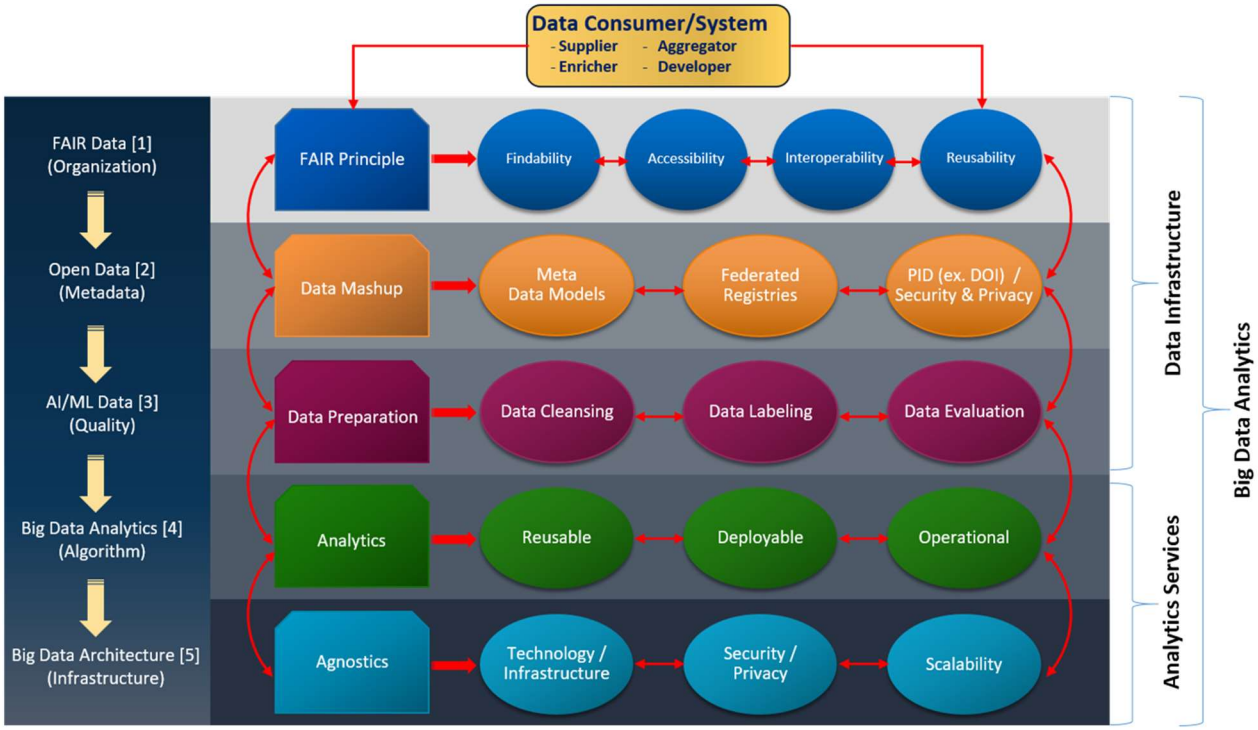
NIST Big Data Analytics and Beyond Roadmap (DRAFT)

Wo Chang, NIST/ITL

September 28, 2019



[1] FAIR Data (https://www.rd-alliance.org/groups/go-fair-ig)
Utilizes FAIR Principles to enable Findability, Accessibility, Interoperability, and Reusability between datasets.

Key technologies

- Metadata – provides data information such as location, access rights, formats, structures, etc.

Group Involved

- Research Data Alliance (RDA) GO FAIR Interest Group

[2] Open Data (https://ieeesa.io/BDGMM)
Enables data mashup among heterogeneous datasets from diversified domain repositories through machine readable and actionable standard data infrastructures.

Key technologies

- Meta Data Models – provide mechanisms to map diversified data models or concepts between datasets.
- Federated Registries – provide distributed databases that store data information such as:
    - Data Catalog Registry: PID-based administration metadata information to provide high-level dataset description. Information may include producer name, location, access rights, licensing, etc.
    - Data Types Registry: PID-based domain specific metadata information to describe low-level basic data types, data properties, derived types, and derived properties. Information may include data description, format, structure, types, properties, versioning, unit of measurement, etc.

- Persistent Identifier (PID) – global unique identifier for digital objects that is resolvable, accessible, and actionable over the Internet.
- Security and Privacy – Provides multiple levels of protection: (a) end-to-end over the net, (b) at repository, (c) at dataset, (d) at data record/element, etc.

Group Involved

- IEEE Big Data Governance and Metadata Management (BDGMM)

[3] AI/ML Data (https://www.iso.org/committee/6794475.html)

Provides data preparation processes via data cleansing, data labeling, and data evaluation so that high-quality datasets can be produced, used, and/or trained by algorithms such as legacy statistical methods, machine learning, and deep learning.

Key technologies

- Data Cleansing – Utilizes various scripting tools to handle input data for alignment, noisy, missing values, outliers, etc. and utilizes output standard formats (XML, JSON, CSV, etc.) to support variety of applications.
- Data Labeling – Apply standard metadata and ontology (mapping) solutions to support variety of applications.
- Data Evaluation – Apply data evaluation metrics to understand the quality of the produced data compared from the original/raw data.

Group Involved

- ISO/IEC JTC 1/SC 42 (AI)/WG 2 (Big Data)

[4] Big Data Analytics (https://bigdatawg.nist.gov/)

Applies DevOps practices to package analytic algorithms/tools with well-defined input and output parameters as analytics payload services/libraries so that they can be reusable, deployable, and operational across multi-cores, many CPUs, and many GPUs computing platforms.

Key technologies

- DevOps Toolchains – Utilizes virtual machine containers to package system configuration and analytics tools as services (i.e., microservices) which can be deployed via NBD-RA Interfaces.

Group Involved

- NIST Big Data Public Working Group (NBD-PWG)

[5] Big Data Architecture (https://bigdatawg.nist.gov/)

Provides vendor-neutral, technology- and infrastructure-agnostic architecture which comprises functional components connected by interoperability interfaces (i.e., services). The goal is to enable system engineers, data scientists, software developers, data architects, and senior decision makers to deploy Big Data reusable analytics packages within an interoperable Big Data ecosystem.

Key technologies

- NIST Big Data Interoperability Framework (NBDIF) – same description as above. With total of 9 volume documents.
- System Orchestration – Provides system resources infrastructure management via NBDIF Vol. 8 Reference Architecture Interfaces to enable scalable analytics deployment across various distributed platforms and clusters.
- System Workflow – Provides configurable constructs including sequential and parallel execution, looping, and conditionals between analytics applications via services.

Group Involved

- NIST Big Data Public Working Group; Some NBDIF documents been adopted by SC 42/WG 2.