Big Data Challenges

Big Data Value Association (BDVA)

ET4HPC SRA preparation
Munich, 20.3.2017

www.bdva.eu

Big Data Value PPP (BDV PPP):

- ✓ Create **value** out of the **data**!
- ✓ Boost **European Big Data research** and **innovation**
- ✓ Strengthen **competitiveness** and ensuring **industrial leadership**

- Work together to organize an **IMPACTFUL program**

- **Efficiency**: Further develop the EU big data ecosystem (**i-Spaces**)

- Proof points**: Demonstrate the impact of Big Data** and show the value in **Lighthouse** projects

- **Innovation:** Create impactful **research** results

BDV BIG DATA VALUE

www.bdva.eu

Big Data PPP aims at

the development of an interoperable data-driven ecosystem as a source for new businesses and innovations using Big Data.

To achieve the BDV SRIA has defined

four implementation mechanisms

i-Spaces	**i-Spaces** (Innovation Spaces) are cross-organization, cross-sector and interdisciplinary Spaces to anchor targeted research and innovation projects. They offer secure accelerator-style environments for experiments for private data and open data, bringing technology and application development together. I-Spaces will act as incubators for new businesses and the development of skills, competence and best practices.
Lighthouse projects	**Lighthouse** projects are large-scale data-driven innovation and demonstration projects that will create superior visibility, awareness and impact.
Technical priorities	**Technical priorities:** These will take up specific Big Data issues addressing targeted aspects of the technical priorities
Cooperation & coordination projects	**Cooperation & coordination projects:** These projects will foster international cooperation for efficient information exchange and coordination of activities

BDV BIG DATA VALUE

www.bdva.eu

Who is behind PPP BDVA?

And many more....
An Industry-led growing European community with over 160 members

www.bdva.eu

BDVA is taking care…
of many different aspects of the big data technology

TF1: Programme

TF2: Impact

TF3: Community

HPC

TF4: Communication

TF5: Policy & Societal

Policy & Societal

TF6: Technical

- TF6-SG1: Data Management
- TF6-SG2: Data Processing Architectures
- TF6-SG3: Data Analytics
- TF6-SG4: Data Protection and Pseudonymisation Mechanisms
- TF6-SG5: Advanced Visualisation and User Experience
- TF6-SG6: Standardisation

TF7: Application

- TF7-SG1: Emerging Application Areas
- TF7-SG2: Telecom
- TF7-SG3: Healthcare
- TF7-SG4: Media
- TF7-SG5: Earth observation & geospatial
- TF7-SG6: Smart Manufacturing Industry
- TF7-SG7: Mobility and Logistics
- TF7-SG8: Smart Cities

TF8: Business

- TF8-SG1: Data entrepreneurs (SMEs and startups)
- TF8-SG2: Transforming traditional business (Large Enterprise)
- TF8-SG3: Observatory on Data Business Models

TF9: Skills and Education

- TF9.SG1: Skill requirements from European industries
- TF9SG2: Analysis of current curricula related to data science
- TF9.SG3: Liaison with existing educational projects

BDVA is taking care…
of many different aspects of the big data technology

TF1: Programme

TF2: Impact

TF3: Community

TF4: Communication

TF5: Policy & Societal

TF6: Technical

- TF6-SG1: Data Management
- TF6-SG2: Data Processing Architectures
- TF6-SG3: Data Analytics
- Protection and Pseudonymisation Mechanisms
- TF6-SG5: Advanced Visualisation and User Experience
- TF6-SG6: Standardisation

TF7: Application

- TF7-SG1: Emerging Application Areas
- TF7-SG2: Telecom
- TF7-SG3: Healthcare
- TF7-SG4: Media
- observation & geospatial
- TF7-SG6: Smart Manufacturing Industry
- TF7-SG7: Mobility and Logistics
- TF7-SG8: Smart Cities

TF8: Business

- TF8-SG1: Data entrepreneurs (SMEs and startups)
- TF8-SG2: Transforming
- (Large Enterprise)
- TF8-SG3: Observatory on Data Business Models

TF9: Skills and Education

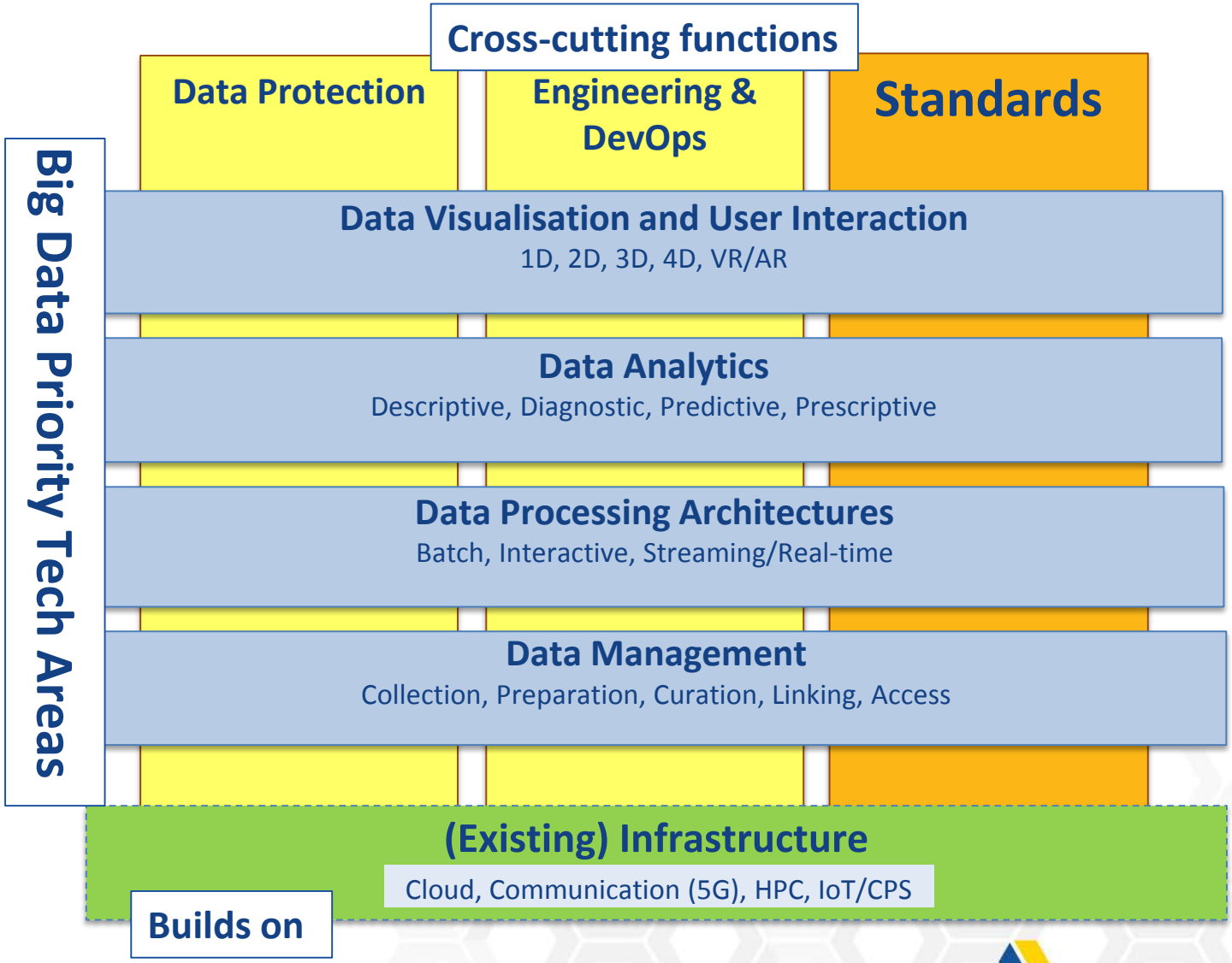
- TF9.SG1: Skill requirements from European industries
- TF9SG2: Analysis of
- related to data science
- TF9.SG3: Liaison with existing educational projects

HPC – Big Data

ETP4HPC – BDVA cPPPs Position Paper (Autumn 2016)

- Cooperation axis:
  - Technology roadmap interaction
  - Development of HPC usage for extreme data exploitation
  - Democratisation of HPC for data intensive applications
- Interaction and exchange in between associations. The BDVA and ETP4HPC agreed to collaborate on a number of strategic issues:
  - Align their SR(I)As to highlight the complementary nature of their technical roadmaps
  - Provide an update to the relevant governance body of each association at least once a year (meeting BDVA – ETP4HPC) Pdts or/and Se
  - Actively participate in each other's events (e.g. BDVA Summit, HPC Summit) including workshops
  - to focus on stakeholder interactions
  - Appoint liaisons to support interaction and to increase cooperation
  - Annual Technical interlock workshop between both associations

Big Data Value Reference Model



www.bdva.eu

BDV BIG DATA VALUE

Sectors: Manufacturing, Health, Energy, Media, Telco, Finance, EO, SE ..

SEMANTICS

DATA

USER

Data Visualisation and User Interaction
1D, 2D, 3D, 4D, VR/AR

Data Analytics
Descriptive, Diagnostic, Predictive, Prescriptive
Machine Learning and AI, Statistics,

Data Processing Architectures
Batch, Interactive, Streaming/Real-time

Data Protection,
Anonymisation, …

Data Management
Collection, Preparation, Curation, Linking, Access
DB types: SQL, NoSQL (Document, Key-Value, Coloum, Array,Graph, …)

ENABLERS

(Existing) Infrastructure, other PPPs
Cloud, Communication (5G), HPC, IoT/CPS

BDV BIG DATA VALUE

Sectors: Manufacturing, Health, Energy, Media, Telco, Finance, EO, SE ..

SEMANTICS

Quality of results
of processing
of experience

QoS

D A T A

U S E R

Data Visualisation and User Interaction
1D, 2D, 3D, 4D, VR/AR

Data Analytics
Descriptive, Diagnostic, Predictive, Pr
Machine Learning and AI, S

Data Processi ures
Batch, Inte g/Real-time

ata Protection,
Anonymisation, …

Data Management
Collection, Preparation, Curation, Linking, Access
DB types: SQL, NoSQL (Document, Key-Value, Coloum, Array,Graph, …)

ADDED VALUE

ENABLERS

(Existing) Infrastructure, other PPPs
Cloud, Communication (5G), HPC, IoT/CPS

BDV BIG DATA VALUE

CHALLENGES

Quality of results
Quality of processing
Quality of experience

- New algorithms
- New (more efficient) implementations
- New (Data Science as a) Service
  - Semantics of use / Automated Decision Making
- New (Data Science as a) Methodology
  - Knowledge modeling and representation (for users)
- Self-adaptive big data systems

CHALLENGES

INFLUENCED BY

Quality of results
Quality of processing
Quality of experience

DOMAIN

- New algorithms

INFRA STRUCTRE

- New (more efficient) implementations
- New (Data Science as a) Service
  - Semantics of use / Automated

DOMAIN

- New (Data Science as a) Methodology
  - Knowledge modeling and representation

INFRA STRUCTRE

- Self-adaptive big data systems

AI

CHALLENGES RELEVANT FOR HPC

PERFORMANCES

DOMAIN
- New algorithms

INFRA STRUCTRE
- New (more efficient) implementations
- New (Data Science as a) Service
  - Semantics of use / Automated Decision Makin

SEMANTICS
- New (Data Science as a) Methodology
  - Knowledge modeling and representation

INFRA STRUCTRE
- Self-adaptive big data systems

INTELLIGENCE

EPDA / HPDA
Extreme-Performance Data Analytics

*Table 2: Big Data and HPC Workloads**

Lower <----->higher Data Intensive	**Typical Big Data** (Big Data + Small Compute): Analytics intense workloads focusing on inferring new insights from big data-sets e.g. search, streaming, data preconditioning, pattern recognition, etc	**Extreme-Performance Data Analytics** (Big Data + Big Compute): Extreme analytics and simulation workloads e.g. near/real-time anomaly detection such as national security or financial services, modelling brain connections or drug efficacy, etc
	Traditional Enterprise IT (Small Data + Small Compute): Traditional enterprise IT workloads focusing on typical database and application hosting e.g. cloud, etc	**Typical HPC** (Small Data + Big Compute): Modelling and simulating intense workloads focusing on interaction amongst parts of a system and the system as a whole e.g. physics, biology, chemistry, etc
	Compute Intensive	
	Lower <-------->higher	

*'Small' and 'Big' are relative (and not absolute) terms.

HPDA strong growth over forecasted period

- The High Performance Data Analytics (HDPA) market is estimated to grow from USD 25.71 Billion in 2016 to USD 78.26 Billion by 2021, at a Compound Annual Growth Rate (CAGR) of 24.9%.
- The proliferation of open source frameworks for big data analytics and the ability of powerful High Performance Computing (HPC) systems to process data at higher resolutions are the key driving forces of HPDA market.
- HPDA applications are used for computational analysis, information escalated research, rich media, 3D-modelling, seismic handling, data mining, and large scale simulation.
- By: marketsandmarkets.com
  Publishing Date: November 2016

However, **high investment costs and government rules & regulations** are acting as the restraining factors for the HPDA market.

HPDA Market Drivers

- **More input data (ingestion)**
  - More powerful scientific instruments/sensor networks
  - More transactions/higher scrutiny (fraud, terrorism)
- **More output data for integration/analysis**
  - More powerful computers
  - More realism
  - More iterations in available time
- **The need to pose more intelligent questions**
  - Smarter mathematical models and algorithms
- **Real time, near-real time requirements**
  - Catch fraud before it hits credit cards
  - Catch terrorists before they strike
  - Diagnose patients before they leave the office
  - Provide insurance quotes before callers leave the phone

Semantics-based

Big Data opportunities

Predictive systems

IDC Analyze the Future

Use Case: PayPal Fraud Detection
The Problem

Detecting fraud in 'real time' as millions of transactions are processed between disparate systems at volume.

Finding suspicious patterns that we don't even know exist in related data sets

Knowledge-based

Ability to create and deploy new fraud models into event flows quickly and with minimal effort.

Self-adaptive systems

Provide environment for fraud modeling, analytics, visualization, M/R, dimensioning and further processing.

Big Data opportunities

IDC Analyze the Future

There Are New Technologies That Will Likely Cause A Mass Explosion In Data – Requiring HPDA Solutions

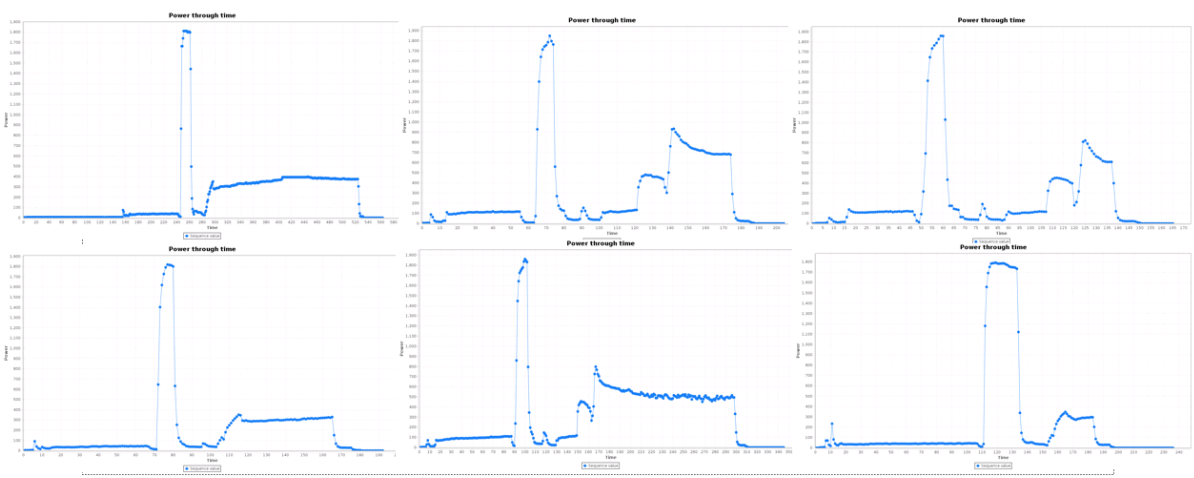
IDC
Analyze the Future

HPDA – analytics challenges

- **Affordable Deep Learning analytics**
  - Algorithms to extract complex data models. Training becomes very complex due to the significant increase in the number required layers in the network.
- **Large scale predictive analytics**
  - Find Unknown Unknown in highdimensional, heterogenous spaces
- **Continuous real-time analytics**
  - Ensure freshens of data in analytics
- **Edge Analytics**
  - Autonomous analytics

Examples
Continuous real-time analytics Challenges

- Continuous analytics deals with unknown unknowns (proper models are not available).
- Continuous analytics is sliding window based data analytics which should be supported by appropriate data staging.
- Scaling out continuous data analytics with parallel and distributed infrastructure (e.g. to calculate distance).



Which metrics can find similarites?

BDV BIG DATA VALUE

Emerging HPDA Segments

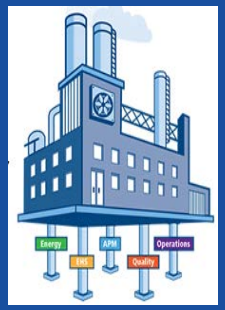
1. **Fraud and anomaly detection.** This "horizontal" workload segment centers around identifying harmful or potentially harmful patterns and causes using graph analysis, semantic analysis, or other high performance analytics techniques.
2. **Marketing.** This segment covers the use of HPDA to promote products or services, typically using complex algorithms to discern potential customers' demographics, buying preferences and habits.
3. **Business intelligence.** The workload segment uses HPDA to identify opportunities to advance the market position and competitiveness of businesses, by better understanding themselves, their competitors, and the evolving dynamics of the markets they participate in.
4. **Other Commercial HPDA.** High-potential workload is the use of HPDA to manage large IT infrastructures, ranging from on premise data centers to public clouds and Internet-of-Things (IoT) infrastructures.

(instead) CONCLUSION

Quality of results
of processing
of experience

QoS

Traditional BigData



Data-intensive workloads
Infrastructure and software are **optimised for cost (IOPS) first**, rather than maximum performance

EPDA

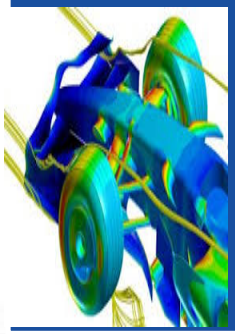
Compute- and Data intensive workloads:
Design principles for infrastructure and software are optimised for minima cost **(IOPS)** and maxima performance **(FLOPS)**

Enterprise IT

'Regular' workloads

Infrastructure and software are **optimised for lower-cost, lower-performance**

HPC



Compute-intensive workloads:
Design principles for infrastructure and software are **optimised for performance (FLOPS) first**, rather than for minimal cost.

Lower <------->higher
Data Intensive

Compute Intensive
Lower <-------->higher

USER

www.bdva.eu

BDV BIG DATA VALUE

ETP4HPC – BDVA cPPPs cooperation

- Existing BDV-PPP Projects: Inviting an ETP4HPC representative to attend Technical Board meetings which will encompass all BDV-PPP funded projects, including Lighthouses and Innovation Spaces.
  - The presence of an ETP4HPC representative can assist with identifying where HPC capabilities may assist big data analytics within these projects.
- Cross SR(I)A Influence: Identify and align research themes common to HPC and Big Data research interests in the area of Extreme Performance Data Analytics (EPDA).
- Extreme-Scale Demonstrator (EsDs): Make a joint recommendation for an 'extreme Big Data' related theme to upcoming Extreme-Scale Demonstrator (EsDs) call for input.
- Centre of Excellence (CoE) in Computing Applications: Make a joint recommendation for a themed Big Data/HPC investigation in forthcoming call for input.
- Additionally, both associations will collaboratively schedule a set of forthcoming interactions involving both ecosystems (e.g. joint events, workshops, and conferences) to advance understanding and definition in these areas.

THANKS

Questions?