Applied GIS

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

Analysis of the various industries in the Pune area and their use of big data processing technologies

Mr. Brijesh Y.Joshi.[1]
Research Scholar IICMR, Pradhikaran, Pune
Dr. Poornashankar[2]
Research Guide
Professor, Indira College of Engineering & Management, Pune

ABSTRACT

Although there are a number of data processing tools at our disposal, data has evolved in fundamental ways. Big data refers to the rapid expansion of stored information that has been noticed in businesses over the last several years. Data production and processing prior to a given year often included structured data. Social media platforms, new types of mobile devices, and other factors have all contributed to a shift in the data's fundamental character. Data sets considered "big" might be organized, semi-structured, unstructured, or a hybrid of these. There is a large volume of data being generated, and it must be processed. However, semi-structured and unstructured data processing are beyond the scope of typical data processing technology. Technologies for processing large amounts of data include Apache Hadoop and its many distributions; Cloudera; SAS; R; NoSQL databases; MongoDB; Amazon Elastic MapReduce; Kalfa; and many more.

This research looks into the impact of data processing on the decision-making of industry, focusing on the companies that operate in the Pune area. Topics covered include the types of companies that generate data, the types of data those companies process, the combinations of the two, the big data processing technologies that are currently available, the availability of skilled personnel to operate those technologies, and more.

Keywords: Big Data, Hadoop, Cloudera, SAS, R, NoSQL (not just SQL), MongoDB, Amazon Elastic MapReduce (EMR), and Kalfa.

Introduction

Today's world is data centric .We can just guess that big data will get only massive. Organization those which are dealing with remarkable amounts of heavy data, know that at present such data is a big challenge.

This can be termed as the massive Information explosion. Daily over a billion bytes of data is generated. This number is increasing as a day passes. On any given day we generate 294 billion or more average number of emails. Internet users are more than 2.2 billion throughout the world. There are more than 555 million of websites. Facebook users are growing day by day. Every minute a 48 hours of videos are uploaded on youtube. More than 465 accounts are existing on twitter. These are some examples which indicates that how the data is growing.[2]

This massively growing data can be termed as Big Data . Big data can be defined as data sets whose size is beyond the capability of generally used software tools which are used to capture , manage and process data.

Big has characteristics such as Data Volume, Data Velocity, Data Variety, Data Veracity.

Data Volume : Big data size can be termed as large. But what do you mean by Large. Large stands for the data is in terms of petabytes, zeta bytes and still increasing.

Data Velocity : Data is of time-sensitive. It can be categorized in terms of Batch ,Near Real Time, Real Time Streams.

Data Veracity : It stands for reality or actuality. The big data comes in bad, good, Undefined, inconsistency, incomplete ,ambiguous forms.

Data Variety : Big data is of different types such as structured ,semi structured and unstructured data.Means data can be text, audio, video, click streams log files etc. Big data can be of following types :

Unstructured Data : It can be defined as that data which does not have predefined model or organized in predefined manner.

Structured Data :It can be defined as that data which comes with predefined model.

Semi-Structured Data: It can be defines as that data which uses such a data structure which has no clear distinction between data and schema.

Big Data is having its impact every where such as social media, marketing, government, individual

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

. According to Manoj Singh chief Operating Officer, Deloitte Touche Tohmatsu by 2020 in internet 40 billion new connected devices will be connected to internet .By 2015 big data will create 4. 4 jobs globally .Big data market is expected to reach $23.8 billion by 2016.

According to NASSCOM in Big data market in India will grow $1B in 2015. It is estimated that global big data opportunity is estimated to grow at 45% annually to reach $25B by 2015, from the at present size about $8B.[2]

EMC Corporation published survey report which was carried out among 10,700 decision makers throughout 50 countries . Big data is increasing the decision making capabilities. It is observed that 79% of respondents says that better use of big data will lead to the better decision making.

Many companies are focusing more on big data and analytics because they are seeing positive results from trial projects as well as anecdotal evidence from industry colleagues or competitors. Perhaps most importantly, senior executives are realizing that good data can yield good decisions, if captured, analyzed, communicated, and acted upon in a timely and efficient fashion.

Nearly half of survey respondents (49 percent) assert that the greatest benefit of using data analytics is that it is a key factor in better decision-making capabilities. The use of data in decision- making has been driven in part by economic necessity[3]

Peer Research, Big Data Analytics survey performed on 200 IT professionals recently in 2013 in United States draws conclusion that Big Data is getting the top priority in organizations. Big data
is speeding up the decision making regarding new strategic initiatives and moving at big data analytics projects. Almost all processing structured and unstructured data .Stakeholder making request of big data analytics strongly understands big data. Big data is used to generate the competitive intelligence and also help in determination of staffing levels. IT managers in US are developing strategies regarding Batch and Real-Time processing. Various organizations are using Apache Hadoop framework for big data processing purpose. Big data processing challenges are facing such as data security, data storage and analytics. It is also observed that there is shortage of skilled data science professionals.[6]

Gartner carried a worldwide survey and published on March 12, 2013 regarding Trends in Big Data. From this survey it was found that 42% percent of IT leaders have invested in Big Data or plan to do so within a year. Doug Laney, research vice president at Gartner "Most organizations are still in the early stages, and few have thought through an enterprise approach or realized the profound impact that big data will have on their infrastructure, organizations and industries." Organizations are found to be in early stages of adoption of Big data processing. As organizations has understood the big data initiatives are critical because of necessity and conviction, also they are not able to meet the business opportunities with traditional data sources and practices
,organizations are stepping into Big data processing Gartner predicts that by 2015, 20 percent of Global 1000 organizations will have established a strategic focus on "information infrastructure" equal to that of application management.[7]

Gartner Report published in May 13, 2013 says that India IT infrastructure spending will reach
$2.3 Billion by 2014. “Despite global economic challenges, India provides strong growth opportunities across segments including infrastructure. Infrastructure alone is expected to surpass
$2.9 billion in 2017,” said Naveen Mishra, research director at Gartner.” Big Data, cloud, social and mobility are real time business drivers.[8]

According to research performed by MGI and McKinsey & Company’s Business Technology Office ,by 2018 the United State alone can could face a shortage of 140,000 to 190,000 people with the deep analytical skills and 15 Million managers and analysts with required knowledge to use big data analysis for decision-making.

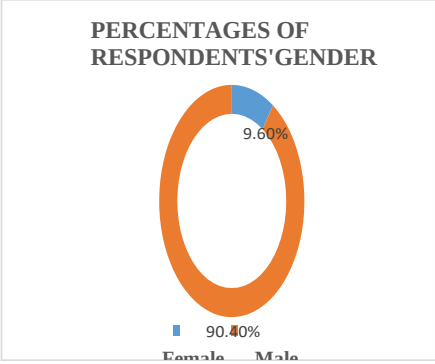
According to A.C. Kearney IT innovation study, more than 45 percent of companies have implemented business intelligence or big data initiative in past two years. Further studies estimate more than 90 % of fortune 500 companies will have at least one big data initiative underway withinyear.[13]

Each and every company interviewed mentioned that ,in 2017 big data adoption attained 53% which was 17% in 2015 .These companies were in telecom and financial services. According to this survey top use cases for big data were Internet of Things, Customer/Social Analysis,

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

Predictive maintenance, Clickstream Analysis , Fraud detection .Big data uses by vertical industries involve Financial services ,telecommunications, Technology, Education ,Health Care
. Those companies mentioned that technologies used for big data analysis encompasses Cloudera, Hortonworks, MAP/R, etc. Frameworks used for big data analysis are Spark
,MapReduce and YARN.[20]
In 2017 NewVantage carried out survey in which it was observed that 95% of executives mentioned that their firms from last five years invested in big data technology. [22]
It was mentioned that in this report that in 2017 many experts will increase their use of Hadoop and NOSQL technologies. Also it is mentioned according to reference of Forrester that usage of Hadoop is increasing 32.9% per year.[23]
In this report it is mentioned that in 2018 , 73% of business processes their big data in cloud , this number is up by 58 % as compared to year 2017. It is also mentioned that Apache Spark has gain in all big data processing technologies. In this report projection of Hadoop and big data market growth projection mentioned for the year 2017($17.1 B to $ 99.31B) and mentioned that in 2022 it will attain to 28.5% GAGR. [24].
Objectives of the Study

i)To study significant Big data processing done in



Pune region. ii)To study use of big data processing in Decision making.

Research Methodology

To make this study more reliable primary data is collected through a survey carried out through self-administered questionnaire. The primary data required for the study has been collected from the respondents working in data processing companies existing in Pune region .The sample size of 300 has been taken from population working the IT Industry. Random sampling technique has been used. Questions related to big data were asked and data so collected has been arranged in a tabular form. Keeping in view the objectives of the study, statistical tools are used for analysis andreach the conclusion.
Analysis and Finding
The study targeted 300 respondents from data processing companies which are intensive in big data processing. From findings of the statistical study of data gathered from respondents, it can be said that all the respondents who filled the questionnaire are from intensive data processing industry, were considered to be aware of big data and relevant things.

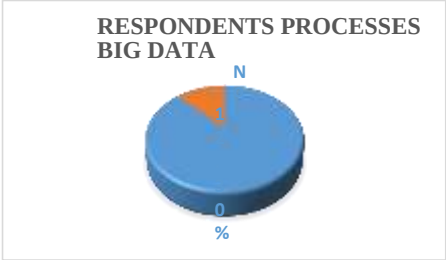


Chart I

Organizations Processes Big Data	Percentages
Yes	87.70%
No	12.30%

Table 1
Chart I and Table 1 represents the 87.7% respondents' processes big data and 12.3% respondents does not process big data. It means 236 respondents processes big data and 37 respondents process big data. It is observed that from the selected population maximum respondents are involved in big data processing.
Chart No II

Sr No	Gender of Respondents	Percentages Of Respondents
1	Female	9.60%
2	Male	90.40%

Table 2
Chart No II and Table 2 represents that 90.40% respondents were male while 9.60% female. It is observed that from the population maximum respondents are male.

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

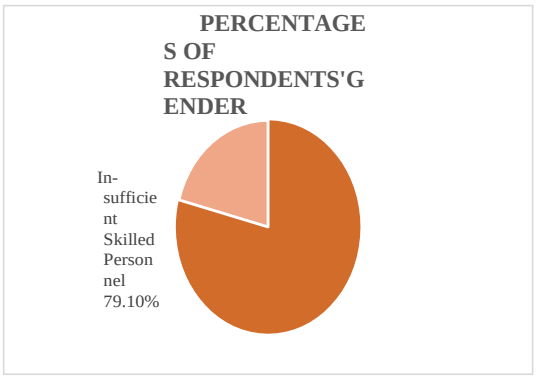


Chart No III

Sr No		Percentages
1	Sufficient Skilled Personnel	79.10%
2	In-sufficient Skilled Personnel	20.90%

Table 3

Chart No III and Table 3 represents that 79.10% of respondents responded that their organization has sufficient skilled personnel, while (20.90%) has responded that their organization has insufficient skilled personnel to deal with big data processing.

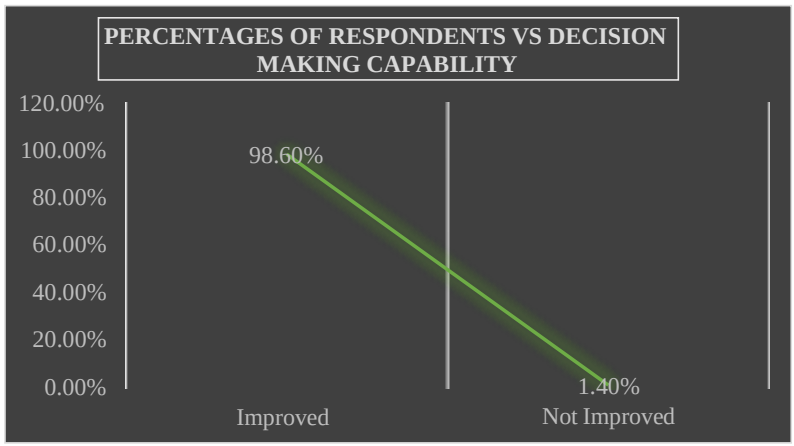


Chart No IV

Decision Making	Percentages of respondents
Improved	98.60%
Not Improved	1.40%

Table 4

Chart No IV and Table 4 represents that 98.6% respondents responded that big data processing improved decision making in their organization while 1.40% respondents responded that it has not improved decision making in their organization.

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

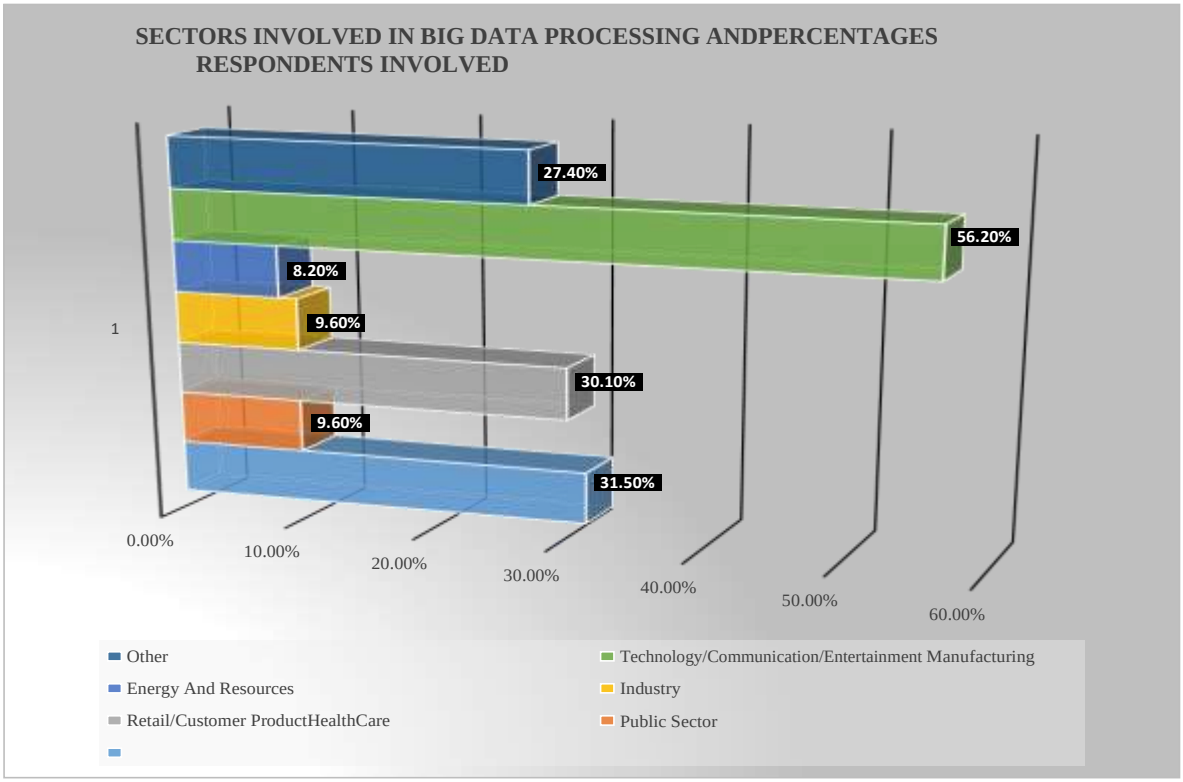


Chart No V

Sector Vs Percentages of respondents		
Sr No	Sector	Percentages of Respondents (%)
1	HealthCare	31.50%
2	Public Sector	9.60%
3	Retail/Customer Product	30.10%
4	Manufacturing Industry	9.60%
5	Energy And Resources	8.20%
6	Technology/Communication/Entertainment	56.20%
7	Other	27.40%

Table 5

Chart No V and Table 5 represents industry and respondents processing data for those industries. 31.50% respondents are for Healthcare, 30.10% respondents are for Retail/Customer Products, 56.20% respondents are for Technology/ Communications / Entertainment, 8.20% respondents are for Energy and Resources 9.60% respondents are for Manufacturing and 5% respondents for OtherIndustries.

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

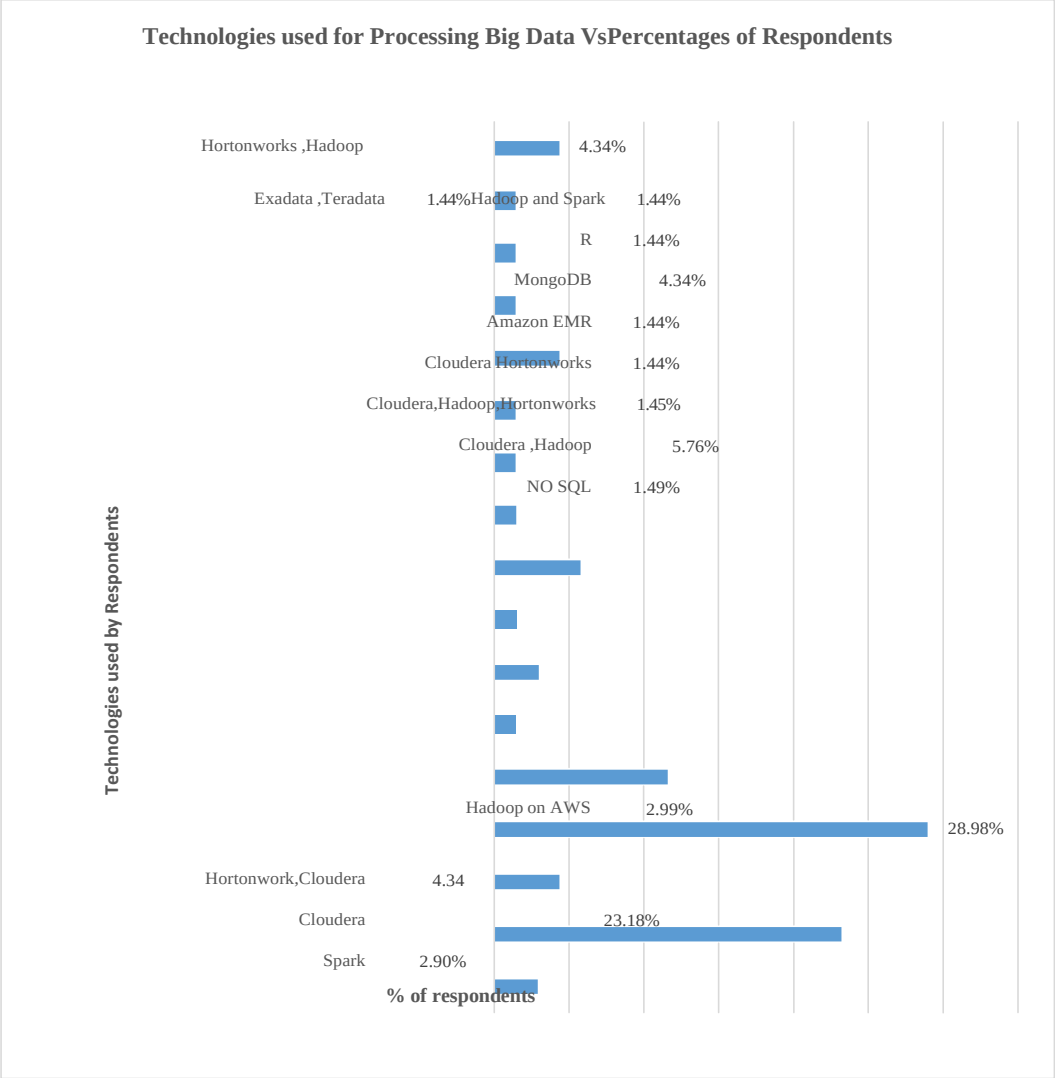


Chart No VI

Sr No	Technology	Percentages
1	Spark	2.90%
2	Cloudera	23.18%
3	Horton work, Cloudera	4.34%
4	Hadoop	28.98%
5	Hortonworks	11.59%
6	SAS,R	1.44%
7	Hadoop on AWS	2.99%
8	NO SQL	1.49%
9	Cloudera ,Hadoop	5.76%
10	Cloudera,Hadoop,Hortonworks	1.45%
11	Cloudera Hortonworks	1.44%
12	Amazon EMR	1.44%
13	MongoDB	4.34%
14	R	1.44%
15	Hadoop and Spark	1.44%
16	Exadata ,Teradata	1.44%
17	Hortonworks ,Hadoop	4.34%

Table 6

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

Chart No VI and Table 6 represents various Big data processing technologies used by respondents. Hadoop is used by 28.98% respondents , Cloudera is used by 23.18%,Hortonworks is used by 11.59%, Cloudera and Hadoop 5.76%, MongoDB 4.34% , Hortonworks and Cloudera 4.34%,Hadoop on AWS 2.99%, Spark 2.90%,NOSQL is used by 1.49%, SAS and R used by 1.44%, (Cloudera , Hadoop , Hortonworks) used by 1.45%,Amazon EMR used by 1.44 % ,R used by 1.44 %, Hadoop and Spark used by 1.44 % ,Exadata used by 1.44 %,Teradata used by 1.44%.

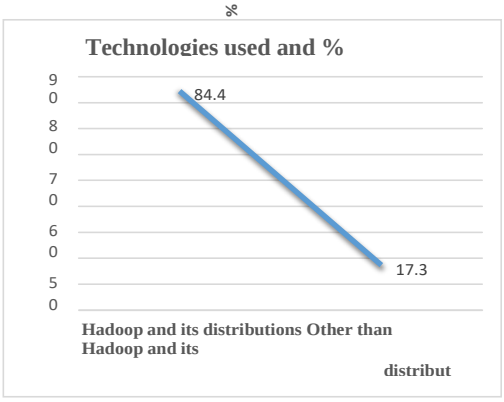


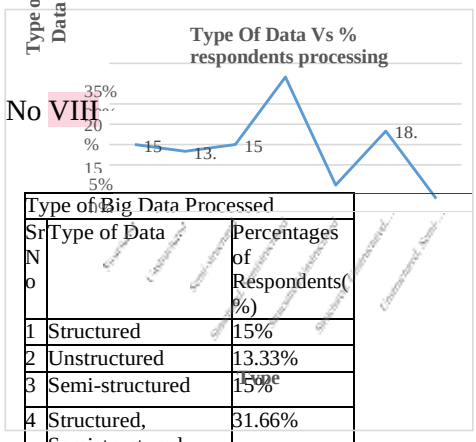
Chart No VII

Technologies Used	Percentages of Respondents
Hadoop and its distributions	84.47%
Other than Hadoop	17.33%

Table 7

Chart No VII and Table 7 shows that Hadoop and its distributions are used by 84.47% respondents while other than hadoop and its distributions are used by 17.33%. It is observed that Hadoop and its distributions are used widely to process big data.

Chart No VIII



Type of Big Data Processed		
Sr No	Type of Data	Percentages of Respondents(%)
1	Structured	15%
2	Unstructured	13.33%
3	Semi-structured	15%
4	Structured, Semistructured	31.66%
5	Strucutred,Unstructured	5%
6	Structured, Unstructured, Semi-structured	18.33%
7	Unstructured, Semi-structured	1.66%

Table Bo 8

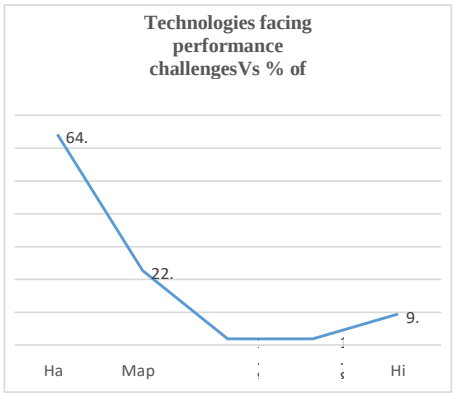


Chart No VIII and Table No 8 represents that structured data is processed by 15% of respondents, Unstructured data is processed by 13.33% of respondents, Structured, Semi-structured data is processed by 31.66% of respondents, structured, unstructured data is processed by 5% of respondents, Structured, Unstructured, Semi-structured data is processed by 18.33% of respondents , Unstructured, Semi-structured data is processed by 1.66% of respondents.

Chart No IX

Sr No	Technology	Percentages
1	Hadoop	64.15%
2	MapReduce	22.64%
3	Hbase	1.88%
4	Kalfa	1.88%
5	Hive	9.43%

Table No 9

Chart No IX represents that Hadoop Ecosystem is facing maximum performance challenges. In Hadoop ecosystem MapReduce i.e processing component is facing major performance challenges. Hive and Hbase components are

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

facing performance challenges but comparatively less.

Conclusion

The purpose of our research was to learn more about the practices of data-heavy businesses in the Pune area. In addition, these businesses rely heavily on big data, which needs careful deliberation, the right people, and cutting-edge technology to handle. Several industries, including healthcare, government, retail, manufacturing, energy, communications, and entertainment, and others, all do Big Data processing in the Pune area. Based on the statistics, it can be concluded that the TCE sector generates and processes big data at a higher rate than the healthcare and retail/customer product sectors. Structured, semi-structured, and unstructured forms of big data exist. More work is done using organized and semi-structured data. Companies handle a wide range of big data permutations. Hadoop technology and its distributions are frequently utilized for such data processing because to their versatility and efficiency. Hadoop is available in several flavors, including those from Apache, Cloudera, and Hortonworks. Organizational decision making has been boosted by the analysis of big data. Because some businesses are already suffering from a lack of qualified workers, it is clear that more will be needed in the future, especially if big data processing technologies are to be applied.
Long-Term Impact

Research might be directed toward learning more about Hadoop, a popular Big data processing tool. Questions such, "How many businesses will use big data processing in the future?" might be answered through study. Where would you expect them to originate? What challenges do businesses confront when they try to put big data ideas into action? When do we expect to see a greater need for qualified workers? Many more details are also possible.

References

Global Services, 22 February 2013, http://www.globalservicesmedia.com/global-services/analysis/175085/big-data-making-big-impact, Global Services, Nasscom India Leadership Forum 2013.

Global Services Media, 22 February 2013, http://www.globalservicesmedia.com/global-services/analysis/22074/nasscom-big-data-market-india-reach-usd1b, NASSCOM: Big Data Market in India to Reach $1 Billion, Conference. Sharma, Smriti.

The EMC Corporation Conducted a Worldwide Survey on December 12, 2013 (available online at http://india.emc.com/about/news/press/2013/20131212-01.htm).

According to a recent report, 60% of Indian businesses want to increase their IT spending on big data analytics in the next year. More than 60% of Indian businesses see big data as the most important area for IT investment.

PwC Luxembourg published Big Data Analytics & Decision Making by Laurent Probst, Erica Monfardini, Laurent Frideres, Steven Clarke, Dawit Demetri, Lina Schnabel, and Alain Kauffmann in September 2013.
Intel's 2013 IT Manager Survey Explores How Businesses Are Leveraging Big Data, August 2013.

42% of IT executives, according to a Gartner survey, have either invested in Big Data or plan to do so over the next 12 months. Monday, March 11, 2013 STAMFORD, Conn.

According to Gartner, India will spend $2.3 billion on its IT infrastructure this year. Mumbai, India, in 2014
May 13, 2013
You Hadoop, right? The State of Big Data Experts,Graham, Bradley M. R. Rangaswami. 29 Oct 2013.

Open Integration Solutions, Talend, 2012. How Big Is Big Data Adoption? [11]Website: www.atkearney.com, "IT Innovation Drives Revitalized Growth."

According to Smart Planet on February 21, 2012, "Big Data Market Set to Explode This Year, But What is Big Data?"

ISSN: 1832-5505
Vol-05 Issue-01 Jan 2017

The Big Data Revolution and the Abrupt Death of the Conventional Business Model, The 2013 report by A.T. Kearney, C. Hagen, M. Ciobo, D. Wall, A. Yadav, K. Khan, J. Millor, and H. Evans

ZDNet, 2 March 2012, MapReduce vs MPP: Opposite Sides of Big Data.

Wipro Council for Industry Research and the Wipro Outsourcing Center: Reaping the Benefits of Big Data

, WIPRO.

About Half of Businesses Are Either Already Doing or Planning to Do Big Data Projects, Press Release, StrongNewsletter.com, 03 February 2014.

White paper on the future of data connection in 2014, available at www.progress.com.

Executive Survey on Big Data, NewVantage, September 2013, wp-content/uploads/2013/09/Big-Data-Executive-Survey.pdf.

Survey-2013-Executive-Summary.pdf
Date: February 8, 2017, https://www.zionmarketresearch.com/report/hadoop-market

Datamation, Big Data Trends, 24 January 2018, https://www.datamation.com/big-data/big-data-trends.html

Click here for the executive summary: https://newvantage.com/wp-content/uploads/2017/01/Big-Data-Executive-Survey-2017.pdf

http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-

Findings-Updated-010219-1.pdf

https://globenewswire.com/news-release/2018/09/11/1569178/0/en/Qubole-Announces-2018- Big-Data-Trends-and-Challenges-Report.html (September 11, 2018) Qubole has announced the publication of its 2018 Big Data Trends and Challenges Report.