

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



- ▣ LAMBDA¹ (Learning, Applying, Multiplying, Big Data Analytics) is a twinning² 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** 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
- ▣ 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



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 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

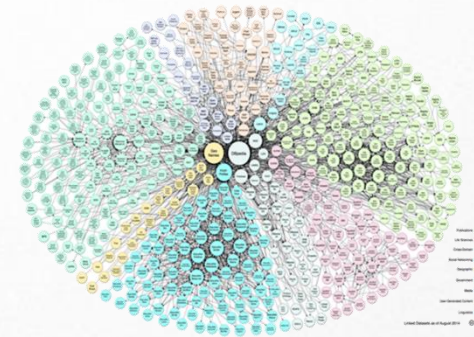
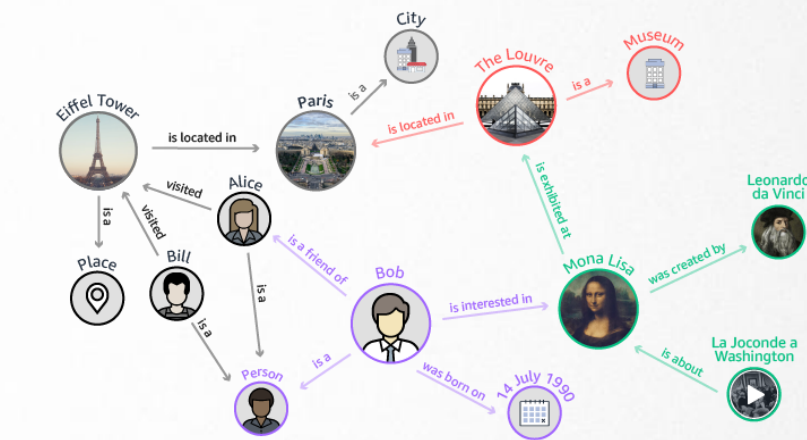


■ Graph oriented

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



■ Knowledge graphs



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

- 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 | + | + | + | + | | | | | + | | | | + | | | | | | | |

Systematization of regression and classification learning algorithms in Big data tools

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

Questions?

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!