Power of Big Data Analytics: Transforming Government

Gopala Krishna Behara & Prasad Palli

Abstract— Leveraging of social media for the better citizen service delivery is continuously growing in government and public sectors. This leads to huge growth in transaction data volume per day in e-Governance demanding for Big Data Analytics. Big Data Analytics provide insights into how government schemes and policies are performing, and Why (Descriptive and Decisive Analyses). It helps in determining future scenarios and recommend best action plan (Predictive and Prescriptive Analysis) to measure sentiments of citizens and understand their perceptions and attitudes towards government policies. Big Data Analytics allows government to covert raw data into concentric diagrams, reports, map patterns for making better decisions by taking action based on patterns revealed by analyzing large volumes of data related and unrelated, structured and unstructured. Structured data includes information such as property tax, building permissions, criminal records and motor vehicle records which are usually stored in legacy systems or data warehouses. Unstructured data includes social media, email, videos and images and so on. This paper covers the Big Data Analytics ecosystem and its architecture to manage huge volume of data that will be generated by various sources. It also covers drivers, opportunities and benefits of Big Data Analytics implementation applicable to real-world.

Index Terms—**Big Data, Framework, Service Delivery, ETL, Structured and Unstructured Data, Data Lake**

I. Introduction

More than 500 million photos are uploaded and shared every day, along with more than 200 hours of video every minute. Since 2005, business investment in hardware, software, talent, and services has increased as much as 50 percent, to $4 trillion [1].

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 [2].

Big data in e-governance 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 [3]. 75% of Big Data is helping government departments to improve the quality of citizen's life style [4, 5].

Big Data Analytics application is an integrated Business Intelligence and Data Analytics system which includes conventional and Big Data. The system is proposed to be a government wide Analytical Engine which takes in data from various government department databases, internet, sensors, machine logs and other sources, transforms them, and presents them in an analyzable format.

The role of Big Data Analytics in enabling Government is gaining increasing significance, as indicated below:

1. **Insight**: Big Data Analytics opens up opportunity to use unlimited amount of data, particularly data to understand citizen sentiment and improve governance. Data from newspapers, blogs, social media, news channels, emails, reports, satellites, images etc. can be scanned, related, and structured in a manner that can be used for analysis. The insights provided by big data analytics enables the Government to predict the future and plan appropriate interventions or make mid-course corrections in all societally important sectors like healthcare, education, welfare and Urban/rural development and infrastructure. Regulatory functions of the government can be more effective with such insights.
2. **Timeliness**: With the use of Data Analytics Government can be PROACTIVE rather than REACTIVE. It is possible to achieve timely and agile response to opportunities, threats and challenges that the government faces. Today, traditional data could be used to identify these opportunities and threats, it will take time to collect, cleanse, standardize, and then analyze.
3. **Breadth**: Provide single view of diverse data sources – People, Entities, Land, etc. By combining traditional and big data under one platform, a true 360 degree analytical view can be achieved.

II. Business Drivers of BIG Data Analytics

Easy and timely retrieval and analysis of related and unrelated information is crucial for government to meet and improve mission requirements that are varied across departments and agencies. Data continues to be generated and digitally archived at increasing rates driven by Open Government initiatives, sensors, citizen interactions and program transactions. Government organizations are beginning to deploy Big Data technologies to analyze massive data sets as well as mining data to prevent bad actors

from committing acts of terror and/or to prevent waste, fraud, and abuse.
The Business Drivers of Big Data are described below:

A. *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 through less government

B. *Improved Government Performances*

- Insights into performance with respect to Department or Government KPIs
- Recommendations and interventions to improve performances

C. *Advanced Analytical capability*

- Capability to foresee key events and take appropriate and timely actions
- Capability to understand socio-economic events and conditions and align government policies to suit the same

D. *Utility of Data*

- Better utilisation of data – not merely for producing statistical reports on the past but intelligent reports that throw light on the future.
- Utilisation of data from multiple departments and sources, leading to creation of a holistic picture of events and happenings, and thereby enabling a more balanced administration.

III. Key Stakeholders

The Table below lists the stakeholders of Big Data Analytics and their expectations from it:

Stakeholder	**Requirement**	**Expectations from Big Data Analytics**
Government Departments	Analytical insights	1. Insights into what has happened, is happening, and why? 2. Insights on what is likely to happen (scenarios) 3. Guidance on best course of action to take.
Educational Institutes & Research bodies	A platform for research and discovery	A platform that enables to perform statistical and other forms of data science research
Government	Tool for good governance	Understand sentiments, perceptions and attitudes and mould Governance in a proactively

Table 1. Expectations from Stakeholders

IV. Opportunities and Benefits to Government

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

A. *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(drought only), 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 Portal 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.

B. *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 & 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

C. 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

D. Business Benefits

- Demand-Supply Analytics and Optimization of generation and power purchase planning
- Analyse the Trends of assessment and collection of property taxes in various counties/divisions and the average collection per property, and provide suitable advisories to the counties to improve tax collections
- Analyse the trends of prices of agricultural produce (top 5 crops for the last 5 years) predict the prices over the next 6 to 12 months and advise the department on the market interventions required

E. Productivity Gain

- To monitor the condition of the roads and provide advance recommendations on optimal resource utilisation for producing best impact on taxpayers.
- Compliance with SOPs

  - Hours of Power Supply to Across various sectors (ex. Agriculture, Industries, Domestic etc.,)

  - Release of new services within SOP norms (Standard of Performance)

  - Consumer grievances analysis within SOP norms

- 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.

V. BIG DATA ANALYTICS FRAMEWORK FOR GOVERNMENT

This section covers high level Big Data Analytics architecture including terminology used, application architecture. The main objective is to propose a framework for e-Governance BIG data platform in order to handle real time reporting from large set of Government data. This data may be taken from single or multiple applications; primarily those are storing data in various Hubs, Data Center databases and various data stores across the departments and agencies. The following diagram shows logical application architecture of Big Data Analytics System with key components and layers. A brief description of these components and layers is provided in this section.

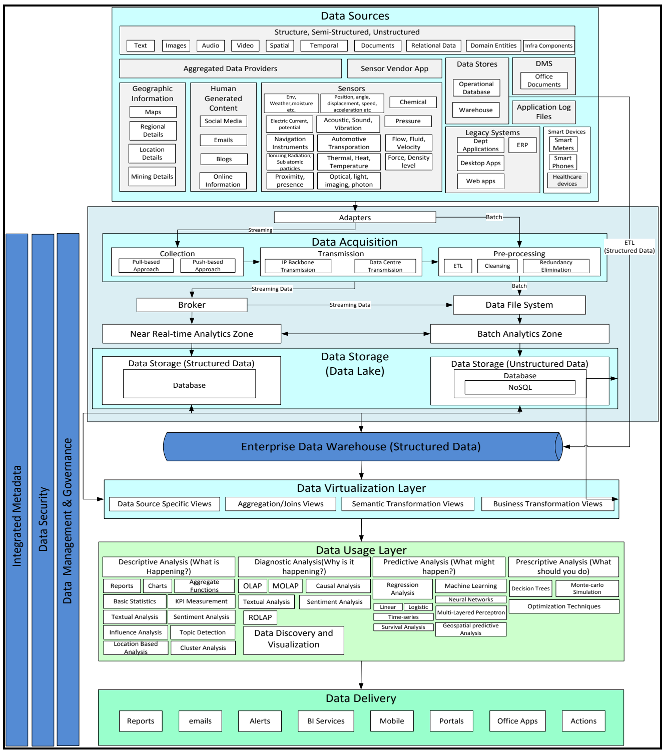

Figure 1: Logical Application Architecture View

A brief description of each of the above logical layers is provided below:

A. *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.

Category	Internal
Structured	Departmental Database, Data Hubs, Data Warehouse, Data Marts
Semi/Unstructured	e-Mails, Documents, XML documents

Category	External
Structured	
Semi/Unstructured	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

B. *Data Acquisition and Enrich*

Relevant push or pull based mechanisms are used to collect 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).
- **Batch Data** – Batch data is normally extracted from Government departments 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. Data Storage and Manage
- **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.

C. *Data Storage*

- **Data Lake -** It has the capabilities required to make it easy for developers, data scientists and analysts to store data of any size, shape and speed and do all types of processing and analytics. It stores all data while making it faster to get up and running with batch, streaming and interactive analytics.
- **Data Warehouse** – A Government Data warehouse stores whole of government data, comprising of structured data from departmental 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.

D. *Data Virtualization*

Data Virtualization provides a layer of abstraction and hide complexity of data storage and retrieval underneath. It hides cryptic names of tables and columns from users and provide business friendly definitions of data which can be used to create reports even by non-technical people. Also, 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 optimiser to tune query execution, for response time as well as throughput

E. *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 Anlytics 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

F. Data Management

Policies, processes, practices and technologies are used to manage data from source to destination. Data compression is an important aspect of Information Life Cycle Management. And compression ratio for data defined for Analytical system shall be applied. Big Data Analytics should have an in built Backup, Archive, and Restore (BAR) Solutions to protect data and ensure availability.

G. Data Security

Data security considerations specific to big data include,

- Increased value of the information asset as government agencies 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.

H. Data Usage Layer

Different users may want different types of outputs based on their role, responsibilities, and functions. 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.

1.1Data Consumers and Delivery

It describes how government 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 thru 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

VI. Conclusion

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 in e-Governance space,

- To provide insights into how government schemes and policies are performing, and Why(Descriptive and Causal Analyses)
- Design of Better Schemes and Projects by being more citizen-centric and effective
- To determine likely future scenarios and recommend best courses of action (Predictive and Prescriptive Analyses)
- To gauge sentiments of people of the government, and understand their perceptions of and attitudes towards government policies
- To provide a system of dashboards that enable administrators monitor and implement Government programs effectively
- To improve collaboration among departments
- To provide a tool for research in Data Sciences and statistical analysis
- To enhance the effectiveness of regulatory and tax collection systems of the government
- Enhanced Citizen 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 Government & Citizen 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

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

Acknowledgment

The authors 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 article as part of Global Enterprise Architecture Practice efforts.

REFERENCES

[1] IBM Big Data Hub (2012) http://www.investopedia.com/terms/c/certificateless municipals.asp#ixzz3a6KOjgFV

[2] McKinsey Global Institute, Big Data: The next frontier for innovation, competition, and productivity, May 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[3] National Science Foundation, Solicitation 12-499: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), 2012, http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

[4] http://www.sap.com/bin/sapcom/hu_hu/downloadasset.2013-09-sep-23-16.closing-the-big-data-gap-in-publicsector-pdf.html

[5] http://www.mondaq.com/australia/x/317214/Constitutional+Administrative+Law/Big+data+and+the+public+se ctor+strategy+and+guidance

Dr. Gopala Krishna Behara is a Senior Enterprise Architect in the Enterprise Architecture division of Wipro. He has a total of 19 years of IT experience.

Prasad Palli is a Practice Partner in the Enterprise Architecture division of Wipro. He has a total of 18 years of IT experience.

DISCLAIMER

The views expressed in this article are that of authors and Wipro or any other organization does not subscribe to the substance, veracity or truthfulness of the said opinion