datos en el Contexto de Ambientes Big Data

Powered By: CLOUDERA

PABLO QUIÑONES G.

Solutions Engineer @ Cloudera LATAM

+7 años trabajando con temas de Big Data
Parte del Capítulo Dama Colombia
Parte Meetup Apache Spark Bogota
Apasionado por la Tecnología
@pquinonesg

Meetup #QuarantineMeetup

AGENDA

01
CONTEXTO
DATA LAKES

02
MARCO
DAMA - DMBoK2

03
CONCEPTO:
“STATE”

04
CLOUDERA
SDX

05
DEMO

06
Q&A

CONTEXTO DATA LAKES & MARCO DAMA.

Una Nueva Generación .

Evolution of the Architecture

Gen I -
On Premise
Co-located Storage/Compute
Large, Shared Clusters
(CDH / HDP)

Gen III -
Multi-Public/Private/Hybrid Cloud
Disaggregated Storage/Compute
Multi-Tenant, Containerized "SaaS Experiences"
(CDP)

Gen II -
Public Cloud
Disaggregated Storage/Compute
Multiple Clusters
(EMR)

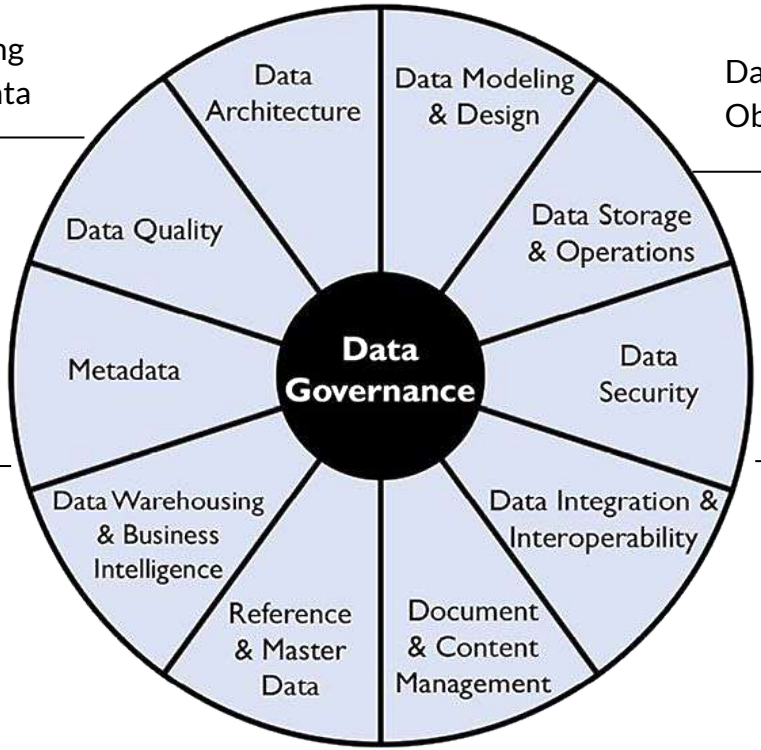
Hadoop: Decade Two, Day Zero*

Data Profiling
Sensitive Data

Data Replication
Object Storage

Technical Metadata
Schemas
Data Lineage
Tags

Data Access Policies (RBAC/ABAC)
Identity Mgmt.
Data Encryption
Audit Trace & Regulations



DAMA-DMBOK2 Data Management Framework

CONCEPTO: “STATE”

Por qué es Importante?

Cloud Creates a New Tension...

How?

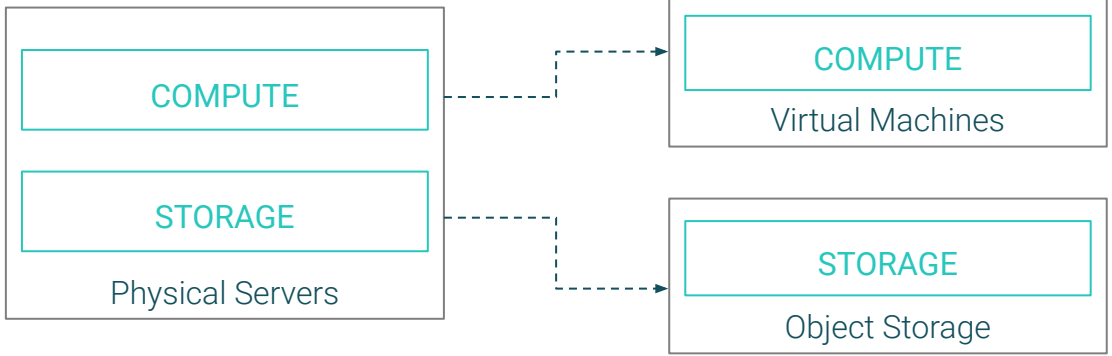
Cloud Agility & Efficiency

Security & Governance

CONVENTIONAL WISDOM: “SEPARATE STORAGE & COMPUTE”

Separate storage & compute

Basic on-prem architecture



Cloud architecture*
(EMR & HDI)

- Agility
- Elasticity
- Isolation
- Transience

* Only works for simple use cases

- Single-tenant
- Single-workload
- Non-sensitive data
- Non-regulated use case

CLOUD DATA LAKES REQUIRE ...

Identities

How do we easily connect to **corporate** identity?

Schema

How do multiple personas **find** and **share** datasets across different compute engines?

Policy

Is there a central place to protect sensitive data at a **field** and **row** level?

Audits

Can I comply to the regulatory requirements with comprehensive **audits** and **lineage**?

Privacy, compliance & regulation

… AND *ACTIVE* MANAGEMENT

Data Profiling & Stewardship

Can I detect **sensitive** data automatically, **tag** and **curate** for easy discovery and security controls?

Replication

What happens to metadata and lineage, when I move data **across environments**?

Workload Management

How do I track workloads for **troubleshooting** and optimization with elastic computes?

Encryption

Is the data **protected** at rest and in-transit?

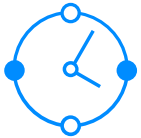
State is required for all but the simplest use cases

- Multi-tenant
- Multi-workload
- Sensitive data
- Regulated use case
- Long Running Apps

DATA LAKES REQUIRE STATE AKA METADATA, BUSINESS LOGIC, …

Identities

Owners



History

Policy

Roles

Quality

Tags

Structure

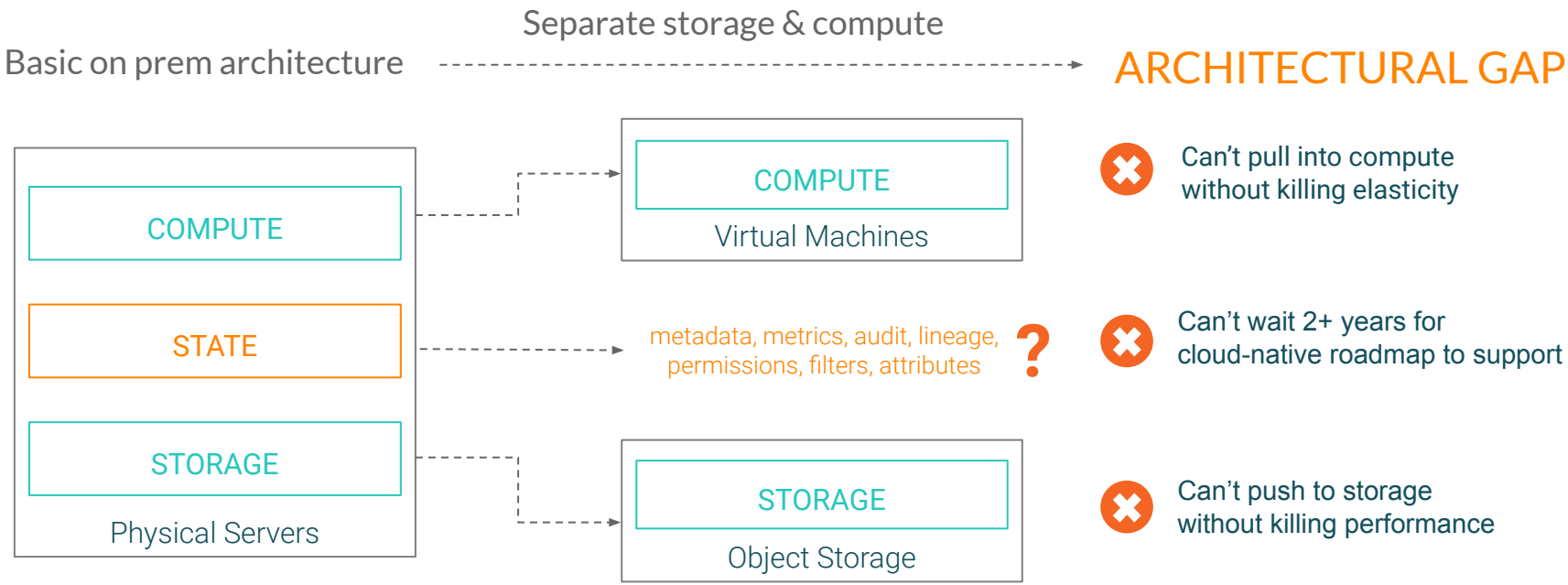
In order to provide:

- Discoverability
- Trust
- Compliance
- Reuse
- Sharing

State is required for all but the simplest use cases

- Multi-tenant
- Multi-workload
- Sensitive data
- Regulated use case
- Long Running Apps

NO STATE, NO DATA LAKE



THREE DATA TECHNOLOGISTS WALK INTO A MEETING...

Engineering Owner

Improve cost efficiency

- Breakup large clusters
- Move to object storage
- Pause idle clusters
- Auto-scale resources
- Infrastructure as code

Architect

Protect the company

- Block anonymous logins
- Encrypt everywhere
- Enforce least privilege
- Capture audit & lineage
- Comply with regulation

Business Owner

Grow the business

- Scale usage
- Add tenants
- Add policies
- Add workloads
- Add geographies

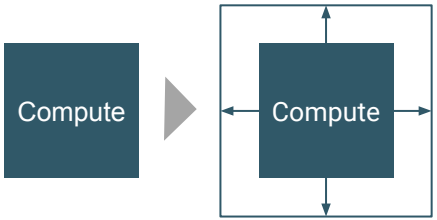
NOT REALIZING THEIR ROADMAPS ARE IN CONFLICT

Engineering Owner

Remove state from the cluster in order to reduce cloud spend

(HDFS, metadata, metrics, …)

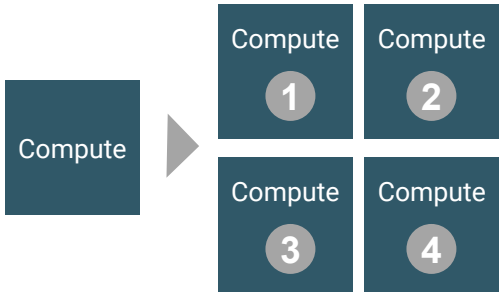
Architect



Capture more state to improve security and governance

(kerberos, audit, lineage, ...)

Business Owner



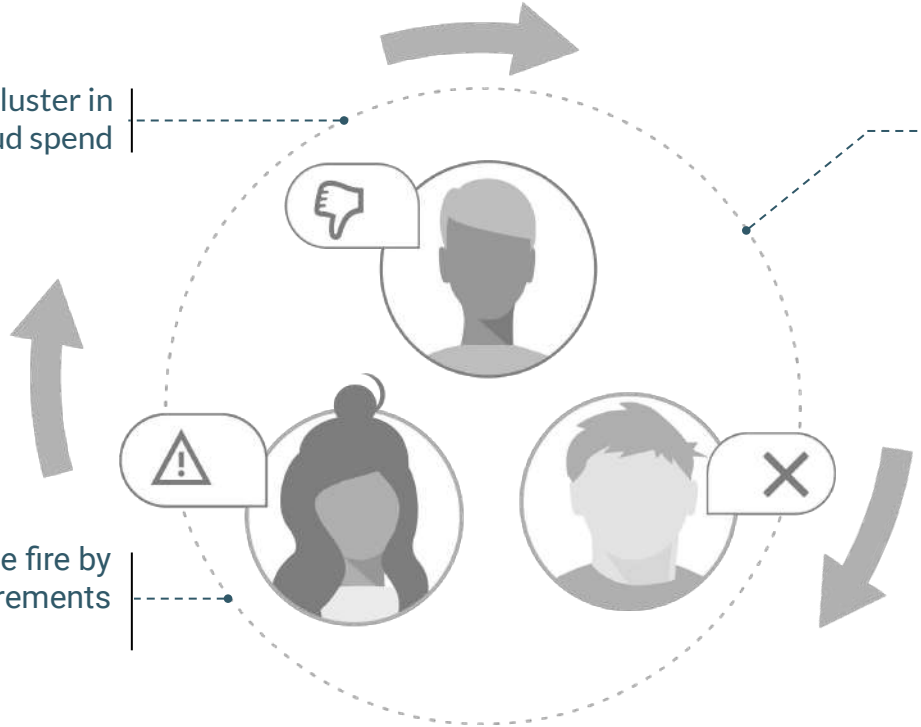
Add fuel to the fire by increasing state requirements

(permissions, filters, attributes, …)

CONFLICTING ROADMAPS CREATE DEADLOCK

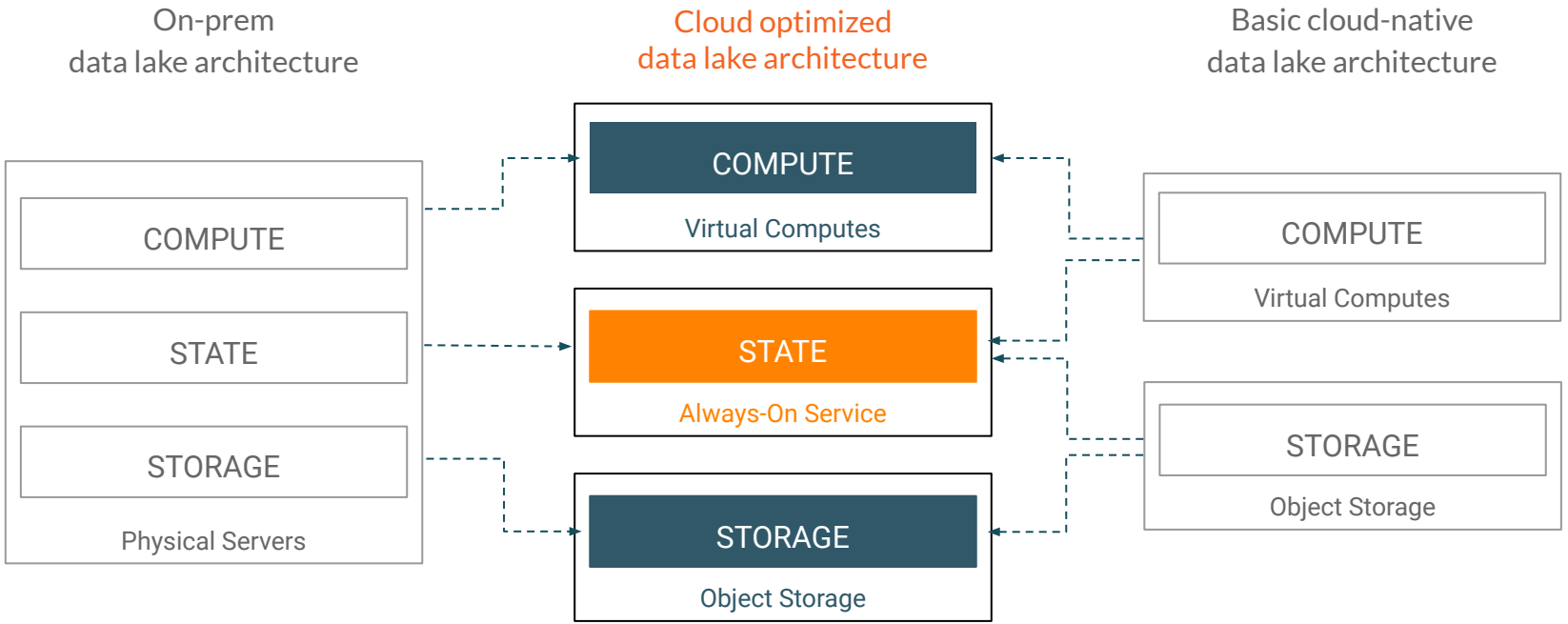
Remove state from the cluster in order to reduce cloud spend

Capture more state to improve security and governance



Add fuel to the fire by increasing state requirements

ANSWER: SEPARATE STORAGE AND COMPUTE *AND STATE*

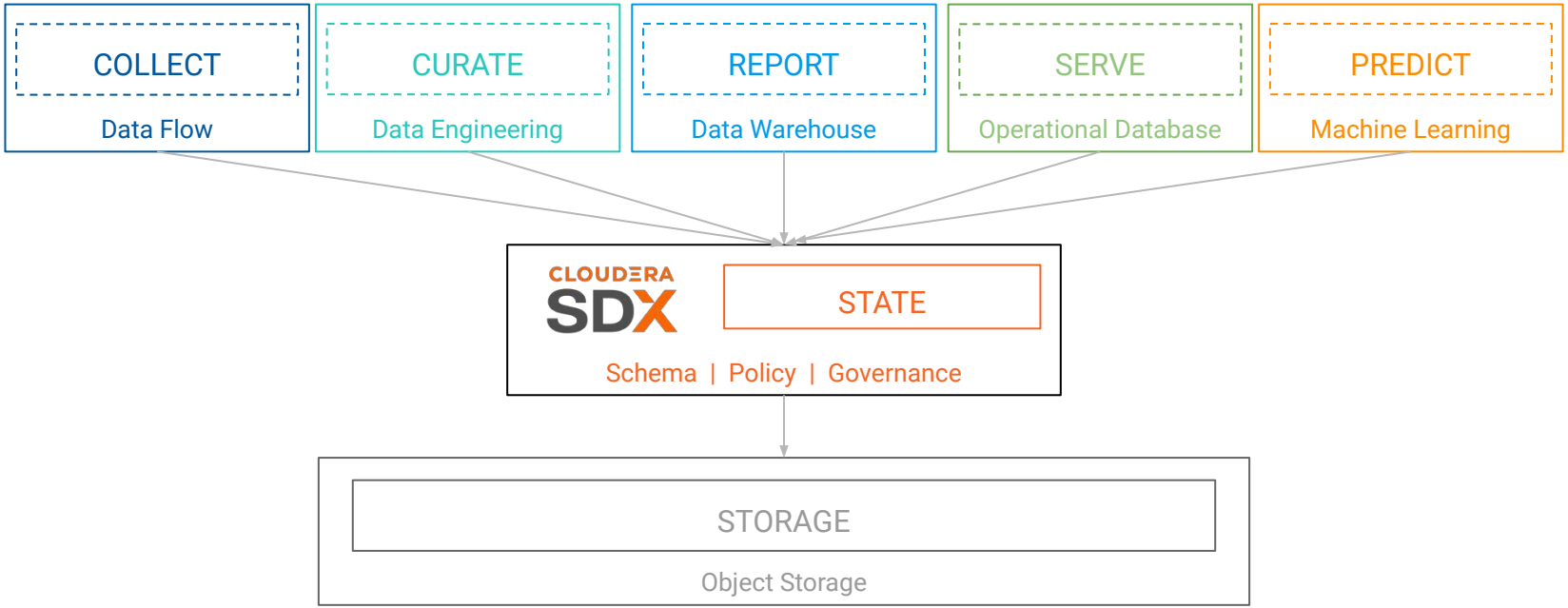


SDX IS BIGGER THAN SECURITY AND GOVERNANCE

It's the architectural innovation that makes the Enterprise Data Cloud possible

Powered by: CLOUDERA

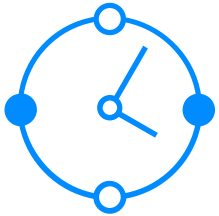
THE NEW GENERATION DATA LAKE ARCHITECTURE



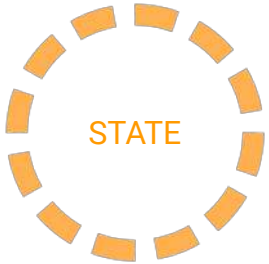
DEMO

#QuarantineMeetup

Final Thoughts...



Evolution of the Data Lake Architectures powered by the Multi Cloud Strategy.



Importance of the **state** in a Data Lake Architecture regarding Security & Data Governance

CLOUDERA SDX

How Cloudera implemented this concept into a new brand Data Plataform through **SDX.**

Q&A.
Dudas.

Muchas Gracias!

Meetup #QuarantineMeetup