A State-of-the-art Review on Big Data Technologies

Semantic technologies for Big Data: Volume, Velocity, Variety and Veracity @ ICIST 2019

Marko Jelić, *BSc EE & CS*

Junior researcher, The Mihajlo Pupin Institute

Dea Pujić, *BSc EE & CS*

Junior researcher, The Mihajlo Pupin Institute

Hajira Jabeen, *PhD*

Senior researcher, Computer Science Institute, University of Bonn

institut MIHAJLO PUPIN

Acknowledgment/Context

Institute Mihajlo Pupin Serbia
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

- **LAMBDA**[1] (Learning, Applying, Multiplying, Big Data Analytics) is a twinning[2] H2020 project
- The main goal of the project is to **provide** different **knowledge transfer instruments** (mentorships, brainstorming sessions, school type activities) and different types of **twinning relationships** (institution to institution, institution to network)
- The **specific focus** of the knowledge transfer process is placed on the **Big data domain** and corresponding **technologies and services**

[1] https://project-lambda.org/
[2] https://ec.europa.eu/neighbourhood-enlargement/tenders/twinning_en

What is 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
- The introduction of US social security in **1937** is considered by some as the **start of the Big data era** but this term has gained **most of its popularity just recently** following the development of data heavy applications

[*] Illustrations by https://www.freepik.com/macrovector

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

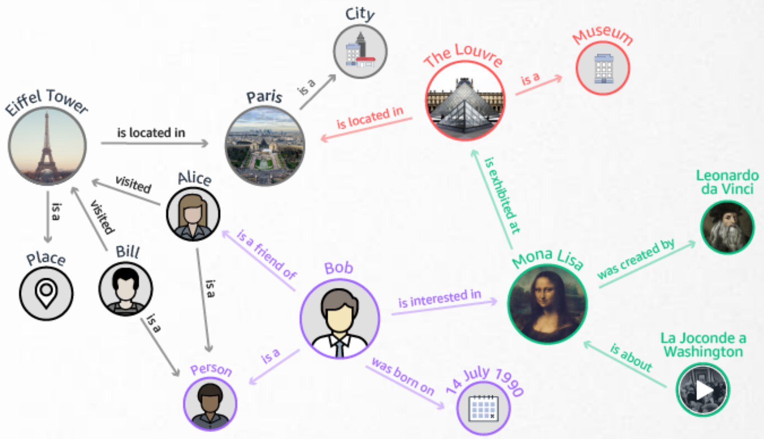
- The **core technological challenges** working with Big data that **stem from** its **complex nature** are:
  - Heterogeneity – differences in structure
  - Uncertainty – data reliability
  - Scalability – sizing the workflow and infrastructure
  - Timeliness – real-time requirements
  - Fault tolerance – sensitivity to errors
  - Data security – privacy issues, data leaks
  - Visualization – displaying of information

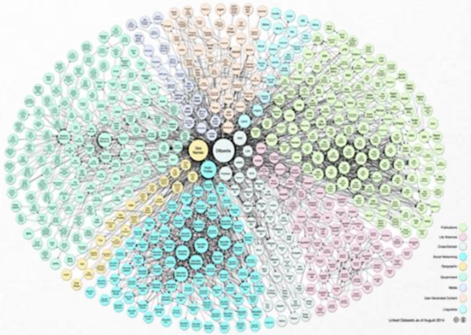
	Storing	Processing	Analytics	Visualization
Heterogeneity	+	+		
Uncertainty of data		+	+	
Scalability	+	+	+	
Timeliness	+	+	+	
Fault tolerance		+	+	
Data security	+	+		
Visualization				+

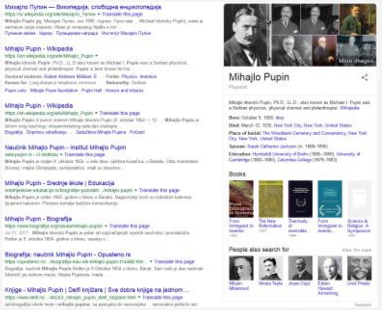
Big data Storage

- 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

- Knowledge graphs







Михајло Пупин — Википедија, слободна енциклопедија

Mihajlo Pupin - Wikipedia

Mihajlo Pupin - Wikipedia

Naučnik Mihajlo Pupin - Institut Mihajlo Pupin

Mihajlo Pupin - Srednje škole | Edukacija

Mihajlo Pupin - Biografija

Biografija: naučnik Mihajlo Pupin - Opusteno.rs

Knjige - Mihajlo Pupin | Delfi knjižare | Sve dobre knjige na jednom ...

Mihajlo Pupin

Physicist

Mihajlo Idvorski Pupin, Ph.D., LL.D., also known as Michael I. Pupin was a Serbian physicist, physical chemist and philanthropist. Wikipedia

Born: October 9, 1858, Idvor

Died: March 12, 1935, New York City, New York, United States

Place of burial: The Woodlawn Cemetery and Conservancy, New York City, New York, United States

Spouse: Sarah Catharine Jackson (m. 1888–1896)

Education: Humboldt University of Berlin (1885–1889), University of Cambridge (1883–1885), Columbia College (1879–1883)

Books

From Immigrant to Inventor · The New Reformation · Thermodynamics of reversible... · From Immigrant to Inventor... · Science & Religion: A Symposium

People also search for

Milutin Milanković · Nikola Tesla · Jovan Cvijić · Edwin Howard Armstrong · Uroš Predić

* Illustrations by https://aws.amazon.com/neptune/ and https://lod-cloud.net/versions/2014-08-30/lod-cloud.svg

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	+	+	+	+					+				+							

Systematization of **regression and classification** learning algorithms in Big data tools

Big data analytics

- If the data is **not already labeled** i.e. separated into appropriate classes, **clustering algorithms** need to be applied first in order to determine adequate class limits

	K-means	G-means	Gaussian mixture	PIC	LDA	aggregator	PAM	CLARA	Fuzzy clustering	Model-based	Hierarhical	Dencity based	Afinity propagation
Apache Spark	+		+	+	+						+		
H2O	+					+							
R	+						+	+	+	+	+	+	
Giraph	+												+
BigML	+	+			+								

Systematization of **clustering learning algorithms** in Big data tools

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

Questions?

institut MIHAJLO PUPIN

Thank you for your attention!

Look for the full paper “**A State-of-the-art Review on Big Data Technologies**” in the ICIST 2019 proceedings after April 15th!