

Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure

Yuri Demchenko

SNE Group, University of Amsterdam

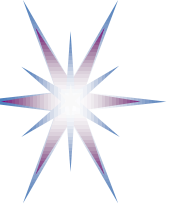
On behalf of

Yuri Demchenko, Canh Ngo, Cees de Laat, Peter Membrey, Daniil
Gordijenko

SDM'13 - Secure Data Management Workshop

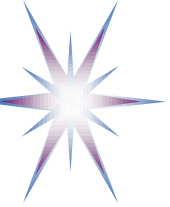
Part of VLDB2013 Conference

30 August 2013



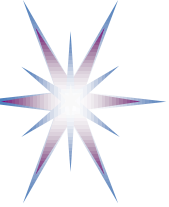
Outline

- Big Data definition
 - 5 + 1 V's of Big Data: Volume, Velocity, Variety + Veracity, Value, Variability
 - 5 parts Big Data Definition
- Paradigm change and new challenges
 - Big Data Infrastructure and Big Data Security
 - CSA's Top 10 Big Data Security Challenges
- Defining Big Data Architecture Framework (BDAF)
 - From Architecture to Ecosystem to Architecture Framework
- Big Data Infrastructure (BDI) and Security Infrastructure components
 - Federated Access and Delivery Infrastructure
 - Trusted Infrastructure Bootstrapping Protocol
- Big Data Security Research topics



Big Data and Security Research at System and Network Engineering, University of Amsterdam

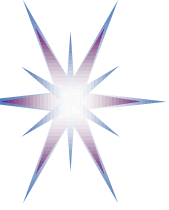
- Long time research and development on Infrastructure services and facilities
 - High speed optical networking and data intensive applications
 - Application and infrastructure security services
 - Collaborative systems, Grid, Clouds and currently Big Data
- Focus on Infrastructure definition and services
 - Software Defined Infrastructure based on Cloud/Intercloud technologies
 - Dynamically provisioned security infrastructure and services
- **NIST Big Data Working Group**
 - Active contribution Reference Architecture, Big Data Definition and Taxonomy, Big Data Security
- **Research Data Alliance**
 - Interest Group on Education and Skills Development on Data Intensive Science
- **Big Data Interest Group at UvA**
 - Non-formal but active, meets two-weekly
 - Provides input to NIST BD-WG and RDA



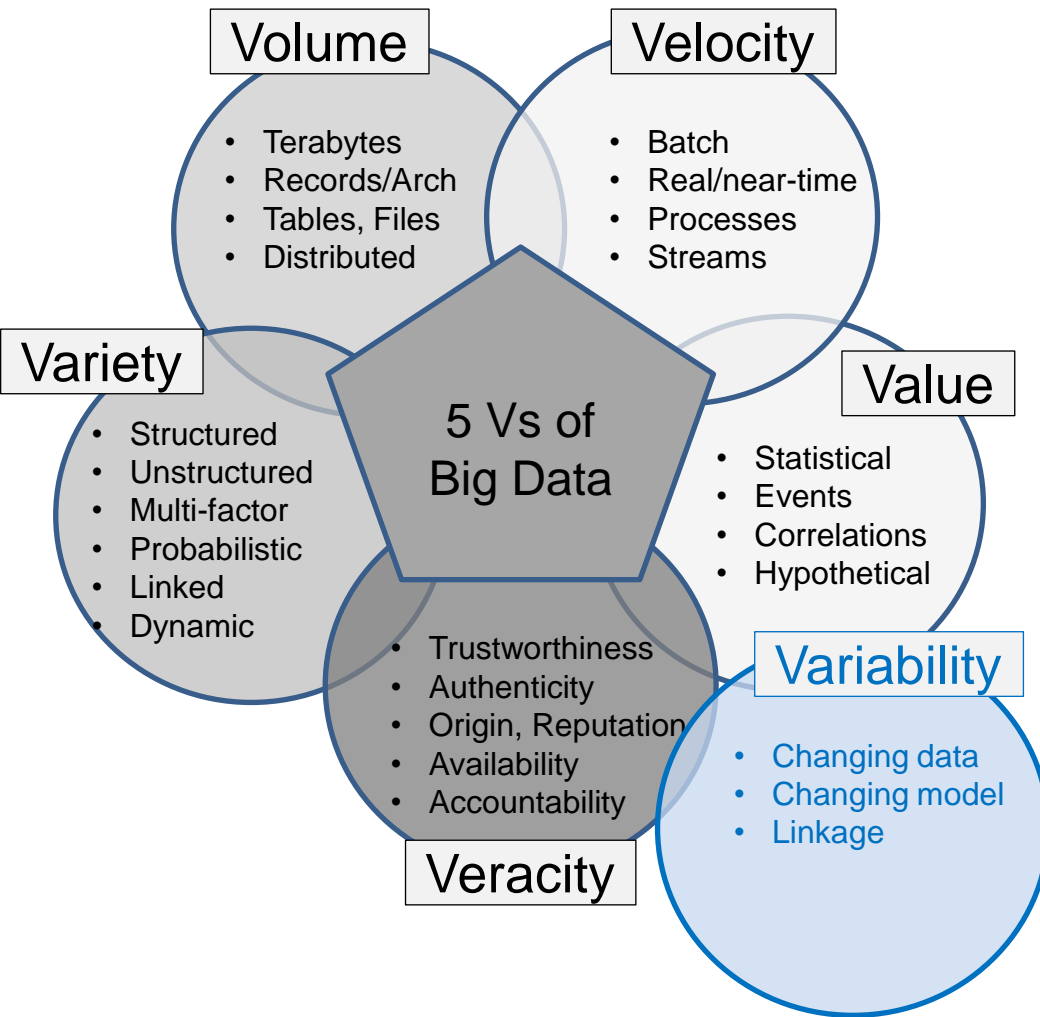
Big Data Definition Collection

- IDC definition (conservative and strict approach) of Big Data:
"A 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"
- Gartner definition
Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. <http://www.gartner.com/it-glossary/big-data/>
 - Termed as 3 parts definition, not 3V definition
- Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.
 - From "The Big Data Long Tail" blog post by Jason Bloomberg (Jan 17, 2013). <http://www.devx.com/blog/the-big-data-long-tail.html>
- "Data that exceeds the processing capacity of conventional database systems. *The data is too big, moves too fast, or doesn't fit the structures of your database architectures.* To gain value from this data, you must choose an alternative way to process it."
 - Ed Dumbill, program chair for the O'Reilly Strata Conference
- Termed as the Fourth Paradigm *)
"The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration." (Jim Gray, computer scientist)

*) *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft, 2009.



Big Data Security: Veracity and other factors



- Trustworthiness and Reputation -> -> Integrity
- Origin, Authenticity and Identification
 - Data authenticity and trusted origin
 - Identification both Data and Source
 - Source: system/domain and author
 - Data linkage (for complex hierarchical data, data provenance)
 - Computer and storage platform trustworthiness
 - Accountability and reputation
- Availability
 - Timeliness
 - Mobility (mobile/remote access; from other domain – roaming; federation)
- Accountability
 - As pro-active measure to ensure data veracity



Big Data Definition: From 5V to 5 Parts (2)

Refining Gartner definition:

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”

- Big Data (Data Intensive) Technologies are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand (3) cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) (2) new data models (supporting all data states and stages during the whole data lifecycle) and (4) new infrastructure services and tools that allows also obtaining (and processing data) from (5) a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

(1) Big Data Properties: 5V or (3+3) V

(2) New Data Models

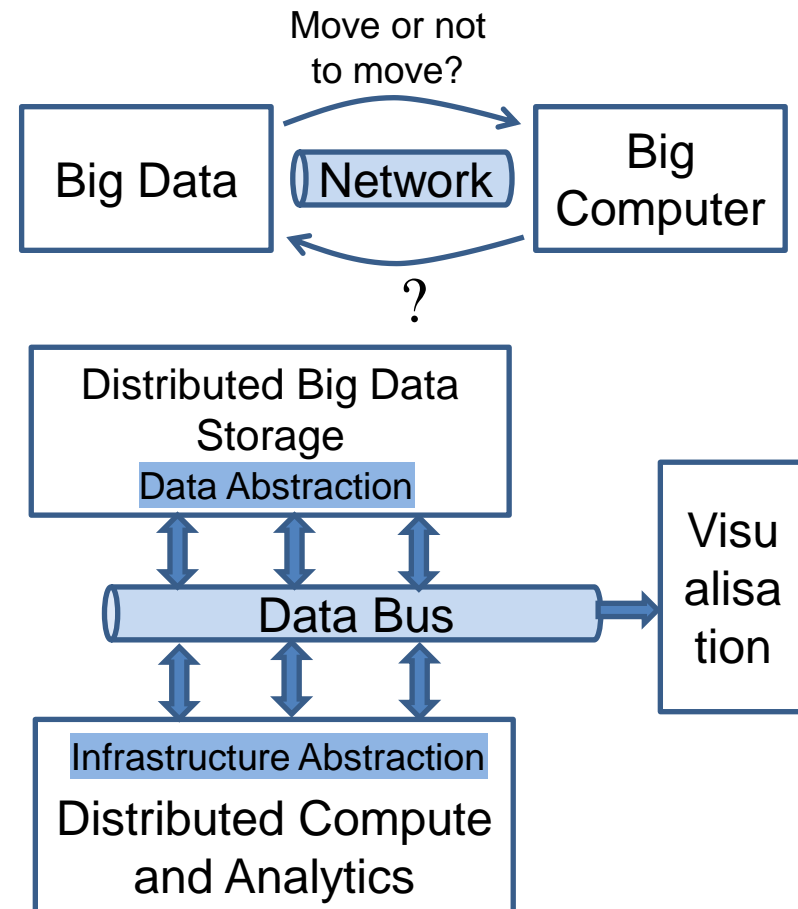
(3) New Analytics

(4) New Infrastructure and Tools

(5) Source and Target

From Big Data to All-Data – Paradigm Change

- Breaking paradigm changing factor
 - Data storage and processing
 - Security
 - Identification and provenance
- Traditional model
 - BIG Storage and BIG computer with FAT pipe
 - Move compute to data vs Move data to compute
- New Paradigm
 - Continuous data *production*
 - Continuous data *processing*
 - *DataBus as Data container and Protocol*





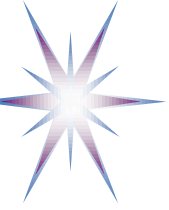
Data Centric Security and Challenges (1)

- Paradigm shift to data centric security model
 - Current and previous security models are host or domain based
 - Any communication or processing is bound to host/computer that runs software, especially in security (PKI as an example)
- Paradigm changing factors
 - **Big Data properties: 5+1 V's**
 - **Data aggregation:** multi-domain, multi-format, variability, linkage, referral integrity
 - **Policy granularity:** variety and complex structure, for their access control processing
 - **Virtualization:** Can improve security of data processing environment but cannot solve data security “in rest”
 - **Mobility** of the different components of the typical data infrastructure: data, sensors or data source, data consumer



Data Centric Security and Challenges (2)

- New security models and new challenges
 - Data confidentiality, integrity and identification
 - Data linkage and referral integrity
 - Data variability and transformation/evolution
 - Data ownership (as related to distributed and evolving data)
 - Data centric access control
 - Encryption enforced access control
 - Personally identified data, privacy, *opacity*
 - Data location, search, access
 - Trusted virtualisation platform
 - Dynamic trust bootstrapping



CSA Top Ten Big Data Security and Privacy Challenges

A. Infrastructure security

- 1) Secure computations in distributed programming frameworks
- 2) Security best practices for non-relational data stores
- 3) Secure data storage and transactions logs
- 4) End-point input validation/filtering

B. Access control and policy

- 5) Granular access control and data centric access policies
- 6) Cryptographically enforced access control and secure communication

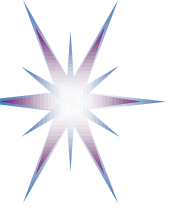
C. Data Management

- 7) Real-time security/compliance monitoring
- 8) Granular audits
- 9) Data provenance

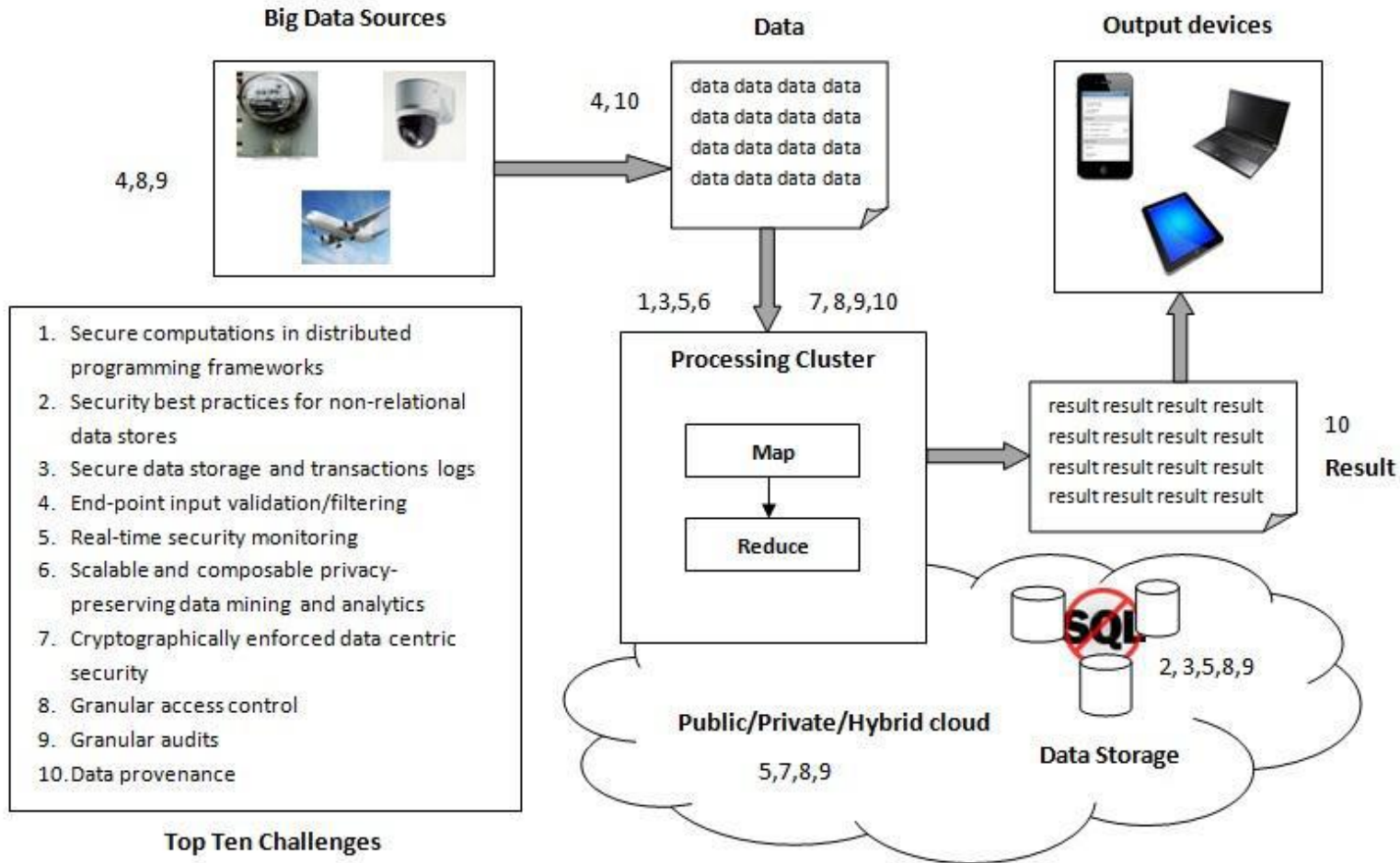
D. Privacy and Confidentiality

- 10) Scalable and composable privacy-preserving data mining and analytics

https://downloads.cloudsecurityalliance.org/initiatives/bdwdg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf



Top Ten Big Data Security and Privacy Challenges

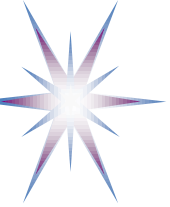


Expanded Top Ten Big Data Security and Privacy Challenges. CSA Report, 16 June 2013.
https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf



Defining Big Data Architecture Framework

- Existing attempts don't converge to something consistent: ODCA, TMF, NIST
 - See http://bigdataawg.nist.gov/uploadfiles/M0055_v1_7606723276.pdf
- **Architecture vs Ecosystem**
 - Big Data undergo a number of transformations during their lifecycle
 - Big Data fuel the whole transformation chain
 - Data sources and data consumers, target data usage
 - Multi-dimensional relations between
 - Data models and data driven processes
 - Infrastructure components and data centric services
- **Architecture vs Architecture Framework (Stack)**
 - Separates concerns and factors
 - Control and Management functions, orthogonal factors
 - Architecture Framework components are inter-related



Big Data Architecture Framework (BDAF) – Aggregated (1)

(1) Data Models, Structures, Types

- Data formats, non/relational, file systems, etc.

(2) Big Data Management

- Big Data Lifecycle (Management) Model
 - Big Data transformation/staging
- Provenance, Curation, Archiving

(3) Big Data Analytics and Tools

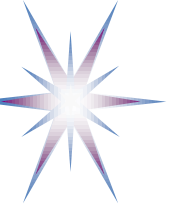
- Big Data Applications
 - Target use, presentation, visualisation

(4) Big Data Infrastructure (BDI)

- Storage, Compute, (High Performance Computing,) Network
- Sensor network, target/actionable devices
- Big Data Operational support

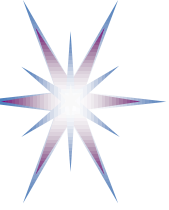
(5) Big Data Security

- Data security in-rest, in-move, trusted processing environments

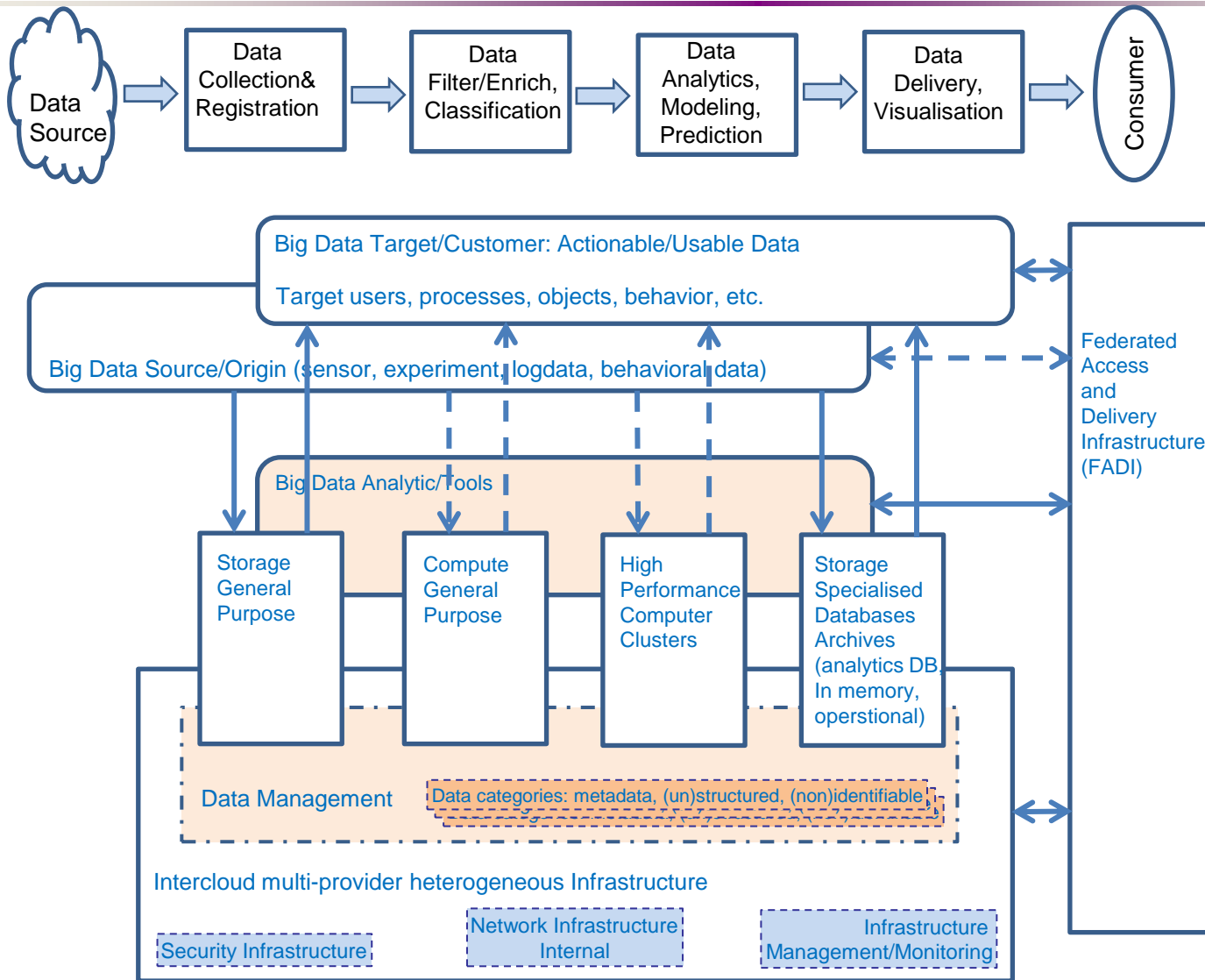


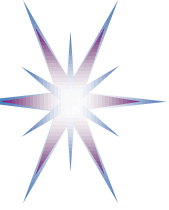
Big Data Architecture Framework (BDAF) – Aggregated – Relations between components (2)

Col: Used By Row: Requires This	Data Models Structrs	Data Managmnt & Lifecycle	BigData Infrastr & Operations	BigData Analytics & Applicatn	BigData Security
Data Models & Structures		+	++	+	++
Data Managmnt & Lifecycle	++		++	++	++
BigData Infrastruct & Operations	+++	+++		++	+++
BigData Analytics & Applications	++	+	++		++
BigData Security	+++	+++	+++	+	

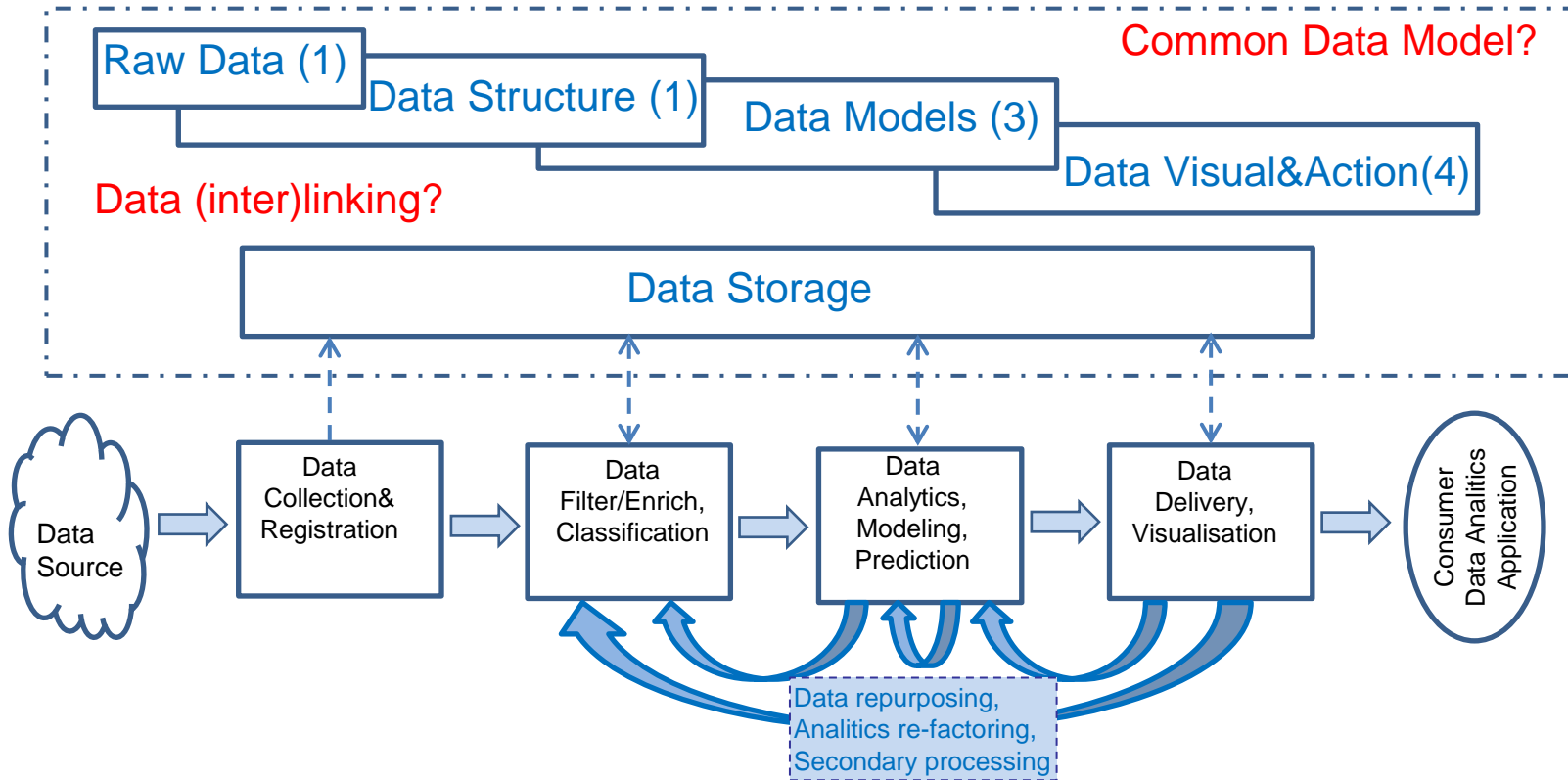


Big Data Ecosystem: Data, Lifecycle, Infrastructure

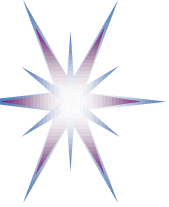




Data Transformation/Lifecycle Model

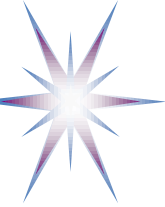


- Data Model and Data structures change along lifecycle
- Data identification and linking
 - Persistent identifier
 - Referral integrity
 - Traceability vs Opacity

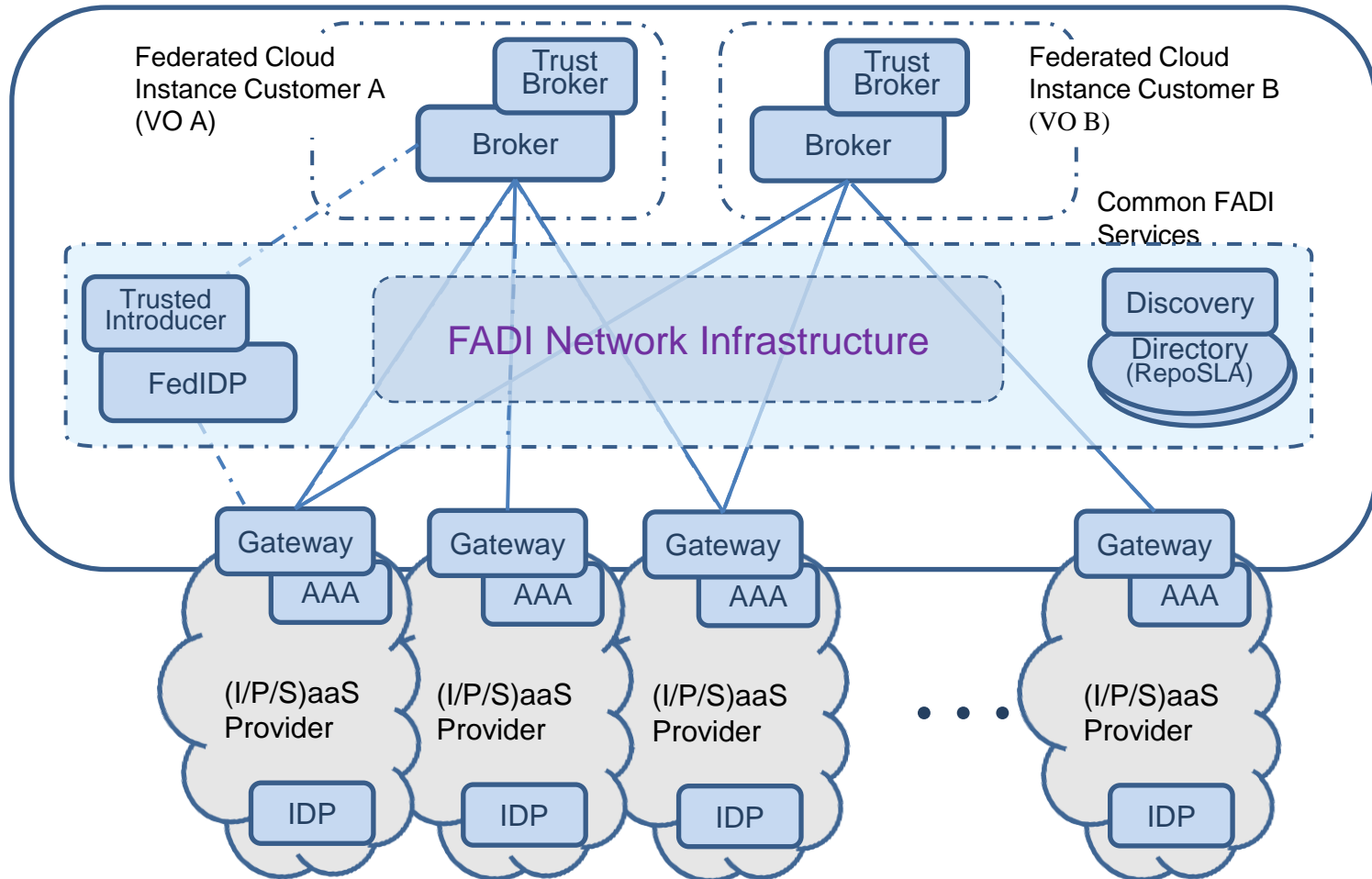


Big Data Infrastructure Security

- Federated Access and Delivery Infrastructure (FADI)
 - Access Control Infrastructure for cloud based Big Data Infrastructure
 - Trusted Virtual Infrastructure Bootstrapping Protocol
- Data centric Access control models
 - Current models with centralised key management
 - Enterprise or Trusted Third Party (TTP)
 - Future/prospective models
 - Looking for solutions from VLDB/SDM community
- Fine grained access control
 - Cell based access control
 - Content based access control



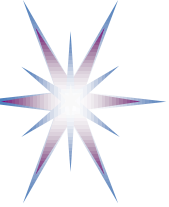
Federated Access and Delivery Infrastructure (FADI)





Virtualised Infrastructure Security: Trusted Infrastructure Bootstrapping Protocol

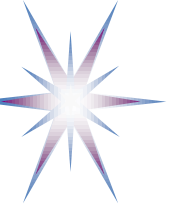
- **Domain Authentication Server (DAS)** provides a trusted root for the third party's domain.
- **Bootstrap Initiator (BI)** is the application that is transferred to the remote machine in order to confirm the machine's status before any infrastructure or software is deployed.
- **Bootstrap Requester (BREQ)** is a client application that runs on the machine responsible for provisioning remote infrastructure. It communicates with its counterpart on the remote machine and handles the first/initial stage of the bootstrapping process.
- **Bootstrap Responder (BRES)** is the counterpart server application. It is responsible for authenticating the machine to a remote client and verifying that the client is authorized to bootstrap the machine. Once each end point has been authenticated, the BRES will receive, decrypt and decompress the payload sent by the client.



Big Data Security: Research directions

- Fined-grained data encryption
 - Key-Policy Attribute-Based Encryption (KP-ABE)[1]: policy \subset keys; attributes \subset ciphertext
 - Ciphertext-Policy Attribute-Based Encryption (CP-ABE) [2]: policy \subset ciphertext, attributes \subset keys
 - Issues: performance, key management model
- Fine-grained access control
 - ABAC access control for structured big-data
 - Fine-grained levels: cell level (like Accumulo), content-based level
 - Configurable authorization policies
 - Challenges: performance; distributed enforcements

-
1. V. Goyal et. al. "Attribute-based encryption for fine-grained access control of encrypted data", CCS '06.
 2. Bethencourt et. al., "Ciphertext-Policy Attribute-Based Encryption," IEEE S&P 2007.



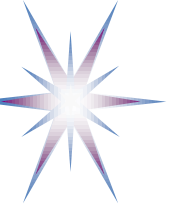
Accumulo cell-level security

Key				Value	
Row ID	Column				Timestamp
	Family	Qualifier	Visibility		

Current Accumulo policy expression (cell based)

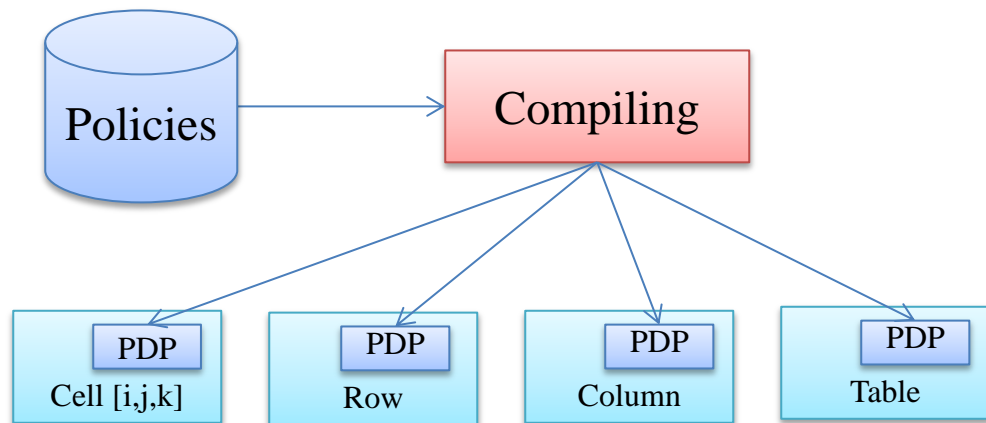
```
username@instance> createtable vistest
username@instance vistest> insert row f1 q1 v1 -1 A
username@instance vistest> insert row f2 q2 v2 -1 A&B
username@instance vistest> insert row f3 q3 v3 -1 (apple&carrot)|broccoli|spinach

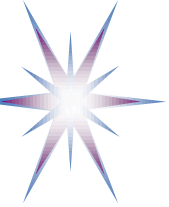
root@instance vistest> setauths -s A,B,broccoli -u username
root@instance vistest> user username
Enter password for user username: ****
username@instance vistest> scan
row f1:q1 [A] v1
row f2:q2 [A&B] v2
row f3:q3 [(apple&carrot)|broccoli|spinach] v3
```



Fined-grained data/content centric access control

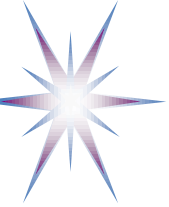
- Using ABAC policies (e.g. XACML)
 - Define authz policies for data granularity: table, columns, rows, cells
 - Compile policies & disperse compiled pieces of policies at each data granularities
 - XACML allows addressing particular policy target with the Resource locator element in a form of URI/URL





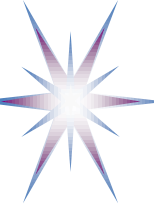
Summary and Discussion

- Initial take on the security issues in Big Data (Infrastructure)
- Required work on Big Data definition and Big Data Architecture Framework
 - Good news: industry via NIST is very active to define Big Data as a new technology and produce Technology Roadmap
 - First deliverable is planned for Sept-Oct 2013
 - 30 Sept 2013 – NIST Big Data Workshop
- Future work: Wide range of topics on defining new security challenges and related solutions



Additional Information

- Existing proposed Big Data architectures
- e-Science and Scientific Data Infrastructure (SDI)



NIST Big Data Working Group (NBD-WG)

- Deliverables target – September 2013
- Activities: Conference calls every day 17-19:00 (CET) by subgroup - <http://bigdatawg.nist.gov/home.php>
 - Big Data Definition and Taxonomies
 - Requirements (chair: Jeffrey Fox)
 - Big Data Security
 - Reference Architecture
 - Technology Roadmap
- BigdataWG mailing list and useful documents
 - Input documents http://bigdatawg.nist.gov/show_InputDoc2.php
 - Brainstorming summary and Lessons learnt (from brainstorming) http://bigdatawg.nist.gov/uploadfiles/M0010_v1_6762570643.pdf
 - Big Data Ecosystem Reference Architecture (Microsoft) http://bigdatawg.nist.gov/uploadfiles/M0015_v1_1596737703.docx



Big Data Definition: From 5V to 5 Parts (1)

(1) Big Data Properties: 5V

- Volume, Variety, Velocity, Value, Veracity
- Additionally: Data Dynamicity (Variability)

(2) New Data Models

- Data linking, provenance and referral integrity
- Data Lifecycle and Variability/Evolution

(3) New Analytics

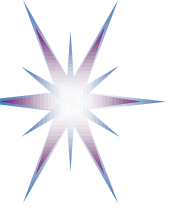
- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools

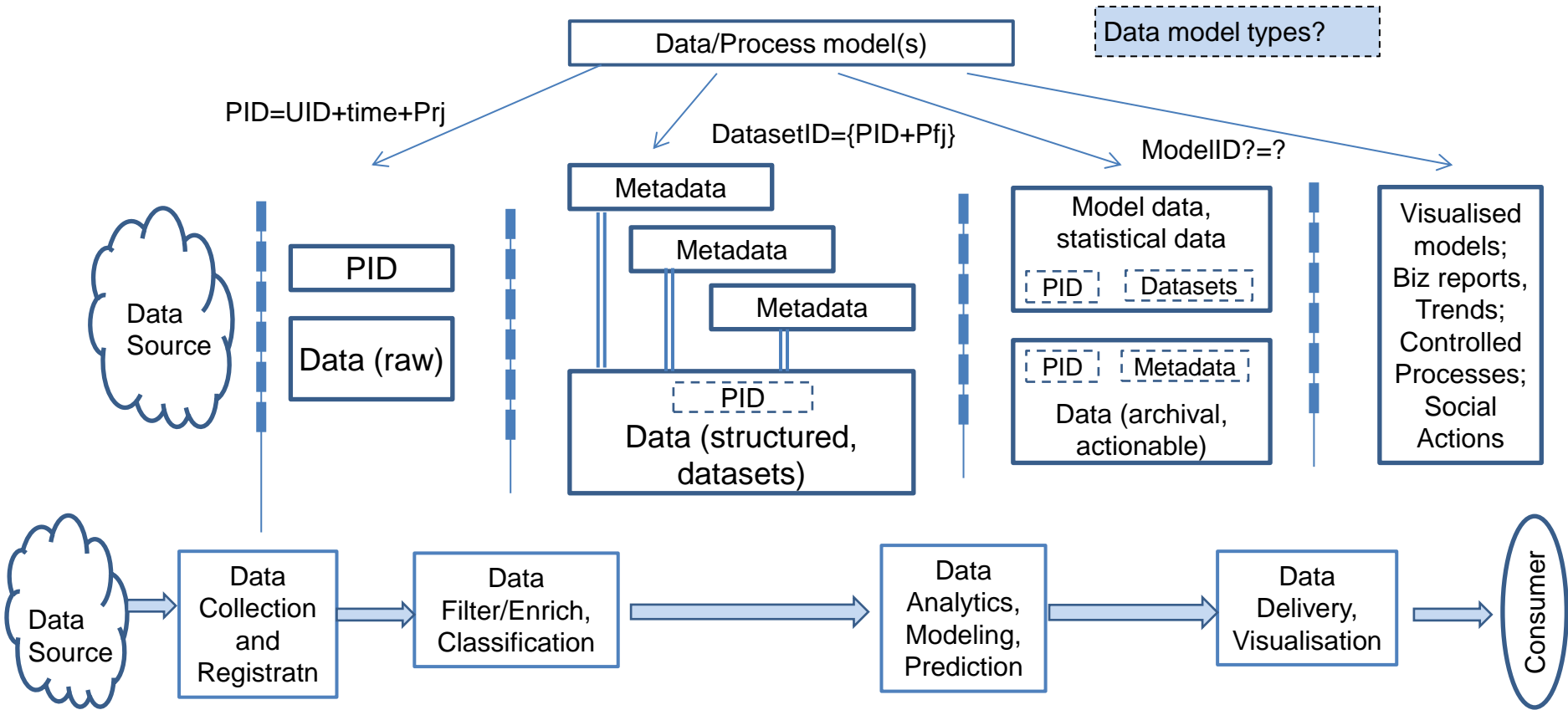
- High performance Computing, Storage, Network
- Heterogeneous multi-provider services integration
- New Data Centric (multi-stakeholder) service models
- New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target

- High velocity/speed data capture from variety of sensors and data sources
- Data delivery to different visualisation and actionable systems and consumers
- Full digitised input and output, (ubiquitous) sensor networks, full digital control



Data Transformation Model



Security issues

- CIA and Access control

- Referral integrity
- Traceability
- Opacity

Scientific Data Lifecycle Management (SDLM) Model

