Chapter 10
Big-Data-Based Architectures and Techniques:
Big Data Reference Architecture

Gopala Krishna Behara
Wipro Technologies, India

ABSTRACT

This chapter covers the essentials of big data analytics ecosystems primarily from the business and technology context. It delivers insight into key concepts and terminology that define the essence of big data and the promise it holds to deliver sophisticated business insights. The various characteristics that distinguish big data datasets are articulated. It also describes the conceptual and logical reference architecture to manage a huge volume of data generated by various data sources of an enterprise. It also covers drivers, opportunities, and benefits of big data analytics implementation applicable to the real world.

INTRODUCTION

In Information Age, we are overwhelmed with data, ways to store, process, analyze, interpret, consume and act upon the data. The term Big Data is quite vague and ill defined. The word “Big” is too generic and the question is how “Big” is considered as “Big” and how “Small” is small (Smith, 2013) is relative to time, space and circumstance. The size of “Big Data” is always evolving and the meaning of Big Data Volume would lie between Terabyte (TB) and Zettabyte (ZB) range. The concept of big data is the explosion of data from the Internet, cloud, data center, mobile, Internet of things, sensors and domains that possess and process huge datasets. Cisco claimed that humans have entered the ZB era in 2015 (Cisco, 2017).

Based on social media statistics 2018, the face book claimed that, there are over 300 million photos uploaded to Facebook every day (Nowak & Spiller, 2017). On an average 300 hours of videos are uploaded every minute on You Tube (YouTube, 2017). Approximately, 42 billion texts are sent and 1.6

DOI: 10.4018/978-1-6684-3662-2.ch010

billion photos shared through Whatsapp daily (Stout, 2018). Since 2005, business investment in hardware, software, talent, and services has increased as much as 50 percent, to $4 trillion (Rijmenam, 2018).

In 2005, Roger Mougalas from O'Reilly Media coined the term Big Data for the first time. It refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools. During the same year, Yahoo created Hadoop. This was built on top of Google's MapReduce. Its goal was to index the entire World Wide Web (Rijmenam, 2018).

In 2009, the Indian government decides to take an iris scan, fingerprint and photograph of all of its 1.2 billion inhabitants. All this data is stored in the largest biometric database in the world (Chandra, 2018).

In 2010, at Technomy conference, Eric Schmidt stated, "There were 5 Exabyte's of information created by the entire world between the dawn of civilization and 2003. Now that same amount is created every two days." (Schmidt, 2010).

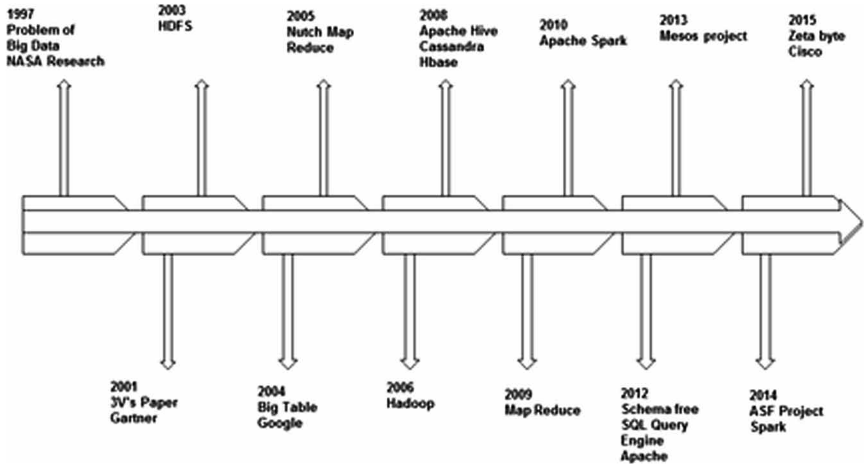
In 2011, McKinsey released a report on Big Data which claimed that, the next frontier for innovation, competition, and productivity, states that in 2018 the USA alone will face a shortage of 140.000 – 190.000 data scientist as well as 1.5 million data managers (Manyika, 2011).

Another detailed review was contributed by Visualizing.org (Hewlett Packard Enterprise, 2017) in Big Data. It is focused on the time line of how to implement Big Data Analytics. Its historical description is mainly determined by events related to the Big Data push by many internet and IT companies such as Google, YouTube, Yahoo, Facebook, Twitter and Apple. It emphasized the significant impact of Hadoop in the history of Big Data Analytics.

In the past few years, there has been a massive increase in Big Data startups, trying to deal with Big Data and helping organizations to understand Big Data and more and more companies are slowly adopting and moving towards Big Data.

Figure 1 shows the history of Big Data and its eco system.

Figure 1. History of Big Data



The data sources and their formats are continuous to grow in variety and complexity. Few list of sources includes the public web, social media, mobile applications, federal, state and local records and databases, commercial databases that aggregate individual data from a spectrum of commercial transactions and public records, geospatial data, surveys and traditional offline documents scanned by optical character recognition into electronic form. The advent of the more Internet enabled devices and sensors expands the capacity to collect data from physical entities, including sensors and radio-frequency identification (RFID) chips. Personal location data can come from GPS chips, cell-tower triangulation of mobile devices, mapping of wireless networks, and in-person payments (Manyika, 2011). The big challenge is, how do we consume those data sources and transform them into actionable information Big Data describes a data management strategy that integrates many new types of data and data management alongside traditional data.

There exist many sources, which predict exponential data growth toward 2020 and beyond. Human- and machine-generated data is experiencing an overall 10x faster growth rate than traditional business data, and machine data is increasing even more rapidly at 50x the growth rate.

IDC predicts that by 2020, 50% of all business analytics software will incorporate prescriptive analytics built on cognitive computing technology, and the amount of high-value data will double, making 60% of information delivered to decision makers actionable (Hewlett Packard Enterprise, 2017). 75% of Big Data is helping government departments to improve the quality of citizen's life style (Mullich, 2013; Wedutenko & Keeing, 2014).

The objective of this chapter is to describe the aspects of big data, definition, drivers, principles, scenarios, best practices and architectures.

BACKGROUND

Today, the data that we deal with is diverse. Users create content like blog posts, tweets, social network interactions, etc. To tackle the challenges of managing this data, a new breed of technologies has emerged.

Data architecture earlier designed primarily for batch processing of mostly structured data and created to address specific Business Units, Enterprise needs. It lacks in the ability to support data democratization, ad-hoc analytics, machine learning/artificial intelligence (ML/AI), complex data governance and security needs - all of which are critical to building a true data driven enterprise.

These new technologies are more complex than traditional databases. These systems can scale to vastly larger sets of data, but using these technologies effectively requires a fundamentally new set of techniques.

Enterprises that do not actively transform themselves to become data driven are left behind, their basic existence questioned. CXOs recognize the threat and the opportunity, and are eager to deploy a modern data architecture that can not only help them store a wide variety of large amounts volume of data but also provides an enterprise-wide analytic platform. This can empower every single employee in the organization to take data driven decisions in real-time, with little or no support from IT.

Big data addresses large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future (National Science Foundation, 2012). It describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data. The boundaries of what constitutes a Big Data problem are also changing due to the ever shifting and advancing landscape of software and hardware technology.

Thirty years ago, one gigabyte of data could amount to a Big Data problem and require special purpose computing resources. Now, gigabytes of data are commonplace and can be easily transmitted, processed and stored on consumer-oriented devices. Data within Big Data environments generally accumulates from enterprise applications, sensors and external sources. Enterprise applications consume this processed data directly or can fed into a data warehouse to enrich existing data there.

The results obtained through the processing of Big Data can lead to a wide range of insights and benefits, such as,

- Operational optimization
- Actionable intelligence
- Identification of new markets
- Accurate predictions
- Fault and fraud detection
- Improved decision-making

Concepts and Terminology

The following are the fundamental concepts and terms used in Big Data Architectures and Techniques.

Datasets: Collections or groups of related data. Each group or dataset member shares the same set of attributes or properties as others in the same dataset. Some examples of datasets are:

- Tweets stored in a flat file
- A collection of image files in a directory
- An extract of rows from a database table stored in a CSV formatted file
- Historical weather observations that are stored as XML files

Data Analysis: Process of analyzing data to find facts, relationships, patterns, insights and trends. The overall goal of data analysis is to support better decision making.

Big Data: A massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques

Data Analytics: A discipline that includes the management of the complete data lifecycle, which covers collecting, cleansing, organizing, storing, analyzing and governing data. Data analytics enable data-driven decision-making with scientific backing so that the decisions based on factual data and not just on experience.

Highlighted below are the four categories of analytics,

1. **Descriptive Analytics**: It addresses to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.
2. **Diagnostic Analytics**: It aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event. The goal is to determine the information related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.
3. **Predictive Analytics**: It helps to determine the outcome of an event that might occur in the future. Enriched information with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations form the basis of models to generate future predictions based on past events.

4. **Prescriptive Analytics**: Built upon the results of predictive analytics by prescribing actions be taken. This type of analytics used to gain an advantage or mitigate a risk.

Types of Data: The data processed by Big Data solutions can be human-generated or machine-generated. Human-generated data is the result of human interaction with systems, such as online services and digital devices. Software programs and hardware devices in response to real-world events generate machine-generated data. The primary types of data are:

- Structured data
- Unstructured data
- Semi-structured data

Structured Data: Conforms to a data model or schema and is often stored in tabular form. Used to capture relationship between different entities. The data is often stored in a relational database. Generally, generation of structured data is by enterprise applications and information systems like ERP and CRM systems.

Unstructured Data: Data that does not conform to a data model or data schema. Unstructured data has a faster growth rate than structured data. This form of data is either textual or binary and often conveyed via files that are self-contained and non-relational. A text file may contain the contents of various tweets or blog postings. Binary files are often media files that contain image, audio or video data.

Semi Structured Data: This type of data has a defined level of structure and consistency, but is not relational in nature. Instead, semi-structured data is hierarchical or graph-based. This kind of data is commonly stored in files that contain text. XML and JSON files are common forms of semi-structured data.

Metadata: It provides information about a dataset's characteristics and structure. This type of data is mostly machine-generated data. The tracking of metadata is crucial to Big Data processing, storage and analysis because it provides information about the pedigree of the data and its provenance during processing. Examples of metadata include:

- XML tags providing the author and creation date of a document
- Attributes providing the file size and resolution of a digital photograph governments

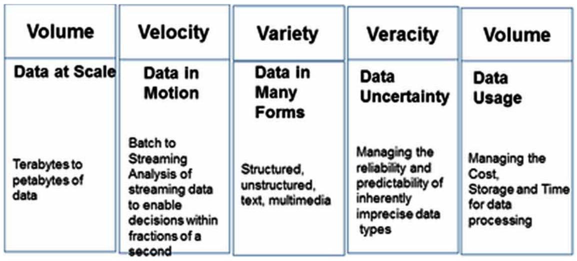
CHARACTERISTICS OF BIG DATA

Big Data Analytics is an integrated Business Intelligence and Data Analytics, which includes conventional and Big Data. Comparing to the traditional data, big data has five major characteristics: Volume, Velocity, Variety, Veracity and Value (Normandeau, 2013; Hilbert, 2016; Hilbert, 2015).

- **Volume:** It indicates more data. The volume of data that processed by Big Data solutions is substantial and ever growing. High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation and management processes. Big Data requires processing high volumes of data, that is, data of unknown value. For example, twitter data feeds, clicks on a web page, network traffic, sensor-enabled equipment capturing data and many more. Typical data sources that are responsible for generating high data volumes are:

- Online transactions
- Scientific and research experiments
- Sensors, such as GPS sensors, RFIDs, smart meters and telematics
- Social media, such as Facebook and Twitter
- **Velocity**: In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods. From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data processing, once it enters the enterprise's perimeter. Coping with the fast inflow of data requires the enterprise to design highly elastic and available data processing solutions and corresponding data storage capabilities. Some Internet of Things (IoT) applications have health and safety ramifications that require real-time evaluation and action. Other internet-enabled smart products operate in real-time or near real-time. As an example, consumer e-commerce applications seek to combine mobile device location and personal preferences to make time sensitive offers.
- **Variety**: New unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video require additional processing to both derive meaning and the supporting metadata. Unstructured data has many of the same requirements as structured data, such as summarization, lineage, auditability, and privacy. Further complexity arises when data from a known source changes without notice.
- **Veracity**: Quality or fidelity of data. Assessment of Data that enters Big Data environments for quality, which can lead to data processing activities to resolve invalid data and remove noise. Data with a high signal-to-noise ratio has more veracity than data with a lower ratio. Data that is acquired in a controlled manner, for example via online customer registrations, usually contains less noise than data acquired via uncontrolled sources, such as blog postings
- **Value**: There is a range of quantitative and investigative techniques to derive value from data. The cost of data storage and compute has exponentially decreased, thus providing an abundance of data from which statistical sampling and other techniques become relevant, and meaningful.

Figure 2. Big Data Characteristics



Volume	Velocity	Variety	Veracity	Volume
Data at Scale	Data in Motion	Data in Many Forms	Data Uncertainty	Data Usage
Terabytes to petabytes of data	Batch to Streaming Analysis of streaming data to enable decisions within fractions of a second	Structured, unstructured, text, multimedia	Managing the reliability and predictability of inherently imprecise data types	Managing the Cost, Storage and Time for data processing

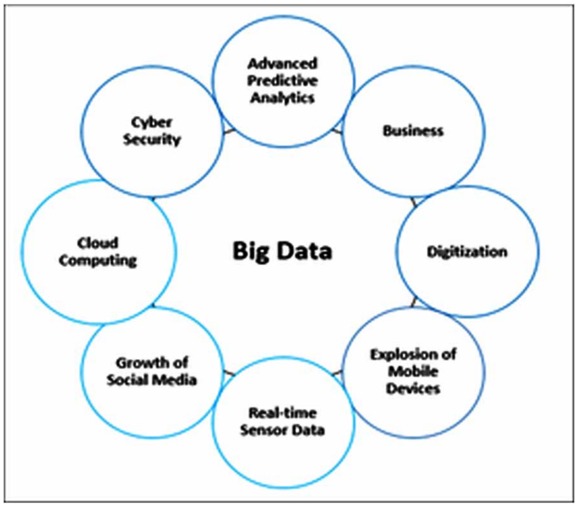
DRIVERS OF BIG DATA

Easy and timely retrieval and analysis of related and unrelated information is crucial for enterprise to meet and improve mission requirements that vary across business units. Enterprises are collecting, procuring, storing and processing increasing quantities of data. This is occurring in an effort to find new insights that can drive more efficient and effective operations, provide management the ability to steer the business proactively and allow the C-suite to better formulate and assess their strategic initiatives. Ultimately, enterprises are looking for new ways to gain a competitive edge. Thus, the need for techniques and technologies that can extract meaningful information and insights has increased. Computational approaches, statistical techniques and data warehousing have advanced to the point where they have merged, each bringing their specific techniques and tools that allow the performance of Big Data analysis. The maturity of these fields of practice inspired and enabled much of the core functionality expected from contemporary Big Data solutions, environments and platforms.

Data continues to be generated and digitally archived at increasing rates driven by customer initiatives, sensors, customer interactions and program transactions. For example, Government organizations are beginning to deploy Big Data technologies to analyze massive data sets as well as mine data to prevent bad actors from committing acts of terror and/or to prevent waste, fraud, and abuse (Kalil, 2012; Department of Defense, 2012).

Figure 3 depicts the drivers of Big Data. Big Data drivers are explained below.

Figure 3. Drivers of Big Data and Analytics



Business

Enterprises today are looking for improvement in marketing, enhance customer experience, improve operational efficiencies, identify fraud and waste, prevent compliance failures and achieve other outcomes that directly affect top- and bottom-line business performance. Big Data analytics helps in discovering new business initiatives. This is the opportunity to enable innovative new business models.

Digitization

Today for all businesses, digital mediums have replaced physical mediums as the de facto communications and delivery mechanism. The use of digital artifacts saves both time and cost as distribution is supported by the vast pre-existing infrastructure of the Internet. As consumers connect to a business through their interaction with these digital substitutes, it leads to an opportunity to collect detailed data by leveraging Big Data Analytics. Collecting detailed data can be important for businesses because mining this data allows customized marketing, automated recommendations and the development of optimized product features.

Explosion of Mobile Devices

The increased use of smart phones Users expect to be able to access their information anywhere and anytime. To the extent that visualizations, analytics, or operationalized big data/analytics are part of the mobile experience.

Real-Time Sensor Data

The coverage of Internet and Wi-Fi networks has enabled more people and their devices to be continuously active in virtual communities. Usage of Internet based connected sensors, Internet of Things and Smart Internet connected devices has resulted in massive increase in the number of available data streams demanding the need for Big Data Analytics. These data streams are public and channeled directly to corporations for analysis.

Growth of Social Media

Customers today are providing feedback on product/item to enterprise, in near real time through various channels. This leads the businesses to consider customer feedback on their service and product offerings in their strategic planning. As a result, businesses are storing increasing amounts of data on customer interactions within their customer relationship management systems (CRM) and from harvesting customer reviews, complaints and praise from social media sites. This information feeds Big Data analysis algorithms that surface the voice of the customer in an attempt to provide better levels of service, increase sales, enable targeted marketing and even create new products and services. Businesses have realized that branding activity no longer managed by internal marketing activities. In addition, enterprises and its customers are co-creating the product brands and corporate reputation. For this reason, businesses are increasingly interested in incorporating publicly available datasets from social media and other external data sources.

Cloud Computing

Cloud computing plays an essential role in data analytics. In many scenarios, it act as a data source, providing real-time streams, analytical services, and as a device transaction hub. Businesses have the opportunity to leverage highly scalable, on-demand IT resources for storage and processing capabilities provided by cloud environments in order to build-out scalable Big Data solutions that can carry out large-scale processing tasks. The ability of a cloud to dynamically scale based upon load allows for the creation of resilient analytic environments that maximize efficient utilization of ICT resources. Cloud computing can provide three essential ingredients required for a Big Data solution: external datasets, scalable processing capabilities and vast amounts of storage.

Cyber Security

Big Data security strategy should be align with the enterprise practices and policies already established, avoid duplicate implementations, and manage centrally across the environments.

Enterprise security management seeks to centralize access, authorize resources, and govern through comprehensive audit practices. Adding a diversity of Big Data technologies, data sources, and uses adds requirements to these practices.

Advanced Analytical Capability

Technological advancement in data collection, storage, analytics and visualization allows the enterprises to increase the amount of data they generate and produce actionable intelligence to support real time decision making. It helps the capability to foresee key events and take appropriate and timely actions. Better utilization of data, not merely for producing statistical reports on the past but intelligent reports that throw light on the future.

PRINCIPLES OF BIG DATA

Architecture principles provide a basis for decision making when developing Big Data solutions and design. These principles will be extended with Organization specific architecture principles and requirements (Blockow, 2018; Forrest, 2016). They form a structured set of ideas that collectively define and guide development of a solution architecture, from values through to design and implementation, harmonizing decision making across an organization.

The following are the principles to guide enterprises in their approach to big data.

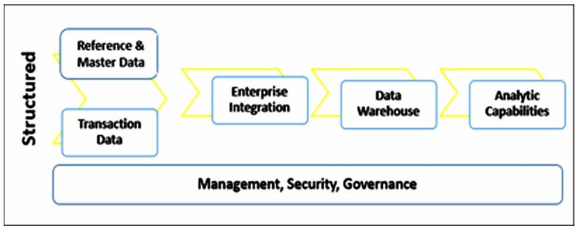
- **Data is an Asset**: Data is an asset that has a specific and measurable value to the Enterprise and managed
- **Data is Shared**: Users have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organizations
- **Data Trustee**: Each data element has a trustee accountable for data quality.
- **Common Vocabulary and Data Definitions**: Data definition is consistent throughout Enterprise, and the definitions are understandable and available to all users.
- **Data Security**: Data is protected from unauthorized use and disclosure.
- **Data Privacy**: that privacy and data protection is considered throughout the entire life cycle of a big data project. All data sharing will conform to relevant regulatory and business requirements
- **Data Integrity and the Transparency of Processes**: Each party to a big data analytics project must be aware of, and abide by their responsibilities regarding: the provision of source data and the obligation to establish and maintain adequate controls over the use of personal or other sensitive data
- **Data Skills and Capabilities**: Skills and expertise in data analytics were shared amongst enterprises and industry, where appropriate. Resources such as data sets and the analytical models used to interrogate them, as well as the infrastructure necessary to perform these computations and shared amongst business units where appropriate and possible to do so.

- **Collaboration with Industry and Academia**: The industry, research and academic sectors have been working on big data analytics projects for some time and continue to invest heavily in the skills, technologies and techniques involved with big data analysis.

BIG DATA ARCHITECTURE FRAMEWORK

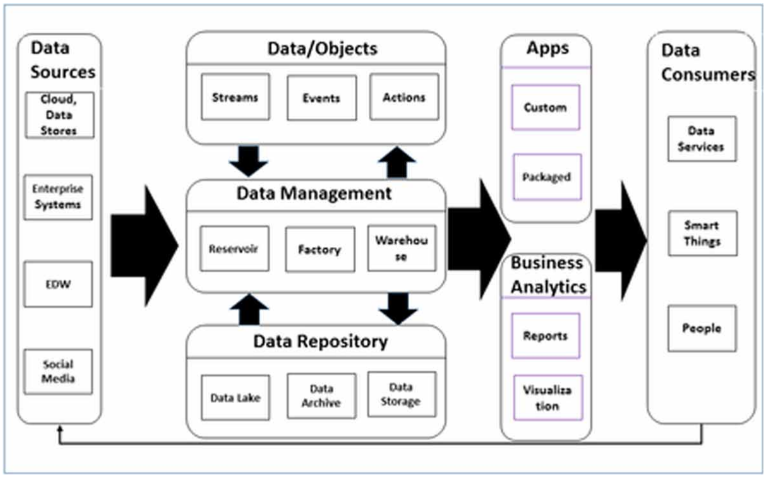
The diagram below depicts the high-level architecture framework of traditional data, structured in nature. It has two data sources that use integration (ELT/ETL/Change Data Capture) techniques to transfer data into a DBMS data warehouse or operational data store, and then offer a wide variety of analytical capabilities to reveal the data. Some of these analytic capabilities include dashboards, reporting, EPM/BI applications, summary and statistical query, semantic interpretations for textual data, and visualization tools for high-density data.

Figure 4. Traditional Architecture Components



A big data architecture, designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems. Conceptual Architecture of the Big Data Analytics shown in Figure below. It illustrates key components and flows and highlights the emergence of the Data repository and various forms of new and traditional data collection.

Figure 5. Conceptual Reference Architecture



Big data solutions typically involve one or more of the following types of workload:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

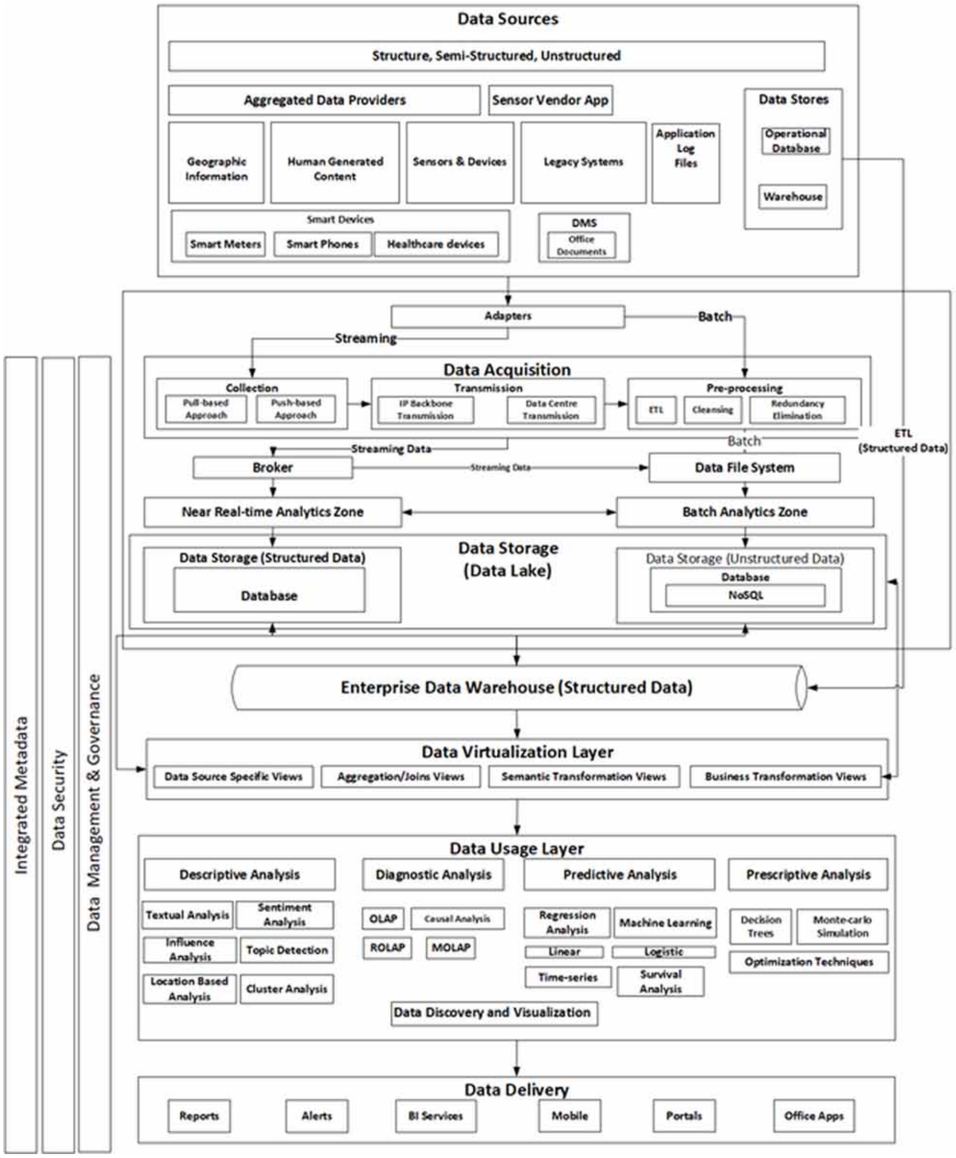
Description of these primary components:

- **Stream Data:** Components that process data in-flight (streams) to identify actionable events and then determine next best action based on decision context and event profile data and persist in a durable storage system. The decision context relies on data in the data reservoir or other enterprise information stores.
- **Reservoir:** Economical, scale-out storage and parallel processing for data that does not have stringent requirements for data formalization or modelling. Typically manifested as a Hadoop cluster or staging area in a relational database.
- **Factory:** Management and orchestration of data into and between the Data Reservoir and Enterprise Information Store as well as the rapid provisioning of data.
- **Warehouse**: Large scale formalized and modelled business critical data store, typically manifested by a Data Warehouse or Data Marts.
- **Data Repository:** A set of data stores, processing engines, and analysis tools separate from the data management activities to facilitate the discovery of new knowledge. Key requirements include rapid data provisioning and sub setting, data security/governance, and rapid statistical processing for large data sets.
- **Business Analytics:** A range of end user and analytic tools for business Intelligence, faceted navigation, and data mining analytic tools including dashboards, reports, and mobile access for timely and accurate reporting.
- **Applications:** A collection of prebuilt adapters and application programming interfaces that enable all data sources and processing directly integrated into custom or packaged business applications.

BIG DATA LOGICAL ARCHITECTURE

The following diagram shows logical application architecture of Big Data Analytics System with key components and layers. A detailed description of these components and layers are provided in this section. While there exist many standard logical architectures for the Big Data Analytics (Wu, 2014; Angelov, 2012; Chen, 2014; Ahmed & Karypis, 2012; Klein, 2017), the author tried to arrive a detailed and concise Big Data Logical Reference Architecture based on practical experience across various domains and technologies.

Figure 6. Logical Application Architecture View



Below is the brief description of each of the logical application architecture layers,

Data Sources and Types

The data sources provide the insight required to solve the business problem. The data sources are structured, semi-structured, and unstructured, and it comes from many sources. Big Data Analytics solution shall support processing of all types of data from a variety of sources. Given below is an indicative list of data sources, and categories. All big data solutions start with one or more data sources. Examples include,

- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT device.

Table 1. Structured and Unstructured data for Internal and External systems

Category	**Internal**
Structured	Departmental Database, Data Hubs, Data Warehouse, Data Marts
Semi/Unstructured	e-Mails, Documents, XML documents
Category	**External**
Structured	Sensor Data, Log Stream Data, Web sites, Satellite Data, Social media, Bioinformatics, Blogs/Articles Documents, E-mails, Audio-visuals, Stream data and Web Analytics data
Semi/Unstructured	

Data Acquisition and Enrich

Relevant push or pull-based mechanisms are used for collecting data from various data sources. Data acquisition provide the capability to hold and transmit raw data collected from various sources to data. The acquisition layer provides the mechanism to cleanse different types of data like traditional, sensor based, log data and data from internet.

- **Streaming Data**: Streaming data comprises of unstructured data coming in from various sources. The data shall be held in a buffer area and when a set limit is reached, it shall be transmitted to Data Analytics system (Hold-Transmit). After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is written to an output sink.
- **Batch Data**: Batch data is normally extracted from enterprise systems using ETL or ELT processes. Structured data may be loaded directly to Data Warehouse, and unstructured/semi structured data to Hadoop or equivalent(or better) unstructured data processing platform. Both ETL and ELT support complex transformations such as cleansing, reformatting, aggregating, and converting large volumes of data from many sources. Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis.
- **Near Real-Time Data Analytics Zone**: It process incoming stream data in real time to provide quick insights into the data. This data may then be persisted on Hadoop system. Near real-time analytics shall provide capabilities like log stream analysis, sensor data analysis etc. The real-time analytics system must be able to quickly identify useful data and that is not useful. Near real-time data shall augment insights obtained from batch analysis.
- **Batch Data Analytics Zone**: Batch data zone ingest large amount of data in batch mode, and also insights obtained in Real-time analytical zone.

Data Storage

Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This type of data store is called Data Lake.

- **Data Lake**: A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure and requirements are not defined until the data is needed. It stores all data while making it faster to get up and running with batch, streaming and interactive analytics. The lake can serve as a staging area for the data warehouse, the location of more carefully treated data for reporting and anlaysis in batch mode. The data lake accepts input from various sources and can preserve both the original data fidelity and the lineage of data transformations.

The features of the Data lake are:

- **Collect Everything:** A Data Lake contains all data, both raw sources over extended periods of time as well as any processed data.
- **Dive in Anywhere:** A Data Lake enables users across multiple business units to refine, explore and enrich data on their terms.
- **Flexible Access:** A Data Lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

Enterprise Data Warehouse

A Data warehouse stores whole of enterprise data, comprising of structured data from enterprise database and data hubs. The data warehouse supports massively Parallel processing and share-nothing architecture and provide optimal performance considering structured and unstructured data. It designed in such a way; it has no single point of failure.

Data Virtualization Layer

Data Virtualization acts as the intermediary and abstraction layer between the information consumers and all sources of data that may contribute to the interaction. It hides cryptic names of tables and columns from users and provide business friendly definitions of data that be used to create reports even by non-technical people. In addition, the data abstraction layer has the capability to access structured, unstructured, or both data in a single query. The query language is standard RDBMS, and query initiated at any level should have ability to process data from all data stores (structured and unstructured). The layer supports a strong optimizer to tune query execution, for response time as well as throughput.

Data Consumers and Delivery

It describes how enterprise users and applications consume output from Big Data Analytics system. This may be in the form of Big Data Analytics Services, alerts on emails and phones, actions, integration with office applications like word, excel etc., collaboration(discussion threads etc.), mobile and so on. The Delivery layer supports delivery through the following mechanisms:

1. **Big Data Analytics Services**: It offers ability to embed actions, alerts, and reports in other application, tool or UI. They shall have ability to refresh automatically based on predefined schedule.
2. **Alerts**: This is to notify stakeholders if a certain event has occurred. Alerts may be delivered in the form of email, reports, or messages.
3. **Actions**: Enable users take some action based on alerts or reports. For example: removing a duplicate record or fixing a corrupted data.
4. **Portal**: Portals provide mechanism to catalogue and index, classify, and search for Big Data Analytics objects such as reports or dashboards. All Big Data Analytics reports to be made available to department users on the portals, based on the roles and responsibilities.
5. **Mobile:** Reports, dashboards, and portals shall be accessible on Mobile devices too.
6. **Office Applications**: The system should integrate with Standard Office products at the minimum. The data and reports should be importable and exportable from/to Office products.

Integrated Meta Data Management

Metadata repository needs to be created for both Structured and Unstructured data. Whether it is for structured data or unstructured, metadata contain enough information to understand, track, explore, clean, and Transform data. Big Data Analytics has the capability to apply metadata on incoming data without any manual intervention.

- **Metadata for Structured Data (DWH)**: It includes Technical, Business, and Process metadata. Besides these, rules of precedence such as which source tables can update which data elements in which order of precedence must be defined and stored.
- **Metadata for Unstructured Data**: Contains rules, definitions, and datasets that help filter out valuable data from incoming data streams or batch load, and persist only such data that are useful. Metadata should enable lineage tracking of data that is loaded into Big Data Analytics system
- **Reusing Data Objects**: Standard queries, models, and metadata can be moved into one layer and virtualise it so that these objects may be reused

Data Security

Data security considerations specific to big data include:

- Increased value of the information asset as enterprises enrich their data, its aggregation and the insights derived from it.
- The increasing range of data acquisition channels and their potential vulnerability.
- The unknowability of the content of unstructured data sources upon acquisition.
- Increased distribution of physical and virtual locations of data storage.

More details about the Big Data Security is explained in the Secuirty Architecture Section of this chapter.

Data Usage Layer

Different users may want different types of outputs based on their role, responsibilities, and functions. The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyze the data, the architecture may include a data modeling layer, such as a multidimensional OLAP cube or tabular data model in Azure Analysis Services. It might also support self-service BI, using the modeling and visualization technologies in Microsoft Power BI or Microsoft Excel. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts. Big Data Analytics shall provide the following usage capabilities:

- **Reports and Ad-hoc Queries**: Analytical reporting (based on data warehouse/Datamart). The system shall provide scripting language, ability to handle complex headers, footers, nested subtotals, and multiple report bands on a single page.
- The system shall support simple, medium, and complex queries against both structured and unstructured data.
- **Online Analytical Processing (OLAP)**: Slicing and dicing, measuring dependent variables against multiple independent variables. It enables users regroup, re-aggregate, and re-sort by dimensions.
- **Advanced Analytics**: This includes predictive, prescriptive, descriptive, causal, statistical, spatial, and mathematical analysis, using structured and unstructured data
- **Dashboards**: Displays variety of information in one page/screen. Typically they display Key Performance Indicators visually.
- **Textual Analytics**: Textual analytics refers to the process of deriving high-quality information from text in documents, emails, Government orders, web, etc. This is useful in sentiment analysis, understand hot topics of discussion in public, and maintaining government image.
- **Performance Management**: Analytical data can be used by departments to understand their performance, and reasons for current levels of performance measured in terms of KPIs.
- **Data Mining, Discovery, and Visualisation**: It is about searching for patterns and values within data streams such as sensor based data, social media, satellite images etc. Data exploration is primarily used by Data scientists or statisticians to create new Analytical models and test them so that they can be used for Analytics.

Data Management and Governance

Data Governance is a process of managing data assets of an enterprise. It includes the rules, policies, procedures, roles and responsibilities that guide overall management of an enterprise's data. It provides the guidance to ensure that data is accurate and consistent, complete, available and secure. Governance is not a onetime event - it is a continual process of maintaining, monitoring and improving the enterprise important asset.

Enterprises have the governance responsibility to align disparate data types and certify data quality. Governance provides the structure to enable:

- The decision making processes that an enterprise uses to ensure the integrity of its key data items.
- Fast and effective decision making in times of ambiguity.

- Adherence to policies, standards and alignment to overall data management approach.

Improved Governance

- Data-analysis-based insights improving quality of governance.
- Data-analysis-driven decisions leading to right planning and right targeting.
- Insights leading to effective regulation and better governance.
- Recommendations and interventions to improve performances.

Principles of Data Governance are:

- **Enterprise Asset**: Data is recognized as a key business asset and will be organized, stored, distributed and managed to allow sharing across the Enterprise.
- **Conformance**: The logical structure of data will be independent of Applications and will conform to defined Logical Data Models and common data formats.
- **Stewards and Owership**: Each data element has a corporate steward accountable for data ownership and data quality.
- **Shared**: Users have access to the data necessary to perform their duties; therefore, data is shared across Enterprise functions and Business Units.
- **Accessible**: Data is accessible for users to perform their functions from any location and by any approved mechanisms.
- **Secure**: Data is protected from unauthorized use and disclosure.
- **Timely:** The Systems, Applications and Databases will be designed to make data available anytime and anyplace.
- **Available:** All shared data will have the capability to be continuously available based on agreed business need.
- **Interoperable Information**: Data must be managed in such a way as to achieve information interoperability.
- **Common Vacabulary**: Data is defined consistently throughout the Enterprise, and the definitions are understandable and available to all users.
- **Meta Data Repository**: An integrated, centralized Metadata Repository.
- **Data Capture**: All primary data will be captured once at the point of creation and stored and managed to enable appropriate levels of sharing across the Enterprise.

BIG DATA SECURITY ARCHITECTURE

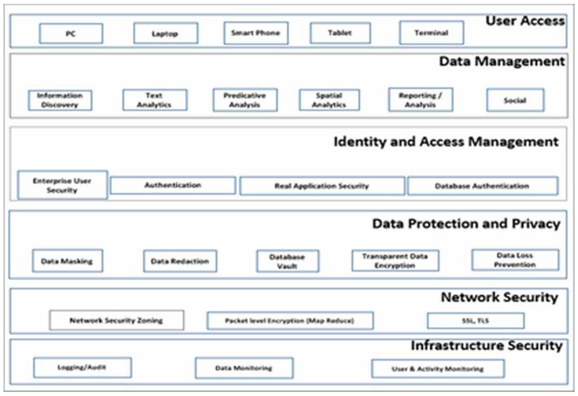
Security today involves far more than just password protection, anti-malware solutions, and network encryption. It requires a continuous application of security measures to manage and control access to valuable electronic assets of an enterprise. Big Data security approach shall ensure that the right people, internal or external, get access to the appropriate data and information at right time and place, within the right channel (National Security Agency, Central Security Services, 2011; Smith & Hallman, 2013). The security prevents and safeguards against malicious attacks and protects enterprise data assets by securing and encrypting data while it is in-motion or at-rest. It also enables organizations to separate roles and responsibilities and protect sensitive data without compromising privileged user access.

Based on the author experience in Big Data Architectures, the core components of the Big Data Security Framework are classified as:

- **Data Management**: Secure data storage and transaction logs
- **Identity and Access Management:** Role based access control for data components
- **Data Protection and Privacy:** Scalable privacy preserving data mining and analytics
- **Network Security:** Network access control allows traffic to approved levels
- **Infrastructure Security and Integrity:** Secure computations in distributed programming

Figure 7 is the logical architecture for the big data security approach:

Figure 7. Big Data Security Architecture View



The various data security capabilities are:

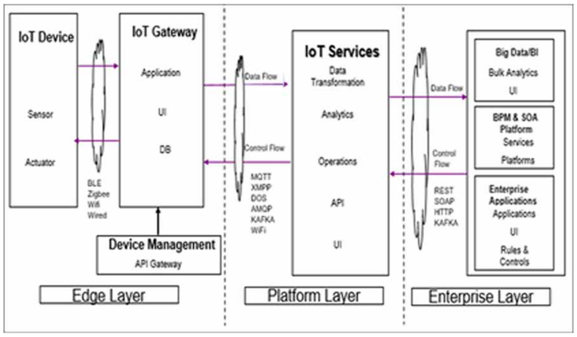
- Authentication and authorization of users, applications and databases
- Privileged user access and administration
- Data encryption and redaction, application level cryptographic protection
- Data masking and sub setting, performed in batch or real time
- Separation of roles and responsibilities, role based access control
- Transport security
- Network security, data protection in transit and network zoning and authorization components
- Database activity monitoring, alerting, blocking, auditing and compliance reporting

IOT ANALYTICS ARCHITECTURE

Imagine a world where billions of objects can sense, communicate and share information and are inter connected over public or private Internet Protocol (IP) networks. These interconnected objects have data regularly collected, analyzed and used to initiate action, providing a wealth of intelligence for planning, management and decision-making. This is called Internet of Things (IOT) (Morris, 2014; Wang, 2017; Greenough, 2014; Rohling, 2014).

These devices are producing Zetabytes of data every month. The transmissions typically consist of high velocity semi-structured data streams that must land in highly scalable data management systems. The following diagram depicts the IoT logical reference architecture as three tiers: 1.) Edge; 2.) Platform, and; 3.) Enterprise. These tiers process the data flows and control flows based on usage activities across the enterprise systems.

Figure 8. IoT Analytics Architecture View



1. **Edge**: Consists of IoT devices and the IoT gateway. The architectural characteristics of this tier, including its breadth of distribution and location, depend on the specific use cases of the enterprise.

It is common for IoT devices to communicate using a relatively short range and specialized proximity network, due to power and processing limitations.

The IoT gateway contains a data store for IoT device data, one or more services to analyze data streaming from the IoT devices or from the data store, and control applications.

The IoT gateway provides endpoints for device connectivity, facilitating bidirectional communication with the enterprise systems. It also implements edge intelligence with different levels of processing capabilities.

2. **Platform**: Receives, processes and forwards control commands from the Enterprise tier to the Edge tier. The Platform tier consolidates, processes and analyses data flows from the Edge tier, and provides management functions for devices and assets. It also offers non-domain-specific services such as data operations and analytics.
3. **Enterprise**: Receives data flows from the Edge and Platform tiers, and issues control commands to these tiers. The Enterprise tier implements enterprise domain-specific applications and decision support systems, and provides interfaces to end users, including operations.

The different networks used to connect these three tiers are:

1. The ***proximity network*** connects the sensors, actuators, devices, control systems and assets, collectively called edge nodes. It typically connects these edge nodes in one or more clusters to a gateway that bridges to other networks.
2. The ***access network*** enables data and control flows between the Edge and the Platform tiers. It may be a corporate network, or a private network overlaid over the public Internet or a 4G/5G network.
3. The ***service network*** enables connectivity between the services in the platform tier and the enterprise tier, and the services within each tier. It may be an overlay private network over the public Internet, or the Internet itself, allowing enterprise-grade security between end-users and various services.

Users of the IoT system include both humans and digital users. Humans typically interact with the IoT system using one or more kinds of user devices – smartphones, personal computers, tablets or specialized devices. In all cases, the IoT system provides some form of application that connects the human user with the rest of the IoT system.

In some scenarios, immediate action must be taken when data is first transmitted (as when a sensor reports a critical problem that could damage equipment or cause injury) or where it would be possible alleviate some other preventable situation (such as relieving a highway traffic jam). Event processing engines designed to take certain pre-programmed actions quickly by analyzing the data streams while data is still in motion or when data has landed in NoSQL database front-ends or Hadoop. The rules applied usually based on analysis of previous similar data streams and known outcomes.

BIG DATA ANALYSIS TECHNIQUES

Big Data analysis blends traditional statistical data analysis approaches with computational ones. In any fast moving field like Big Data, there are always opportunities for innovation. An example of this is the question of how to blend statistical and computational approaches for a given analytical problem. Statistical techniques are commonly preferred for exploratory data analysis, after which computational techniques that advantage the insight gleaned from the statistical study of a dataset can be apply (Buhler, 2016; We, 2014; Oracle Corporation, 2015; Lopes & Ribeiro, 2015; NIST, 2015; Labrinidis & Jagadish, 2012).

The shift from batch to real-time presents other challenges, as real-time techniques need to leverage computationally efficient algorithms. The following are the basic types of data analysis:

- Quantitative analysis
- Qualitative analysis
- Data mining
- Statistical analysis
- Machine learning
- Semantic analysis
- Visual analysis
- **Quantitative Analysis**: Quantitative analysis is a data analysis technique that focuses on quantifying the patterns and correlations found in the data. Based on statistical practices, this technique involves analysing a large number of observations from a dataset. Since the sample size is large,

the results can be applied in a generalized manner to the entire dataset. Quantitative analysis results are absolute in nature and used for numerical comparisons.

- **Qualitative Analysis**: Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words. It involves analysing a smaller sample in greater depth compared to quantitative data analysis. Extending these results to entire dataset is not possible due to the small sample size. The analysis results state only that the figures were “not as high as,” and do not provide a numerical difference.
- **Data Mining**: Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets. In relation to Big Data analysis, data mining generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends. Specifically, it involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns. Data mining forms the basis for predictive analytics and business intelligence (BI).
- **Statistical Analysis**: Statistical analysis uses statistical methods based on mathematical formulas as a means for analysing data. Statistical analysis is most often quantitative, but can also be qualitative. This type of analysis commonly used to describe datasets via summarization, such as providing the mean, median, or mode of statistics associated with the dataset. In addition, it is used to infer patterns and relationships within the dataset, such as regression and correlation.
- **Machine Learning**: Humans are good at spotting patterns and relationships within data. Unfortunately, we cannot process large amounts of data very quickly. Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how. If human knowledge combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning.
- **Semantic Analysis**: A fragment of text or speech data can carry different meanings in different contexts, whereas a complete sentence may retain its meaning, even if structured in different ways. In order for the machines to extract valuable information, text and speech data needs to understand by the machines in the same way as humans do. Semantic analysis represents practices for extracting meaningful information from textual and speech data.
- **Visual Analysis**: Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception. Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data. The objective is to use graphic representations to develop a deeper understanding of the data being analysed. Specifically, it helps identify and highlight hidden patterns, correlations and anomalies. Visual analysis, directly related to exploratory data analysis as it encourages the formulation of questions from different angles.

INDICATIVE BUSINESS SCENARIOS OF BIG DATA IN GOVERNMENT

The following lists summarizes representative categories where Big Data Analytics system will be used to improve Government and department processes.

Integrated Services

- Analyzing the content in electronic and social media and other sources to understand public sentiment on the programs of the Government, conducting a root-cause analysis and suggesting appropriate interventions and mid-course corrections to improve the delivery of the programs.
- Predicting a disaster and identifying the areas likely to be affected, and suggesting advance interventions required to mitigate the adverse impact on the population.
- Analyzing the Text inputs (unstructured data) in the Grievance system and the popular print media, identifying of key problem areas (Region / Type of Problem / Frequency/Severity) and suggesting suitable remedial action.
- Designing a Happiness Index, appropriate to the socio-economic profile of the Government agency, supporting the Government in conducting approriate sample surveys, Analyzing the results and making suitable recommendations for enhancement of the Index.

Service Delivery

- Analyzing the medium-term impact of development and welfare schemes, identifying the gaps and realigning the schemes for enhanced effectiveness.
- Analyzing the geographical spread of various schemes and making corrections for even distribution.
- Conduct sentiment analysis based on social media and electronic media, and provide appropriate inputs for action by the municipality.
- Qualitative and Quantitative analysis of potable drinking water supplied to the rural people in the habitations as per defined norms through implementation of various water supply schemes under different programs in the government.

Statistics

- Analyzing the patterns of public expenditure on top 10 sectors of the economy, identifying the correlations with the progress in achieving the relevant Sustainable Development Goals and suggesting the desired areas and sectors for intervention.
- Analyzing the trends of growth of GSDP, geographically and sector-wise, identifying causal factors for high and low growth rates and suggesting the right mix of interventions required to optimize the growth rate of the economy of the Government.
- Analysis of trends of cropped areas and economics of various crops area wise over the last 5 years, and the demand-supply position for different agricultural produce across the country and to arrive at the optimised crop area planning for various crops in different agro-climatic regions of the Government and giving decision support to agricultural planners.
- Analysis of soil health records of the last 5 years, along with the crops grown during the period, rainfall, irrigation, yield and other parameters, to arrive at a plan for maximising micro-nutrient corrections, through focused interventions.

Productivity Gain

- To monitor the condition of the roads and provide advance recommendations on optimal resource utilisation for producing best impact on taxpayers.
- Identify leakages of taxes and other major revenues, conduct causal analysis and provide decision support.
- Monitor the sanitary conditions, analyse w.r.t climatic and othe rconditions and predict the outbreak of communicable diseases to enable the department to take corrective action.
- Analysis of global commodity prices and provision of advisories to farmers on the export markets to be preferred for exporting grain and horticultural products.
- Integrating climatic, economic, and social data along with quality of healthcare provided, identify geographic regions that are vulnerable to Viral diseases and providing decision support to the department (realtime).
- Usage of IOT for Smart City to improve the quality of the life of the Citizen.

BIG DATA BEST PRACTICES

The following are the best practices for the Big Data Architectures:

Business

- **Align Big Data with Business Goals**: Advice business of an enterprise on how to apply big data techniques to accomplish their goals. For example, understand e-commerce behaviour, derive sentiment from social media and customer support interactions and understand statistical correlation methods and their relevance for customer, product, manufacturing, or engineering data. Even though Big Data is a newer IT frontier and there is an obvious excitement to master something new, it is important to base new investments in skills, organization, or infrastructure with a strong business-driven context to guarantee ongoing project investments and funding. Determine how Big Data support and enable enterprise business architecture and top IT priorities.
- **Consolidate Enterprise Data**: Today enterprises have an overwhelming amount of data available in the form of structured and unstructured application data (documents, files, logs, click streams, events, social media, images, videos and more). All this data is either poorly captured or not easily accessible by employees due to the siloed nature of old data architectures. Big Data Analytics helps in establishing an enterprise-wide common data platform that makes data available from a central location.
- **Align with the Cloud Operating Model**: Big Data processes and users require access to broad array of resources for both iterative experimentation and running production jobs. Data across the data realms (transactions, master data, reference, and summarized) is part of a Big Data solution. Private and Public cloud provisioning and security strategy plays an integral role in supporting these changing requirements.

- **Manage Operations**: Operationalizing insights requires a repeatable and scalable process for developing numerous analytic models and a reliable architecture for deploying these models into production applications. Ease of operationalization is an important characteristic of a successful modern data architecture.

Technical

- **Unstructured and Structured Data**: It is certainly valuable to analyse Big Data on its own. However, by connecting and integrating low density Big Data with the structured data you are already using today, you can bring even greater business clarity. For example, there is a difference in distinguishing all sentiment from that of only your best customers. Whether you are capturing customer, product, equipment, or environmental Big Data, an appropriate goal is to add more relevant data points to your core master and analytical summaries, which can lead to better conclusions. For these reasons, many see Big Data as an integral extension of enterprise existing business intelligence and data warehousing platform and information architecture.
- **Partition Data**: Partition data files and data structures such as tables, based on temporal periods that match the processing schedule. That simplifies data ingestion and job scheduling, and makes it easier to troubleshoot failures.
- **Schema-on-Read Semantics**: Use *schema-on-read* semantics, which project a schema onto the data when the data is processing, not when the data is stored. This builds flexibility into the solution, and prevents bottlenecks during data ingestion caused by data validation and type checking.
- **Cloud**: Incorporate on premise and cloud Organizations have different criteria for determining which workloads run on premise vs cloud. The criteria could involve internal or external policies regarding location of data stored; availability of an application/system in the cloud; availability of capacity for running a specific workload, etc. It is important for a modern architecture to support a hybrid environment as it is fast becoming the new operating reality for enterprises.
- **Process Data In-Place**: Traditional BI solutions often use an extract, transform, and load (ETL) process to move data into a data warehouse. With larger volumes data, and a greater variety of formats, big data solutions generally use variations of ETL, such as transform, extract, and load (TEL). With this approach, the data processed within the distributed data store, transforming it to the required structure, before moving the transformed data into an analytical data store.
- **Orchestrate Data Ingestion**: In some cases, existing business applications may write data files for batch processing directly into data storage, where Data Lake Analytics consume it. To orchestrate the ingestion of data from on-premises or external data sources into the data lake. Use an orchestration workflow or pipeline, to achieve a predictable and centrally manageable fashion.
- **Automate Processes:** At a time when data volume, variety and number of sources are ever-increasing, automation plays a key role in keeping the data driven culture alive. Data pipeline automation, automation of data cataloging (using ML/AI) and such, help in near real-time availability of consumable data.

Governance

- **Support from Management**: Get C-suite support Building a data driven enterprise needs deep collaboration between various functions across the organization. Many challenges pertaining to people, policies, data ownership and sharing will arise. For the initiative to succeed, a top-down mandate with a clear mission and approval framework to resolve logjams is required.
- **Culture Cultivation**: Develop a data driven culture Building a data driven enterprise is more about people and culture than technology. To ensure employees replace their gut-based decision making with a more thoughtful, data driven approach, it is important to sensitize employees to the need for, and advantages of, being data driven. Orientation and training sessions on how to use data and analytics as part of their daily operations will play a key role in the successful implementation and adoption of a modern data architecture.
- **Data Governance**: Data is a shared asset for any enterprise. Data governance assumes an important role and needs a well thought out enterprise wide strategy coupled with strong execution. It needs a framework that transcends enterprise silos to establish how data assets managed, accessed by employees. Data quality, lineage, security, discovery, self-serve access, compliance, legal hold and information lifecycle management need to be given due importance as part of the data governance strategy. To achieve a comprehensive governance strategy, put together a strategy team representing the legal and compliance departments, IT operations, line of business stakeholders, and application/ information owners. Further, enterprises need to implement a comprehensive communication program to sensitize employees about the need for the governance policy and their adherence to it.
- **Data Community**: Build a data community Data democratization needs trusted guardians who can help data consumers use the right data in its relevant form. Building a data community comprising IT managers, data engineers, data scientists and functional experts is crucial for enabling a trusted data environment for business users. This community is responsible for the availability of centralized data dictionaries, MDM, data enrichment, data preparation, pre-prepared data models, business formulae, algorithms and such to the business users. This greatly helps business users to quickly extract insights without having to worry about the data quality or the trustworthiness of the data available.
- **Skills and Governance**: Organizations implementing Big Data solutions and strategies should assess skills requirement early and often and should proactively identify any potential skills gaps. Skills gaps can be addressed by training / cross-training existing resources, hiring new resources, or leveraging consulting firms.
- **No Big Bang**: Do not try to do everything at once. Deploying a modern data architecture is a big initiative and is heavily influenced by technology, policies and people. Though it is important to approach this in a holistic manner, it is not necessary to do it all at once. Enterprises can realize benefits by implementing modern data architecture even for a single function and use the lessons learned for the next phase.

FUTURE RESEARCH DIRECTIONS

Big data analytics is gaining so much attention these days and there were number of research problems that need to be addressed going forward. Few research directions for the future are highlighted below.

Many different models like fuzzy sets, rough sets, soft sets, neural networks, their generalizations and hybrid models are used in analyzing the data. The challenges in analyzing the data may affect performance, efficiency and scalability of the data intensive computing systems. Fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue. Expressing data access requirements of applications and designing programming language abstractions to exploit parallelism are an immediate need.

- **Data Life Cycle of Big Data Analytics:** Most of the customer requirements today demanding real-time performance of the big data analytics. This leads for the definition of data life cycle, the value it can provide and the computing process to make the analytics process real time. This increases the value of the analysis (Boyd & Crawford, 2012). A proper data filtering techniques need to be developed to ensure correctness of the data (Nielsen & Chuang, 2000) in Big Data Analysis. The availability of data that is complete and reliable is a big challenge. In most of the cases, data is very limited and do not show clear distribution, resulting to misleading conclusions. A method to overcome these problems needs proper attention and sometimes handling of unbalanced data sets leads to biased conclusion.
- **Storage and Retrieval Data:** Multidimensional data should be integrated with analytics over big data. With the explosion of smart phones, the Images, Audios and Videos are being generated at an unremarkable pace. However, storage, retrieval and processing of these unstructured data require immense research in each dimension.
- **Big Data Computations:** Apart from current big data paradigms like Map-Reduce, other paradigms such as YarcData (Big Data Graph Analytics) and High-Performance Computing cluster (HPCC explores Hadoop alternatives), are being explored.
- **Algorithms for Real Time Processing:** The pace at which data is being generated and the expectations from these algorithms may not be met, if the desired time delay is not met.
- **Smart Storage Devices:** The demand for storing digital information is increasing continuously. Purchasing and using available storage devices cannot meet this demand. Research towards developing efficient storage device that can replace the need for HDFS systems that is fault tolerant can improve the data processing activity and replace the need for software management layer.
- **Quantum Computing for Big Data Analytics:** A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously (Hashem, 2015). Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantumly. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale

quantum computers compared with classical computers. Hence it is a challenge for this generation to built a quantum computer and facilitate quantum computing to solve big data problems.

- **Cloud Computing for Big Data Analytics:** Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting results. This can help to solve large applications that may arise in various domains (Chen, 2012). In addition, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques. The major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.
- **IoT for Big Data Analytics:** An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understanding these streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective. Key technologies that are associated with IoT are also discussed in many research papers.
- **Machine Learning for Big Data Analytics:** Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it. Author, argue that while each of the tools has their advantages and limitations, more efficient tools can be developed for dealing with problems inherent to big data. These efficient tools to be developed must have provision to handle noisy and imbalance data, uncertainty and inconsistency, and missing values.

CONCLUSION

Modern businesses are evolving and are constantly demanding more from their Information Management systems. No longer satisfied with standardized reporting by a limited set of users, modern businesses manage by fact, demanding faster and more pervasive access to information on which to base critical business decisions. This change to the volume, velocity and reach of the information is in turn forcing changes to the solution architecture and technology that underpins the solutions.

Big Data employs the tenet of “bringing the analytical capabilities to the data” versus the traditional processes of “bringing the data to the analytical capabilities through staging, extracting, transforming and loading,” thus eliminating the high cost of moving data.

In Big Data world, data storage platforms are not restricted to a predefined rigid data model and data systems are capable of handling all kinds of structured and unstructured data. Big data offers capabilities such as deploying data storage/processing from new sources such as external social media data, market data, communications, interaction with customers via digital channels, etc. with unconstrained scalability and flexibility to adapt to constantly changing data landscape.

The following are the Outcome and recommendations on the usage of Big Data Analytics:

- To provide insights into how current business scenario's are performing, and Why(Descriptive and Causal Analyses).
- Design of Better Projects by being more customer centric and effective.
- To determine likely future scenarios and recommend best courses of action (Predictive and Prescriptive Analyses).
- To gauge sentiments of customers, and understand their perceptions of and attitudes towards enterprise products, policies.
- To provide a system of dashboards that enable administrators monitor and implement enterprise programs effectively.
- To improve collaboration among various stakeholders.
- To provide a tool for research in Data Sciences and statistical analysis.
- Enhanced customer Satisfaction through participation in decision-making.
- Formulation of the Right Policies that factor the needs of the people.
- Enhanced transparency of public institutions through feedback & social audit.
- Increased Trust between enterprise & customer allows the free flow of the information.
- Real-time fraud monitoring can be done by integrating large amounts of diverse, structured and unstructured high-velocity data.
- Real-time location information to provide more accurate traffic and drive-time information by analyzing the commute patterns, drive times to and from work.

Big Data also opens up a range of new design and implementation patterns that can make Information Management solutions less brittle, speed development reduce costs and generally improve business delivery. When designed appropriately, they can combine these benefits without giving up the things the business has come to expect and value such as good governance, data quality and robustness.

Finally, Big Data Analytics is not about adopting a technology solution. It is about leveraging tools that enable enterprise to operate more effectively through making informed decisions and where needed, in real time.

ACKNOWLEDGMENT

The author would like to thank Hari Kishan Burle, Raju Alluri of Architecture Group of Wipro Technologies for giving us the required time and support in many ways in bringing this chapter as part of Global Enterprise Architecture Practice efforts. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Ahmed, R., & Karypis, G. (2012). Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks. *Knowledge and Information Systems*, *33*(3), 603–630. doi:10.100710115-012-0537-2

Angelov, S., Grefen, P., & Greefhorst, D. (2012). A framework for analysis and design of software reference architectures. *Journal of Information and Software Technology*, *54*(4), 417–431. doi:10.1016/j.infsof.2011.11.009

Blockow, D. (2018). *Big Data Architecture Principles*. Data to Decision CRC. Retrieved from https://www.d2dcrc.com.au/blog/big-data-architecture-principles/

Boyd & Crawford. (2012). *Six Provocations for Big Data. Proceeding of A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall.

Chandra, S. (2018). *India's Biometric Identity Program Is Rooting Out Corruption*. Retrieved from: https://slate.com/technology/2018/08/aadhaar-indias-biometric-identity-program-is-working-but-privacy-concerns-remain.html

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. doi:10.100711036-013-0489-0

Chen, X. (2012). Article. *Research on Key Technology and Applications for Internet of Things*, *33*, 561–566.

Cisco. (2017). *The Zettabyte Era: Trends and Analysis*. Retrieved from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

Department of Defence. (2012). *Big Data Across the Federal Government*. Executive Office of the President. Retrieved from: https://www.hsdl.org/?view&did=742609

Enterprise, H. P. (2017). *The Exponential Growth of Data*. Retrieved from: https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/

Forrest, C. (2016). 5 architectural principles for building big data systems on AWS. *TechRepublic*. Retrieved from: https://www.techrepublic.com/article/5-architectural-principles-for-building-big-data-systems-on-aws/

Greenough, J. (2014). *The 'Internet of Things' Will Be The World's Most Massive Device Market And Save Companies Billions Of Dollars. BI Intelligence reports*. Retrieved form: https://www.businessinsider.in/The-Internet-of-Things-Will-Be-The-Worlds-Most-Massive-Device-Market-And-Save-Companies-Billions-Of-Dollars/articleshow/44766662.cms

Hashem, I., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. doi:10.1016/j.is.2014.07.006

Hilbert, M. (2015). What is Big Data. *YouTube*. Retrieved from: https://www.youtube.com/watch?v=XRVIh1h47sA

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34*(1), 135–174. doi:10.1111/dpr.12142

Kalil, T. (2012). *Big Data is a Big Deal*. The White House. Retrieved from: https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal

Klein, J. (2017). *Reference Architectures for Big Data Systems*. Carnegie Mellon University Software Engineering Institute. Retrieved from: https://insights.sei.cmu.edu/sei_blog/2017/05/reference-architectures-for-big-data-systems.html

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceeding of VLDB Endowment*, *5*(12), 2032–2033.

Lopes & Ribeiro. (2015). GPUMLib: An Efficient Open-source GPU Machine Learning Library. *Machine Learning for Adaptive Many-Core Machines - A Practical Approach, 7*, 15–36.

Manyika. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Morris, H. (2014). *A Software Platform for Operational Technology Innovation*. International Data Corporation. Retrieved from: https://www.predix.com/sites/default/files/IDC_OT_Final_whitepaper_249120.pdf

Mullich, J. (2013). *Closing the Big Data Gap in Public Sector*. SAP, Bloomberg Inc.

National Science Foundation. (2012). *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)* (Publication Number: 12-499). Retrieved from: http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

National Security Agency, Central Security Services. (2011). *Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center*. NSA Press. Retrieved from: https://www.nsa.gov/news-features/press-room/press-releases/2011/utah-groundbreaking-ceremony.shtml

Nielsen & Chuang. (2000). Quantum Computation and Quantum Information. Cambridge University Press.

NIST. (2015). *NIST Big Data Interoperability Framework: Use Cases and General Requirements*. NIST Big Data Public Working Group (Publication number 1500-3). Retrieved from: https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-3.pdf

Normandeau. (2013). *Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity*. Big Data Innovation Summit. Retrieved from: https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

Nowak & Spiller. (2017). *Two Billion People Coming Together on Facebook*. Facebook News.

Oracle Corporation. (2015). *An Enterprise Architect's Guide to Big Data*. Oracle Corporation. Retrieved from: https://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

Rijmenam, M. (2018). *A Short History Of Big Data*. Retrieved from: https://datafloq.com/read/big-data-history/239

Rohling, G. (2014). *Facts and Forecasts: Billions of Things, Trillions of Dollars*. Siemens - Internet of Things: Facts and Forecasts. Retrieved from: https://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/internet-of-things-facts-and-forecasts.html

Schmidt, E. (2010). *Techonomy*. Retrieved from: https://www.youtube.com/watch?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq&v=UAcCIsrAq70

Smith, T. P. (2013). *How big is big and how small is small, the size of everything and why*. Oxford University Press.

Smith & Hallman. (2013). NSA Spying Controversy Highlights Embrace Of Big Data. *The Huffington Post*. Retrieved from: https://www.huffingtonpost.in/entry/nsa-big-data_n_3423482

Stout. (2018). *Social Media Statistics 2018: What You Need to Know*. Retrieved from: https://dustn.tv/social-media-statistics/

Wang, J. (2017). Big Data Driven Smart Transportation: the Underlying Story of IoT Transformed Mobility. *The WIOMAX SmartIoT Blog*. Retrieved from: http://www.wiomax.com/big-data-driven-smart-transportation-the-underlying-big-story-of-smart-iot-transformed-mobility/

We, H. (2014). *SAP and Hortonworks Reference Architecture*. SAP AG.

Wedutenko & Keeing. (2014). *Big data and the public sector: strategy and guidance*. Clayton Utz Insights.

Wu, X. (2014). Data Mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. doi:10.1109/TKDE.2013.109

YouTube. (2017). *YouTube by the Numbers*. Retrieved from: https://www.youtube.com/yt/about/press/

KEY TERMS AND DEFINITIONS

Cloud Computing: Cloud computing is an ICT sourcing and delivery model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Data Exhaust: Data exhaust (or digital exhaust) refers to the by-products of human usage of the internet, including structured and unstructured data, especially in relation to past interactions.

ETL: Extract, transform, load.

OLAP: Online analytical processing.

OLTP: Online transaction processing.

Open Data: Data which meets the following criteria: accessible (ideally via the internet) at no more than the cost of reproduction, without limitations based on user identity or intent. In a digital, machine readable format for interoperation with other data; and free of restriction on use or redistribution in its licensing conditions.

Structured Data: The term-structured data refers to data that is identifiable and organized in a structured way. The most common form of structured data is a database where specific information is stored based on a methodology of columns and rows. Structured data is machine readable and efficiently organized for human readers.

Unstructured Data: The term unstructured data refers to any data, that has little identifiable structure. Images, videos, email, documents, and text fall into the category of unstructured data.

This research was previously published in Predictive Intelligence Using Big Data and the Internet of Things; pages 19-50, copyright year 2019 by Engineering Science Reference (an imprint of IGI Global).