HOBBIT
A Brief Overview

Michael Röder, Axel-Cyrille Ngonga Ngomo

DICE research group
Institute of applied Informatics, Leipzig, Germany
University Paderborn, Germany
(Horizon 2020, GA No 688227)

BDV PPP Summit
Riga, June 27th, 2019
Introduction
A Lot of Data

Volume
SCALE OF DATA
- 40 ZETTABYTES (43 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005.
- 6 BILLION PEOPLE have cell phones.
- WORLD POPULATION: 7 BILLION

The FOUR V’s of Big Data

Variety
DIFFERENT FORMS OF DATA
- 30 BILLION PIECES OF CONTENT are shared on Facebook every month.
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users.
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month.
- By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS.

Velocity
ANALYSIS OF STREAMING DATA
- The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session.
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure.
- By 2016, it is projected there will be 18.8 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth.

Veracity
UNCERTAINTY OF DATA
- 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate.

By 2015
4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States.

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES (161 GIGABYTES).

By 2020,
It’s estimated that
2.5 QUINTILLION BYTES (2.3 TRILLION GIGABYTES) of data are created each day.
Most companies in the U.S. have at least 100 TERABYTES (100,000 GIGABYTES) of data stored.

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technologies and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety, and Veracity.

- Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

http://www.ibmbigdatahub.com/infographic/four-vs-big-data
Which tool(s) should I use for my use case?
Introduction
Many Questions

- Where are the current bottlenecks?
- Which steps of the data lifecycle are critical?
- Which solutions are available on the market?
- Which key performance indicators are relevant?
- How well should tools perform?
- How do existing solutions perform w.r.t. relevant indicators?
Research project from 2015 – 2018 (Horizon 2020, GA No 688227)
Focus on Big Linked Data
Cover the business-critical steps of the Linked Data lifecycle
Used by a growing number of companies
Mature and maturing technologies
Gathered real requirements
- Focussed on industrial requirements
- Gathered relevant performance indicators
- Gathered relevant performance thresholds
- Gathered real datasets
Introduction

Aims

1. Gathered real requirements
   - Focussed on industrial requirements
   - Gathered relevant performance indicators
   - Gathered relevant performance thresholds
   - Gathered real datasets

2. Developed benchmarks based on real data

---

Michael Röder (InfAI)
Introduction

Aims

1. Gathered real requirements
   - Focussed on industrial requirements
   - Gathered relevant performance indicators
   - Gathered relevant performance thresholds
   - Gathered real datasets

2. Developed benchmarks based on real data

3. Provided universal benchmarking platform
   - Comparable results
   - Hosted as a free-to-use online instance

Michael Röder (InfAI)
Introduction

Aims

1. Gathered real requirements
   - Focussed on industrial requirements
   - Gathered relevant performance indicators
   - Gathered relevant performance thresholds
   - Gathered real datasets

2. Developed benchmarks based on real data

3. Provided universal benchmarking platform
   - Comparable results
   - Hosted as a free-to-use online instance

4. Periodic benchmarking challenges and reporting
Introduction

Aims

1. Gathered real requirements
   - Focussed on industrial requirements
   - Gathered relevant performance indicators
   - Gathered relevant performance thresholds
   - Gathered real datasets

2. Developed benchmarks based on real data

3. Provided universal benchmarking platform
   - Comparable results
   - Hosted as a free-to-use online instance

4. Periodic benchmarking challenges and reporting

5. Created an association (Special Group 7 of Task Force 6)
Section 1

Project Highlights
Scalable open-source benchmarking platform
Local, distributed and remote (AWS) deployment
First FAIR platform for benchmarking Big Linked Data in a holistic manner
Project Highlights

- 5 mimicking algorithms
- 52 benchmarks
- 200+ systems
- 14 challenges
- DEBS GC 2017 and 2018
- 300+ users
- 13K+ experiments
Section 2

Benchmarking Machine Learning
- The task: find anomalies in molding machine sensor data to predict maintenance intervals (predictive maintenance).
- Mimicking algorithm based on real data
- Data was streamed as in the real world
- Participants had to use Markov Models to identify anomalies
- 14 Participants, 7 made it into the last round
Benchmarking Machine Learning
SML Benchmark v1 for DEBS GC 2017
The task: predictions about ship routes based on AIS data
- Spatio-temporal streaming data
- Predictions for vessels’ destinations and arrival times
<table>
<thead>
<tr>
<th>Team</th>
<th>Earlyness rate</th>
<th>A</th>
<th>Working time (sec)</th>
<th>B</th>
<th>Total Q1</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Iasi</td>
<td>0.685</td>
<td>1</td>
<td>99</td>
<td>2</td>
<td>1.25</td>
</tr>
<tr>
<td>University of Illinois</td>
<td>0.672</td>
<td>2</td>
<td>86</td>
<td>1</td>
<td>1.75</td>
</tr>
<tr>
<td>Jean Monnet University</td>
<td>0.668</td>
<td>3</td>
<td>149</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>Chungnam National University</td>
<td>0.653</td>
<td>4</td>
<td>102</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td>University of Iasi (2nd)</td>
<td>0.647</td>
<td>5</td>
<td>157</td>
<td>6</td>
<td>5.25</td>
</tr>
<tr>
<td>Israel Institute of Technology</td>
<td>0.5</td>
<td>6</td>
<td>129</td>
<td>4</td>
<td>5.5</td>
</tr>
<tr>
<td>Dresden University of Technology</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Insight Centre</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>University of Carthage</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Project Overview

<table>
<thead>
<tr>
<th>Team</th>
<th>Mean Absolute Error (min.)</th>
<th>A</th>
<th>Working time (sec)</th>
<th>B</th>
<th>Total Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Iasi</td>
<td>959.839</td>
<td>1</td>
<td>100</td>
<td>2</td>
<td>1.25</td>
</tr>
<tr>
<td>Jean Monnet University</td>
<td>1099</td>
<td>2</td>
<td>145</td>
<td>4</td>
<td>2.5</td>
</tr>
<tr>
<td>Chungnam National University</td>
<td>1251.15</td>
<td>3</td>
<td>100</td>
<td>2</td>
<td>2.75</td>
</tr>
<tr>
<td>Israel Institute of Technology</td>
<td>1493.18</td>
<td>4</td>
<td>133</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td>University of Illinois</td>
<td>5425.53</td>
<td>5</td>
<td>86</td>
<td>1</td>
<td>4.75</td>
</tr>
<tr>
<td>University of Iasi (2nd)</td>
<td>1705.35</td>
<td>6</td>
<td>164</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Dresden University of Technology</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Insight Centre</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>University of Carthage</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Benchmarking Machine Learning

**SML Benchmark v2 for DEBS GC 2018**

**Team** | **Q1** | **Q2** | **Total Score**
--- | --- | --- | ---
University of Iasi | 1.25 | 1.25 | 2.5
Jean Monnet University | 3.5 | 2.5 | 6
Chungnam National University | 3.75 | 2.75 | 6.5
University of Illinois | 1.75 | 4.75 | 6.5
Israel Institute of Technology | 5.5 | 3.75 | 9.25
University of Iasi (2nd) | 5.25 | 5 | 10.25
Dresden University of Technology | - | - | -
Insight Centre | - | - | -
University of Carthage | - | - | -
Section 3

Future Directions
Future Directions

KnowGraphs (Innovative Training Networks (ITN))
- 4 years, starting in October 2019
- 15 Early-Stage Researchers (ESRs)
- HOBBIT will be used as central benchmarking platform
- Further datasets will be integrated (e.g., UICML datasets)

RAKI (BMWi project)
- 3 years, starting in September 2019
- HOBBIT will be used for evaluation

More projects pending
Future Directions

KnowGraphs (Innovative Training Networks (ITN))
- 4 years, starting in October 2019
- 15 Early-Stage Researchers (ESRs)
- HOBBIT will be used as central benchmarking platform
- Further datasets will be integrated (e.g., UICML datasets)

RAKI (BMWi project)
- 3 years, starting in September 2019
- HOBBIT will be used for evaluation

More projects pending

→ Further development of the HOBBIT platform

HOBBIT is open for the community! Benchmarks, systems, datasets can be added

Not limited to linked data
Thank you

HOBBIT offers

- Scalable benchmarking
- Based on real world data in an
- Extendable,
- Open source platform
- Following the FAIR data principles

http://project-hobbit.eu/
https://dice-research.org/about/