An Integrated framework of Big Data, Business Intelligence and Analytics for taking efficient business strategic decisions

Dr. Mukesh Negi
Sr. Technical Delivery Manager TechMahindra Ltd, Noida (U.P), India

Abstract - Preparing an efficient business strategy is a vital goal for any organization. Consistent growth and future of an organization are completely depending on the strategic decisions taken by the upper management. Data is the backbone for every organization and plays a critical role from every front. Especially, when its historical data of the past few years, then its works like a catalyst in defining the strategy based on the previous patterns. Traditional ways were not sufficient due to the complexity of manual interventions. Technologies have been evolved rapidly in past few years. Today we have sufficient technologies available to automate the complete strategic decision process, but the integration between all to achieve maximum benefits from all is still a challenge [1]. In this paper, I have demonstrated an integrated framework of big data, business intelligence, and analytics which will help organizations to integrate all technologies to get maximum benefits by defining efficient strategic decisions for consistent growth of the organization.

Keywords—*IoT; Cloud Computing; Big Data; Apache Spark; Business Intelligence and Analytics*

I. INTRODUCTION

Data is a subset of Big Data. Today, irrespective of a the specific sector, data is growing expeditiously every day and converting into Big Data. Big data required a big storage, and to process a big data and covert into useful information there is a need for strong business intelligence and Analytics system. Cloud computing is a an great option to store big data instead of to store it in tradition on-premise data center storage, where it's not possible to grow it exponentially and vertically. Once stored, a strong big data processing like Hadoop is required along with related eco system software to process data offline as well as online [2]. There would be a need to the intermediate system before its ready to display on business intelligence and analytics dashboard like ETL (extract, transformation, and load) due to multiple sources of data and in multiple formats. Finally, after the filtering process, it ready to display on a centralized management the dashboard.

II. PROBLEM STATEMENT

Customer data is the foundation for every organization. It gives a great insight of business and processes to define long-term strategies. It helps to understand where business is lacking and what all areas need improvements. Where we are doing good, where we are not doing good, what went wrong and how we can prevent it in future, so that same process gaps should not be repeated. It was a big challenge since long to get a consolidated view of historical data due to manual process and unavailability of technologies. Today, almost all organizations are maintaining digital data, and numerous technologies allowing to get a great insight of business patterns. Today, one of the biggest challenges is data is growing in multiple folds daily and increasing day by day in the same patterns. Terabyte to a petabyte of data is generating each day. It is stored in multiple staging areas, such as data warehouse and used to prepare a centralized management dashboard to analyze the different business patterns [3]. Apart from such offline processing, an uncountable number of areas required real-time processing, and it's another big challenge with such big data is to process such data at real time. For example, an aircraft generates a terabyte of data each day, which is collected via different sensors deployed in aircraft and processed at real-time to monitor the health if each part and equipment. Any delay in the real-time processing of such big and critical data can be an invitation to a big disaster. We are still lacking in the field of offline and real-time processing of big data and to integrate it with a centralized business intelligence and analytics system to leverage the complete benefit of technologies.

III. A GLANCE ON CLOUD COMPUTING, IOT, BUSINESS INTELLIGENCE, AND ANALYTICS

We are living in the golden era of the digital world, where human brains are developing the technologies by taking the horizon of the next few years. Cloud computing concept has been evolved from telecom industry virtual private network concept around in 1990's. initially, it was moving around between specific source and destinations, but later considering the growth and future of technologies in human life, the scope of cloud computing expanded to public network (internet) to provide services to end users [4]. Now, as of today, from small to large scale of business sectors utilizing the digital services due to cheap rental service model of cloud computing. Later, organizations have started to understand of historical data of their customer transactions to grow their business. They started retaining the data of the past few years to understand the business trends and customer like and dislikes. That was the time, when big data technology was discovered for processing of such big data. Cloud computing is the best model to store historical data due to uncostly storage. Later, upper management started facing a challenge to have some technology, which can display consolidated business trends from historical data on a single screen or dashboard [6]. This was the time when Business Intelligence and Analytics started capturing the market from different software vendors. Today SAS, Tableau, Qlik, and Pentaho are the leaders in the field of Business Intelligence and Analytics. Recently the concept of IoT (Internet of Things) has been evolved, with the concept of

connecting machine with the machine, and machine with humans. It's all about connecting all thing together where the output of one source could be the input of another target source. No doubt, it has opened unlimited opportunities to make the life of humans better, but there are lots of challenges still persist, which doesn't allow it to be adopted by all organizations and sectors. Again, the big challenge with IoT is, it is one of the biggest sources of big data, where the unlimited type of sensors generating gigabyte to a terabyte of data in few seconds to the minute[7]. I have presented an integrated framework of Business Intelligence & Analytics and Big Data for efficient utilization of all capabilities both technology and related eco system.

IV. HIGH-LEVEL FRAMEWORK

High-level framework depicted below, where small data collected from different sources converted into big data, and there is a need to store it in a big data processing engine like Hadoop. HDFS work as a distributed storage and MapReduce as processing engine for stored data. It produced the results on end user's data dashboard. Hadoop is mainly for the batch processing, hence generate the reports from historical data whereas the real-time data processed from the same big data set and visualized using some Hadoop eco system utilities. A high-level framework is shown below [9].

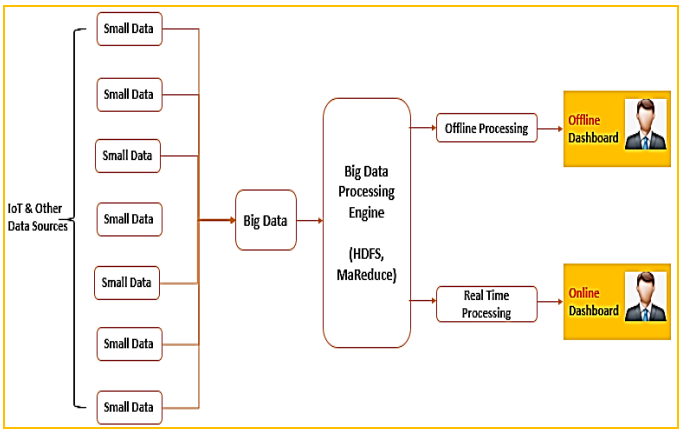

Fig 1. Multiple Small Data Sources

V. IMPLEMENTATION METHODOLOGY

There could be multiple data sources with structured, semi-structured and unstructured data. Large structured is also a challenge, but processing and converting unstructured data into useful information and visualizations is still a big challenge [8]. Due to unstructured nature of data, there have lots of algorithms were developed, and still, multiple data scientist struggling to convert organization unstructured data into meaningful and desired format.

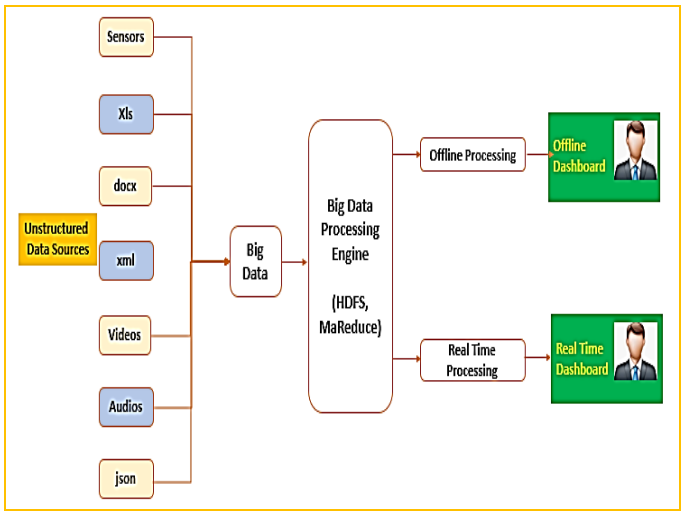

Fig 2. Different Structured Data Sources

Apache Spark is the best option in comparison with Hadoop MapReduce to process real data at real-time as it's an In-memory data processing engine [10]. It's not an alternate of Hadoop, but just a processing engine which can be used in place of MapReduce. When it was tested and compared then it was found almost 100x faster than MapReduce.

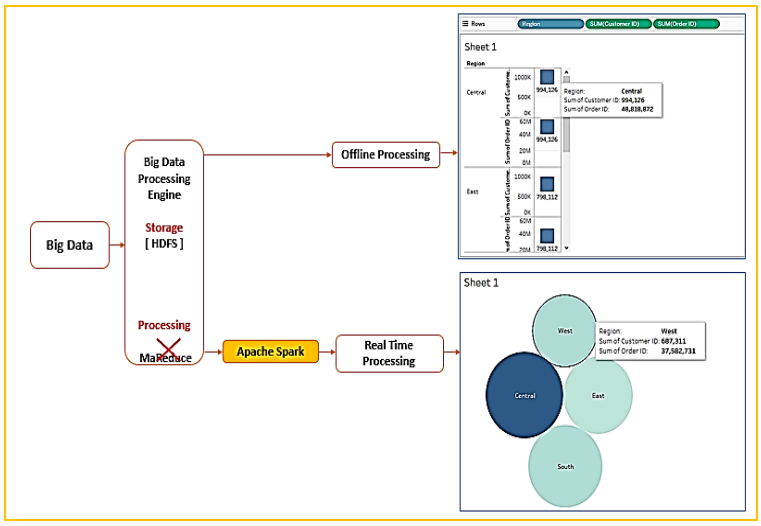

Fig 3. Apache Spark instead of MapReduce for Real-Time Data Processing

VI. IMPLEMENTATIONS AND RESULTS ANALYSIS

One of the U.S super market data is used for the simulation as big data and analysis have been done using Tableau tool. It contains the sales data of a data store with maximum details related to each transaction [11]. There were almost 24 fields in the dataset and different reports were generated by combining different fields.

U.S Super Store Data

Row ID	Order Priority	Discount	Unit Price	Shipping Cost	Customer ID	Customer Name	Ship Mode	Customer Segment
20847	High	0.01	2.84	0.93	3	Bonnie Potter	Express Air	Corporate
20228	Not Specified	0.02	500.98	26	5	Ronnie Proctor	Delivery Truck	Home Office
21776	Critical	0.06	9.48	7.29	11	Marcus Dunlap	Regular Air	Home Office
24844	Medium	0.09	78.69	19.99	14	Gwendolyn F Tyson	Regular Air	Small Business
24846	Medium	0.08	3.28	2.31	14	Gwendolyn F Tyson	Regular Air	Small Business
24847	Medium	0.05	3.28	4.2	14	Gwendolyn F Tyson	Regular Air	Small Business
24848	Medium	0.05	3.58	1.63	14	Gwendolyn F Tyson	Regular Air	Small Business
18181	Critical	0	4.42	4.99	15	Timothy Reese	Regular Air	Small Business
20925	Medium	0.01	35.94	6.66	15	Timothy Reese	Regular Air	Small Business
26267	High	0.04	2.98	1.58	16	Sarah Ramsey	Regular Air	Small Business
26268	High	0.05	115.99	2.5	16	Sarah Ramsey	Regular Air	Small Business
23890	High	0.05	26.48	6.93	18	Laurie Hanna	Regular Air	Small Business
24063	Not Specified	0.07	12.99	9.44	19	Jim Rodgers	Regular Air	Small Business
5890	High	0.05	26.48	6.93	21	Tony Wilkins Winters	Regular Air	Small Business

Fig 4. U.S Super Store Data part 1

Product Category	Product Sub-Category	Product Container	Product Name
Office Supplies	Pens & Art Supplies	Wrap Bag	SANFORD Liquid Accent™ Tank-Style Highlighters
Furniture	Chairs & Chairmats	Jumbo Drum	Global Troy™ Executive Leather Low-Back Tilter
Furniture	Office Furnishings	Small Pack	DAX Two-Tone Rosewood/Black Document Frame, Desktop, 5 x 7
Furniture	Office Furnishings	Small Box	Howard Miller 12-3/4 Diameter Accuwave DS™ Wall Clock
Office Supplies	Pens & Art Supplies	Wrap Bag	Newell 321
Office Supplies	Pens & Art Supplies	Wrap Bag	Newell 351
Office Supplies	Rubber Bands	Wrap Bag	OIC Colored Binder Clips, Assorted Sizes
Office Supplies	Envelopes	Small Box	Grip Seal Envelopes
Office Supplies	Envelopes	Small Box	Tyvek ® Top-Opening Peel & Seel ® Envelopes, Gray
Office Supplies	Rubber Bands	Wrap Bag	Staples Gold Paper Clips
Technology	Telephones and Communication	Small Box	StarTAC 7797
Furniture	Office Furnishings	Small Box	DAX Natural Wood-Tone Poster Frame
Technology	Office Machines	Medium Box	Hewlett Packard 6S Scientific Calculator
Furniture	Office Furnishings	Small Box	DAX Natural Wood-Tone Poster Frame

Fig 5. U.S Super Store Data part 2

Product Base Margin	Country	Region	State or Province	City	Postal Code	Order Date	Ship Date	Profit	Quantity ordered new	Sales	Order ID
0.54	United States	West	Washington	Anacortes	98221	1/7/2015	1/8/2015	4.56	4	13.01	88522
0.6	United States	West	California	San Gabriel	91776	6/13/2015	6/15/2015	4390.367	12	6362.85	90193
0.45	United States	East	New Jersey	Roselle	7203	2/15/2015	2/17/2015	-53.8096	22	211.15	90192
0.43	United States	Central	Minnesota	Prior Lake	55372	5/12/2015	5/14/2015	803.4705	16	1164.45	86838
0.56	United States	Central	Minnesota	Prior Lake	55372	5/12/2015	5/13/2015	-24.03	7	22.23	86838
0.56	United States	Central	Minnesota	Prior Lake	55372	5/12/2015	5/13/2015	-37.03	4	13.99	86838
0.36	United States	Central	Minnesota	Prior Lake	55372	5/12/2015	5/13/2015	-0.71	4	14.26	86838
0.38	United States	East	New York	Smithtown	11787	4/8/2015	4/9/2015	-59.82	7	33.47	86837
0.4	United States	East	New York	Smithtown	11787	5/28/2015	5/28/2015	261.8757	10	379.53	86839
0.39	United States	East	New York	Syracuse	13210	2/12/2015	2/15/2015	2.63	6	18.8	86836
0.55	United States	East	New York	Syracuse	13210	2/12/2015	2/14/2015	652.7331	10	945.99	86836
0.49	United States	West	Montana	Helena	59601	5/15/2015	5/16/2015	314.4813	17	455.77	90081
0.39	United States	West	Montana	Missoula	59801	5/21/2015	5/23/2015	-114.6399	18	231.79	90082
0.49	United States	East	New York	New York City	10012	5/15/2015	5/16/2015	384.38	70	1876.69	41793

Fig 6. U.S Super Store Data part 3

Test results are explained below. Few fields have been pulled from the data set to generate offline and real-time data processing. The region, and corresponding cities were selected along with respect count of customer ID's and purchased orders as shown in the figure below.

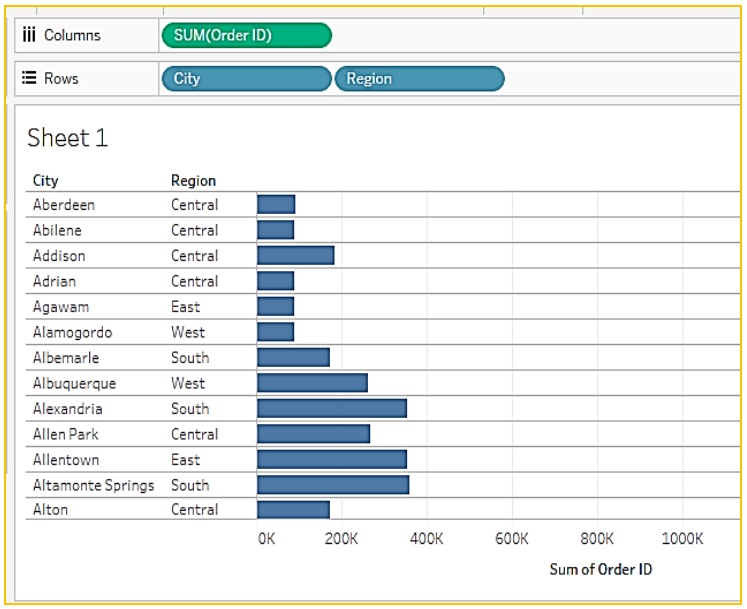

Fig 7. Region and city wise customer and order count details.

The simulation was done to test real-time data processing as shown in the figure below. Initially, the number of customers in the west region was 687311 and sum of order id was 37, 582, 731.

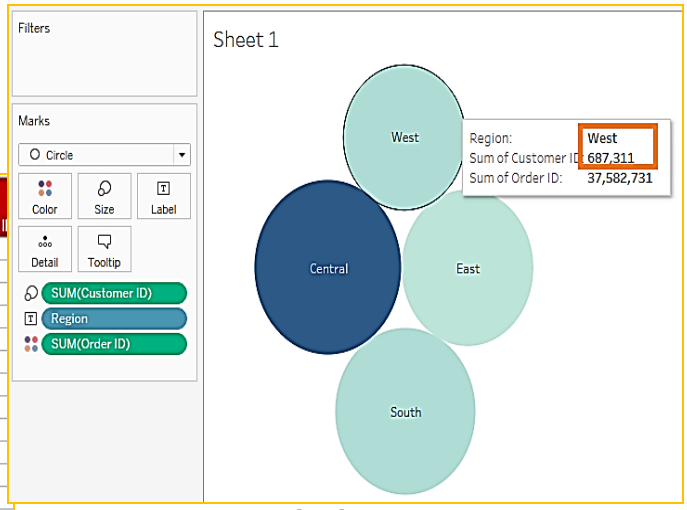

Fig 8. Customer and order count in wet region 1

As soon as order started increasing, the count was starting reflecting on the dashboard as shown in the figure below. Initial count was 687, 311 and after some time it was 694, 405 and the sum of order id was 38, 476, 898, and it was reflected on the dashboard after few seconds.

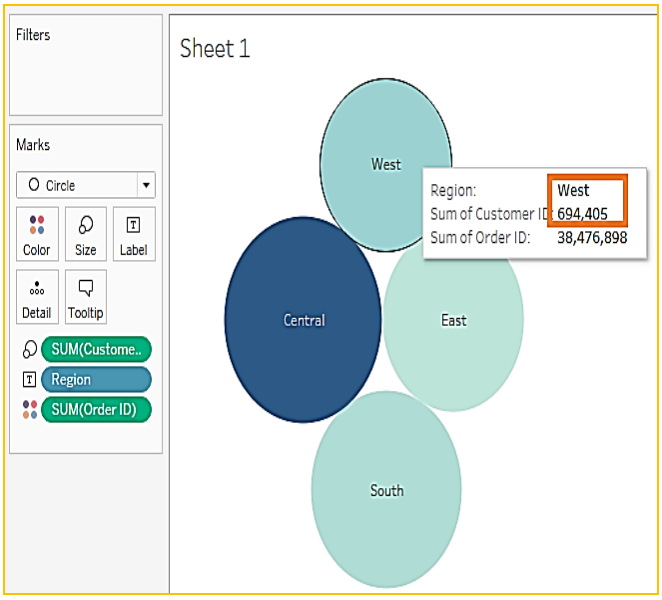

Fig 9. Customer and order count in wet region 2

Similar simulation tested for the central region, initially, the sum of orders was 48, 818, 872.

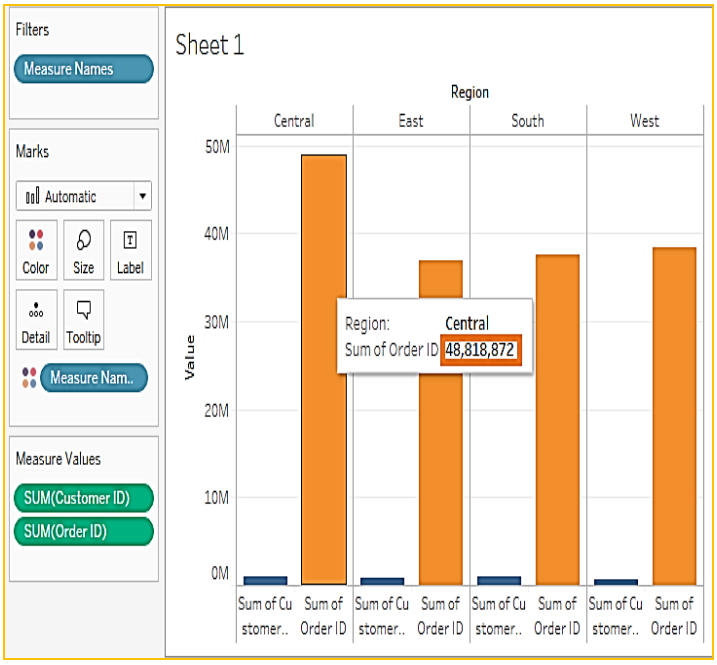

Fig 10. Customer and order count in central region 1

And after simulation of a few more order in the same region, it was 50, 227, 523, and it was reflected on the dashboard after a few seconds.

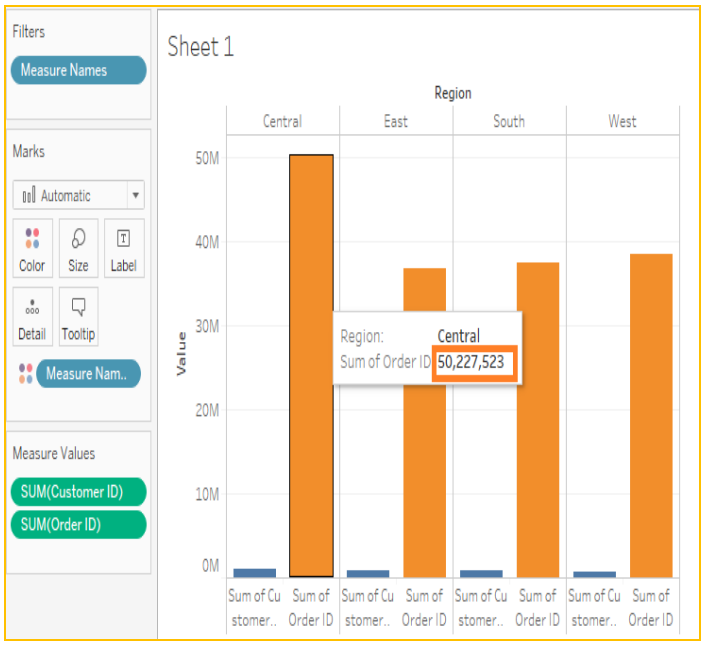

Fig 11. Customer and order count in central region 1

VII. CONCLUSION

Although, Hadoop and Apache spark is sufficient as of today for batch and real-time processing of data [13], but according to Forbes, by 2025 there would be more than 75 billion connected devices in the world. It's just one field, IoT. Apart from that, there wouldn't be any count on the terabyte to a petabyte of data generated in seconds to milliseconds. So there would a need of some advance and enhance technique in future to sustain current big data procession environment and technologies [14].

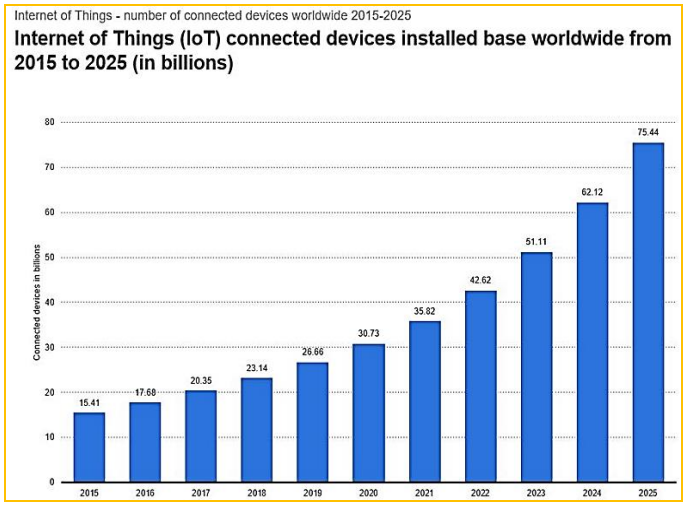

Fig 12. Forbes prediction on connected IoT devices

After few simulation testing and analysis, it has been observed that, for offline and real-time data processing Hadoop, apache spark and some related technologies are sufficient as of today, but in near future when data will start growing with unexpected rate, these technologies will not be able to sustain.

There would be need of an alternate advance solution, which can make them sustain. However, it's just an envisage model as of now where I am proposing some intermediate high performance storage to work as a Caching Engine, as shown in the figure below to reduce the processing load on big data and related eco system technologies.

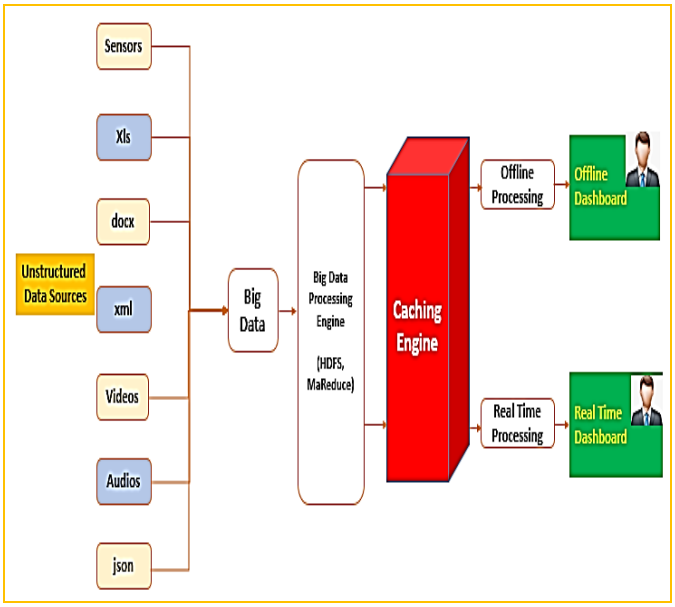

Fig 13. Proposed Model

VIII. REFERENCES

[1] Zappi, P., Lombriser, C., Benini, L., & Tröster, G. (2010). Collecting datasets from ambient intelligence environments. International Journal of Ambient Computing and Intelligence, 2(2), 42–56.

[2] Bureš, V. (2016). Application of ambient intelligence in educational institutions: Visions and architectures. International Journal of Ambient Computing and Intelligence (IJACI), 7(1), 94–120.

[3] Building Smarter Manufacturing With The Internet of Things (IoT), Lopez Research LLC 2269 Chestnut Street #202 San Francisco, CA 94123 T(866) 849–5750W.

[4] Huang, D., Zhou, Z., Xu, L., Xing, T., Zhong, Y.: Secure data processing framework for mobilecloud computing. In: Proceeding IEEE INFOCOM Workshop on Cloud Computing, INFOCOM 2011, Shanghai, China (2011)

[5] hen, D., Zhao, H.: Data security and privacy protection issues in cloud computing. In: IEEE International Conference on Computer Science and Electronics Engineering (2012)

[6] Munir, K.: Security model for cloud database as a service (DBaaS). In: IEEE Proceedings of the International Conference on Cloud Computing Technologies and Applications - CLOUDTECH 2015, pp. 1–5 (2015). ISBN 978-1-4673-8148-2

[7] AlZain, M.A., Li, A.S., Soh, B., Pardede, E.: Multi-cloud data management using Shamir's secret sharing and quantum byzantine agreement schemes. Int. J. Cloud Appl. Comput. **5**(3), 35–52 (2015)

[8]T.H. DavenportBusiness Intelligence and Organizational Decisions
International Journal of Business Intelligence Research, 1 (1) (2010), pp. 1-12

[9] M.Z. Elbashir, P.A. Collier, M.J. DavernMeasuring the Effects of Business Intelligence Systems: The Relationship between Business Process and Organizational Performance
International Journal of Accounting Information Systems, 9 (3) (2008), pp. 135-153

[10] B. Hocevar, J. JaklicAssessing Benefits of Business Intelligence Systems - a Case Study
Management, 15 (1) (2010), pp. 87-120

[11] D. Sommer, B. SoodMarket Share Analysis: Business Intelligence and Analytics Software 2013,Gartner Research Report (2014)

[12] B.H. Wixom, H.J. Watson, A. Reynolds, J. HofferContinental Airlines Continues to Soar with Business Intelligence Information Systems Management, 25 (2) (2008), pp. 102-112

[13] C.H. Hsinchun, R.H.L. Chiang, V.C. StoreyBusiness Intelligence and Analytics: From Big Data to Big Impact
MISQ, 36 (4) (2012), pp. 1165-1188

[14] M.Z. Elbashir, P.C. Collier, S.G. Sutton, M.J. Davern, S.A. LeechEnhancing the Business Value of Business Intelligence: The Role of Shared Knowledge and Assimilation
Journal of Information Systems, 27 (2) (2013), pp. 87-105

[15] Yeoh, W., Critical Success Factors for Implementation of Business Intelligence Systems in Engineering Asset Management Organizations. 2008

[16] H. Chen, R.H.L. Chiang, V.C. StoreyBusiness Intelligence and Analytics: From Big Data to Big Impact
Mis Quarterly, 36 (4) (2012), pp. 1165-1188

[17]S. Chaudhuri, U. Dayal, V. NarasayyaAn overview of business intelligence technology
Communications of the ACM, 54 (8) (2011), pp. 88-98

[18] C.F. Cheung, F.L. LiA quantitative correlation coefficient mining method for business intelligence in small and medium enterprises of trading business, Expert Systems with Applications, 39 (7) (2012), pp. 6279-6291

Dr. Mukesh Negi is an Oracle, IBM, ITIL & Prince2 Certified Engineer with 16+ years of rich experience in multiple Advance and Emerging IT Technologies such as DBMS & Big Data, Cloud Computing, Virtualization, Internet of Things, Artificial Intelligence, Machine Learning, Business Intelligence & Analytics, IT Security etc. He is currently working at TechMahindra Ltd (Part of Mahindra & Mahindra Group), an Indian multinational provider of Information Technology, Networking Technology solutions and Business Process Outsourcing to various industry verticals and horizontals as a Sr. Technical Delivery Manager. In the Education field, He is serving as an Editorial Board Member of many international journals. He has conducted several Faculty Development Programs and serving as a Guest & Visiting Faculty in many reputed University and Colleges in India.