BIG DATA STATE OF THE ART
Sr. Josep Lluís Larriba-Pey
CEO | Director
Sparsity | DAMA-UPC
www.sparsity-technologies.com
www.dama.upc.edu
The Linked Data Benchmark Council: Big Data Benchmarking and Generation

Josep Lluís Larriba-Pey
Sparsity & DAMA-UPC
The Linked Data Benchmark Council

Linked Data Benchmark Council:
- Benchmarking organization
- Specializes in Graph and RDF technologies
- Formed by

ORACLE

IBM

SAP

HUAWEI

Neo4j

OPENLINK SOFTWARE

ontotext

*Sparsity
Four main elements in a benchmark:

• **Data & Schema.** Describe a business.

• **Workloads.** Shows the type of interaction between users and the system. Transactional, BI, Analytics.

• **Performance metrics.** How we compare systems fairly.

• **Auditing rules.** We all must proceed in the same way to be comparable.
Big Data Generation

• Data is becoming the cornerstone of many business models and applications:
  • It contains sensitive and critical information,
  • Holders reluctant to share data
    • Privacy at risk
    • Competitive advantage

• This can:
  • Limit the growth of the data economy
  • Truncate business opportunities
Need for anonymized Big Data generation

Use cases:
• Medical data: necessary to analyse, difficult to release.
• City data: understanding how people move for the benefit of society.
• Migrant data: how to alleviate absorption.
BigData-EnGen platform

BigData-EnGen is a joint effort
BigData-EnGen platform

- Original Data + Schema
  - Forwards:
    - Schema
    - Correlations
    - Distributions
    - Structural Properties
    - Anonymized/Encrypted Dictionaries

- Analysis + Encryption Tool

- Data Generator (DataSynth)

- Big Data Cluster Platform (i.e. Spark)

- BigData platform to generate large amounts of data
DataSynth

Generic Data generator

• Accepts achema specifications and data properties, via DSL
• Extensible via plugins
• Designed to scale on clusters
• Completely Open Source (https://github.com/DAMA-UPC/DataSynth)

Still on pre-alpha state, subject to changes

We are open to contributions!!!
DataSynth Domain Specific Language

- JSON
- We can express entities, edges and attributes
- Example of code and the property graph generated
- The compiler traverses the property graph and generates IR code
DataSynth Backend

Backend implemented in Apache Spark 2.x.x

- Currently using DataSet API (in Scala) taking advantage of built-in query optimizer
- Can incorporate other programming models and APIs, such as GraphX or Streaming

Backends in other technologies can be implemented
Conclusions

• We strongly believe in the future of Big Data generation

• There are many business models that can grow out of it:
  • SMEs providing services
  • Companies growing their analytics businesses

• We are open to collaborate with third parties