THE COST OF CLOUD BIG DATA PLATFORMS

Dr Christopher Royles
Principle Systems Engineer | Cloudera
I work with organisations on both AWS and Azure, several use both.
DEPLOYMENT OPTIONS

- **Bare Metal**
  - Applications
  - Clusters
  - Operating System
  - Network
  - Storage
  - Servers

- **Private Cloud**
  - Applications
  - Clusters
  - Operating System
  - Network
  - Storage
  - Servers

- **Cloud IaaS**
  - Applications
  - Clusters
  - Operating System
  - Network
  - Storage
  - Servers

- **Cloud PaaS**
  - Applications
  - Clusters
  - Operating System
  - Network
  - Storage
  - Servers

- **Customer Managed**
- **Vendor Managed**
AN ANALYST VIEW

Cloud deployments will be the dominant environment in every category

Every cloud deployment environment will see increases in every workload category

Analytics and App Development areas expected strong gains

TRADITIONAL APPLICATIONS

Many data silos, each with its own proprietary tools and infrastructure
Different vendors, products, and services on-premises versus in cloud
A fragmented approach is difficult, expensive, and risky
TRADITIONAL SIZING

\[ H = \frac{crS}{(1 + i) \times 120\%} \]

\[ n = \frac{H}{d} = \frac{c*r*S}{(1-i)} / d \]
PARADOX, SCALABLE TRANSIENT CLUSTERS

- 15 x D12v2 / 2hr / ~$8.98
- 30 x D12v2 / 1hr / ~$8.97
- 6 x D15v2 / 1hr / ~$8.98
STAKEHOLDERS

**Knowledge Workers**
- Instant, self-service access to data and IT resources
- Application performance
- Job-oriented tools
- Choice and integration

**Infrastructure Team**
- Secure, controlled provisioning of data and IT resources
- Predictable infrastructure costs
- Systems-oriented tools
- Standardization and portability
CONSIDER THE PROFILE AND SLA

Data engineering - Batch oriented
Data Availability, inter-hour reporting

Analytic Database grows with concurrency, spikey reporting cycles. Business Visibility

Data science - Model building vs Model scoring. Scoring can be mission critical
“Think big, start small, iterate often”
CLOUDERA ENTERPRISE DATA PLATFORM

The modern platform for machine learning and analytics optimized for the cloud.
CLOUD STORAGE SELECTION

Consider the whole ecosystem. Due to significant differences in file system semantics and consistency it is hard to compare like-for-like.

Consider

- IOPS (analytics) v Throughput (ETL)

Storage Types

- Ephemeral & attached & data lake
- Consider latency & consistency & resilience
- IOPS and throughput can scale with cluster size

STORAGE

Amazon S3

Microsoft ADLS
NATIVE CLOUD INTEGRATION

Optimized compute engines integrated and tuned for commodity native cloud services

**2x** More efficient

**0.5x** The price of Native PaaS service
CROSS CLOUD PORTABILITY

No change in code, schema or data model.

No duplication of data and reduces data movement costs.

Many organizations have a duel cloud strategy with cross cloud use-cases.
INSTANCE SELECTION TYPE

Altus Data Engineering
Hive on Spark
( TP C-DS 3TB)
SUMMARY

Instance type selection

**c4 Family**
- Fast for Hive-on-Spark and Hive-on-MR for TPC-DS and Insert
- Failed on certain queries due to lack of memory

**r4 Family**
- Fastest for Spark wordcount
- More reliable for Spark workloads
INSTANCE SELECTION SIZE

Altus Data Engineering
Hive on Spark
R4 Instance Type
( temperament DS 3TB)
PRICE V CLUSTER SIZE

Altus Data Engineering
Hive on Spark
R4.2xl Instance Type
(TPC-DS 3TB)
Understand how BOTH performance and cost scale

Choose the best point to fit your goals and SLAs
HA
BDR

Consider a primary/secondary in place upgrade.

High resilience for Storage

Standby Cluster
SUMMARY

Scale up vs Scale out

Larger clusters, smaller nodes
- Scaling out with smaller instances is best for read-only workloads
- Better aggregate S3 throughput
- r4.2xlarge, r4.4xlarge are a good choice for Spark iterative workloads

Scale up can benefit write intensive workloads
- S3 metadata handling benefits larger master size

Test your own workloads
“Think big, start small, iterate often”

Amy O’Connor | CDIO
THANK YOU

Chris Royles
royles@cloudera.com
@royles