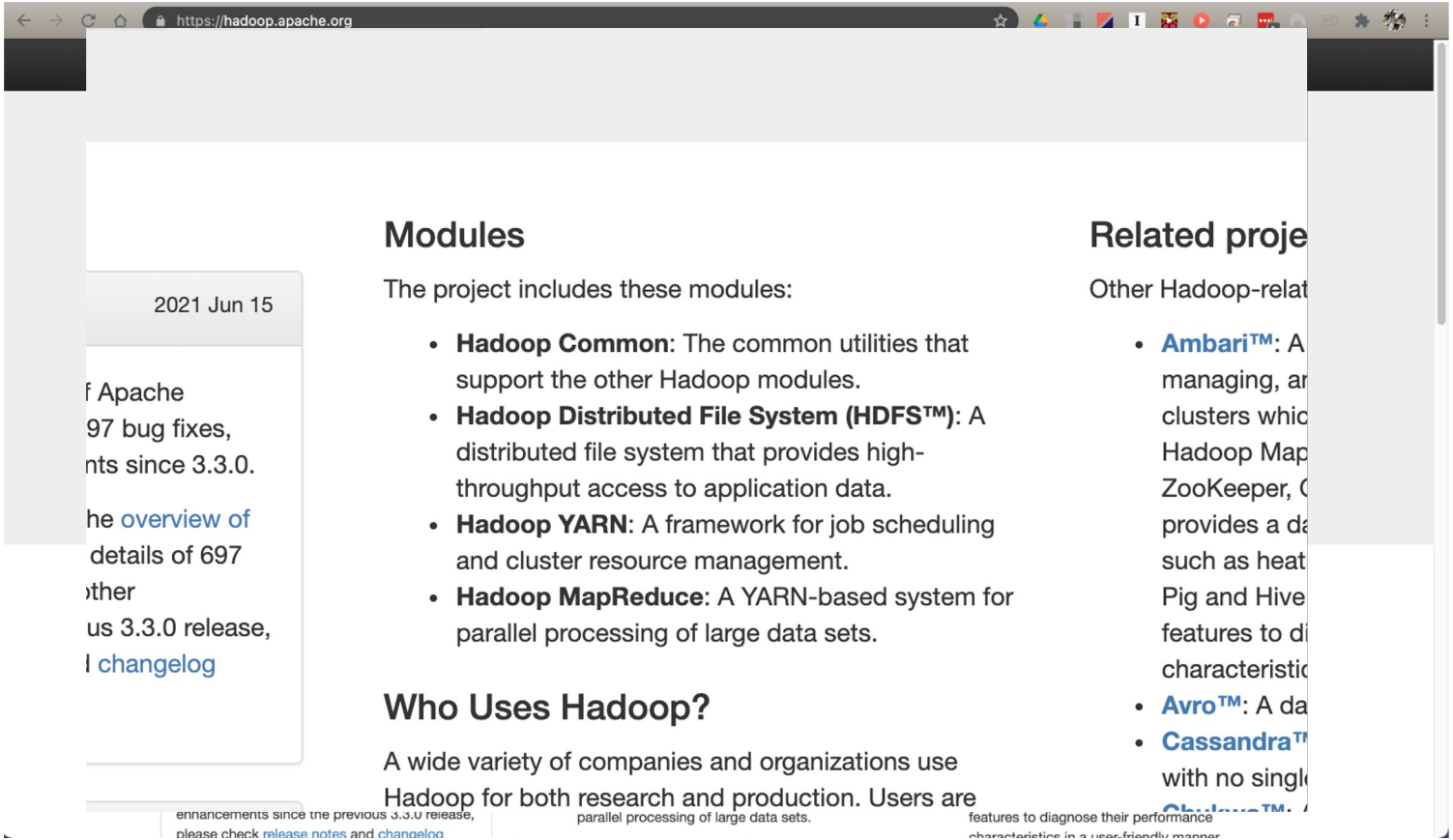
Cloud & Data Management

CMPT 732 - Fall 2023

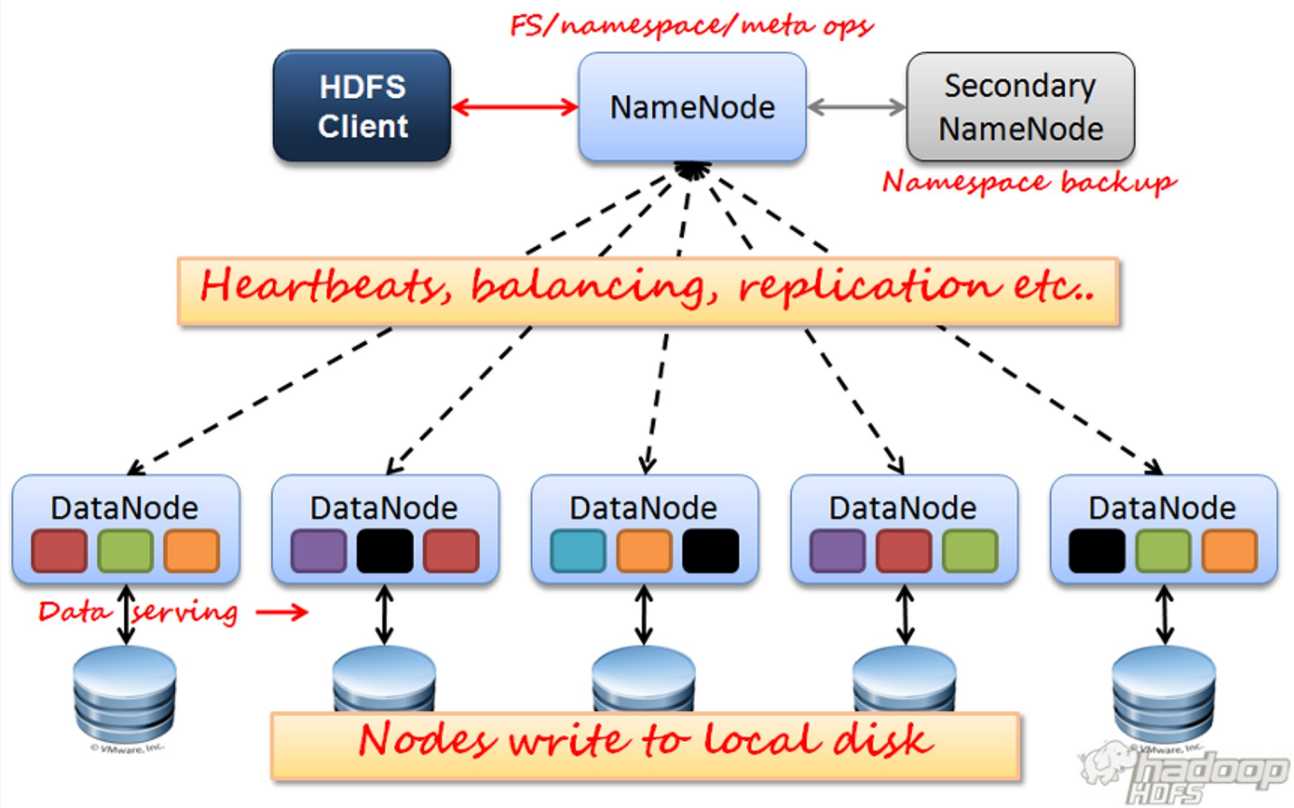
Agenda

- Hadoop & Spark recap
- Cloud Computing
- Data Management practices

Hadoop: overview

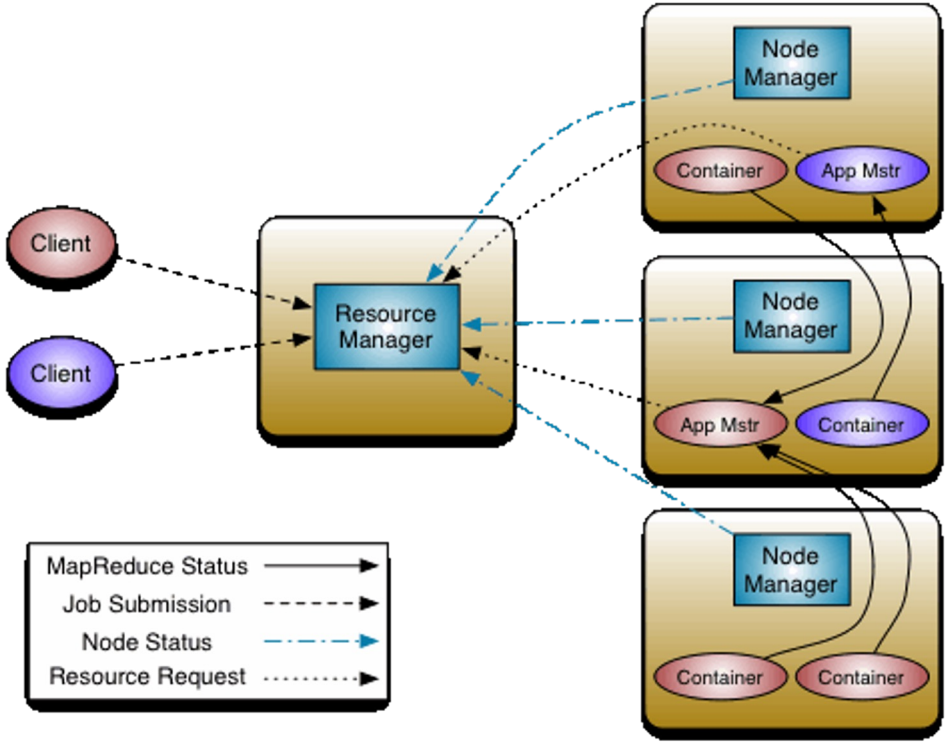


Hadoop: storage

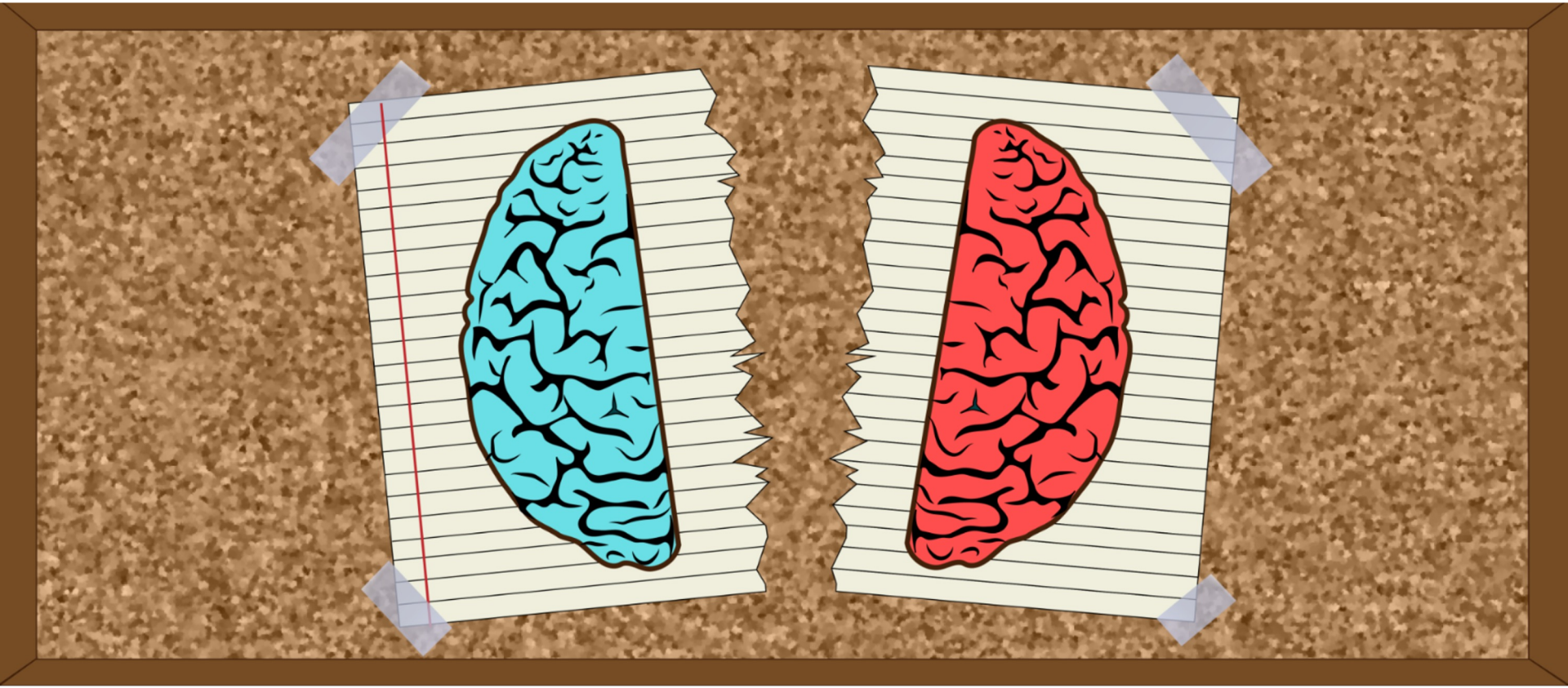


https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-

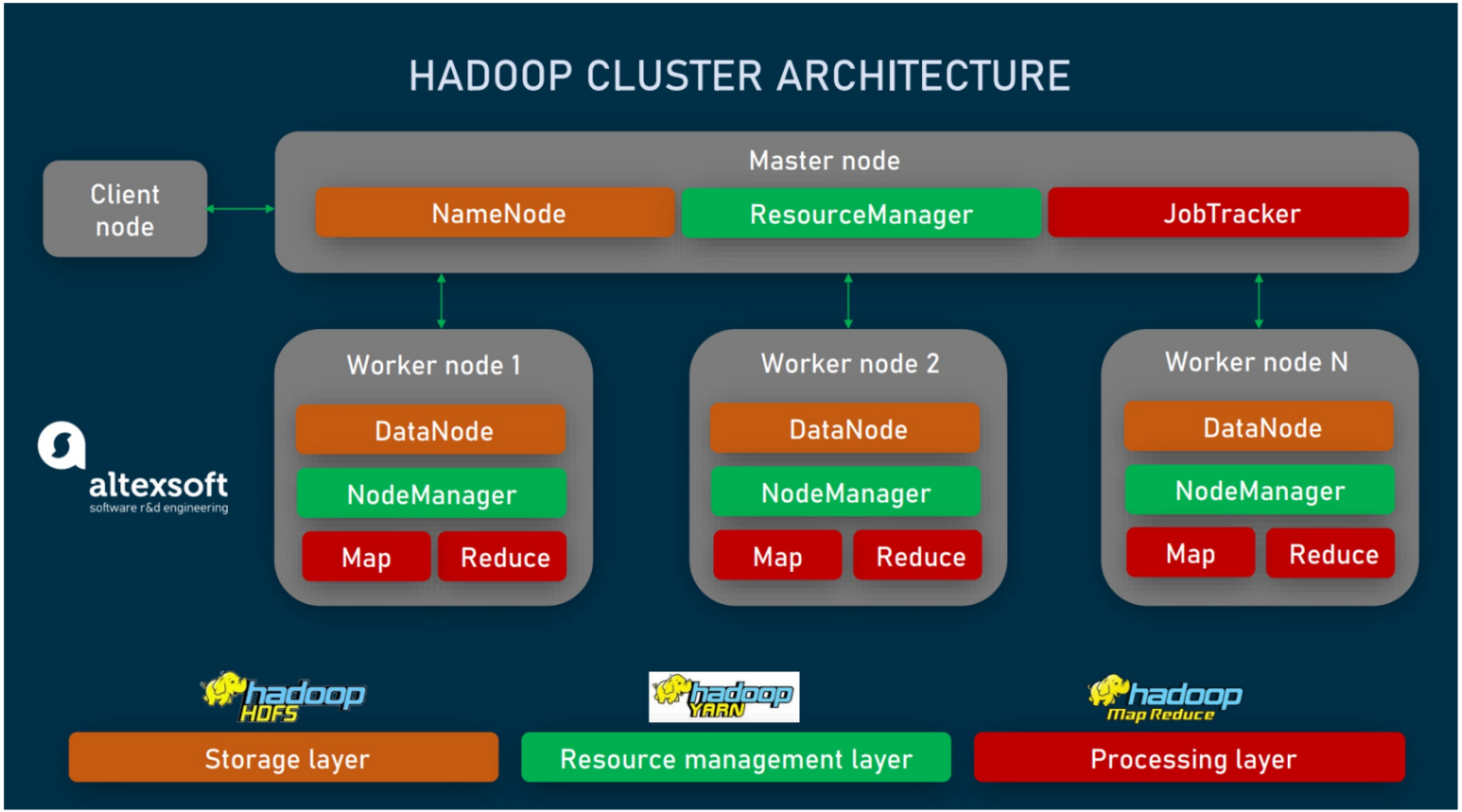
Hadoop: compute



https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-

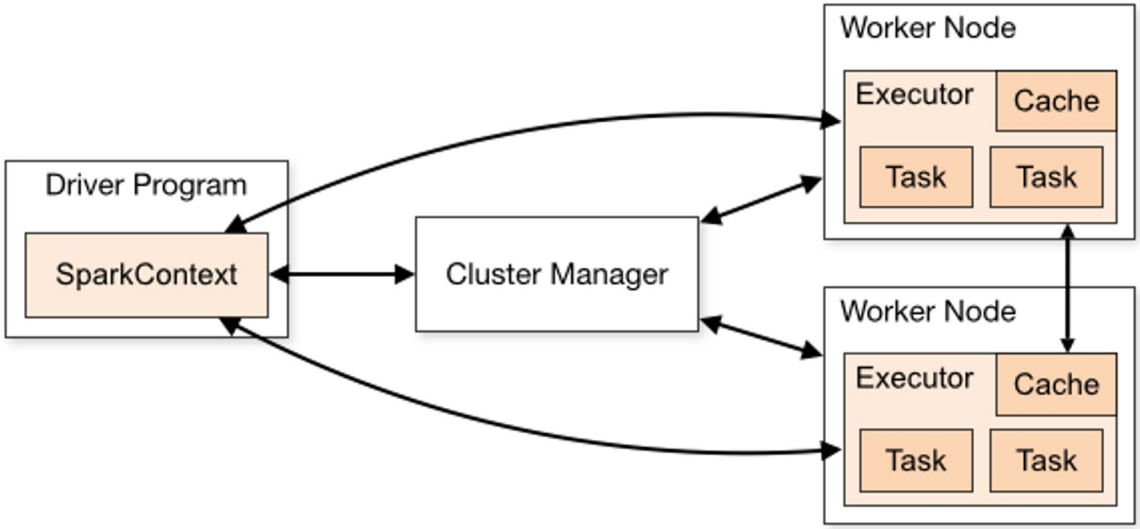


Hadoop: together



https://www.altexsoft.com/blog/hadoop-vs-spark/

Spark cluster architecture

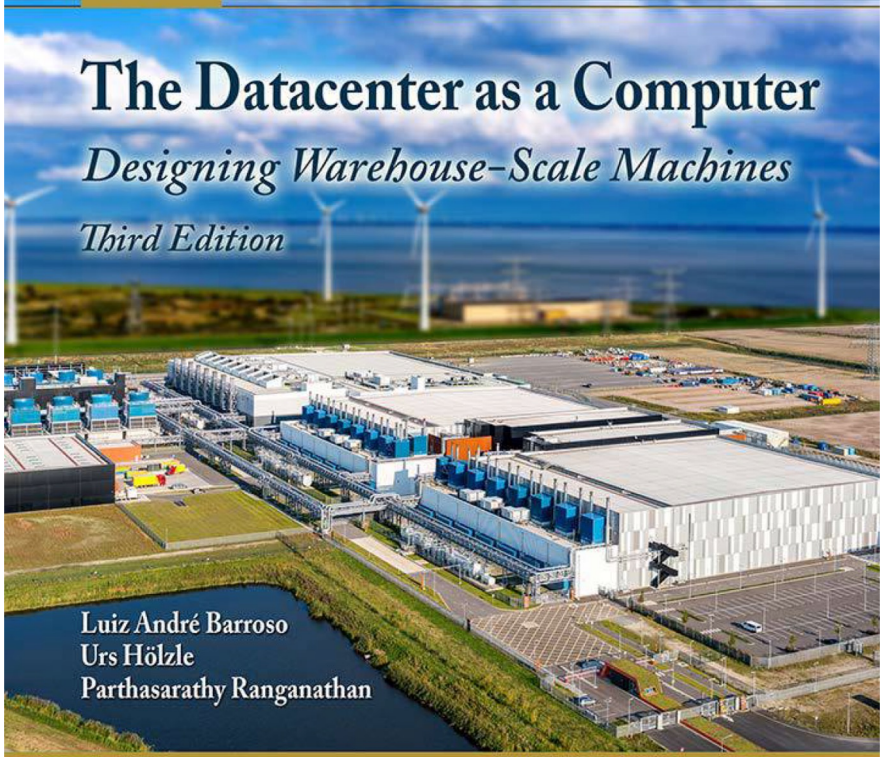


MORGAN & CLAYPOOL PUBLISHERS

The Datacenter as a Computer

Designing Warehouse-Scale Machines

Third Edition



Luiz André Barroso
Urs Hölzle
Parthasarathy Ranganathan

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

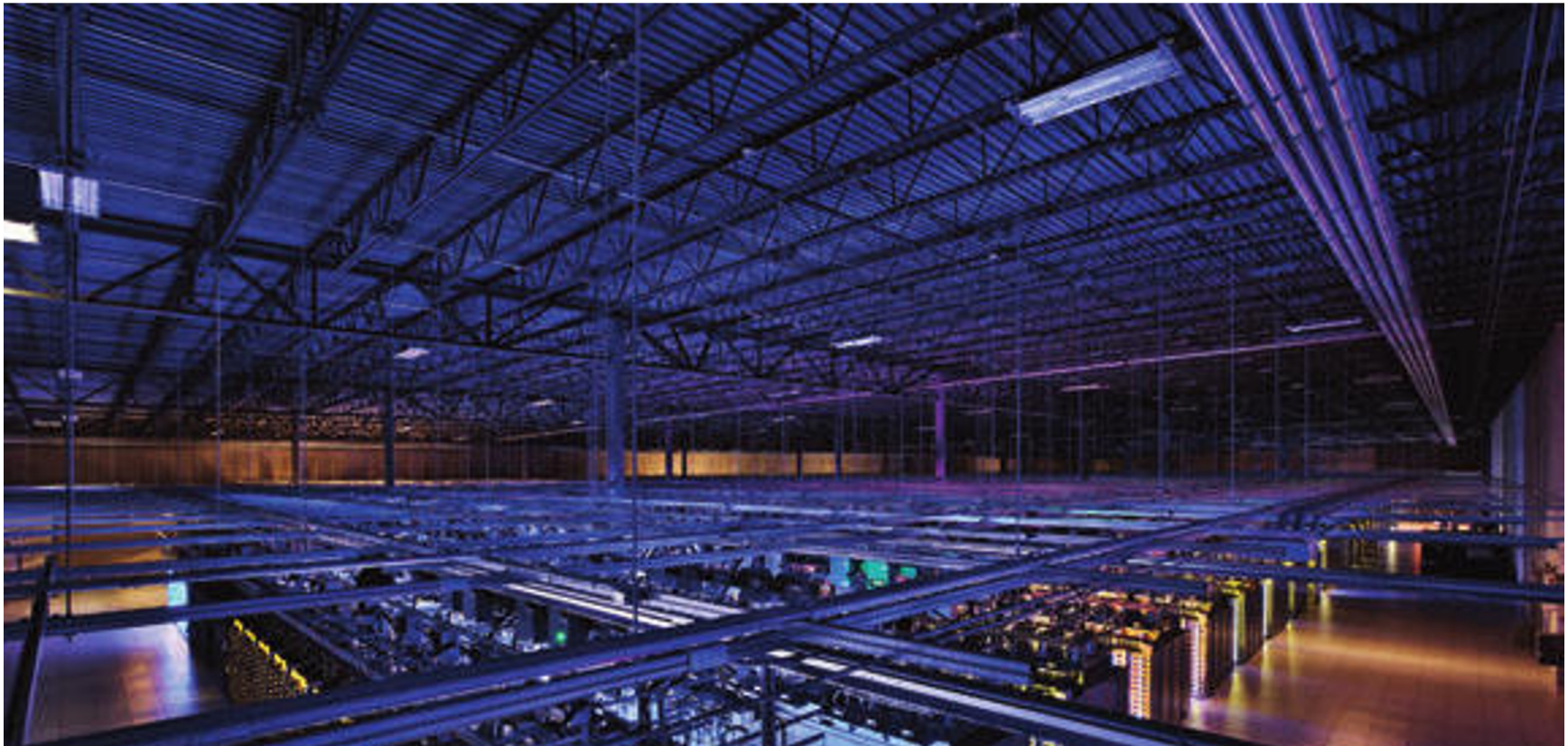
Margaret Martonosi, *Series Editor*

Not one computer

https://www.trentonsystems.com/blog/what-is-a-

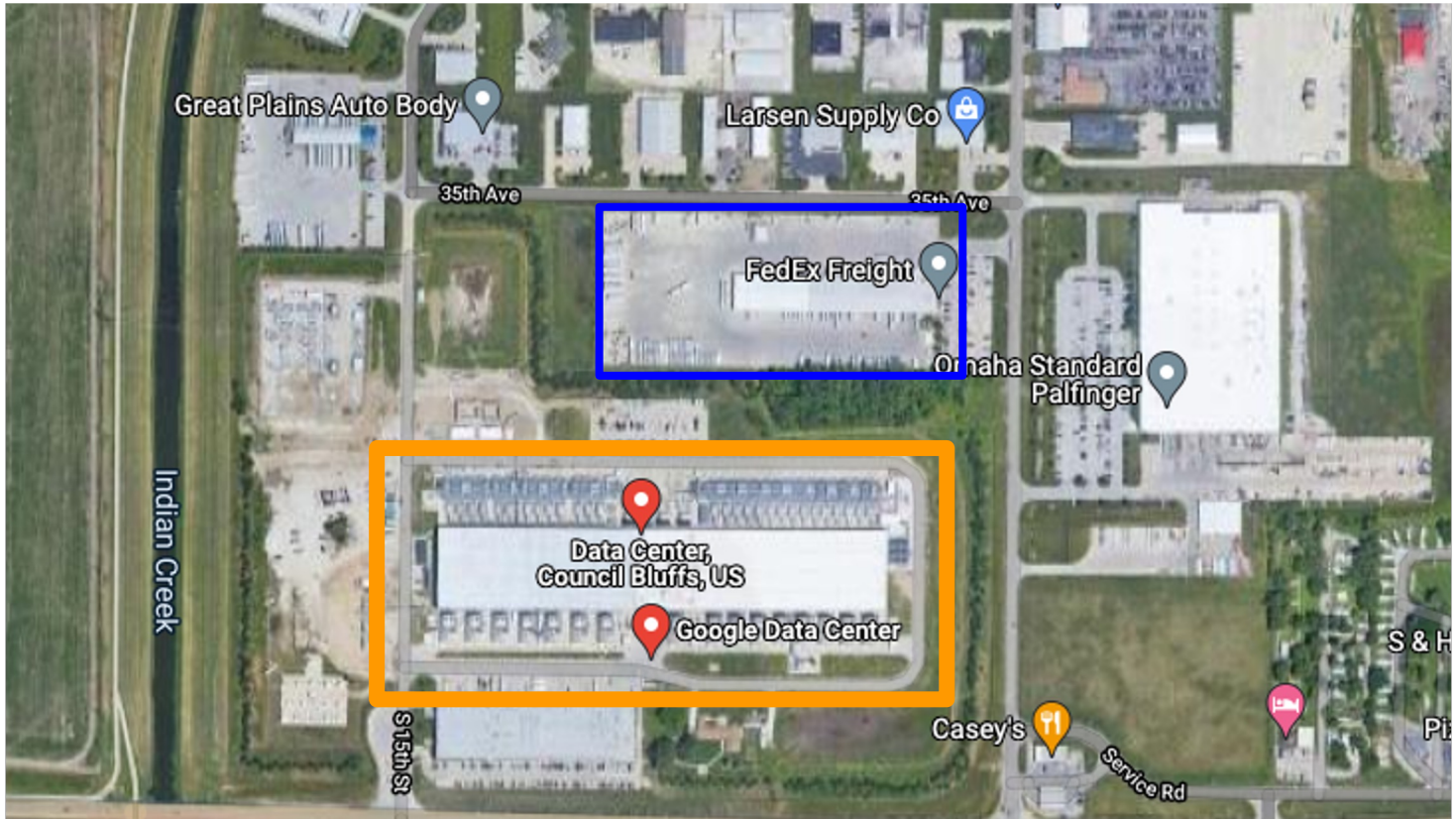
Datacenter

Datacenter: Power distribution



Datacenter: Cooling

The scale of a datacenter



Cloud computing...

Infrastructure as a Service (IaaS)

IaaS contains the basic building blocks for cloud IT. It typically provides access to networking features, computers (virtual or on dedicated hardware), and data storage space. IaaS gives you the highest level of flexibility and management control over your IT resources. It is most similar to the existing IT resources with which many IT departments and developers are familiar.

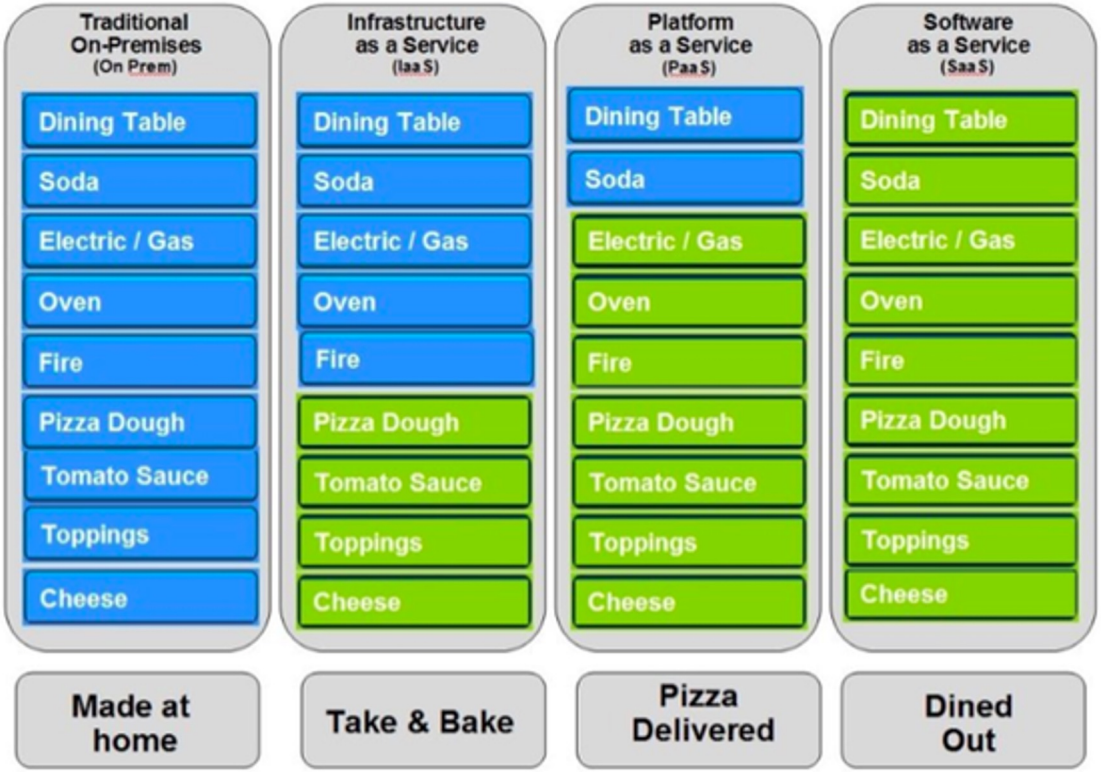
Platform as a Service (PaaS)

PaaS removes the need for you to manage underlying infrastructure (usually hardware and operating systems), and allows you to focus on the deployment and management of your applications. This helps you be more efficient as you don't need to worry about resource procurement, capacity planning, software maintenance, patching, or any of the other undifferentiated heavy lifting involved in running your application.

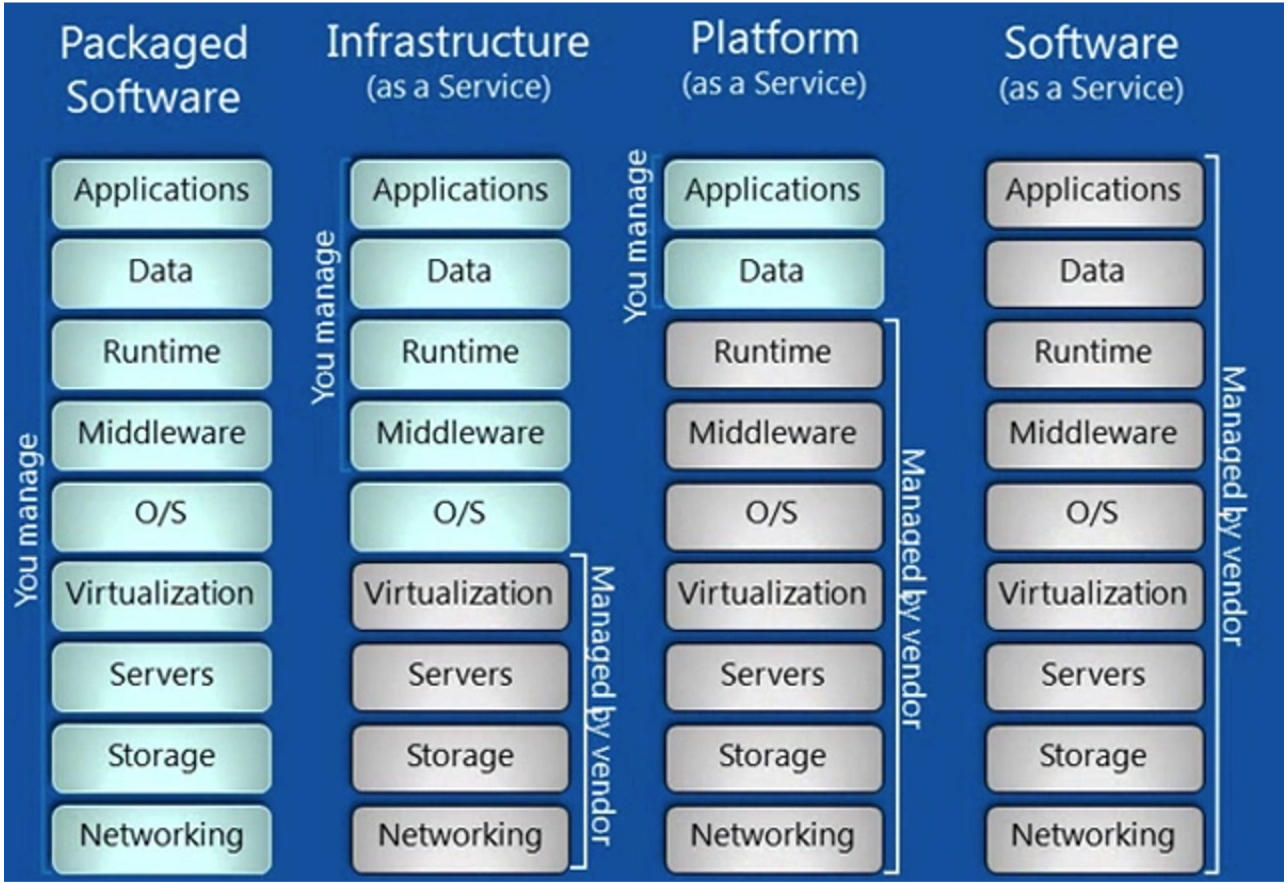
Software as a Service (SaaS)

SaaS provides you with a complete product that is run and managed by the service provider. In most cases, people referring to SaaS are referring to end-user applications (such as web-based email). With a SaaS offering, you don't have to think about how the service is maintained or how the underlying infrastructure is managed. You only need to think about how you will use that particular software.

Pizza as a Service

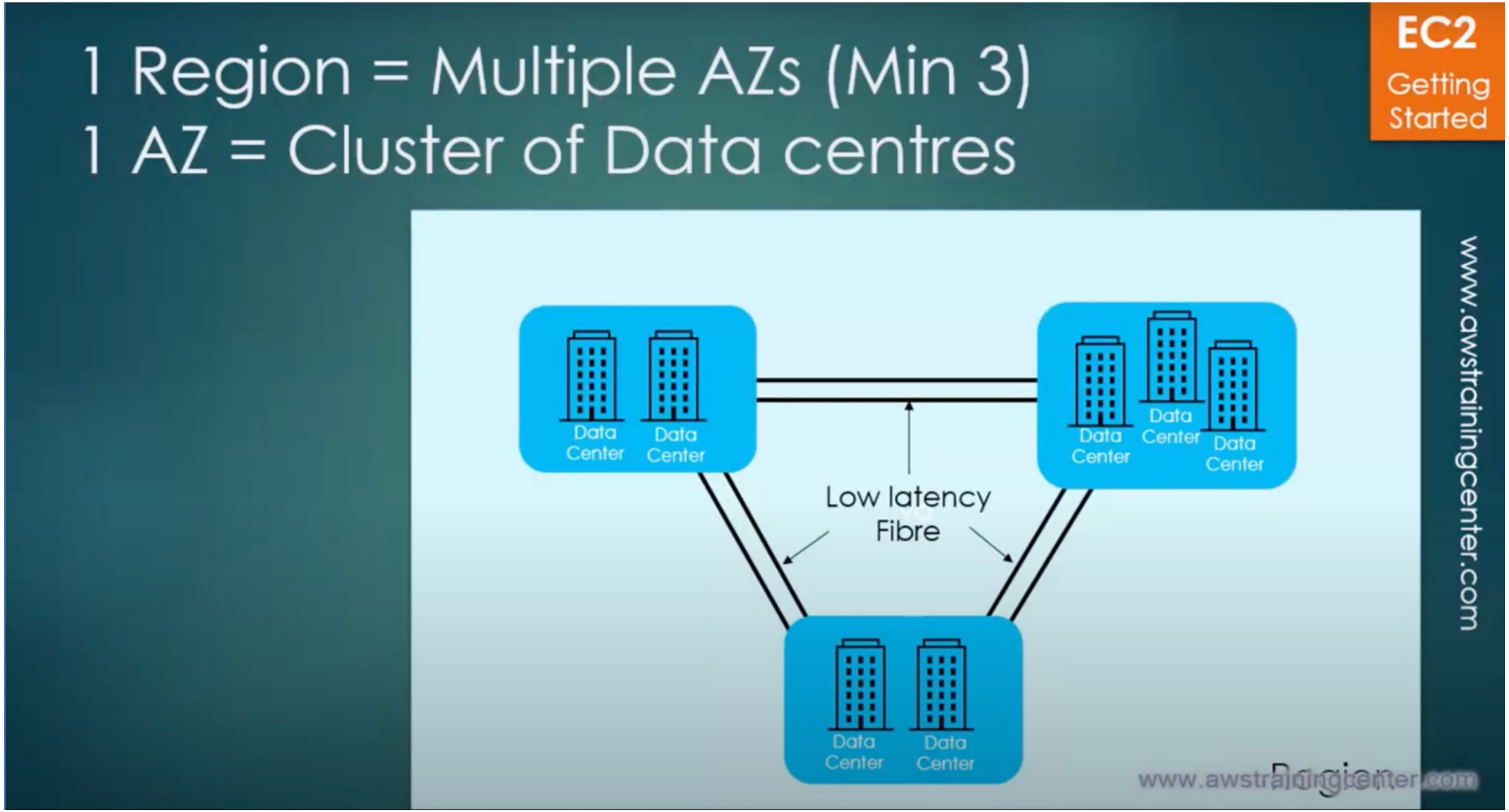


Source: "Pizza as a service"



https://venturebeat.com/2011/11/14/cloud-iaas-

Region & AZ


1 Region = Multiple AZs (Min 3)
1 AZ = Cluster of Data centres
EC2
Getting Started
Data Center
Data Center
Data Center
Data Center
Data Center
Data Center
Data Center
Low latency Fibre
Region
www.awstrainingcenter.com
www.awstrainingcenter.com

AWS services

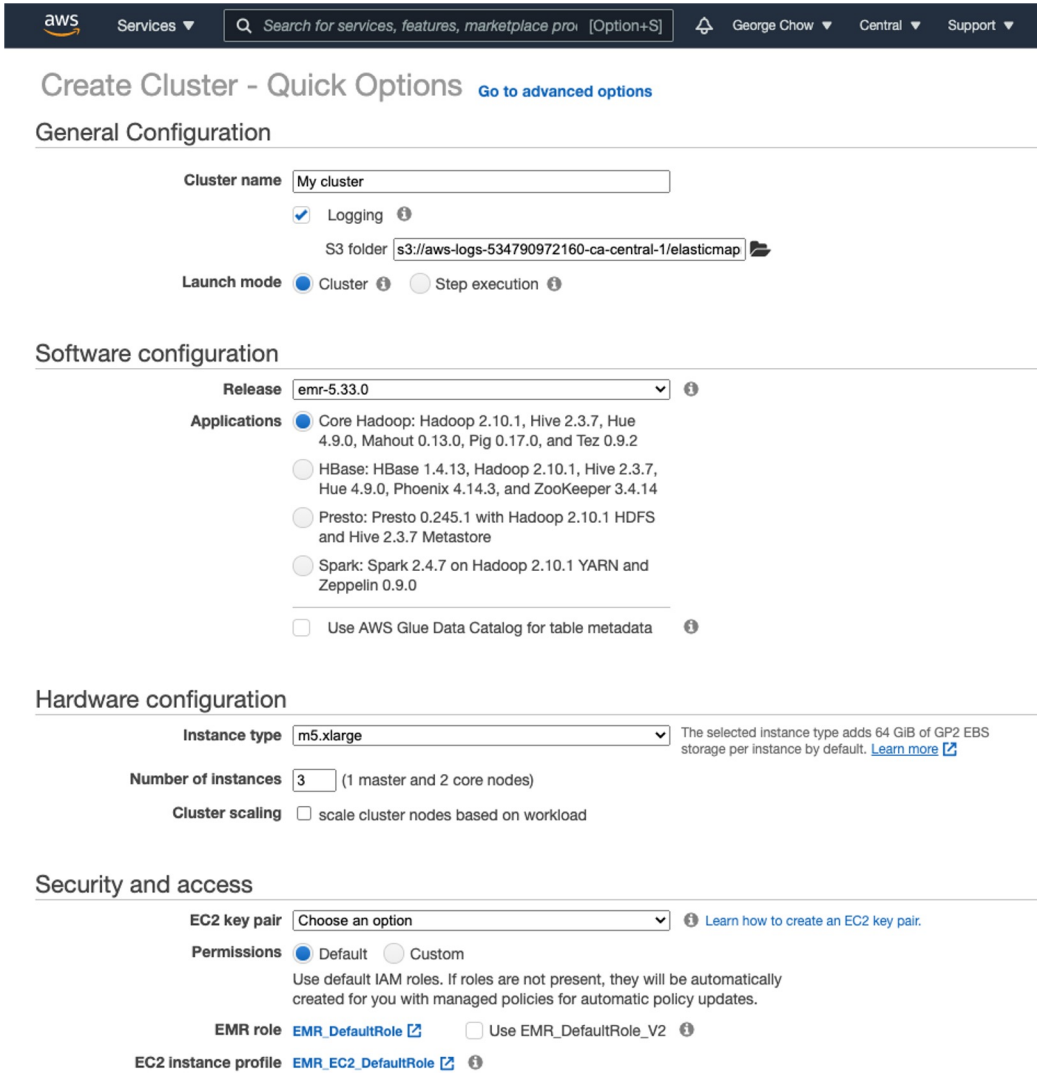
EMR (Elastic MapReduce)

S3 (Simple Storage Service)

EC2 (Elastic Compute Cloud)

EMR

- Not just Hadoop
  - HBase, Presto, Spark, custom…
- Configurable:
  - Nodes
  - Processor
  - Memory
  - Storage
  - Networking
  - Access control



S3

- Replicated/redundant storage (1
  9's!)
- Object-store
- Region-specific
- Granular access control
- Used throughout AWS
- S3 API a standard for storage (EMC, NetApp, etc)
- Metered usage

Search for services, features, marketplace products, and docs [Option+S] Global

Amazon S3 > aws-us-west-2-5[redacted]e

aws-us-west-2-5[redacted]e Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Bucket overview

AWS Region	Amazon Resource Name (ARN)	Creation date
US West (Oregon) us-west-2	arn:aws:s3:::aws-us-west-2-5[redacted]	May 21, 2021, 11:04:47 (UTC-07:00)

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. Learn more

Edit

Bucket Versioning

Enabled

Multi-factor authentication (MFA) delete

An additional layer of security that requires multi-factor authentication for changing Bucket Versioning settings and permanently deleting object versions. To modify MFA delete settings, use the AWS CLI, AWS SDK, or the Amazon S3 REST API. Learn more

Disabled

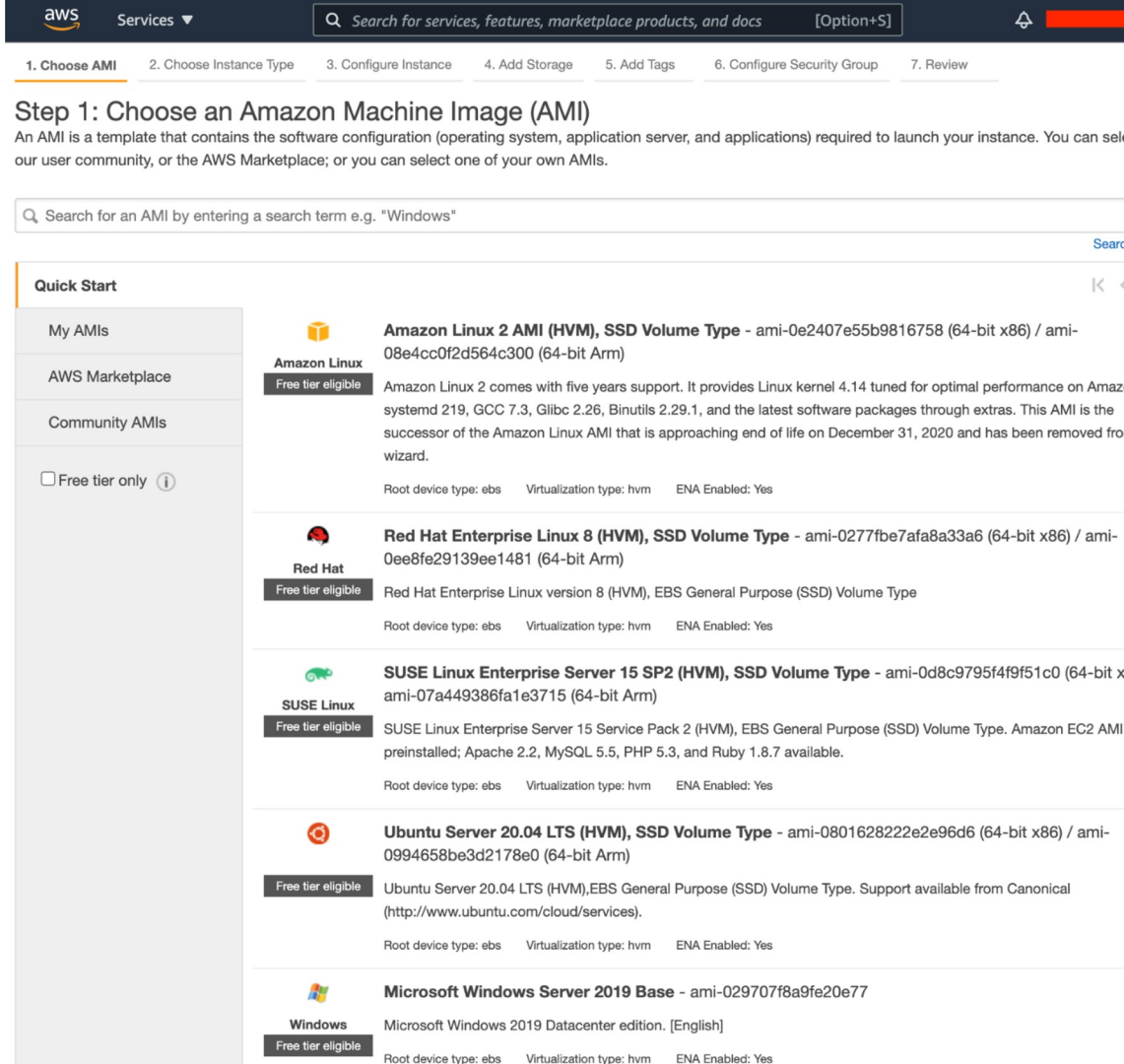
Tags (6)

Edit

Track storage cost or other criteria by tagging your bucket. Learn more

EC2

- “VM on demand”
- Based on AMI images



EC2

- Metered usage
- Configurable:
  - Processor
  - Memory
  - Storage
  - Networking
  - Access control

aws Services ▼ Search for services, features, marketplace products, and docs [Option+S] Cent

1. Choose AMI | 2. Choose Instance Type | 3. Configure Instance | 4. Add Storage | 5. Add Tags | 6. Configure Security Group | 7. Review

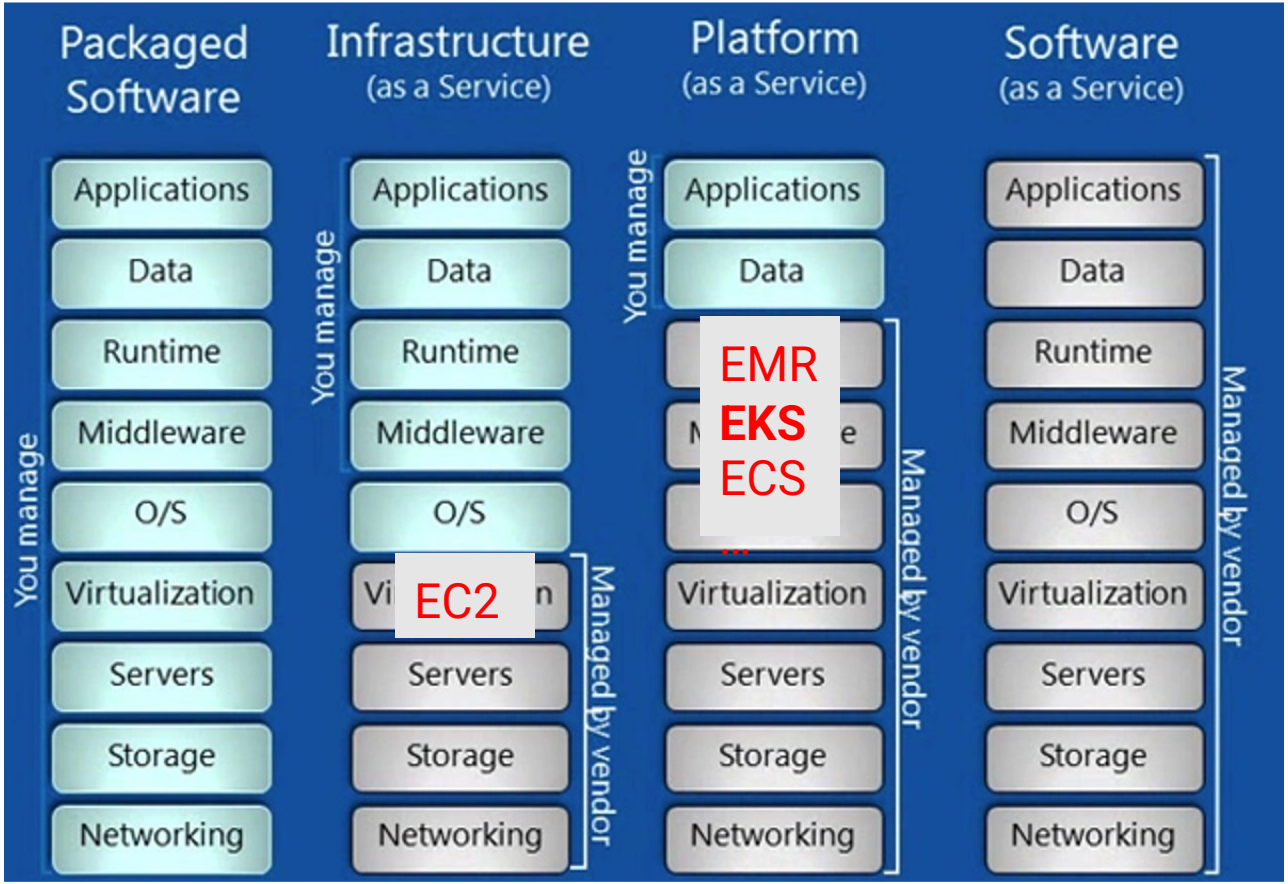
Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combina memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. Learn more about instance types and h your computing needs.

Filter by: All instance families | Current generation | Show/Hide Columns

Currently selected: t2.micro (- ECUs, 1 vCPUs, 2.5 GHz, -, 1 GiB memory, EBS only)

Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performanc
t2	t2.nano	1	0.5	EBS only	-	Low to Moderate
t2	t2.micro Free tier eligible	1	1	EBS only	-	Low to Moderate
t2	t2.small	1	2	EBS only	-	Low to Moderate
t2	t2.medium	2	4	EBS only	-	Low to Moderate
t2	t2.large	2	8	EBS only	-	Low to Moderate
t2	t2.xlarge	4	16	EBS only	-	Moderate
t2	t2.2xlarge	8	32	EBS only	-	Moderate
t3	t3.nano	2	0.5	EBS only	Yes	Up to 5 Gigabit
t3	t3.micro	2	1	EBS only	Yes	Up to 5 Gigabit
t3	t3.small	2	2	EBS only	Yes	Up to 5 Gigabit
t3	t3.medium	2	4	EBS only	Yes	Up to 5 Gigabit
t3	t3.large	2	8	EBS only	Yes	Up to 5 Gigabit
t3	t3.xlarge	4	16	EBS only	Yes	Up to 5 Gigabit
t3	t3.2xlarge	8	32	EBS only	Yes	Up to 5 Gigabit
t3a	t3a.nano	2	0.5	EBS only	Yes	Up to 5 Gigabit



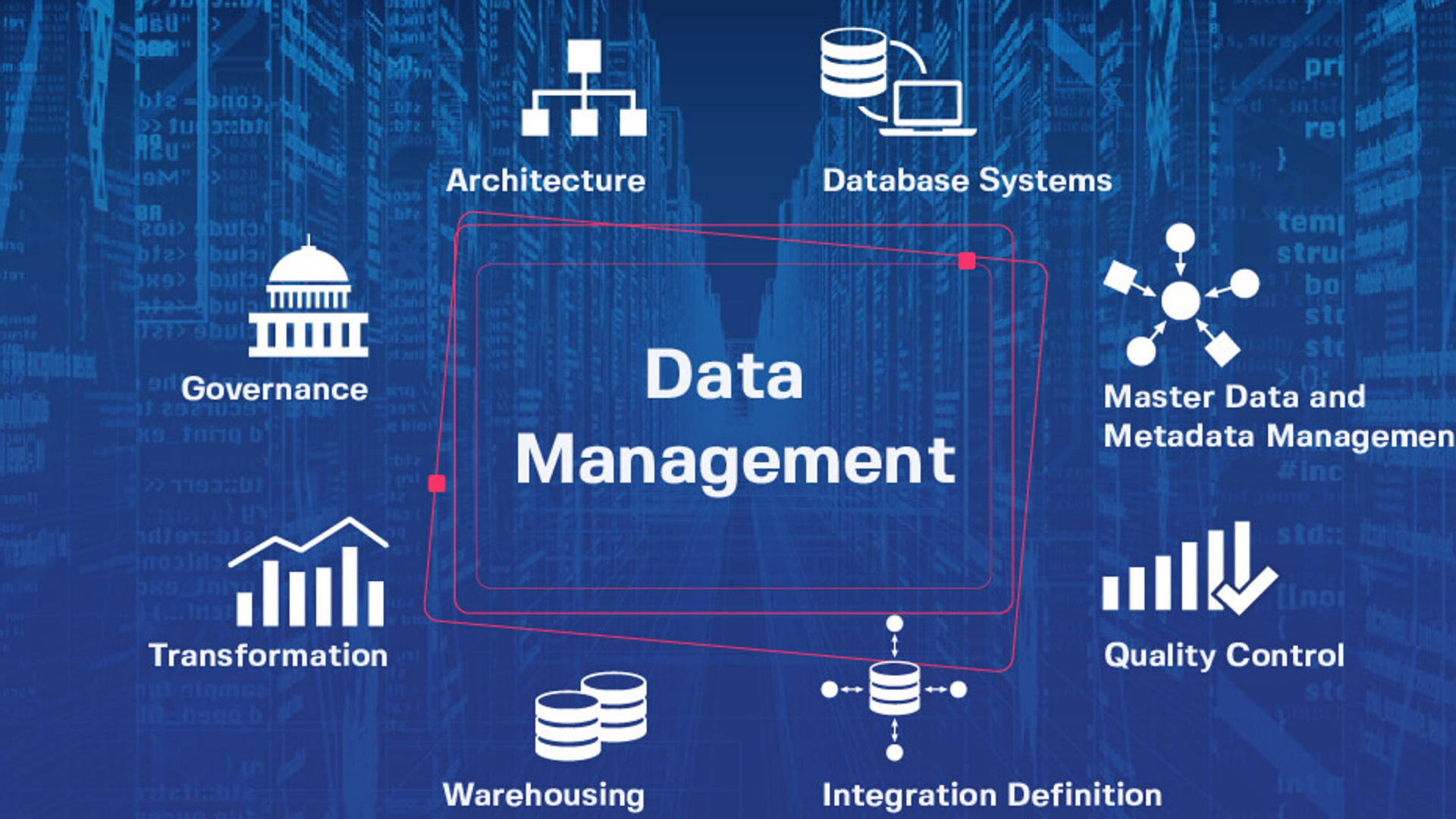
https://venturebeat.com/2011/11/14/cloud-iaas-

Wal-Mart's data center remains mystery

May 28, 2006 8 min to read

Globe File The Wal-Mart Data Center in McDonald County is deemed so secret the county assessor was required to sign a non-disclosure statement before entering the site to determine property value. The photo was taken in 2004, when the center was nearly complete.

The Joplin Globe, Joplin, MO


Architecture
Database Systems
Governance
Data Management
Master Data and Metadata Managemen
Transformation
Quality Control
Warehousing
Integration Definition

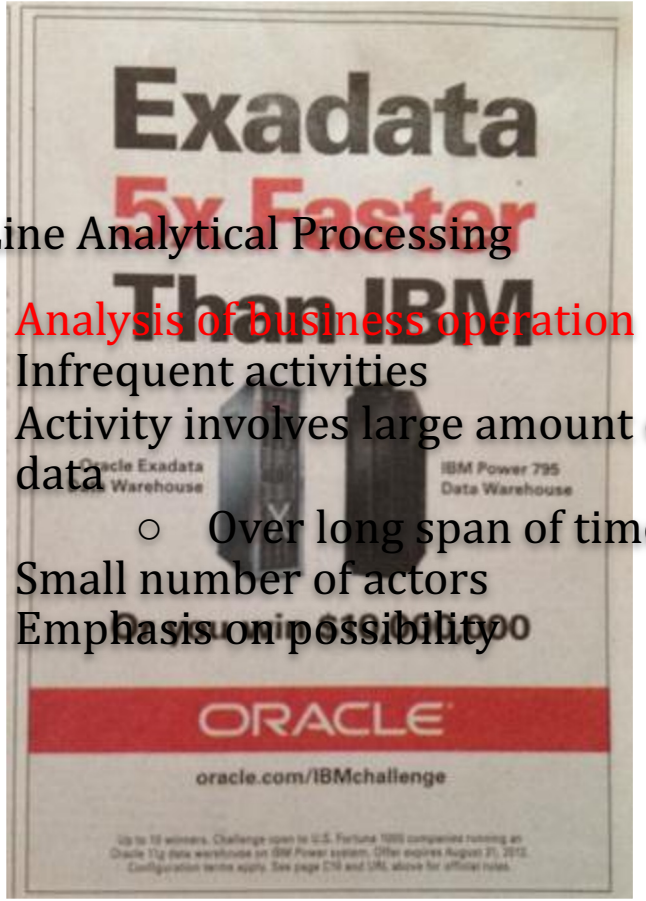
OLTP vs OLAP

OnLine Transaction Processing

- Moment-to-moment business operation
- Continuous streams of activities
- Activity involves small amounts of data
    - Over short span of time
- Large number of actors
- Emphasis on throughput & correctness

OnLine Analytical Processing

- Analysis of business operation
- Infrequent activities
- Activity involves large amount of data
    - Over long span of time
- Small number of actors
- Emphasis on possibility



https://storageioblog.com/who-will-be-winner-with-oracle-10-

~~OLAP~~ Analytics

Descriptive Analytics

Reporting

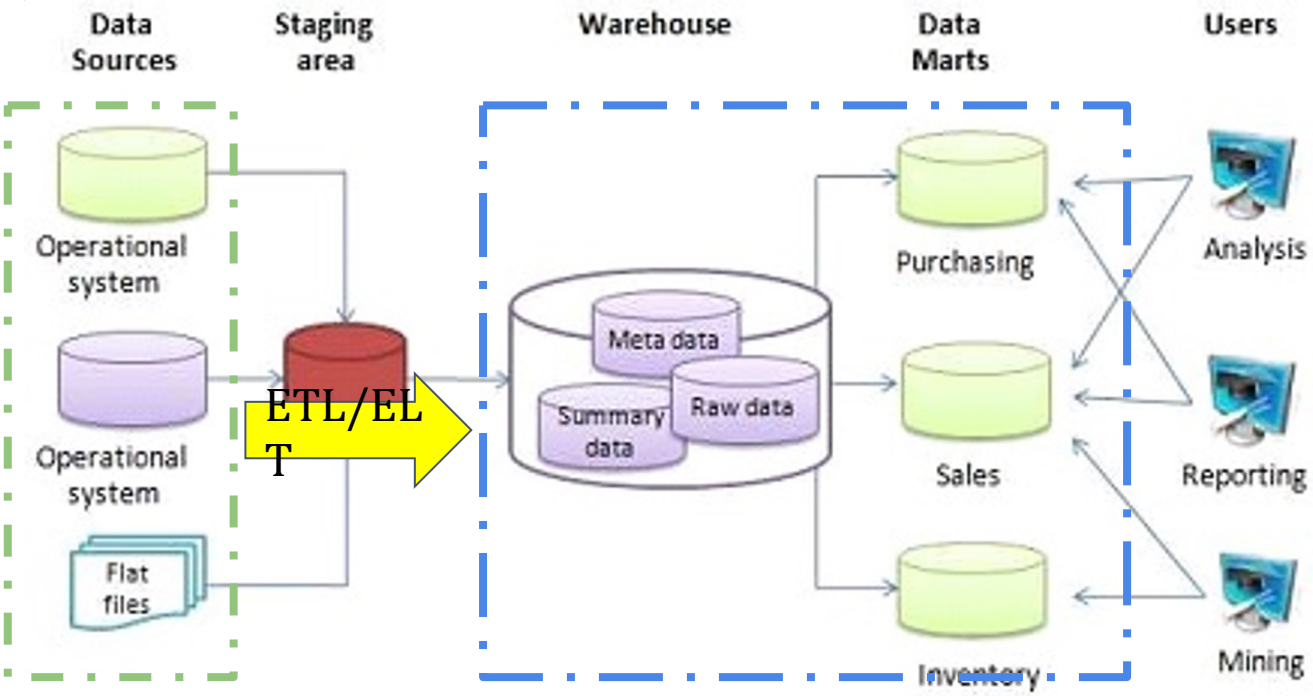
"Business Intelligence"

Predictive Analytics

Recommendation system

"AI" & "ML"

Data Warehouse



https://en.wikipedia.org/wiki/Data_

ETL

Extract-Transform-Load

- Target: data warehouse
- Source: structured data
- Recurring event
  - Extract
  - Transform
  - Load
- or ELT

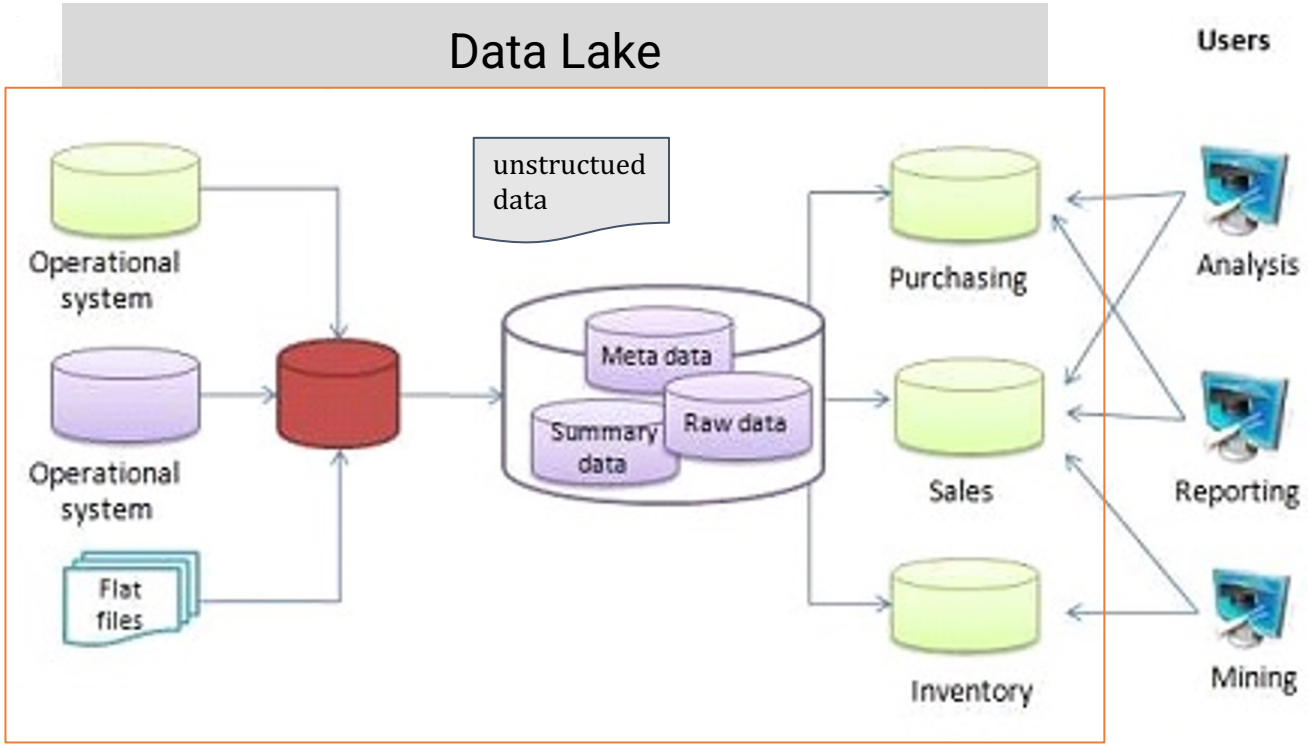
vs Data Wrangling

- Target: anything
- Source: Structured, unstructured, or unmanaged data
- Adhoc event
- “Data science”

Data Management components

- Technology
- Process
- People
- Governance

Data Lake


Data Lake
Users
unstructued data
Operational system
Operational system
Flat files
Meta data
Summary data
Raw data
Purchasing
Sales
Inventory
Analysis
Reporting
Mining

Data Warehouse vs Data Lake

Data Warehouse

- Structured data only (database-based)
- Limited access
- “on-prem”

Data Lake

- Structured ***and*** unstructured data (filesystem-based)
- Broadly accessible
- Tend to be “cloud-based”

Data Lake options

Open Source

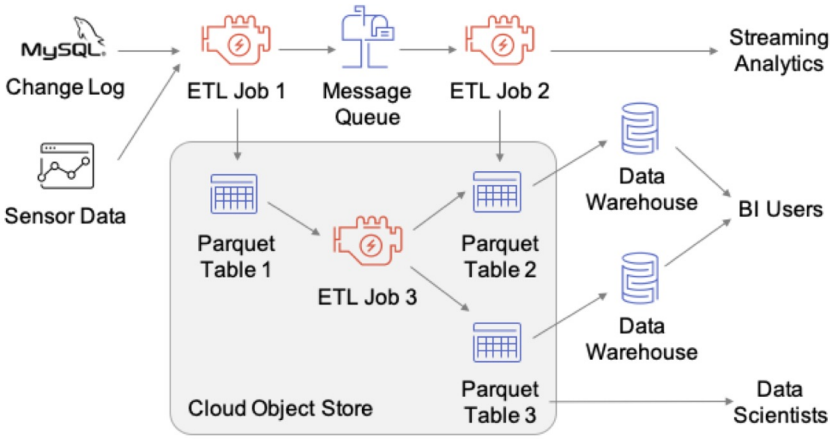
- Delta Lake
  https://delta.io
- Iceberg
  https://iceberg.apache.org

Cloud/Commercial

- Azure Data Lake
- Databricks (Delta Lake)
- Dremio
- Tabular (Iceberg)

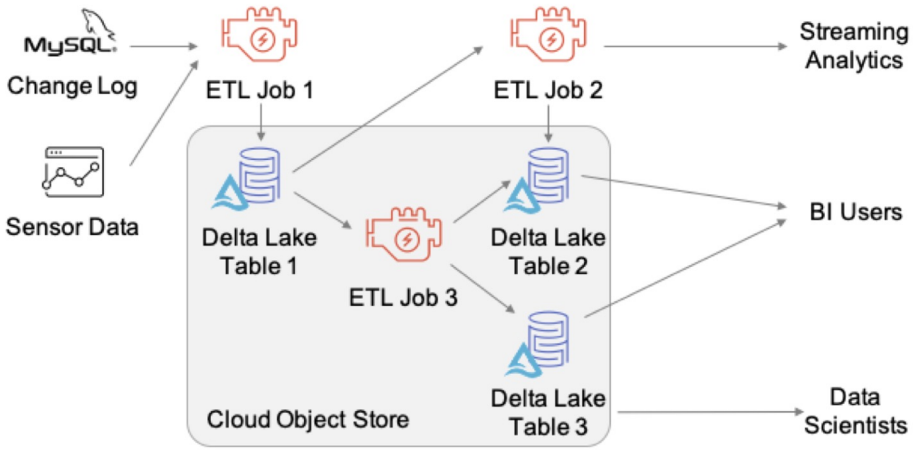
Example: Delta Lake

Data Warehouse



(a) Pipeline using separate storage systems.

Delta Lake



(b) Using Delta Lake for both stream and table storage.

https://delta.i

Agenda

- Hadoop and Spark recap ✅
- Cloud Computing ✅
- Data Management practices ✅