Design Basis Concerns And Challenges For The Data Environment

T.Chandana Gouri
Assistant Professor-CSE,
Indo American Institutions Technical Campus
tchandana@iaitc.in

Abstract
Big Data are becoming a new technology focus both in science and in industry and motivate technology shift to data centric architecture and operational models. There is a vital need to define the basic information/semantic models, architecture components and operational models that together comprise a so-called Big Data Ecosystem. This paper discusses a nature of Big Data that may originate from different scientific, industry and social activity domains and proposes improved Big Data definition that includes the following parts: Big Data properties (also called Big Data 5V: Volume, Velocity, Variety, Value and Veracity), data models and structures, data analytics, infrastructure and security. The paper discusses paradigm change from traditional host or service based to data centric architecture and operational models in Big Data. The Big Data Architecture Framework (BDAF) is proposed to address all aspects of the Big Data Ecosystem and includes the following components: Big Data Infrastructure, Big Data Analytics, Data structures and models, Big Data Lifecycle Management, Big Data Security. The paper analyses requirements to and provides suggestions how the mentioned above components can address the main Big Data challenges. The presented work intends to provide a consolidated view of the Big Data Phenomena and related challenges to modern technologies, and initiate wide discussion.

Keywords: Big Data, Big Data Challenges, Big Data Ecosystem, Big Data Framework, Big Data Infrastructure

I. Introduction
Big Data, also referred to as Data Intensive Technologies, are becoming a new technology trend in science, industry and business. Big Data are becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer. Current technologies such as Cloud Computing and ubiquitous network connectivity provide a platform for automation of all processes in data collection, storing, processing and visualization.

The goal of our research at current stage is to understand the nature of Big Data, their main features, trends and new possibilities in Big Data technologies development, identify the security issues and problems related to the specific Big Data properties, and based on this to review architecture models and propose a consistent approach to defining the Big Data architecture/solutions to resolve existing challenges and known issues/problems.

In this paper we continue with the Big Data definition and enhance the definition given in that includes the 5V Big Data properties: Volume, Variety, Velocity, Value, Veracity, and suggest other dimensions for Big Data analysis and taxonomy, in particular comparing and contrasting Big Data technologies in e-Science, industry, business, social media, healthcare. With a long tradition of working with constantly increasing volume of data, modern e-Science can offer industry the scientific analysis methods, while industry can bring advanced and fast developing Big Data technologies and tools to science and wider public. In Big Data, data are rather a “fuel” that “powers” the whole complex of technical facilities and infrastructure components built around a specific data origin and their target use. We will call it a Big Data Ecosystem (BDE). By defining BDE we contrast its data centric character to traditional definition of the architecture that is more applicable for facility or service centric technologies. We discuss the major (architecture) components that together constitute the Big Data Ecosystem: 5V Big Data properties, Data Models and Structures, Big Data Infrastructure, Big Data lifecycle management (or data transformation flow), Big Data Security Infrastructure.

There are not many academic papers related to Big Data; in most cases they are focused on some particular technology or solution that reflect only a small part of the whole problem area. The same relates to the Big Data definition that would provide a conceptual basis for the further technology development. There is no well-established
This Reference Architecture can be applied to a range of big data solutions that may be tightly-integrated enterprise systems or loosely-coupled vertical industrial systems. In the architecture the two big value chains are represented namely: the Information Value Chain and the IT Value Chain [7]. Information value chain is represented along the vertical axis. Along the vertical information flow axis value creation is done by tasks such as data collection, integration, analysis and the results are used down the value chain [7]. The IT Value Chain is represented as the horizontal IT axis. Along IT Value Chain the value is created using information regarding networking, platforms, application tools, and IT services needed for hosting, operating and transforming big data for implementing a specific application. At the junction of both axes a transformation block is present indicating special value of big data analytics and implementation for stakeholders in both value chains [7]. The five major building blocks of NIST reference architecture representing technical roles existing in almost every big data system are [7, 9]:

- **Data Provider:** An important characteristic of a Big-Data system is its ability to import and use data from a variety of data sources such as online or offline applications, tapes, images, audio and videos, Web log data, HTTP cookies, etc. Data Provider acts as an abstraction layer for different data sources. It introduces new information into the big data system for discovery, access, and transformation purposes [7]. In case of raw data the provider can cleanse, correct, and store the data in internal format. Thus, the data provider activities can be summarized as:
  a. Creating metadata information describing attributes such as data sources, access rights, usage information etc.
  b. Enabling data accessibility to other components of the architecture using appropriate programmable interfaces.
  c. Collecting and Persisting the data
  d. Keeping track of the availability of the information, its reference and the means to access it
  e. Enforcing access rights on data access and establishing formal or informal contract for data access authorizations.

- **Data Consumer:** Data Consumer refers to the end users or other systems that use the output of data application provider component. In order to access information suited to its own interest, data consumer uses the interfaces (or services) exposed by Big Data Application Provider. Services provided include data reporting, data retrieval, and data
rendering support [7]. Data Consumer activities can be summarized as follows:
  a. Data visualization software can be used for exploring data.
  b. Settling data to appropriate work for the business instance converting information produced by the transformers to business rule.
  c. Importing and processing data for storage or later use into their own system.
  d. Data Query, search and retrieval tasks
  e. Using business intelligence software for creating and analyzing reports.

- **Big Data Application Provider**: The next component is the Big Data Application Provider enables execution of a specific set of data life cycle processes to satisfy the requirements of System Orchestrator along with the Security and Privacy requirements [7]. These processes include data collection from various sources, multiple data transformations using traditional and new technologies, and using data for varied purposes. The Big Data Application Provider activities can be listed as under:
  a. **Data Collection**: Application provider connects to Data Provider APIs to establish connection with local system to collect similar data sets with uniform security considerations. Initial metadata is created for successive aggregation [7].
  b. **Data Preparation**: The following tasks are accomplished under data preparation activity [7]:
     i. Data preparation through cleansing, outlier removal, and standardization of data.
     ii. Data aggregation from various data providers with correlated metadata into a single data set.
     iii. Data matching from different data providers with unrelated metadata into a single data set.
     iv. Data optimization by devising appropriate data manipulations and indexing to optimize transformation processes.
  a. **Data Analysis and Analytics**: Implementation of the techniques to mine knowledge from the data based on the requirements by using specific algorithms [7].
  b. **Visualization**: The following tasks are included under data visualization [7]:
     i. Integration of pertinent visualization tools into data life cycle.
     ii. Formatting and presenting data so that it is clearly understandable.
     iii. Developing statistical charts and diagrams to present analysis outcomes.
  e. **Data Access**: The following tasks are accomplished under data preparation activity [7]:
     i. Identifying and storing data in data repositories for usage, sharing, and re-use purposes.
     ii. Descriptive, administrative, and preservative metadata schemes are developed and used.
  iii. Guarantees secure access to data.

- **Big Data Framework Provider**: Big Data Framework Provider provides a basic computing structure or framework including system hardware, networking structure, storage structure, computing platform etc. to carry out certain transformation tasks, while preserving data privacy and integrity [7].

- **System Orchestrator**: The role of System Orchestrator is to define and integrate the data applications activities into a set vertical system [7]. It represents a collection of more specific roles performed by one or more actors to manage and the operations of the Big Data System. In an enterprise environment, the System Orchestrator role is typically centralized while in a loosely-coupled environment, the System Orchestrator role is typically decentralized [7].

System Orchestrator has the following responsibilities:
  a. Translating business goal(s) to technical requirements.
  b. Integrating both external and internal Data Providers.
  c. Defining requirements for the collection, preparation, and analysis of data.
  e. Establishing System architecture requirements
  f. Auditing data applications activities
  g. Defining data dictionaries and data storage models

The two additional “Security & Privacy” and “Management” blocks are shown enclosing all subsystems and thus providing services and functionality to the rest of the system components in the areas specific to “big data” [7]. These two key considerations are vital and hence are included into any Big Data solution. The data flows between the above listed components may take place either physically by value or by providing its location and the means to access it i.e. by reference and is shown using “DF” arrows. The “ST” arrows represent transfer of software tools for processing of big data and the “SU” arrows represent software programmable interfaces [7].

The design of the Big Data reference architecture has the following limitations:
1. It does not take into consideration the detailed specifications of organizational operational systems
2. Detailed specifications of information exchanges or services are not considered.

standards for integrating infrastructure products are not taken into account in reference architecture.

III. Paradigm change in Big Data and Data Intensive Science and Technologies
The recent advancements in the general ICT, Cloud Computing and Big Data technologies facilitate the
paradigm change in modern e-Science and industry that is characterized by the following features:

- Transformation of all processes, events and products into digital form by means of multidimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content.
- Automation of all data production, consumption and management processes including data collection, storing, classification, indexing and other components of the general data curation and provenance.
- Possibility to re-use and repurpose the initial data sets for new and secondary data analysis based on the model improvement.
- Global data availability and access over the network for cooperative group of researchers, including wide public access to scientific data.
- Existence of necessary infrastructure components and management tools that allow fast infrastructures and services composition, adaptation and provisioning on demand for specific research projects and tasks.
- Advanced security and access control technologies that ensure secure operation of the complex research and production infrastructures and allow creating trusted secure environment for cooperating groups of researchers and technology specialists.

The following are additional factors that will create new challenges and motivate security paradigms change in Big Data ecosystem/technology:

- 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: sensors or data source, data consumer, and data themselves (original data and staged/evolutional data). This in its own cause the following problems o On-demand infrastructure services provisioning o Inter-domain context communication
- Big Data aggregation that may involve data from different administrative/logical domains and evolutionally changing data structures (also semantically different).
- Policy granularity: Big Data may have complex structure and require different and high granular policies for their access control and handling.

The future Big Data Infrastructure (BDI) should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time. Important is that this infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), and data ownership protection. With current needs to process big data that require powerful computation, there should be a possibility to enforce data/dataset policy that they can be processed on trusted systems and/or complying other requirements. Customers must trust the BDI to process their data on BDI facilities and be ensured that their stored research data are protected from nonauthorised access. Privacy issues are also arising from distributed remote character of BDI that can span multiple countries with different local policies. This should be provided by the Access Control and Accounting Infrastructure (ACAI) which is an important component of SDI.

IV. Data Management and Big Data Lifecycle

With the digital technologies proliferation into all aspects of business activities, the industry and business are entering a new playground where they need to use scientific methods to benefit from the new opportunities to collect and mine data for desirable information, such as market prediction, customer behavior predictions, social groups activity predictions, etc. Refer to numerous blog articles suggesting that the Big Data technologies need to adopt scientific discovery methods that include iterative model improvement and collection of improved data, re-use of collected data with improved model.

![Figure. Big Data Lifecycle in Big Data Ecosystem.](image-url)
management and processing in Big Data industry is reflected in the Big Data Lifecycle Management (BDLM) model as a result of analysis of the existing practices in different scientific communities. New BDLM requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in BDI. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed BDLM and must also be done in a secure and trustworthy way.

V. Big Data Security Framework Components

This section discusses the Big Data Security Framework that supports a new paradigm of the data centric security. The following components are included:

- Security lifecycle
- Fine-grained access control
- Encryption enforced access control
- Trusted environment
- FADI for cooperation and services integration

Federated Access and Delivery Infrastructure (FADI)

Federated Access and Delivery Infrastructure (FADI) is defined as Layer 5 in the generic SDI Architecture model for e-Science (e-SDI). It includes federation infrastructure components, including policy and collaborative user groups support functionality.

When implemented in clouds, the FADI and SDI in general may involve multiple providers and both cloud and non-cloud based infrastructure components. Our vision and intention is to use for this purpose the general Intercloud Architecture Framework (ICAF) proposed in our works [19]. ICAF provides a common basis for building adaptive and on-demand provisioned multi-provider cloud based services.

Figure illustrates the general architecture and the main components of the FADI (that corresponds to the ICAF Access and Delivery Layer C5) that includes infrastructure components to support inter-cloud federations services such as Cloud Service Brokers, Trust Brokers, and Federated Identity Provider. Each service/cloud domain contains an Identity Provider IDP, Authentication, Authorisation, Accounting (AAA) service and typically communicate with other domains via service gateway.

SDI/BDI will incorporate standards and if needed advance access control services and mechanisms at the level of FADI and users/services level. However consistent data centric security and access control will require solving the following problems:

- Fine-granular access control policies.
- Encryption enforced attribute based access control

Depending on the data type and format, the two basic access control and policy models can be defined: resource and/or document based access control, including intra-document; and cell or record based access control for data stored in databases. We identify XACML policy language as appropriate for document/intra-document access control. For databases we need to combine their native access control mechanisms and general document based access control.

XACML policies for fine-granular access control

The policies for data centric access control model should provide the fine-grained authorization features, based not only on the request context attributes such as subjects/users, data identifiers, actions or lifetimes, but also on the structured data
content. A prospective direction is to design and apply attribute based access control mechanisms with policies incorporate along with data granularity. Such policies may contain complex logic expressions of attributes. Based on input attribute values from users, their queries could return either authorized data or errors. In this respect, managing SDI/BDI big data using attribute-based policy languages like XACML is applicable. However, for large documents or complex data structures XACML policies evaluation may create a significant performance overhead.

We refer to our experience in developing dynamically provisioned Access Control Infrastructure (DACI) for complex infrastructure services and resources [44]. It uses advanced features of the XACML based policies that allow describing access control rules for complex multi-domain resources, including domain, session context, multi-domain identity and trust delegation [45, 46]. The proposed in [47] the Multi-datatypes Interval Decision Diagrams (MIDD) policy decision request evaluation method allows for significant performance gain for massively large policy sets.

**Access control in NoSQL databases**

The popular NoSQL databases for structured data storage MongoDB, Cassandra, HBase, and Accumulo provide different levels of security and access control. Most of them have coarse-grain authorization features, both on user management and on protected data granularity like table-level or rowlevel security. Accumulo provides the most advanced features to allow cell-level security with which accesses from keys to values are only granted when the submitted attributes satisfy predefined Boolean expressions provided as a security label of the cell key index. However, the current policy language in Accumulo is at early development stage and lacks of features for distributed, multi-domains environments.

**Encryption enforced access control**

Described above solutions are capable to address majority of the problems for data access, transfer and processing stages, however data in-rest when stored on remote facilities may remain unprotected. The solution to this problem can be found with using the encryption enhanced access control policies that in addition to the traditional access control, use also attributes based encryption to allow data decryption only to the targeted subject or attribute owner. We admit such approach as potentially effective and applicable to many data protection use cases in Big Data, in particular, healthcare or targeted broadcast of streaming data that make take place when using distributed sensor networks.

**Trusted Infrastructure Bootstrapping Protocol**

To address the issues with creating trusted remote/distributed environment for processing sensitive data, in our earlier papers. We proposed a generic Dynamic Infrastructure Trust Bootstrapping Protocol (DITBP). This includes supporting mechanisms and infrastructure that takes advantage of the TCG Reference Architecture (TCGRA) and Trusted Platform Module (TPM) [56, 57]. The TPM is used to provide a root of trust that extends from the physical hardware itself. The TPM is used to generate a key pair in hardware where the private key is never revealed (the key pair is non-migratable). The key is only available when the machine is in a known and trusted state. The key pair is used to authenticate the machine and to decrypt the payload which is then executed to bootstrap the rest of the virtual infrastructure.

**VI. Opportunities And Challenges For Big Data**

Big Data environment has started to influence almost all types of organizations, since it has the potential power to extract useful knowledge from huge volumes of data and operate upon it as per the requirements on real time basis [1]. The opportunities provided by Big Data systems can be explained by analyzing its applicability in different areas as below:

1. **Healthcare:** The healthcare industry is quickly moving to electronic medical records and images that it may use for public health monitoring and in epidemiological research programs [11]. In healthcare Big Data is also associated with the massive volume of patient-specific data. A valid example is in medical imaging where small pathological features measuring only a few millimeters can be detected in magnetic resonance imaging and in CT scans [1, 4].

2. **Mobile Networks:** The amount of mobile data traffic is expected to grow to 10.8 Exabyte per month by 2016 due to increased usage of smart phones and tablets [1]. Big Data is needed for managing and operating mobile networks and with the aim of improving network quality and considering issues such as isolation and correlation of network faults, security breach detection, traffic planning, hardware maintenance predictions etc. [1]

3. **Video surveillance:** Video surveillance is in a transition phase from CCTV to IPTV cameras and recording systems that organizations want to analyze for behavioral patterns. Big data can be used to analyze huge volumes of data so generated for security and service enhancement [11].

4. **Media and Entertainment:** The media and entertainment industry has shifted to digital recording, production, and delivery in recent times and Big Data approach could be used for collecting the huge volumes of rich content to find and analyze user viewing behaviors [11].

5. **Life sciences:** In field Life Sciences the low-cost gene sequencing can produce tens of terabytes of
information required to be analyzed to find genetic variations, DNA sequencing, treatment success rate etc. [11].

6. Transportation: Sensor data is being generated at an accelerating rate from fleet GPS transceivers, RFID tag readers, smart meters, and cell phones and this data is being used to optimize business operations to realize incoming business opportunities [11, 12].

7. Environment Study: Efficient environment study requires collecting and analyzing data from thousands of sensors that monitor air and water quality and meteorological conditions [4]. Careful analysis of collected data can then be used to direct simulations of climate and groundwater models for predicting longterm trends and changes in environment such as increased CO₂ emissions, ground water table level etc. [4]

Big data environments create significant opportunities with some of them listed above. However, organizations must find ways to cope with security and other technological challenges they introduce. As big data environments deal with massive volumes of data, organizations have to face significant risks and threats to these data repositories. As compared to previous records nowadays organizations are generating more data but the point of concern is that they don’t realize its importance, context and ways to protect it from any kind of risks. Some such risks and challenges faced by Big Data systems are listed and described below:

8. Heterogeneity and Incompleteness of Data: Data for analysis may be collected from differently structured data sources which possess a great deal of heterogeneity. Human beings can tolerate data heterogeneity but machine analysis algorithms expect homogeneous data [13]. Due to this carefully structured data is needed for data analysis. Issues such as representation, access, and analysis of unstructured and semi-structured data require further processing to be done [1, 13]. Even after undergoing data cleaning and error correction there may still exist some incompleteness and errors in data. Data analysis activity must deal with such data incompleteness and errors [13]. Doing this correctly poses a challenge. However, managing probabilistic data offers a way to tackle the problem to some extent [13].

9. Scale or Shrink: Managing large and rapidly increasing volumes of data is a challenge for any system. Previously, this challenge was handled by processors getting faster, as suggested by Moore’s law to provide resources needed to cope with increasing volumes of data [13]. But, the scenario now is that the data volume is scaling faster than compute resources and CPU speeds are also static. The question then arises is how to scale up or shrink as required. NoSQL can partly address this issue and is more flexible to adopt to new business processes [16].

10. Context awareness: The increasing mobility of users and devices has given rise to the concept of context awareness and the need of suitable and efficient context aware data routing [1]. In view of Big Data systems several available solutions process and route all data at once but the goal should be to take into account the data context information that and provide the right data to the right people and devices [1]. Context-aware Big Data solutions can use context awareness to concentrate on some portions of data only by keeping high probability of hit for all application-relevant events [1]. This approach provides benefits of cost reduction and less complexity. However, research on industrial applicability on how to achieve the best trade-off between limitedness of the considered portions and probability of hits needs significant investigation [1].

11. Performance Issues: Data management performance and scalability are core technical issues to deal with the huge volume of data stored and processed by Big Data environment [1]. This gives rise to the need for effective solutions to deal with the issue of massive data volume and enable feasible, scalable and lucrative storage and processing of massive amounts of data [1].

12. Security and Privacy: The less structured and informal nature of Big Data approaches generates the problems of security and privacy. If the data involved is sensitive then using such approaches may represent a serious security breach. The privacy of data is another huge concern that further increases in the Big Data systems. Privacy security is a technical as well as a sociological problem that must be addressed from both perspectives to effectively exploit the benefits of big data.

Other Challenges: Some other challenges such as timeliness of data analysis i.e. how much time it will take to analyze data, the problems related Distributed Storage structures, content validation, Stream processing and real time analytics etc. also need further exploration in context of Big Data environment. Big Data deployments also require new skill sets for IT administration and application development, and people possessing these skills may be in short supply.

VII. Future Research and Development
The future research and development will include further Big Data definition initially presented in this paper. At this stage we tried to summarize and rethink some widely used definitions related to Big Data, further research will require more formal approach and taxonomy of the general Big Data use.
cases in different Big Data origin and target domains. The authors will continue contributing to the NIST Big Data WG targeting both goal to propose own approach and to validate it against industry standardization process. Another target research direction is defining a Common Body of Knowledge (CBK) in Big Data to provide a basis for a consistent curriculum development. This work and related to the Big Data metadata, procedures and protocols definition is planned to be contributed to the Research Data Alliance (RDA). The authors believe that the proposed paper will provide a step toward the definition of the Big Data Architecture framework and Common Body of Knowledge (CBK) in Big Data and Data Intensive technologies.

VIII. Conclusion

Big Data technologies represent a new generation of architectures and technologies developed in order to extract value from a very huge volumes of a wide variety of data by innovatively enabling high-velocity data capture, discovery, and analysis. It is a term used for large and complicated data sets that are difficult to be processed by standard traditional data processing applications and tools. Big data has established the ability to improve performance, save cost, efficient data processing and better decision-making in diverse fields of application such as traffic control, healthcare weather forecasting, fraud control, media and entertainment, disaster prevention, education etc. Big data poses opportunities and challenges in its application areas that need further significant research efforts. This paper presented a review on the recent efforts dedicated to big data, NIST proposed reference architecture and opportunities and challenges posed by Big Data environment.

References:

[8] “Data Modeling for Big Data”, by Jinbao Zhu, Principal Software Engineer, and Allen Wang, Manager, Software Engineering, CA Technologies