Big Data Management Infrastructures and Analytics

Introduzione

9-12 Dic. 2019, Bologna

Barbara Martelli

Organizzazione del corso

- Sala Venturi + Prendiparte dalle 9 alle 18 con pausa pranzo dalle 13 alle 14
- Giovedi' si termina alle 16
- I docenti sono in realta' dei facilitatori
  - varie comunita' e competenze presenti, le lezioni frontali sono l'occasione per scambiarci reciprocamente conoscenza ed esperienze
  - Pause caffe' piuttosto lunghe (30 minuti), per consentirci di interagire
- Hands-on per provare sul campo le tecnologie
- https://indico.infn.it/e/bigdata

Istruzioni per la connessione

- Accesso alla rete wifi
  - Gli utenti **INFN** si devono connettere utilizzando la rete wifi “CNAF-dot1x”.
  - Per tutti gli altri utenti usare il Voucher per la rete ospiti «CNAF»
- Connessione SSH
  - installare un client SSH, per esempio Putty:
    - https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html

Cloud

The pillars of cloud

https://governance.openstack.org/tc/reference/technical-vision.html

- There are at least as many different opinions of what ‘cloud’ means as there are software developers. However, we can all agree that cloud does mean *something*.
- Cloud computing promotes more efficient utilization of resources by reducing the transaction costs involved in provisioning and deprovisioning infrastructure to near zero, and it is able to do so because it differs in qualitative ways from previous models of computing (including virtualization).

The pillars of cloud according to OpenStack

https://governance.openstack.org/tc/reference/technical-vision.html

- Self-service¶
  Clouds are self-service. They provide users with the ability to deploy applications on demand without having to wait for human action or review in the loop. Consequences:
  - cloud services must provide robust multi-tenancy.
  - cloud services must also have some mechanism to ensure that capacity is only utilized when the value to the user of doing so exceeds the opportunity cost to the operator of providing it.
    - In public clouds this is typically accomplished by charging users for the resources consumed. Private clouds will often require the same sorts of monitoring and reporting capabilities

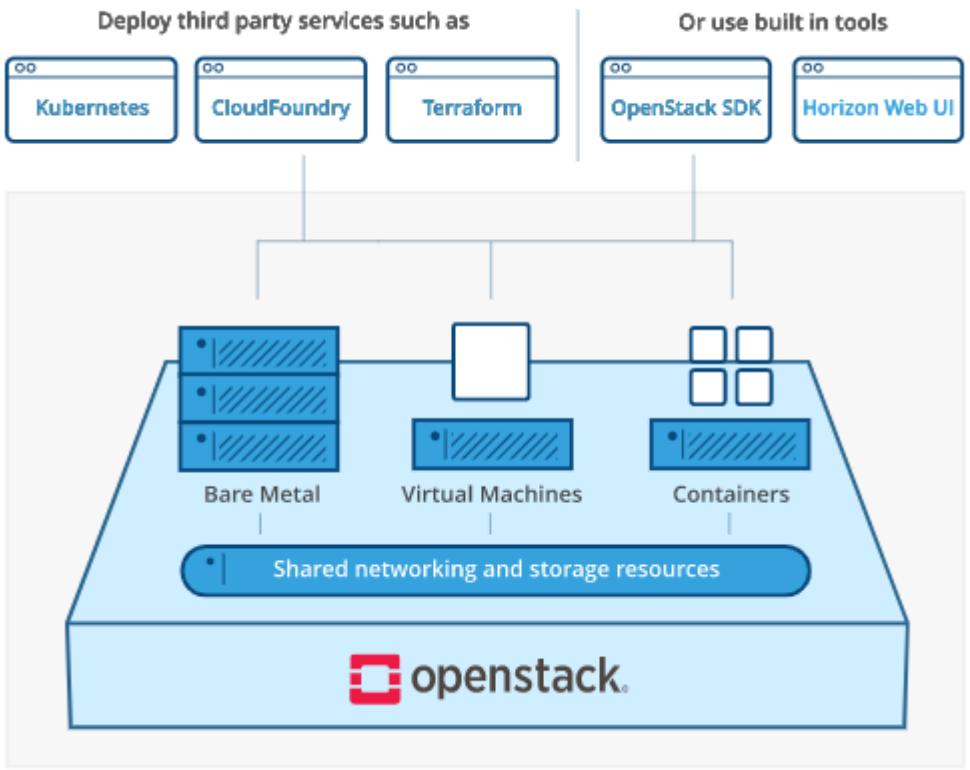
The pillars of cloud

https://governance.openstack.org/tc/reference/technical-vision.html

- Application Control
  - Clouds allow control of an application's infrastructure to be vested in the application itself. clouds also eliminate the need for a human user to be in the loop.
  - While a cloud ***may*** have a user interface (graphical or otherwise), it ***must*** have an application programming interface. It should supply operationally relevant information in a form that is legible to applications, including event notifications where appropriate.
  - It should also be designed to facilitate secure access to its APIs for applications that are running within the cloud itself, because *no part of the application should need to reside outside of the cloud*.

openstack

- OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed and provisioned through APIs with common authentication mechanisms.
- A dashboard is also available, giving administrators control while empowering their users to provision resources through a web interface.
- Beyond standard infrastructure-as-a-service functionality, additional components provide orchestration, fault management and service management amongst other services to ensure high availability of user applications.



Containers

When to chose containers vs VM?

- *If the point of the shift is at the operating system level — to provide each user or user population with its own operating environment while requiring as few physical servers as possible — then hardware virtualization is a logical choice. If the focus is on the application, with the operating system hidden or irrelevant to the user, then Docker or a similar container-based system becomes a realistic option for deployment.*
  - https://rancher.com/learning-paths/containers-vs-virtualization/

- See Cesini, Duma, Costantini presentations on Tuesday

Big Data

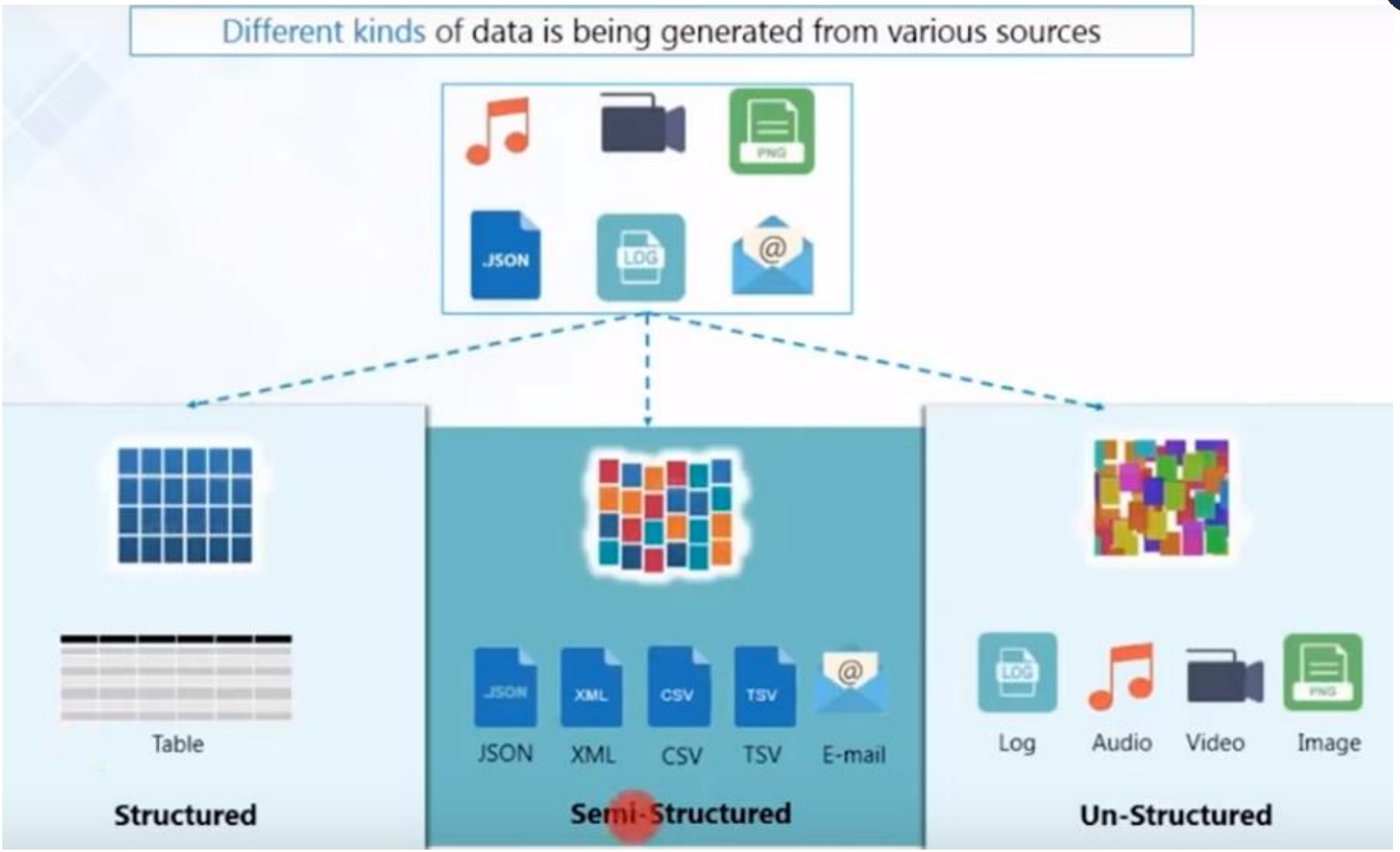
Big Data 5Vs (7, 10…)

- **V**olume, **V**elocity, **V**ariety, **V**eracity, **V**alue.
- **V**isualization, **V**ariability

Volume

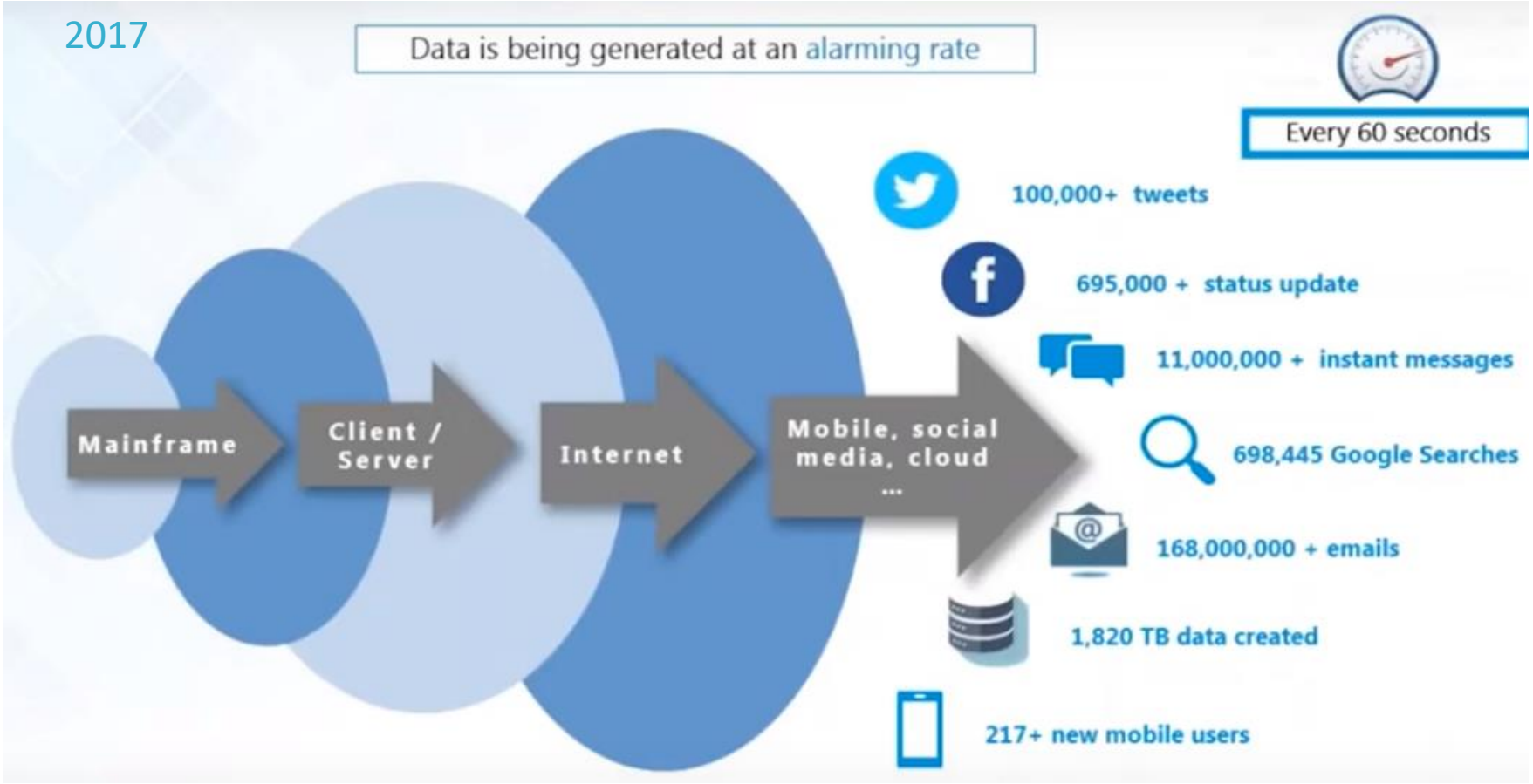
- By 2020 accumulated digital data wil grow from 4.4 Zettabytes (Sep 2017) to 44 Zettabytes (www.edureka.co)
- **90% of all data has been created in the last two years**
  - https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#717adc4760ba

Variety



- https://www.edureka.co/

Velocity



2019

Social Media numbers for things happening EVERY MINUTE of the day:

- 456,000 tweets sent via Twitter
- Facebook:510,000 comments posted and 293,000 statuses updated
- 120 professionals join LinkedIn
- 4,146,600 users watch YouTube videos (yes, that is over 4 million videos watched every minute!)
- 46,740 photos posted by Instagram users
- 156 million Emails are sent + 130 million spam Emails sent

- https://www.edureka.co/

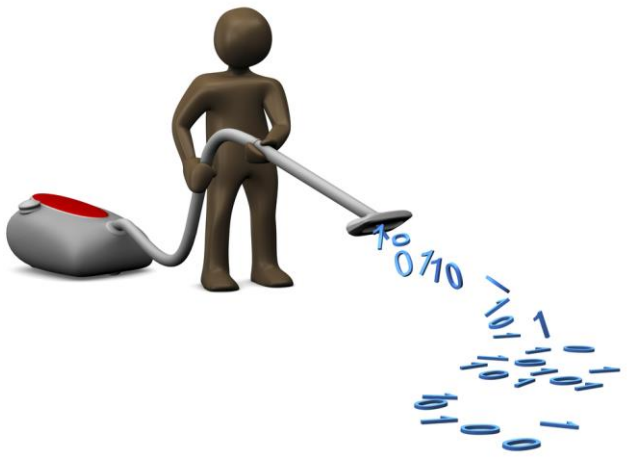
https://iorgforum.org/case-study/some-amazing-statistics-about-online-data-creation-and-growth-rates/

Value, Veracity

- **Value:** How to extract useful information from the big-data ocean

strictly connected to

- **Veracity:** data are often dirty, with NA values, errors, thats why one of the first step for a data scientist is *Data Cleaning:* the process of detecting and correcting (or removing) corrupt or inaccurate entries from a dataset
  - *See Elisabetta's presentation on Thursday*



Variability

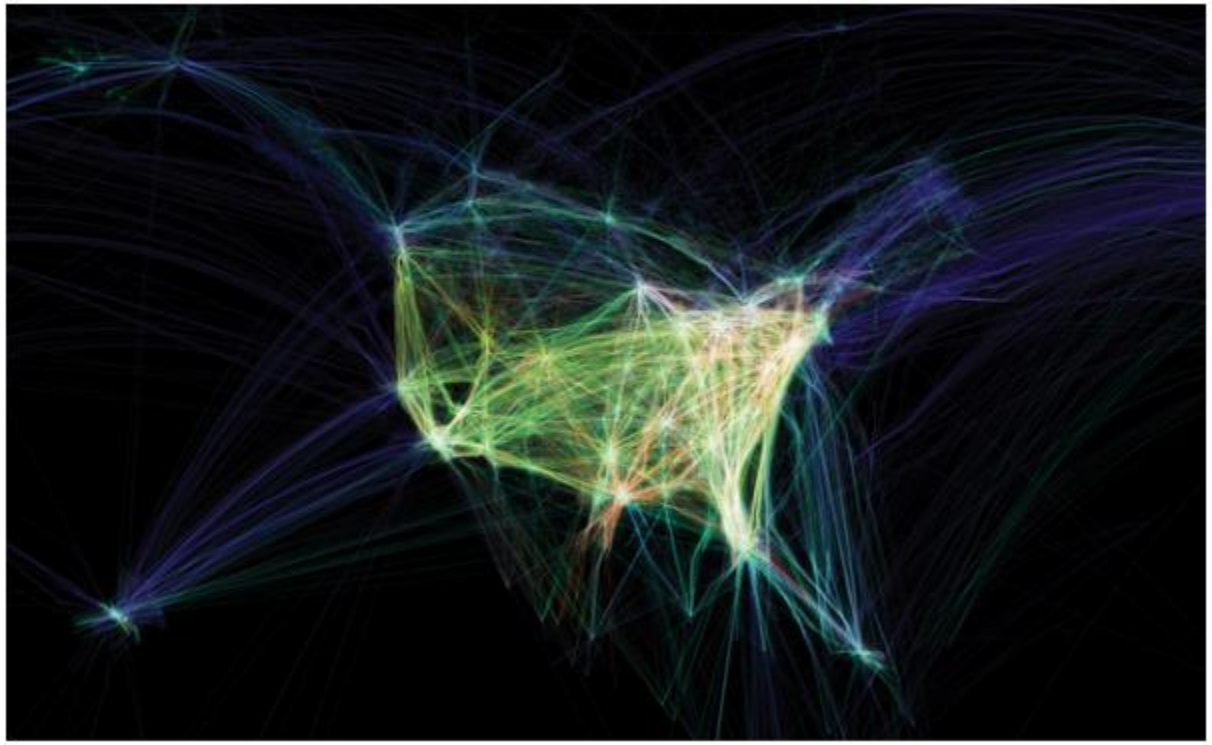
- Variability refers to data whose meaning is constantly changing.
  - Ex: gathering data in the field of language processing.
  - Variance in meaning, in lexicon.
  - Brian Hopkins (a principal analyst at Forrester) cited the supercomputer Watson as a prime example of this. To participate in the gameshow Jeopardy, Watson had to “dissect an answer into its meaning and […] to figure out what the right question was”.
  - https://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety
  - Words don’t have static definitions, and their meaning can vary wildly in context.

Visualization

- **Big data visualization** refers to the implementation of more contemporary **visualization** techniques to illustrate the relationships within **data**. **Visualization** tactics include applications that can display real-time changes and more illustrative graphics, thus going beyond pie, bar and other charts.

- See Aaron Koblin (data artist)TED talk about *taking real world and community-generated data and using it to reflect on cultural trends and the changing relationship between humans and technology*

https://www.ted.com/talks/aaron_koblin_visualizing_ourselves_with_crowd_sourced_data



Buon divertimento!