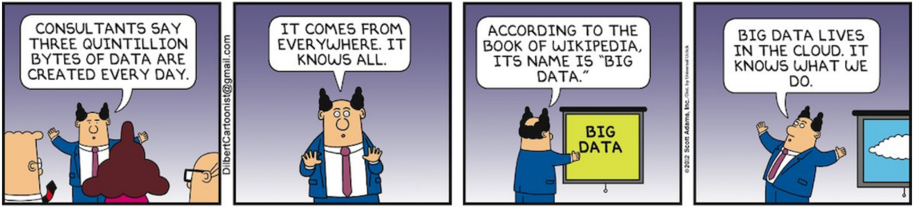
Introduction

Big Data and Cloud Computing (CC4093)

Eduardo R. B. Marques, DCC/FCUP

Big Data? Cloud Computing?


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc.
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

From Dilbert 2012-07-29 by Scott Adams.

So this is a course on **Big Data** *and* **Cloud Computing** ...

To start with, let us then discuss:

- What is Big Data? What makes it *big* ?
- What is Cloud Computing? What makes it *cloudy* ?
- How to they relate?

Big Data? Cloud Computing?

> ***Quintillion bytes of data [...] every day. It comes from everywhere.***

Big data has great **volume** and is generated at great **velocity** and from a **variety** of sources.

> ***It knows all.***

We must **learn** in order to **know**. **How do we learn from big data**?

> ***Big Data lives in the cloud.***

Dilbert's cartoon provides few clues about the "cloud". What do you think the "cloud" is? Why does big data "live" in the cloud? How is the cloud used to handle big data?

Big Data – the NIST definition

The NIST definition:

> ***"Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis."***

Source: "NIST Big Data Interoperability Framework: Volume 1, Definitions", Final Version 1, 2015.

You can find an interesting reference/discussion of other definitions and perspectives on Big Data in this document (Table 1, "Sampling of Definitions Attributed to Big Data", p. 10-11).

The Vs of Big Data

Big Data is "Vig"! :)

Big data stems from (usually the combination of) one or more of the following characteristics, known as the *the four Vs*:

- **Volume** : has a high volume;
- **Velocity** : is generated / needs to be processed at a high rate;
- **Variety** : can originate from multiple sources / be of multiple types;
- **Variability** : exhibits variability in respect to the other Vs or other factors;

In many alternative definitions, two other important "Vs" are also considered:

- **Veracity** : the need to infer (measurable, principled) veracity;
- **Value** : the inherent value of data / its potencial to generate value.

Sources of Big Data

Everyday web/mobile apps

Google

search, maps, mail, …

amazon

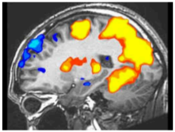
online shopping

facebook

social networking

Science

Hubble telescope



fRMI brain-scan

CERN's Large Hadron Collider

Technologies in rapid progress

IoT

mobile crowd-sensing

smart cities

It comes from everywhere.

Handling Big Data

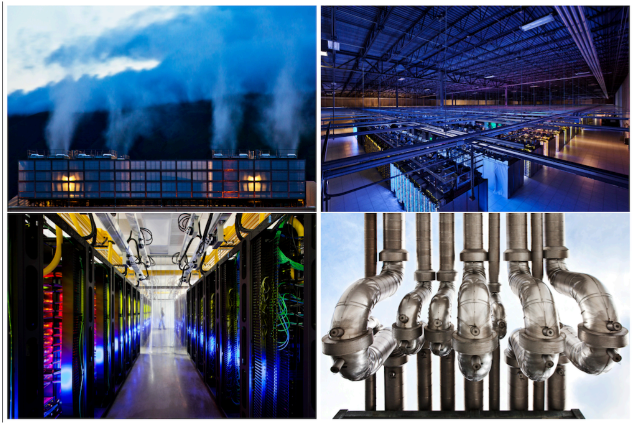
- There is a deluge of Big Data, collected, stored and queried “every second” from “everywhere”.
- Big data holds valuable information from a scientific, societal, and/or economical viewpoint.
- “Raw” big data is not self-explanatory: we must *mine* value from it, to seek for emerging patterns and correlations, validate scientific theories, etc.

Massive data ingestion, storage and processing requires also massive computing infra-structures that are provided through **cloud computing**.

Google searches

> *"Google data centers process an average of 40 million searches per second, resulting in 3.5 billion searches per day and 1.2 trillion searches per year."*

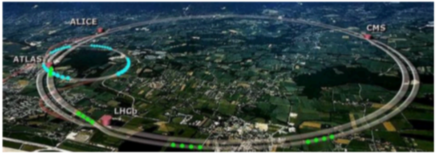
From: Data Center Knowledge; images from: Google Data Centers

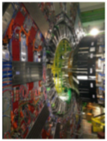


CERN's Large Hadron Collider

Particles collide in the LHC ~1 billion times per second, generating ~1 petabyte of data per second. Most of it is filtered/discarded, still CERN's data centre processes ~1 petabyte of data on average each day.

For more read: CERN Data Centre passes the 200-petabyte milestone, Mélissa Gaillard, CERN News, 2017





Cloud Computing – historical perspective

John McCarthy, an AI pioneer and Turing award winner (1971), said the following at MIT's centennial celebration in 1961:

> ***"Computing may someday be organized as a public utility just as the telephone system is a public utility ... Each subscriber needs to pay only for the capacity he actually uses, but he has access to all programming languages characteristic of a very large system ... Certain subscribers might offer service to other subscribers ... The computer utility could become the basis of a new and important industry"***

The prophecy has been fullfiled! We all now rely on (and to some significant extent even *live*) in the *cloud* !

Cloud computing – the NIST definition

NIST's Big Data definition refers to the need of ***"a scalable architecture for efficient storage, manipulation, and analysis"***. In line with these needs, NIST (also) defines Cloud Computing as:

> ***"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."***

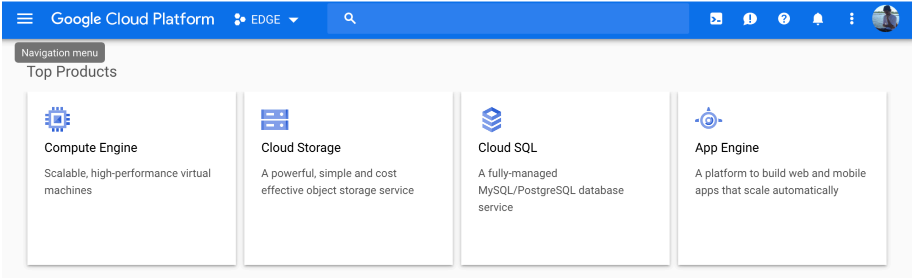
- **What does cloud computing provide**: a shared pool of configurable computing resources.
- **How does it provide it**: on-demand scalable services available through the network, in automated and rapid manner.

Cloud computing - major players

amazon web services

Google Cloud Platform

Microsoft Azure



Top Products

Compute Engine

Scalable, high-performance virtual machines

Cloud Storage

A powerful, simple and cost effective object storage service

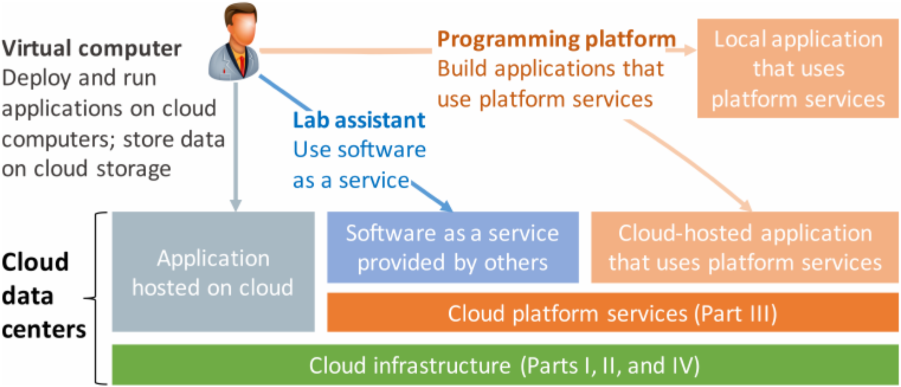
Cloud SQL

A fully-managed MySQL/PostgreSQL database service

App Engine

A platform to build web and mobile apps that scale automatically

Using the cloud for big data applications



Virtual computer
Deploy and run applications on cloud computers; store data on cloud storage

Lab assistant
Use software as a service

Programming platform
Build applications that use platform services

Local application that uses platform services

Cloud data centers

Application hosted on cloud

Software as a service provided by others

Cloud-hosted application that uses platform services

Cloud platform services (Part III)

Cloud infrastructure (Parts I, II, and IV)

Image source: Cloud Computing for Science and Engineering, Creative Commons License