
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](#)

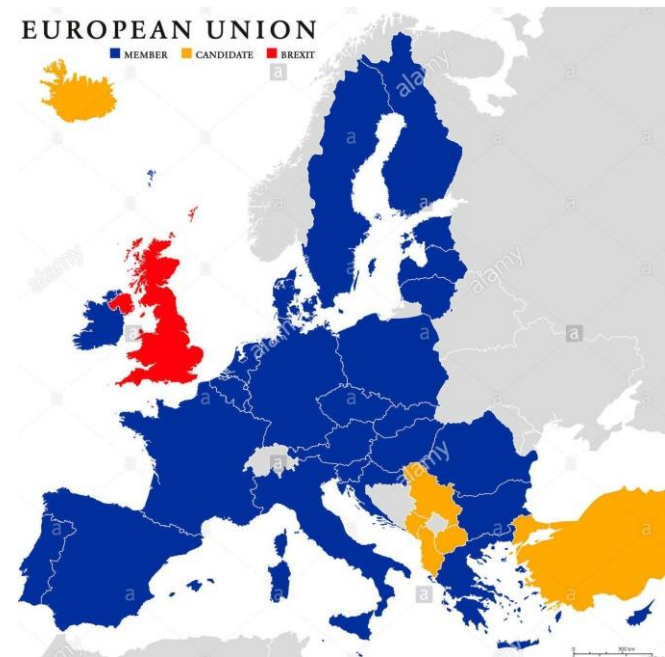


Vision and Primary Objectives



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



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

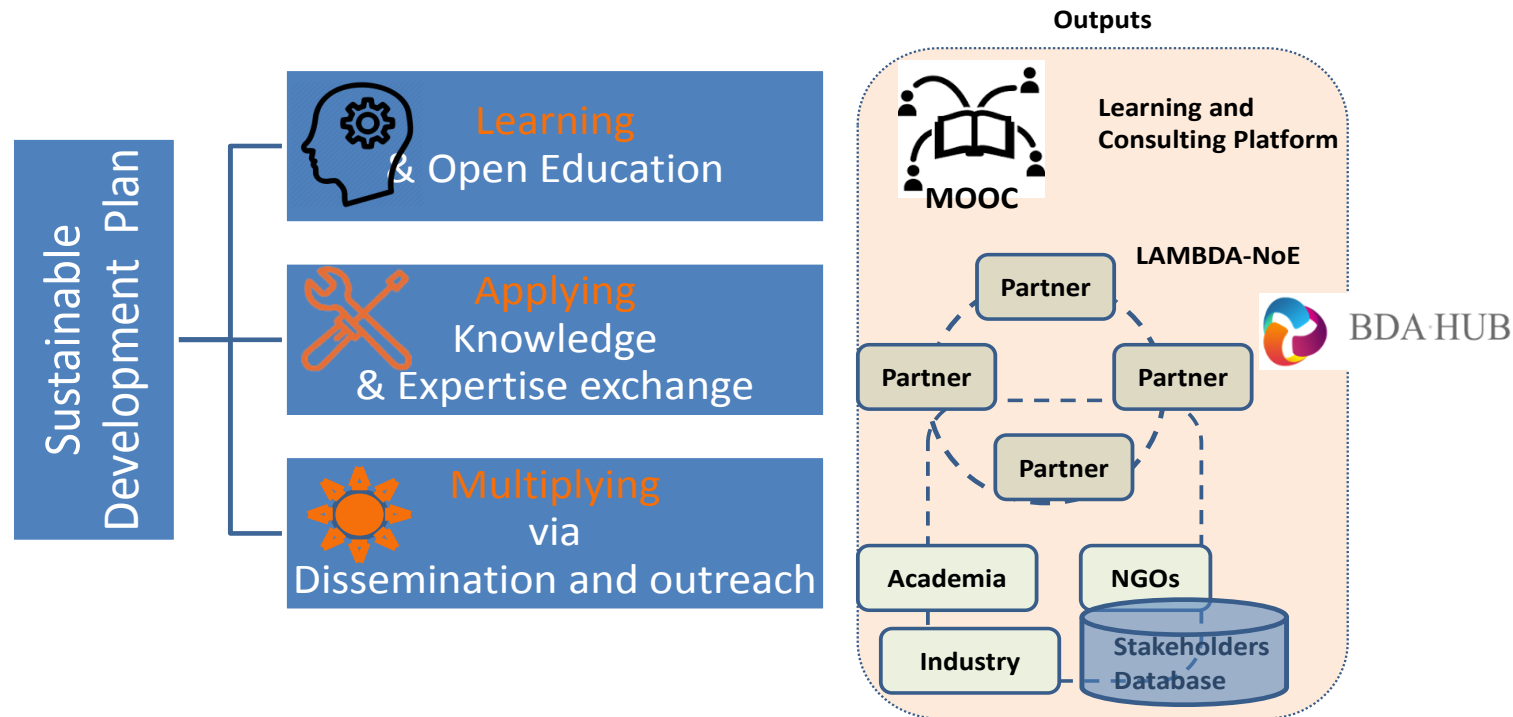
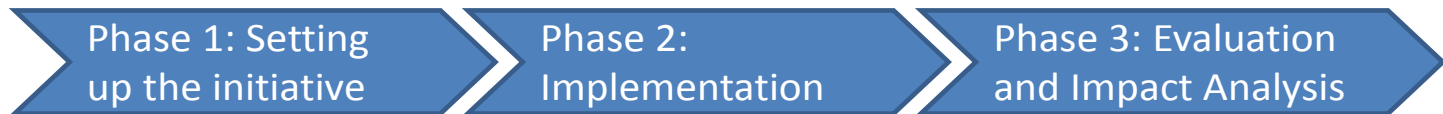


Methodology

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	<p>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:</p> <ul style="list-style-type: none">• https://project-lambda.org/Learning• https://project-lambda.org/Knowledge-repository/Lectures
Applying Knowledge & Cooperation	<p>LAMBDA Experts Exchange Program for teachers, researchers and developers) will open possibilities for collaborative research on open issues in Big Data related areas:</p> <ul style="list-style-type: none">• Industry 4.0• ICT for Energy
Multiplying Dissemination and outreach	<p>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</p> <p>Sustainable Development Plan for PUPIN (2021-2025)</p> <p>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.</p>



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**



LAMBDA Platform



[Home](#) [Project](#) [Methodology](#) [eLearning](#) [News & Events](#) [Results](#) [Join Us](#)

[Stakeholders Section](#) [Summer School](#) [Knowledge Repository](#)

[Home](#) » [Big Data Analytics School 2019](#)

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 EU** 2020. The event will bring together researchers and professionals from respectable EU Universities and other stakeholders from the West Balkan countries to discuss state-of-the-art in Big Data research and applications. The Big Data Analytics Summer School will allow the invited PhD students (participants of the School from Serbia) to 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)
- 3rd day Big Data Analytics Summer School - Lectures from LAMBDA partners (UBO and UOXF)

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

		1st Day	2nd Day	3rd Day	
8 30		Registration	Registration	Registration	

Invited Teachers -
Keynotes - Lectures

Education & Research
Organizations

Faculties/ Departments

Industry

Public Administration &
NGO

- [My account](#)
- [Log out](#)

bda-school@mail.project-lambda.org

LAMBDA - Learning, Applying, Multiplying Big Data Analytics

Big Data Analytics
State-of-the-art Review



LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

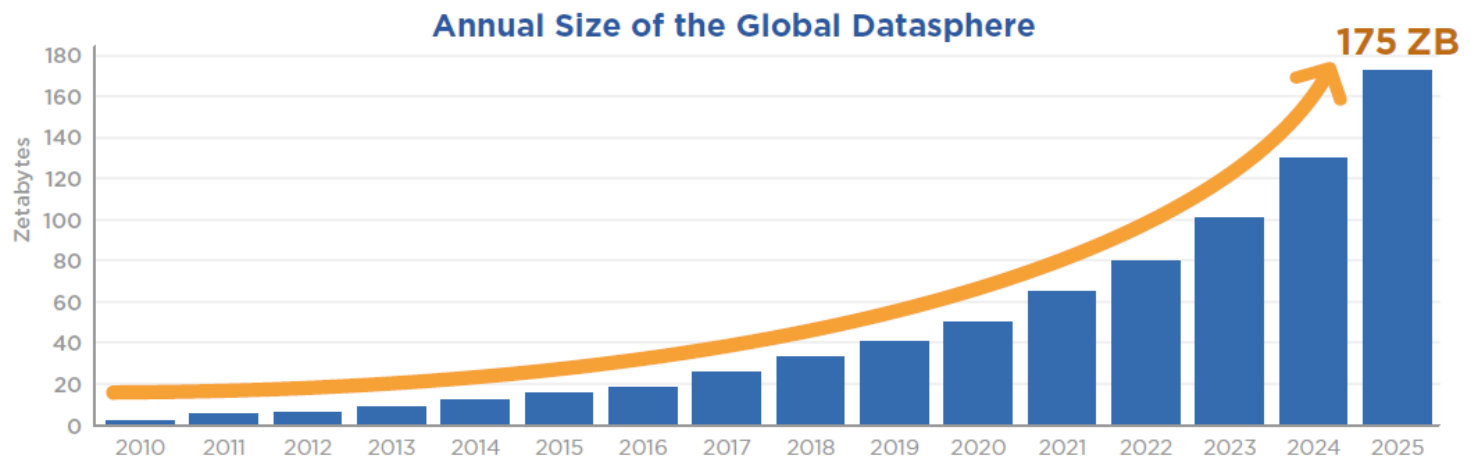
This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 809965.



Big Data

- **Big Data** is used more as a **buzzword** than 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 through **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
- Veracity – truthfulness (**uncertainty**) of data
- Validity – **suitability** of the selected dataset for a given application
- Volatility – **temporal validity** and fluency of the data
- Value – (useful) **information** extracted from the data
- Visualization – properly **displaying** and showcasing information
- Vulnerability – **security** and **privacy** concerns associated
- Variability – the **changing meaning** of data

3V's

5V's

7V's

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



Log Data Apps



Ad/Media Apps



Business Intelligence



Analytics and Visualization



Data As A Service



Analytics Infrastructure



Operational Infrastructure



Infrastructure As A Service



Structured Databases



Technologies



Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
cloudera Hortonworks MAPR Pivotal IBM InfoSphere splice bluedata jethro

Hadoop in the Cloud
amazon Microsoft Azure Google Cloud Platform IBM InfoSphere CAZENA TREASURE DATA altiscale Duoble xplenty

Spark
databricks GridGain TACHYON NEXUS

Cluster Services
amazon web services kubernetes HPCC SYSTEMS docker MESOSPHERE CoreOS pepperdata StackIQ

Analytics

Analyst Platforms
Palantir AYASDI Quid enigma ORBITALINSIGHTS

Analytics Platforms
Microsoft guavus Datameer interana

Data Science Platforms
context relevant DataRobot Alpine ADATA MODE plotly dataiku plotian DOMINO sense yhat ALGORITHMIA

Visualization
Google Cloud Platform Roambi Qlik CHARTIO

Applications

Sales & Marketing
RADIUS Gainsight bloomreach Zeta livefyre blueyonder kahuna Lattice SAILTHRU persado infer sense AVISO ACTIONIQ QUANTIFIND ENGAGIO

Customer Service
MEDALLIA ATTENITY CLARABRIDGE STELLASERVICE NGDATA Preact DigitalGenius Wiseio appurri fuse/machines

Human Capital
gild Connectifier textic entelo hiQ

Legal
RAVEL JUDICATA Everlaw Brevia PREMORITION

NoSQL Databases
amazon DynamoDB Google Cloud Platform Microsoft Azure ORACLE mongoDB MarkLogic DATASTAX Couchbase Aerospike SequoiaDB redislabs influxdata

NewSQL Databases
SAP HANA Clustrix Pivotal paradigm4 memsql nuODB MariaDB VOLTDB citusdata deepdb Trafodion Cockroach LABS

BI Platforms
Power BI amazon web services Domo Wave Analytics GoodData birst GoodData platform looker atscale arcadia BUSINESS

Statistical Computing
sas SPSS MATLAB

Log Analytics
splunk sumologic kibana CLOUD PHYSICS loggly

Social Analytics
NETBASE DATASIFT trolox bitly synthetio bottlenose simplereach

Ad Optimization
MediaMath Integral Ad Science OpenX rocketfuel theTradeDesk Algorithms Liveintent distillery DataXu Clipper TAPAD

Security
CYLANCE CounterTack cybereason ThreatMetrix AREA 1 SECURITY SentinelOne Recorded Future Guardian Analytics FORTSCALE sift science Kaybase feedzai SICNIFYD

Vertical AI Applications
facebook Clara KASISTO lumiata

Graph Databases
neo4j OrientDB InfiniteGraph

MPP Databases
TERADATA VERTICA Netezza Vertica Kognitio dremio

Cloud EDW
amazon web services Google Cloud Platform Microsoft Azure Pivotal snowflake WATERLINE DATA Infoworks

Data Transformation
alteryx TRIFACTA tamr Paxata StreamSets Alation

Data Integration
informatica Put potential to work MuleSoft snapLogic BedrockData

Real-Time
amazon web services METAMARKETS confluent DATATORRENT dataArtisans

Machine Learning
Azure Machine Learning H2O OIL SKYTREE Dato rapidminer DATAARMA deepinsight VIZENZ PredictionIO glowfluh

Speech & NLP
NarrativeScience api.ai NUANCE Gridspace semantic machines contextual Mind Meld iDIBON yscop

Horizontal AI
IBM Watson Cortana sentient VIV nervana nora SI Numenta MetaMind clarifai

Publisher Tools
Outbrain mixpanel Chartbeat yieldbot Yieldmo

Govt/Regulation
Socrata OPENGOV FN FiscalNote enigma mark43 PREDPOL OpenDataSoft

Finance
affirm LendingClub OnDeck Kreditech LendUp Kabbage tidemark PAYFI INSIGHT ZUORA Dataminr Lenddo KENSHC AIDYIA ISENTIUM Quantopian

Management / Monitoring
New Relic APPDYNAMICS amazon web services actifio Numerity splunk DATADOG Yocana Anodot

Security
TANIRUM Illumio CODE42 DataGravity CipherCloud VECTRA sqrl BlueTalon

Storage
amazon web services Google Cloud Platform Microsoft Azure Pivotal panasas nimblestorage Qumulo

App Dev
apigee CASK Kaven IO Typesafe CONCURRENT

Crowd-sourcing
amazon mechanical turk CrowdPower WorkFusion

Search
hp ORACLE ENDECA EXALEAD Lucidworks elastic ThoughtsSpot MAANA swifttype Algolia SINEQUA

Data Services
UG OPERA Mu Sigma DATA SCIENCE kaggle datacscope DataKind

For Business Analysts
OrigamiLogic ClearStory CIRRO import io

SMB / Commerce
Google Analytics AMPUTRUE RJMetrics BLUECORE sumal granify Airtable retention custora

Education / Learning
KNEWTON Clever Geclara PANORAMA knowre

Life Sciences
23andMe Counsyl ReCombine KYRUS FLATIRON oozymergen HealthTop METABIOTA ZEPHYR HEALTH ovio Gingerio transcriptic Glow enlith AlCure Atomwise

Industries
OPOWER eHarmony RetailNext duetto STITCH FIX WorkFusion TACHYUS SwiftKey Seeq FarmLogs HowGood select BBOXEVER

Cross-Infrastructure/Analytics

amazon web services Google Microsoft IBM SAP SAS hp Autonomy vmware talend TIBCO TERADATA ORACLE NetApp

Open Source

Framework
Hadoop HADOOP HADOOP HADOOP YARN Spark MESOS TEZ Flink CDAP

Query / Data Flow
SLAMDATA ADVANTAGE DRILL Google Cloud Dataflow

Data Access
cassandra HBASE mongoDB kafka riak OPENTSDB CouchDB

Coordination
talend Zookeeper Apache Ambari

Real-Time
STORM Spark APEX Flink TACHYON druid

Stat Tools
R Scala NumPy SciPy

Machine Learning
mllib Apache SINGA MADlib Aerosolve Caffe FeatureFu DUMSUM jupyter DL4J

Search
elasticsearch Solr Lucene

Security
Apache Ranger Visualization Zeppelin

Data Sources & APIs

Health
Apple JAWBONE GARMIN practicefusion fitbit Withings VALIDIC netatmo kinsa Human API

IOT
UPTAKE ThingWorx helium samsara samsara estimate

Financial & Economic Data
Bloomberg DOW JONES YODLEE PREMISE S&P CAPITAL IQ quandl xignite CBINSIGHTS mattermark estimate PLAID

Air / Space / Sea
PLANET LABS WINWARD CRUISE SKYWATCH Airware DroneDeploy

Location/People/Entities
GARMIN foursquare InsideView esri STREETLINE CARTODB factual PlaceIQ CRIMSON Hexagon placemeter BASIS Sense

Other
qualtrics panjiva DATA.GOV

Incubators & Schools
DataCamp INSIGHT DataElite METIS The Data Incubator

BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

The collage is organized into three main sections, each with a title and a collection of company logos:

- HADOOP ON-PREMISE:** Includes logos for Cloudera, Hortonworks, MapR, Pivotal, IBM InfoSphere, Aludata, and jethro.
- HADOOP IN THE CLOUD:** Includes logos for AWS, Microsoft Azure, Google Cloud, IBM InfoSphere BigInsights, Treasure Data, Dataloq, Altilscale, CAZEN, and CenturyLink.
- STREAMING / IN-MEMORY:** Includes logos for AWS, databricks, Stream, Confluent, GridGain, EMC Data Direct, dataArtisans, Hazelcast, TERRACOTTA, Kox, Wallaroo, and FASTDATA.

ANALYTICS

DATA ANALYST PLATFORMS

- Microsoft
- Pentaho
- alteryx
- Digital Learning
- quavus
- AYASDI
- ATTIVIO
- Datameer
- Quid
- incorta
- inter ana
- ClearStory
- Origami
- EVIDIO

DATA SCIENCE PLATFORMS

- IBM
- KNIME
- data iku
- DOMINO
- rapidminer
- CONTINUUM ANALYTICS
- ALGORITHMIA
- DATALAB
- SAS

APPLICATIONS – ENTERPRISE

The collage displays logos for several database technologies, organized into five categories:

- NoSQL DATABASES:** Google Cloud, AWS, Oracle, Microsoft Azure, MongoDB, MarkLogic, Aerospike, DataStax, ArangoDB, Couchbase, Redis Labs, and Scylla.
- NewSQL DATABASES:** SAP, Clustrix, Pivotal, nuodb, Cockroach Labs, InfluxData, MemSQL, IBM, Oracle, VoltDB, Ceph, and Splice.
- GRAPH DBs:** Neo4j, Amazon Neptune, IBM, Oracle, and Gremlin.
- MPP DBs:** Teradata, Vertica, IBM Data Warehouse Systems, Cloudera, Kognitio, and Exasol.
- CLOUD EDW:** AWS, Google Cloud, Microsoft Azure, Pivotal, and Snowflake.

The collage is organized into three columns, each with a red header:

- BI PLATFORMS:** Includes logos for Microsoft, AWS, Domo, Wave Analytics, Looker, Qlik, Tableau, Alteryx, and others.
- VISUALIZATION:** Includes logos for Tableau, SAP, Google Cloud, Celonis, Qlik, Periscope Data, Alteryx, and others.
- MACHINE LEARNING:** Includes logos for AWS, Google Cloud, DataRobot, Element, Vizeen, and others.

DATA TRANSFORMATION

- talend
- pentaho
- alteryx
- TRIFACTA
- tmrm
- Paxata
- StreamSets
- UNIFI

DATA INTEGRATION

- SAP Data Services
- Informatica
- Microsoft
- TEALUM
- inprolog
- enigma
- podium data
- Segment
- alooma
- spicely
- ZALONI
- Stitch
- import.io
- InfoWorks
- ATTUNITY

DATA GOVERNANCE

- Informatica
- SailPoint
- IBM
- Mobile Security Shield
- colibra
- Waterline Data
- Alation
- OKERA

MGMT / MONITORING

- AWS
- New Relic
- actifio
- rubik
- ADAPPOYANICS
- WAVEFORM
- by VERTIGO
- dynatrace
- sparkfun
- SignalFX
- druuvo
- Moogsoft
- pagerduty
- Numery
- unwired
- Andros

COMPUTER VISION

- Microsoft Azure
- Amazon Rekognition
- Clarifai
- Cloud Vision API
- EVER AI
- deepomatic
- twentybn
- neurata

HORIZONTAL AI

- IBM Watson Cortana
- Facer 人脸识别
- Sentient
- Voyager
- Affective
- Numenta
- PETUM
- NOLOGICS
- OSARG
- BILAL VISION

SPEECH & NLP

- Google Cloud
- twilio
- amazon alexa
- naturaler voice
- semantic solutions
- Motiv8
- Eigen Technologies
- PRIMER
- SoundHound Inc.
- Voxtek
- Medfield
- snips
- vncor

The collage is organized into six vertical columns, each representing a different category of technology:

- STORAGE:** Includes logos for Google Cloud, Microsoft Azure, IBM Cloud, Pure Storage, Alluxio, Hadoop, and Cohesity.
- CLUSTER SVCS:** Includes logos for AWS, IBM Watson, Docker, Keen IO, Mesosphere, and Core OS.
- APP DEV:** Includes logos for GitHub, Keen IO, and Rainforest.
- CROWD-SOURCING:** Includes logos for Amazon Mechanical Turk, Upwork, and Figure Eight.
- HARDWARE:** Includes logos for Google GPU, ARM, Intel AI, Graphcore, MYTHIC, NVIDIA, Movidius, and Wave Computing.
- GPU DBs:** Includes logos for Kinetic, Numenta, Sift Science, BLAZEIO, and BriteLyt PG-Store.

SEARCH

- elasticsearch
- ORACLE
ENERGIA
- EXALENGO
- COVEO
- Searchmetrics
- ATTIVO
- swiftype
- algolia
- alphasense
- MAANA
- omni-us
- SINEQUA

LOG ANALYTICS

- splunk
- sumologic
- LOGGLY
- THIMARA
- kbano
- logz.io

SOCIAL ANALYTICS

- Hootsuite
- sprinklr
- NETBASE
- synthesio
- etrack
- smarterreach
- bitty
- predata
- SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS

- Google Analytics
- mixonpanel
- Amplitude
- sumall
- Airtale
- RESCI
- SIGOPT
- granify
- custora

HEALTHCARE
 flatiron Clover Atrius HealthTap
 ViMethica Ginkgo Glow babylon
 3DME Zebra PAPA ovia
 TEMPUX patientshome AIcure
 recursion prognos @natic mapQ

LIFE SCIENCES
 GlaxoSmithKline color Cytosine
 Bioventurox verily
 WalkMeXcode ZEPHYR HEALTH
 Zebra Clear Labs
 Fimmata DYNAREUS
 HANDBOXED P Resonance

TRANSPORTATION
 uber TESLA
 CLEARPATH
 drive.ai
 nauto PILOTAI NIRO
 OPTIMUS moovit
 FarmLogs

AGRICULTURE
 FARMERS
 Granular
 John Deere BLUE RIVER
 Farmers Edge
 FarmLogs

COMMERCE
 iStockphoto STITCH FIX
 Dac & Co
 Heartbeat
 Hiveworks
 TACHYON
 Alluvium SCOUT24

INDUSTRIAL
 AVIVA SIEMENS
 iStockphoto PREDICI
 UPTAKE
 TACHYON
 Alluvium SCOUT24

OTHER
 vcharmony storm rafes neta Ampere
 and beyond

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA vmware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera

OPEN SOURCE

The diagram illustrates a comprehensive ecosystem of data science and engineering tools, organized into 11 functional categories:

- FRAMEWORK**: Includes logos for TensorFlow, PyTorch, Keras, Flink, YARN, TFS, MESOS, and CDAP.
- QUERY / DATA FLOW**: Includes Spark SQL, Presto, SLAMDATA, Google Cloud Dataflow, and Databricks.
- DATA ACCESS**: Includes Cassandra, MongoDB, CouchDB, OpenTable, SciDB, and HBase.
- COORDINATION**: Includes Talend, Apache Zookeeper, Apache Ambari, and Apache Airflow.
- STREAMING**: Includes Spark, Flink, Beam, Kafka, Druid, and Storm.
- STAT TOOLS**: Includes Python, R, Scalalab, and SciPy.
- AI / MACHINE LEARNING / DEEP LEARNING**: Includes TensorFlow, Theano, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, Keras, FeatureFu, MXNet, Chainer, VEGAS, DMSUM, and DL4J.
- SEARCH**: Includes Elasticsearch, Solr, and Lucene.
- LOGGING & MONITORING**: Includes Kibana, Elasticsearch, Sentry, Logstash, and Prometheus.
- VISUALIZATION**: Includes BeakerX, Rodeo, and Anaconda.
- COLLABORATION**: Includes Jupyter, Leppin, and Anaconda.
- SECURITY**: Includes Apache Ranger, KNOX, and Sentry.

DATA SOURCES & APIs

HEALTH

- Apple
- VALIDIC
- practicefusion
- fitbit
- GARMIN
- HUMAN API
- kinsa

IOT

- GE Digital
- UPTAKE
- thingworx
- helium
- samsara
- ENVESTNET | YODLEE
- PREMISE
- ALTIMETER
- SECOND MEASURE
- Eagle Alpha
- StockTwits
- PLAID
- Thinkbox
- earnest

FINANCIAL & ECONOMIC DATA

- Bloomberg
- THOMSON REUTERS
- DOW JONES
- SEI CAPITAL IQ
- CB INSIGHTS
- xignite
- Quandl
- WINDWARD
- INDUSTRY
- DRONEDROP
- TELLUS LABS
- MARKETCALL

AIR / SPACE / SEA

- Orbital Insight
- Pionet
- AIRBOTICS
- spire
- kespry
- AXIOM
- EXPERIAN
- ESLON
- INSIDEVIEW
- SENSE360
- JETWAY SYSTEMS
- HEXAGON
- PLACEIQ
- ESRI
- FACTUAL
- ENIGMA
- CRUX

PEOPLE / ENTITIES

- axiom
- experian
- ESLON
- INSIDEVIEW
- SENSE360
- JETWAY SYSTEMS
- HEXAGON
- PLACEIQ
- ESRI
- FACTUAL
- ENIGMA
- CRUX

LOCATION INTELLIGENCE

- FOURSQUARE
- Mapbox
- MOPOBOX
- SENSE360
- JETWAY SYSTEMS
- HEXAGON
- PLACEIQ
- ESRI
- FACTUAL
- ENIGMA
- CRUX

OTHER

- QUALTRICS
- DATA.GOV
- DATAWORLD
- ENIGMA
- CRUX
- STREETVIEW

DATA RESOURCES

DATA SERVICES

- Palantir
- Uber
- OPERA
- DATA SCIENCE
- fractal
- kaggle
- DataKind
- iEXL
- innoCOPUS

INCUBATORS & SCHOOLS

- FLURIA:8IGHT
- GA
- yalvanize
- DataCamp
- DataElite
- INSIGHT
- The Data Incubator
- METIS

RESEARCH

- facebook research
- OpenAI
- MIRI
- GML
- VECTOR INSTITUTE
- CSAIL
- MIT
- ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE

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	
<ul style="list-style-type: none"> Frameworks 	Hadoop MapReduce, Apache Spark, Apache Storm, Apache FLink
<ul style="list-style-type: none"> Tools 	Apache Pig, Apache Hive
<ul style="list-style-type: none"> Libraries 	SparkR, Apache Mahout, MLlib, etc
Data Integration	
<ul style="list-style-type: none"> Message Passing Managing data heterogeneity 	Apache Kafka SemaGrow, Strabon
Operational Framework	
<ul style="list-style-type: none"> Monitoring 	Apache Ambari



Big Data Analytics

- **Processing** the data and applying **inference** (i.e. through **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/Couchbase
 - ▣ Riak
 - ▣ Voldemont
 - ▣ Infinispan



- ▣ Wide-column

- ▣ Apache Hbase
 - ▣ Hypertable
 - ▣ Apache Cassandra



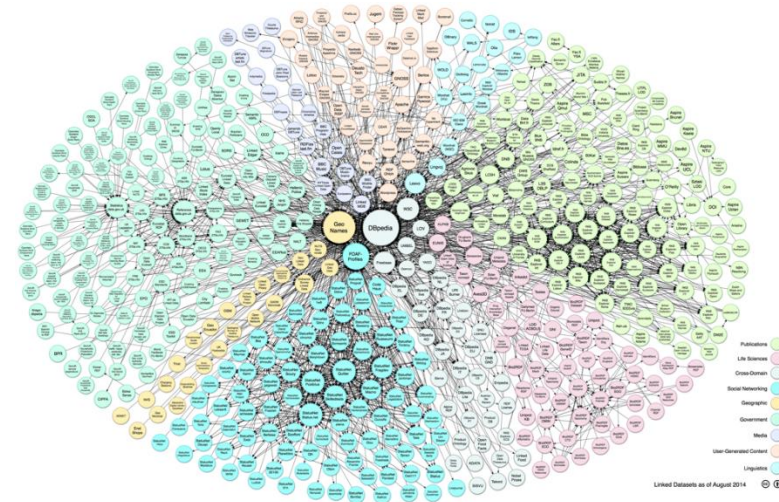
- ▣ Document oriented

- ▣ MongoDB
 - ▣ Apache CouchDB
 - ▣ Terrastore
 - ▣ RavenDB



- ▣ Graph oriented

- ▣ Neo4j
 - ▣ Infinite-Graph
 - ▣ InfoGrid
 - ▣ HypergraphDB
 - ▣ AllegroGrap
 - ▣ BigData

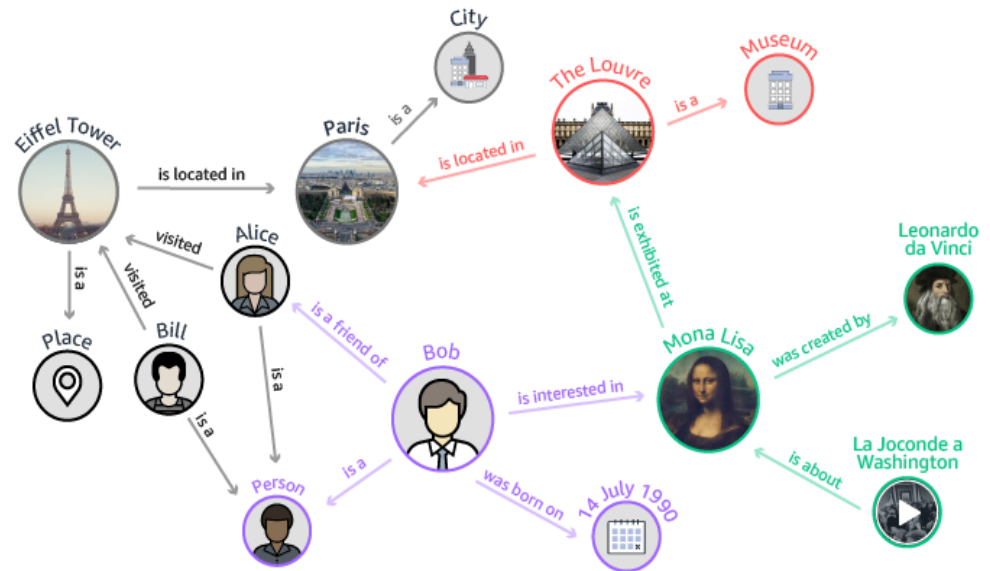


38 Billion triples



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 unique identifier, a set of outgoing edges and/or incoming edges and a set of properties expressed as key/value pairs. Each edge is defined by a unique identifier,



1994



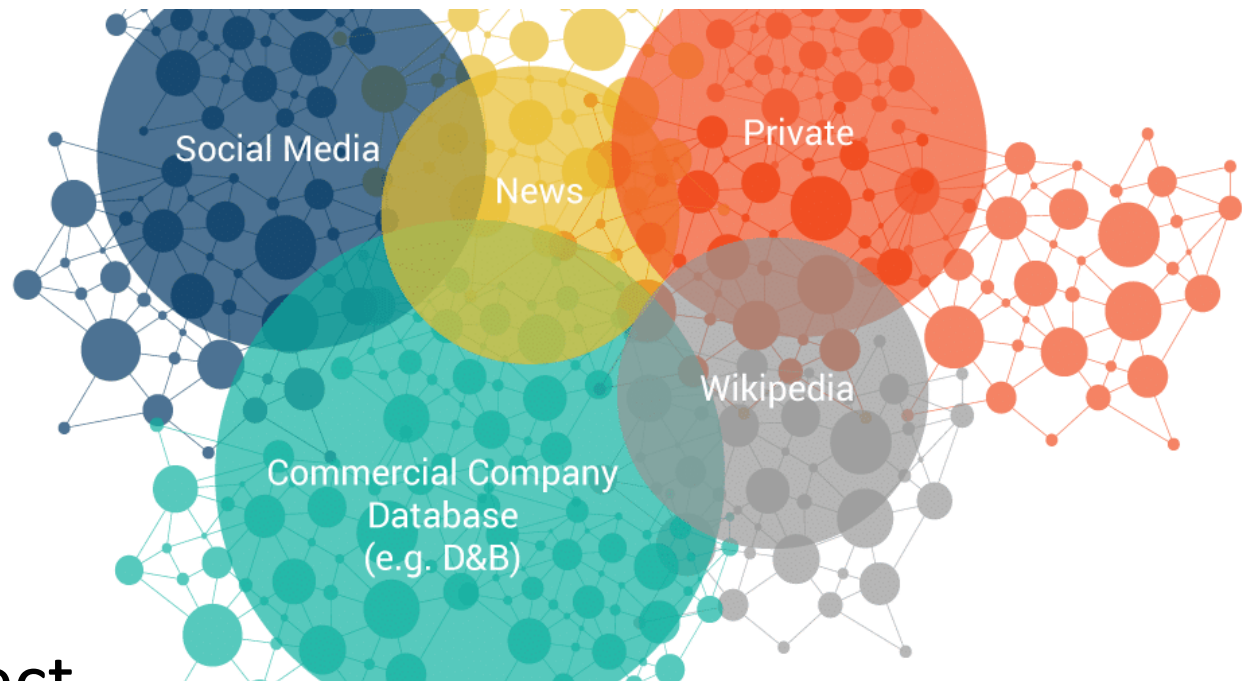
2001

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

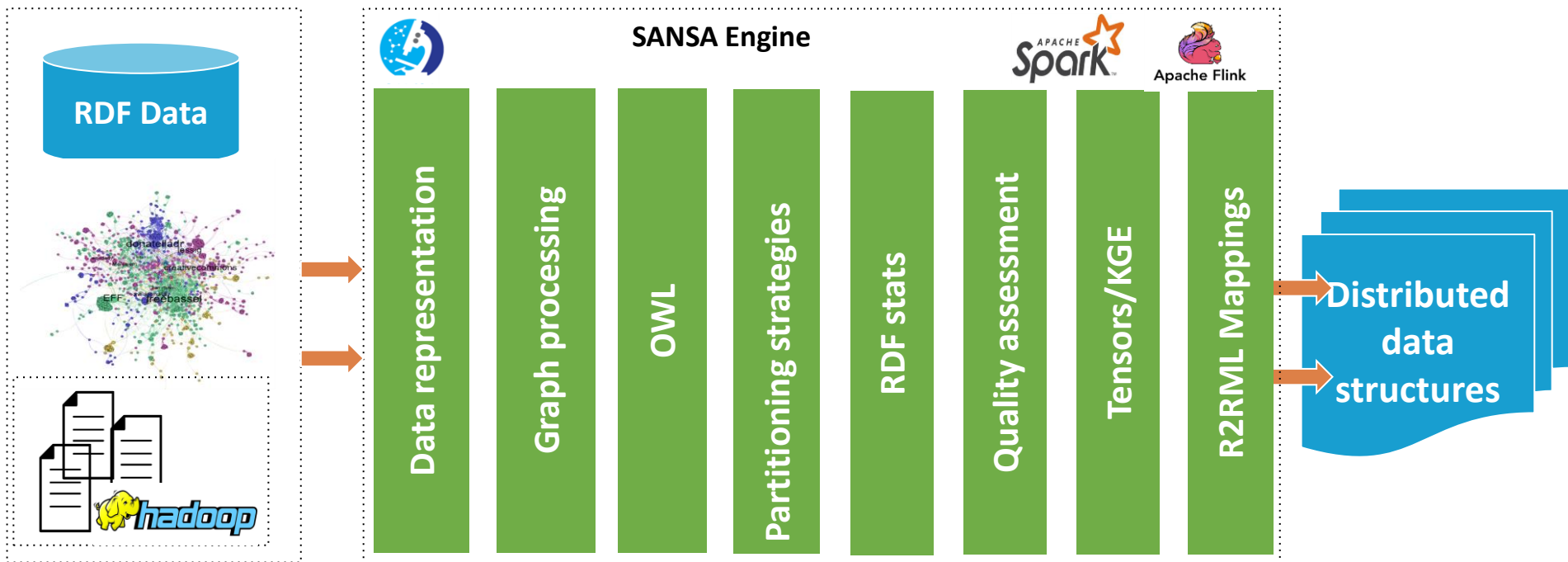


Enterprise Knowledge Graphs

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.

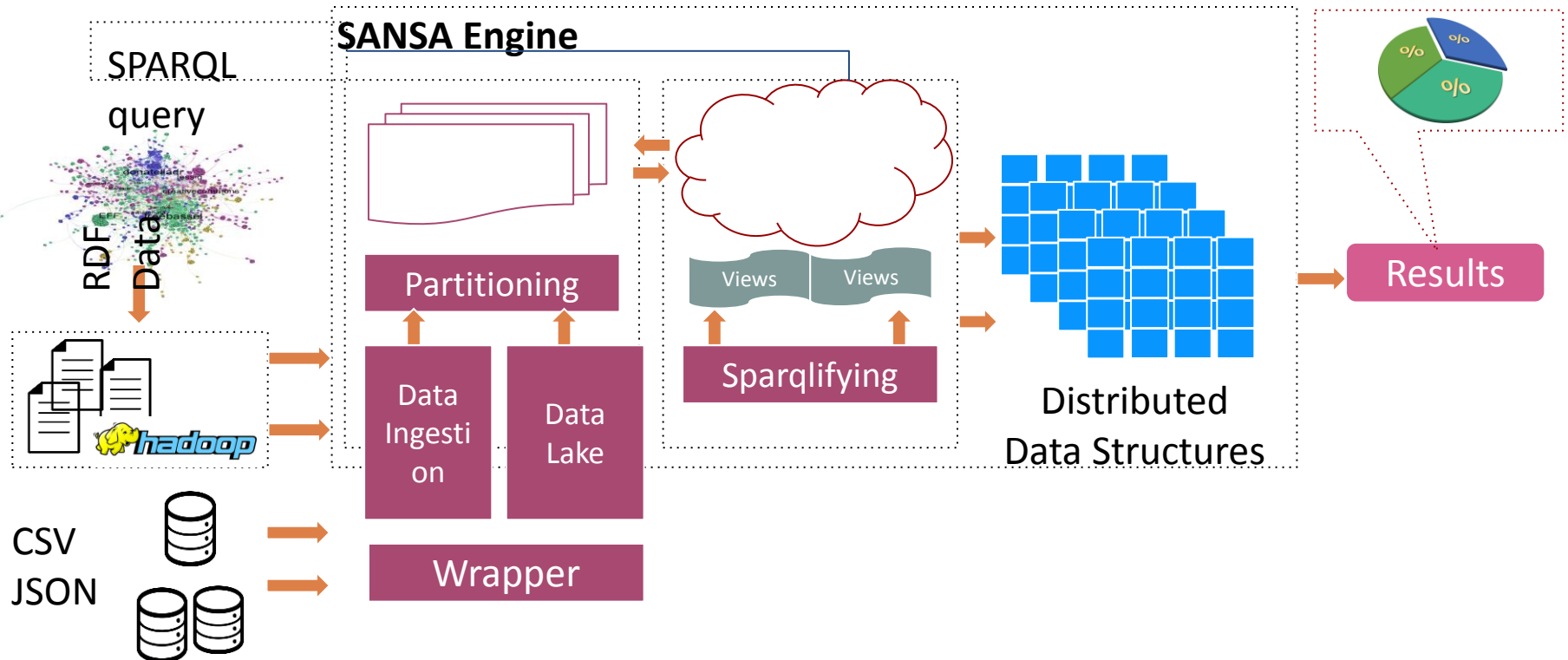


RDF Querying and Processing



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

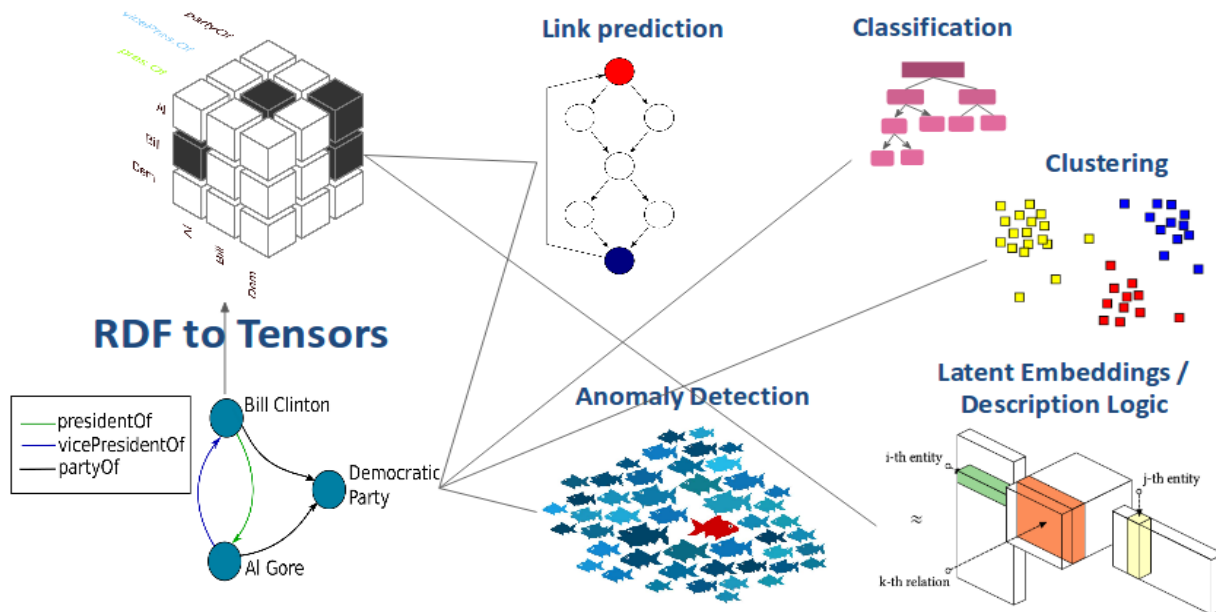
Querying via SPARQL & Partitioning



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)



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
- ▣ Commercial (desktop)
 - ▣ Tableau
 - ▣ Infogram



LAMBDA Consortium



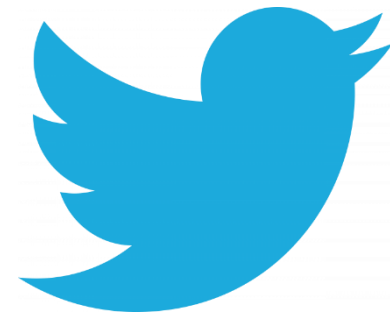
LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS



Networking

LAMBDA Network of Experts

[@Net4LAMBDA](#)



LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS



Fraunhofer

IAIS

UNIVERSITÄT



BONN





WELCOME

