RESEARCH ARTICLE OPEN ACCESS

A Survey on Big Data

[1]NIHARIKA SAHU,
Gandhi Institute of Excellent Technocrats, Bhubaneswar, India

[2]SUBHASMITA BEHERA,
Black Diamond College of Engineering & Technology, Jharsuguda, Odisha, India

ABSTRACT –Big data is a term used to describe data or data sets that are so massive or complicated that distributed databases are required instead of standard data processing software. Big data has been the foundation of companies like Google, eBay, LinkedIn, and Facebook from the start. It consists of a collection of enormous and intricate data sets, including vast amounts of data, social media analytics, data management tools, real-time data, etc. The design of sensors, data collection, data duration, sharing, storage, analysis, visualization, and information privacy are among the difficulties. Big data refers to datasets that are rapidly growing and have a large degree of variability, making them challenging to manage with conventional tools andmethods.Big data analytics is the study of enormous amounts of data to uncover hidden correlations. Big Data is a type of data that is so complex that managing it and gleaning value and untapped knowledge from it calls for new management strategies, algorithms, and analytics. In order to structure Big Data and address the issue of making it relevant for analytics, a different platform, called Hadoop, is required.

Key Words: Big Data, Parallel programming, MapReducetechnique.

I. INTRODUCTION

Every digital process and social media exchange producesBigdata.TheSystems,sensorsandmobiledevicestransmit. The arrival of big data is from multiple sourcesatafrighteningvelocity,volumeandvariety.Weneedoptimal processing power, analytics capabilities and skillsto extract meaningful value from big data. More confidentdecisionmakingcandowithaccuratebigdata.Gooddecisionsleadtogreateroperationalefficiency,costreduction and reduced risk. Analysis of data sets can findnewcorrelations,to"spotbusinesstrends,preventdiseases,andcombatcrimeandsoon"[1].Scientists,businessexecutives,practitionersofmedia,andadvertisingandgovernmentsalikeregularlymeetdifficulties with large data sets in areas including Internetsearch,financeandbusinessinformatics.Datasetsgrow in sizeinpartbecausetheyareincreasinglybeinggathered bycheapandnumerousinformation-sensingmobiledevices, aerial (remote sensing), software logs, cameras,microphones,radio-frequencyidentification(RFID)readers,andwirelesssensornetworks[2][3][4].Bigdata"size" is a constantly moving target, ranging from a fewdozen terabytes to many petabyte of data.(1 petabyte is1000terabytes)

Fig-1:Animage ofBigdata

Herearesomereal-worldexamplesofBigData inaction:

- Consumerproductcompaniesandretailorganizationsaremonitoringsocialmedialike Face book and Twitter to get an unprecedentedviewintocustomerbehavior,preferences,andproduct perception.
- Manufacturersareabletomonitorminutevibrationdatafromtheirequipment,whichchanges slightly as it wears down, to predict theoptimal time to replace or maintain. Replacing ittoo soon wastes money and replacing it too latetriggersan expensivework stoppage.
- Manufacturers are also monitoring social networks,but with a different goal than marketers: They areusing it to detect aftermarket support issues before awarrantyfailurebecomespubliclydetrimental.
- Thegovernmentismakingdatapublicatthenationallevel, statelevel, and citylevelfor userstodevelopnewapplicationsthatcangeneratepublicbetter.
- FinancialServices organizations are taking dataminedfromcustomerinteractionstosliceanddicetheirusersintofinelytunedsegmentsand enablesthesefinancialinstitutionstocreateincreasinglyrelevantandsophisticatedoffers.
- Advertising and marketing agencies are trackingsocial media to Insurance companies are using BigDataanalysistoseewhichhomeinsuranceapplicationscan be immediatelyprocessed, andwhichonesneed avalidatingin-personvisit.
- Retailorganizationsareengagingbrandadvocates,changingtheperceptionofbrandantagonists,andevenenablingenthusiasticcustomers to sell their products. All these thingsaredoingbyembracingsocialmedia.
- Hospitals predict those patients that are likely toseekreadmissionwithinafewmonthsofdischarge by analyzing medical data and patientrecords.Thehospitalcanthenpreventinganothercostlyhospital stay.
- Tooffermoreappealingrecommendationsandmoresuccessfulcouponprogramsthe,Web-basedbusinessesaredevelopinginformationproductstthatcombinedatagatheredfromcustomers
- Sportsteamsareusingdatafortrackingticketsalesandareusingbigdatafortrackingteamstrategiesalso.

a. Three Vs of big data: volume, velocity andvariety[5].(Big Data Parameters)

Fig-2:AnimageofBigdata

Volume. Volume of data stored in enterprise repositorieshavegrownfromgigabytestopetabytes.Manyfactorscontribute to the increase in data volume like transaction-based data stored throughthe years, unstructureddatastreaminginfromsocialmediaetc.Hugeamountsofsensorandmachine-to-machinedatabeing collected. Inthe past days, excessive data volume was a storage issue.Butwithdecreasingstorage costs,other issuesemerge,includinghowtodeterminerelevancewithinlargedata volumesandhowtouseanalyticstocreatevaluefromrelevantdata. Volumereferredasamount ofdata.

Velocity. Data is streaming in at extraordinary speed andmust be dealt with in a timely manner. RFIDsensors andsmartmeteringaredrivingtheneedtodealwithfast-movingof data in near-real time. It is a challenge for mostorganizations to reacting quickly enough to deal with datavelocity. Velocityreferredthe speed of dataprocessing.For time-sensitive processes such as catching fraud, bigdatamustbeused.Itstreamsintoyourenterpriseinorderto maximizeitsvalue.

Variety. Today data comes in different types of formats.Structuredandnumericdataintraditionaldatabases.Informationcreatedfromline-of-

businessapplications.Unstructuredtextdocuments,email,video,audioandfinancial transactions. Managing, merging and governingdifferentvarietiesofdataaresomethingmanyorganizationsstillstrugglewith.Differenttypesandsourcesofdataarethere.Datavarietyexplodedfromstructuredandlegacydatastoredinenterprisestoragestounstructured, semistructured, audio, videoetc.

We consider two additional dimensions when thinkingabout bigdata:

Variability. With the increasing velocities and varieties ofdata, data flows can be highly inconsistent with periodicpeaks. It istrending in social media.Everyday seasonalandevent-triggeredpeakdataloadscannotbeabletomanage.Even moreunstructureddatainvolved.Theinconsistencythe datacanshowattimes—-whichcanhamper the process of handling andmanaging the dataproperly.Theinconsistencythedatacanshowattimescanhampertheprocessofhandlingandmanagingthe dataproperly.

Complexity. Today's data comes from different types ofsources.itisstillanundertakingtolink,matchandtransform data across systems. Anyway, it is necessary toconnectandcorrelaterelationships,hierarchiesand multiple data linkages. Otherwise your data can quicklyspiral out of control. When large volumes ofdata comefrommultiplesources,thedatamanagementisverycomplex.Especially data must be linked, connected, andcorrelatedtheuserscangrasptheinformationormessagesthe dataissupposedtoconvey.

VeracityThe qualityofcaptureddata,whichvarysohigh.TheAccurateanalysisofdatadependsontheveracityofsource data.

II. PARALLEL PROGRAMMING & MAP REDUCE

Data analysis software parallelizes fairly naturally. Manyprogrammers are interested to building programs on theparallel model. The parallel research had the most successin the field of parallel databases. Rather than requiring theprogrammer to unknot an algorithm into separate threadsto be run on separate cores, parallel databases let thembreak up the input data tables into pieces, and pump eachpiece through the same single-machine program on eachprocessor. This "parallel dataflow" model makes parallelprogramming as easy as programming a single machine.And it works on "shared-nothing" clusters of computers ina data center: The machines involved can communicate viasimple streams of data messages, without a need for anexpensiveshared RAMor diskinfrastructure.[6]

Famous big data analysis tool is Hadoop.Apache Hadoopis an open-source software framework.it iswritteninJava for distributed storage and distributed processing ofbigdataoncomputerclustersbuiltfromcommodityhardware. All the modules in Hadoop are designed with afundamentalassumptionthathardwarefailures(ofindividualmachinesorracksofmachines)arecommonplace and thus should be automatically handled insoftwarebythe framework.[7]

TheheartofHadoopis**MapReduce**.Itisthisprogramming paradigm that allows for massive scalabilityacrossthousandsofserversinaHadoopcluster.Itisuseful for batch processing on petabytes or zeta bytes ofdatastoredinApacheHadoop.Ifwearefamiliarwithclusteredscale-outdataprocessingsolutions.ThentheMapReduce concept is simple to understand. MapReduceprogrammingmodelhastwistedanewpageintheparallelism story. The MapReduce framework is a paralleldataflowsystemthatworksbydividingdataacrossmachines. Each of which runs the same single-node logic.MapReduce asks programmers to write traditional code, inlanguages like C, Java, Python and Perl. In addition to itsfamiliar syntax, MapReduce allows programs to be writtento and read from traditional files in a file system, ratherthanrequiringdatabaseschemadefinitions.

MapReducerefers to two separate and distinct tasks.Thefirstisthejobofmap,whichtakesasetofdataandconvertsitinto another setof data.Individualelementsare broken down into value pairs. The reduce job takes theoutputfromamapasinputandcombinesthosedatavaluesintoasmallersetofvalues.Thereducejobisalways performed after the map job.So the sequence ofthenameMapReduce.

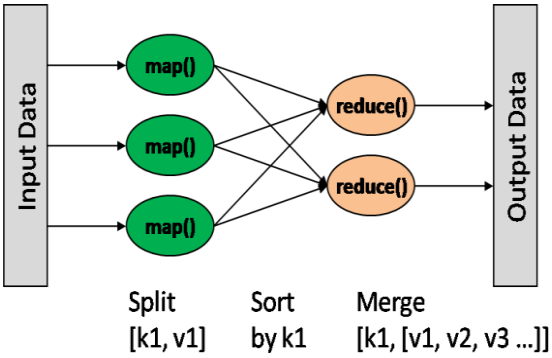

Fig-2:MapReduce

III. BEST BIG DATA ANALYTICS USE CASES

SentimentAnalysis

Sentiment analysis offers powerful business intelligence toenhance the customer experience, revitalize a brand, andgaincompetitiveadvantage.Thekeytosuccessfuls entimentanalysisliesintheabilitytodigformulti-structured data pulled from differentsources into a singledatabase.

360-DegreeViewofCustomer

A360-degreecustomerviewoffersadeeperunderstandingofcustomerbehaviorandmotivations.Obtaining a 360-degree customer review requires analysisofdatafromdifferentsourceslikesocialmedia, datacollecting sensors, mobile devices etc. From there, moreeffective micro-segmentation and real-time marketing aregettingasresult.

AdHoc DataAnalysis

Ad-hocanalysisonlylooksatthedatarequestedorneeded, providing another layer of analysis for data setsthat are becoming larger and more varied. Big data ad-hocanalytics can help in the effort to gain greater insight intocustomersbyanalyzingtherelevantdatafromunstructuredsources, both externaland internal.

Real-TimeAnalytics

Systems that offer real-time analytics quickly decipher andanalyze data sets, providing results even as data is beinggeneratedandcollected.Thishigh-velocitymethodofanalyticscanleadtoimmediatereactionandchanges.It allowsforbettersentimentanalysis,splittesting,andimproved targetedmarketing.

Multi-ChannelMarketing

Multi-channelmarketingcreatesaseamlessacrossdifferenttypesofmedialikecompanywebsites,socialmedia, and physical stores. During all stages of the buyingprocessmulti-channelmarketingrequiresanintegratedbig dataapproach.

CustomerMicro-Segmentation

Customer micro-segmentation provides more tailored andtargeted messaging for smaller groups. This personalizedapproach requires analysis of big data collected throughsources like customers' online interactions, social mediaetc.

AdFrauddetection

Adfrauddetectionrequiresdataanalysisoffraudstrategiesbyrecognizingpatternsandbehaviors.Datathat shows irregularity of group behavior make it so adfraudis find outandblockedbeforeit isspread.

Clickstreamanalysis

Click stream analysis helps to grow the user experience byoptimizing company websites, and offering better insightintocustomersegments.clickstreamanalysishelpstopersonalizethebuyingexperience,gettinganimprovedreturnoncustomer visits withbig data.

DataWarehouseModernization

Integratebigdataanddatawarehousecapabilitiestoboostoperationalefficiency.Optimizeyourdatawarehouse to enable fresh types of analysis. Use big datatechnologies to set up a staging area or landing zone foryournewdatabeforeformativewhatdatashouldbemovedtothedatawarehouse.divestinfrequentlyaccesseddorageddatafromwarehouseandapplicationdatabasesusinginsequenceintegrationsoftwareandtools.

BigDataandPredictiveModeling

The most common uses of big data by companies are fortrackingbusinessprocessesandoutcomes,andforbuilding awidearray of predictivemodels.
Amazon and Netflix recommendations rely on predictivemodelsofwhatbookormovie an individualmightwantto
purchase. Google's search results and news feed rely onalgorithms that predict the significance of particular webpages or articles. Apple's auto-complete function tries toforecasttherestofone'stextoremailbasedonpastconventionpatterns.Onlineadvertisingandmarketingrely greatlyonautomatedpredictivemodelsthataimindividuals who might be particularly likely to answer tooffers.

Theapplicationofpredictivealgorithmsextendswellaheadoftheonlineworld.Inhealthcare,itisnowcommonforinsurerstoadjustpaymentsandqualitymeasures based on "risk scores," which are resulting frompredictivemodelsofhumanbeinghealthexpensesandoutcomes.Anindividual'sriskscoreisnaturallyaweighted sum of health indicators that recognize whetheran individual has different persistent conditions, with theweights chosen based on a statistical analysis. Credit cardcompaniesusepredictivemodelsofdefaultandrepaymenttoguidetheirunderwriting,pricing,andmarketingactions.

IV. BIGDATA CHALLENGES

Heterogeneity, scale, timeliness, complexity, and privacyproblems with Big Data impede progress at all phases ofthe pipeline that can create value from data. The problemsstart right away during data acquirement, when the datatsunami requires us to make decisions, currently in an adhocmanner,aboutwhatdatatomaintainandwhattoreject, and how to store what we keep unfailingly with theright metadata. a great deal data today is not
nativelyinstructuredformat;forexample,tweetsandblogsareweakly ordered pieces of text, while images and video arestructured for storage and display. But not for semanticcontentandlookfor.Transformingsuchcontentintoastructuredformatforlateranalysisisamaintest.
Thevalueofdataexplodeswhenitcanbeassociatedwithotherdata.Thusdataintegrationisamajorcreatorofvalue.Themajoritydataisdirectlygeneratedindigitalformat today; we have the opportunity and the challengeboth to influence the creation to facilitate later linkage andto automatically link before created data. Data analysis,organization,recovery,andmodelingareotherfoundational challenges. Data analysis is a clear bottleneckinalotofapplications,bothduetobesmallofscalabilityof the original algorithms and due to the complexity of thedata that needs to be analyzed. Lastly, presentation of theresultsanditsclarificationbynon-technicaldomainexpertsisvital to extracting actionable
Knowledge

Volumeofdata

The volume of data, especially machine-generated data, isexploding, how fast that data is growing every year, withnew sources of data that are emerging. For instance, in theyear 2000, 800,000 petabytes (PB) of data were stored inthe world. According to IBM it is anticipated to reach 35zettabytes (ZB) by 2020. Social media plays a key role ieTwittergenerates7+terabytes(TB)ofdataeveryday. Face book, 10 TB. Mobile devices alsoplay an importantrole.

Bigdataskillsareinshortsupply

There's already a shortage of data scientists in the market.This includes a shortage of people who know how to laborwellwithlargevolumesofdataandbigdatasets.Companies need the right merge of people to help makesenseofthedatastreamsthatarecomingintotheirorganizations. This includes skills for applying propheticanalyticstobigdata,askillsetthatevenmostdatascientistsbeshortof.

V. CONCLUSIONS

TheavailabilityofBigData,low-costcommodityhardware,andanalyticsoftwarehasshapedauniquemoment in the history of data analysis. The union of thesetrendsmeansthatwehavethecapabilitiesrequiredtoanalyze amazing data sets quickly and cost-effectively forthe first time in history. All these capabilities are neitherthetheoretical nor trivial. They represent a real leap forwardand a clear chance to realize enormous gains in terms ofefficiency,productivity, income,andprofitability.

Requirements for dealing out that may seem
unbelievabletodaywillsoonberoutinewhenbigdatasystemsareavailable. We learn how to exploit them.Not very manyyearsago,systemsthescaleofFacebookandGooglewould have seemed like science fiction.At that time 100transactions per second for airline and banking
systemswasastretch.Severalnewrequirementswillalsocombine data from many sources, not all of which will becompany-

owned.Forinstance,somewillmakeuseof“opendata”fromgovernment.Lotsofopeningforinnovators!

REFERENCES

[1] "Data,dataeverywhere".TheEconomist.25February2010.Retrieved9December 2012.
[2] "Data, data everywhere". The Economist. 25 February2010.Retrieved9December 2012.
[3] "Community cleverness required". Nature **455** (7209): 1.4September2008.doi:10.1038/455001a.
[4] "Sandiaseesdatamanagementchallengesspiral".HPCProjects.4August2009.
[5] METAGroup."3DDataManagement:ControllingDataVolume, Velocity,and Variety."February2001.
[6] Hellerstein, Joe (9 November 2008). "ParallelProgrammingintheAgeofBigData".Gigaom Blog.
[7] "Welcome to Apache™ Hadoop®!".hadoop.apache.org.Retrieved 2015-09-20.