

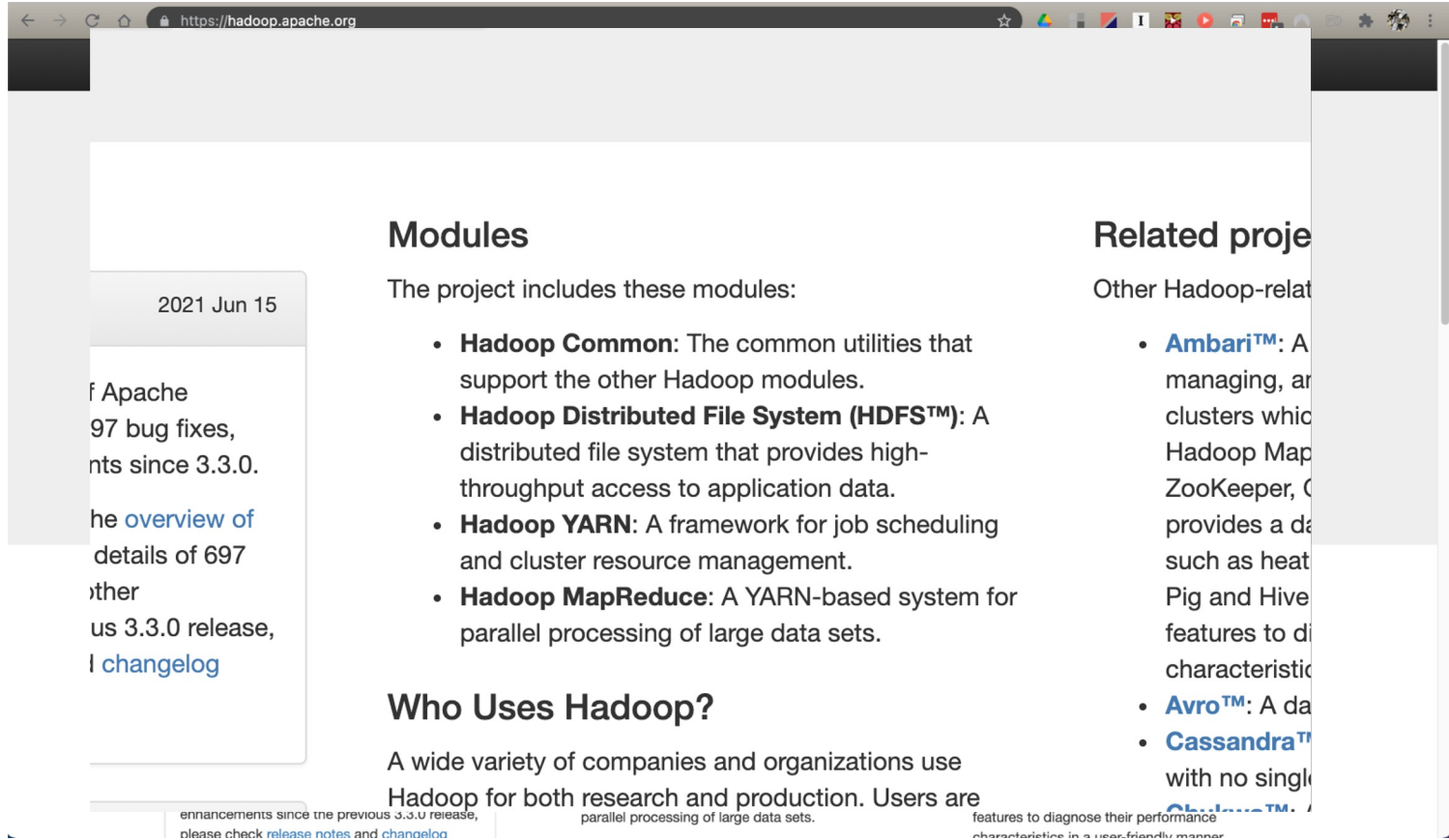
Cloud & Data Management

CMPT 732 - Fall 2023

Agenda

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

Hadoop: overview



The screenshot shows the Apache Hadoop website. The browser address bar displays <https://hadoop.apache.org>. The page content includes a navigation sidebar on the left, a main content area with sections for 'Modules' and 'Who Uses Hadoop?', and a 'Related projects' section on the right. The 'Modules' section lists Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce. The 'Who Uses Hadoop?' section states that a wide variety of companies and organizations use Hadoop for both research and production. The 'Related projects' section lists Ambari, Avro, and Cassandra.

2021 Jun 15

of Apache
97 bug fixes,
nts since 3.3.0.

he [overview](#) of
details of 697
ther
us 3.3.0 release,
l [changelog](#)

Modules

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

Who Uses Hadoop?

A wide variety of companies and organizations use Hadoop for both research and production. Users are

Related projects

Other Hadoop-related projects include:

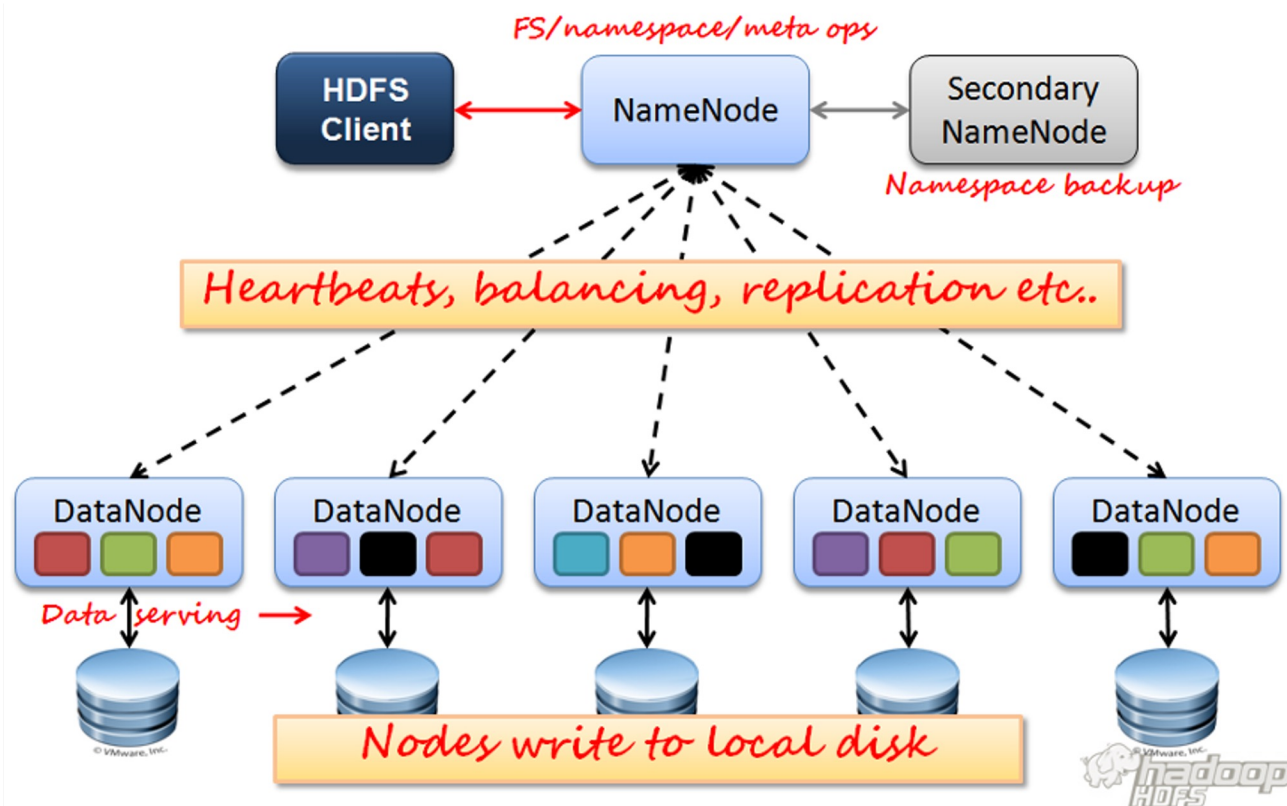
- **Ambari™:** A tool for managing, monitoring, and configuring Hadoop clusters which integrates with Hadoop MapReduce, ZooKeeper, and other services. It provides a dashboard for monitoring and managing clusters such as heat maps and alerts.
- **Avro™:** A data interchange format that provides a rich set of features to diagnose their performance characteristics in a user-friendly manner.
- **Cassandra™:** A distributed database with no single point of failure.

enhancements since the previous 3.3.0 release, please check [release notes](#) and [changelog](#).

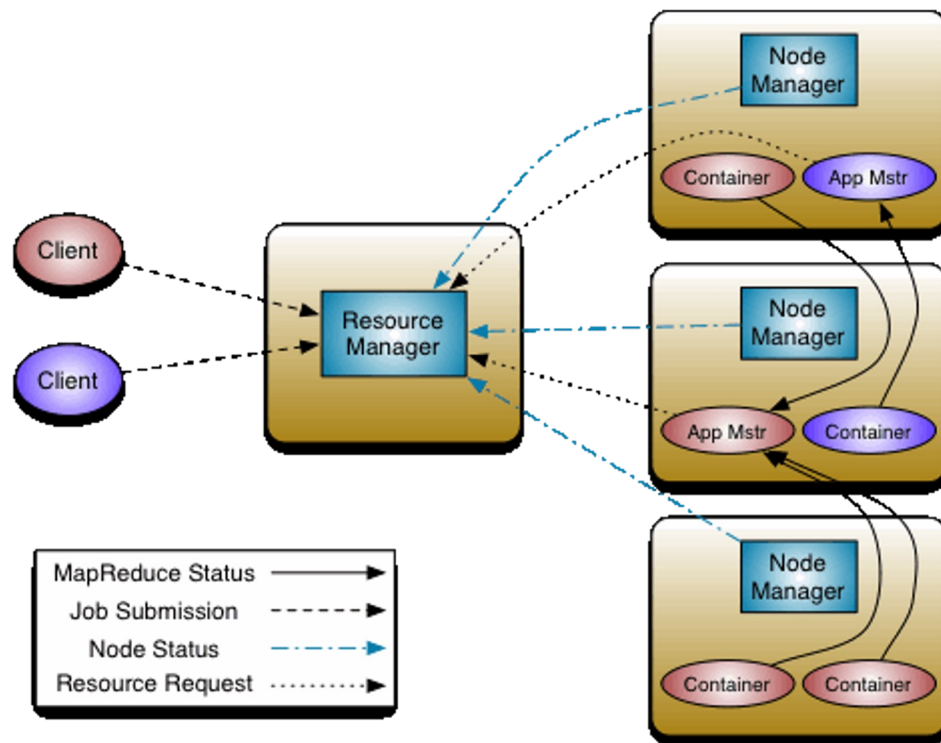
parallel processing of large data sets.

features to diagnose their performance characteristics in a user-friendly manner.

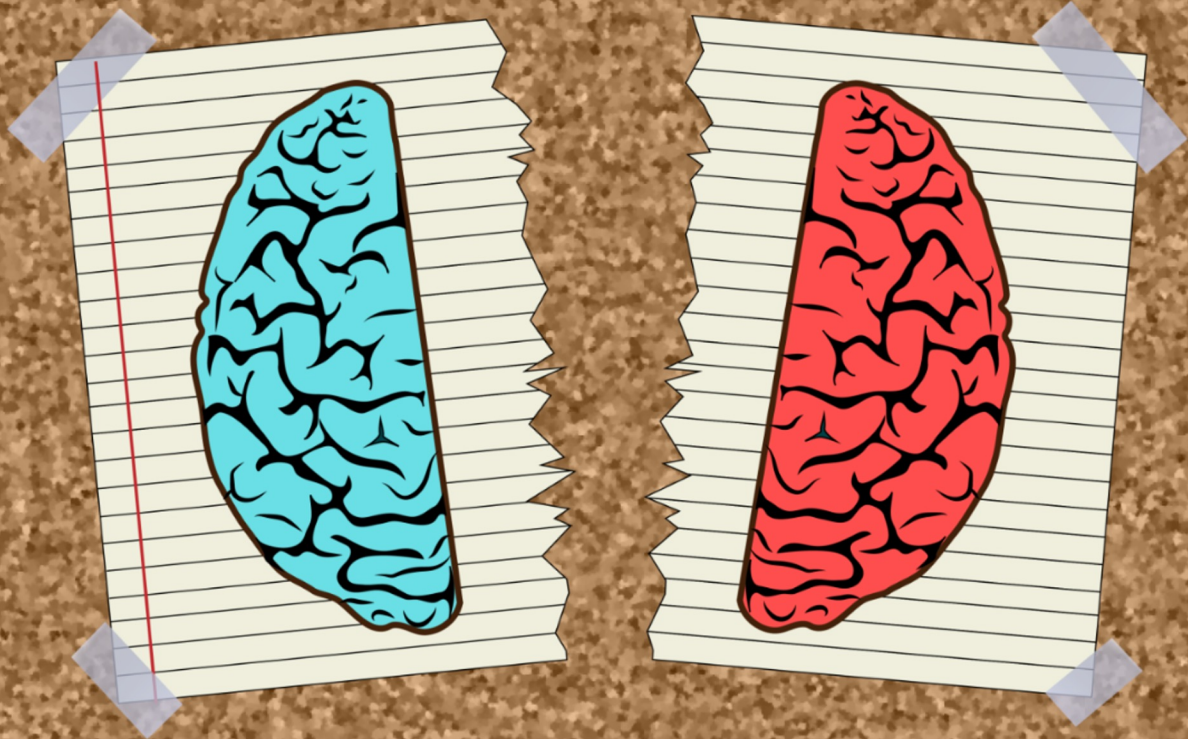
Hadoop: storage



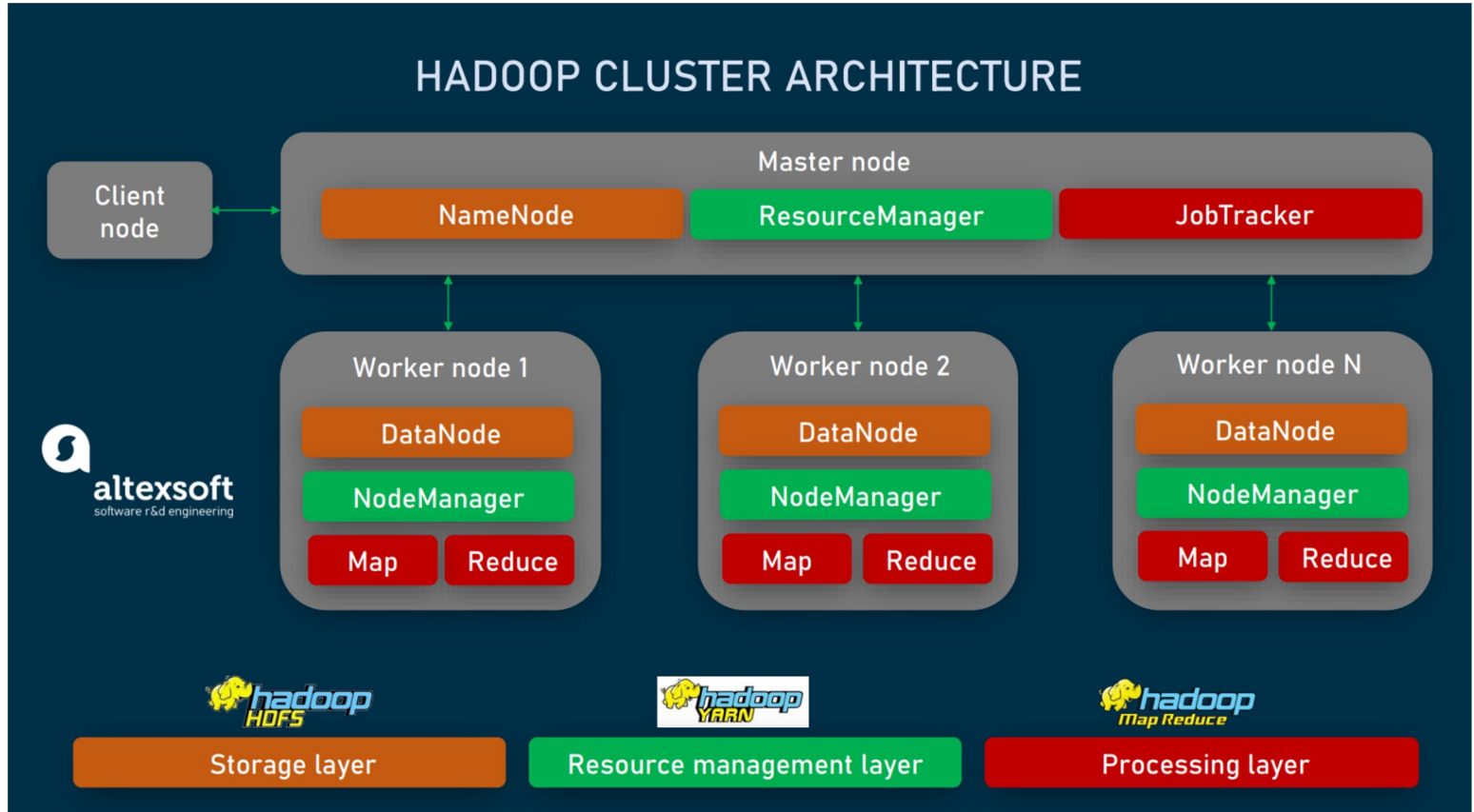
Hadoop: compute



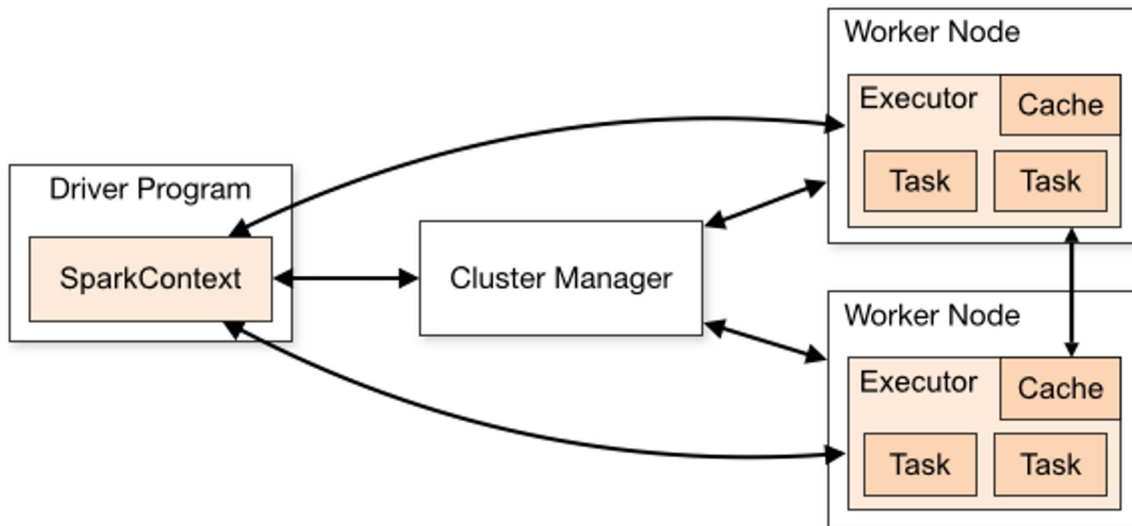
<https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn->



Hadoop: together



Spark cluster architecture





MORGAN & CLAYPOOL PUBLISHERS

The Datacenter as a Computer

Designing Warehouse-Scale Machines

Third Edition



Luiz André Barroso
Urs Hölzle
Parthasarathy Ranganathan

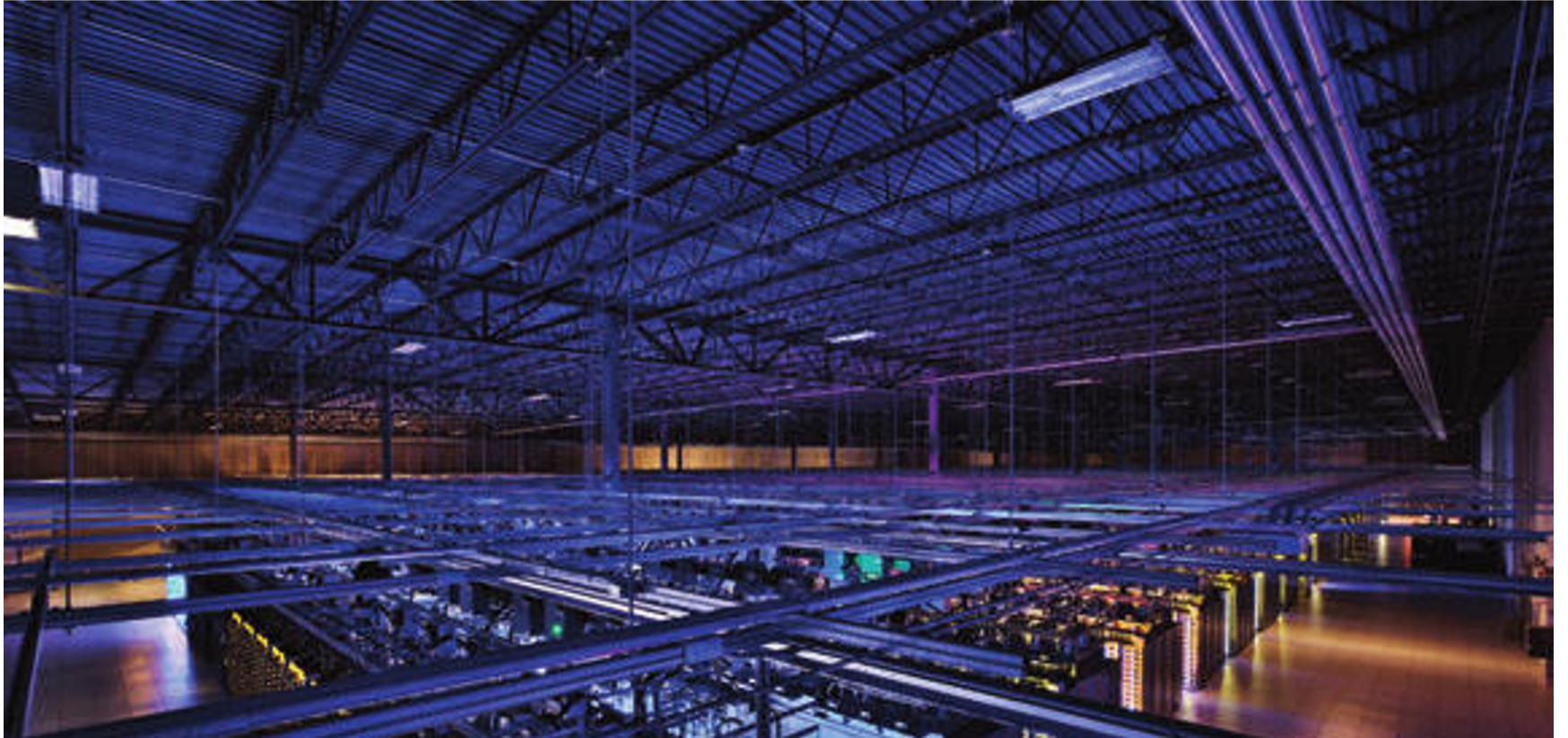
*SYNTHESIS LECTURES ON
COMPUTER ARCHITECTURE*

Margaret Martonosi, *Series Editor*

Datacenter



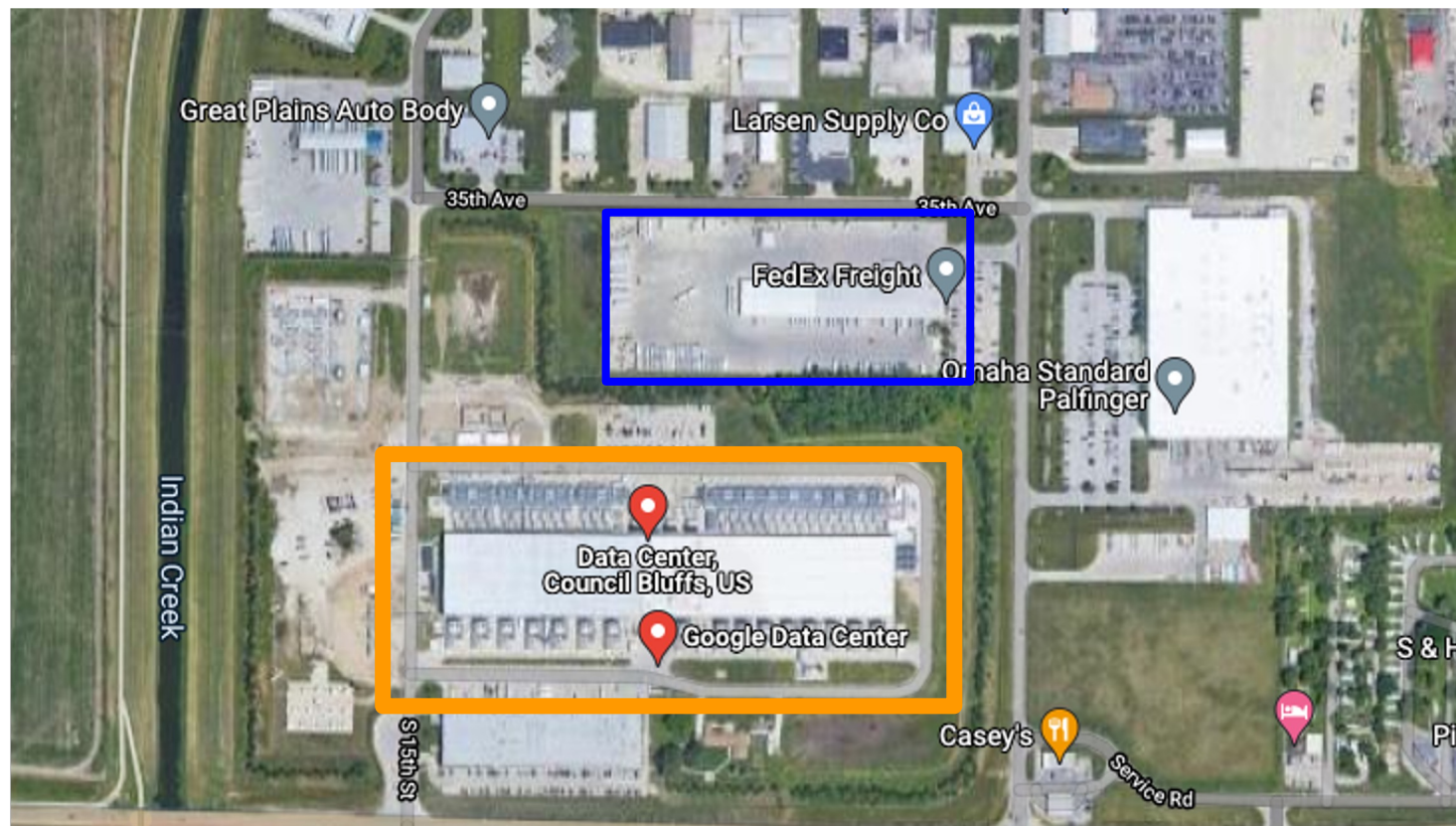
Datacenter: Power distribution



Datacenter: Cooling



The scale of a datacenter





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)

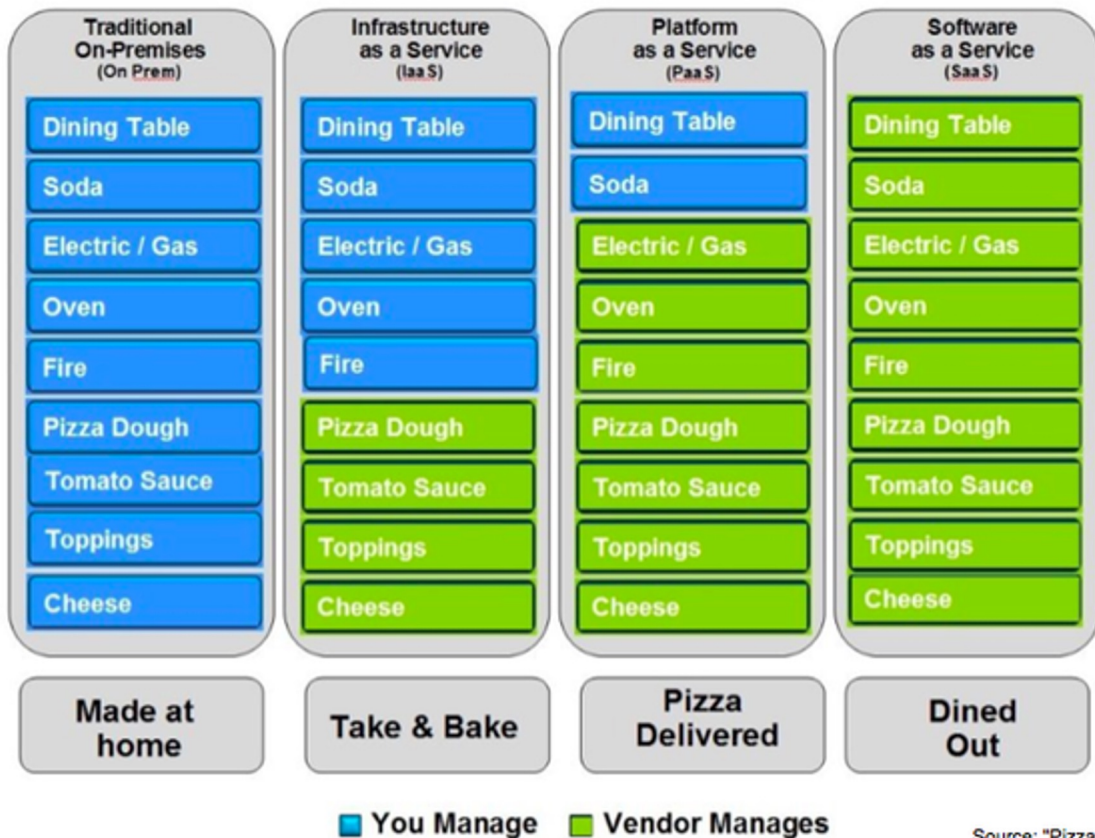
PaaS removes the need for you to manage underlying infrastructure (usually hardware and operating systems), and allows you to focus on the deployment and management of your applications. This helps you be more efficient as you don't need to worry about resource procurement, capacity planning, software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



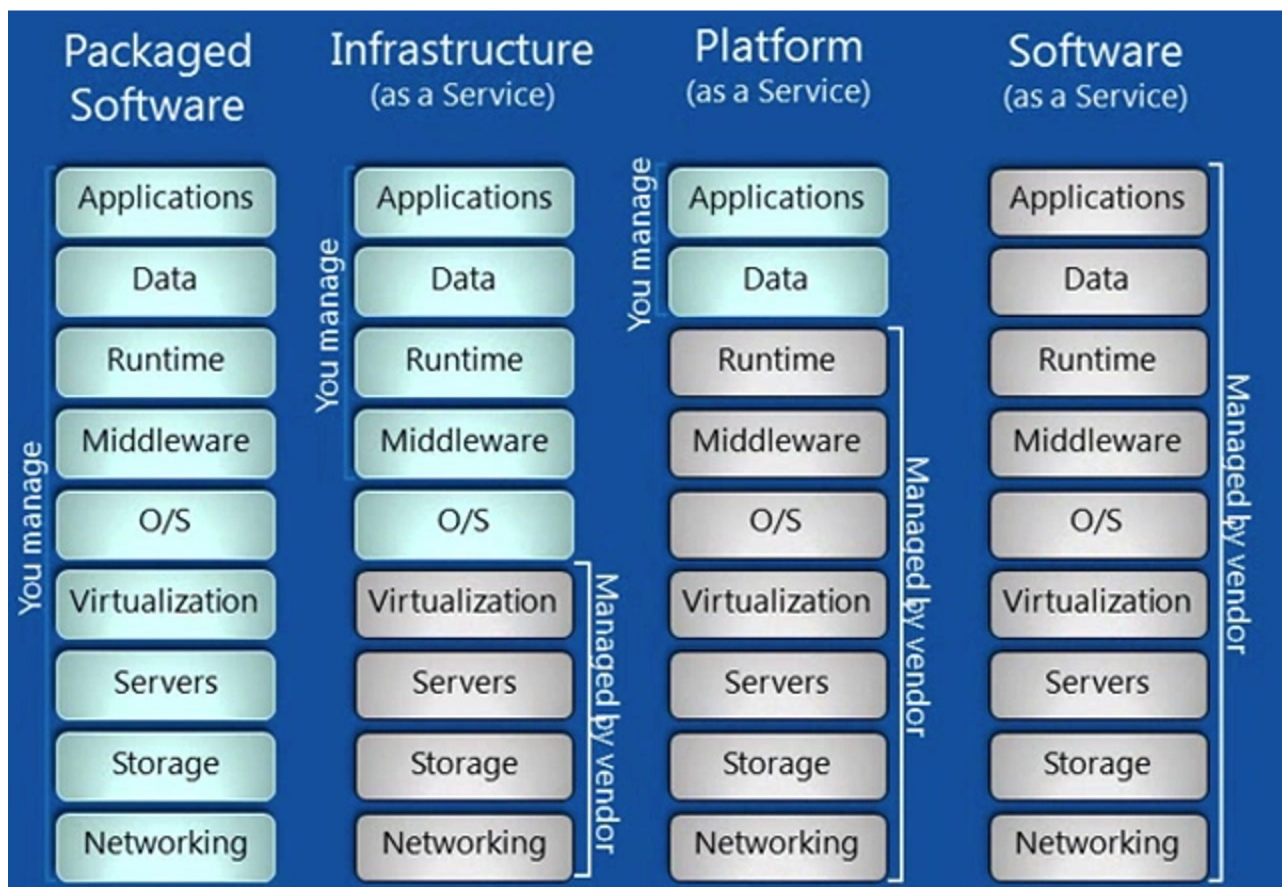
Software as a Service (SaaS)

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"

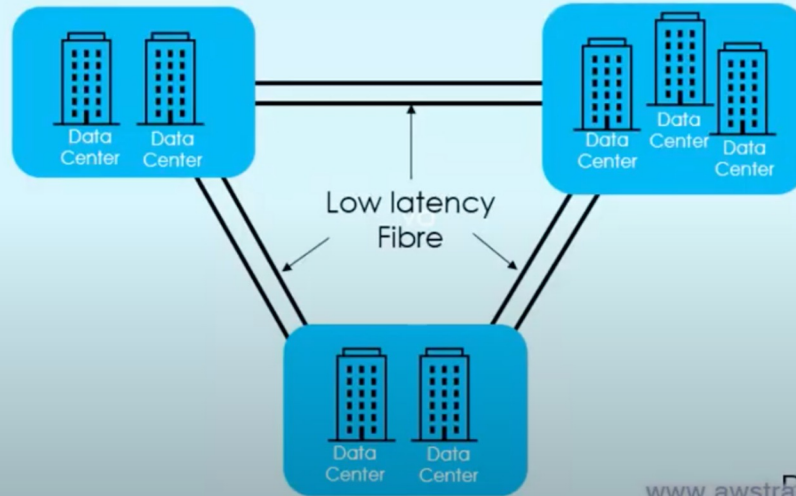


Region & AZ

1 Region = Multiple AZs (Min 3)
1 AZ = Cluster of Data centres

EC2

Getting
Started



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

The screenshot shows the AWS EMR console interface for creating a cluster. At the top, there's a navigation bar with the AWS logo, 'Services', a search bar, and user information. The main heading is 'Create Cluster - Quick Options' with a link to 'Go to advanced options'. Below this is the 'General Configuration' section, which includes a text input for 'Cluster name' (set to 'My cluster'), a checked 'Logging' option, an 'S3 folder' dropdown (set to 's3://aws-logs-534790972160-ca-central-1/elasticmap'), and 'Launch mode' radio buttons (selected as 'Cluster'). The 'Software configuration' section features a 'Release' dropdown (set to 'emr-5.33.0') and several 'Applications' radio button options: 'Core Hadoop' (selected), 'HBase', 'Presto', and 'Spark'. There is also an unchecked checkbox for 'Use AWS Glue Data Catalog for table metadata'. The 'Hardware configuration' section shows an 'Instance type' dropdown (set to 'm5.xlarge'), a note about storage, a 'Number of instances' input (set to 3), and an unchecked 'Cluster scaling' checkbox. The 'Security and access' section includes an 'EC2 key pair' dropdown (set to 'Choose an option'), 'Permissions' radio buttons (selected as 'Default'), and 'EMR role' and 'EC2 instance profile' dropdowns (both set to their default roles).

aws Services [Option+S] George Chow Central Support

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

Logging [?](#)

S3 folder

Launch mode Cluster [?](#) Step execution [?](#)

Software configuration

Release [?](#)

Applications Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore

Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

Use AWS Glue Data Catalog for table metadata [?](#)

Hardware configuration

Instance type [?](#) The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#)

Number of instances (1 master and 2 core nodes)

Cluster scaling scale cluster nodes based on workload

Security and access

EC2 key pair [?](#) [Learn how to create an EC2 key pair.](#)

Permissions Default Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) [?](#) Use EMR_DefaultRole_V2 [?](#)

EC2 instance profile [EMR_EC2_DefaultRole](#) [?](#)

S3

- Replicated/redundant storage (12 9's!)
- Object-store
- Region-specific
- Granular access control
- Used throughout AWS
- S3 API a standard for storage (EMC, NetApp, etc)
- Metered usage

The screenshot shows the AWS Management Console interface for an Amazon S3 bucket. At the top, there is a search bar and a navigation menu. The breadcrumb trail indicates the location: Amazon S3 > aws-us-west-2-5 [redacted] > aws-us-west-2-5 [redacted] > Info. The main content area has several tabs: Objects, Properties (selected), Permissions, Metrics, Management, and Access Points. Below the tabs is the 'Bucket overview' section, which displays key information in a table-like format:

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

Below the overview is the 'Bucket Versioning' section, which includes a description of versioning and an 'Edit' button. The status is 'Enabled'. Underneath, there is a section for 'Multi-factor authentication (MFA) delete', which is currently 'Disabled'. At the bottom, there is a 'Tags (6)' section with an 'Edit' button and a link to learn more about tagging.

EC2

- “VM on demand”
- Based on AMI images

aws Services [Option+S]

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


Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select from our user community, or the AWS Marketplace; or you can select one of your own AMIs.

Search for an AMI by entering a search term e.g. "Windows"


Quick Start

- My AMIs
- AWS Marketplace
- Community AMIs
- Free tier only ⓘ

 **Amazon Linux 2 AMI (HVM), SSD Volume Type** - ami-0e2407e55b9816758 (64-bit x86) / ami-08e4cc0f2d564c300 (64-bit Arm)


Free tier eligible Amazon Linux 2 comes with five years support. It provides Linux kernel 4.14 tuned for optimal performance on systemd 219, GCC 7.3, Glibc 2.26, Binutils 2.29.1, and the latest software packages through extras. This AMI is the successor of the Amazon Linux AMI that is approaching end of life on December 31, 2020 and has been removed from the wizard.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

 **Red Hat Enterprise Linux 8 (HVM), SSD Volume Type** - ami-0277fbe7afa8a33a6 (64-bit x86) / ami-0ee8fe29139ee1481 (64-bit Arm)


Free tier eligible Red Hat Enterprise Linux version 8 (HVM), EBS General Purpose (SSD) Volume Type

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

 **SUSE Linux Enterprise Server 15 SP2 (HVM), SSD Volume Type** - ami-0d8c9795f4f9f51c0 (64-bit x86) / ami-07a449386fa1e3715 (64-bit Arm)


Free tier eligible SUSE Linux Enterprise Server 15 Service Pack 2 (HVM), EBS General Purpose (SSD) Volume Type. Amazon EC2 AMI preinstalled; Apache 2.2, MySQL 5.5, PHP 5.3, and Ruby 1.8.7 available.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

 **Ubuntu Server 20.04 LTS (HVM), SSD Volume Type** - ami-0801628222e2e96d6 (64-bit x86) / ami-0994658be3d2178e0 (64-bit Arm)

Free tier eligible Ubuntu Server 20.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

 **Microsoft Windows Server 2019 Base** - ami-029707f8a9fe20e77

Free tier eligible Microsoft Windows 2019 Datacenter edition. [English]

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

EC2

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

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

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

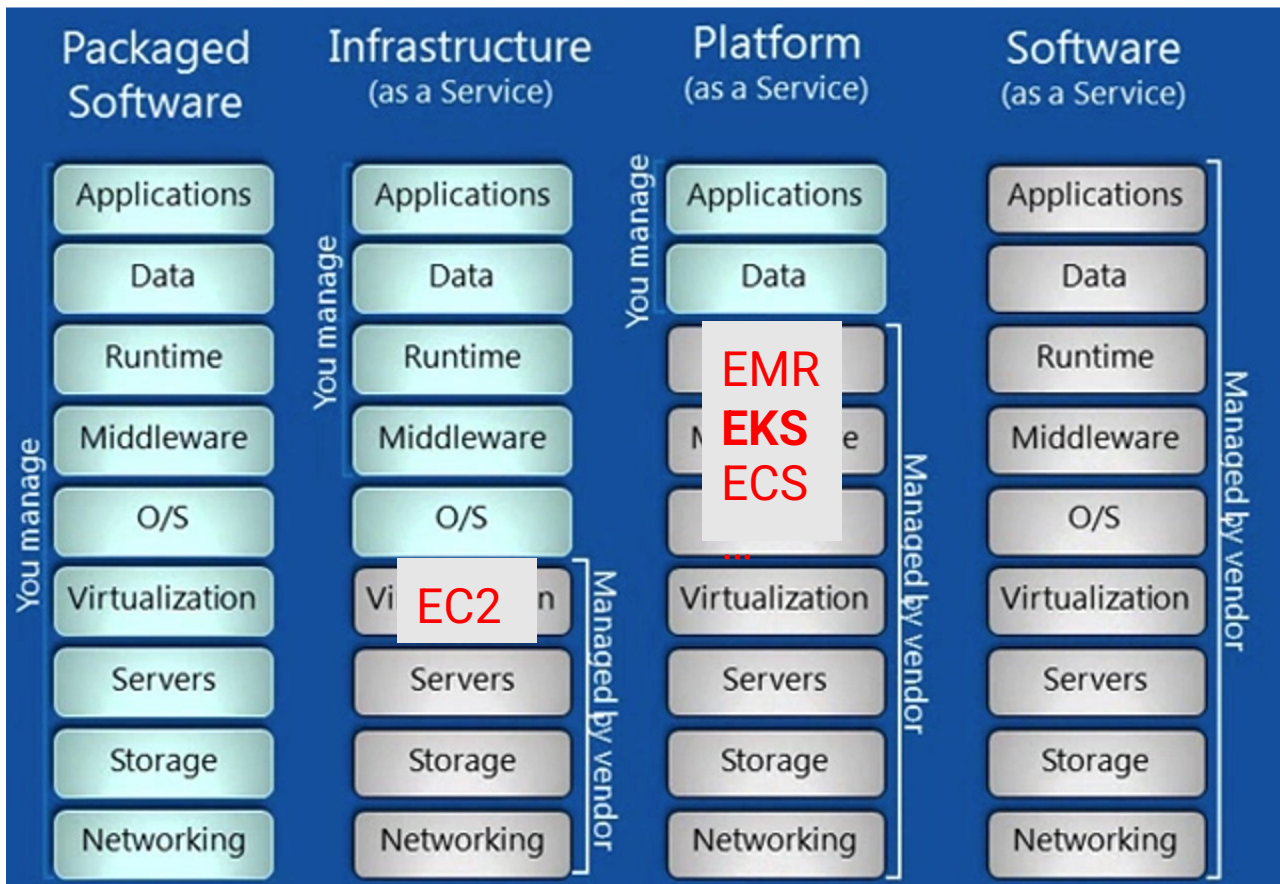
Step 2: Choose an Instance Type

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 combinations of vCPUs, 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 how to choose the right instance type for your computing needs.

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

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

	Family ▾	Type ▾	vCPUs ⓘ ▾	Memory (GiB) ▾	Instance Storage (GB) ⓘ ▾	EBS-Optimized Available ⓘ ▾	Network Performance ⓘ
<input type="checkbox"/>	t2	t2.nano	1	0.5	EBS only	-	Low to Moderate
<input checked="" type="checkbox"/>	t2	t2.micro <small>Free tier eligible</small>	1	1	EBS only	-	Low to Moderate
<input type="checkbox"/>	t2	<u>t2.small</u>	1	2	EBS only	-	Low to Moderate
<input type="checkbox"/>	t2	<u>t2.medium</u>	2	4	EBS only	-	Low to Moderate
<input type="checkbox"/>	t2	<u>t2.large</u>	2	8	EBS only	-	Low to Moderate
<input type="checkbox"/>	t2	t2.xlarge	4	16	EBS only	-	Moderate
<input type="checkbox"/>	t2	t2.2xlarge	8	32	EBS only	-	Moderate
<input type="checkbox"/>	t3	t3.nano	2	0.5	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.micro	2	1	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.small	2	2	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.medium	2	4	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.large	2	8	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.xlarge	4	16	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3	t3.2xlarge	8	32	EBS only	Yes	Up to 5 Gigabit
<input type="checkbox"/>	t3a	t3a.nano	2	0.5	EBS only	Yes	Up to 5 Gigabit





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 Management**



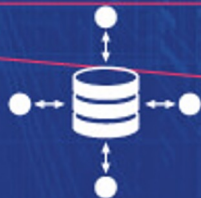
Transformation



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

The image is a vertical advertisement for a data processing challenge between Oracle and IBM. At the top, the text reads "Exadata 5x Faster Than IBM" in large, bold letters. Below this, there are two server racks: an Oracle Exadata Data Warehouse on the left and an IBM Power 795 Data Warehouse on the right. A large "X" is superimposed over the racks. Below the racks, the text says "Challenge: 100 million records in 10 minutes". At the bottom, the Oracle logo is displayed in a red bar, followed by the URL "oracle.com/IBMchallenge".

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

~~OLAP~~ Analytics

Descriptive Analytics

Reporting

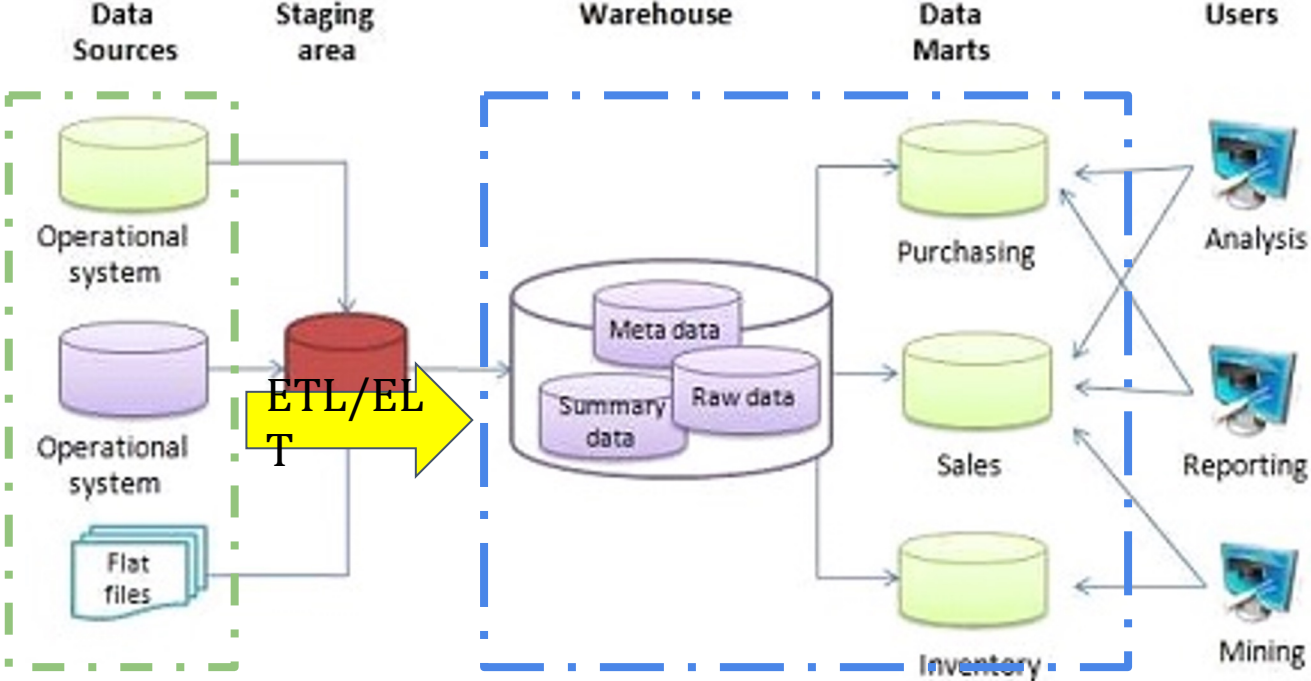
“Business Intelligence”

Predictive Analytics

Recommendation system

“AI” & “ML”

Data Warehouse



ETL

Extract-Transform-Load

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

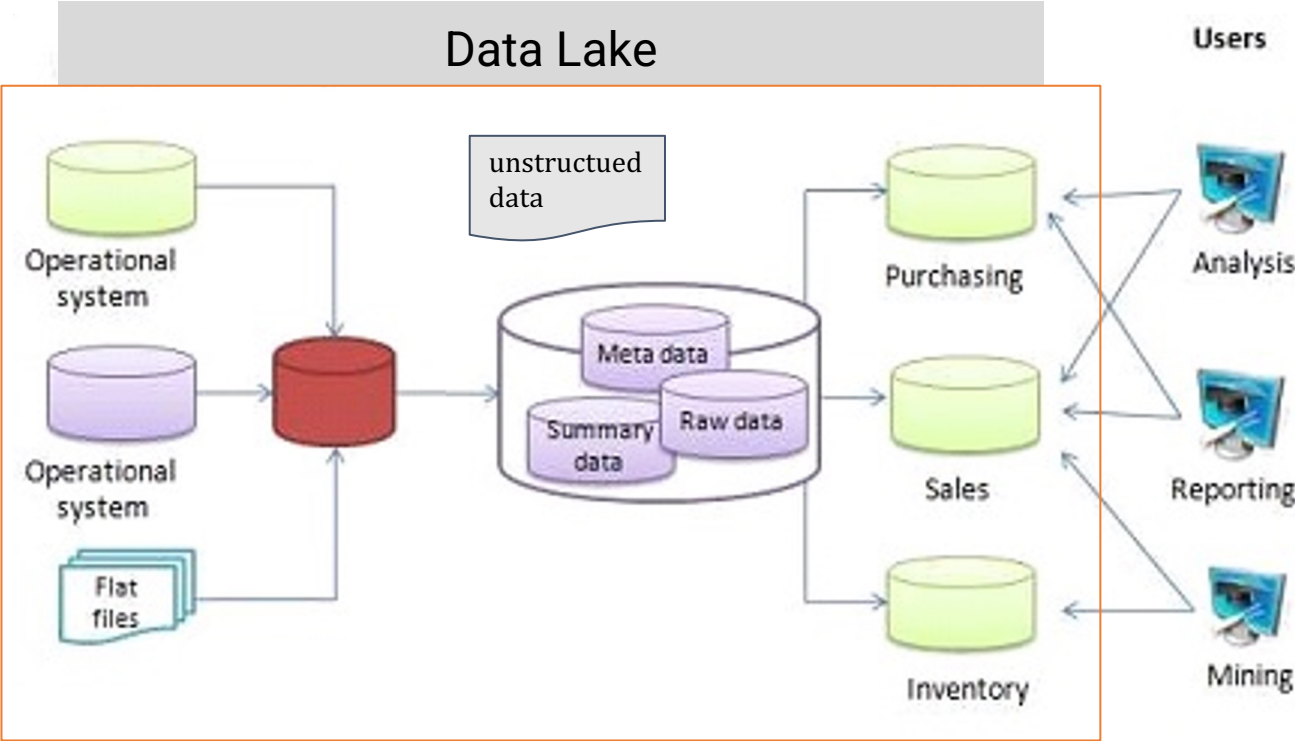
vs Data Wrangling

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

Data Management components

- Technology
- Process
- People
- Governance

Data Lake



Data 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

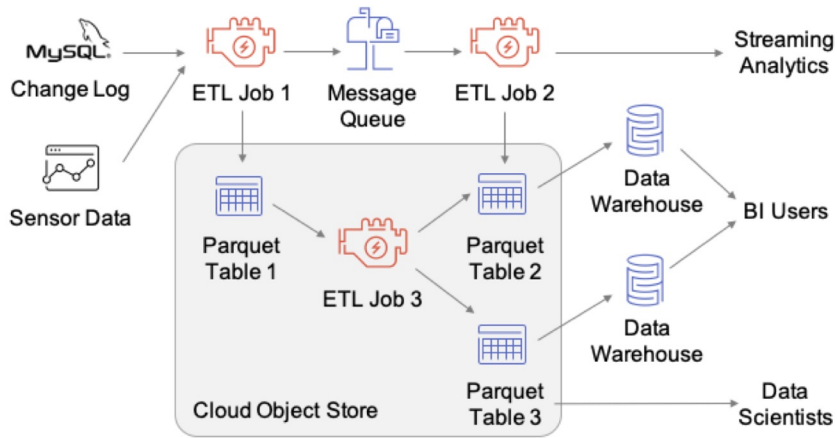
- Delta Lake
<https://delta.io>
- Iceberg
<https://iceberg.apache.org>

Cloud/Commercial

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

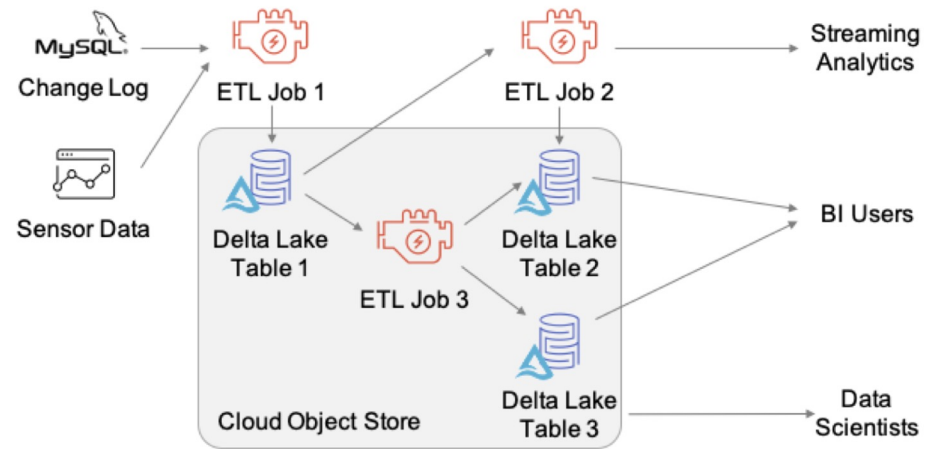
Example: Delta Lake

Data Warehouse





(a) Pipeline using separate storage systems.

Delta Lake



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

Agenda

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