BIG DATA ANALYTICS

The phenomenal growth in data and the endless possibilities that it promises in all spheres of life has led to the coining of the term Big Data. Gartner defines Big Data as "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". Big Data is not merely about the size of data; it embodies a whole new concept of opportunities waiting to be unveiled with the right tools, algorithms and techniques. In short, it represents the astronomical volume of both structured and unstructured data that is not attuned to process using the traditional database and software techniques.

Trends and Studies have shown the inroads that big data have made into almost all sectors, the consequences of which are most predominant in the domains of business, advertising, government, health care, social sector.

KAPIL KUMAR SHARMA
Principal Systems Analyst
kapilks@nic.in

YERUR SIRAJ AHMED
Scientific Officer/Engineer-SB
siraj.ahmed@nic.in

Edited by
MOHAN DAS VISWAM

The big data analytics is all about the techniques used for putting to use the large Volume of data and make cost effective analysis and prediction. Here it not only the sheer volume of data that need to be reconciled but also the different Varieties of structured data in traditional databases and unstructured data from sources like XML, JSON, eMails, Organizational Intranets and Enterprise Social networks. This data is coming at a very fast rate(Velocity). The characteristics of Big Data are now known popularly as the three Vs: Volume, Variety and Velocity. A fourth V has also been added which is Veracity. Veracity is the quality and authenticity of received data.

HOW BIG DATA ANALYSIS WORKS?

Data Preparation and Cleaning: The extracted data is subjected to statistical analysis for cleansing & verification and identifying outliers and applying appropriate treatment.

Transformation: Advanced algorithms or approaches are applied that extract information from a data sets both structured and unstructured and transform it into an understandable structure for further use.

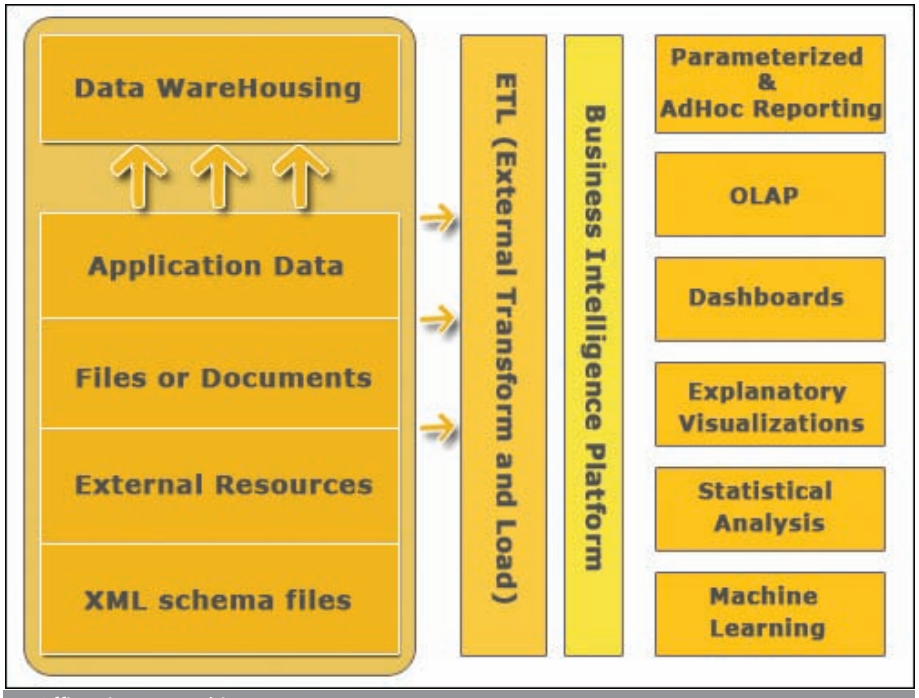
Business Intelligence (BI): Various BI tools for analyzing structure data and tools, Hadoop being one of the most popular among them, for unstructured data are applied for predictive analysis.

Visualization: A vast numbers of visuals/graphs/dashboards are generated that provide insights into the data. These are very exhausted visualizations and actually very complex (but quick) processing at the backend. Some really complex statistical analysis can be done quickly that till now was not achievable.

It may be mentioned that at each and every step a domain expert is required to support the accuracy of the analysis.

TECHNICAL ASPECTS

A template into which the extracted



e-Office Big Data Architecture

information should be contained has to be defined. Patterns are then defined on the transactional history. After identifying patterns we fit/find the models in the patterns. For various mathematical operations we may use numPy, R, python etc for calculations. We use predictions wherever required, using probability for getting most probable chances of an action, defining simple mathematical formula for making decisions.

Data Warehousing, Machine learning, Pattern recognition etc. play a major role in Big Data analysis. Predictive Analysis is one of the areas which analyze the current and historical facts to make predictions about future or otherwise unknown events. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions.

APPLICATION OF BIG DATA ANALYSIS IN THE GOVERNMENT: FEW SCENARIOS

1. "Aadhaar" Cards in India: Based on the data transformed, it is possible to project the total number of people below poverty line, average economy of a person etc. Based on the findings government may take a decision to improve the life of poor people, take women and child care schemes to enhance woman population etc. It may be correlated with rural health statistical parameters to predict onset of a disease.

2. Analysis of National health records for identifying the spread of epidemic or efficacy of clinical trials, with respect to region, gender age etc.

3. Employment exchanges may help unemployed job seekers find jobs by combining their job qualifications, place of residence, age etc. with an analysis of available job opportunities.

4. Data forensics of Logs or other data, logs data play a major role in understanding the customer demand and his usage of the applications. Based on the clicks that the user makes, the points through which he navigates provide us information that helps in enhancing the usability of the web pages and the application access.

BIG DATA: IMPROVING GOVERNMENT FUNCTIONING

Much emphasis is being laid by the Government on transparency and accountability in its internal functioning. e-Office, a Mission Mode Project is being implemented in the Government . In e-Office, the data of Ministries and Departments stored is vast enough to derive the decision making ability through analysis of transaction patterns.

• Turnaround times for decision making and pattern of decision making.

• Efficiency on the response to Citizen Centric case processing matters of Grievances, RTI and related heads.

Such a analysis may be used at some point to open such file for public viewing for openness and transparency.

• Based on the transaction history from e-Connect or e-Talk the social medium tools within eOffice, the Government user behavior and interests may be analyzed. Analysis may be made as to why some users log less frequently than others.

• Based on the books taken from Library, titles searched, trainings attended, projects worked and qualifications of employee, we identify their area or interest of the user and we may suggest books and training programs/seminars to the employee.

• The data from Personnel Information management Systems (PIMS) may be analysed for training requirements, resource allocation for various projects, future vacancies in the Government.

• The analysis of Personnel data can be extremely useful to the government for estimating the expenditure when declaring the Dearness Allowance(DA) for serving and retired government employees.

• The Personnel Information can be used for analysis to predict the posts getting vacant in near future, the government can make decisions about allotting these locations to the most suitable employees. These analysis may further be worked for understanding/ analyzing the employee motivation needs (for eg. Maslow's Hierarchy)

PRIVACY AND SECURITY CONCERNS

Examples given above are subjected to privacy and security concerns and data is to be made available subject to the existing rule and regulations.

BIG DATA SKILL SETS

The Big data analytics require various skill sets like statistics, operational research mathematics, Programming, machine learning etc. The combined skill sets is being now called as Data Science and anyone possessing these skills sets, the term Data Scientist is being increasingly used. However, even if these skill sets are possessed, the big data analytics is incomplete without domain expertise. Some of the technologies / languages / methodology used in Big Data analysis are Hadoop, Map Reduce, Machine learning, SAS, R, Python etc.

BIG DATA: AS A CHANGE AGENT

Government with its huge repository of data needs to devise and effective strategy framework and appropriate governance mechanisms for integrating with BIG data in terms of technology and skill sets. Followings Steps that can be considered in the overall move:

1. Formulate a scheme for common metadata management;

2. Policy on privacy and security for Big Data solutions;

3. Enabling deployment of required infrastructure for supporting the initiative;

The possibilities with Big Data within the Government structure are waiting to be unfolded and with the right strategy, policies and procedures combined with effective implementation, it can mark a whole new era of Governance.

FOR FURTHER INFORMATION:
Kapil Kumar Sharma
Scientist-D
e-Office Project
NIC, New Delhi
E-mail: kapilks@nic.in