LAMBDA - Learning, Applying, Multiplying Big Data Analytics

Project presentation

LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.

Project Funding

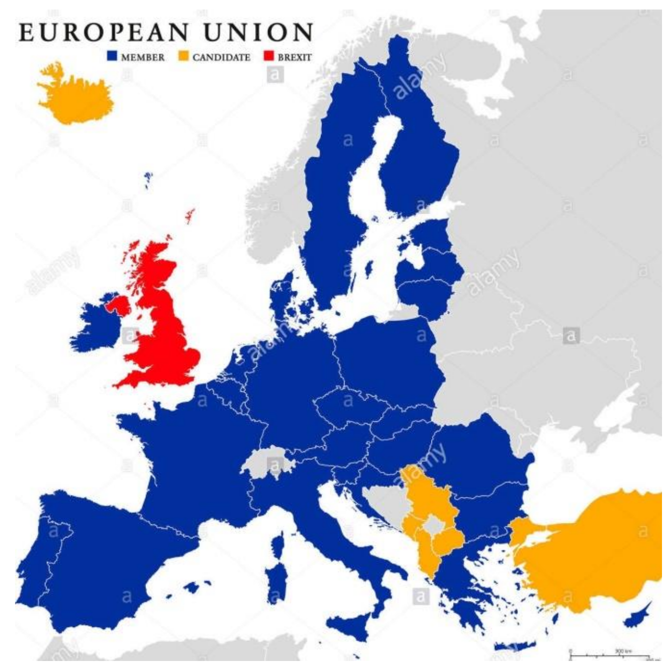
- This project has received funding from the European Union's Horizon 2020 research and innovation programme, GA No 809965
- **Twinning** Coordination and Support Action, [H2020-WIDESPREAD-2016-2017]

- Project Partners
  - Institute Mihajlo Pupin, Serbia (Coordinator)
  - Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
  - Institute for Computer Science - University of Bonn, Germany
  - Department of Computer Science - University of Oxford, UK

Vision and Primary Objectives

institute MIHAILO PUPIN

Strengthening the Human capital and Education, Research and Development capacities of "Mihajlo Pupin" Institute the leading Serbian R&D institution in information and communication technologies in order to serve as a **Big Data & Analytics HUB** that connects and integrates scientists and professionals from the West Balkans and the entire region into the European Research Area.



Decreasing the existing European regional R&I disparity by Fostering excellence in the Big Data Ecosystem areas unlocking and raising the scientific profile of academics institutions from Serbia and the region while contributing to European progress beyond the state-of-the-art of related research and technology, as well as establishing productive and fruitful long-term cooperation.

Specific Objectives

OBJ 1: Strategic Partnership - Establishment and development of productive and fruitful long-term cooperation that continues after project completion

- Sustainable Development Plan for PUPIN (2021-2025)

OBJ 2: Boosting scientific excellence of the linked institutions and capacity building of the widening country and the region in Big Data Analytics and semantics

- Different capacity building activities (Big Data Analytics Summer School)

OBJ 3: Spreading excellence and disseminating knowledge throughout the West Balkan and South-East European countries

- Workshops at International conferences in the region

ICIST2019 ESWC19 CECIIS 30TH EDITION

OBJ 4: Sustainability of research related to key societal challenges (sustainable transport, sustainable energy, security, societal wellbeing) and financial autonomy in the long run

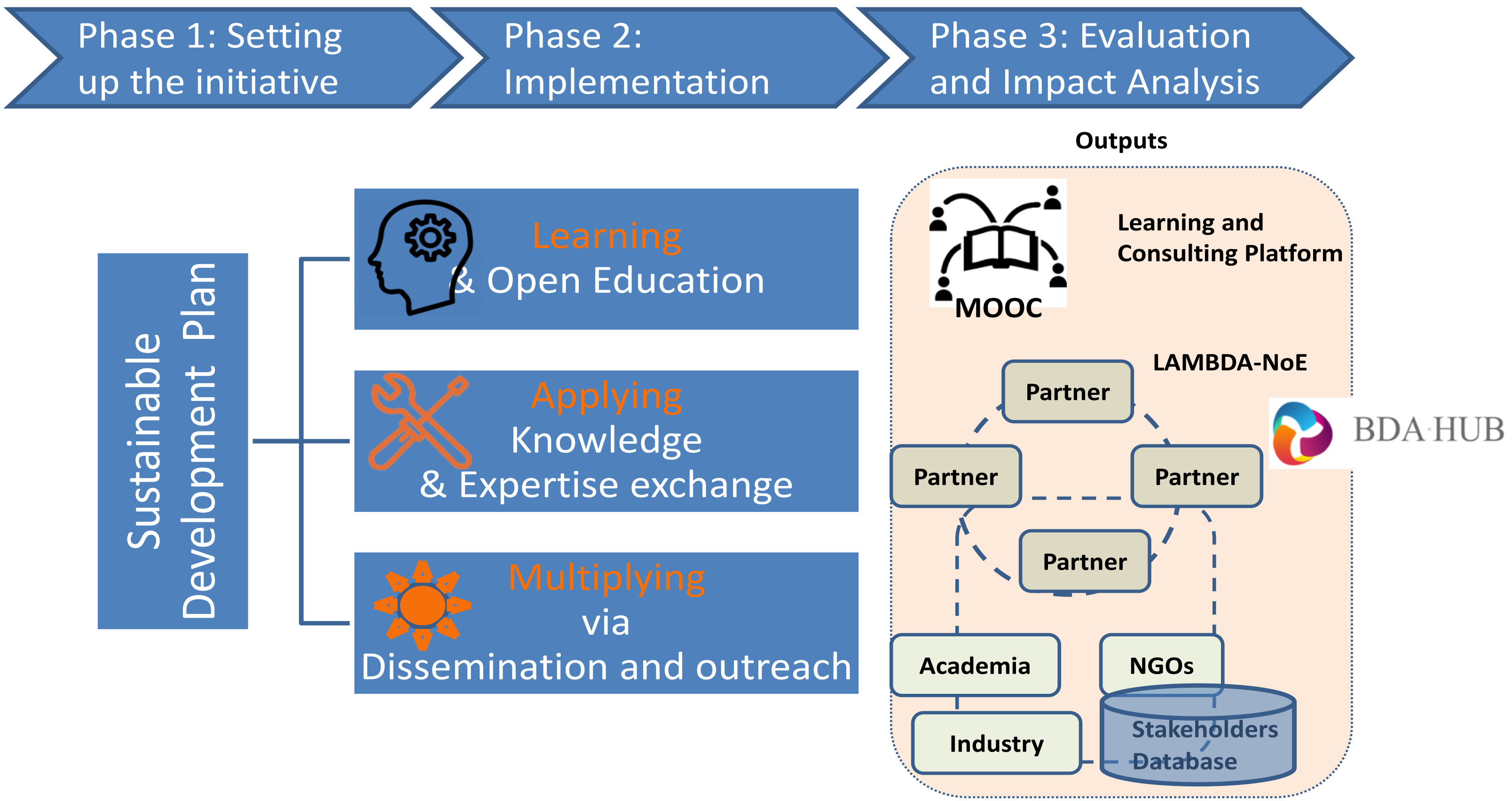
- Brainstorming sessions on key societal challenges

Methodology

Phase 1: Setting up the Initiative and preparing the Twinning Strategy and Action Plan for 2018-2020,

Phase 2: Execution / Implementation and

Phase 3: Closure / Evaluation and Impact Analysis and delivery of the Strategy and Action Plan for 2021-2025.



Key Pillars

Component	Description
Learning & **Open Education**	Knowledge repository as part of the **LAMBDA Learning and Consulting Platform** will be established to facilitate spreading learning materials, as well as exchange of best practice between research institutions from South-Eastern Europe and leading EU partners: • https://project-lambda.org/Learning • https://project-lambda.org/Knowledge-repository/Lectures
Applying **Knowledge** & **Cooperation**	**LAMBDA Experts Exchange Program** for teachers, researchers and developers) will open possibilities for collaborative research on open issues in Big Data related areas: • Industry 4.0 • ICT for Energy
Multiplying **Dissemination and outreach**	Raising awareness about **future trends in Big Data, Emerging Tools and Technologies**, and standards by organization of events at international (e.g. DEXA, ESWC, SEMANTiCS) and regional (e.g. ICIST, ICT Innovations) conferences, organization of the Belgrade Big Data Analytics Summer/Winter School, https://project-lambda.org/Announcement-1 **Sustainable Development Plan for PUPIN (2021-2025)** Strategy development and monitoring activities; Self-assessment of research accomplishments at PUPIN aimed at increasing the shared awareness about the research capacities, primarily human resources.

Open Education (June 2019)

- Enterprise Knowledge Graphs (University of Oxford)
  - Introduction to Knowledge Graphs
  - Extraction for Knowledge Graphs
  - Reasoning in Knowledge Graphs
- Semantic Big Data Architectures (Fraunhofer Institute)
  - Introduction to Big Data Architecture
  - Big Data Solutions in Practical Use-cases
  - Distributed Big Data Frameworks
- Smart Data Analytics (University of Bonn)
  - Distributed Big Data Libraries
  - Distributed Semantic Analytics I
  - Distributed Semantic Analytics II

Staff Exchange Activities

- ✓ Analysis of Big Data Tools
- ✓ Writing position papers / proposals
- ✓ Writing joint papers
- ✓ Organizing events
- ✓ **Other knowledge transfer instruments**

✓https://project-lambda.org/Past-Events
✓https://project-lambda.org/Staff-Exchange

LAMBDA Platform

Learning, Applying, Multiplying Big Data Analytics

Home Project Methodology eLearning News & Events Results Join Us

Stakeholders Section Summer School Knowledge Repository

- Invited Teachers - Keynotes - Lectures
- Education & Research Organizations
- Faculties/ Departments
- Industry
- Public Administration & NGO

Home » Big Data Analytics School 2019

- My account
- Log out

Posted on: Fri, 10/26/2018 - 15:37 **By:** valentina.janev

One of the objectives of the project is organization of a **Big Data Analytics Summer School in E**
2020. The event will bring together researchers and professionals from respectable EU Universit
stakeholders from the West Balkan countries to discuss state-of-the-art in Big Data research and ta
Analytics Summer School will allow the invited PhD students (participants of the School from Se
learn about the newest technologies and trends in this and related fields.

The 3-day event is scheduled as follows

- 1st day LAMBDA Research-Industry Forum - Keynotes + Presentations from Companies
- 2nd day Big Data Analytics Summer School - Invited Lectures and Lectures from LAMBDA partners (UBO and UOXF)
- 3nd day Big Data Analytics Summer School - Lectures from LAMBDA partners (UBO and UOXF)

Advisory Board Meeting: TIB, OntoText, UPM, SZTAKI, UVT

bda-school@mail.project-lambda.org

		1st Day	2nd Day	3rd Day	
8.30		Registration	Registration	Registration	

LAMBDA - Learning, Applying, Multiplying Big Data Analytics

Big Data Analytics

State-of-the-art Review

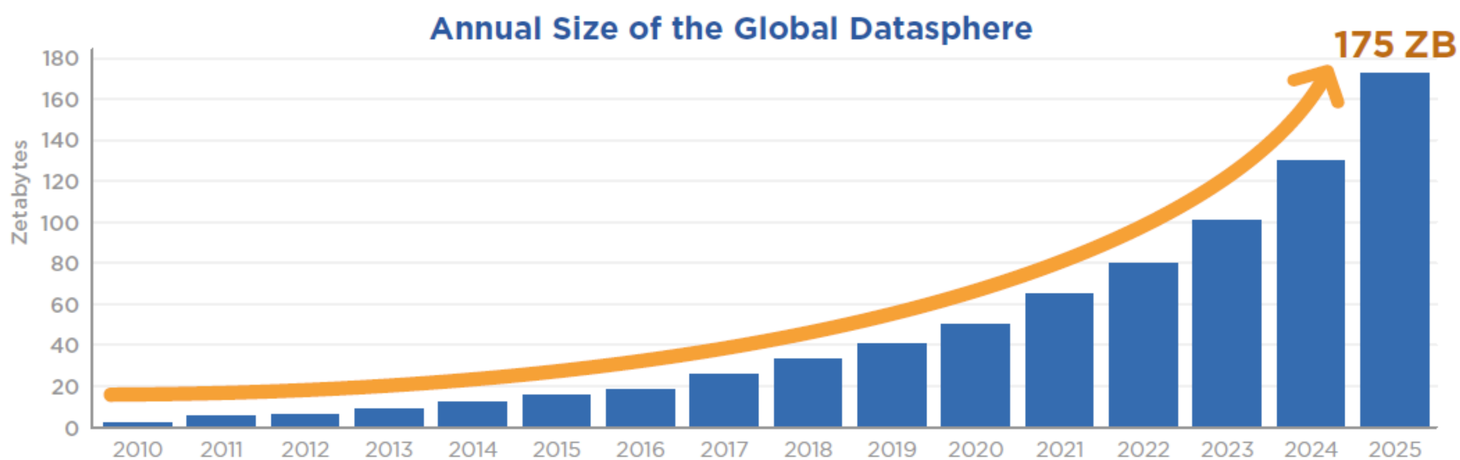
LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.

Big Data

- **Big Data** is used more as a **buzzword** then a **precisely defined** scientific **object** or **phenomena**
- Generally used when referring to **data loads** that the **modern-day IT infrastructure** cannot cope with at all or **in an efficient manner**
- More precisely, Big data is usually used when referring to **data sets** that are sized in the **order of magnitude** of **exabytes** (10^{18} B) or greater (10^{21} ZB)
- [International Data Corporation](), Expect 175 zettabytes of data worldwide by 2025

Figure 1 – Annual Size of the Global Datasphere



Nature of Big Data

Big data is often characterized trough **so-called V's of Big data** that capture its complex nature

- Volume – **amount** of data that has to be captured, stored, processed and displayed
- Velocity – the **rate** at which the data is being generated, or analyzed
- Variety – **differences in** data **structure** (format) or **differences in** data **sources** themselves

3V's

- Veracity – truthfulness (**uncertainty**) of data
- Validity – **suitability** of the selected dataset for a given application

5V's

- Volatility – **temporal validity** and fluency of the data
- Value – (useful) **information** extracted from the data

7V's

- Visualization – properly **displaying** and showcasing information
- Vulnerability – **security** and **privacy** concerns associated
- Variability – the **changing meaning** of data

10V's

Big Data challenges

The **core technological challenges** working with Big data that **stem from its complex nature** are:

- Heterogeneity – differences in structure
- Uncertainty – data reliability
- Scalability – sizing the workflow and infrastructure
- Timeliness – real-time requirements
- Fault tolerance – sensitivity to errors
- Data security – privacy issues, data leaks
- Visualization – displaying of information

	Storing	Processing	Analytics	Visualization
Heterogeneity	+	+		
Uncertainty of data		+	+	
Scalability	+	+	+	
Timeliness	+	+	+	
Fault tolerance		+	+	
Data security	+	+		
Visualization				+

Big Data Landscape

Vertical Apps
- PREDICTIVE POLICING
- bloomreach
- MYRRIX

Log Data Apps
- splunk>
- loggly
- sumologic

Ad/Media Apps
- rocketfuel
- collective[i]
- bluefin
- Recorded Future
- LuckySort
- Media Science
- TURN
- DataXu

Data As A Service
- factual.
- kaggle
- knoema beta
- GNIP
- DATASIFT
- Windows Azure Marketplace
- INRIX
- LexisNexis
- Space Curve
- LOQATE Everything Location

Business Intelligence
- ORACLE | Hyperion
- SAP
- Business Objects
- RJMetrics
- Microsoft Business Intelligence
- IBM
- COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- Chart.io
- GoodData

Analytics and Visualization
- tableau
- Palantir
- OPERA
- metaLayer
- METAMARKETS
- TERADATA ASTER
- dataspora
- centrifuge
- sas
- TIBCO
- KARMASPHERE
- panopticon Real-Time Visual Data Analysis
- pentaho
- Datameer
- ClearStory
- CIRRO
- platfora
- alteryx
- visual.ly
- AYATA

Analytics Infrastructure
- Hortonworks
- VERTICA An HP Company
- MAPR
- cloudera
- INFOBRIGHT
- ParAccel
- EMC²
- GREENPLUM
- NETEZZA
- kognitio
- DATASTAX
- EXASOL
- calpont

Operational Infrastructure
- Couchbase
- 10gen | the MongoDB company
- TERADATA
- HADAPT
- TERRACOTTA
- VoltDB
- MarkLogic
- INFORMATICA

Infrastructure As A Service
- amazon web services
- Windows Azure
- infochimps
- Google BigQuery

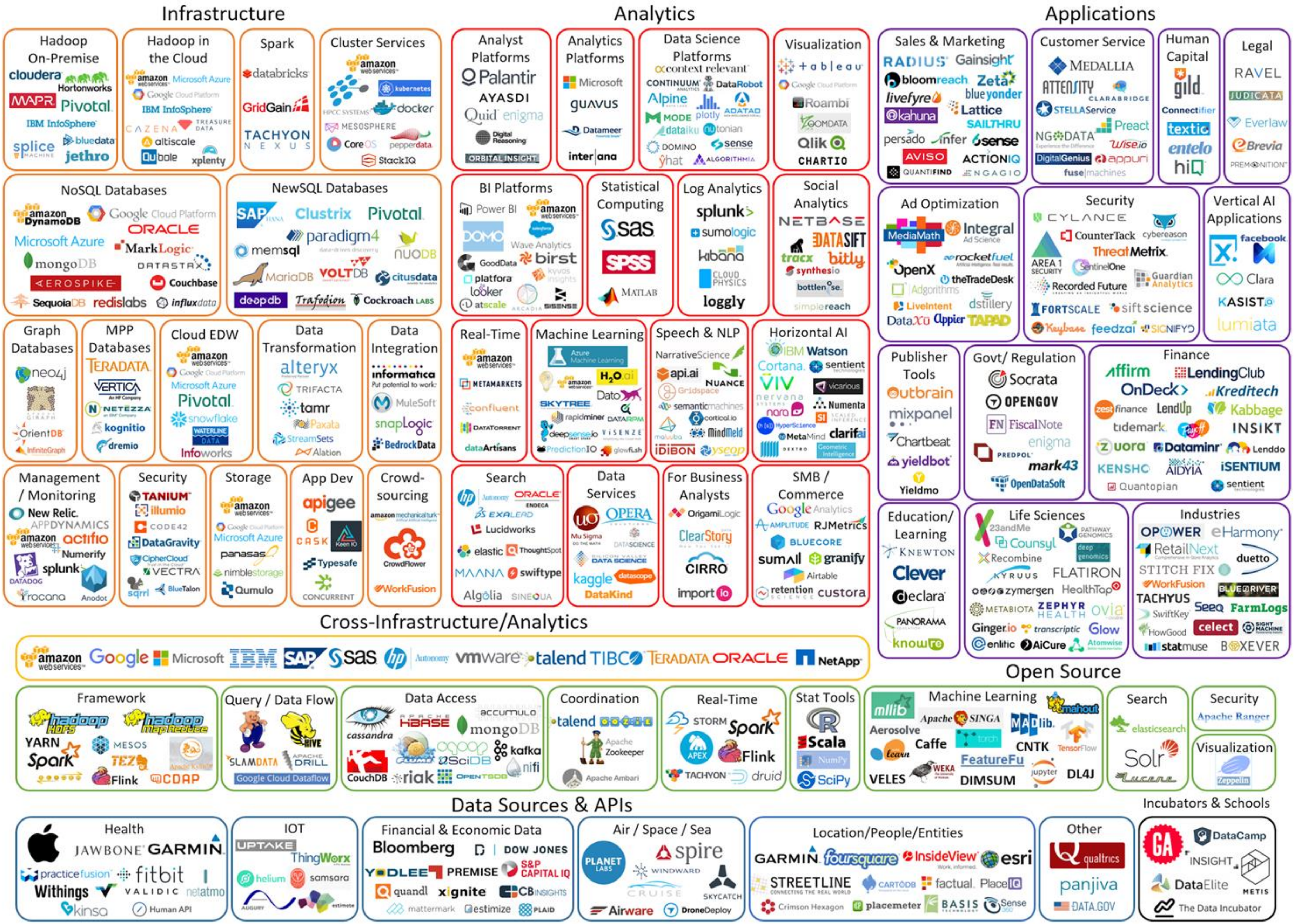
Structured Databases
- ORACLE
- MySQL
- Microsoft SQL Server
- PostgreSQL
- IBM DB2
- SYBASE
- memsql

Technologies
- hadoop
- hadoop MapReduce
- mahout
- APACHE HBASE
- Cassandra

dave@vcdave.com

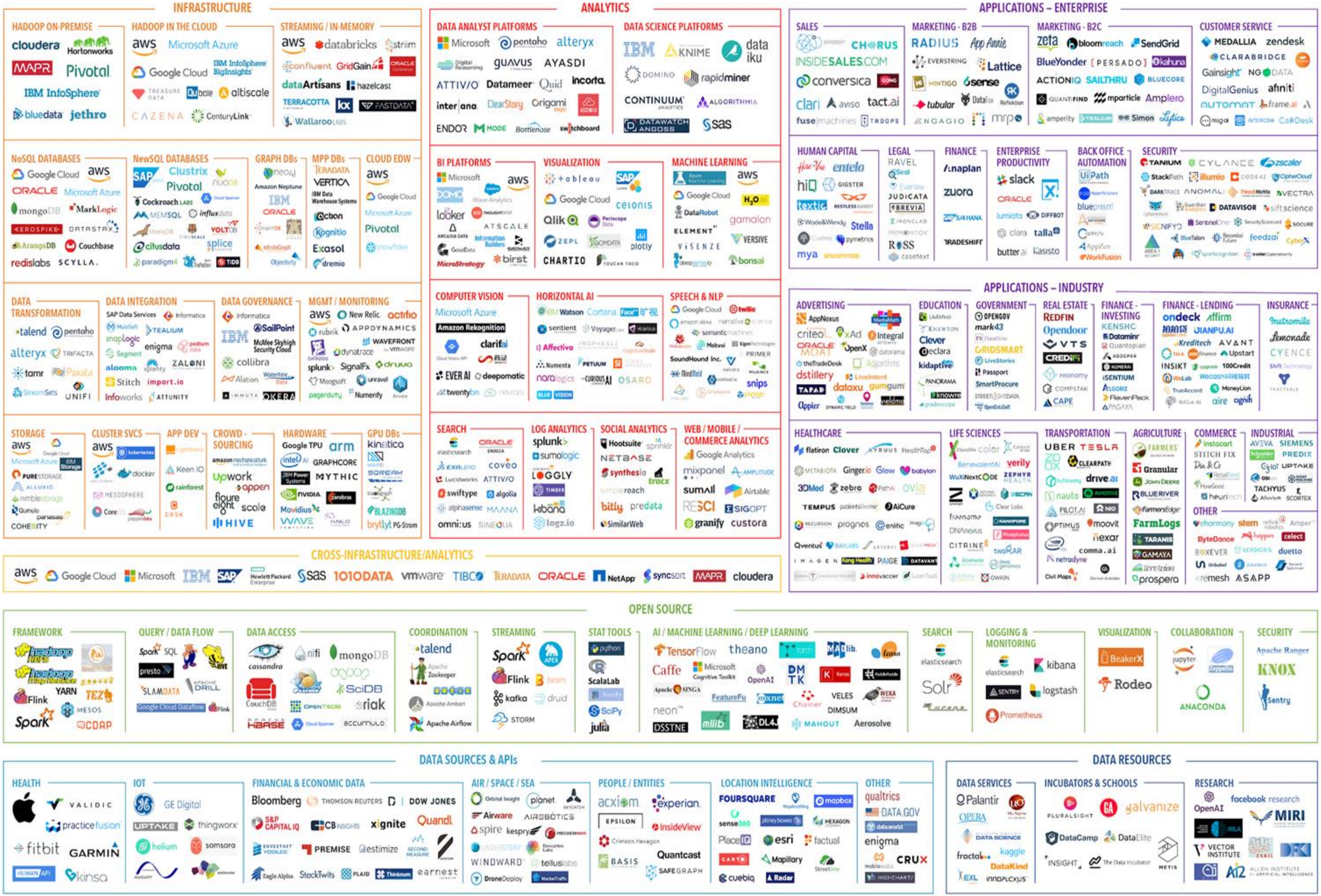
blogs.forbes.com/davefeinleib

Big Data Landscape 2016



© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

BIG DATA & AI LANDSCAPE 2018



Final 2018 version, updated 07/15/2018

© Matt Turck (@mattturck), Demi Obayomi (@demi_obayomi), & FirstMark (@firstmarkcap)

mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Data Ecosystem

File system	HDFS, NFS
Resource managers	Mesos, Yarn
Coordination	Zookeeper
Data Acquisition	Apache Flume, Apache Sqoop
Data Stores	MongoDB, Cassandra, Hbase
Data Processing	
• Frameworks	Hadoop MapReduce, Apache Spark, Apache Storm, Apache FLink
• Tools	Apache Pig, Apache HIve
• Libraries	SparkR, Apache Mahout, MlLib, etc
Data Integration	
• Message Passing • Managing data heterogeneity	Apache Kafka SemaGrow, Strabon
Operational Framework	
• Monitoring	Apache Ambari

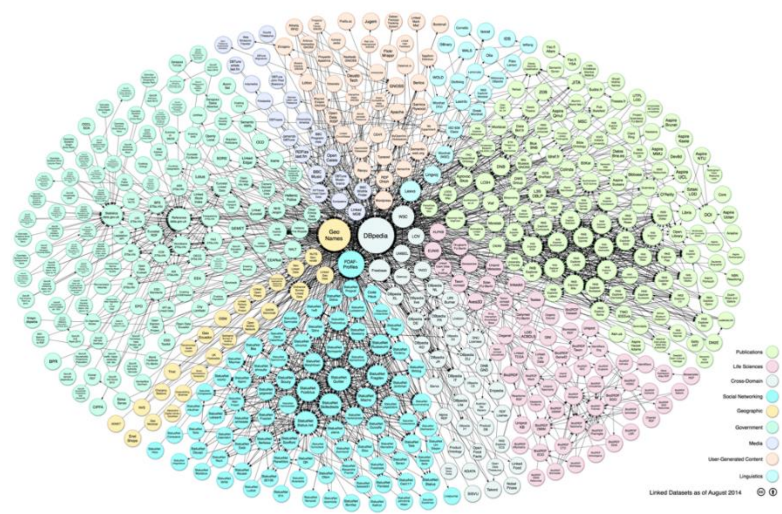
Big Data Analytics

- **Processing** the data and applying **inference** (i.e. trough **machine learning**) on Big data is key for proper **knowledge** (value) **extraction**

	linear regression	logistic regression	SVM	naive Bayes	discriminant analysis	survival regression	isotonic regression	decision trees	random forest	gradient boosting tree	isolation forest	bagging CART	C4.5	generalized linear model	ensembles	XGboost	NN	kNN	drift classifier	model-fitting
Apache Spark	+	+	+	+		+	+	+	+	+							+			
H2O				+					+	+	+			+	+	+	+			
R		+	+	+	+			+	+	+		+	+				+	+		
MOA				+				+									+		+	
Scikit - Learn	+	+	+	+	+		+	+	+	+	+			+	+		+	+		+
Bigml	+				+			+	+	+	+						+			
Weka	+	+	+	+					+				+							

Big Data Storage

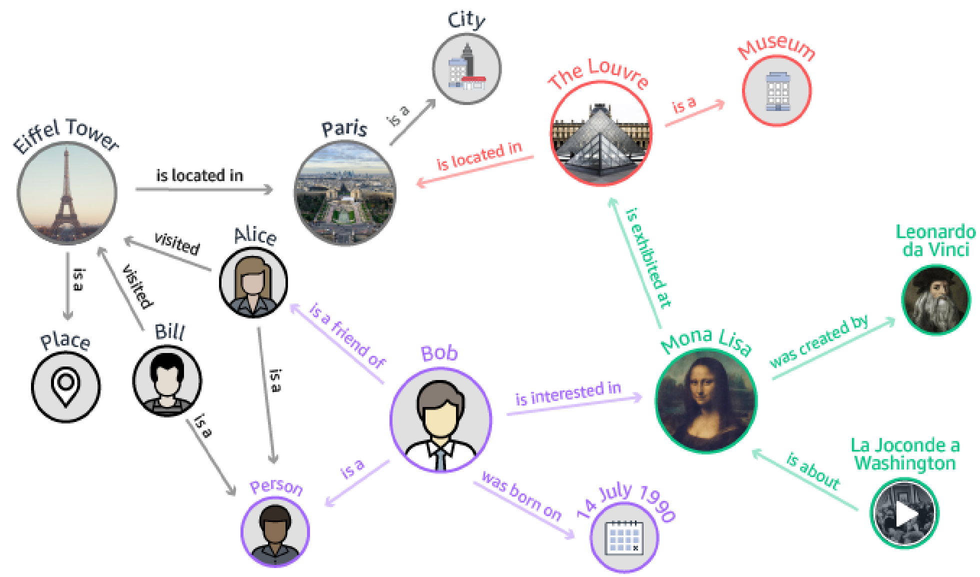
- No-SQL (not only SQL) databases
  - Key-value stores
    - Hazelcast
    - Redis
    - Membrane/Cocuhbase
    - Riak
    - Voldemort
    - Infinispan
  - Wide-column
    - Apache Hbase
    - Hypertable
    - Apache Cassandra
  - Document oriented
    - MongoDB
    - Apache CouchDB
    - Terrastore
    - RavenDB
  - Graph oriented
    - Neo4J
    - Infinite-Graph
    - InfoGrid
    - HypergraphDB
    - AllegroGrap
    - BigData



38 Billion **triples**

Graph Database

A graph database is essentially a collection of nodes and edges. Each node represents an entity (such as a person or business) and each edge represents a connection or relationship between two nodes. Every node in a graph database is defined by a <u>unique identifier</u>, a set of outgoing edges and/or incoming edges and a set of properties expressed as <u>key/value pairs</u>. Each edge is defined by a unique identifier,

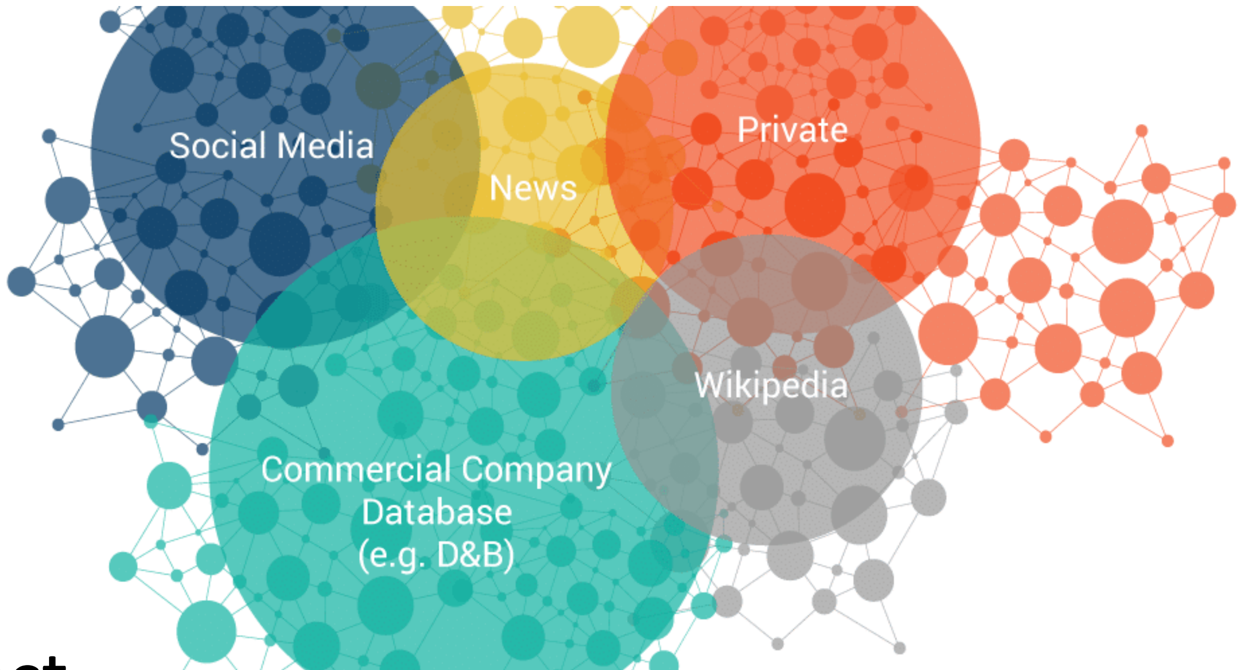


1994 2001

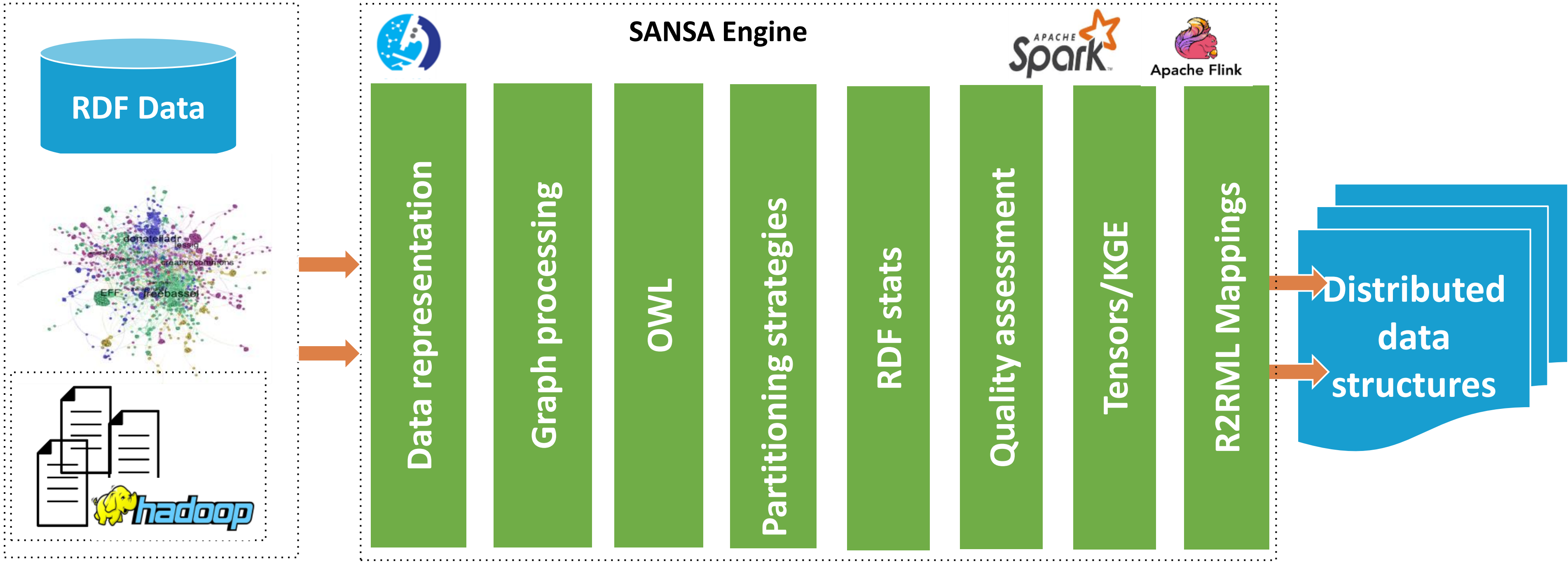
- The information is stored using **spo**-triples: (Subject, Predicate, Object) or as **spo = (s, p, o)**

Enterprise Knowledge Graphs

A knowledge graph structure not only allows an enterprise to organize, manage and discover internal data, but also to link these data to external data sources and benefit from the network effect.

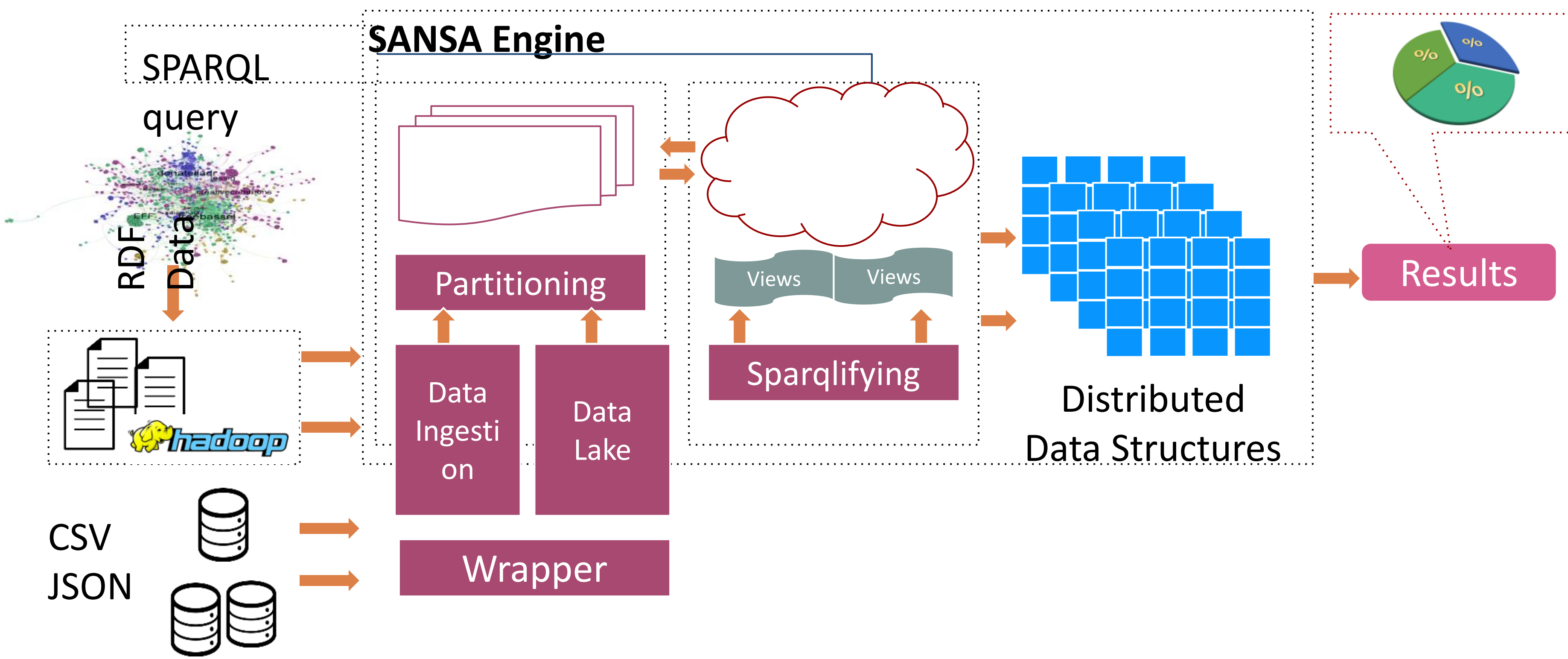


RDF Querying and Processing



- SANSA: Its core is a **data flow processing** engine that provides data distribution, and fault tolerance for distributed computations over RDF large-scale datasets

Querying via SPARQL & Partitioning



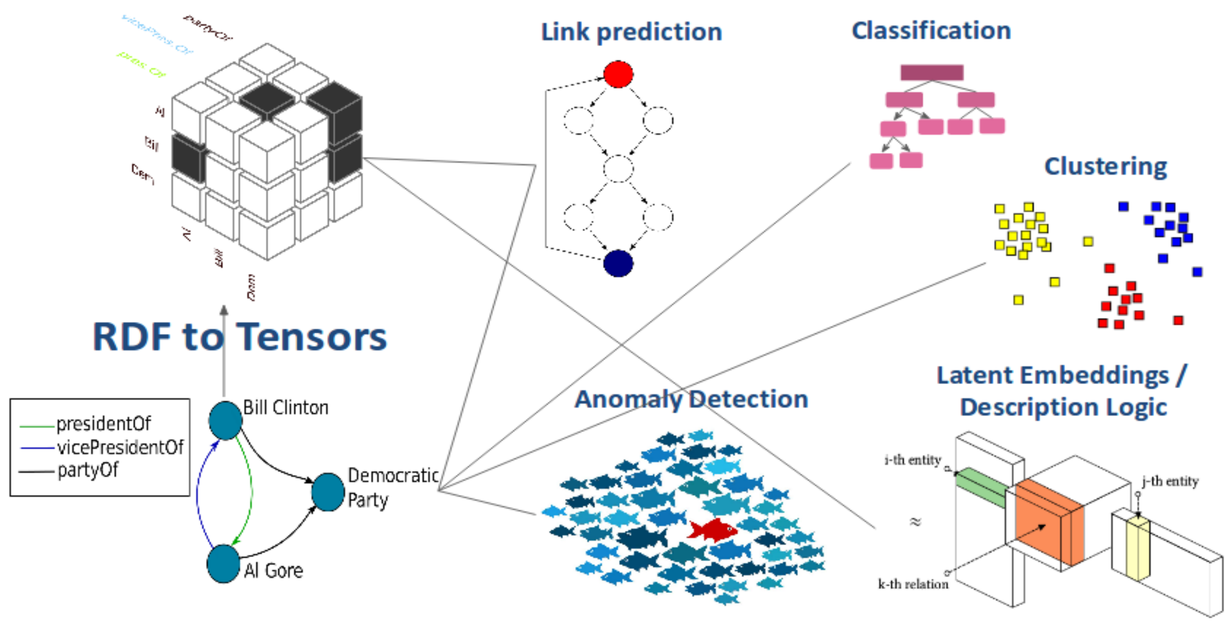
Machine Learning Layer

- ❖ Distributed ML algorithms using structure / semantics
- ❖ Algorithms:
  - ➢ Knowledge graph embeddings for e.g. KB completion, link prediction
  - ➢ Graph Clustering
  - ➢ Association rule mining (AMIE+ = mining horn rules from RDF data using partial completeness assumption and type constraints)
  - ➢ Anomaly Detection
  - ➢ Semantic Decision trees (in progress)

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

UNIVERSITÄT BONN

LEARNING, APPLYING, MULTIPLYING



Big Data Visualization

- JavaScript libraries (open source)
  - Chart.js
  - Leaflet
  - Chartist.js
  - n3-charts
  - Sigma JS
  - Polymaps
  - Processing.js
  - Dyagraph
- Timelines
  - Timeline JS

- Chart tools
  - Fusion Charts
  - Chart.js
  - Chartist.js
  - n3-charts
  - Canvas
- Map tools
  - Leaflet
  - Polymaps
- Images
  - Processing.js

- Graphs and networks
  - Sigma JS
- Multi-purpose
  - D3.js
  - Ember-charts
  - Google charts
- Non-web
  - Cuttlefish
  - Cytoscape
  - Gephi
  - Graphwiz
  - Graph-tool

- Cross-platform
  - NodeXL
  - Pajek
  - SocNetV
  - Sentinel Visualizer
  - Statnet
  - Tulip
  - Visone
- Commertial (desktop)
  - Tableau
  - Infogram

LAMBDA Consortium

Learning, Applying, Multiplying Big Data Analytics

Networking

LAMBDA Network of Experts

@Net4LAMBDA


WELCOME