Big Data: noves eines i estratègies per a la gestió de grans volums de dades

Big data i valorització de les dades

Ajuntament de Barcelona
Barcelona Activa

Barcelona Activa Cibernàrium

El Cibernàrium és l'espai de Barcelona Activa-Ajuntament de Barcelona per a la formació i la divulgació tecnològica.

Situat a la 1a planta de l'edifici MediaTIC, ofereix una àmplia oferta formativa per donar resposta a les demandes en formació TIC dels professionals i pimes de la ciutat.

Cibernàrium
Edifici Media TIC
Roc Boronat, 117 – 08018 Barcelona

Accés
Metro: L1 Glòries i Clot / L4 Llacuna
Bus: 71 / 90 / 192
Bicing: 143 / 342 / 393

Segueix-nos a les xarxes socials

barcelonactiva
elcibernarium
company/barcelona-activa

bcn.cat/cibernarim

Índex

1 DEFINICIÓ OFICIAL DE BIG DATA

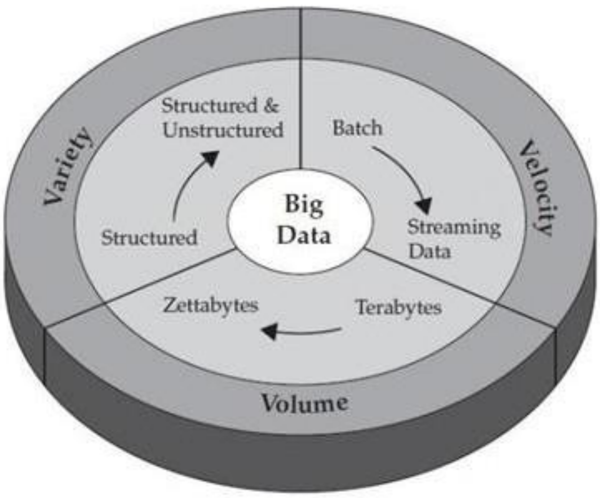
"Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

http://www.gartner.com/it-glossary/big-data/

Gartner, 2012

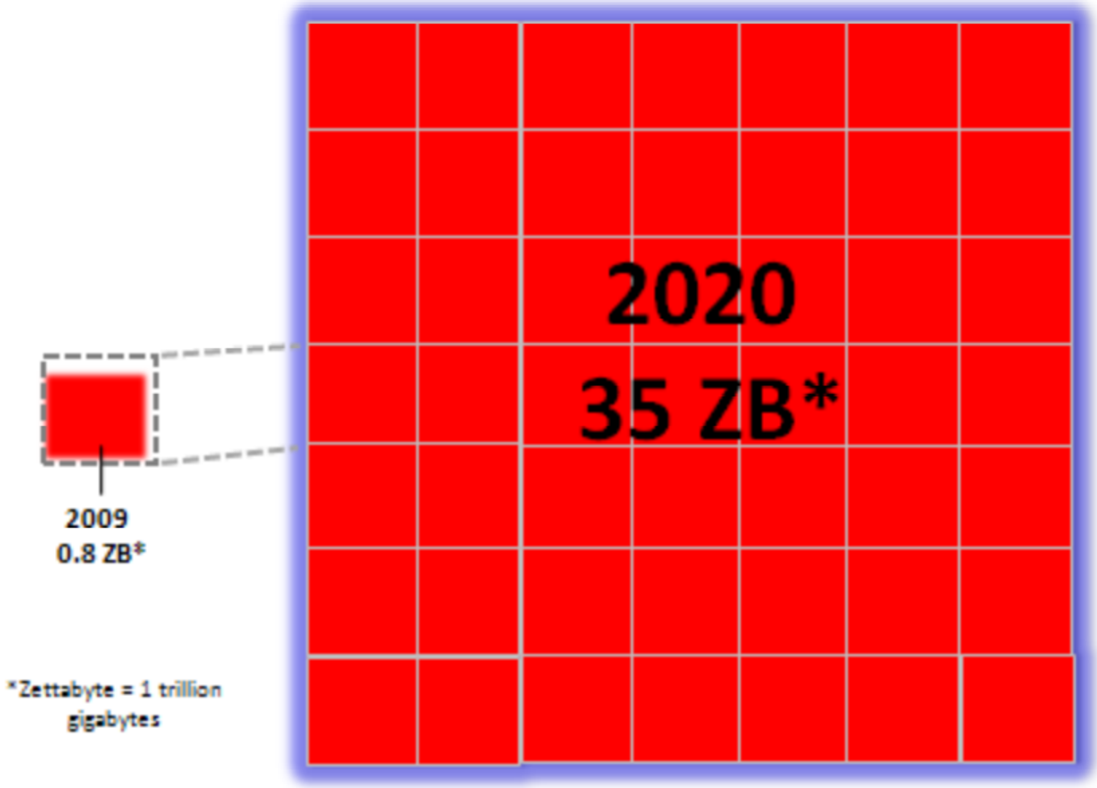
Com podem veure Big Data és bàsicament un concepte econòmic i de gestió. De fet, Gartner ho va definir com: ***el conjunt de propietats d'alt Volum, alta Velocitat i alta Varietat de les dades, que fan imprecindibles la busqueda de noves formes de processament de l'informació eficients i assumibles en cost per així millorar la comprensió de les dades. la presa de decisions i el procés automàtic.***

1.1 LES TRES "V'S"



1.1.1 VOLUM

Fa referència a la grandària dels datasets a manipular. Actualment és habitual haver de processar quantitats de dades en l'escala dels Gigabytes, Terabytes o superiors de manera que les tècniques d'emmagatzematge i processat clàssiques són viables.



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Volum, probablement la característica principal amb la que tothom associa el Big Data. D'acord amb un estudi de la universitat d'Stanford i EMC (2010), la producció de dades es dobla cada 40 mesos, això vol dir que es generen més dades en un dia que els que han existit en els últims 20 anys. Aquest estudi també calcula que a principis del 2010 portàvem acumulats 0.8ZB de dades digitals al món i aquestes en multiplicaran per 40 en 10 anys.

1.1.2 VELOCITAT

Es refereix no només a l'alta freqüència amb què es generen noves dades, sinó a la necessitat de donar resposta a la informació en temps real.

La velocitat en les dades diferencia principalment les tipologies d'anàlisi de dades que tenim actualment.

- El processament per lots es refereix a l'execució dels treballs sense intervenció manual. Els treballs s'estableixen de manera que puguin ser completats sense cap interacció humana sobre volums de dades històriques emmagatzemades. Per exemple, un programa que llegeix un arxiu gran i genera automàticament un informe, això és un tipus de treball per lots.

- El processament interactiu es refereix a programari que rep la intervenció humana en la forma d'ordres o dades. La interacció pot ser directa o indirecta, a través d'una línia de comandaments de la interfície, interfície gràfica d'usuari o un sensor.
- El processament en temps real garanteix una resposta dins dels límits estrictes de temps. La latència es refereix al temps transcorregut entre una entrada rebuda i la corresponent sortida visible. En el món big data, la baixa latència es refereix al fet que el retard es redueix al mínim, i sembla i se sent instantània o temps-real.

1.1.3 VARIETAT

Es refereix a la naturalesa diversa de la informació a manejar. Venim d'informació estructurada que encaixava perfectament en el model relacional però ara ens trobem amb informació semi i des-estructurada (vídeo, àudio, imatges, xarxes socials, etc.) que requereix de nous mètodes de persistència i consulta.

La veritat de dades es classifica en tres tipologies:

- Dades estructurades (Structured Data): Dades que tenen ben definits la seva longitud i el seu format, com les dates, els números o les cadenes de caràcters. S'emmagatzemen en taules. Un exemple són les bases de dades relacionals i els fulls de càlcul.
- Dades no estructurades (Unstructured Data): Dades en el format tal com van ser recol·lectats, no tenen un format específic. No es poden emmagatzemar dins d'una taula ja que no es pot desgranar la seva informació a tipus bàsics de dades. Alguns exemples són els PDF, documents multimèdia, e-mails o documents de text, però també tots els documents *media*.
- Dades semiestructurades (Semistructured Data): Dades que no es limiten a camps determinats, però que conté marcadors per separar els diferents elements. És una informació poc regular com per ser gestionada d'una forma estàndard. Aquestes dades posseeixen els seus propis metadades semiestructurados16 que descriuen els objectes i les relacions entre ells, i poden acabar sent acceptats per convenció. Un exemple és l'HTML, l'XML o el JSON.

1.2 CONCLUSIÓ BIG DATA

Un cop coneguda la definició més teòrica podem treure la conclusió que **tenim un problema de Big Data quan el volum, la velocitat i/o la varietat de les nostres dades fa impossible tractar-les amb les tecnologies informatiques (software i hardware) habituals.**

1.3 PUBLICACIONS DE REFERÈNCIA DEL MÓN BIG DATA

1.3.1 PAPERS OF 2022

2022- The bearable Lightness of Big Data: Towards Massive Public Datasets in Scientific Machine Learning

2022- Big Data and Education: using big data analítics in Language learning

2022- Fitting Semiparametric Cumulative Probability Models for Big Data

1.3.2 PAPERS OF 2021

2021- Toeplitz Least Squares Problems, Fast Algorithms and Big Data

2021- Intelligent Route Planning for Electric Buses Based on urban Big Data

2021- Differential Privacy in Privacy-Preserving Big Data and Learning: Challenge and Opportunity

1.3.3 PAPERS OF 2020

2020 – SDN helps Big Data to optimize access to data

2020 – Toward Compact Data from Big Data

2020 – Big Data

1.3.4 PAPERS OF 2019

2019- Next- Generation Big Data Federation Access Control: A Reference Model

2019- Big Data in Cloud Computing Review and Opportunities

2019- Is Big Data Performance Reproducible in Modern Cloud Networks?

1.3.5 PAPERS OF 2018

2018 – Two- Phase Dynamic Throughput Optimization Model for Big Data Transfers

2018 – Greening Big Data Networks: The Impact of Veracity

2018 – Resource Management and Scheduling for Big Data Applications in Cloud Computing Environments

1.3.6 PAPERS OF 2017

2017 – A Real-Time Autonomous Highway Accident. Detection Model Based on Big Data Processing and Computational Intelligence

2017 – Big Data and Fog Computing

2017 – An Artificial Neural Network – based Stock Trading System Using Technical Analysis and Big Data Framework

1.3.7 PAPERS OF 2016

2016 – Fog Computing for Geospatial Big Data Analytics

2016 – Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples

2016 – Understanding Deep Convolutional Networks

1.3.8 PAPERS OF 2015

2015 – A Neural Algorithm of Artistic Style

2015 – Deep Image: Scaling up Image Recognition

2015 – Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

2015 – Deep Speech: Scaling up end-to-end speech recognition

2015 – Fast Convolutional Nets With fbfft: A GPU Performance Evaluation

2015 – G-OLA: Generalized On-Line Aggregation for Interactive Analysis on Big Data

2015 – Giraffe: Using Deep Reinforcement Learning to Play Chess

2015 – Hidden Technical Debt in Machine Learning Systems

2015 – Klout Score: Measuring Influence Across Multiple Social Networks

2015 – Large-scale cluster management at Google with Borg

2015 – Machine Learning Classification over Encrypted Data

2015 – Machine Learning Methods for Computer Security

2015 – Neural Networks with Few Multiplications

2015 – Self-Repairing Disk Arrays

2015 – Spark SQL: Relational Data Processing in Spark

2015 – SparkNetwork: Training Deep Network in Spark

2015 – Succinct: Enabling Queries on Compressed Data

2015 – Taming the Wild: A Unified Analysis of HOGWILD!-Style Algorithms

2015 – The Missing Piece in Complex Analytics: Low Latency, Scalable Model Management and Serving with Velox

2015 – Trill: A High-Performance Incremental Query Processor for Diverse Analytics

2015 – Twitter Heron: Stream Processing at Scale

1.3.9 PAPERS OF 2014

2014 – 3D Object Manipulation in a Single Photograph using Stock 3D Models

2014 – A Partitioning Framework for Aggressive Data Skipping

2014 – A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data

2014 – A Self-Configurable Geo-Replicated Cloud Storage System

2014 – All File Systems Are Not Created Equal: On the Complexity of Crafting Crash-Consistent Applications

2014 – Arrakis: The Operating System is the Control Plane

2014 – Automatic Construction of Inference-Supporting Knowledge Bases

2014 – Bayesian group latent factor analysis with structured sparse priors

2014 – Chinese Open Relation Extraction for Knowledge Acquisition

2014 – Coordination Avoidance in Database Systems

2014 – DeepFace: Closing the Gap to Human-Level Performance in Face Verification

2014 – Diagram Understanding in Geometry Questions

2014 – Discourse Complements Lexical Semantics for Non-factoid Answer Reranking

2014 – Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

2014 – Eidetic Systems

2014 – Execution Primitives for Scalable Joins and Aggregations in Map Reduce

2014 – Extracting More Concurrency from Distributed Transactions

2014 – f4: Facebookâ□™s Warm BLOB Storage System

2014 – Fast Databases with Fast Durability and Recovery Through Multicore Parallelism

2014 – Fastpass: A Centralized "Zero-Queue" Datacenter Network

2014 – First-person Hyper-lapse Videos

2014 – GloVe: Global Vectors for Word Representation

2014 – GraphX: Graph Processing in a Distributed Dataflow Framework

2014 – Guess Who Rated This Movie: Identifying Users Through Subspace Clustering

2014 – In Search of an Understandable Consensus Algorithm

2014 – Learning Everything about Anything: Webly-Supervised Visual Concept Learning

2014 – Learning to Solve Arithmetic Word Problems with Verb Categorization

2014 – Log-structured Memory for DRAM-based Storage

2014 – Logical Physical Clocks and Consistent Snapshots in Globally Distributed Databases

2014 – MapGraph: A High Level API for Fast Development of High Performance Graph Analytics on GPUs

2014 – Mesa: Geo-Replicated, Near Real-Time, Scalable Data Warehousing

2014 – Modeling Biological Processes for Reading Comprehension

2014 – Orca A Modular Query Optimizer Architecture for Big Data

2014 – Pigeon: A Spatial MapReduce Language

2014 – Project Adam: Building an Efficient and Scalable Deep Learning Training System

2014 – Quantum Deep Learning

2014 – R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics

2014 – Salt: Combining ACID and BASE in a Distributed Database

2014 – Scalable Object Detection using Deep Neural Networks

2014 – Sequence to Sequence Learning with Neural Networks

2014 – Show and Tell: A Neural Image Caption Generator

2014 – Simple Testing Can Prevent Most Critical Failures: An Analysis of Production Failures in Distributed Data-Intensive Systems

2014 – The Mystery Machine: End-to-end Performance Analysis of Large-scale Internet Services

2014 – The Trill Incremental Analytics Engine

1.3.10 PAPERS OF 2013

2013 – A Demonstration of SpatailHadoop: An Efficient MapReduce Framework for Spatial Data

2013 – A Lightweight and High Performance Monolingual Word Aligner

2013 – Answer Extraction as Sequence Tagging with Tree Edit Distance

2013 – Automatic Coupling of Answer Extraction and Information Retrieval

2013 – CG_Hadoop: Computational Geometry in MapReduce

2013 – Consistency-Based Service Level Agreements for Cloud Storage

2013 – Dimension Independent Matrix Square using MapReduce

2013 – Druid A Real-time Analytical Data Store

2013 – Efficient Estimation of Word Representations in Vector Space

2013 – Event labeling combining ensemble detectors and background knowledge

2013 – Everything You Always Wanted to Know About Synchronization but Were Afraid to Ask

2013 – F1: A Distributed SQL Database That Scales

2013 – Fast Training of Convolutional Networks through FFTs

2013 – GraphX: A Resilient Distributed Graph System on Spark

2013 – HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality 2013 Estimation Algorithm

2013 – MillWheel: Fault-Tolerant Stream Processing at Internet Scale

2013 – MLbase: A Distributed Machine-learning System

2013 – Naiad: A Timely Dataflow System

2013 – Omega: flexible, scalable schedulers for large compute clusters

2013 – Online, Asynchronous Schema Change in F1

2013 – Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices

2013 – Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

2013 – Rich feature hierarchies for accurate object detection and semantic segmentation

2013 – Scalable Progressive Analytics on Big Data in the Cloud

2013 – Scaling Memcache at Facebook

2013 – Scuba: Diving into Data at Facebook

2013 – Semi-Markov Phrase-based Monolingual Alignment

2013 – Shark: SQL and Rich Analytics at Scale

2013 – Some Improvements on Deep Convolutional Neural Network Based Image Classification

2013 – Sparrow: Distributed, Low Latency Scheduling

2013 – Sparrow: Scalable Scheduling for Sub-Second Parallel Jobs

2013 – TAO: Facebookâ™s Distributed Data Store for the Social Graph

2013 – Toward Common Patterns for Distributed, Concurrent, Fault-Tolerant Code

2013 – Unicorn: A System for Searching the Social Graph

2013 – Warp: Lightweight Multi-Key Transactions for Key-Value Stores

1.3.11 PAPERS OF 2012

2012 – A Few Useful Things to Know about Machine Learning

2012 – A Sublinear Time Algorithm for PageRank Computations

2012 – Avatara: OLAP for Web-scale Analytics Products

2012 – Blink and It's Done. Interactive Queries on Very Large Data

2012 – BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data

2012 – Building high-level features using large scale unsupervised learning

2012 – Dimension Independent Similarity Computation

2012 – Earlybird: Real-Time Search at Twitter

2012 – Fast and Interactive Analytics over Hadoop Data with Spark

2012 – HyperDex: A Distributed, Searchable Key-Value Store

2012 – ImageNet Classification with Deep Convolutional Neural Networks

2012 – Large Scale Distributed Deep Networks

2012 – Large:Scale Machine Learning at Twitter

2012 – Multi-Scale Matrix Sampling and Sublinear-Time PageRank Computation

2012 – Paxos Made Parallel

2012 – Paxos Replicated State Machines as the Basis of a High-Performance Data Store

2012 – Perspectives on the CAP Theorem

2012 – Processing a Trillion Cells per Mouse Click

2012 – Shark: Fast Data Analysis Using Coarse-grained Distributed Memory

2012 – Spanner: Google's Globally-Distributed Database

2012 – Temporal Analytics on Big Data for Web Advertising

2012 – The Unified Logging Infrastructure for Data Analytics at Twitter

2012 – The Vertica Analytic Database- C-Store 7 Years Later

1.3.12 PAPERS OF 2011

2011 – Consistency, Availability, and Convergence

2011 – CrowdDB: Answering Queries with Crowdsourcing

2011 – CrowdDB: Query Processing with the VLDB Crowd

2011 – Fast Crash Recovery in RAMCloud

2011 – Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent

2011 – It's Time for Low Latency

2011 – Matching Unstructured Product Offers to Structured Product Specifications

2011 – Megastore: Providing Scalable, Highly Available Storage for Interactive Services

2011 – Resilient Distributed Datasets- A Fault-Tolerant Abstraction for In-Memory Cluster Computing

2011 – Scarlett: Coping with Skewed Content Popularity in MapReduce Clusters

1.3.13 PAPERS OF 2010

2010 – A Method of Automated Nonparametric Content Analysis for Social Science

2010 – Dapper, a Large-Scale Distributed Systems Tracing Infrastructure

2010 – Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

2010 – Dremel: Interactive Analysis of Web-Scale Datasets

2010 – Finding a needle in Haystack- Facebook's photo storage

2010 – FlumeJava: Easy, Eff¥cient Data-Parallel Pipelines

2010 – Large:scale Incremental Processing Using Distributed Transactions and Notifications

2010 – Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center

2010 – Pregel: A System for Large-Scale Graph Processing

2010 – S4: Distributed Stream Computing Platform

2010 – Spark: Cluster Computing with Working Sets

2010 – The Learning Behind Gmail Priority Inbox

2010 – ZooKeeper: Wait-free coordination for Internet-scale systems

1.3.14 PAPERS OF 2009

2009 – Cassandra – A Decentralized Structured Storage System

2009 – Feature Hashing for Large Scale Multitask Learning

2009 – HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads

2009 – Vertical Paxos and Primary-Backup Replication

1.3.15 PAPERS OF 2008

2008 – Chukwa: A large-scale monitoring system

2008 – Column:Stores vs. Row-Stores- How Different Are They Really?

2008 – PNUTS: Yahoo!Õs Hosted Data Serving Platform

2008 – Top 10 algorithms in data mining

1.3.16 PAPERS OF 2007

2007 – Architecture of a Database System

2007 – Consistent Streaming Through Time: A Vision for Event Stream Processing

2007 – Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks

2007 – Dynamo: Amazon's Highly Available Key-value Store

2007 – Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments

2007 – Life beyond Distributed Transactions: an ApostateÕs Opinion

2007 – Paxos Made Live – An Engineering Perspective

1.3.17 PAPERS OF 2006

2006 – Bigtable: A Distributed Storage System for Structured Data

2006 – Ceph: A Scalable, High-Performance Distributed File System

2006 – Map-Reduce for Machine Learning on Multicore

2006 – The Chubby lock service for loosely-coupled distributed systems

1.3.18 PAPERS OF 2005

2005 – Fast Paxos

1.3.19 PAPERS OF 2004

2004 – Cheap Paxos

2004 – MapReduce: Simplified Data Processing on Large Clusters

1.3.20 PAPERS OF 2003

2003 – Interpreting the Data: Parallel Analysis with Sawzall

2003 – The Google File System

1.3.21 PAPERS OF 2002

2002 – Brewer’s Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services

1.3.22 PAPERS OF 2001

2001 – Chord: A Scalable Peer-to-peer Lookup Service for Internet

Applications

2001 – Paxos Made Simple

2001 – Random Forrest

1.3.23 PAPERS OF 1999

1999 – Pasting Small Votes for Classification in Large Databases and On-Line

1999 – The PageRank Citation Ranking: Bringing Order to the Web

1.3.24 PAPERS OF 1997

1997 – Application-Controlled Demand Paging for Out-of-Core Visualization

Descobreix tot el que Barcelona Activa t´ofereix

Acompanyament durant tot el procés de recerca de feina

bcn.cat/treball

Suport per posar en marxa la teva idea de negoci

bcn.cat/emprenedoria

Impuls a les empreses per ser més competitives

bcn.cat/empresa

Formació tecnològica i gratuïta per a les persones en recerca de feina, emprenedors, empreses...

bcn.cat/cibernarium

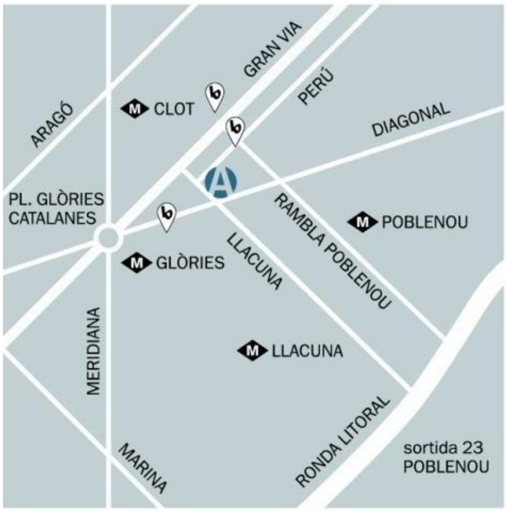
Barcelona Activa

Seu Central

Llacuna, 162-164
08018 Barcelona
+34 934 019 777
barcelonactiva.cat

Accés

Metro: L1 Glòries i Clot / L2 Clot
Bus: 7 / B21 / H12 / 60 / 92 / 192
Rodalies: R1 i R2 Clot
Tramvia: T4 Ca l´Aranyó / T5 i T6 Can Jaumandreu
Bicing: 42 / 133 / 132



Segueix-nos a les xarxes socials:

- barcelonactiva
- barcelonactiva
- bcn_empresa
- elcibernarium
- company/barcelona-

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional