cloudera

Cloudera & The Cloud


Explosion of data and devices (IoT)
30B
connected devices
$200B
total market[1]
440x
more data
Transformation of IT infrastructure
cloud
open source
machine learning
The data-driven enterprise
cloudera
[1] IDC Worldwide Big Data and Business Analytics Market Through 2020

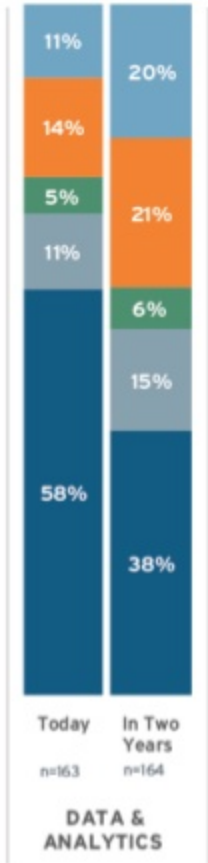
The Big Shift

In 2017

58% on-premises

11% private cloud

25% public cloud



By 2019

38% on-premises

15% private cloud

41% public cloud

Source: 451 Research, Voice of the Enterprise: Workloads and Key Projects, Cloud Transformation, 2017.

SaaS IaaS Hosted Private Cloud On-Premises Private Cloud Non-Cloud

My organization
is moving to the cloud,
why should we
consider Cloudera?

Current State

Stakeholders

KNOWLEDGE WORKERS

Instant, self-service access to data and resources

Application performance

Job-oriented tools

Choice

DATA TEAM

Advance strategic initiatives

Link analytics to business

Reduce admin burden

Integrated solutions

INFRASTRUCTURE TEAM

Secure, controlled provisioning

Predictable costs

Systems-oriented tools

Standards and portability

- Speed of deployment
- Tenant isolation
- Self-service
- Workload elasticity
- Shared storage
- Pay-as-you-go
- Bring your own tools
- Bring your own data

- Disjointed services
- Too many data copies
- Multiple security frameworks
- Difficult to troubleshoot workloads
- No shared metadata
- Unable to track data lineage
- Few on-premises integration services
- Proprietary services
- Cloud lock-in

Deployment options

Bare Metal	Private Cloud	Cloud IaaS	Cloud PaaS
Applications	Applications	Applications	Applications
Clusters	Clusters	Clusters	Clusters
Operating System	Operating System	Operating System	Operating System
Network	Network	Network	Network
Storage	Storage	Storage	Storage
Servers	Servers	Servers	Servers
On-premises			

Customer managed

Vendor managed

The Answer

cloudera®

- The **modern platform** for machine learning and analytics
- with multiple **deployment options**
- and one **shared data experience**

The modern platform for ML and Analytics...

DATA ENGINEERING

DATA PROCESSING

- Cost efficient
- Reliable
- Scalable

- Based on Spark, MapReduce, Hive & Pig
- Supported by Workload Analytics

ANALYTIC DATABASE

FAST BI & SQL

- Flexibility
- Elastic scale
- Go beyond SQL

- Based on Impala & Hive
- SQL dev enviro
- Supported by Workload Analytics

DATA SCIENCE

MACHINE LEARNING

- Fast dev to production
- Secure self-serve

- Based on Python, R, and Spark
- ML dev environment (CDSW)

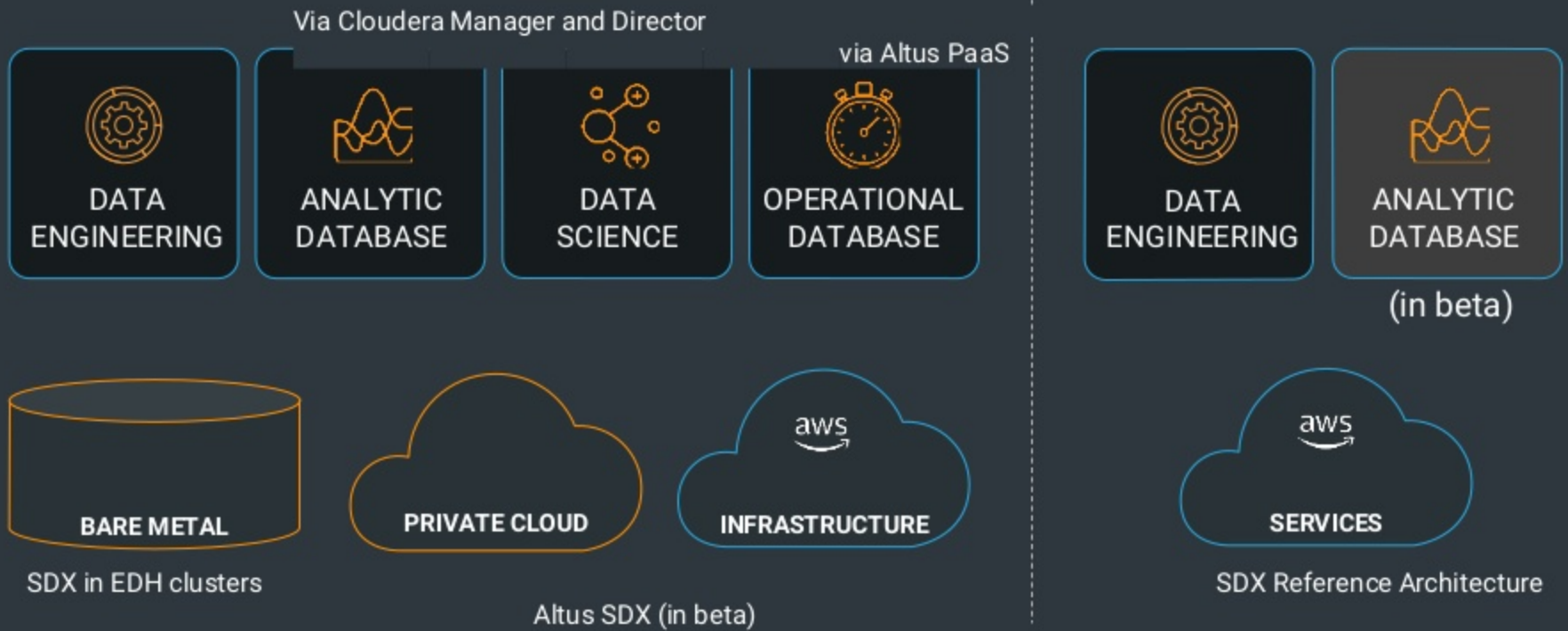
OPERATIONAL DATABASE

ONLINE & REAL-TIME

- High throughput, low latency
- Strongly consistent

- Based on Hbase, Kudu & Spark streaming

...With multiple deployment options...



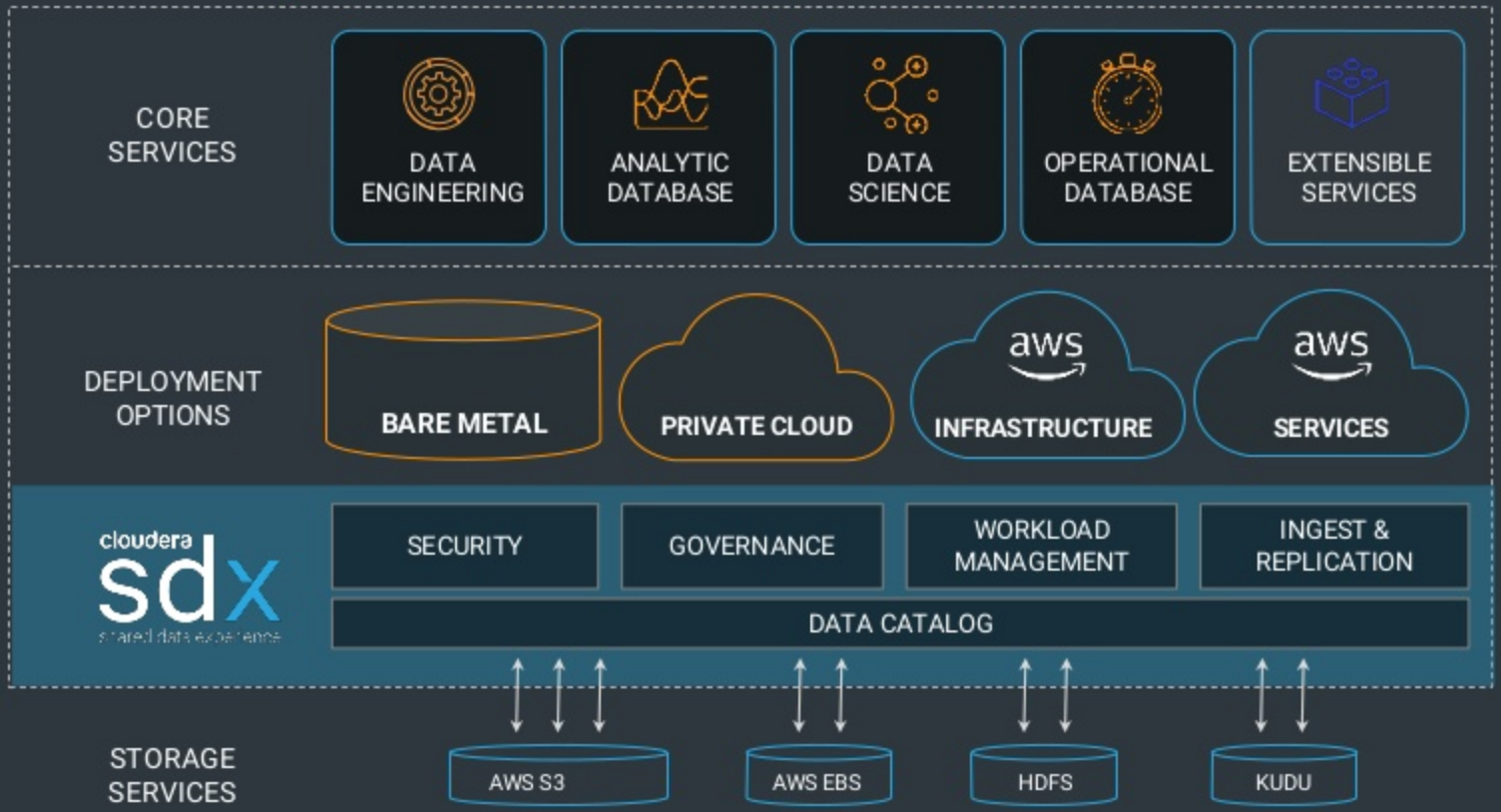
...And one Shared Data Experience

cloudera sdx
shared data experience

- Shared catalog
- Unified security
- Consistent governance
- Easy workload management
- Flexible ingest and replication

Cloudera Enterprise

The modern platform for machine learning and analytics optimized for the cloud



"Better to bet on cloud providers for infrastructure, Cloudera for data, analytics and security fabric, and leave the rest to the ecosystem"

- Eliminate data copies
- Single security framework
- Easy to troubleshoot workloads
- Universally shared metadata
- Easy to track data lineage
- Unified services
- Same solution as on-premises
- Same solution in multi-cloud
- Open source standards

- 2-4x storage savings, faster ramp
- Lower risk of data breach
- Analysts more productive on jobs
- Safe self-service, no shadow IT
- Less effort for audits/compliance
- IT more strategic, less admin time
- Deployment choices
- Portability and price arbitrage
- No proprietary lock-in

Cloud Customer Successes

- Manufacturing Customer
- Analytic DB, Data Science & Engineering on AWS

"We can deliver all of the data with less compute and much less complexity."

- Joe Smith, ACME Manufacturing

- Advisory and tech for finserv
- Cloudera EDH on AWS

"Cloudera gives us the best of both worlds—the ability to capture and process thousands of critical events at scale and the option to deploy on prem, in the cloud, or in a hybrid architecture. And running Cloudera on AWS enables us to leverage a world-class, reliable cloud infrastructure that supports our changing business needs and the needs of our customers."

- Kaushik Deka, CTO and Director of Engineering

NOVANTAS

cloudera

Call to Action

1. **Learn** Cloudera product offerings
2. **Deploy** using AWS Quickstart template
3. **Leverage** Professional Services
4. **Optimize** based on changing business need

Best practices for running your Cloudera Enterprise cluster on AWS

Quick AWS Component Review

EC2 - servers

S3 - object storage

EBS - block storage

RDS - we still need RDBMS

VPC - everything needs network

AWS Service Limits

Default limits on most AWS services (most can be increased)

- EC2 - ten (10) m4.4xlarge instances per region
- EBS - 20 TB of Throughput Optimized HDD (st1) per region
- S3 - 100 buckets per account
- VPC - 5 VPC per region

Defaults might limit your cluster, so plan ahead!

Deployment Topology

cloudera

Deployment Topology

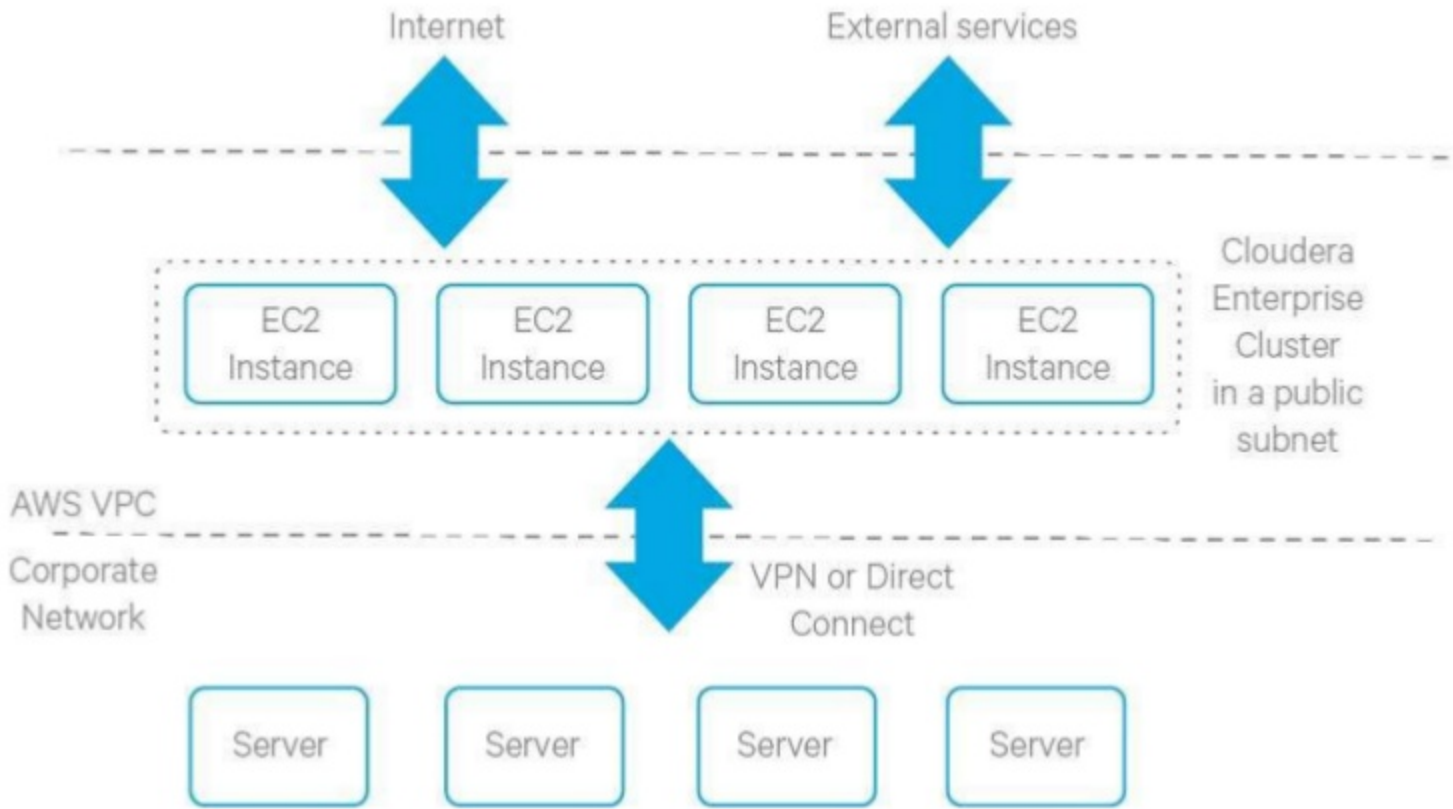
Two types of deployments from network perspective:

- Public Subnet - direct access to Internet & AWS services
- Private Subnet - instances must go through a VPC endpoints to reach AWS services & NAT instances for Internet

If you require high-bandwidth access to data sources on Internet or AWS services in another region, deploy to a public subnet.

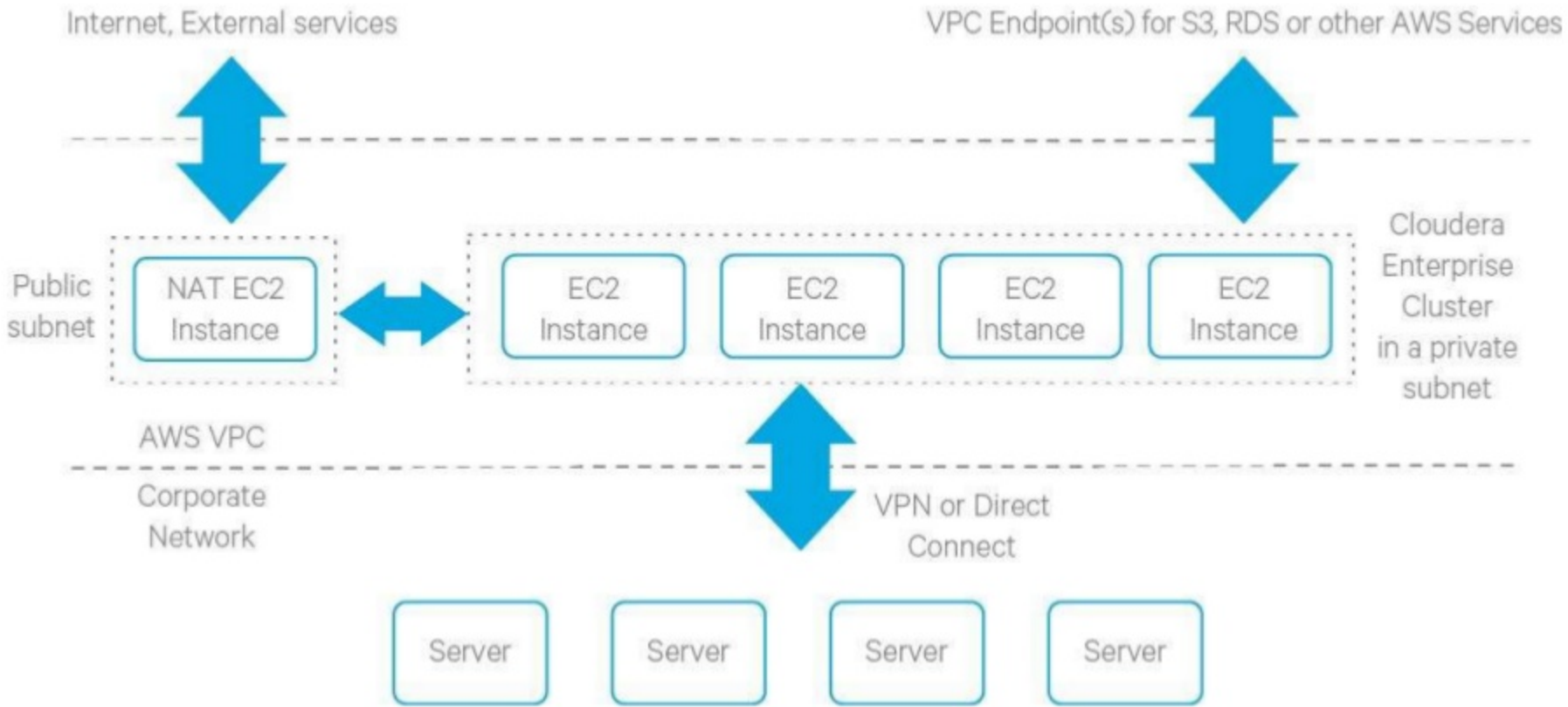
Public doesn't mean open to world. Limit access using Security Groups.

Deployment Topology (Public Subnet)



cloudera

Deployment Topology (Private Subnet)



cloudera

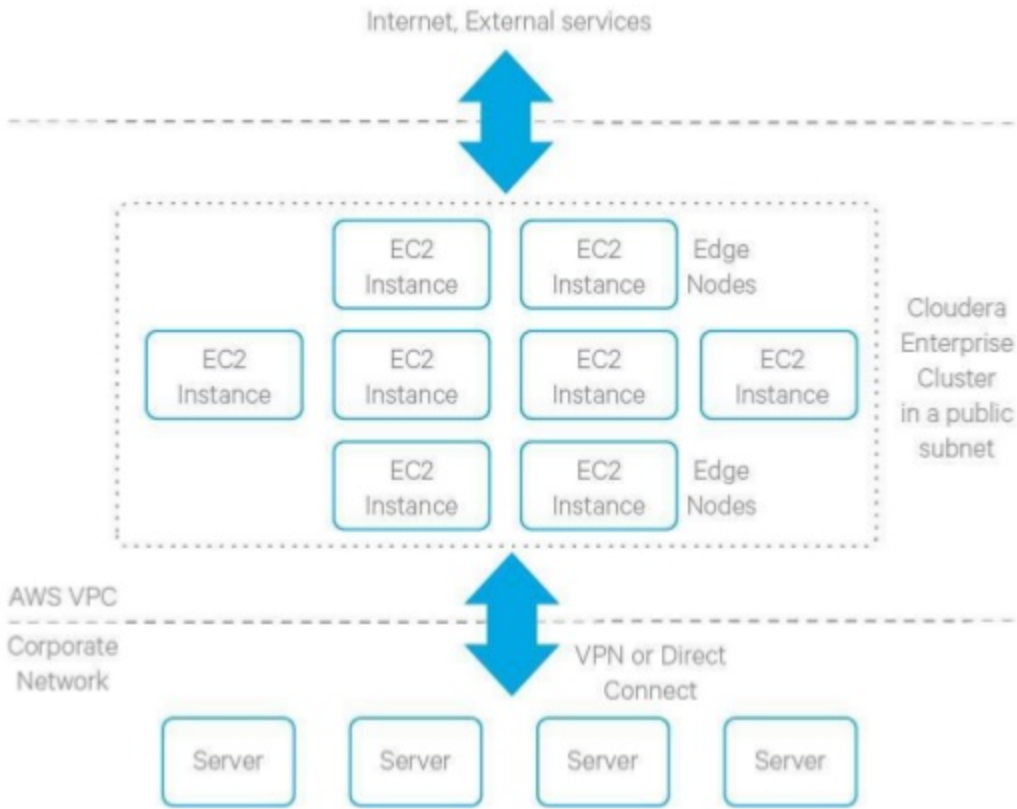
A Word About Edge Nodes

Edge nodes:

- Have direct access to cluster
- Users user these nodes via client applications
- Web applications, BI tools, or just Hadoop command-line client

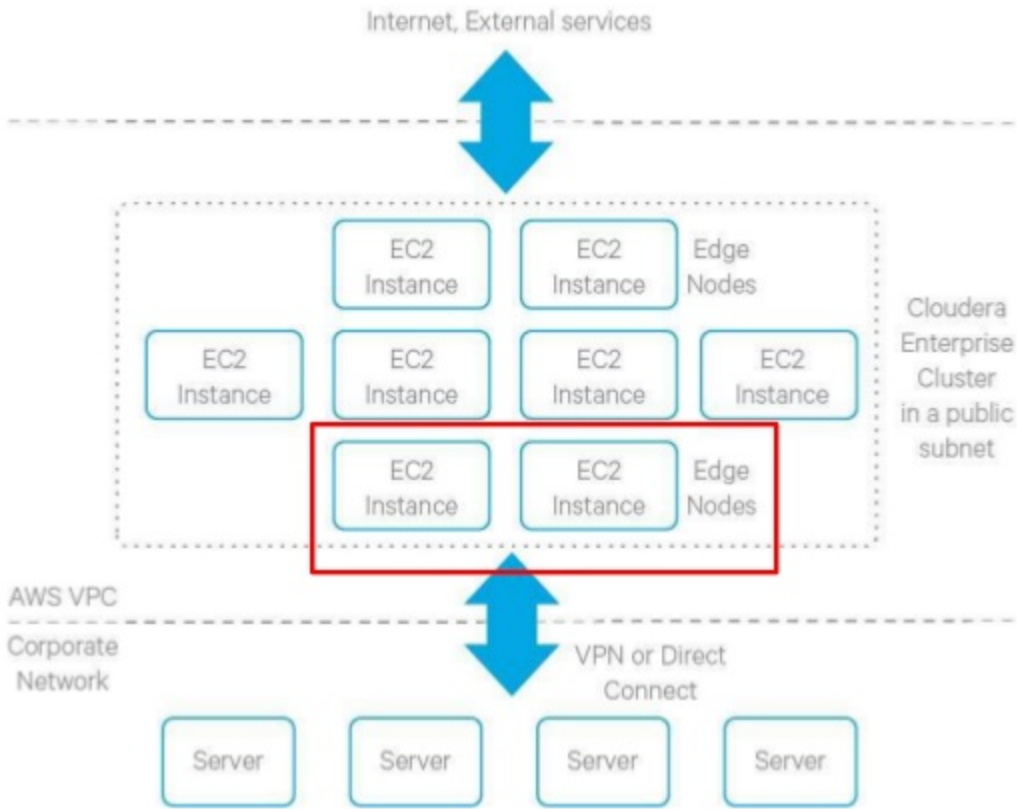
Avoid direct user access to the cluster!

Deployment Topology with Edge Nodes



cloudera

Deployment Topology with Edge Nodes



Roles and Instance Types

cloudera

So you want a cluster, eh?

Deployment Model

Two types of deployments:

- Persistent - long-lived, always on, no spin-up time
- Temporary - short-lived, can be started/stopped to save money

Impacts job setup time, instance types, storage method, and cost.

Instance Types

Ephemeral

- have locally attached storage, HDD or SSD
- on instance termination, data is irrecoverable

EBS-only

- no local storage, must mount EBS volumes
- on instance termination, data is safe in EBS

Networking
Connectivity
Security

cloudera

Storage Configuration

Three types of storage:

- Instance Storage
- Elastic Block Storage (EBS)
- Object Storage (S3)

Storage Configuration - Instance Storage

- Attached to EC2 instances, like physical disks on physical server
- Lifetime of storage == Lifetime of EC2 instance
- Each EC2 instance has different amounts of storage
  c1.xlarge has 4 x 420 GB
  d2.8xlarge has 24 x 2 TB
- For HDFS data directories, we like HDD instance storage
- Backup planning -- multi-instance shutdowns, multi-VM AWS events

Storage Configuration - EBS

- Persistent block level storage volumes
- Can be encrypted at rest w/ negligible impact to latency/throughput
- OS: General Purpose (gp2)
- DFS: Throughput Optimized HDD (st1), Cold HDD (sc1)
- Baseline & burst performance increase with size of provisioned volume
  e.g. 500 GB st1 baseline throughput 20 MB/s; 1000 GB → 40 MB/s

Storage Configuration - EBS Recommendations

Instance selection

- Use EBS-optimized instances OR instances with 10Gb+ network
- Minimum dedicated EBS bandwidth of 1000 Mb/s (125 MB/s)

Volume selection

- Baseline performance, 40 MB/s or better (1000 GB st1, 3200 GB sc1)
- Do not exceed instance's dedicated EBS bandwidth!

Storage Configuration - S3

- Great for cold backup: durable, available, inexpensive
- For hot backup, use a second HDFS cluster
- Hive and Spark can also use S3 directly
- Standard data operations can read from & write to S3 buckets

Storage Configuration

Three types of storage:

- Instance Storage
- Elastic Block Storage (EBS)
- Object Storage (S3)
- Root Device

Storage Configuration - Root Device

- For operating system and logs
- Use EBS gp2 volumes as root devices
- At least 500 GB for OS, CDH software, and logs
- Do not use instance storage for the root device!

Capacity Planning

cloudera

Capacity Planning

- AWS makes expansion easy; advance planning makes things easier
- Consider workloads: how much storage vs compute? balanced?
- Consider data replication (3x), growth, retention
- Low storage density, r3.8xlarge or c4.8xlarge provide less storage but higher compute and memory
- High storage density, d2.8xlarge offers 48 TB per instance with a good amount of compute and memory

Cloudera Enterprise Hardware Requirements Guide

tiny.cloudera.com/hw-reqs

Provisioning Instances

cloudera

Provisioning Instances

- Cloudera Director automates most things
- Manual via EC2 command-line API tool or AWS management console
- Don’t forget your databases (either RDS or self-managed)
- Cloudera Altus

Provisioning Instances

No matter which route you take...

- root device: 500 GB+ gp2 EBS volume
- master metadata: ephemeral or recommended gp2 EBS volumes
- DFS data: ephemeral or recommended st1/sc1 EBS volumes
- use tags to indicate the role instances/volumes will play

Thank You

cloudera

