An Introduction to Data Intensive Computing

Amir H. Payberah
amir@sics.se

KTH Royal Institute of Technology

Course Information

Course Objective

- Introduction to main concepts and principles of cloud computing and data intensive computing.
- How to read, review and present a scientific paper.

Topics of Study

- Topics we will cover include:
  - How to store big data?
  - How to process big data?
  - How to manage cluster resources?

The Course Material

- Mainly based on research papers.
- You will find all the material on the course web page:
  http://www.sics.se/~amir/id2221

The Course Grading

- Four lab assignments: 20%
- Six reading assignments: 30%
- The final presentation: 10%
- The final exam: 40%

The Lab Assignments

- Four lab assignments.
- Java and Scala programming.
- The IPython/Jupyter notebooks.
- Students will work in groups of two.
- `https://github.com/payberah/id2221`

The Reading Assignments

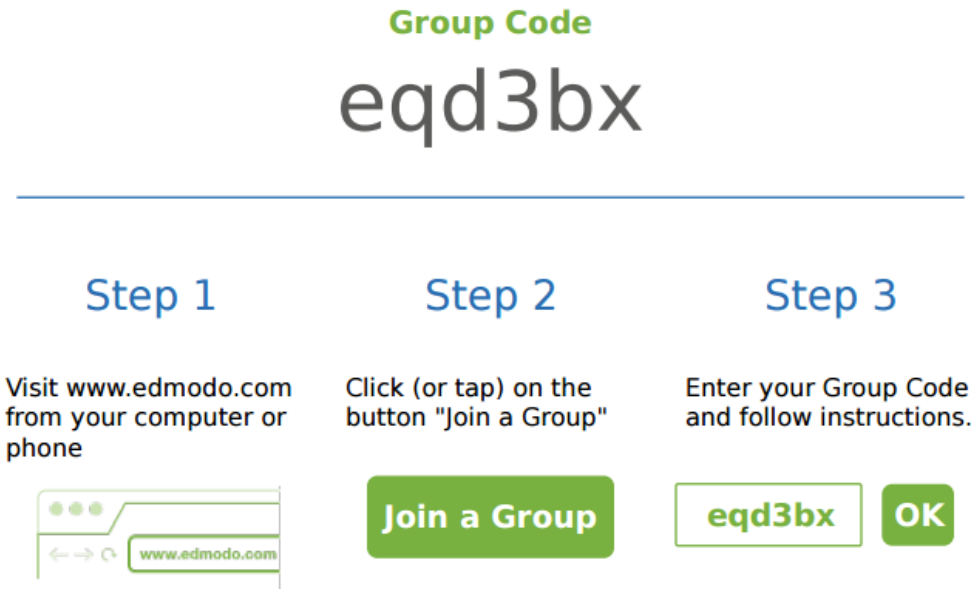
- Six reading assignments.
- Write a review for each paper (at most two pages).
- For each paper you should identify, the motivation, the contribution, the solution, and positive/negative aspects of the solution/paper.
- Students will work in groups of two.

The Final Presentation

- Each group give a 20 minutes talk on a scientific paper.
- The list of papers will be available in the course web page.
- You are also free to choose any other paper, but it should be confirmed.

How to Submit the Assignments?

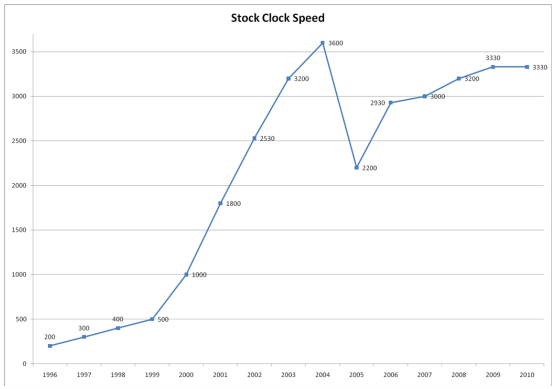
- Through the Edmodo site.
- You need to join the course group first.

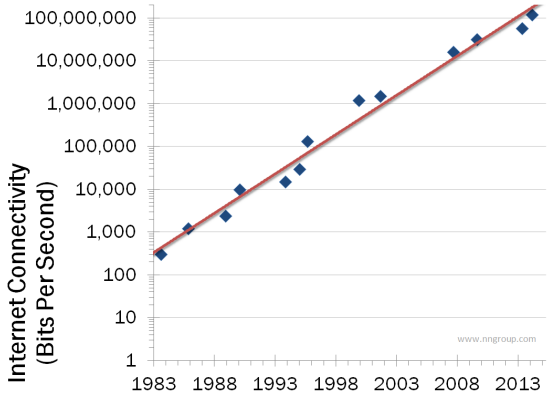


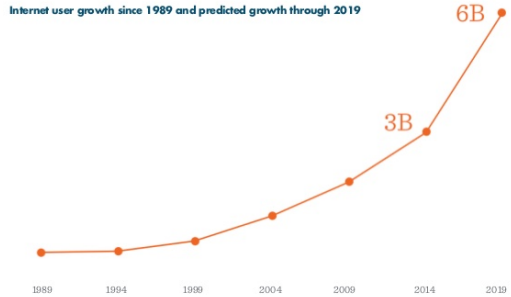
The Course Overview

Cloud Computing and Big Data

- The main trends:
  - Computers not getting any faster
  - Internet connections getting faster
  - More people connected to the Internet







Cloud Computing and Big Data

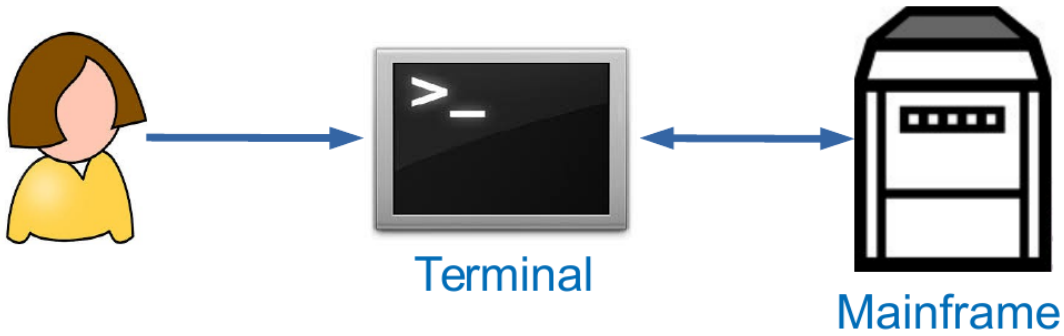
Conclusion

Move the computation and storage of big data to the cloud!

Cloud Computing

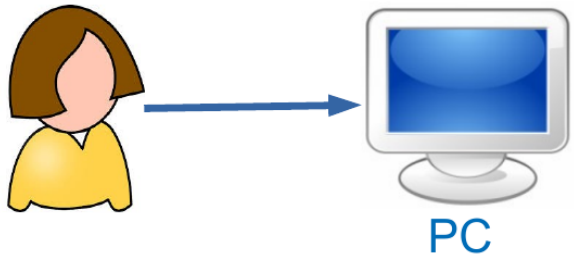
Computing Paradigms - Phase 1

- Many users shared powerful mainframes using dummy terminals.



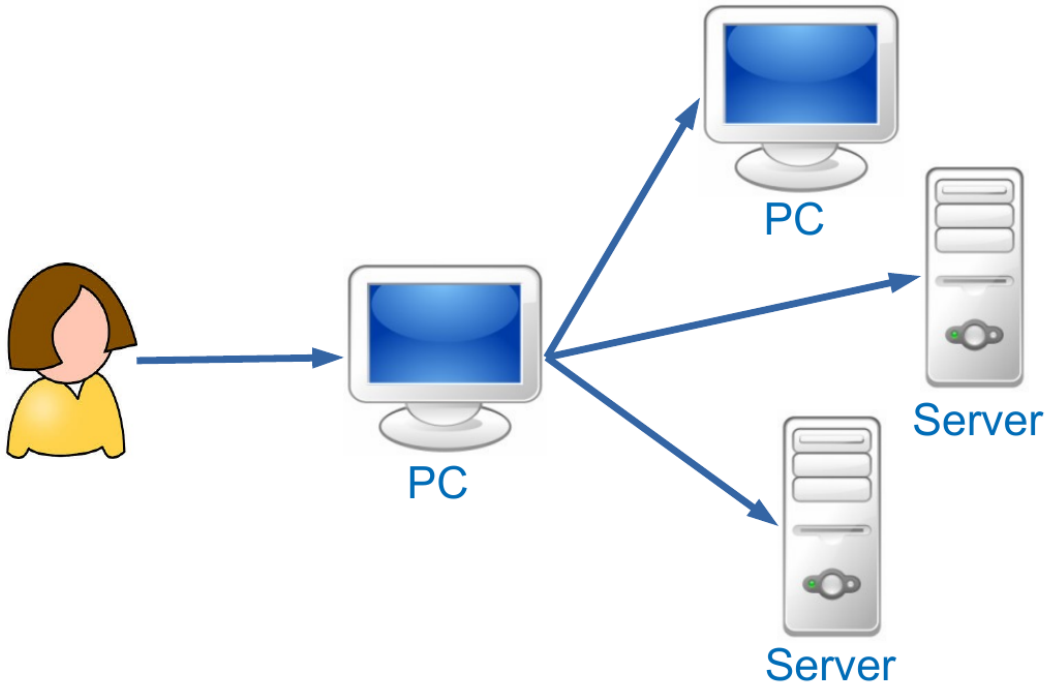
Computing Paradigms - Phase 2

- Stand-alone PCs.



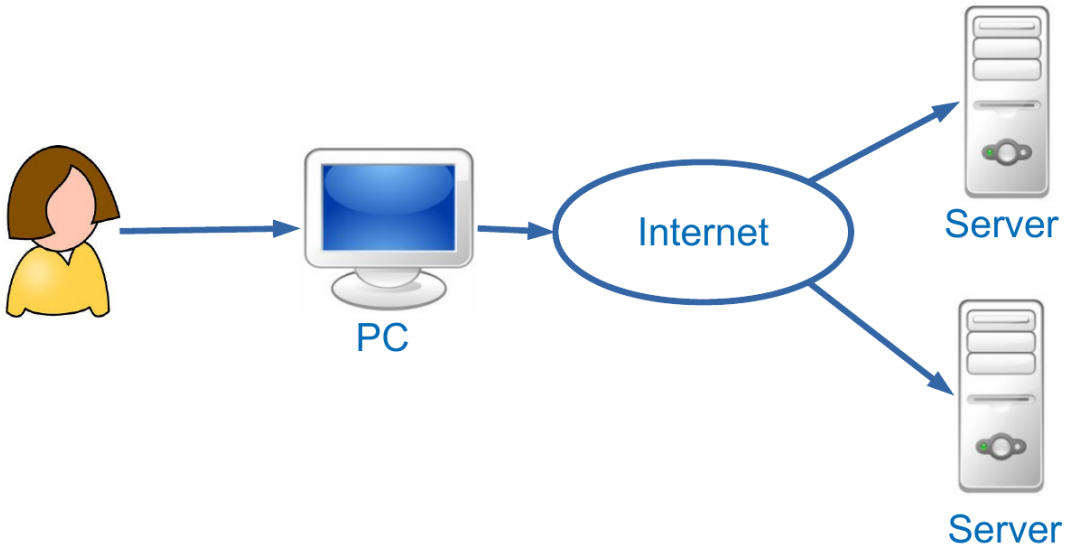
Computing Paradigms - Phase 3

- PCs, laptops, and servers were connected together through local networks.



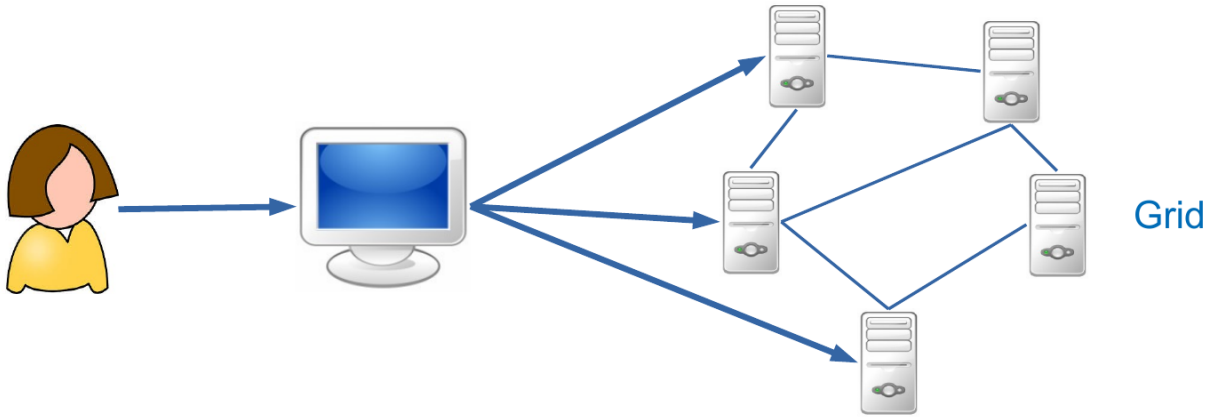
Computing Paradigms - Phase 4

- The Internet: a global network of local networks.



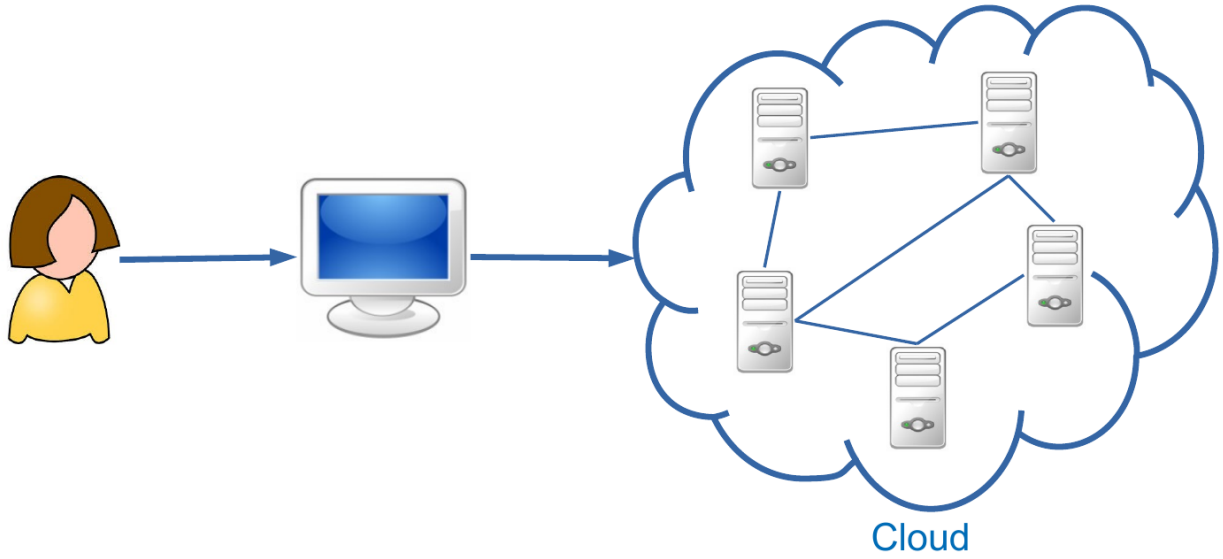
Computing Paradigms - Phase 5

- Grid computing: shared computing power and storage through a distributed computing system.



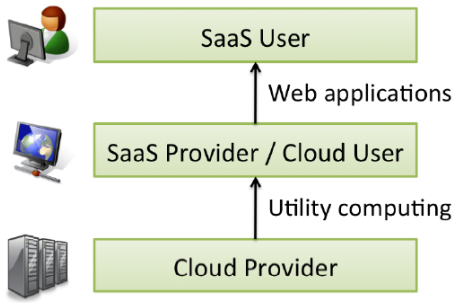
Computing Paradigms - Phase 6

- Cloud computing: shared resources on the Internet in a scalable and simple way.



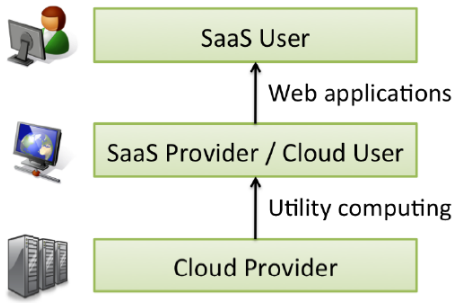
Cloud Computing Definition

- Cloud Computing refers to both:



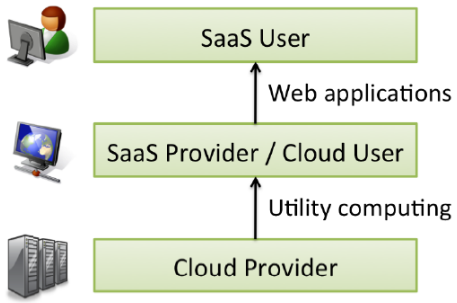
Cloud Computing Definition

- Cloud Computing refers to both:
  1. the applications delivered as services over the Internet



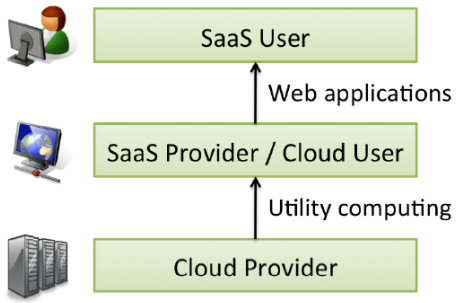
Cloud Computing Definition

- Cloud Computing refers to both:
  1. the applications delivered as services over the Internet
  2. the hardware and systems software in the datacenters that provide those services.



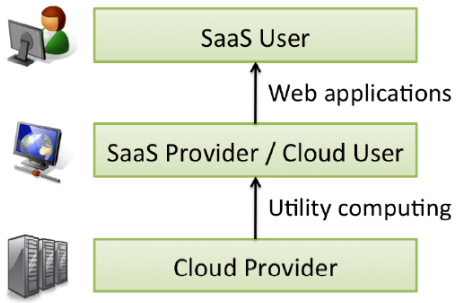
Cloud Computing Definition

- Cloud Computing refers to both:
  1. the applications delivered as services over the Internet
  2. the hardware and systems software in the datacenters that provide those services.
- The services: called Software as a Service (SaaS).



Cloud Computing Definition

- Cloud Computing refers to both:
  1. the applications delivered as services over the Internet
  2. the hardware and systems software in the datacenters that provide those services.
- The services: called Software as a Service (SaaS).
- The datacenter hardware and software: called Cloud



- The NIST definition:
  - Five characteristics
  - Three service models
  - Four deployment models

Cloud Characteristics

Cloud Characteristics

On-demand self-service

Ubiquitous network access

Location transparent resource pooling

Rapid elasticity

Measured service with pay per use

[http://aka.ms/532]

Cloud Characteristics - On-demand Self-Service

- A consumer can unilaterally provision computing capabilities without human interaction with the service provider.

On-demand
self-service

Cloud Characteristics - Ubiquitous Network Access

- Available over the network.
- Accessed through mobile phones, laptops, ...

Ubiquitous network access

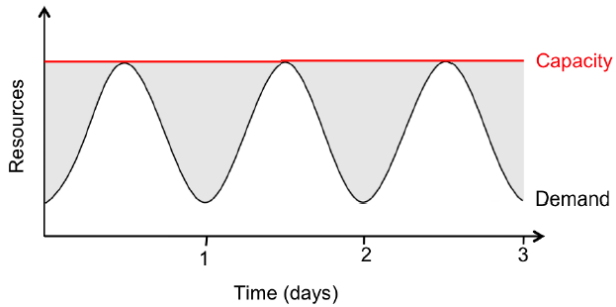
Cloud Characteristics - Resource Pooling

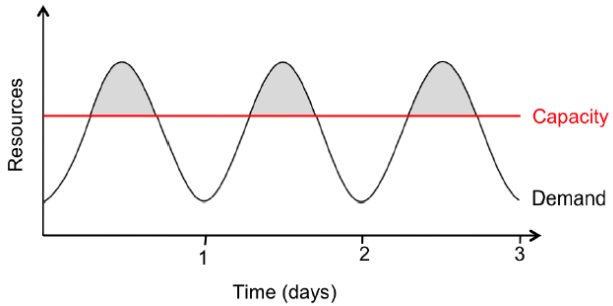
- Provider's computing resources are pooled to serve consumers.
- Location transparent

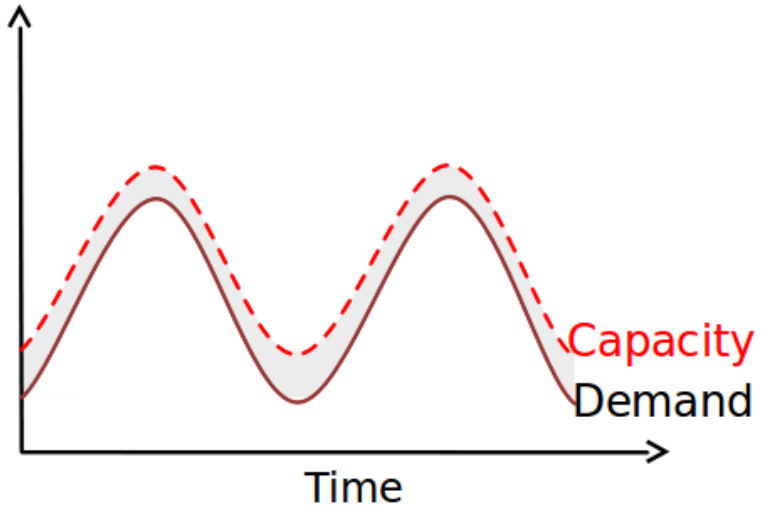
Location transparent resource pooling

Cloud Characteristics - Rapid Elasticity

- Capabilities can be rapidly and elastically provisioned, in some cases automatically.







Rapid elasticity

Cloud Characteristics - Measured Service

- Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer.

Measured service with pay per use

Cloud Service Models

Cloud Service Models

SaaS

PaaS

IaaS

[http://aka.ms/532]

- Assume, you just moved to a city and you are looking for a place to live.

- What is your choice?

- What is your choice?
  - Built a new house?

- What is your choice?
  - Built a new house?
  - Buy an empty house?

- What is your choice?
  - Built a new house?
  - Buy an empty house?
  - Live in a hotel?

- Let's built a new house!

- Let's built a new house!
- You can fully control everything your like your new house to have.
- But that is a hard work.

- What if you buy an empty house?

- What if you buy an empty house?
- You can customize some part of your house.
- But never change the original architecture.

- How about live in a hotel?

- How about live in a hotel?
- Live in a hotel will be a good idea if the only thing you care is enjoy your life.
- There is nothing you can do with the house except living in it.

Let's translate it to Cloud Computing

Service Models

- Infrastructure as a Service (IaaS): similar to build a new house.
- Platform as a Service (PaaS): similar to buy an empty house.
- Software as a Service (SaaS): similar to live in a hotel.

IaaS - (1/2)

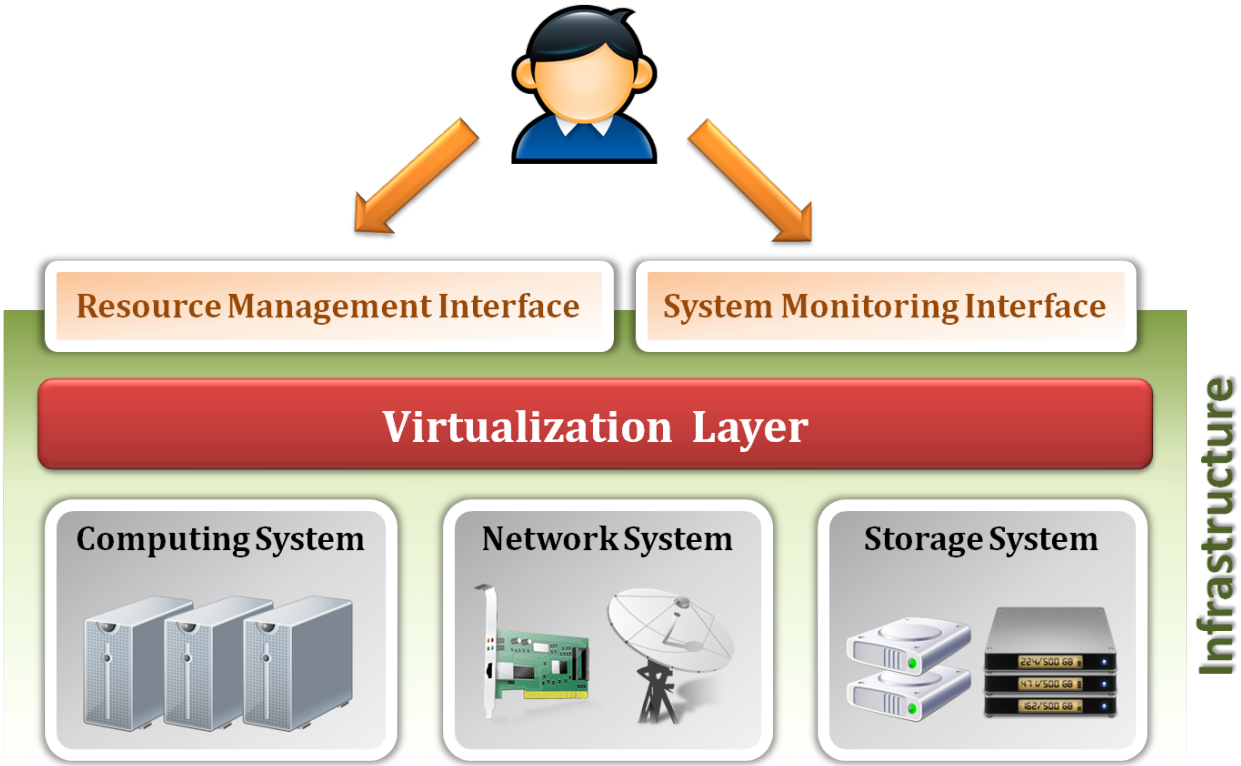
- Vendor provides resources, e.g., processing, storage, network, ...
- Consumer is provided customized virtual machines.
- Consumer has control over the resources.

IaaS - (1/2)

- Vendor provides resources, e.g., processing, storage, network, ...
- Consumer is provided customized virtual machines.
- Consumer has control over the resources.
- Example: Amazon Web Services (AWS)

IaaS - (2/2)

- System architecture



PaaS - (1/2)

- Vendor provides development environment.
  - Tools and technology selected by vendor.
  - Control over data life-cycle.

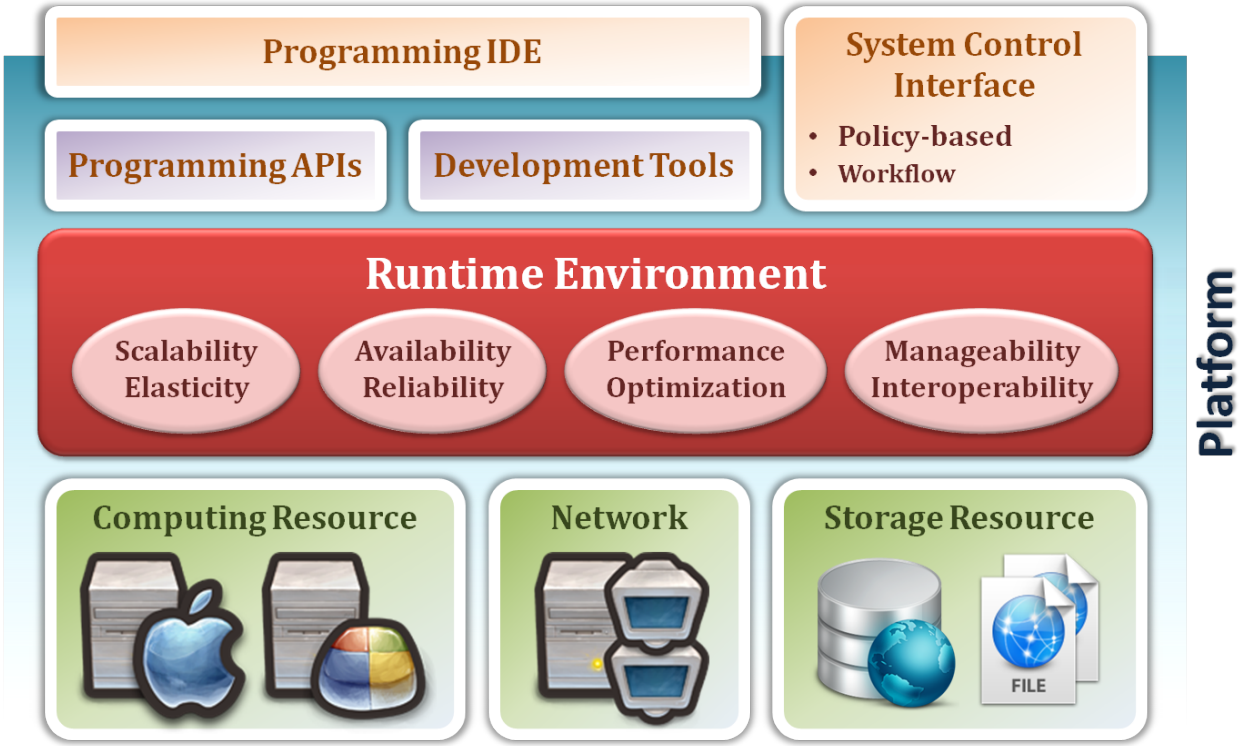
PaaS - (1/2)

- Vendor provides development environment.
  - Tools and technology selected by vendor.
  - Control over data life-cycle.

- Example: Google app engine, Microsoft Azure

PaaS - (2/2)

- System architecture



SaaS - (1/3)

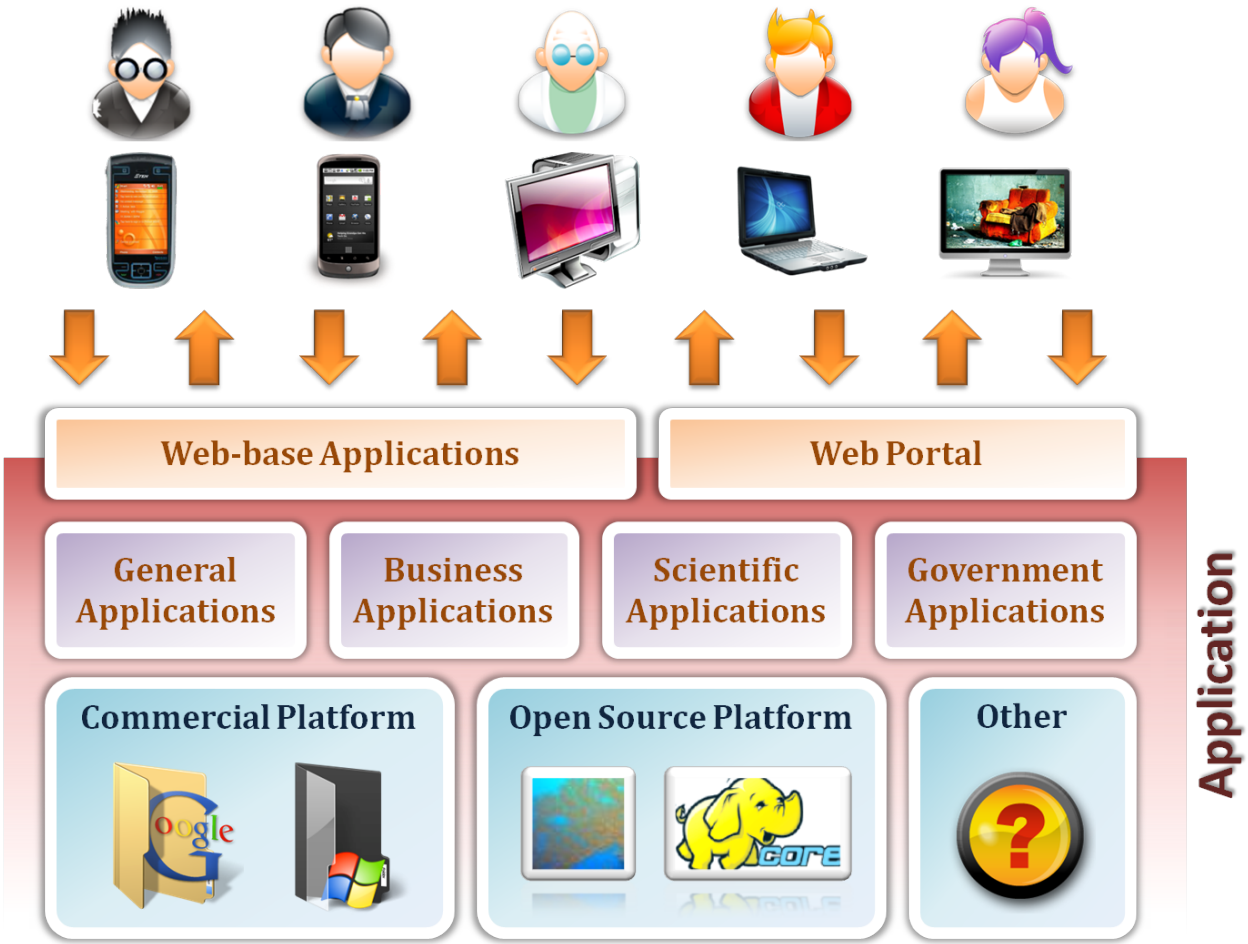
- Vendor provides applications accessed over the network.

SaaS - (1/3)

- Vendor provides applications accessed over the network.
- Example: Google Docs, Salesforce.com

SaaS - (2/3)

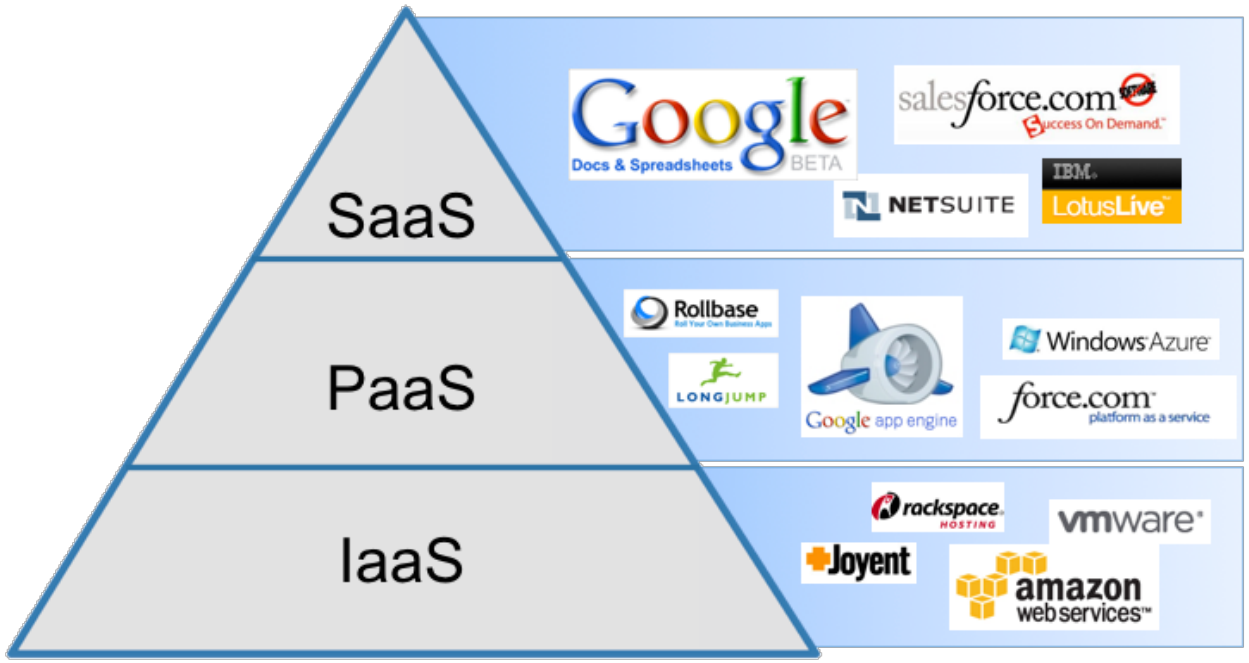
- System architecture



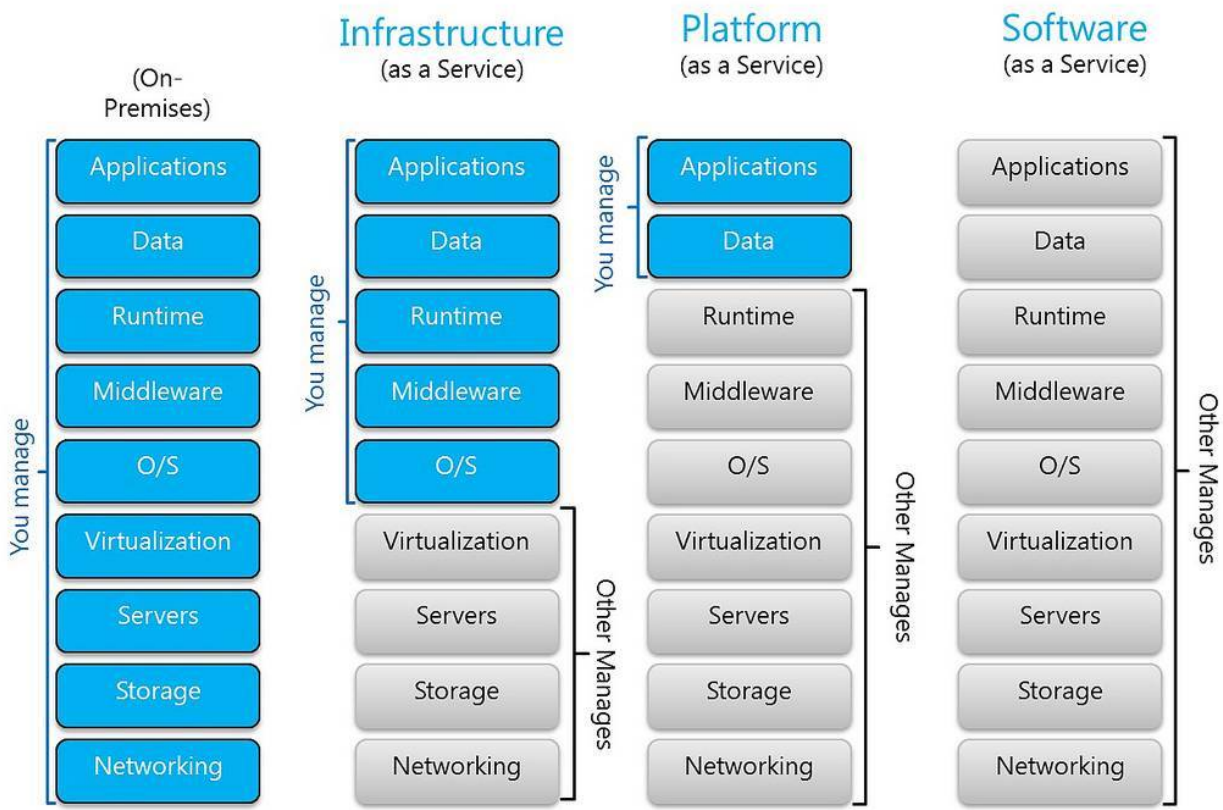
SaaS - (3/3)

- Web Service and Web 2.0
- Viewing the Internet as a computing platform.
- Running interactive applications through a web browser.

IaaS - PaaS - SaaS

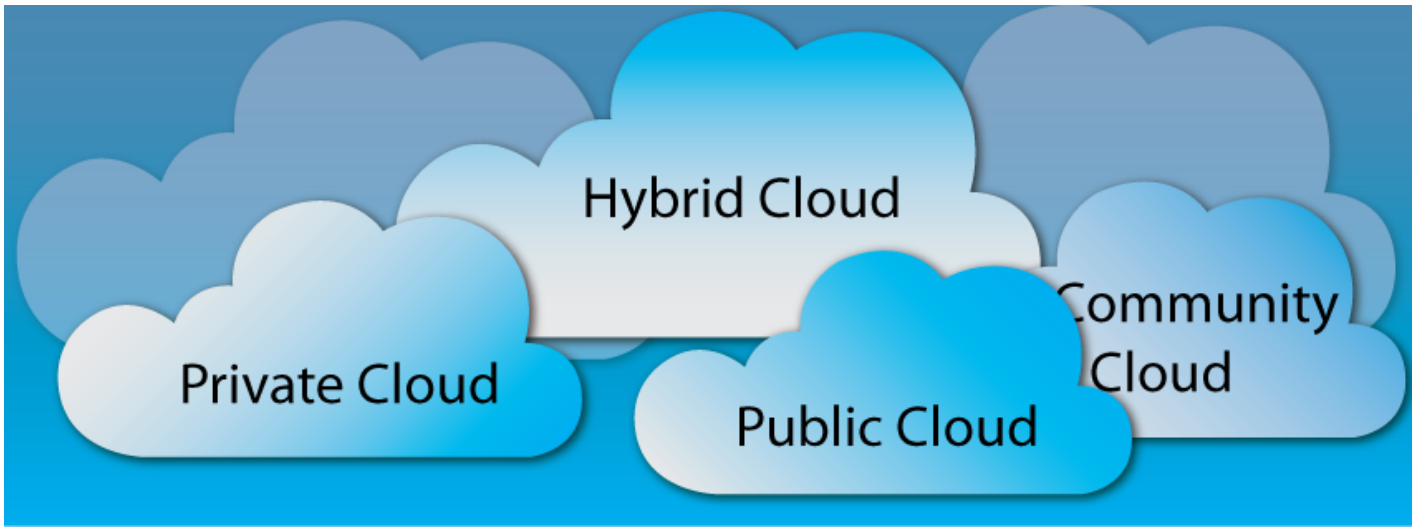


IaaS - PaaS - SaaS



Cloud Deployment Models

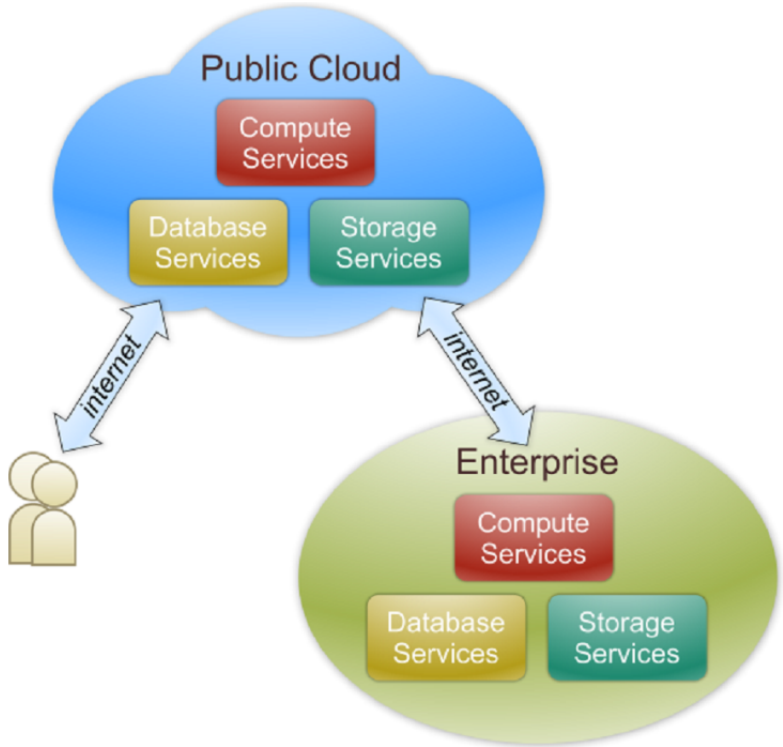
Cloud Deployment Models



[http://www.atomrain.com/it/technology/cloud-deployment-models]

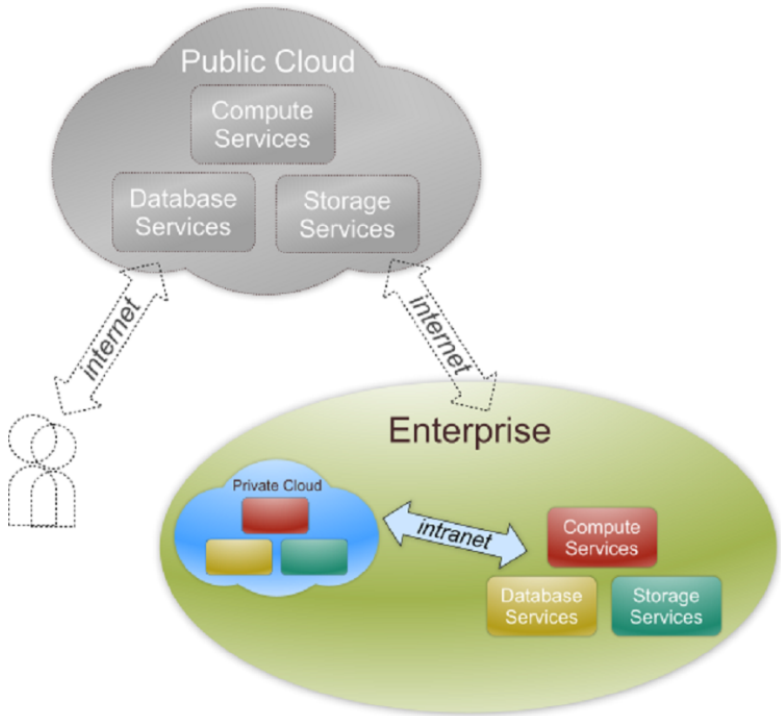
Public Cloud

- Infrastructure is made available to the general public.
- Owned by an organization selling cloud services.



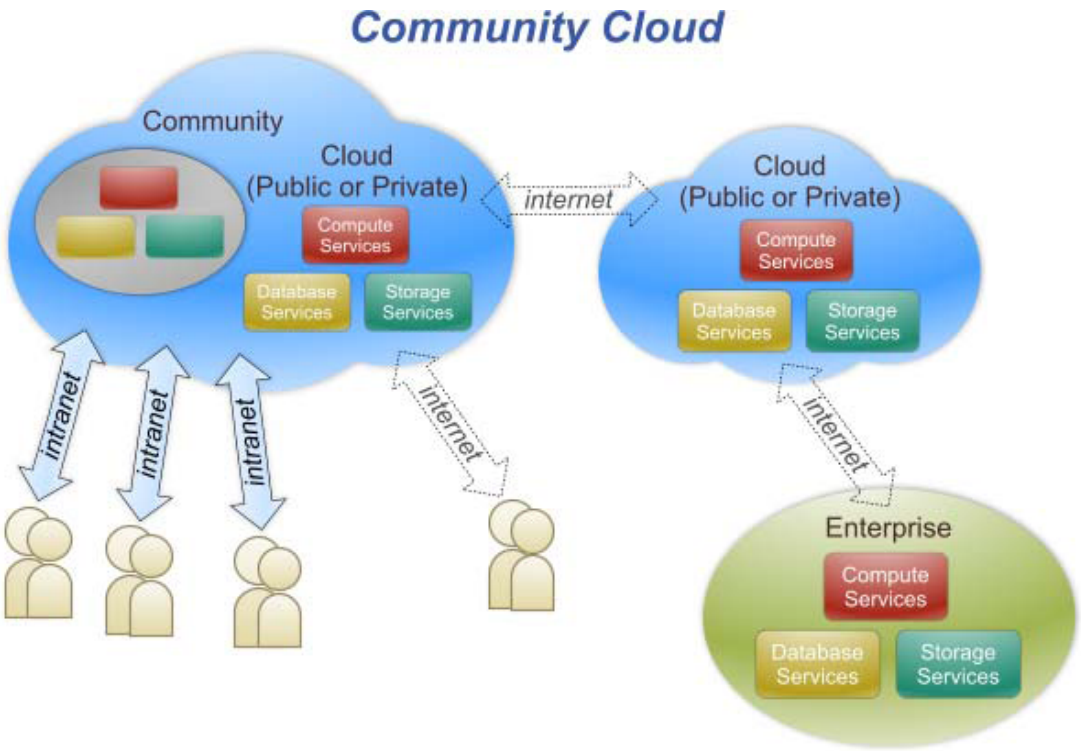
Private Cloud

- Infrastructure is operated solely for an organization.
- Managed by the organization or by a third party.



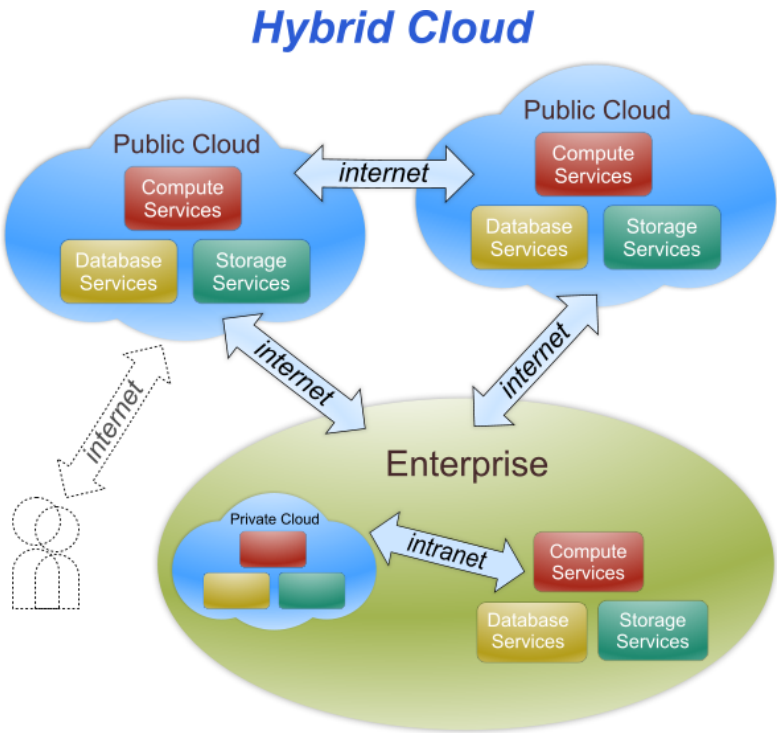
Community Cloud

- Supports a specific community.
- Infrastructure is shared by several organizations.



Hybrid Cloud

- Infrastructure is a composition of two or more clouds deployment models.
- Enables data and application portability.



Big Data

Big Data

> ... everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.
>
> - Dan Ariely

Big Data

> Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.
>
> - O'Reilly

O'REILLY®

Big Data

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

- O'Reilly

Too Many Words

O'REILLY®

Big Data

> Big data is the data characterized by 3 attributes: volume, variety and velocity.
>
> - IBM

Big Data

> Big data is the data characterized by 3 attributes: volume, variety and velocity.
>
> - IBM

Big Data

> Big data is the data characterized by 4 key attributes: volume, variety, velocity and value.
>
> - Oracle

Big Data

> Big data is the data characterized by 4 key attributes: volume, variety, velocity and value.
>
> - Oracle

Buzzwords

Let's Define Big Data In Simple Words

Big Data

DevOps Borat

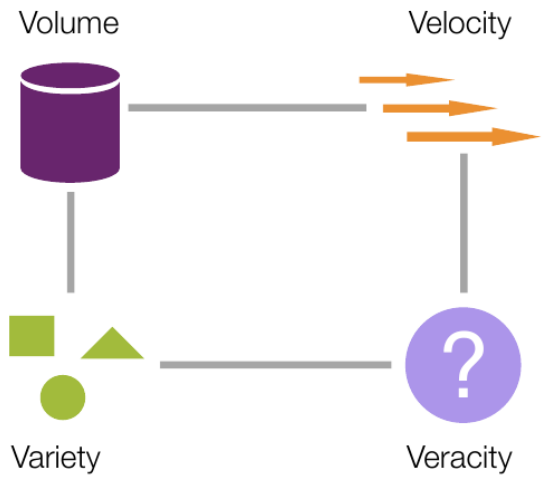
@DEVOPS_BORAT

Small Data is when is fit in RAM. Big Data is when is crash because is not fit in RAM.

2/6/13, 8:22 AM

The Four Dimensions of Big Data

- Volume: data size
- Velocity: data generation rate
- Variety: data heterogeneity
- This 4th V is for Vacillation: Veracity/Variability/Value



Big Data Sources

How Much Data?

- Google: 100 PB/day process, 15000 PB storage
- EBay: 100 PB/day, 90 PB storage
- Baidu: 10-100 PB/day, 2000 PB storage
- Facebook: 600 TB/day, 300 PB storage
- Spotify: 2.2 TB/day, 100 PB storage

[https://followthedata.wordpress.com/2014/06/24/data-size-estimates]

Two Driving Factors

- Cloud computing
- Open source communities

Who Uses Big Data?

- Banking
- Government
- Manufacuring
- Education
- Health care
- ...

How To Store and Process Big Data?

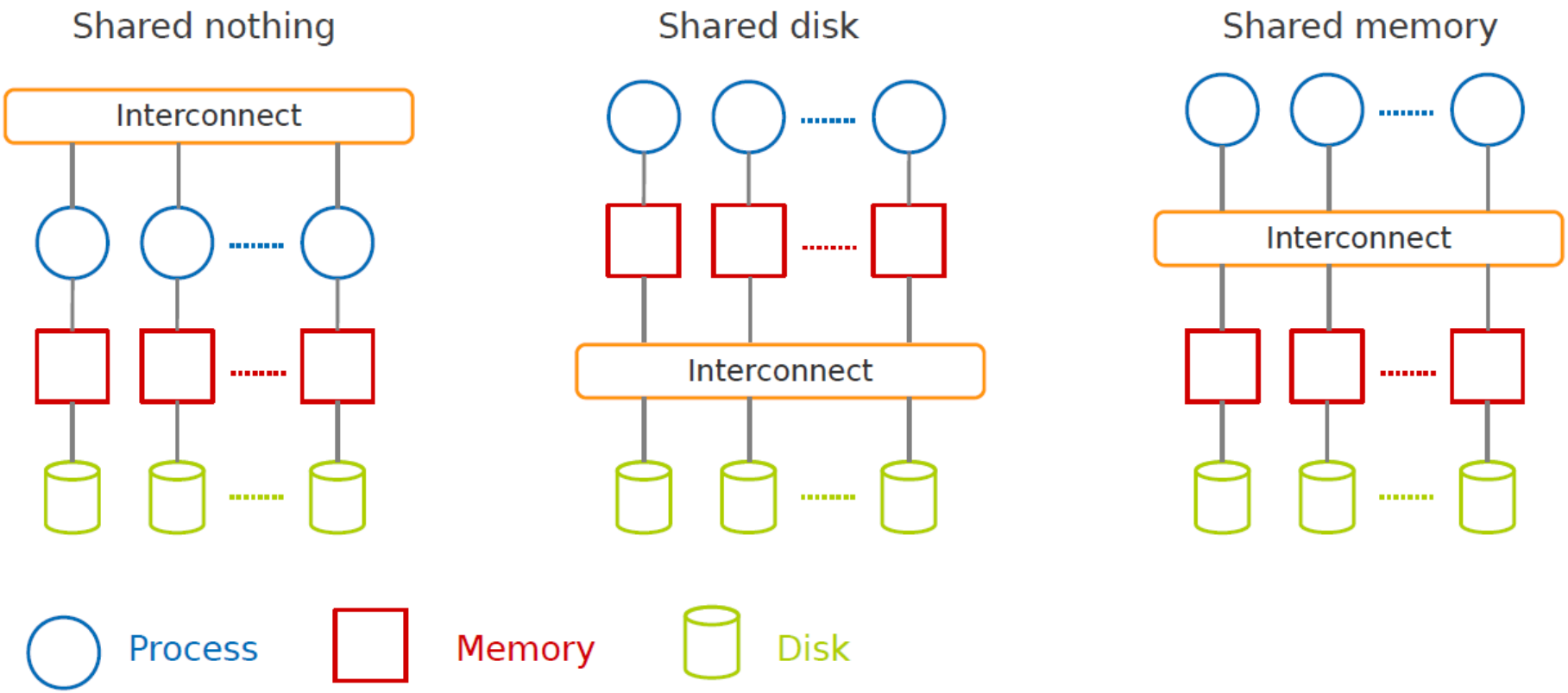
Scale Up vs. Scale Out (1/2)

- Scale up or scale vertically: adding resources to a single node in a system.
- Scale out or scale horizontally: adding more nodes to a system.

Scale Up vs. Scale Out (2/2)

- Scale up: more expensive than scaling out.
- Scale out: more challenging for fault tolerance and software development.

Taxonomy of Parallel Architectures

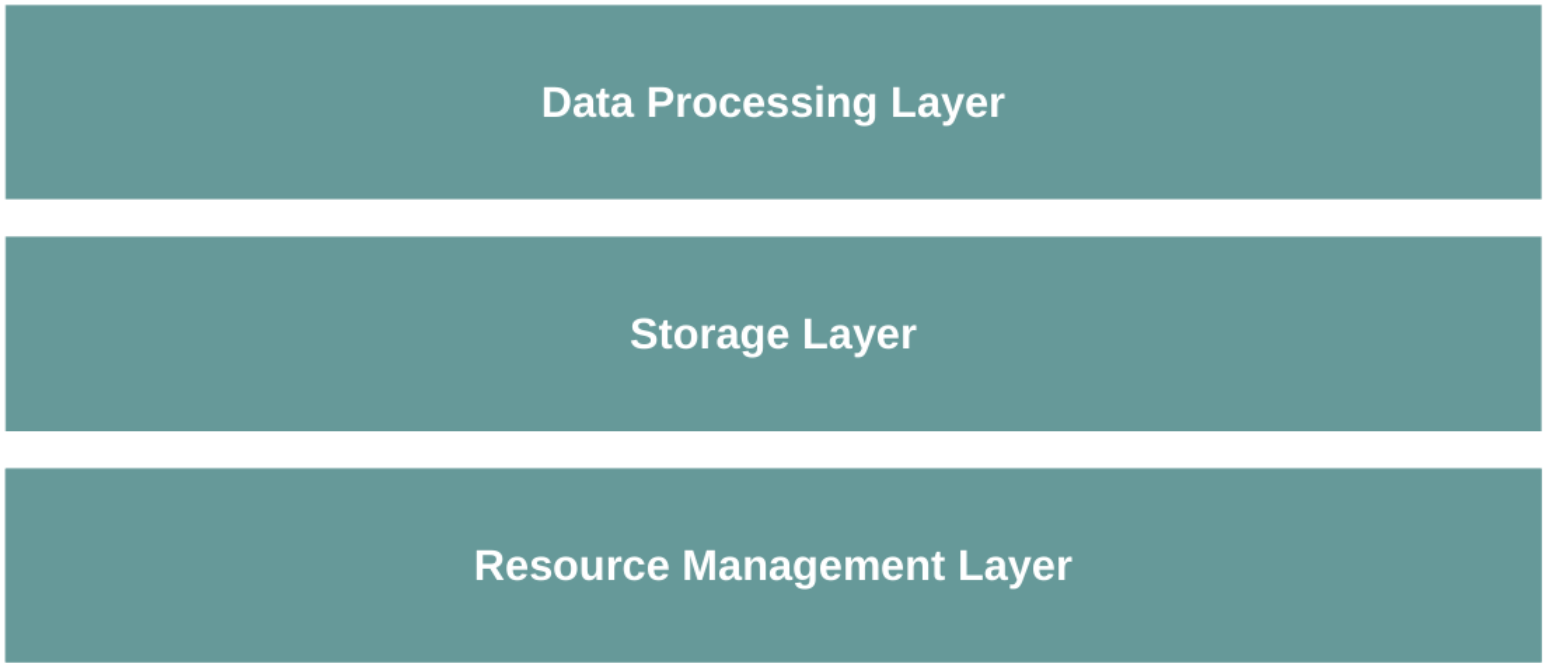


DeWitt, D. and Gray, J. "Parallel database systems: the future of high performance database systems".

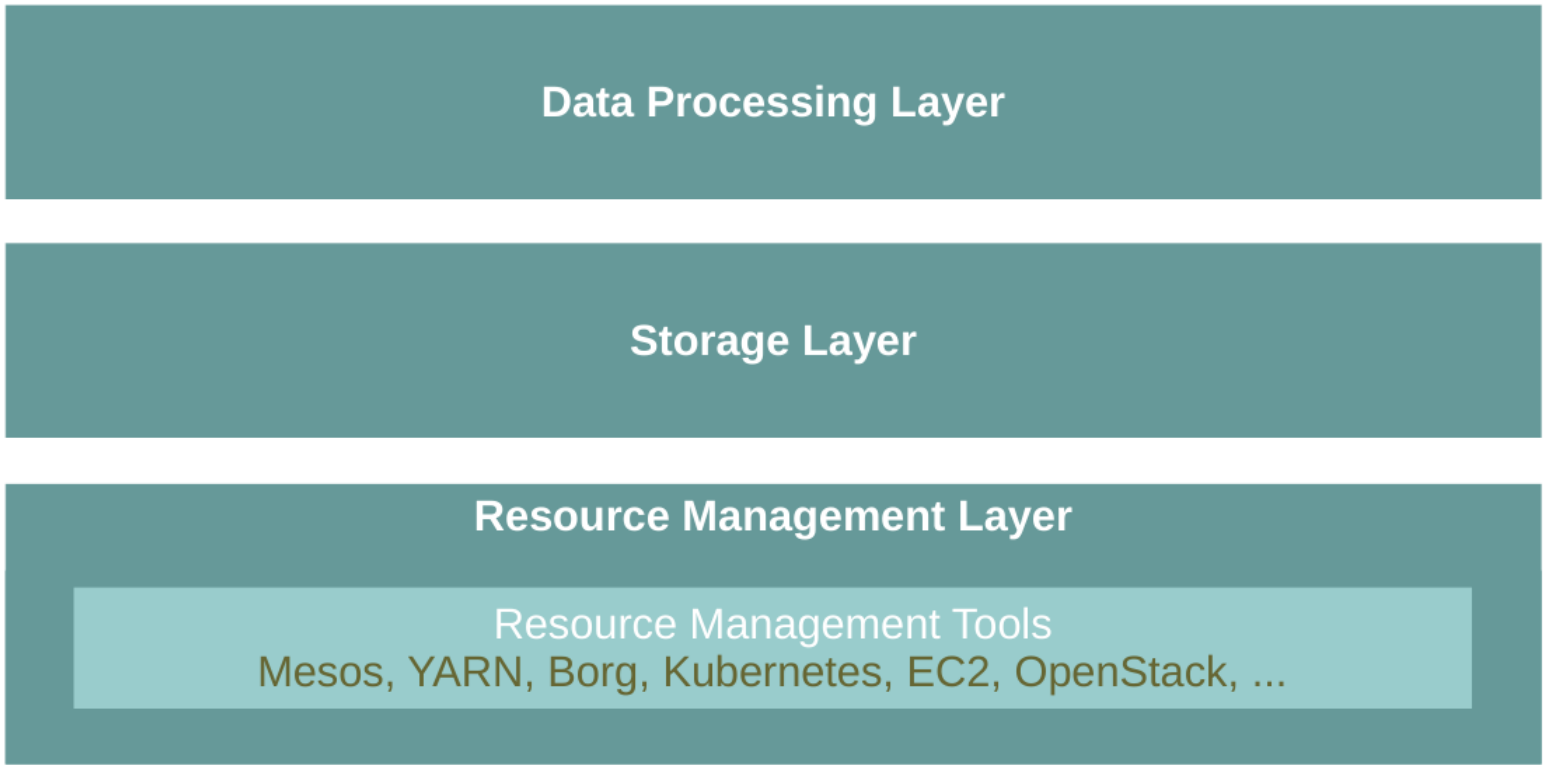
ACM Communications, 35(6), 85-98, 1992


APACHE
HBASE
mahout
GraphX
hadoop
kafka
Storm
Dato
Spark
HIVE
S4 distributed stream computing platform
cassandra
Google Cloud Platform

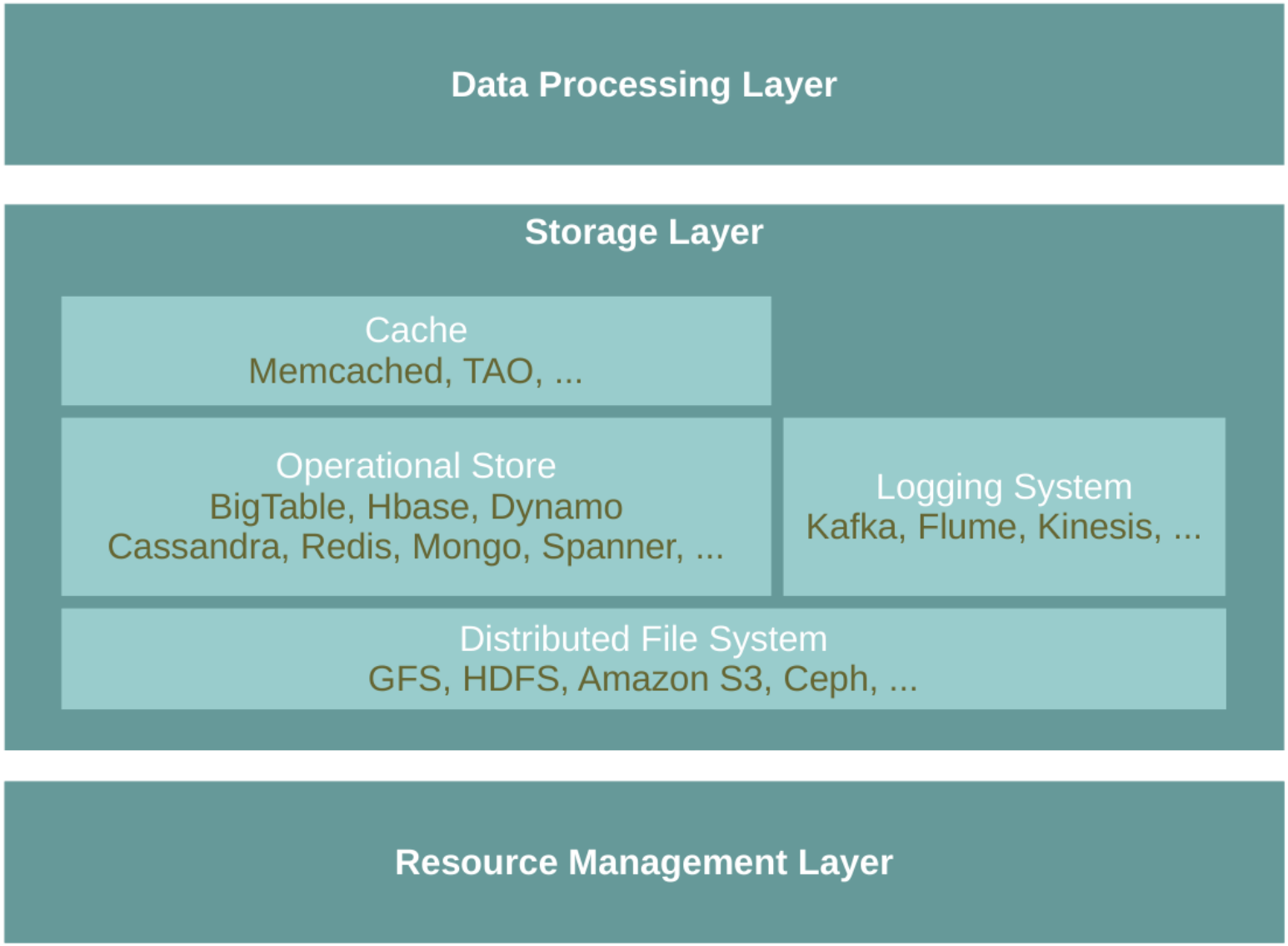
Three Main Layers: Big Data Stack



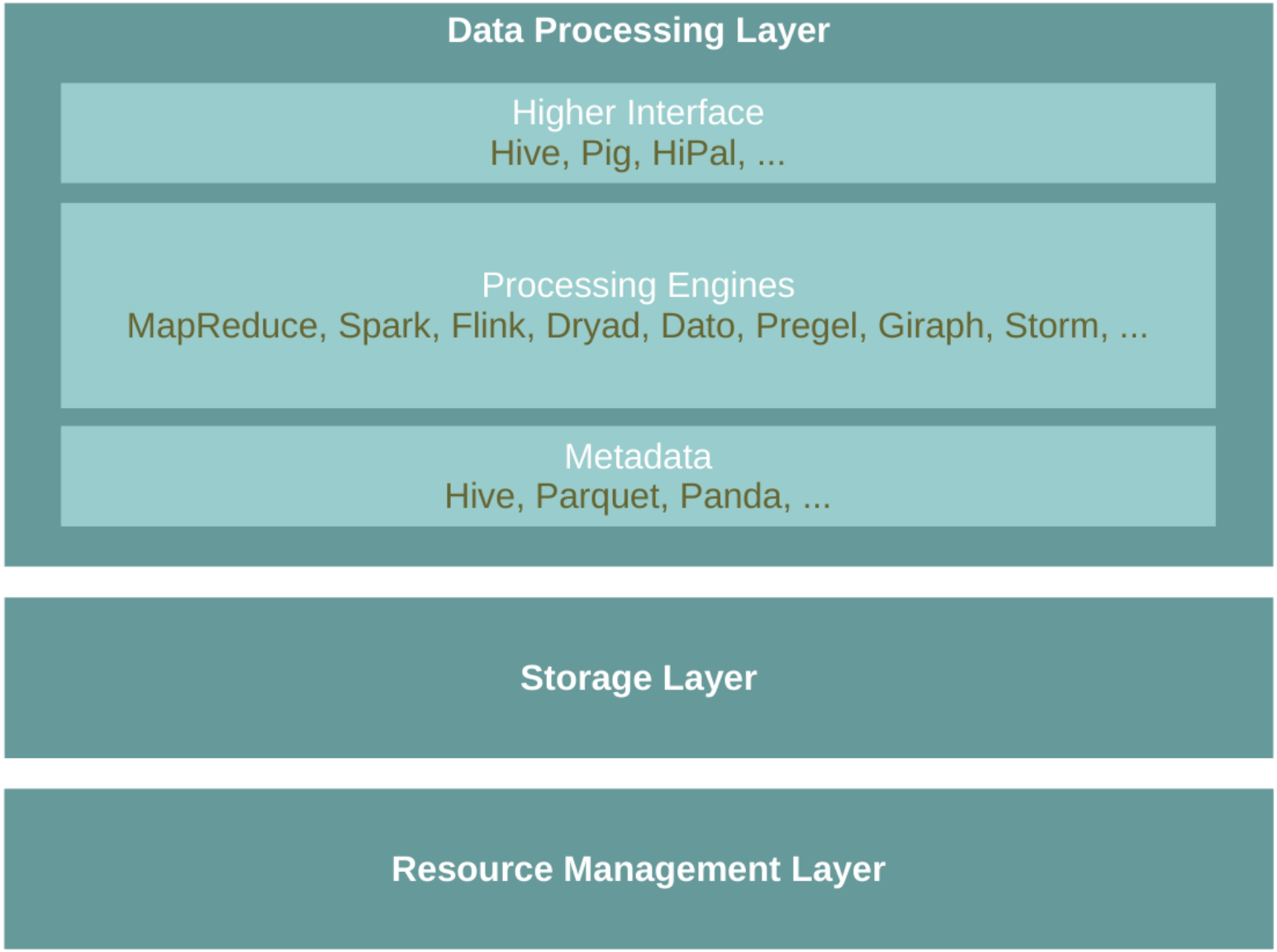
Resource Management Layer



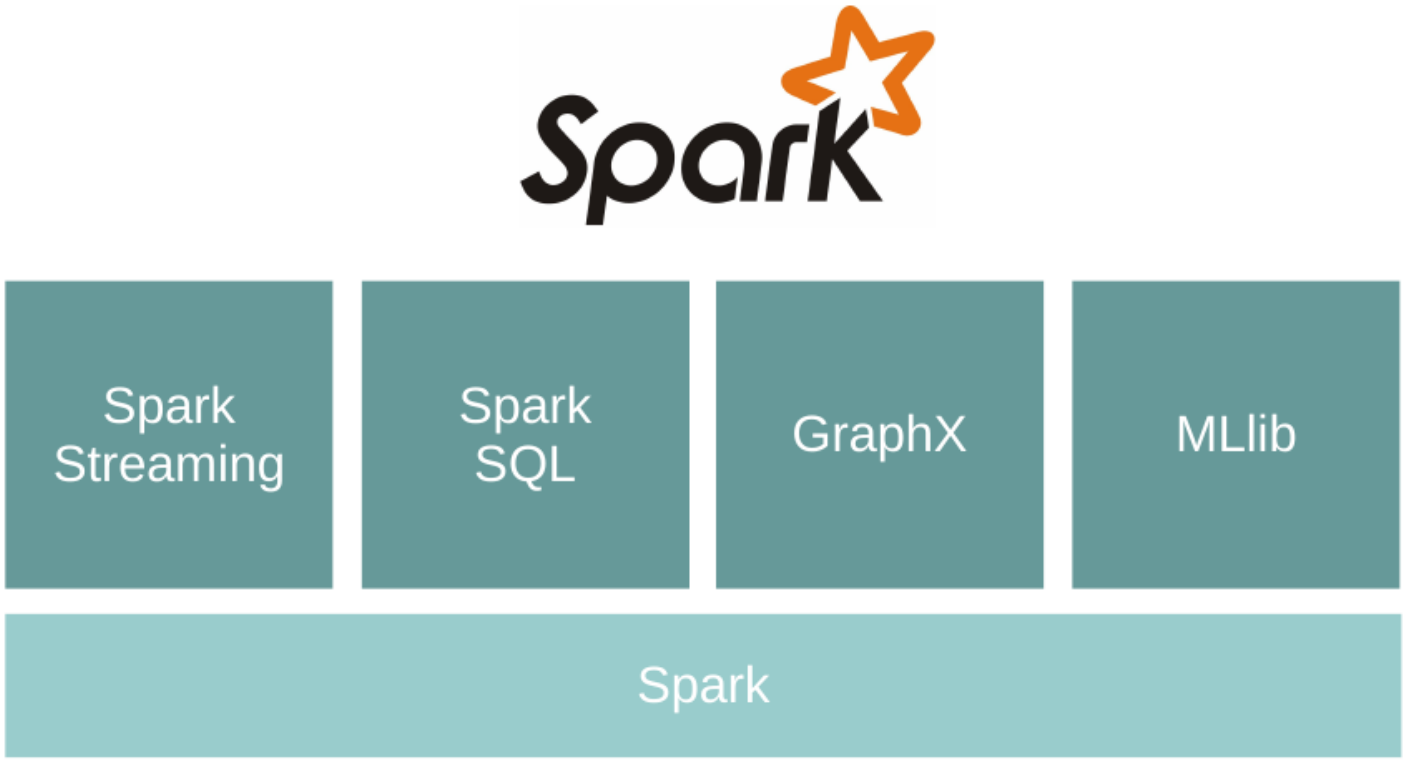
Storage Layer



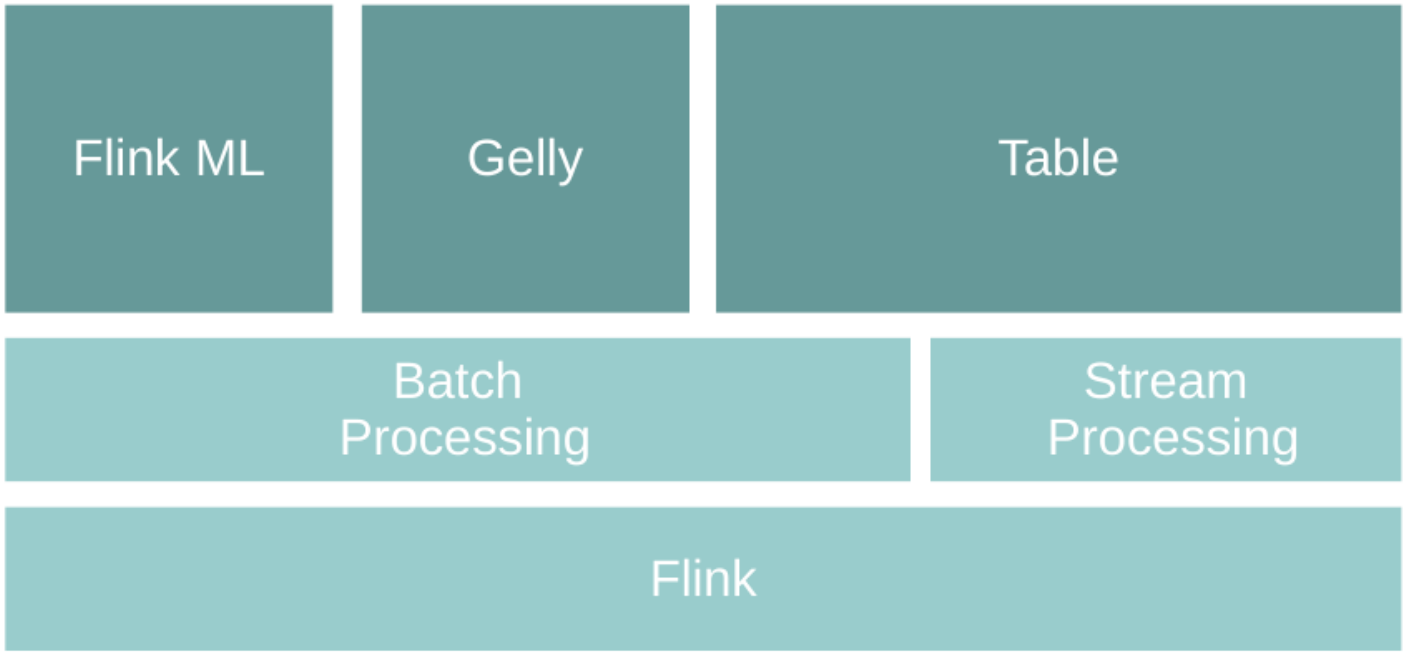
Processing Layer



Spark Processing Engine

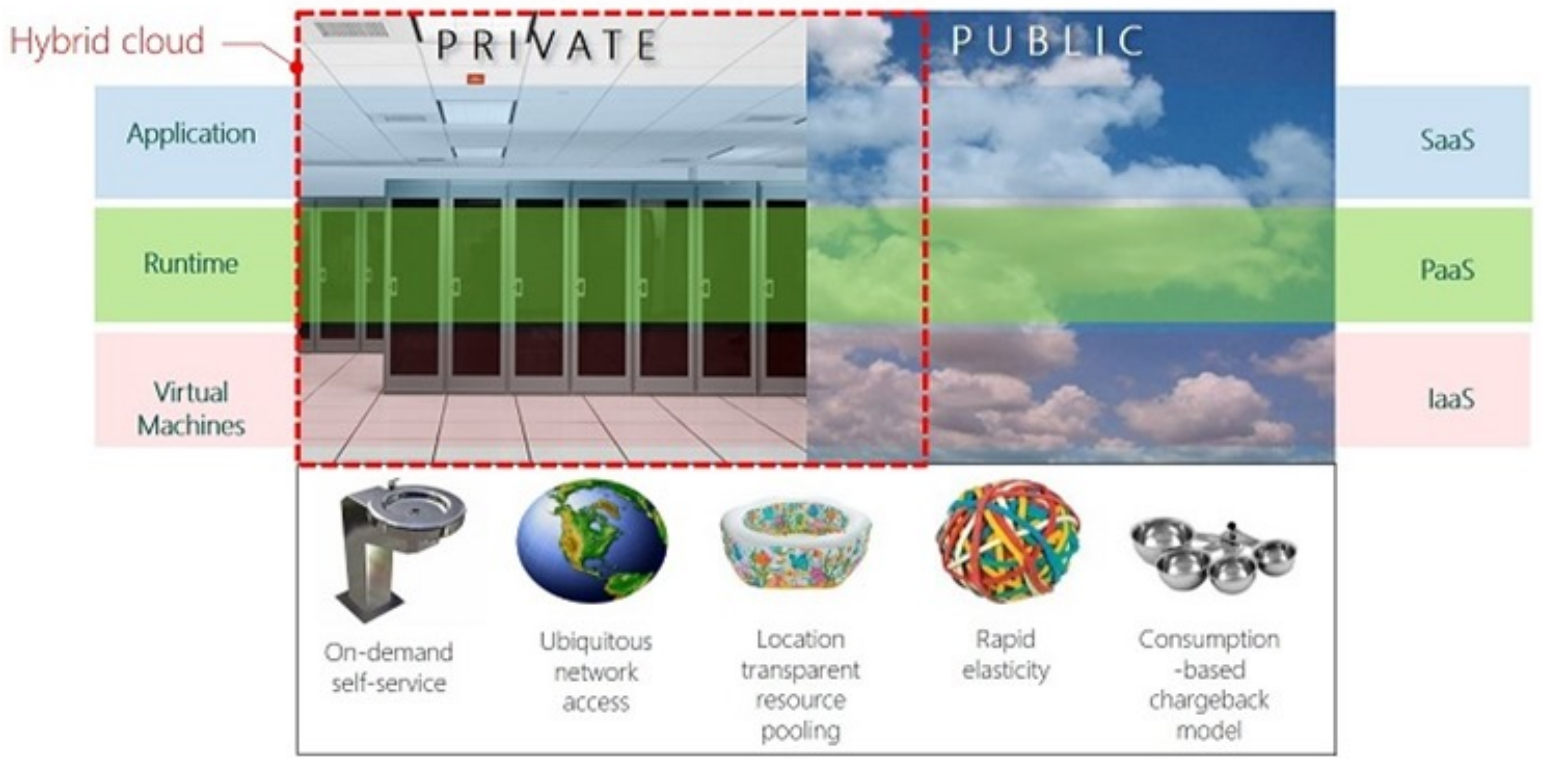


Flink Processing Engine



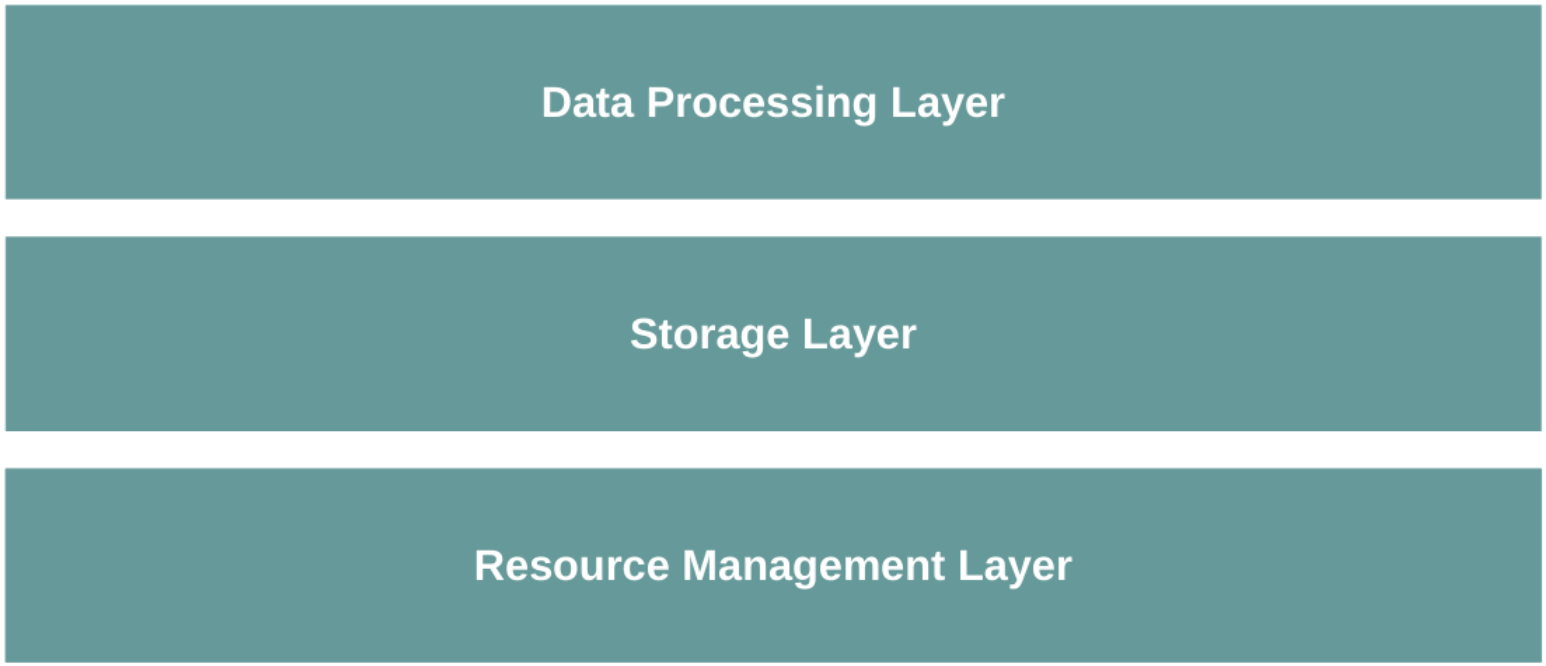
Summary

Summary



[http://aka.ms/532]

Summary


Data Processing Layer
Storage Layer
Resource Management Layer

Questions?