GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

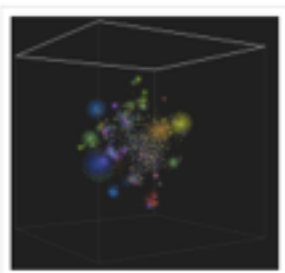
LECTURE 17: CLOUD COMPUTING

Francesco Mosconi
DAT10 SF // Dec 1st, 2014

DATA SCIENCE IN THE NEWS

DATA SCIENCE IN THE NEWS

VISUALIZING HIGH-DIMENSIONAL DATA IN THE BROWSER WITH SVD, T-SNE AND THREE.JS

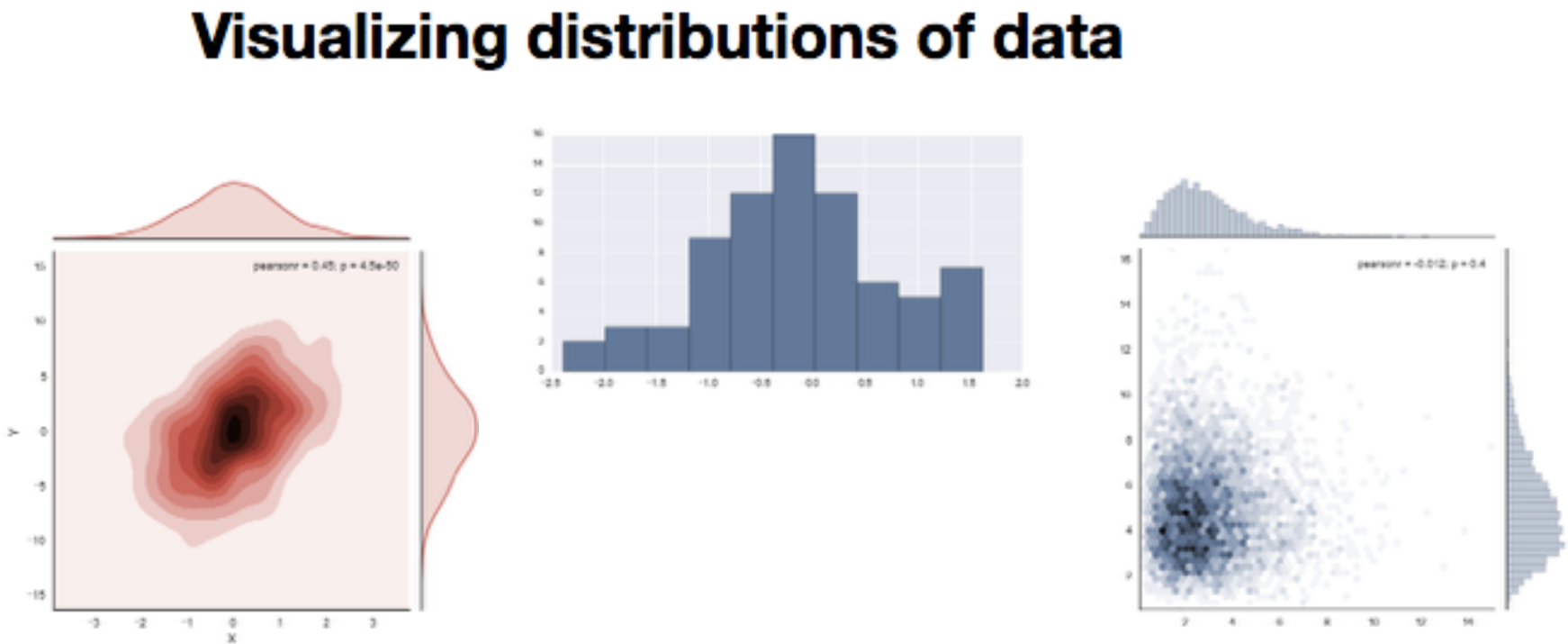


by **Nicolas Kruchten**
Tue, 07/15/2014 - 09:55

Data visualization, by definition, involves making a two- or three-dimensional picture of data, so when the data being visualized inherently has many more dimensions than two or three, a big component of data visualization is dimensionality reduction. Dimensionality reduction is also often the first step in a big-data machine-learning pipeline, because most machine-learning algorithms suffer from the Curse of Dimensionality: more dimensions in the input means you need exponentially more training data to create a good model. Datacratic's products operate on billions of data points (big data) in tens of thousands of dimensions (big problem), and in this post, we show off a proof of concept for interactively visualizing this kind of data in a browser, in 3D (of course, the images on the screen are two-dimensional but we use interactivity, motion and perspective to evoke a third dimension).

source: http://datacratic.com/site/blog/visualizing-high-dimensional-data-browser-svd-t-sne-and-threejs

DATA SCIENCE IN THE NEWS



source: http://nbviewer.ipython.org/github/mwaskom/seaborn/blob/master/examples/plotting_distributions.ipynb

DATA SCIENCE IN THE NEWS

source: http://www.ourworldindata.org/

DATA SCIENCE IN THE NEWS

Stock Price Prediction With Big Data and Machine Learning

Apache Spark and Spark MLLib for building price movement prediction model from order log data.

source: http://eugenezhulenev.com/blog/2014/11/14/stock-price-prediction-with-big-data-and-machine-learning/

LAST TIMES:

I. QUICK REVIEW OF REGULAR EXPRESSIONS
II. DEFINITIONS OF TERMS IN NLP

QUESTIONS?

AGENDA

I. INTRO TO CLUSTER COMPUTING
II. LAB
III. REVIEW OF MIDTERM

I. CLOUD COMPUTING

Q: What is a cloud computing?

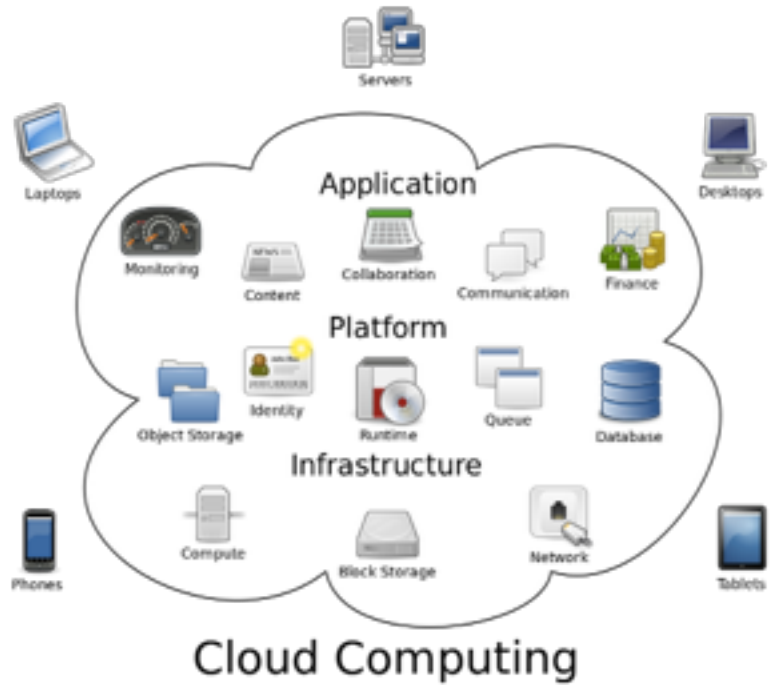
Q: What is a cloud computing?

Computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources.

Q: Why cloud?

Q: Why cloud?

For a user, the network elements representing the provider-rendered services are invisible, as if obscured by a cloud.



source: http://en.wikipedia.org/wiki/Cloud_computing

Q: What is the goal of cloud computing?

Q: What is the goal of cloud computing?
The goal of cloud computing is to allow users to take benefit from existing Information Technologies, without the need for deep knowledge about or expertise with each one of them.

Q: What is the goal of cloud computing?
The goal of cloud computing is to allow users to take benefit from existing Information Technologies, without the need for deep knowledge about or expertise with each one of them.
The cloud aims to cut costs, and help the users focus on their core business instead of being impeded by IT obstacles.

Q: What enables the cloud?

Q: What enables the cloud?

The main enabling technology for cloud computing is virtualization.

Q: What is Virtualization?

the act of creating a virtual (rather than actual) version of something, including but not limited to a virtual computer hardware platform, operating system (OS), storage device, or computer network resources.

source: http://en.wikipedia.org/wiki/Virtualization

Q: What is Virtualization?

With operating system–level virtualization essentially creating a scalable system of multiple independent computing devices, idle computing resources can be allocated and used more efficiently.

source: http://en.wikipedia.org/wiki/Virtualization

Q: Characteristics of cloud computing?

Q: Characteristics of cloud computing?
NIST identifies "five essential characteristics” for cloud computing

Q: Characteristics of cloud computing?
NIST identifies "five essential characteristics" for cloud computing:

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

Let’s go through them one by one…

On-demand self-service

On-demand self-service

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access

Broad network access

Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

Resource pooling

Resource pooling

The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand

Rapid elasticity

Rapid elasticity

Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time

Measured service

Measured service

Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Discussion

How can data science and machine learning use cloud computing?

II. CLOUD TOOLS

CLOUD TOOLS

Q: what kind of services?

Q: what kind of services?

- virtual computers (instances)
- storage
- databases
- queues and pipelines
- and more…

Q: instances

Amazon EC2

Q: instances

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)
General Purpose - Current Generation				
t2.micro	1	Variable	1	EBS Only
t2.small	1	Variable	2	EBS Only
t2.medium	2	Variable	4	EBS Only
m3.medium	1	3	3.75	1 x 4 SSD
m3.large	2	6.5	7.5	1 x 32 SSD
m3.xlarge	4	13	15	2 x 40 SSD
m3.2xlarge	8	26	30	2 x 80 SSD

Q: instances

Compute Optimized - Current Generation

c3.large	2	7	3.75	2 x 16 SSD
c3.xlarge	4	14	7.5	2 x 40 SSD
c3.2xlarge	8	28	15	2 x 80 SSD
c3.4xlarge	16	55	30	2 x 160 SSD
c3.8xlarge	32	108	60	2 x 320 SSD

Q: instances

Memory Optimized - Current Generation

r3.large	2	6.5	15	1 x 32 SSD
r3.xlarge	4	13	30.5	1 x 80 SSD
r3.2xlarge	8	26	61	1 x 160 SSD
r3.4xlarge	16	52	122	1 x 320 SSD
r3.8xlarge	32	104	244	2 x 320 SSD

Q: instances

Standard

Instance type	Virtual Cores	Memory
n1-standard-1	1	3.75GB
n1-standard-2	2	7.5GB
n1-standard-4	4	15GB
n1-standard-8	8	30GB
n1-standard-16	16	60GB

Q: instances

High CPU

Machines for tasks that require more virtual

Instance type	Virtual Cores	Memory
n1-highcpu-2	2	1.80GB
n1-highcpu-4	4	3.60GB
n1-highcpu-8	8	7.20GB
n1-highcpu-16	16	14.40GB

Q: instances

Google Cloud Platform

High Memory

Machines for tasks that require more memo

Instance type	Virtual Cores	Memory
n1-highmem-2	2	13GB
n1-highmem-4	4	26GB
n1-highmem-8	8	52GB
n1-highmem-16	16	104GB

Q: Spot Instances

Amazon EC2 Spot Instances

Spot Instances allow you to name your own price for Amazon EC2 computing capacity. You simply bid on spare Amazon EC2 instances and run them whenever your bid exceeds the current Spot Price, which varies in real-time based on supply and demand. The Spot Instance pricing model complements the On-Demand and Reserved Instance pricing models, providing potentially the most cost-effective option for obtaining compute capacity, depending on your application.

Q: what is an AMI?

Q: what is an AMI?
Amazon Machine Image

Q: what is an AMI?
Amazon Machine Image

An Amazon Machine Image (AMI) provides the information required to launch an instance, which is a virtual server in the cloud. You specify an AMI when you launch an instance, and you can launch as many instances from the AMI as you need.

Q: what is an AMI?
Amazon Machine Image

Basically is a virtual image of the Operating System you would like to run on the remote instance

Q: what is an AMI?
Amazon Machine Image

Basically is a virtual image of the Operating System you would like to run on the remote instance.

Third parties provide pre-packaged AMIs for all purposes, including…

Q: what is an AMI?
Amazon Machine Image

Basically is a virtual image of the Operating System you would like to run on the remote instance.

Third parties provide pre-packaged AMIs for all purposes, including… DATA SCIENCE !

Q: Storage

Amazon S3

Key-Value stores are called buckets

Amazon S3 stores data as objects within resources called "buckets." You can store as many objects as you want within a bucket, and write, read, and delete objects in your bucket. Objects can be up to 5 terabytes in size.

source: http://aws.amazon.com/s3/details/

Storage pricing is based on amount of data stored

Q: Databases

Amazon DynamoDB
Amazon RDS
Amazon EMR

Cloud Datastore
Cloud SQL

Resources

http://aws.amazon.com/products/

https://cloud.google.com/products/

III. COMPUTER CLUSTERS

Q: What is a cluster?

Q: What is a cluster?
A: A **computer cluster** consists of a set of **connected computers** that **work together** so that in many respects they can be viewed as a **single system**.

http://en.wikipedia.org/wiki/Computer_cluster

Q: Why use a cluster?

Q: Why use a cluster?
A: General and Specific reasons

Q: Why use a cluster?

A: General:

- Lower Cost: pay-as-you-need
- Elasticity: add and remove resources
- Availability: launch jobs through API

Q: Why use a cluster?

A: General:

- Lower Cost: pay-as-you-need
- Elasticity: add and remove resources
- Availability: launch jobs through API

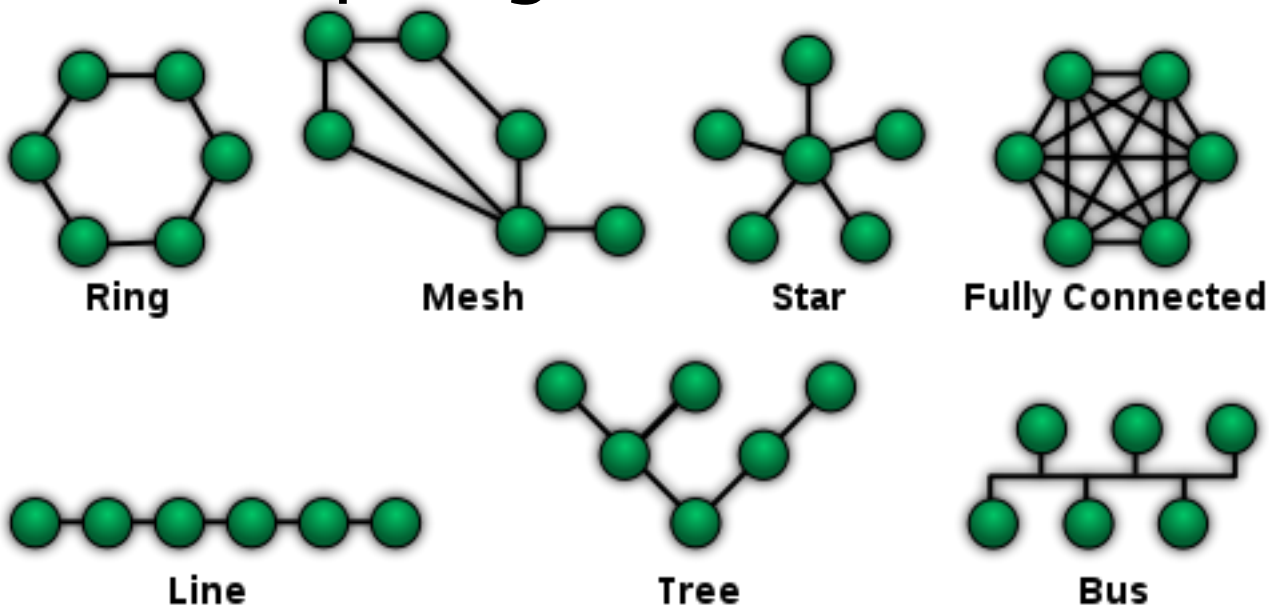
Specific to Data Science:

- Distributed Map Reduce
- Data doesn't fit in memory

Q: How is a cluster formed?

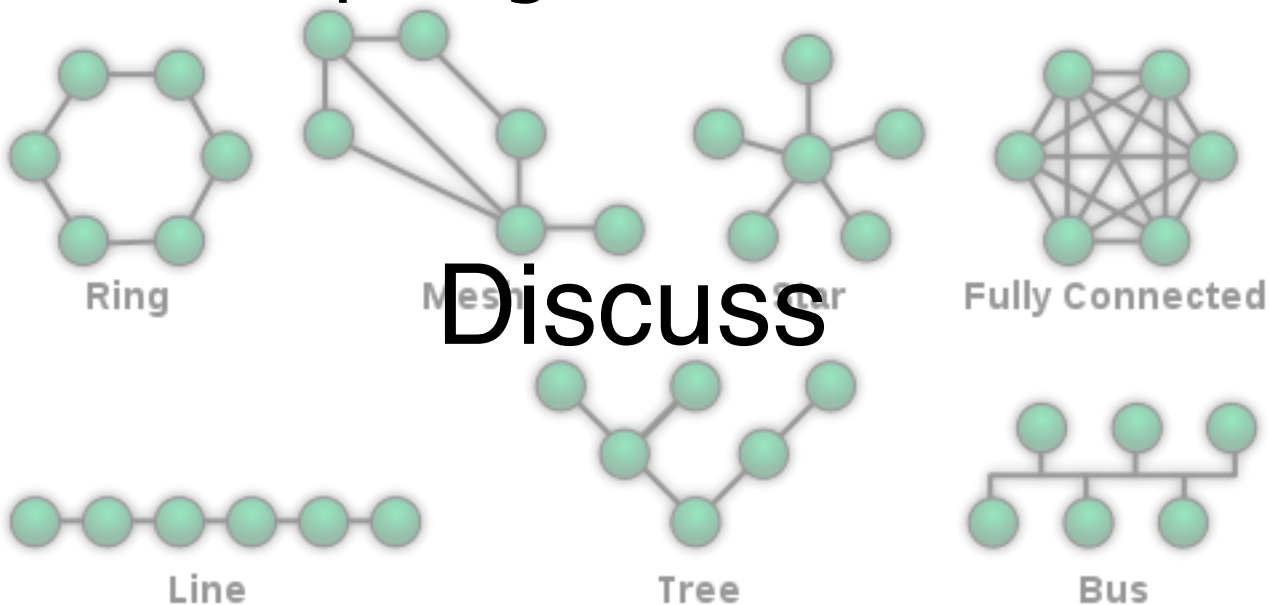
Q: How is a cluster formed?
A: Different topologies

Q: How is a cluster formed?
A: Different topologies



http://en.wikipedia.org/wiki/Computer_cluster

Q: How is a cluster formed?
A: Different topologies



http://en.wikipedia.org/wiki/Computer_cluster

Q: Benefits of Star Topology

Q: Benefits of Star Topology
A:

- Better performance: max 3 dev and 2 links
- Isolation: non-centralized failure no effect
- Centralization: control, fault detection
- Easy Install and config

Q: Features of a cluster?

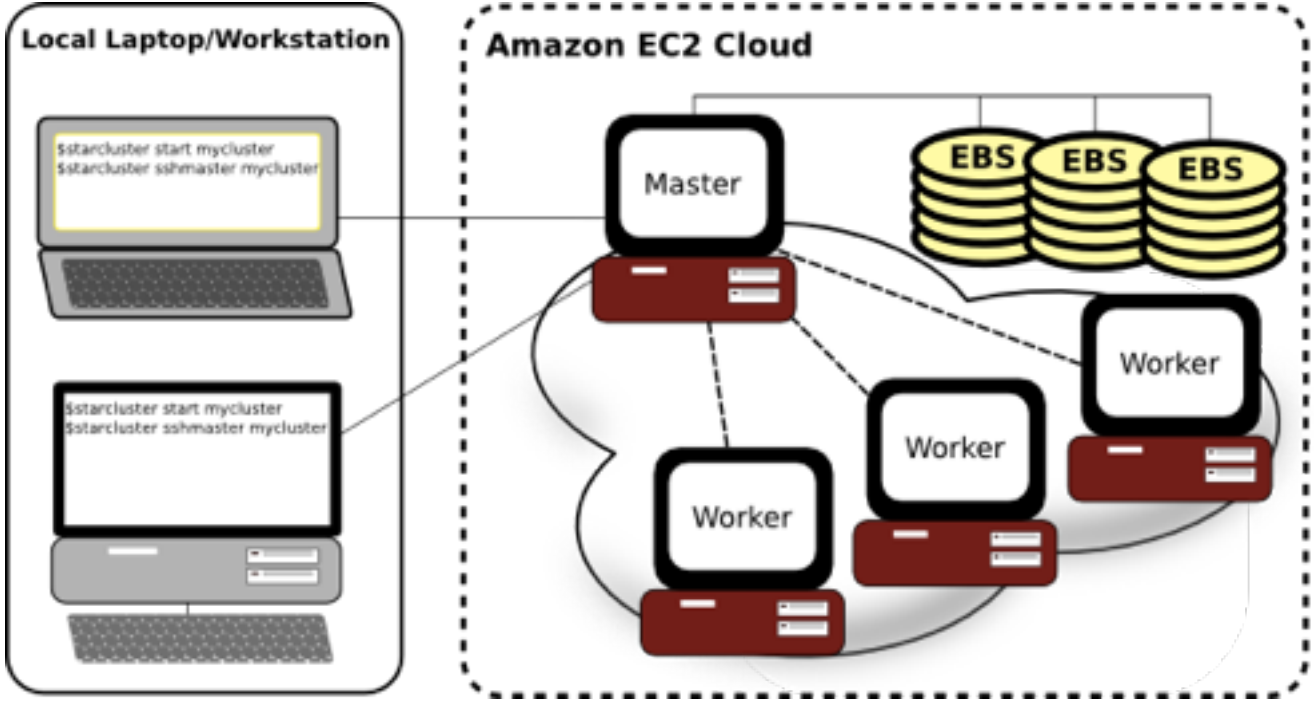
Q: Features of a star cluster?
A:

- Dispatch jobs to whole cluster
- Control individual nodes
- Shared memory

IV. CLUSTER TOOLS: STARCLUSTER & IPYTHON PARALLEL

Q: What is Starcluster?

Q: What is Starcluster?



Q: What is Starcluster?

StarCluster is an open source cluster-computing toolkit for Amazon's Elastic Compute Cloud (EC2). StarCluster has been designed to simplify the process of building, configuring, and managing clusters of virtual machines on Amazon's EC2 cloud.

source: http://star.mit.edu/cluster/docs/latest/overview.html

Starcluster takes care of:

- Security Groups
- Hostnames
- User Accounts
- Password-less SSH
- NFS Share & EBS Volumes
- Scratch Space
- Queueing System

source: http://star.mit.edu/cluster/docs/latest/overview.html

Starcluster also has nice plug-ins:

```
from starcluster.clustersetup import ClusterSetup

class PackageInstaller(ClusterSetup):
    """
    Installs a Debian/Ubuntu package on all nodes in the cluster
    """
    def __init__(self, pkg_to_install):
        self.pkg_to_install = pkg_to_install

    def run(self, nodes, master, user, user_shell, volumes):
        for node in nodes:
            node.ssh.execute('apt-get -y install %s' % self.pkg_to_install)
```

source: http://star.mit.edu/cluster/docs/latest/overview.html

Q: What is iPython parallel?

Q: What is iPython parallel?

iPython module for parallel computing!

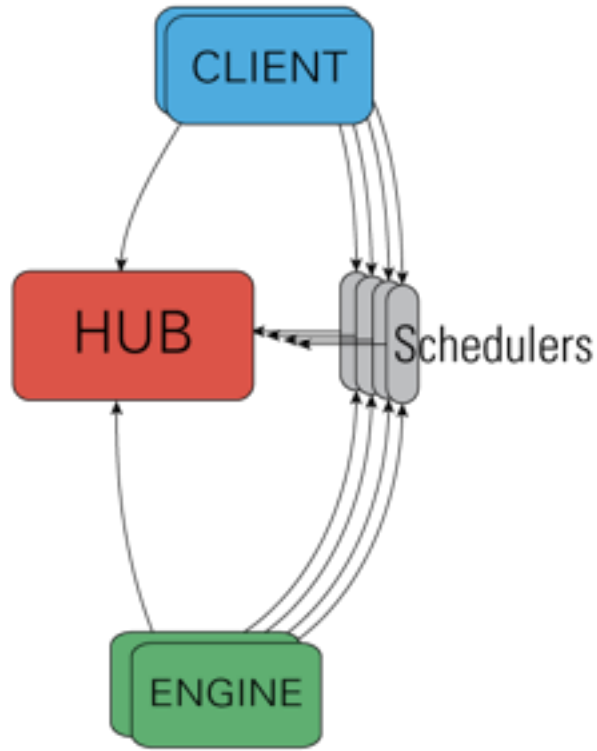
source: http://ipython.org/ipython-doc/2/parallel/parallel_intro.html

iPython parallel support many different parallelisms styles:

- Single program, multiple data (SPMD) parallelism.
- Multiple program, multiple data (MPMD) parallelism.
- Message passing using MPI.
- Task farming.
- Data parallel.
- Combinations of these approaches.

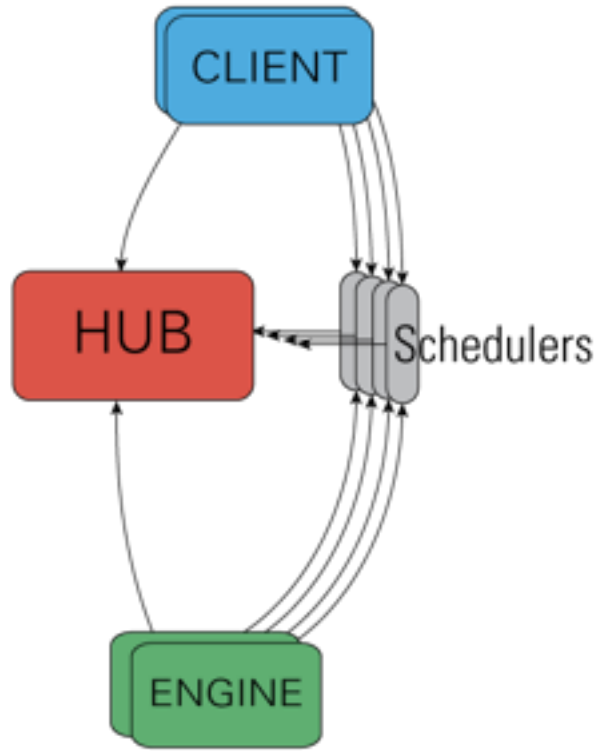
source: http://ipython.org/ipython-doc/2/parallel/parallel_intro.html

iPython architecture



IPython engine

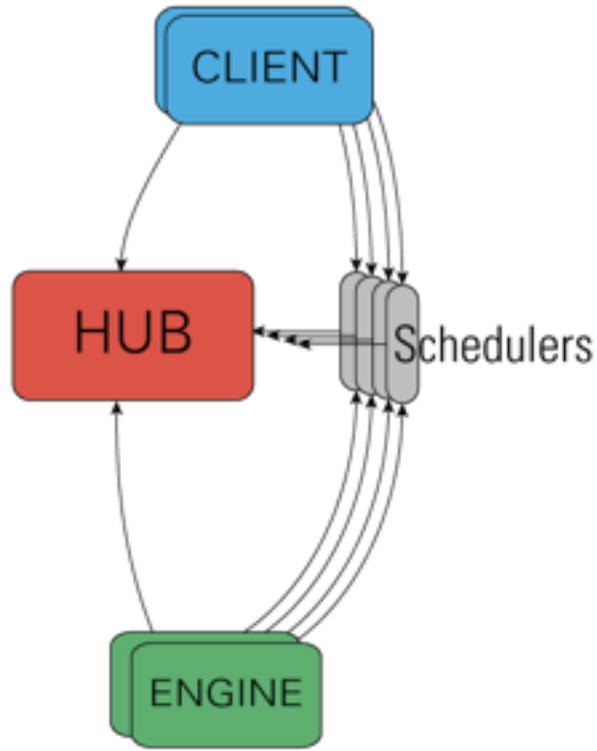
The IPython engine is a Python instance that takes Python commands over a network connection



IPython engine

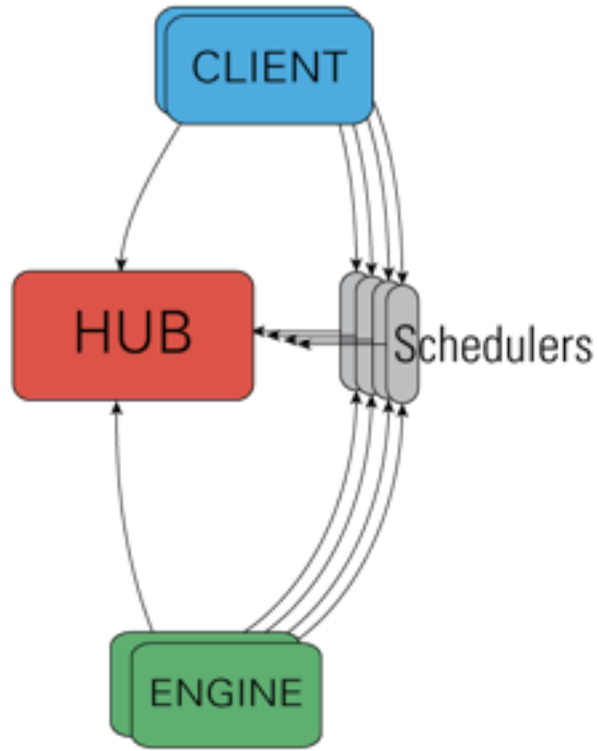
The IPython engine is a Python instance that takes Python commands over a network connection.

Multiple engines => parallel and distributed computing



IPython controller

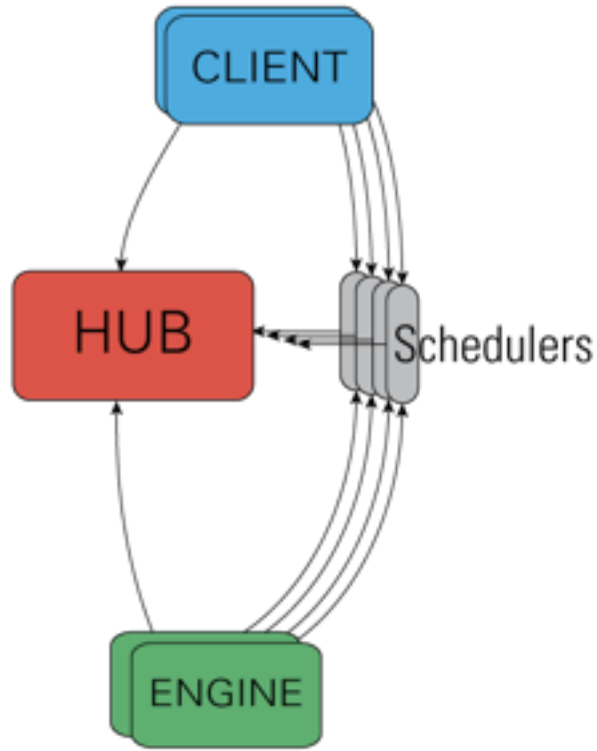
The IPython controller processes provide an interface for working with a set of engines



IPython controller

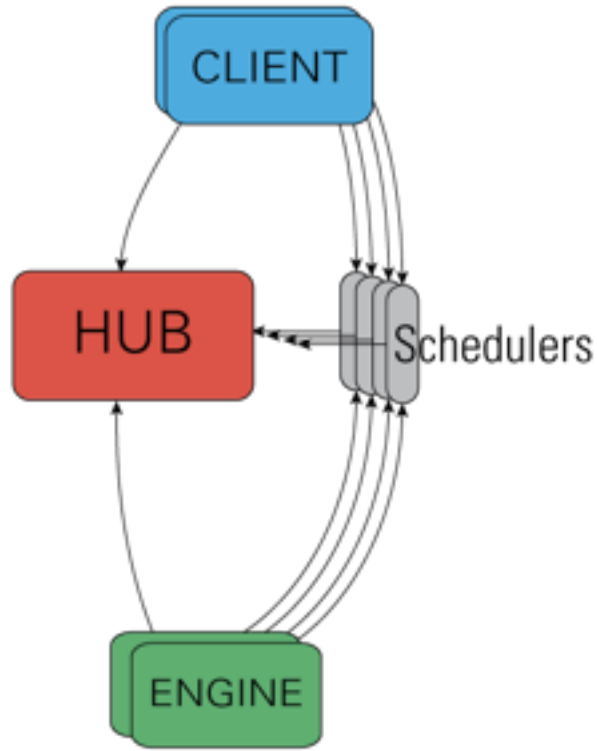
The IPython controller processes provide an interface for working with a set of engines

Composed of a Hub and a collection of Schedulers.



IPython Hub

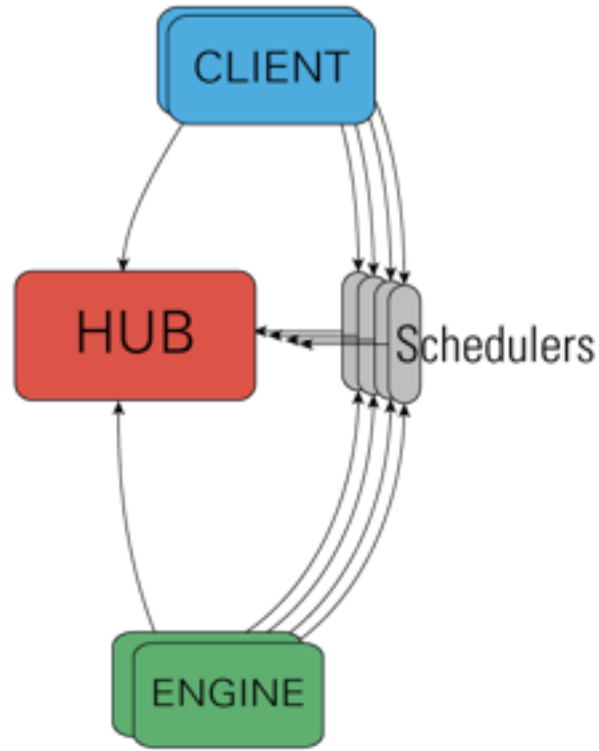
The center of an IPython cluster is the Hub



IPython Hub

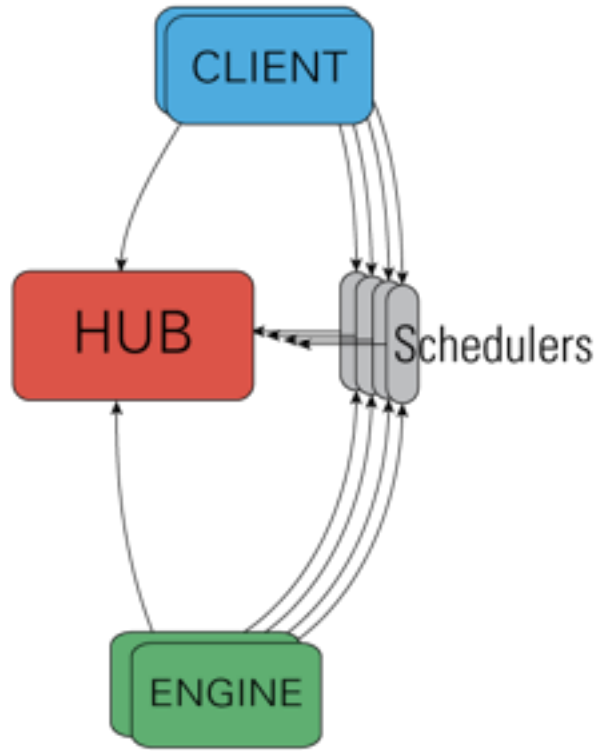
The center of an IPython cluster is the Hub

Process that keeps track of engine connections, schedulers, clients, as well as all task requests and results.



IPython schedulers

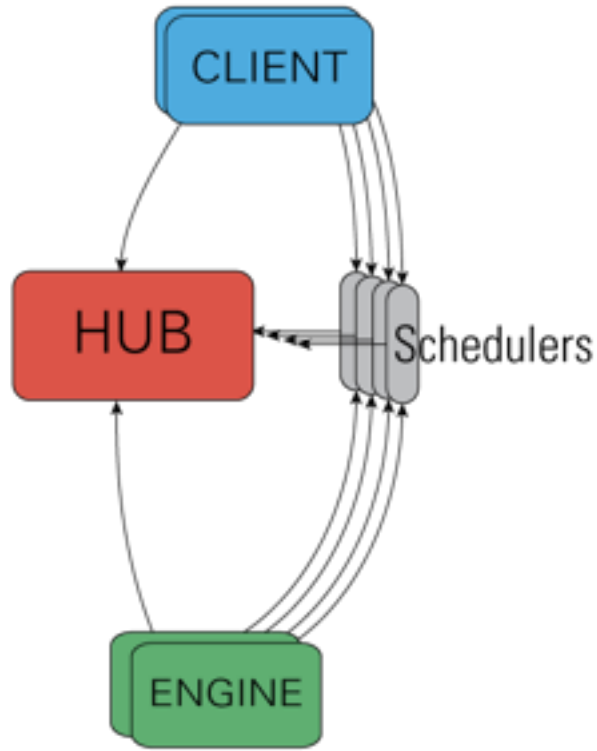
All actions that can be performed on the engine go through a Scheduler



IPython schedulers

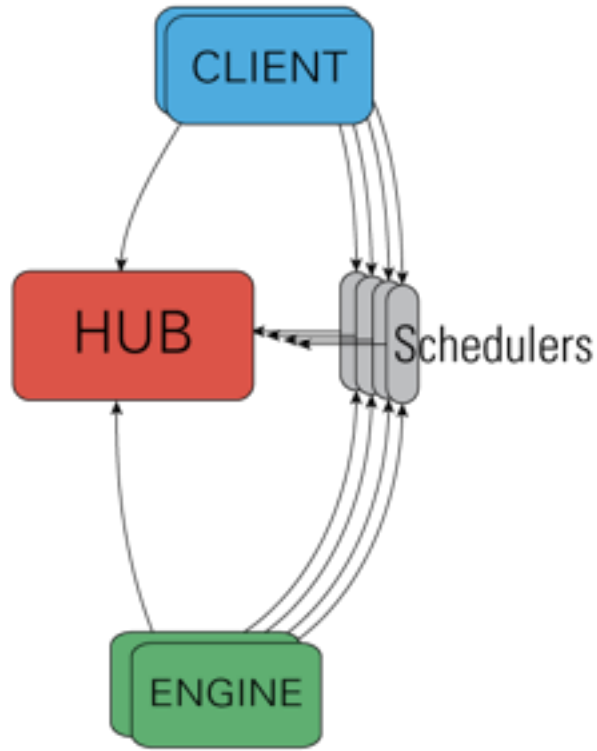
All actions that can be performed on the engine go through a Scheduler

Engines -> block
Schedulers -> asynchronous



IPython client

There is one primary object, the Client, for connecting to a cluster



Getting started

http://ipython.org/ipython-doc/dev/parallel/parallel_intro.html#introduction

http://twiecki.github.io/blog/2014/02/24/ipython-nb-cluster/

https://github.com/ogrisel/parallel_ml_tutorial