International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 3, March 2017

ISSN: 2394-4404

Business Intelligence & Predictive Analytics In Big Data For Big Insights

Sanjeev D. Chaube

Chief Data Scientist & Global Practice Head (Big Data Analytics) a Leading IT Company, Lunkad Abode, Viman Nagar, Pune

Professor (Dr.) P.H Karmadkar

Modern High Campus, Yamunanagar, Nigdi, Pune

Abstract: ***A forecast from International Data Corporation (IDC) sees the big data technology and services market growing at a compound annual growth rate (CAGR) of 23.1 per cent over the 2014-2019 forecast period with annual spending reaching USD 48.6 billion in 2019. This wide range includes domain such as Banking, Government, Manufacturing, Utilities, Telecommunications Education, Healthcare, Retail, Insurance, Securities, Railroads, Customer Service / BPO etc.***

One of the major applications of future generation parallel and distributed systems is in big-data analytics. Data repositories for such applications currently exceed exabytes and are rapidly increasing in size. Beyond their sheer magnitude, these datasets and associated applications' considerations pose significant challenges for method and software development. Datasets are often distributed and their size and privacy considerations warrant distributed techniques. Data often resides on platforms with widely varying computational and network capabilities.

In this paper, we will analyze the current BI & Analytics Solutions for Big Data Analytics world market in order to come up with business plan recommendations for better decision making and opportunities generated by the global BI Software Solutions in open source environment.

Keywords: Decision Support System (DSS), Advanced Analytics, Big Data, Hadoop, Apache Spark, Scala, Sqoop, Flume, Pig, Hive, Hbase, Cassandra, Mongo DB, Business Intelligence Software & Applications, Machine Learning, Statistics, Mathematics, Open Source Environment, Enterprise Resource Planning (ERP), Analytics, Workforce Data, Data Warehousing, MapReduce, Data Visualization, SQL, Open Source, Social Data, Cloud Computing, Data Mining, Data Quality, IT, Software, Business Intelligence, Collaborative Data, Data Management, Unstructured Data Performance Management, Predictive Analytics, Cloud, Data Mining.

I. INTRODUCTION TO BIG DATA ANALYTICS

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves using Statistical & Machine Learning techniques.

Further Analytics which is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight. Firms may apply analytics to business data to describe, predict, and improve business performance. Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modelling, web analytics, sales force sizing and optimization, price and promotion modelling, predictive science, credit risk analysis, and fraud analytics. Since

analytics can require extensive computation, the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics.

II. IMPORTANCE OF BIG DATA WITH ANALYTICS

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyse it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

III. DEFINING THE BIG DATA

Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale. The critical processes & challenge of big data:

Store. Can you capture and store the data?

Process. Can you cleanse, enrich, and analyze the data?

Access. Can you retrieve, search, integrate, and visualize the data?

IV. CHARACTERISTICS OF BIG DATA

VOLUME: The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

VARIETY: The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

VELOCITY: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

VARIABILITY: Inconsistency of the data set can hamper processes to handle and manage it.

VERACITY: The quality of captured data can vary greatly.

TYPES BIG DATA

STRUCTURED DATA: ERP, CRM, Eco/Fin, Lob etc

UNSTRUCTURED DATA: Log Files, Sensors, Social Networks etc

SEMI STRUCTURED DATA: XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured.

BIG DATA ANALYTICS & MODELLING FOR INSIGHTS

- ✓ Classification Models
- ✓ Segmentation Models
- ✓ Association Models
- ✓ Anomaly Detection Models
- ✓ Dimension Reduction Models
- ✓ Forecast Models
- ✓ Probabilistic Scoring Models and so on.

V. COMPONENTS OF BIG DATA ANALYTICS

A. BUSINESS LAYERS IN BIG DATA ANALYTICS SOLUTIONING

Figure 1

A big data solution typically comprises these logical layers

a. BIG DATA SOURCES

All of the data available for analysis, coming in from all channels. Data scientists clarifies what data is required to perform the kind of analyses required by the client. This layer includes all the data sources necessary to provide the insight required to solve the business problem. The data is structured, semi-structured, and unstructured, and it comes from many sources

- ✓ Enterprise legacy systems (Billing Operations, CRM, ERP etc)
- ✓ Data management systems (DMS) (MS Excel, MS Word Documents etc)
- ✓ Data stores (Enterprise Data Warehouse, Operational database and transactional database)
- ✓ Smart devices (e.g smartphones, meters, healthcare / travel tourism devices)
- ✓ Aggregated data providers (Geographical information, human generated content, sensor data etc.)

b. DATA MASSAGING AND STORE LAYER

This layer is responsible for acquiring data from the data sources and, if necessary, converting it to a format that suits

how the data is to be analyzed. For example, an image might need to be converted so it can be stored in an Hadoop Distributed File System (HDFS) store or a Relational Database Management System (RDBMS) warehouse for further processing. Compliance regulations and governance policies dictate the appropriate storage for different types of data. Because incoming data characteristics can vary, components in the data massaging and store layer must be capable of reading data at various frequencies, formats, sizes, and on various communication channels:

DATA ACQUISITION: Acquires data from various data sources and sends the data to the data digest component. It must be able to determine whether the data should be massaged before it can be stored or if the data can be directly sent to the business analysis layer.

DATA DIGEST: Responsible for massaging the data in the format required to achieve the purpose of the analysis. This component can have simple transformation logic or complex statistical algorithms to convert source data.

DISTRIBUTED DATA STORAGE: Responsible for storing the data from data sources. Often, multiple data storage options are available in this layer, such as distributed file storage (DFS), cloud, structured data sources, NoSQL, etc.

c. ANALYSIS LAYER

The analysis layer reads the data digested by the data massaging and store layer. In some cases, the analysis layer accesses the data directly from the data source. Designing the analysis layer requires careful forethought and planning. Decisions must be made with regard to how to manage the tasks to:

- ✓ Analysis-layer entity identification
- ✓ Analysis engine
- ✓ Model management

d. CONSUMPTION LAYER

This layer consumes the business insight derived from the analytics applications and enables monitoring & building, Business process management processes, Real-time monitoring, Reporting engine, Recommendation engine, Visualization and discovery

e. COLLABORATIVE LAYER

Aspects that affect all of the components of the logical layers (big data sources, data massaging and storage, analysis, and consumption) are covered by the vertical layers:

INFORMATION INTEGRATION

Big data applications acquire data from various data origins, providers, and data sources and are stored in data storage systems such as HDFS, NoSQL, and MongoDB. This vertical layer is used by various components (data acquisition, data digest, model management, and transaction interceptor, for example) and is responsible for connecting to various data sources. Integrating information across data sources with varying characteristics (protocols and connectivity, for example) requires quality connectors and adapters.

BIG DATA GOVERNANCE

Big data governance helps in dealing with the complexities, volume, and variety of data that is within the enterprise or is coming in from external sources.

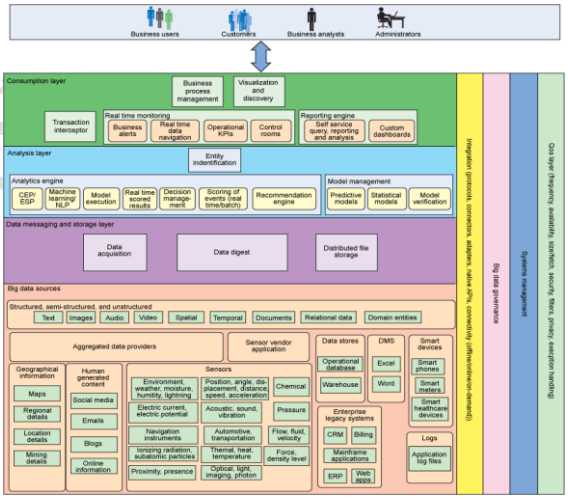
Systems management

Systems management is critical for big data because it involves many systems across clusters and boundaries of the enterprise. Monitoring the health of the overall big data ecosystem.

QUALITY OF SERVICE

This layer is responsible for defining data quality, policies around privacy and security, frequency of data, size per fetch, and data filters

COMPONENTS OF BUSINESS LAYERS IN BIG DATA ANALYTICS SOLUTIONING



Source: IBM

Figure 2

B. TECHNOLOGY LAYERS IN BIG DATA ANALYTICS SOLUTIONING

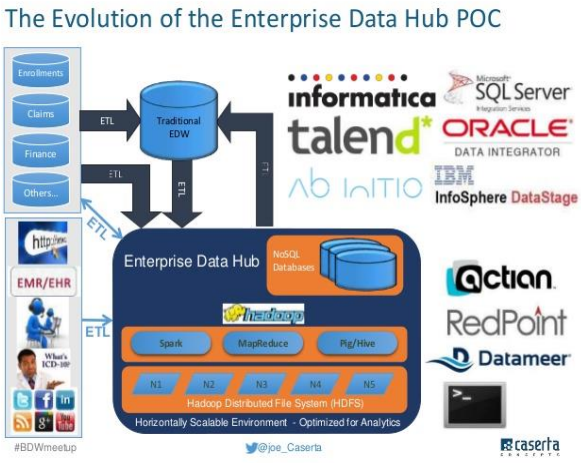


Figure 3

a. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

The Hadoop Distributed File System HDFS is designed to hold large amounts of data. It is a Java based file system and provides faster access to data. It is highly scalable having scalability up to 200 PB of storage. HDFS is also highly reliable and fault tolerant system. A single cluster of HDFS containing 4500 servers can support billion files and blocks.

The master slave architecture of HDFS consists of a NameNode and DataNode. NameNode is a centralized process maintains the directory tree of HDFS .It can performs common operations like rename, delete, openand close It does not store any data, but maintains a map of the blocks in the file, the file name, and the DataNode where the blocks are stored. The data of the file is split into 3 parts, each of size 128 Megabytes, and each part is stored separately at multiple DataNodes. This is done so that the data is not lost even if one DataNode fails. Blocks are created or destroyed at the request of NameNode, which processes the requests from clients. Clients can communicate with the DataNodes directly to read or write data at the HDFS block level. Apache Spark can be used with HDFS in following 3 ways: Standalone deployment, using YARN, and Spark in MapReduce.

b. BIG DATA ANALYTICS USING SPARK

Apache Spark is a high performance framework for analyzing large datasets. It was developed at UC Berkeley AMPLAB as an alternative to Hadoop MapReduce framework. Apache spark consists of a driver program (SparkContext), workers also called executors, cluster manager, and the HDFS. Driver program is the main program of spark. SparkContext is the object that gets created during execution of spark program, and is responsible for entire execution of the job. The SparkContext object connects to cluster manager, which are used to manage the resources across cluster. Cluster managers provide Executors, which are used to run the logic and also storing the app data.

Spark is based on two concepts: Resilient Distributed Datasets (RDD) and execution engine Directed Acyclic Graph (DAG).

- ✓ *RESILIENT DISTRIBUTED DATASETS (RDD)* are Spark's primary abstraction, which are a fault-tolerant collection of elements that can be operated on in parallel. They are immutable once you create an RDD. They can be transformed, or actions can be performed on them, but they cannot be changed. They help with rearranging the computations and optimizing the data processing. They are also fault tolerant because an RDD know how to recreate and compute the datasets. RDD can be constructed by paralyzing existing collections such as lists, or by transforming existing RDD, or from existing files in HDFS. A spark programmer specifies the number of partitions for the RDD and if not specified, a default value is used.

  There are two types of operations that can be performed on RDD's:

  TRANSFORMATION: When transformations are applied on RDD's, they return a new RDD and not a single value. Transformations are lazily evaluated, i.e. they are not computed immediately. They are executed only when an action runs on it. Some of the Transformation functions are map, filter, ReduceByKey, FlatMap and GroupByKey.

  ACTION: When Action operation are applied on RDD's, they evaluate and return a single value. All the queries that process the data are computed when an Action function is called, and return the result value. Some Action operations are first, take, reduce, collect, count, foreach and CountByKey.

- ✓ *DIRECTED ACYCLIC GRAPH (DAG)*

Spark consists of an advanced Directed Acyclic Graph (DAG) engine which supports cyclic data flow. Each Spark job creates a DAG of task stages to be performed on the cluster. DAGs created by Spark can contain any number of stages, as compared to MapReduce, which creates a DAG with two stages - Map and Reduce. This allows simple jobs to complete after just one stage, and more complex tasks to complete in a single run of many stages, rather than splitting it into multiple jobs. Thus, jobs complete faster than they would in MapReduce.

There are 4 core Apache spark component: Spark SQL, Spark Streaming, GraphX and MLlib (Machine learning library) -

Spark Core is the foundation of the framework. It provides basic I/O functionalities, distributed task dispatching, scheduling.

- ✓ *SPARK SQL:* Spark SQL allows running SQL like queries on Spark data using traditional BI and visualization tools. It provides support for structured and semi structured data by introducing a new data abstraction Schema RDD. It also provides SQL language support, with command-line interfaces and ODBC/JDBC server.
- ✓ *SPARK STREAMING:* Spark streaming allows processing the real-time data. It uses DStream, which is a series of RDD's, to process real-time data.
- ✓ *SPARK GRAPHX:* GraphX introduces the Resilient Distributed Property Graph, which is directed multi-graph having properties attached to each edge and vertex. GraphX includes a set of operators like aggregate Messages, subgraph and join Vertices, and optimized variant of Pregel API. It also includes builders and graph algorithms to simplify graph analytics tasks.
- ✓ *MLLIB:* MLlib is Spark's scalable machine learning which consists of utilities and common learning algorithms like regression, classification, collaborative filtering and dimensionality reduction.

VI. CONCLUSION

Data growth today is phenomenal and in the world of Business Intelligence; Data Science & Advanced business

analytics solutions have played key role in turning Big Data into information using Statistical & Machine Learning Algorithms in open source environment. The present research paper attempts to provide an insight into both the Business as well as the Technological perspective in Big Data Analytics & Solutioning.

As we have entered the era of Big Data, new analytics tools like Hadoop MapReduce and Apache Spark have been developed to analyze and process this Big Data. Apache Spark has gained significant momentum and is considered to be a promising alternative to support iterative processing logic and ad-hoc queries by replacing MapReduce. Apache Spark has received a lot of appreciation in many fields like pattern recognition, machine learning, data mining, information retrieval, and image retrieval. However, as the amount of data to be processed grows, many data processing methods have become less efficient. This paper exploits the

Apache Spark framework for efficient analysis of big data in HDFS, and compares it with other data analysis framework-Hadoop MapReduce. It is seen that Apache Spark increases the speed of computation of iterative algorithms and completes them in much less time as compared to Hadoop MapReduce. Apache Spark also provides a high performance, highly scalable and fault tolerant framework for big data analysis.

REFERENCES

[1] Understanding The Various Sources of Big Data, https://datafloq.com/read/understanding-sources-big-data-infographic/338 @ 2016
[2] Big data Analytics, http://www.sas.com/en_us/insights/analytics/big-data-analytics.html. 2016
[3] Hadoop Tutorial,YahooInc., https://developer.yahoo.com/hadoop/tutorial/index.html 2016
[4] Big Data Analytics, http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
[5] J. Shafer, S. Rixner, A.L. Cox, "The Hadoop Distributed Filesystem: Performance versus Portability", IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2015), White
[6] Plains, NY (March 2015).
[7] https://databricks.com/blog/2014/01/21/spark-and-hadoop.html.
[8] Apache Spark, http://spark.apache.org/
[9] Davenport, Thomas H. (2006 January). Competing on Analytics, Harvard Business Review, Prod. #: R0601H-PDF-ENG. Friedman T. L. (2015). The World is Flat New York: Farrar, Straus and Giroux.
[10] Gartner (2016 January). "Business Intelligence ranked Top Technology Priority by CIOs for Fourth Year in a Row". Gartner (2016 January). "Market Share: Business Intelligence, Analytics and Performance management Software Worldwide, 2016".
[11] International Data Corporations (IDC) (2016 Feb). "The Digital Universe Decade – Are you Ready?"
[12] International Data Corp (IDC) (2015 December). "Worldwide Business Analytics Software Forecast and 2015 Vendor Share".