

BIG DATA EUROPE

Empowering Communities with Data Technologies



CESSDA sets the stage for data infrastructure of the future

IASSIST 2016 conference, 2 June 2016



BIG DATA EUROPE

HTTP://WWW.BIG-DATA-EUROPE.EU/

Integrating Big Data, Software & Communities for Addressing Europe's Societal Challenges

Info

- “BigDataEurope - Empowering Communities with Data Technologies”
- 3-year Horizon 2020 CSA project
- integrated stack of tools to manipulate, publish and use large-scale data resources.

Summary



Two clearly defined coordination and support measures:

- Coordination: **Engaging with a diverse range of stakeholder groups** representing particularly the Horizon 2020 societal challenges Health, Food & Agriculture, Energy, Transport, Climate, Social Sciences and Security; Collecting requirements for the ICT infrastructure needed by data-intensive science practitioners tackling a wide range of societal challenges; covering all aspects of publishing and consuming semantically interoperable, large-scale data and knowledge assets;
- Support: **Designing, realizing and evaluating a Big Data Aggregator platform** infrastructure that meets requirements, minimises disruption to current workflows, and maximises the opportunities to take advantage of the latest European RTD developments (incl. multilingual data harvesting, data analytics & visualisation).

BigDataEurope will implement and apply two main instruments to successfully realize these measures:

- Build **Societal Big Data Interest/Community Groups** in the W3C interest group scheme & involving a large number of stakeholders from the Horizon 2020 societal challenges as well as technical Big Data experts;
- Design, integrate and deploy a **cloud-deployment-ready Big Data aggregator platform** comprising key open-source Big Data technologies for real-time and batch processing, such as Hadoop, Cassandra and Storm.

Rationale



- **Show** societal value of Big Data
- **Lower** barrrier for using big data technologies
  - Required effort and resources
  - Limited data science skills
- **Help** establishing cross-lingual/organizational/ domain Data Value Chains

Rationale



CSA Measures

COORDINATION
Stakeholder Engagement
(Requirements Elicitation)

SUPPORT
Design, Realise, Evaluate
Big Data Aggregator Platform

Results

Create and Manage Societal
Big Data Interest Groups

Cloud-deployment ready
Big Data Aggregator Platform

BDE Partners

National and Kapodistrian University of Athens

InfAI Institut für Angewandte Informatik

Fraunhofer IAIS

cessda

SEMANTIC WEB COMPANY linking data to knowledge

SatCen

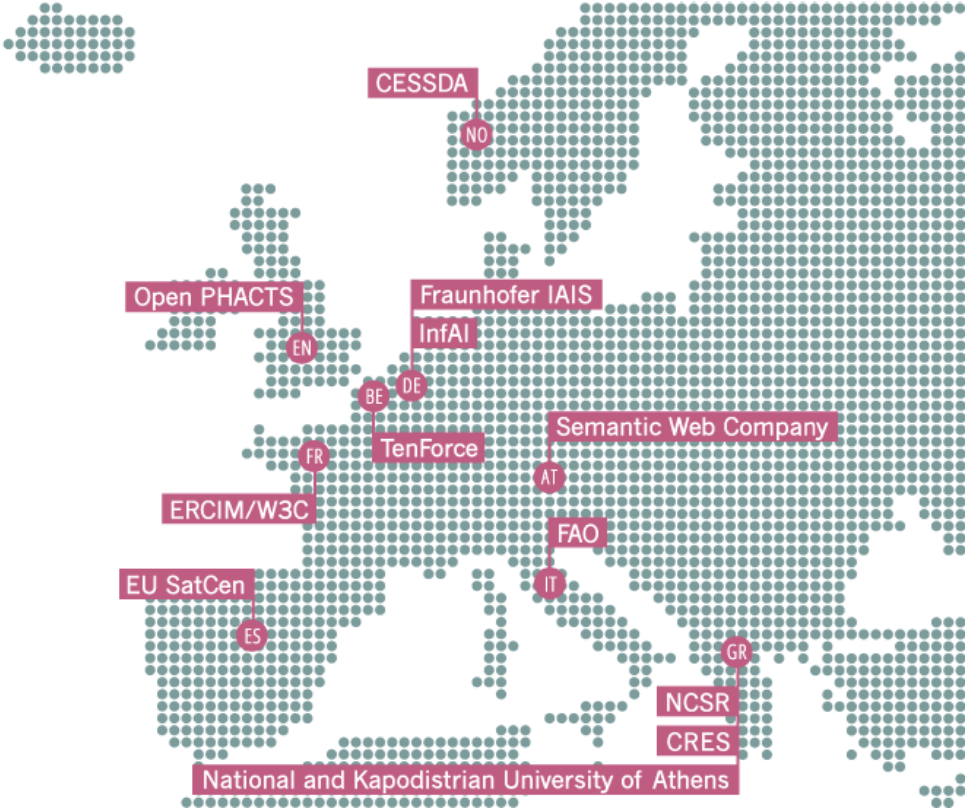
W3C

TenForce The Pragmatic Company

The Open PHACTS Foundation

FAO

ΚΑΠΕ CRES



CESSDA in BDE

- **To coordinate** the Societal Challenge 6 and potential users of big data in the fields of social sciences and humanities (SSH)
- **To build** this interest group, collect requirements, assist the building big data infrastructure access point for SSH

The Motivation – Big Data



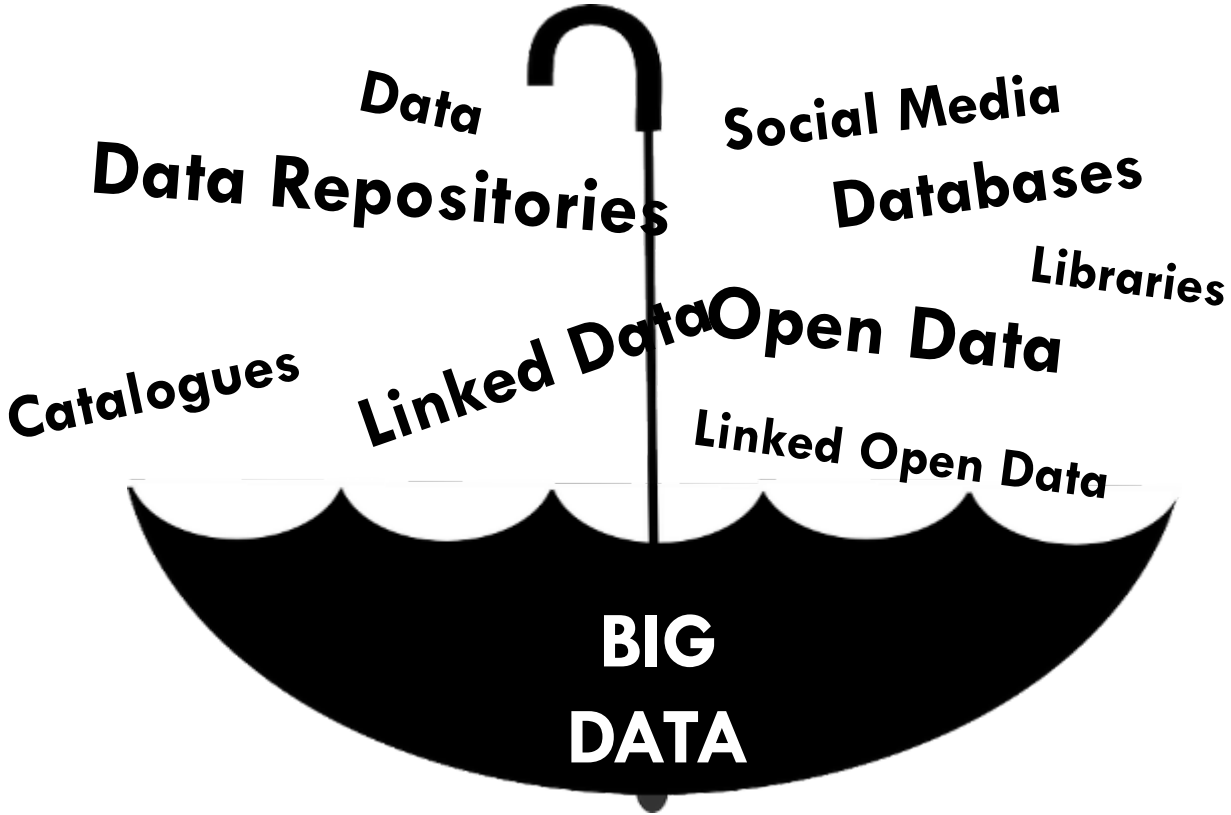
Every day, we create 2.5 quintillion bytes of data — so much that **90% of the data in the world today has been created in the last two years** alone.

This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

This data is **big data.**

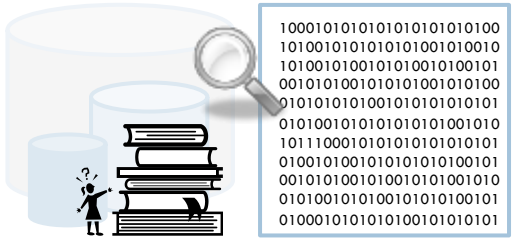
Source: IBM

The Motivation – Big Data

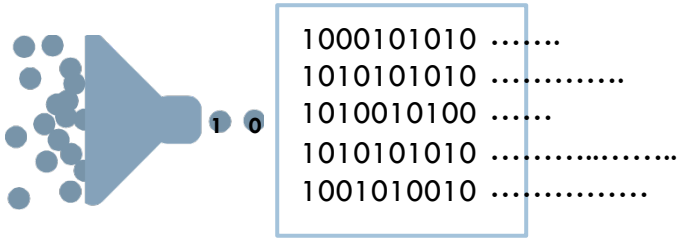


Big Data Dimensions

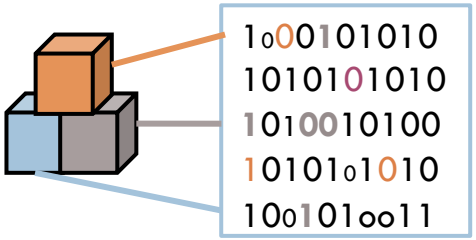
- ***Volume***



- ***Velocity***



- ***Variety***



- ***Veracity!***

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones





Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]





By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



30 BILLION PIECES OF CONTENT
are shared on Facebook every month





400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session





Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

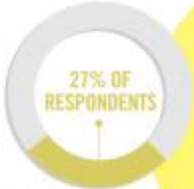
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth



Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions





in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

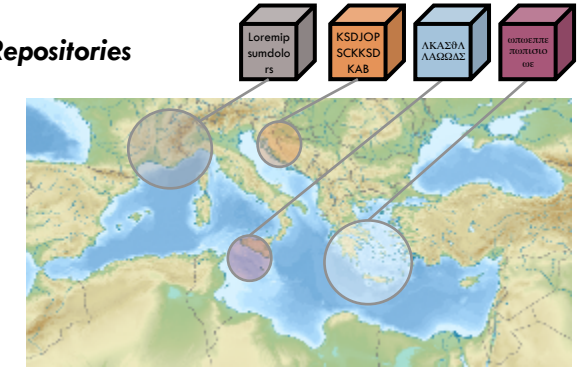


Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Big Data in Europe: Challenges, Opportunities

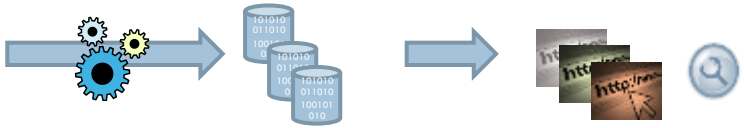
Regional Data Repositories



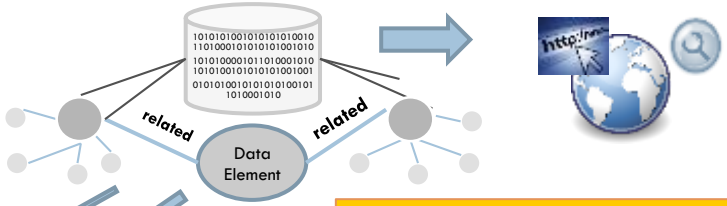
#4: Mashup, Cross-domain Exploitation



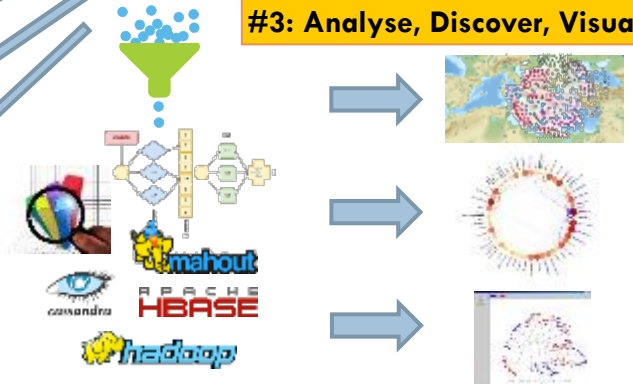
Health
Climate
Energy
Transport
Food
Societies
Security



#2: Interlink, Centralise Access, Explore



#3: Analyse, Discover, Visualize



Big Data in Europe: Obstacles

#1 Big Data "Variety" problem

- Multiple Data Sources
- Required: Integration, Harmonisation



#2 Opening-up Data concerns

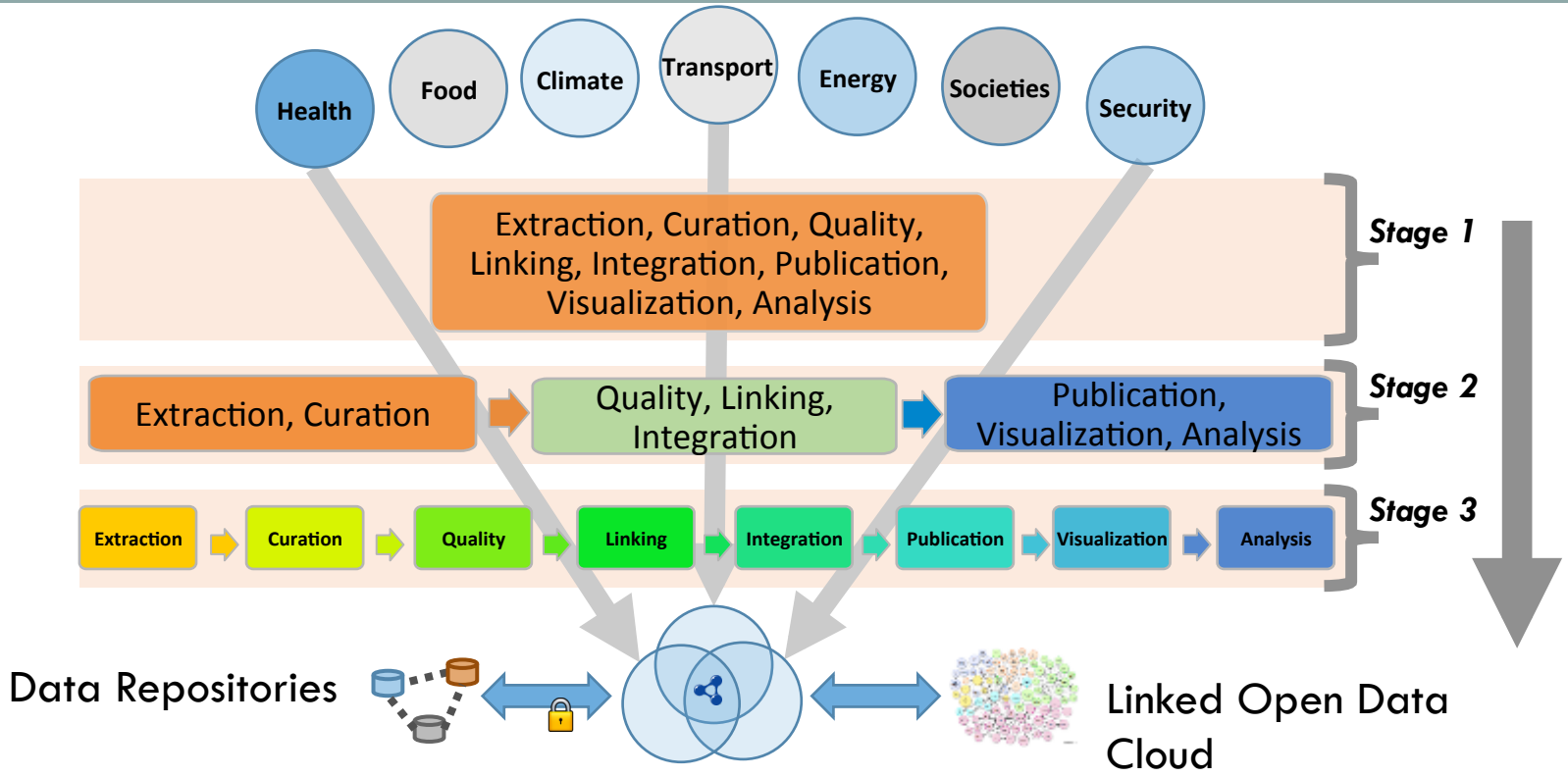
- Loss of control, lack of tracking
- Reservations about large corporations



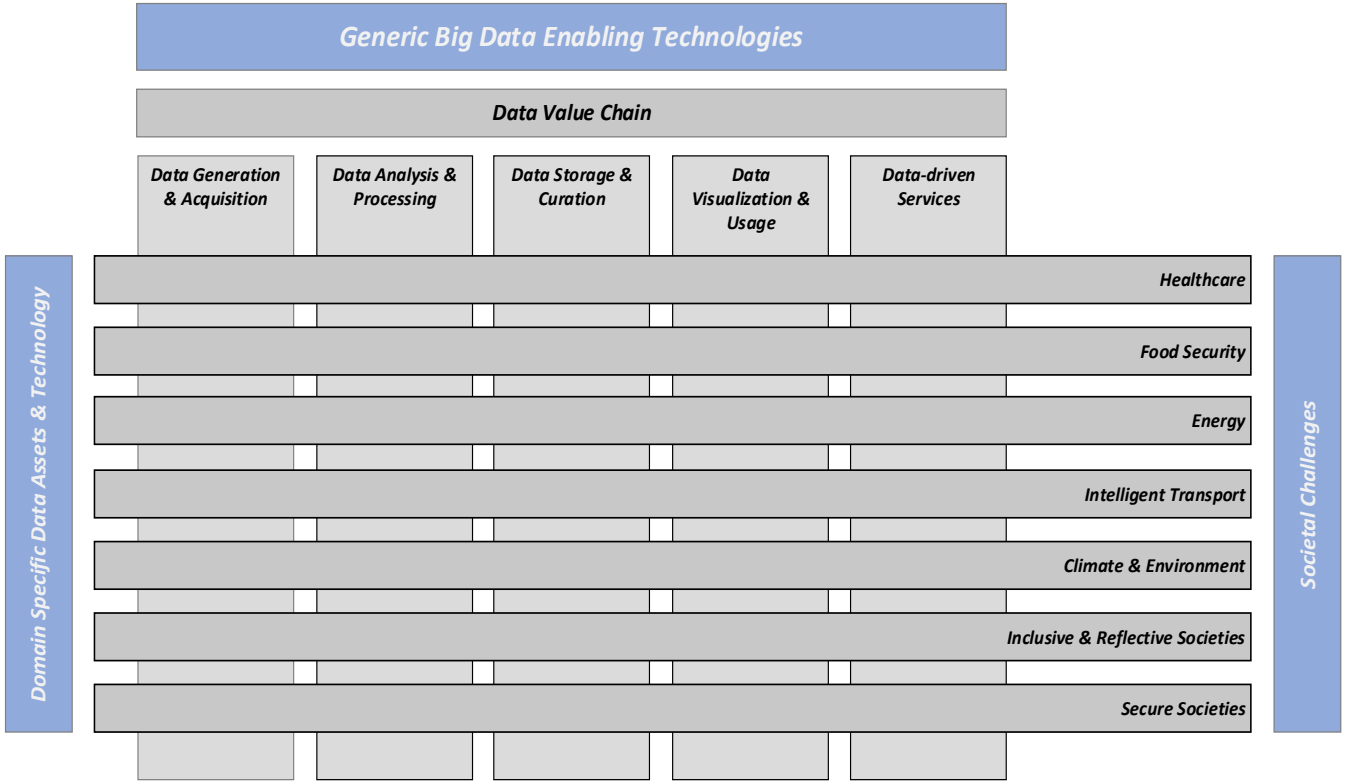
#3 Limited Skills, Training, Technology

- Lack of Data Scientists
- Lack of Generic Architectures, components

Big Data in Europe: Obstacles



Orthogonal Dimensions of Big Data Ecosystems



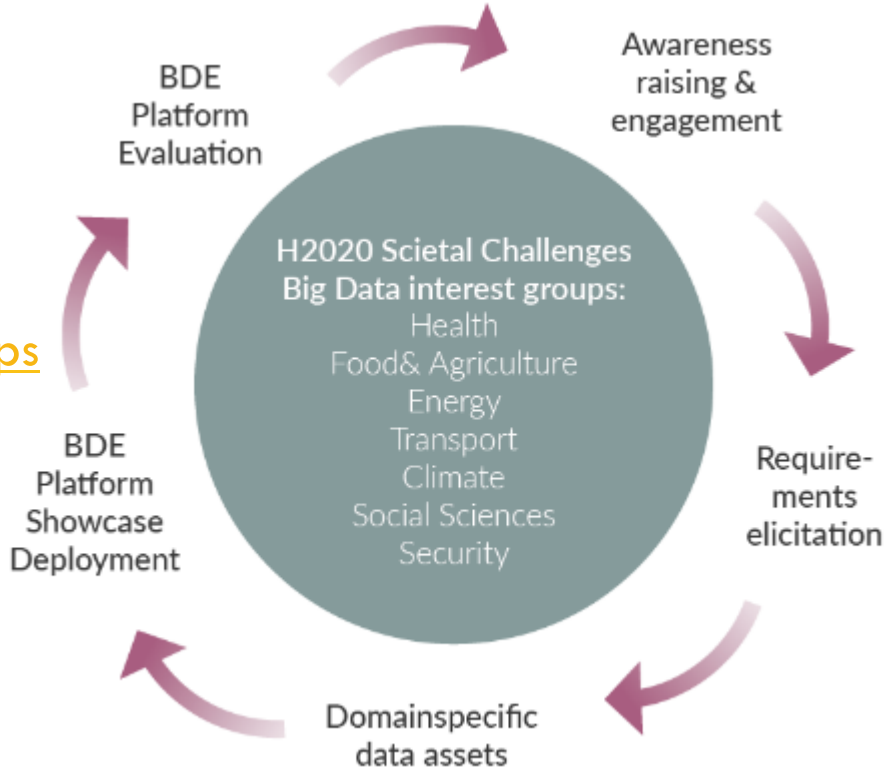
BDE Stakeholder Engagement Approach & Activities

BDE Community Tools – JOIN IN NOW !

- Website: news, events, community, …
- 7 x BDE W3C Community Groups
- 7+1x Mailing Lists
- 7 x SC Workshops/Year = 21 Workshops
- Full set of communication tool-set…

Future Outlook

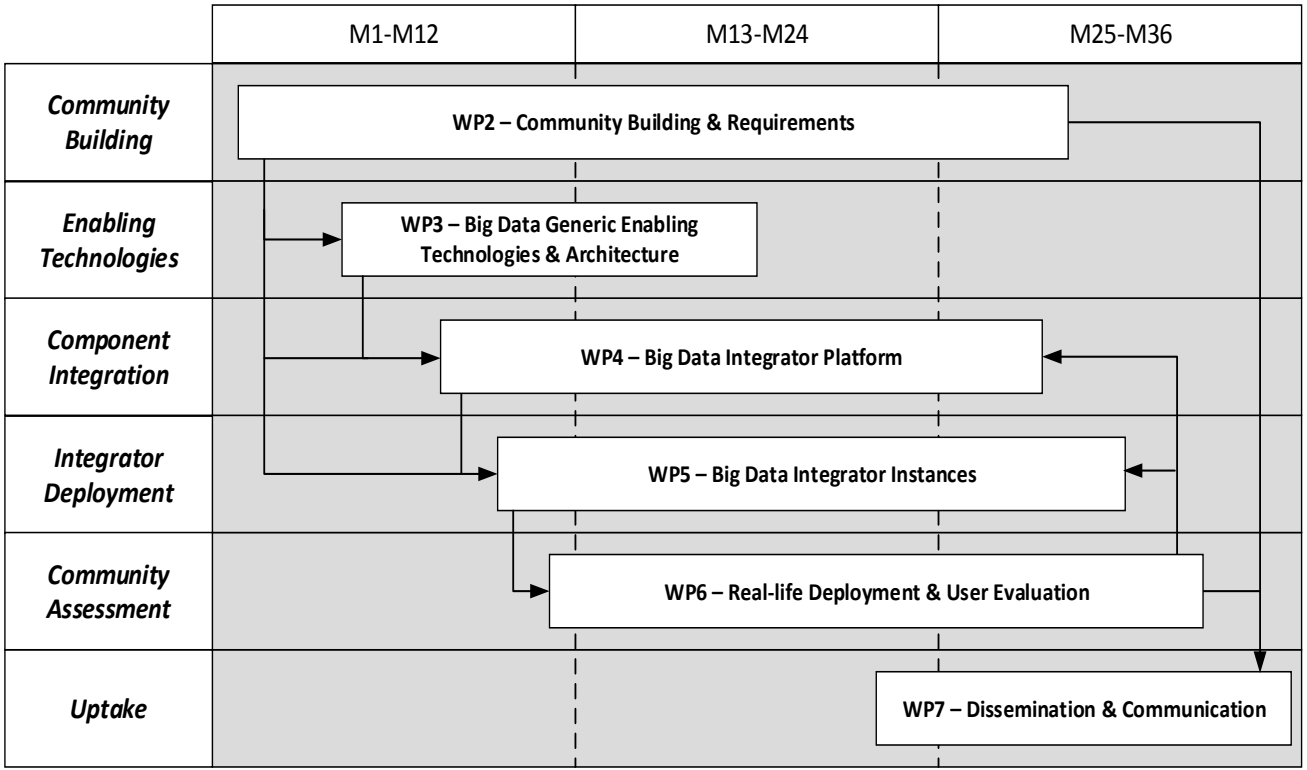
- BDE Aggregator Platform
  - For download / internal use
  - Cloud Version
- Big Data Technology Support Tools



Domains, Focus Areas & Data Assets

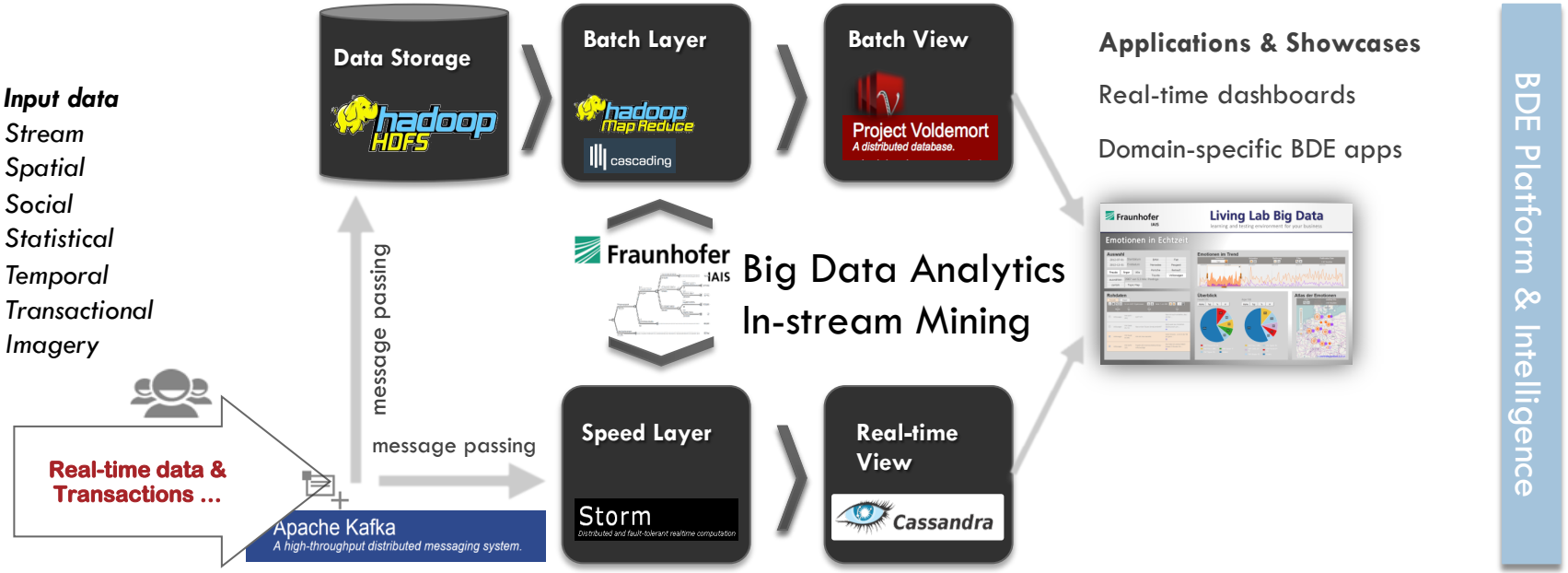
Societal Domain	Preliminary Big Data Focus area	Selected Key Data assets
Life Sciences & Health	Heterogeneous data Linking & integration Biomedical Semantic Indexing & QA	ACD Labs / ChemSpider, ChEBI, ChEMBL, Con-ceptWiki, DrugBank, EN-ZYME, Gene Ontology, GO Annotation, Swis-sProt, UniProt, Wik-iPathways, PubMed, MeSH, Disease Ontology (DO), Joint Chemical Dic-tionary (Jochem), Bio-ASQ datasets
Food & Agriculture	Large-scale distributed data integration	INFOODS, AQUASTAT Green Learning Network (GLN), Agricultural Bibliography Network (ABN), AGRIS, AquaMaps, Fishbase
Energy	Real-time monitoring, stream processing, data analytics, and decision support	European Energy Exchange Data, smart meter measurement data, gas/fuels/energy market/price data, consumption statistics, equipment condition monitoring data)
Transport	Streaming sensor network & geo-spatial data integration	GTFS data, OSM/ LinkedGeoData, MobilityMaps, Transport sensor data, ROSATTE Road safety attributes, European Road Data Infrastructure - EuroRoadS
Climate	Real-time monitoring, stream processing, and data analytics.	European Grid Infrastructure (EGI), Databases hosting atmospheric data. Several software frameworks for simulation, calibration and reconstruction.
Social Sciences	Statistical and research data linking & integration	Federated social sciences data catalogs, statistical data from public data portals and statistical offices (e.g. EuroStats, UNESCO, WorldBank)
Security	Real-time monitoring, stream processing, and data analytics. Image data analysis	Earth Observation data (e.g. Very High Resolution Satellite Imagery acquired from commercial providers and governmental systems) and collateral data for supporting CFSP/CSDP missions and operations, Databases hosting atmospheric Data. Experimental and simulation data concerning dispersion of hazardous substances

Work Packages & Implementation Phases



Blueprint of the Data Aggregator Platform

Lambda Architecture



+ Semantic Layer (Retaining Semantics using LD approach)

Announcements & Pointers….

BIG DATA EUROPE
Empowering Communities with Data Technologies

Website SC6: http://www.big-data-europe.eu/social-sciences/

W3C Community Group SC6: https://www.w3.org/community/bde-societies/

Overall & SC6 Mailing List: http://bit.ly/1K3ZnJ2

Twitter: https://twitter.com/bigdata_europe @BigData_Europe #BigDataEurope

Slideshare: http://slideshare.net/BigData_Europe

flickR: https://www.flickr.com/photos/133018547@N06/

LinkedIN Group: http://bit.ly/1VO5dow

- Thank you!

Ivana Ilijasic Versic, CESSDA AS