Qubole

Big Data in the Cloud

State of the Union and Future Trends

Ashish Thusoo
Qubole CEO & Co-Founder

About Me

Alma Mater

- BA, Computer Science - IIT (India Institute of Technology, Delhi)
- MS, Computer Science - University of Wisconsin - Madison

Background

- Started career at Oracle
- Ran Data Infrastructure team at Facebook from 2007-2011:
  - Built out the Self-service Big Data Platform at Facebook for internal operations
  - Saw huge growth, while chartered to provide all teams unified analytics (Marketing, Analytics, Engineering, Sales Finance, etc.)
  - Spawned developments of big data engines such as Apache Hive and precursors of Presto DB
- Co-created and led Apache Hive project

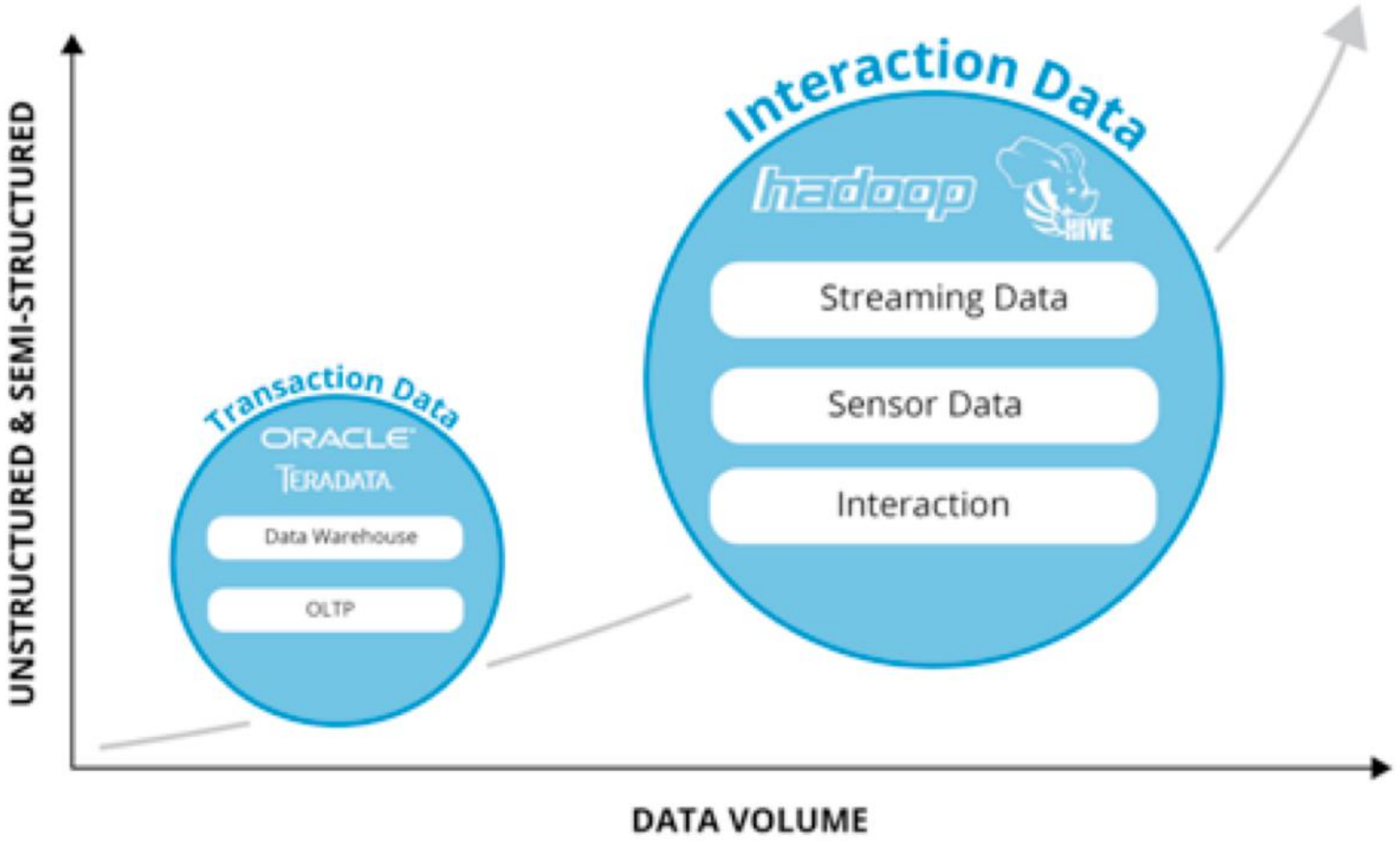
Today - CEO & Co-founder of Qubole

- Cloud-Native Big Data Platform
  - Cloud and workload optimized Spark, Hive, Hadoop and Presto Engines
  - Processes more than an Exabyte of data per month on Cloud Infrastructure (AWS, GCP, Azure, Oracle)
  - Provides Automation and Self-Service for big data jobs (e.g. ETL, Machine Learning, Ad-hoc)

From Data Warehouse to Data Lake

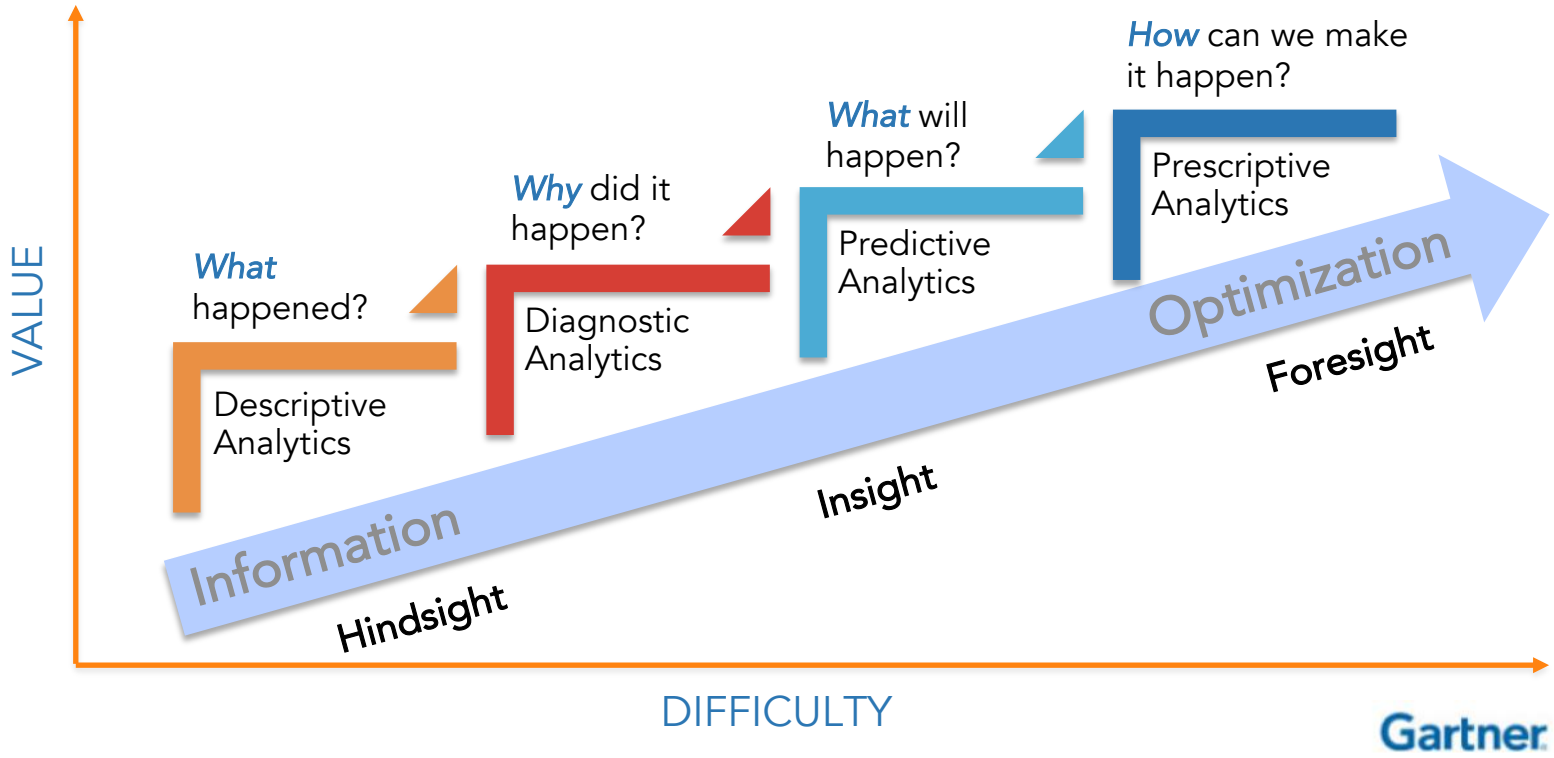
Evolution of Big Data

Changing Nature of Data

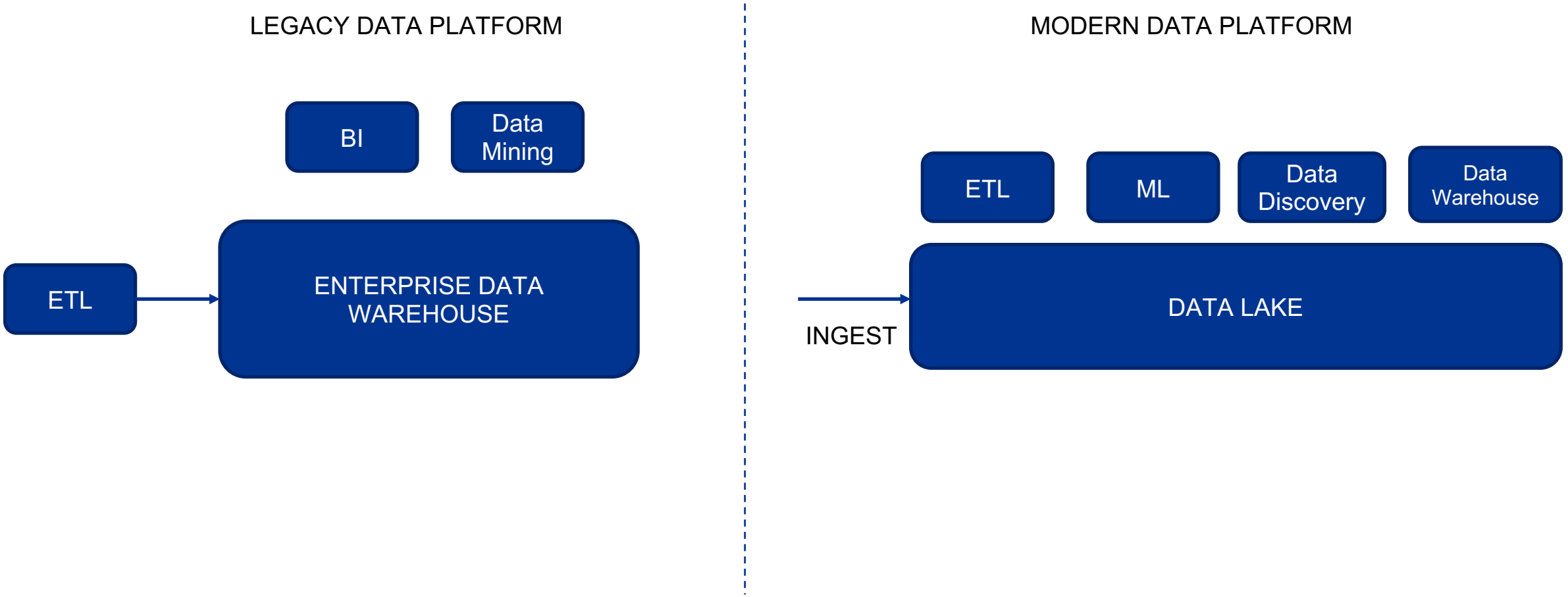


Changing Nature of Analytics

Analytics Value Escalator



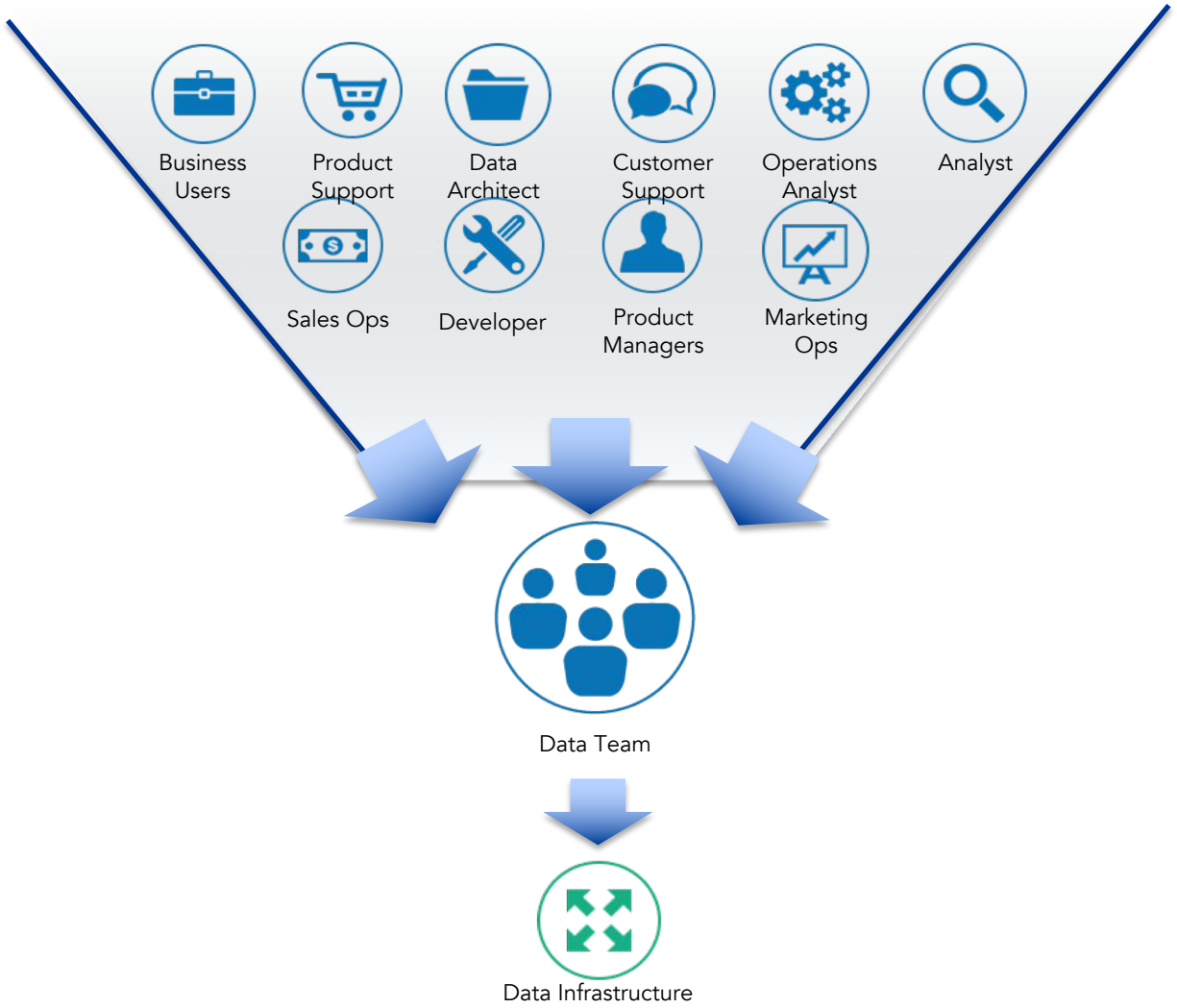
Breakdown of Data Warehouse – Emergence of Data Lake



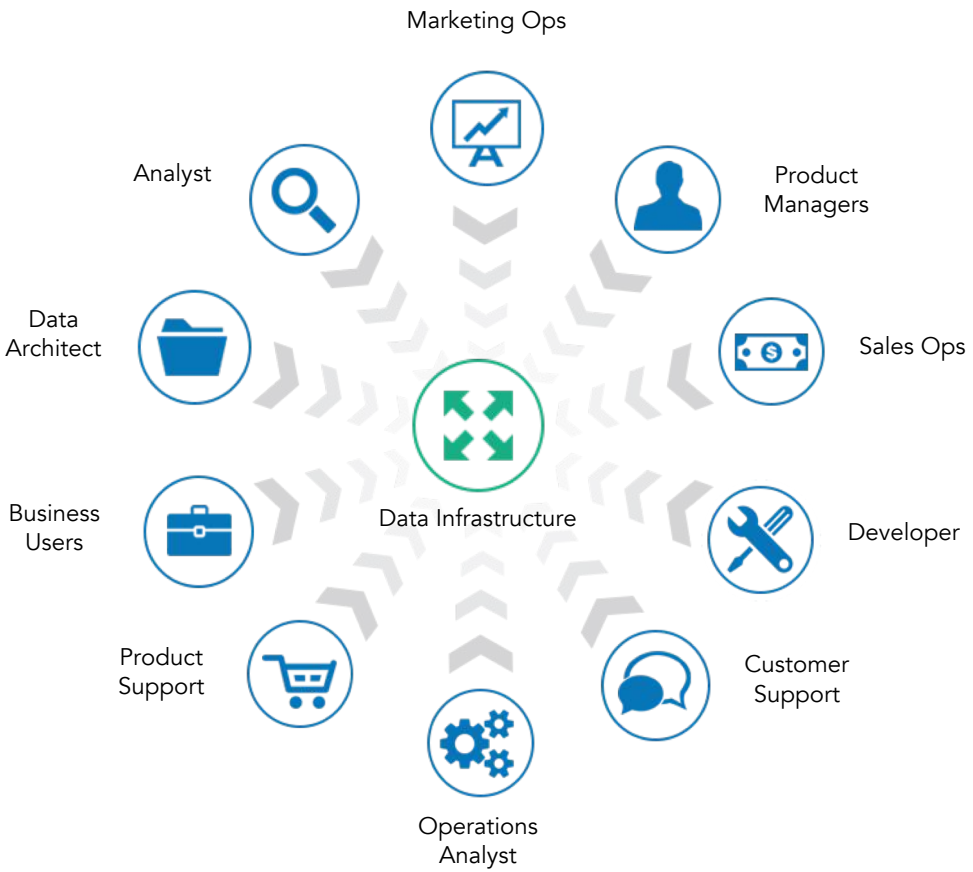
Differences Between Data Lakes and Data Warehouses

DATA LAKE	vs	DATA WAREHOUSE
Semi-structured / unstructured / structured / raw	DATA	Structured data
SQL / Machine Learning / ETL / Graph Analytics etc.	ANALYTICS FLEXIBILITY	SQL
Cheap storage for large volumes of data	VOLUME	Expensive at large volumes of data
High agility with ability to quickly reconfigure for new workloads	AGILITY	Fixed configuration and limited agility
Data Engineers / Data Scientists / Analysts	USERS	Analysts / Business Users

Data Back Office with a Data Warehouse Centric Approach



Data Back Office Transformation with a Data Lake



State of Data Lake Adoption

The five stages of Data Lake Maturity

01 Stage

Aspiration

- Production Reporting/DW
- Researching

02 Stage

Experimentation

- Initial Big Data Deployment
- Targeted Use Case

03 Stage

Expansion

- Multiple Departments
- Multiple Engines
- Top Down Use Cases

04 Stage

Inversion

- Enterprise Transformation
- Bottoms up use cases

05 Stage

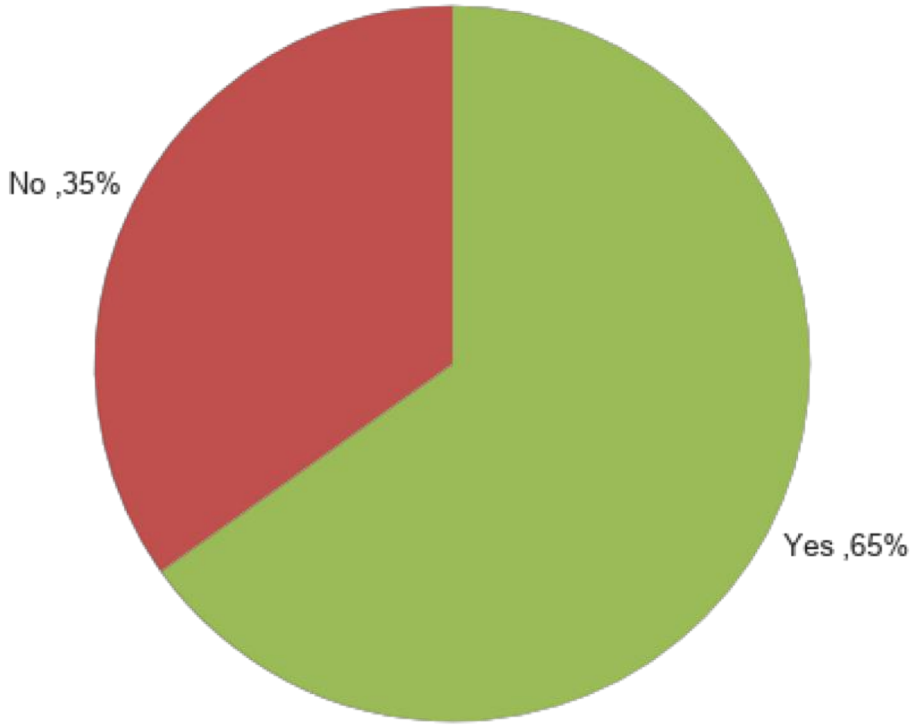
Nirvana

- Digital Enterprise
- Ubiquitous Insights
- True Business Transformation

Data Lake's Reality Gap: Everyone wants self-service nirvana

65% Moving to Self-Service Model to Enable Data Professionals

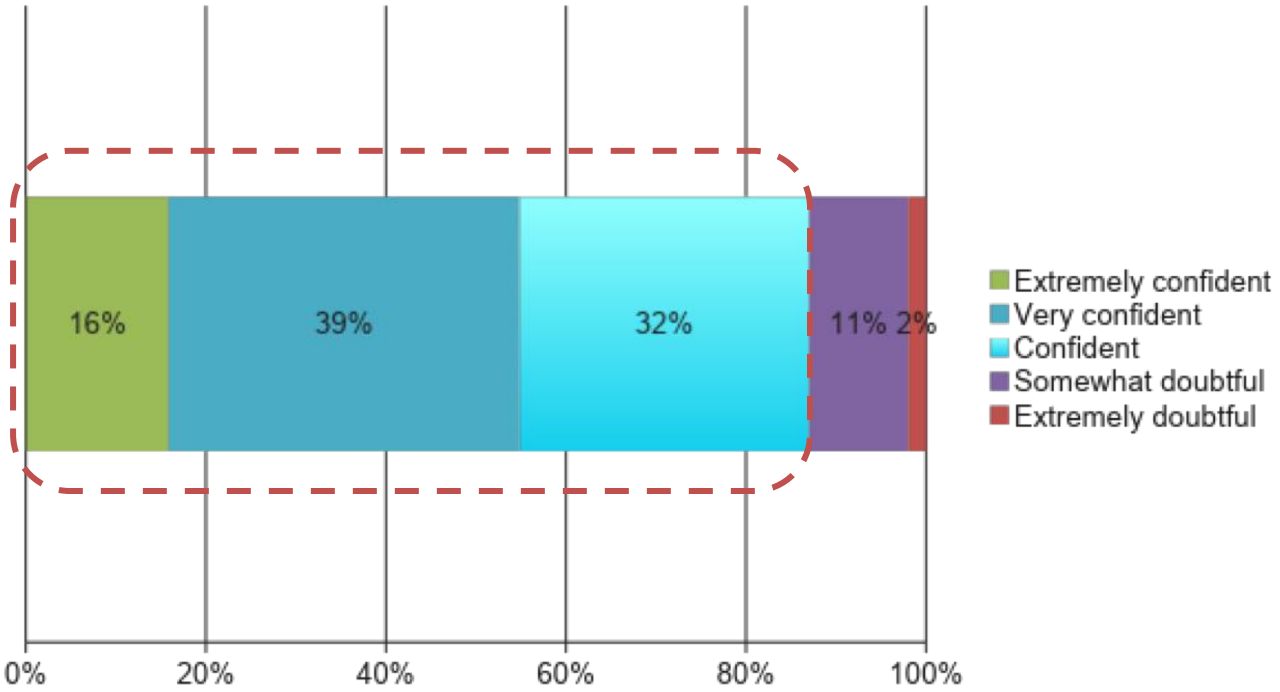
Is there a plan in place to move to a big data self-service analytics model?



Data Lake's Reality Gap: IT is confident they can get there

87% Confident They Can Provide Self-Service Analytics

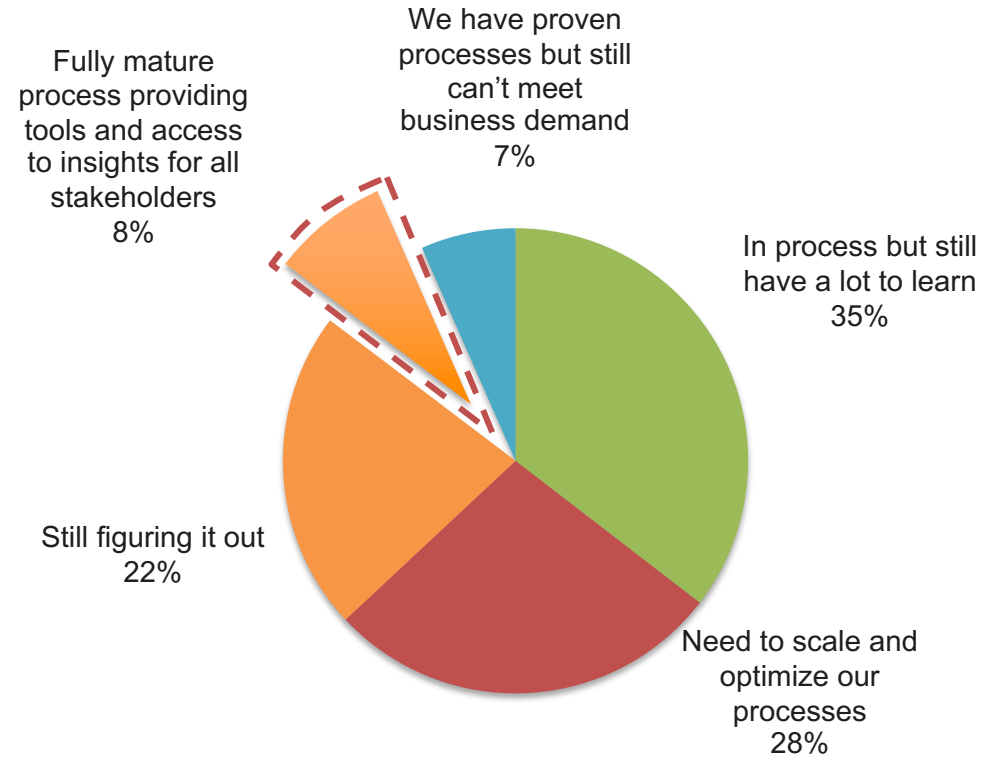
How confident are you that the data team can achieve self-service analytics?



Data Lake's Reality Gap: Only 8% are there today

Only 8% Have Mature Big Data Processes

How do you assess your big data maturity?



A Prescription to Success - Move to the Cloud

Adaptability

- Best machine configuration for the workload
- Best Engine for the workload
- On demand and elastic; automatically scale up or down

Agility

- Initial provisioning in min/hours, not months
- Change configurations dynamically
- Compute and Storage scale independently

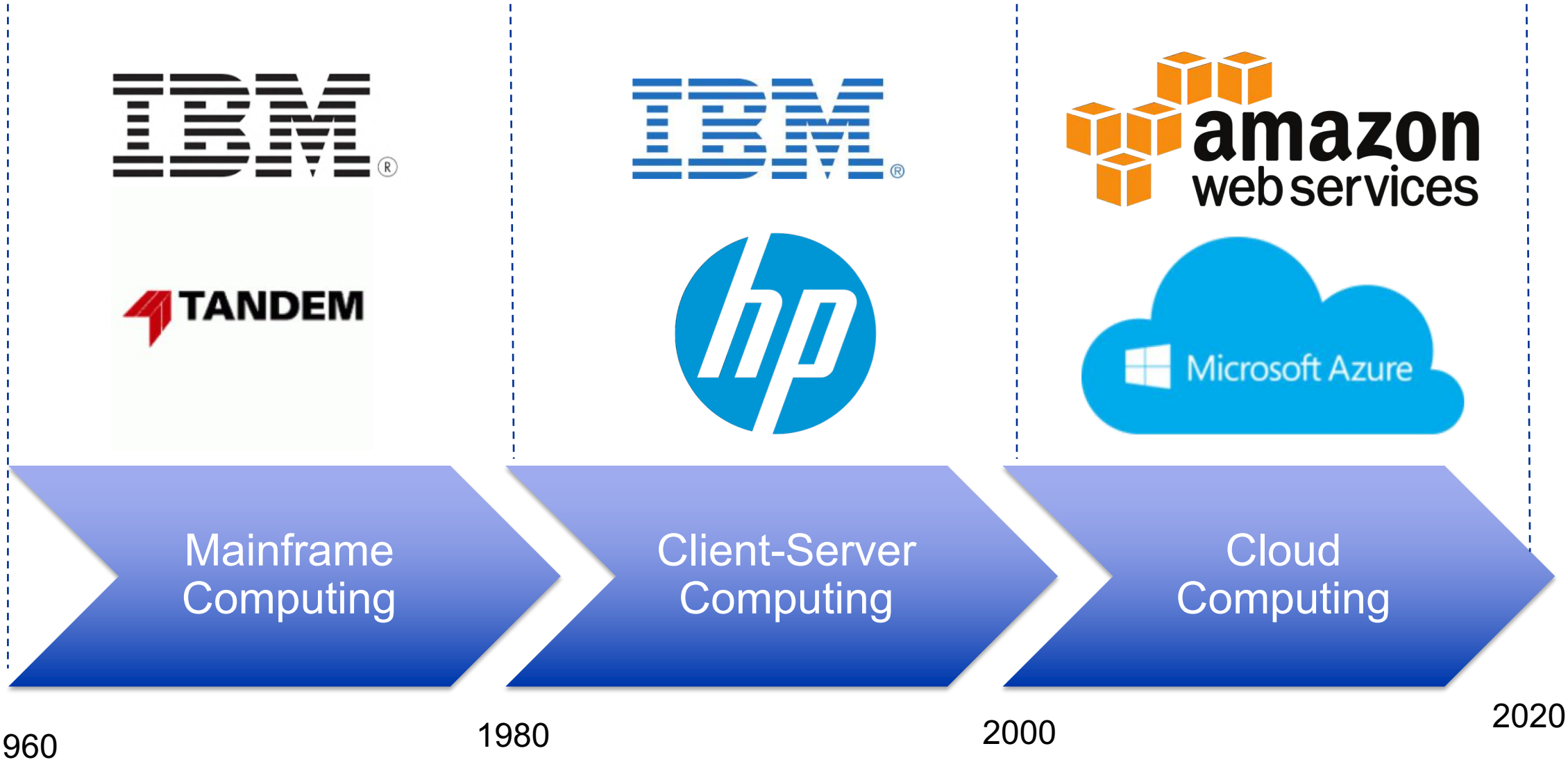
Cost

- Pay only for what you actually use
- Use spot instances to reduce cost by up to 80%

Cloud Computing and Data Lakes

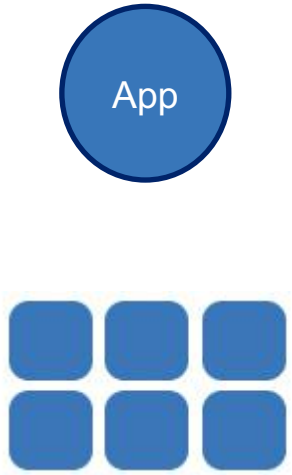
Infrastructure as a Service

Changing Nature of the IT Infrastructure

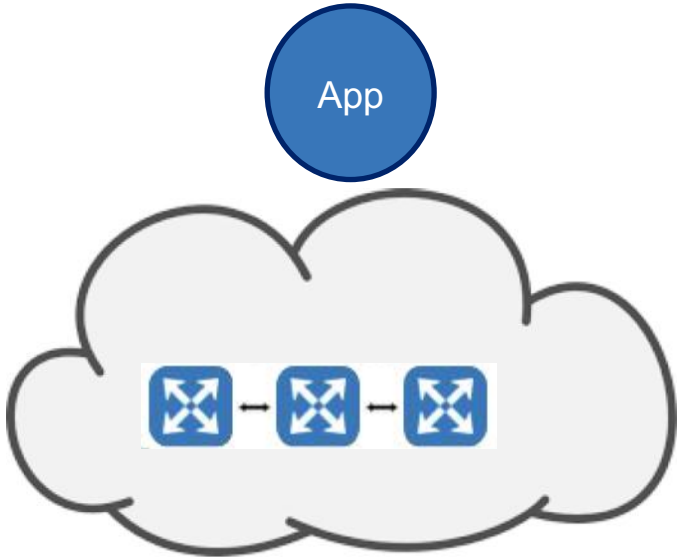


Cloud vs Data Centers

- Infrastructure is an API – Therefore Infrastructure can adapt to the needs of the Application



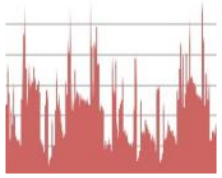
STATIC & RIGID



FLEXIBLE & ELASTIC

Properties of Data Lakes

Data Lakes are



Bursty

e.g. at Qubole we see on an average the minimum to maximum size of infrastructure to vary by 3400%

Ever Expanding

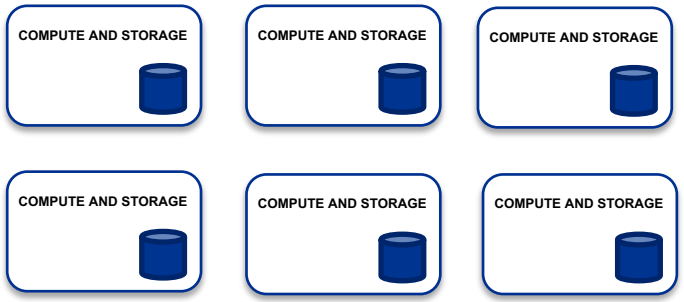
e.g. data processed on Qubole as grown 2.5x in a year

Rapidly Evolving

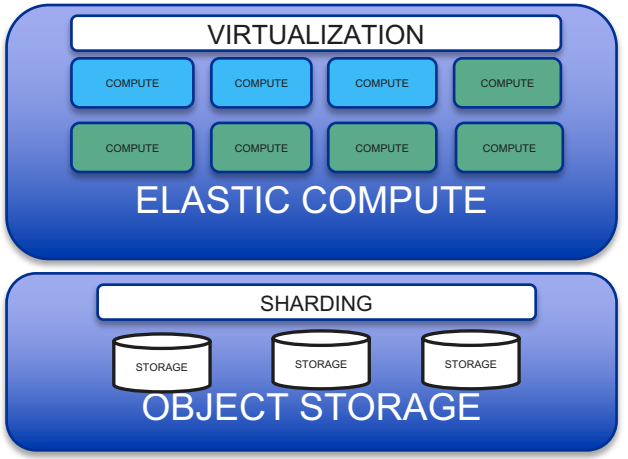
e.g. Spark, Presto and others addressing gaps in technology

Cloud-Native Big Data Platform – Separation of Compute and Storage

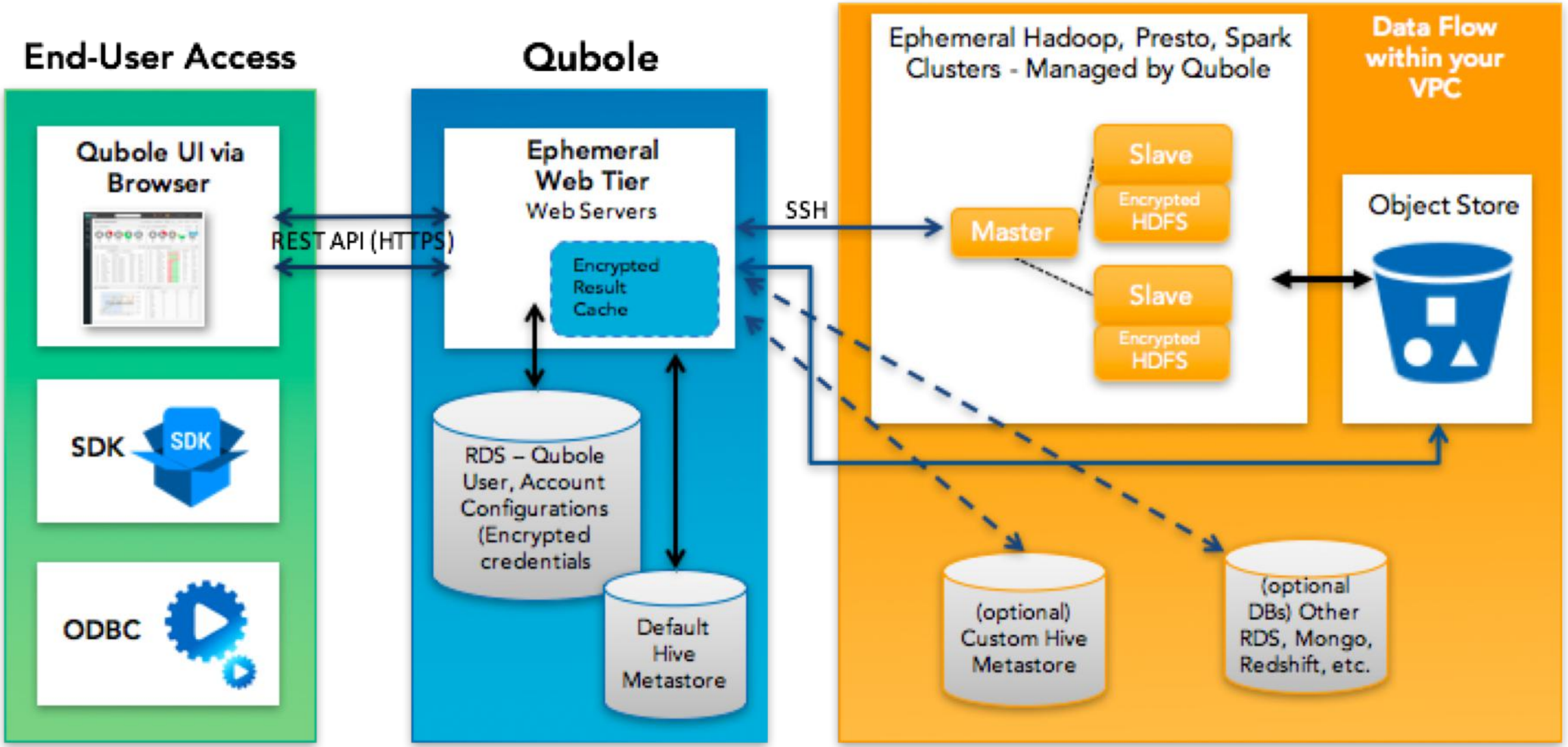
LEGACY DATA CENTER ARCHITECTURE



CLOUD INFRASTRUCTURE ARCHITECTURE



Architecture – Putting Together Cloud Data Lakes



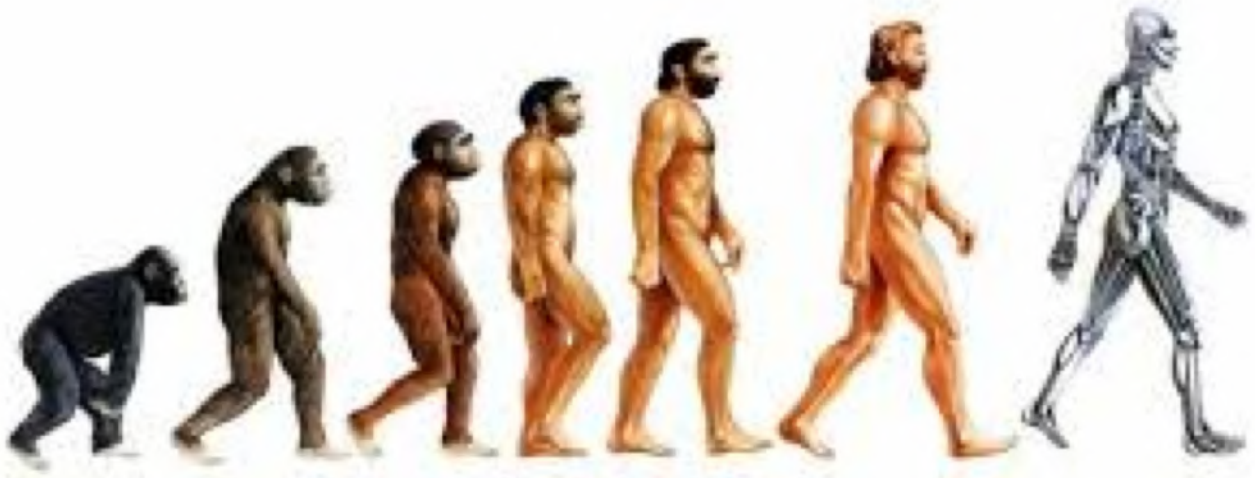
Cloud Data Lakes vs Data Center Data Lakes - Automation

Cluster Lifecycle Management

Auto start/terminate
Auto-scaling up/down

Performance Optimization

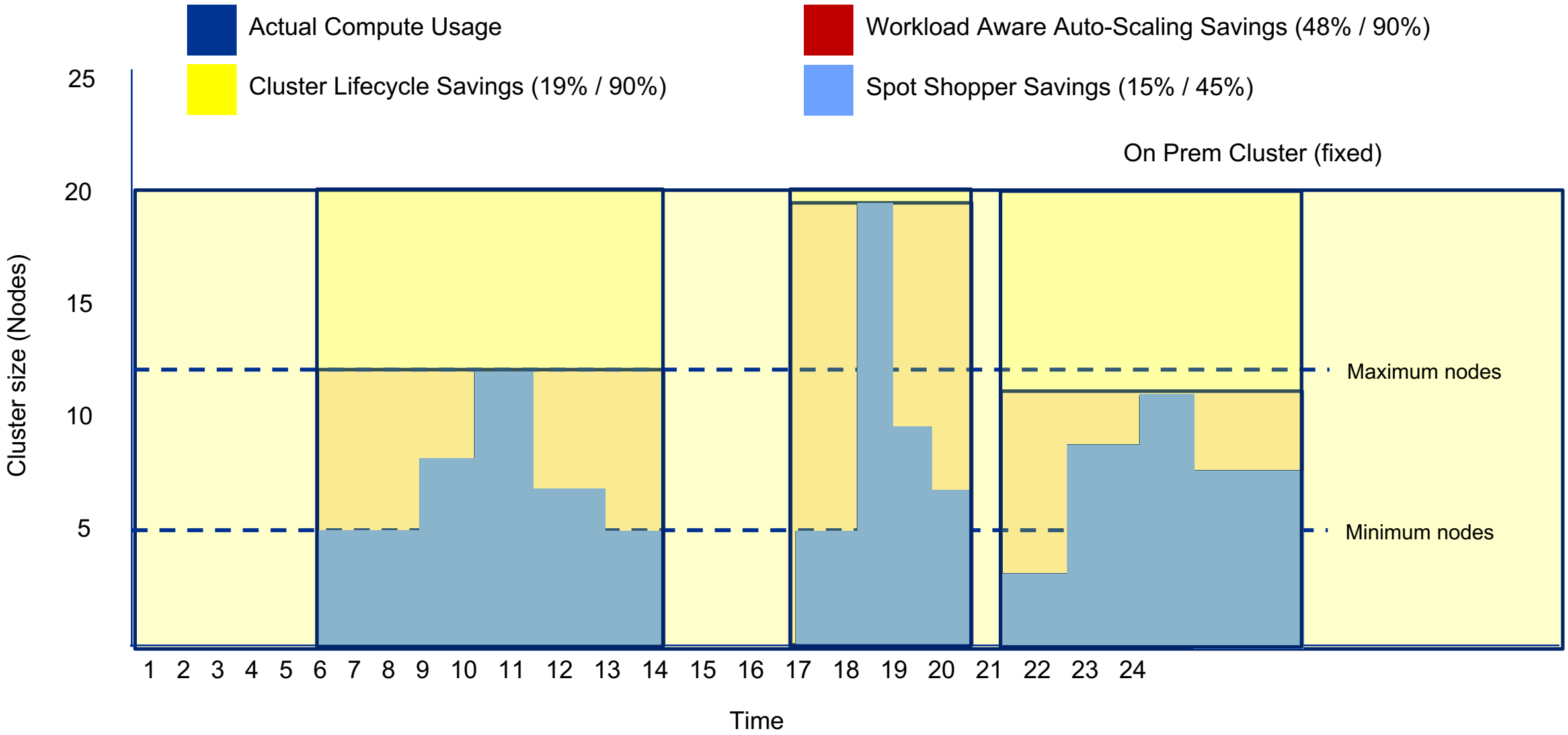
Cluster rebalancing
Performance/Caching



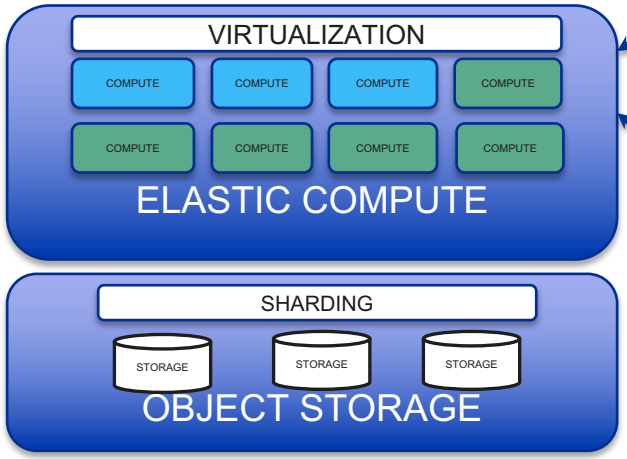
Cost Optimization

Spot node usage
Resource substitution

Cloud Data Lakes vs Data Center Data Lakes - TCO



Cloud Data Lakes vs Data Center Data Lakes – Concurrency and Elasticity




```
select t.county, count(1) from

(select transform(a.zip) using 'geo.py' as
a.county from SMALL_TABLE a) t

group by t.county;
```

```
insert overwrite table dest
select a.id, a.zip, count(distinct b.uid)
from ads a join LARGE_TABLE b on
(a.id=b.ad_id) group by a.id, a.zip;
```

Bringing it Together

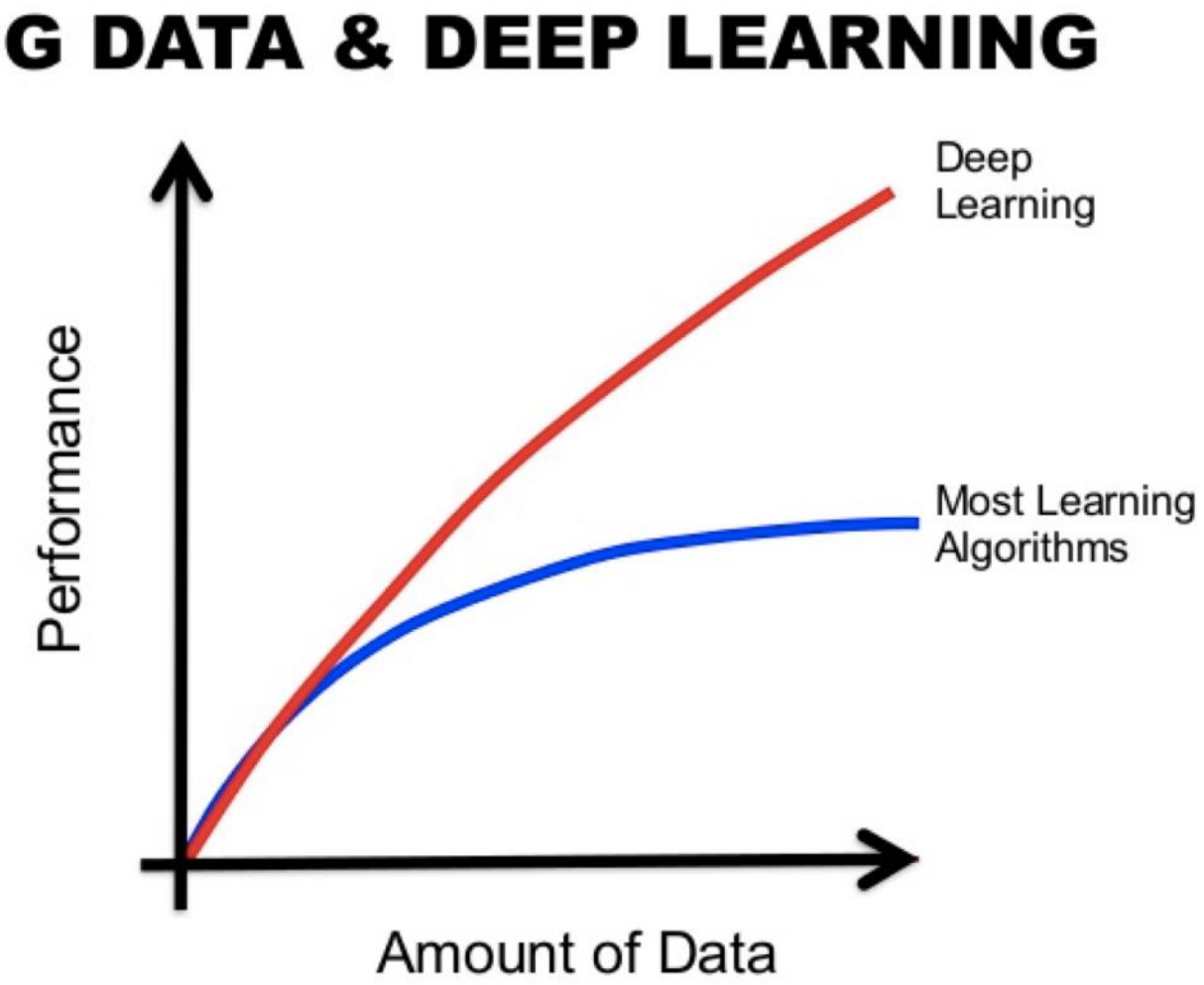
Using the Cloud for Big Data Platforms and Data Science Operations is *Fundamentally Different* from Operating Big Data Platforms On-Premise.

Done Properly This Leads to

1. Faster Time to Value
2. Increased Flexibility and Scale
3. Better Adoption of Analytics
4. Better TCO

Future Directions

(Re)Emergence of Deep Learning



Deep Learning Applications

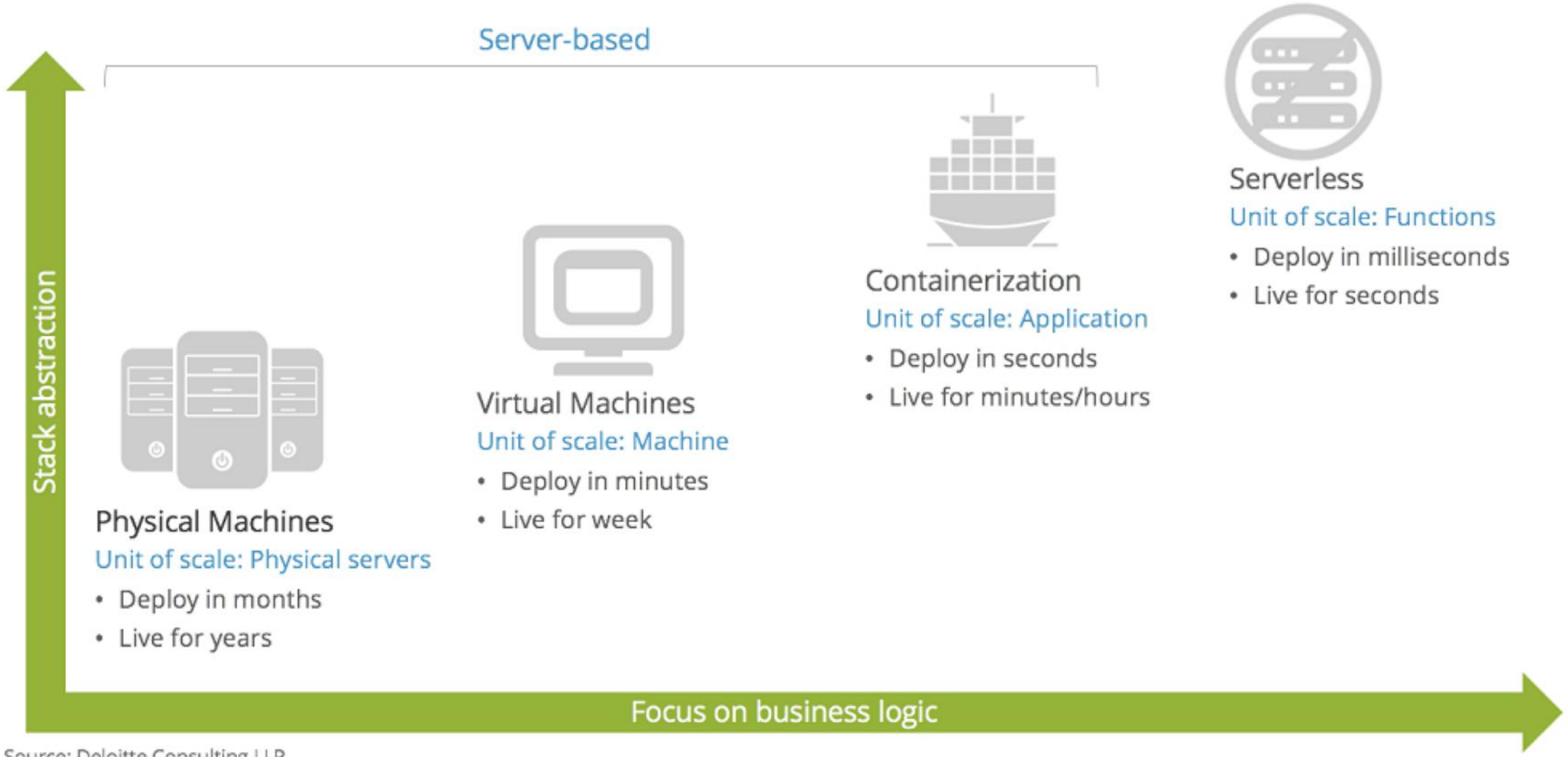
Applications today focused in areas around

- Image Recognition and Processing
- Speech Recognition and Processing
- NLP and Text Analysis

Emergence of New Use Cases and Technology

Deep Learning Platforms are Emerging

Serverless Computing



Source: Deloitte Consulting LLP

Advantages of Serverless

- Zero Administration
- Fast Bursting
- Cost Advantages

Contact Information

athusoo@qubole.com

Learn More at www.qubole.com

@qubole