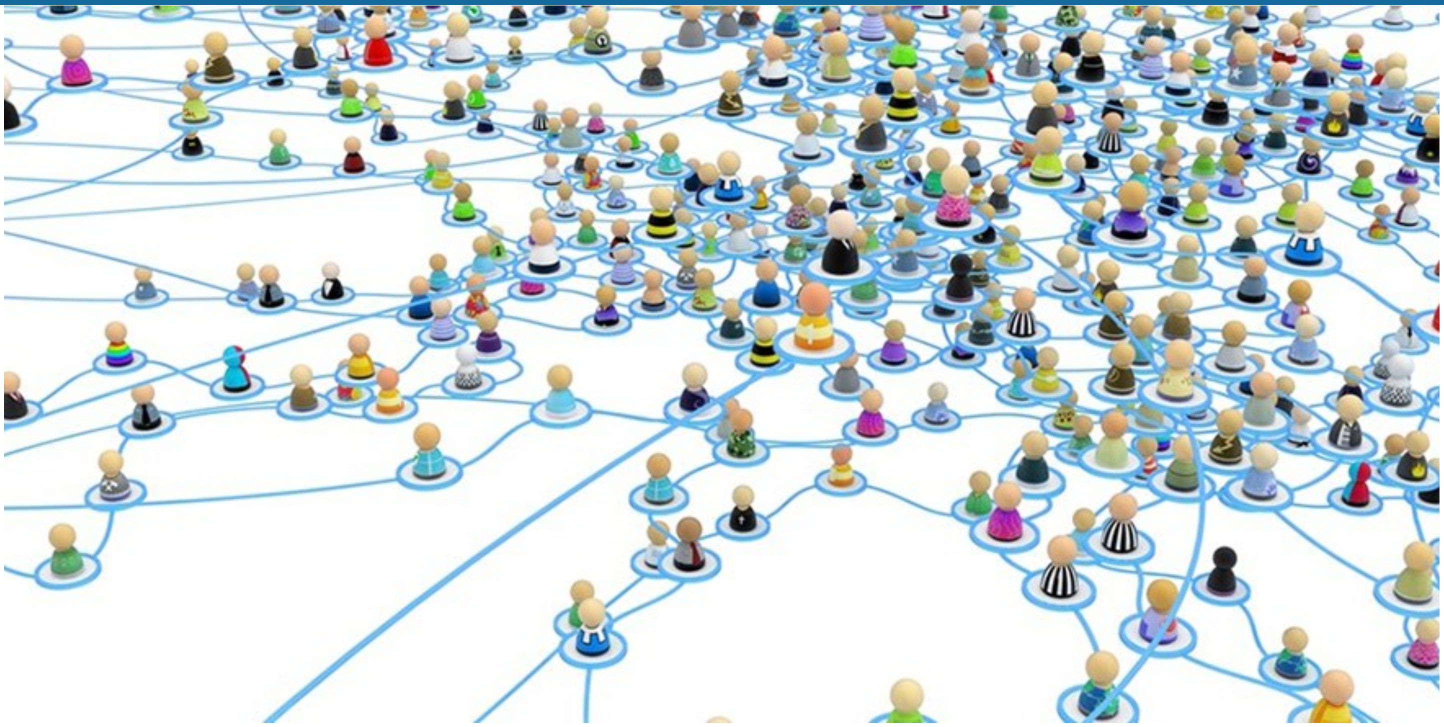
RAZSOFT CANADA | WHITE PAPER



DEVELOPING A BIG DATA STRATEGY

Big Data is changing our world rapidly and it’s impacting how we conduct our business in today's world. Every business today needs to find a way to deal with Big Data -- how to process it, how to use it, and how to monetize it.

RAZI CHAUDHRY
Founder & Principal Consultant at
RAZSOFT CANADA
razsoftcanada.com

Developing Big Data Strategy and modern Data Visualization, Building Data Sciences like Simulation, Regression & Correlation, Trending & Forecasting, Clustering & Segmentation, and Predictive Analysis to seek deeper Data Insights can be vital to your business.

BIG DATA continues to be the biggest trend and topic of discussion in variety of technology forums today, and it’s dramatically changing the way enterprises use information to better their customer experience. Many companies that invested early in big data and pioneered to analyze it are finding real value in big data.

Unlike structured data and analytics that has been traditionally used by marketeers, sales and operations — Big data is unstructured, voluminous and often fuzzy at first look. It wasn’t until recently that technologies evolved allows enterprises to process such large volumes of data at an affordable price tag. That coupled with an ability to build machine learning sciences to develop artificial intelligence to predict and forecast future. This ability is giving enterprises a deeper insights into their customers and partners.

Sources of BIG DATA are far too many. It is everywhere — where humans and machines interact. Internet, Social Media, Sensors, Logs, Machines, and vast range of electronic gadgets. A report suggests that there are 500 billion gigabytes of digital content on Internet, and it's likely to double in a year. Another report suggests there are 3.2 billion people now represented on Internet.

Spread of telecommunication, mobile networks, cloud computing, and electronic devices that are connected has given rise to new technologies to make sense of this in-comprehensively large data. Another field of technology Machine-to-Machine Communication (M2M) got a broader perspective as Internet of Things (IoT), which is a network of physical objects, devices, or any other items that are embedded with electronics, software, sensors, and network connectivity. All of these are generating more and more data — A really BIG DATA. And this data is continuing to grow at an unprecedented rates.

BIG DATA is changing our world rapidly and it's impacting how we conduct our business in today's world. Every business today needs to find a way to deal with BIG DATA -- how to process it, how to use it, and how to monetize it.

Today, decrease in the cost of both storage and compute power have made it feasible to collect and process this data. Using advance data sciences like Simulation, Correlation & Regression, Trending & Forecasting, Clustering & Segmentation, a machine learning algorithms can be developed to ingest Big Data to predict future, behaviors and events. For instance,

⇒ Healthcare industry is using big data to find usefulness of medicine by correlating prescriptions and patient's medical records.

⇒ Enterprises are mining social media's data to better understand customer preferences and concerns to improve its customer experience and craft better marketing strategy.

⇒ A geo-location data from mobile devices and sensors is used to analyze travel movements and shopping patterns.

⇒ Law enforcement agencies are using social media and sensor's data to proactively gather intelligence to reduce crime.

⇒ Cities are using big data to improve it's planning, and to improve management of traffic and commuters.

As an Enterprise Strategist, ignoring BIG DATA can bring fatal consequences to your organization's strategy, while the competitors may be able to find real-time insights to better theirs!

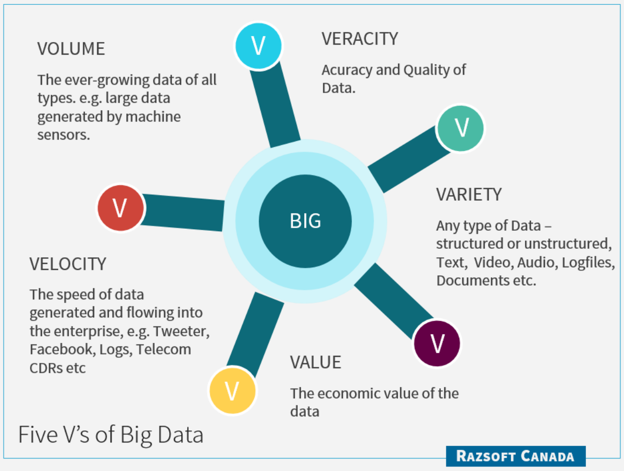
As an Enterprise Strategist, ignoring BIG DATA can bring fatal consequences to your organization's strategy, while the competitors may be able to find real-time insights to better theirs!

WHAT IS BIG DATA?

BIG DATA is a popular term used to describe tools, technologies and practices to process and analyze massive datasets that traditional data warehouse applications were unable to handle. However, Big Data landscape is still rapidly evolving.

The data and related data sciences has existed long before. Hence, big data is a collection of data from traditional and digital sources, either inside or outside the enterprise that represents a source for ongoing discovery and analysis, that can be structured or un-structured.

Big Data is often defined by **Five Vs**:


VOLUME
The ever-growing data of all types. e.g. large data generated by machine sensors.
V
VERACITY
Acuracy and Quality of Data.
V
VARIETY
Any type of Data – structured or unstructured, Text, Video, Audio, Logfiles, Documents etc.
V
BIG
V
VELOCITY
The speed of data generated and flowing into the enterprise, e.g. Tweeter, Facebook, Logs, Telecom CDRs etc
V
VALUE
The economic value of the data
V
Razsoft Canada

Five V's of Big Data

DEVELOPING BIG DATA STRATEGY

Big data holds many promises, such as gaining valuable customer insights, predict future, generate new revenue streams etc. However, formulating a clear business case and strategy for it is not without its challenges.

Big Data is different from traditional analytics in a sense that it is often an Exploratory Data Analysis (EDA). i.e. One questions leads to other, one finding leads to other findings. Hence without identifying a clear business objective its even more difficult to sell in an enterprise.

Often it's a catch-22. Finding and exploring value in data, first requires data! This has created a new role in an enterprise called, "Data Scientist" — a person who is inquisitive, exploring, asking questions, has good business acumen, can find patterns in data using statistical modeling, and then communicate informed conclusions and recommendations across an organization's leadership structure. Data Scientist is often a scarce resource to find!

KEY SUBJECT AREAS

A good Big Data Strategy will explore following subject domain, and align it to their organizational objectives:

1. Identify an Opportunity & Economic Value of Data
2. Defining Big Data Architecture
3. Selecting Big Data Technologies
4. Understanding Big Data Science
5. Developing Big Data Analytics
6. Institutionalizing Big Data

1. IDENTIFY AN OPPORTUNITY & ECONOMIC VALUE

First and foremost, a good strategy need to identify right opportunity for their enterprise, and what data make sense to explore.

a. Catalog existing data sources available inside the organization, tapped or untapped.

b. Invent new ways of capturing data, integrate your data sources with external communities. Develop semantics and metadata for association, clustering, classification and trending.

c. Identify and create opportunities to integrate and fuse data with partner's dataset in industry like Telecom, Travel, Financial, Healthcare, and Entertainment Industries etc.

d. Conceptualize the data insights and possible data sciences to extract valuable data, e.g. associations, simulation, regression, correlation, segmentation, trending, and predictive etc.

e. Identify the scope of data access. i.e. Who can explore data? Who gets access to Data Insights. This creates another deeper and at times a controversial debate that restricted access to data lakes by handful of data scientists and analysts affects the democratization of data, whilst too open access may reduce focused analysis and may compromise a potential competitive advantage, if a value is found.

f. Identify possibility of monetizing data to generate revenue from Data Insights gained, like generating leads, campaigns, upsell/cross sell opportunity, data streaming, data API, improving staff productivity & customer service etc.

g. Identify ethical and legal code associated with data under exploration with respect to industry standards, organizational culture, data policies, data privacy, regulatory and legal requirements.

2. DEFINING BIG DATA ARCHITECTURE

BIG DATA Architecture is an evolving practice for developing an automated data pipeline on a reliable and scalable platform at a lower cost, to gather and maintain large collections of data, and to process and apply advance statistical models for extracting useful data insights and information.

BIG DATA Architecture deals with fairly large landscape and associated challenges:

a. Defining Business Problems & Classification of associated data, such as, Market Sentiment Analysis, Churn Predictions, or Fraud Detection.

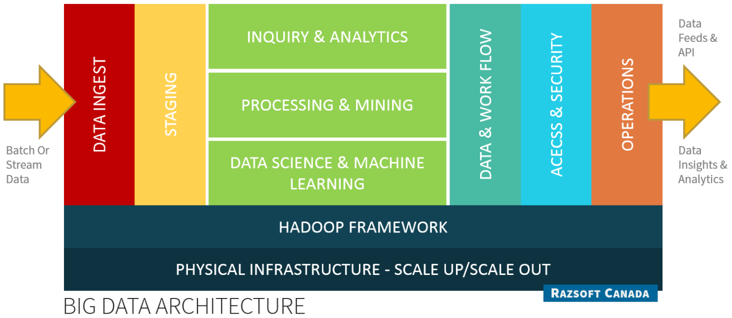
b. Defining Data Acquisition Strategy .

c. Selecting a Hadoop Framework & Distro.

d. Big Data Life Cycle Management Framework.

e. Choosing Big Data stores, traditional or noSQL and Polyglot persistence.

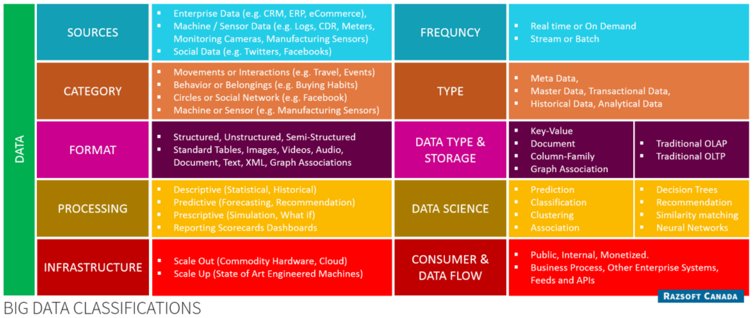
f. Defining Big Data Infrastructure & Platform Taxonomy



BIG DATA ARCHITECTURE

g. Identifying Big Data Analytics Frameworks, and associated Machine Learning Sciences.

h. Lastly, develop Data Monetization Strategy to exploit its value internally within the enterprise, or externally.



DATA	SOURCES	▪ Enterprise Data (e.g. CRM, ERP, eCommerce), ▪ Machine / Sensor Data (e.g. Logs, CDR, Meters, Monitoring Cameras, Manufacturing Sensors) ▪ Social Data (e.g. Twitters, Facebooks)	FREQUNCY	▪ Real time or On Demand ▪ Stream or Batch	
	CATEGORY	▪ Movements or Interactions (e.g. Travel, Events) ▪ Behavior or Belongings (e.g. Buying Habits) ▪ Circles or Social Network (e.g. Facebook) ▪ Machine or Sensor (e.g. Manufacturing Sensors)	TYPE	▪ Meta Data, ▪ Master Data, Transactional Data, ▪ Historical Data, Analytical Data	
	FORMAT	▪ Structured, Unstructured, Semi-Structured ▪ Standard Tables, Images, Videos, Audio, Document, Text, XML, Graph Associations	DATA TYPE & STORAGE	▪ Key-Value ▪ Document ▪ Column-Family ▪ Graph Association	▪ Traditional OLAP ▪ Traditional OLTP
	PROCESSING	▪ Descriptive (Statistical, Historical) ▪ Predictive (Forecasting, Recommendation) ▪ Prescriptive (Simulation, What if) ▪ Reporting Scorecards Dashboards	DATA SCIENCE	▪ Prediction ▪ Classification ▪ Clustering ▪ Association	▪ Decision Trees ▪ Recommendation ▪ Similarity matching ▪ Neural Networks
	INFRASTRUCTURE	▪ Scale Out (Commodity Hardware, Cloud) ▪ Scale Up (State of Art Engineered Machines)	CONSUMER & DATA FLOW	▪ Public, Internal, Monetized. ▪ Business Process, Other Enterprise Systems, Feeds and APIs	

RAZSOFT CANADA

BIG DATA CLASSIFICATIONS

CLASSIFICATION OF BIG DATA

BIG DATA architecture varies by industry, and is dependent upon nature of business problem and type of big data that needs to be sourced to solve that problem.

General characteristics of big data deals with, How to acquire the data? What is the structure and format of the data? How frequently data becomes available? What is the size of data? What is the nature of processing required to transform this data? What algorithms or statistical model is required to mine the data?, etc.

DATA ACQUISITION

It involves ingesting data from variety of sources including sensors, mobile network, internet, social data, existing Enterprise OTLP/OLAP data, or archived data that is not tapped, e.g. log data.

Build partnerships with key industry players to fuse data with industries like Telecom, Travel, or Entertainment etc.

Integrate data sources with aligned and external businesses to derive 360-degree insights.

BIG DATA LIFE CYCLE MANAGEMENT

While Big Data is part of overall Data Management process in an Enterprise, its demands its own unique cycle due to its nature. It can be described in six key steps: 1) Acquire 2) Classify & Organize 3) Store 4) Analyze 5) Share & Act 6) Retire

Industry uses CRISP-DM, a comprehensive data mining methodology and process model that provides a complete blueprint for conducting a data mining project.

BIG DATA STORE

Big Data consists of structured, unstructured and semi-structured data. It includes relational databases (OLTP or OLAP), as well as other non-relational noSQL databases like key-value, document, columnar, graph or GeoSpatial data stores. A typical Big Data implementation will include multiple databases to serve different needs of Big Analytics -- a theme that's knowns as Polyglot persistence.

noSQL

noSQL Stands for NOT ONLY SQL, however, it is not a formal definition. Its an umbrella term for unstructured data stores, though they may support SQL-like query languages. It provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

BIG DATA INFRASTRUCTURE

Big Data requires an Elastic Infrastructure with Compute and Storage architecture that can deal with petabytes of high velocity data on-demand, and to support the scale and complexity of Big Data Architecture.

Virtualization is fundamental to both Cloud Computing and Big Data. It provides high efficiency and scalability for Big Data Platform, and gives MapReduce the desired distributed environment with endless scalability.

Typically, an implementation will have a hybrid infrastructure of in-house and cloud, that supports both **Scale up and Scale out options**. Scaling up is replacing current technology with something more powerful, e.g. 1GbE switch with 10GbE switch. Scale out means taking the current infrastructure, and replicates it to work in parallel in distributed environment.

Cloud provides three key support for Big data, i.e. Scalability, Elasticity and Flexibility. A cloud-based big data solution will complement existing enterprise infrastructure, especially to support real-time on-demand scalability.

Security is of paramount concern in big data environment due to the nature of collected data, privacy concerns, regulations and compliance. Careful Security policy is required to mitigate risks. Numerous techniques are available, such as, Tokenization, Sanitization, Data isolation etc.

3. SELECTING BIG DATA TECHNOLOGIES

There has been a massive amount of innovation in Big Data Tools & Technologies over last few years. Technologies that are coming out in Big Data are based on following key trends:

- **Internet Technologies:** Technologies developed by Internet, Web and Social media companies like Google, Facebook, Amazon developed text processing, log processing, social networks etc.
- **Machine Learning:** Apply Statistical modeling, computational science and machine learning to invent Artificial Intelligence to solve real work problems, and to provide analysis, intelligence, predictions, and deep insights.
- **Commodity Hardware:** Use Commodity hardware that is much more cheaper than traditional hardware, and support distributed processing at much lower price tag.
- **Distributed Processing:** Leverage Distributed processing to power up and Use Scale out option to build infrastructure on demand, and destroy once the outcome is achieved
- **Cloud:** Leverage Cloud based approach to reduce time to market, reduce risk and gain better SLA out of the box.

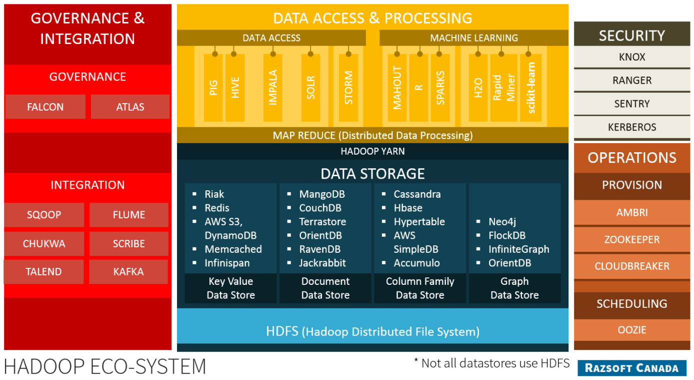
Big Data innovations can be grouped into following subject areas:

- Big Data Software & Platforms
- Hardware Engineered for Big Data
- Cloud based Solutions

BIG DATA PLATFORMS

There are large number of Platforms and tools available for **Big Data Analytics Platform**. Large number of commercial versions have come onto the market in recent years, as vendors have created their own versions designed to be more user-friendly, providing highly distributed architecture, new levels of processing and memory power, and often cloud base.

Some vendors have open source licensing models, others leverage their engineered hardware platform, or engineered cloud platforms, or consulting services.



HADOOP ECO-SYSTEM

* Not all datastores use HDFS

The popular distributions are provided by: Apache Hadoop, Hortonworks, Cloudera, Oracle, MapR, Pentaho, IBM, Amazon AWS, 1010 Data, Microsoft Azure HD Insights, Datastax Enterprise Analytics, Pivotal HD GreenPlum

There are large number of applications and tools available on Big Data, and many of them can be provisioned in the cloud.

HADOOP

Hadoop is a framework and set of tools for processing very large data sets. It was designed to work on cluster of servers using commodity hardware, providing powerful parallel processing on compute and data nodes at a very low price. Technology is rapidly advancing both for software and engineered hardware, making this framework even more powerful. Hadoop implements a computational paradigm known as MapReduce, which was inspired by an architecture developed by Google to implement its search technology, and it's based on "map" and "reduce" function from LISP Programming.

Today large number of Tools & Technologies are available that staples to Hadoop and commonly termed as Hadoop or Big Data Eco-System.

BIG DATA TOOLS IN HADOOP ECO-SYSTEM

ACQUISITION

- Apache Falcon is a data governance engine that defines, schedules, and monitors data management policies.
- Apache Atlas is an extensible set of core governance services that enables enterprises to meet their compliance requirements.

INTEGRATION

- Apache Sqoop provides interface to transfer bulk data between (to/from) non-hadoop or relational databases and Hadoop.
- Apache Flume is a service for efficiently collecting, aggregating, and moving large amounts of log data.
- Apache Chukwa is a data collection system for monitoring large distributed systems. It can collect logs and perform analysis. It includes powerful toolkit for displaying, monitoring and analyzing results.
- Apache Kafka is publish-subscribe messaging rethought as a distributed commit log. It sends large numbers of events from producers to consumers.

ACCESSING DATA

- Apache Pig is SQL-Like environment to access data in Hadoop. It support scripting language called Pig Latin.
- Apache Hive is also SQL-Like. It provides batch-like programming interface to Hadoop. It has its own table, partitions, buckets (files) and supporting meta data.
- Impala is native analytic MPP database for Apache Hadoop and is supported by Cloudera Enterprise. It provides ANSI SQL compatibility.

PROCESSING

- MapReduce (first developed by Google) is a software framework, and it's the core of Hadoop Eco System, that allows to write applications that can process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

- Mahout is a scalable data mining library, that provides machine learning algorithms for clustering, classification, recommendations etc.
- R is a language for statistics and graphics, and it is most used by data scientists. It provides wide variety of statistical and graphical techniques. R-Studio has a free R IDE.
- Apache Spark allows fast in-memory data processing, It provide APIs in Scala, Java, R, and Python for machine learning and data analysis.

DATA STORES

- **KEY-VALUE** stores are Simplest noSQL data stores that can be persistent or run totally in memory. Popular are Riak, Redis, DynamoDB.
- **COLUMNAR** stores are unstructured data store. They store data in tabular format (in columns). Popular are Cassandra, Hbase (Like Google BigTable), HyperTable, Amazon (SimpleDB).
- **DOCUMENT** are semi-structured data stores. There is no pre-defined schema or structure, like XML. Popular are MongoDB, CouchDB, JackRabbit, RavenDB, OrientDB, TerraStore.
- **GRAPH** stores entities and their relationships, like Social Circles, Associations etc. Popular are Neo4j, FlockDB, InfiniteGraph, OrientDB.
- **GEOSPATIAL** stores are optimized to store and query data defined in a geometric space. Popular are Oracle Spatial, GeoMesa, Neo4j, and many columnar data stores now supports it.

SECURITY & OPERATIONS

- Apache Ranger is a framework security across the Hadoop platform.
- Apache Ambari is provisioning, managing, and monitoring platform for Hadoop Cluster.
- Apache ZooKeeper performs distributed configuration, synchronization, naming registry.
- Cloudbreak automates provisioning, managing and monitoring of on-demand clusters in the cloud for elastic Hadoop clusters.
- Apache Oozie is a workflow scheduler system to manage Hadoop jobs.

4. UNDERSTANDING BIG DATA SCIENCE

Data science is transforming the research and commerce. Data Science is the ongoing process of discovering information from data. It is a process that never stops, and often one question leads to another new question. It focuses on real-world problems, and tries to explain it.

Data scientists use mathematical and statistical methods to build decision models to solve complex business and scientific problems. Statistical methods are useful to understand data, to validate hypotheses, to simulate scenarios, and for making predictive forecasts of future events, e.g. linear regression, Monte Carlo simulations, and time series analysis etc.

Data scientists are required to have a strong subject -matter and domain-specific expertise in the area in which they're analyzing, e.g. city planning, crime analysis, energy efficiency, customer churn etc.

Common techniques involves Inductive Models, Deductive Models, Analogical Learnings, and Reinforcement Models.

MACHINE LEARNING

Machine Learning provides ideal technique for exploiting opportunities hidden in big data and finding data insights. It is useful where traditional analytic tools are not adequate to deal with large volumes of data.

Machine Learning can help find correlations and relationship between desperate data, and provide techniques to test all hypothesis to investigate hidden value in the data.

TYPES OF MACHINE LEARNING

There are three types of Machine Learning techniques.

a. Supervised, based on previously known data
b. Unsupervised, no prior data rather learning through exploration and finding patterns
c. Hybrid is combination of both Supervised and Unsupervised.

COMMON ALGORITHMS

- **CLASSIFICATION** — used to predict an outcome, like fraud detection, targeted marketing, prediction, manufacturing diagnosis, credit risk.
- **CLUSTERING** — used to partition data into subsets, like Customer Segmentation, targeted campaigns, Census and Social analysis.
- **ASSOCIATIONS & CORRELATIONS** — uses Frequent Patterns to find associations, like Market Basket Analysis to find Customer buying patterns.
- **TEXT MINING** — is used to find and exploit useful patterns in text or unstructured data, e.g. emails, machine logs, tweets and blogs, text documents, multi-media contents etc.
- **LINEAR REGRESSION** — is the most commonly used statistical method for numeric predictions, like Trending Sales to forecast, predicting election results, insurance claims etc.

5. DEVELOPING BIG DATA ANALYTICS

Big Data applications varies based on industry. Businesses are trying to find value in monetizing data, or use it to improve efficiencies and customer experience. Some of the key trends are:

- Identifying new revenue streams
- Improving Customer Experiences
- Agile Marketing & Campaigning
- Predicting Churn & Customer Behaviours
- Fraud Detection and Prevention
- Security Intelligence and Law Enforcement
- Research, Planning & Development

TYPES OF ANALYTICS

Typically there are four groups of analytics:

a. **DESCRIPTIVE** — based on current and historical data, deals with "What happened", like Quarterly Sales, or Year by year revenue etc.
b. **DIAGNOSTIC** — to diagnose a situations or "What went wrong?" or why, like reviews and customer feedback, actual sales etc.
c. **PREDICTIVE** — models based on current and historical data, to predict future outcome.
d. **PRESCRIPTIVE** — to improve or optimize a process or system, or to avoid a failure through informed action. e.g. denying fraudulent credit card based on fraud control analytics and scoring.

6. INSTITUTIONALIZING BIG DATA

Each enterprise will tailor the Big Data to meet the objectives of their particular vision. However, since Big Data significantly differs from that of traditional data warehouse, hence, a traditional approach with SDLC and ITIL may not always work.

A more suitable approach is of R&D, where empowering employees with their own Sandboxes will yield higher returns on investment. A typical methodology may include following steps:

a. Discovery (Opportunity, Requirements, Best Fit)
b. Proof of Concept (to Evaluate Business Value)
c. Provision Infra (here Big Data Elasticity comes to play)
d. Ingest (Source the data)
e. Process (Transform, Analyze, Data Science)
f. Publish (Share the learnings)