Progress on Big Data challenges

Ralph Niederberger
Jülich Supercomputing Center (FZJ)
r.niederberger@fz-juelich.de

4th Wise Workshop, Amsterdam, The Netherlands
March, 27th 2017

What is Big Data all about

Everyone talks about BIG Data, but what is the real definition of Big data?

Gabler economy encyclopedia:
Large amount of data coming from the areas of Internet, mobile networks, finance, energy, health, and traffic, as well as from intelligent agents, social media, credit cards, smart metering systems, security cameras, credit cards …
being stored, proccessed and analysed by purpose built systems.

Gartner.com
High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Characteristics of Big Data

Fast data insertion – vast amount of data generated every second

Distribute redundant data storage – distributed filesystems

Parallel task processing – unstructured data

DIfferent types of data – conversations, video, images, sensors

Scalable – very large data sets across a vast number of systems

Large scale analysis – reducing amount of data effectively

Hardware agnostic – effectiv analysis independent of underlying hardware

Accessibility - easy and continous access to data

Cost effectiveness – No costly traditional RDBMS

Big Data technologies

Today increasing number of devices, sensors and people generate, share, and access data.

Data volumes have become so large, that conventional processing methods do not scale.

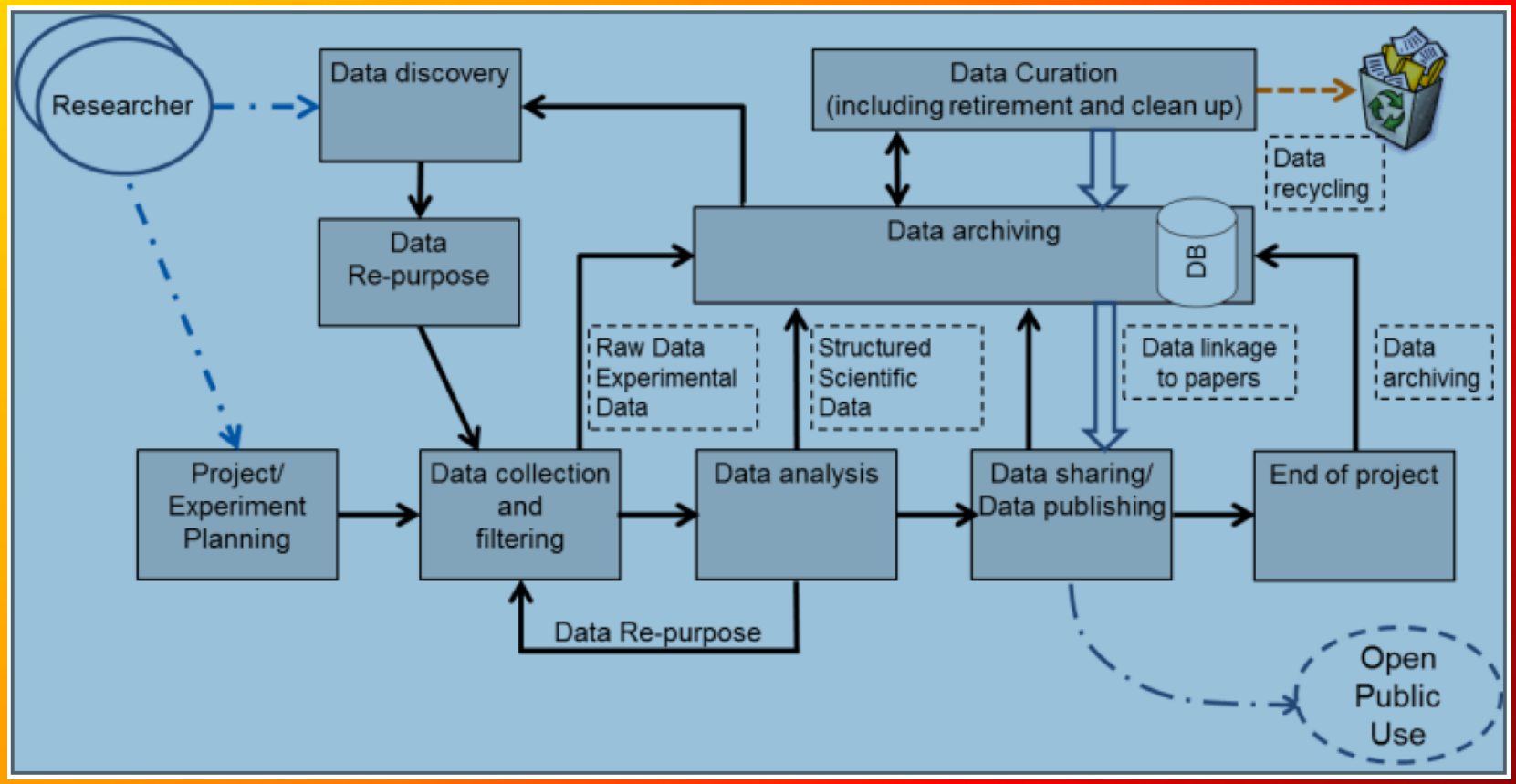
Nowadays we see decreasing storage costs, better storage solutions and algorithms.

Big Data technologies can be defined as

„New generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-Velocity capture, discovery, and/or analysis.“

Source: EMC/IDC, „The Digital Universe“ Study 2014

Scientific Data Lifecycle Management



Source: Demchenko Y., Ngo C., de Laat C., Membrey P., Gordijenko D. (2014) Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. In: Jonker W., Petković M. (eds) Secure Data Management. SDM 2013. Lecture Notes in Computer Science, vol 8425. Springer, Cham

Security impacts on SDLM in collaborating e-infrastructures

- "Generic data lifecycle including data acquisition/collection, filtering and classification, processing and analytics, visualization and delivery" based on UID/GID does not work anymore.
- We face
  - Research projects, communities, different IDs of singe entities
  - Generation systems, networks, data processing systems, storage systems
  - distributed wordwide (instruments, HPCs, clouds, networks)
  - with diverse infrastructure set ups, security policies, national laws
  - widespread administrative personal (different cultures)

  We need
  - Data centric Access Control with
  - fine-granular access control policies

The „V“s of Big Data

Traditionally there are three „V“s of data :

Volume – the size of the data
Velocity – speed as data arrives
Variety – data from various sources, (semi) structured and unstructured

Further „V“s have been added recently:

Veracity – trust into data (concerning source of data and correctness)
Value – inherent economic and social wealth in any data set
Volacity – tendency of data to change structure
Validity – appropriateness of data for its intended use

Areas of big data

Three areas to look into when discussing Security in Big and Open data

The data itself –

- high value of data and therefore valuable target
- Security not fundamentally different to „normal“ data,
- but high speed of aggregation leads to a lot of new security challenges

The analytics of data -

- Integration of different technologies done sluttish often,
- introduce new security challenges, to be addressed properly

The presentation of the results of the analytics

- Systems are complex and heterogeneous.
- Approach must be holistic

Resulting most prominent security challenges in Big Data

- Access control and authentication
- Secure data management
- Source validation and filtering
- Application software security
- Infrastructure security

To address these challenges ENISA describes in its “Big Data Security” report from 2015 several mitigation measures and good practices.

Strong and scalable encryption

- Encrypt data in transit and at rest, to ensure data confidentiality and integrity
- Ensure proper encryption key management solutions
- Consider timeframe for which data to store – could be very long - data protection
- Design databases with confidentiality in mind, i.e. separate confidential data from other data

Application security

- Use regular security testing procedures to re-assure the level of security, specially after patches or functionality changes.
- Ensure tamper resistant devices to avoid misuse.
- Ensure internal or third party security testing procedures for new and updated components are carried out regularly; evaluations, audits and certification are key elements for the confidence and trust in products and actors.

Standards, Certification and commissioning of data processing

- Use devices which comply with desired security standards.
- Ensure obtained certification relates to the use of Big Data.
- Ensure proper
  - risk assessment
  - Service Level Agreements
  - resource isolation and exit strategies
- Especially for commissioning of data processing (e.g. clouds) this is highly recommended

Source filtering

- Use devices with authentication capabilities to ensure that validation of endpoint sources is possible
- Assign confidence levels on the endpoint sources
- Re-evaluate confidence levels of the endpoints regularly, specially after patches or changes in firmware
- If confidence in endpoint source is low, use it in combination with other higher confidence endpoint sources for taking actions

Access control and authentication

- Use authentication and authorization to ensure that Big Data queries are executed by authorized users and entities only
- Use components in the Big Data system that follow same security standards to maintain the desired level of security

Monitoring and logging

- Enable logging on nodes participating in the Big Data computation
- Enable logging on databases (relational or not) , as well as Big Data applications
- Detect and prevent modification of logs
- Regularly test the restoration of Big Data backups considering the vast amount of data being used in the system

So what can be done by WISE

- Publicize proper Authentication methods
- Look into SSO techniques for distributed work on Big Data (generation, analysis, and storage)
- Advice application developers to work on Appropriate Access Control policies on micro-level data structures
- Advertise good practices for Big Data on
  - High speed data transfer techniques (encryted or clear text) and
  - storage solutions especially distributed storage

Some references

- Big Data Security – Good Practice and Recommendations on the Security of Big Data Systems, ENISA, Dec. 2015, ISBN 978-92-9204-142-7, DOI 10.2.2824/13094
- Big Data Threat Landscape and Good Practice Guide, ENISA, January 2016
- Demchenko Y., Ngo C., de Laat C., Membrey P., Gordijenko D. (2014) Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. In: Jonker W., Petković M. (eds) Secure Data Management. SDM 2013. Lecture Notes in Computer Science, vol 8425. Springer, Cham
- NIST Big Data Interoperability Framework: Volume 4, Security and Privacy, Sep. 2015, http://dx.doi.org/10.6028/NIST.SP.1500-4

Questions