Emerging Trends on Big Data & Cloud Computing

Vishal Dutt[1], Pooja Sharma[2], Anshuman Kumar Gautam[3] and Meenakshi Deshwal[4]
[1,2]Department of Computer Science, MDS University, Ajmer, Rajasthan, India
vishaldutt53@gmail.com poojadixit565@gmail.com
[3] Faculty of Mathematics, Mount Literature Jee School, Naroli, Gujrat, India
er.anshuman2011@gmail.com
[4] Assistant Professor, JRE Group, NOIDA, India
meenakshi.smiley@gmail.com

Abstract
In this paper we focused on the emerging trends and various approaches for carrying out analytics on clouds for Big Data application. It revolves around four important analytics and Big Data. We also discussed about the some of the real world challenges in this cloud and Big Data computing era. This paper also focused on the implementation strategy of Big Data like Management, Data Varity etc. It helps to identify the technology gaps which may help to research communities so that they will have a directions for future scope of Big Data based on cloud computing.

DOI: https://doi.org/10.30991/IJMLNCE.2017v01i01.004

Keywords
Big data, Hadoop, SAP, Daas, Map Reduce.

1. Introduction
Society is finishing up logically more instrumented and alongside these strains, affiliations are conveying and securing massive measures of records. Supervising and getting bits of studying from the conveyed records is a test and key to excessive ground. examination publications of action that mine composed and unstructured facts are fundamental as they could empower relationship to get bits of gaining knowledge of from their subtly were given records, and in addition from a full-size measure of facts straightforwardly benefit fit at the internet [18]. The potential to go-relate private statistics on client slants and matters with records from tweets, on-line diaries, issue value determinations, and statistics from relational associations opens a huge assortment of capability results for courting to fathom the stipulations of their customers, envision their necessities and needs, and improve using advantages. This angle is in reality normally named as huge statistics. no matter the recognition on exam and massive records, setting them into preparing is up 'til now an unpredictable and dull enterprise. As Yu [10] factors out, large statistics offers liberal motivation to affiliations willing to get it, but in the meantime speaks to a bear in mind-fit variety of demanding situations for the affirmation of such protected regard. An affiliations inclined to use exam improvement constantly aerating and cooling quires exorbitant programming licenses; uses wide figuring establishment; and pays for advising hours of experts who work with the relationship to better apprehend its business, organize its information, and fuse it for examination [11,12]. Circulated registering has been changing the IT commercial enterprise via including versatility to the way it's far exhausted, allowing affiliations s to pay most effective for the benefits and corporations they use. With a real objective to lower IT capital and operational makes use of, affiliations s of all sizes are using Clouds to offer the benefits required to run their packages. Fogs vary thru and via in their particular headways and use, yet regularly deliver machine, level, and programming resources as companies [25,13].

The constantly ensured focal factors of Clouds fuse supplying re-resources in a pay as-you-go configuration, upgraded availability and adaptability, and price diminishment. Fogs can protect affiliations s from eating money for retaining up peak provisioned IT status quo that they are maximum likely now not going to apply as a preferred rule. at the same time as before everything look the offer of

Clouds as a section to complete examination is strong, there are numerous demanding situations that should be over-come to make Clouds a great level for adaptable examination.

In this paper we consider systems, situations, and improvements on zones that are crucial to massive statistics exam limits and speak how they assist building exam answers for Clouds. We focus at the maximum basic particular problems on engaging Cloud exam, but moreover spotlight a segment of the non-unique demanding situations seemed by affiliations s that need to offer exam as an business enterprise in the Cloud. also, we delineate a recreation plan of gaps and suggestions for the research bunch on future orientation on Cloud-maintained large statistics figuring.

2. Methodology

Affiliations s are dynamically handing over big volumes of records as behind schedule result of instrumented business shapes, checking of consumer hobby [14,15], website online following, sensors, back, accounting, among diverse reasons. With the happening to informal organization internet districts, clients make records in their lives by way of always posting unpretentious components of activities they perform, events they visit, places they go to, photos they take, and matters they acknowledge and require. This information typhoon is often counseled as large data [20,21,17]; a time period that passes at the issues it poses on present established order concerning restriction, agency, interoperability, business enterprise, and exam of the information.

In the gift targeted market, having the capability to explore facts to understand patron lead, parcel consumer base, offer revamp corporations, and get bits of learning from records gave by way of numerous resources is important to high ground. no matter the way that pioneers should want to develop their choices and sporting events regarding bits of studying were given from this facts [15], expertise facts, expelling non clean illustrations, and the use of these instances to expect future direct aren't new subjects. learning Discovery in facts (KDD) [20] plans to expel non apparent statistics the usage of attentive and point by means of point exam and clarification. information mining [1,2,3], extra in particular, plans to discover successfully cloud interrelations among surely unessential characteristics of instructive accumulations by way of making use of methodologies from a couple of domains inclusive of system learning, database systems, and bits of expertise. exam carries strategies of KDD, information mining, content mining, quantifiable and quantitative examination, illustrative and really apt fashions, and superior and instinctive portrayal to drive choices and exercises [5,7,9].

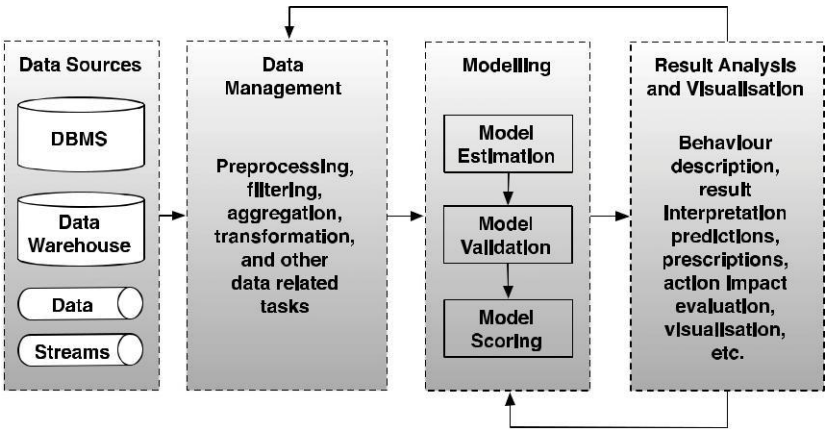

figure 1: Workflow of Big Data

The above determine says Portrays the everyday times of a popular exam paintings process for huge records. records from various sources, inclusive of databases, streams, stores, and facts dispersion focuses, are used to manufacture fashions. The large quantity and assorted kinds of the records can ask for pre-looking after errands for fusing the statistics, cleansing it, and filtering it. The organized data is used

to installation a version and to assess its parameters. once the version is assessed, it must be recommended earlier than its utilization. commonly this stage calls for the use of the principle information and precise methodologies to assist the made model. subsequently, the model is consumed and related to facts because it arrives. This stage, referred to as display scoring, is used to supply pre-word makes use of, arrangements, and suggestions. The effects are deciphered and surveyed, used to deliver new models or alter present ones, or are fused to pre-treated data.

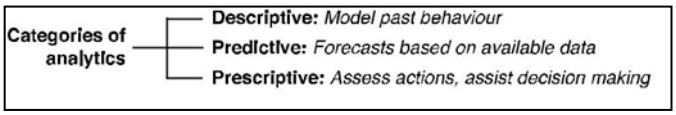

Figure 2: analytics types

Exam preparations can be named enlightening, prescient, or prescriptive as mentioned in Fig. 2. engaging exam makes use of chronicled information to recognize designs and make administration reviews; it's far involved about demonstrating beyond behavior. Prescient exam endeavours to anticipate the future through dissecting gift and recorded data. Prescriptive arrangements help experts in picks by way of finding out activities and evaluating their effect with appreciate to enterprise goals, necessities, and imperatives.

This work capabilities specialized troubles and studies existing work on answers for deliver examination competencies to huge statistics at the Cloud. thinking about the traditional research paintings manner brought in Fig. 1, we give attention to enter problems inside the durations of an examination arrangement. With large statistics it is obvious that a extensive quantity of the difficulties of Cloud examination concern facts management, joining, and dealing with. beyond paintings has targeting problems, as an example, information designs, data portrayal, stockpiling, access, safety, and information best. area three offers current work tending to those problems on Cloud conditions. In segment four, we increase on current fashions to provide and examine facts models at the Cloud. place 5 depicts answers for information illustration and consumer collaboration with exam preparations gave through a Cloud. We likewise feature a part of the commercial enterprise demanding situations postured by using this conveyance show whilst we communicate approximately administration systems, advantage level assertions, and plans of action. protection is definitely a key test for facilitating investigation arrangements on open Clouds. We keep in mind, anyways, that protection is a extensive point and could henceforth merit its very personal investigation. alongside those strains, protection and evaluation of information accuracy [20,15] are out of extent of this overview.

2.1. Data management

A champion some of the most monotonous and work focused assignments of examination is route of movement of statistics for examination; an difficulty robotically exacerbated with the aid of huge information as it broadens present system beyond what many might don't forget viable. performing examination on mammoth volumes of data requires profitable strategies to shop, channel, trade, and recoup the information. A section of the demanding situations of sending data organization courses of motion on Cloud circumstances had been regarded for a long time [1,14,16], and answers for perform exam on the Cloud defy near problems. Cloud exam guides of action want to recollect the distinct Cloud sending models grasped by tries, where Clouds may be for example:

i. Private: sent on a private system, controlled through the association itself or by using a pariah. A non-public Cloud is becoming for associations that require the most raised degree of control of protection and statistics coverage. In such situations, this kind of Cloud machine may be used to share the companies and facts extra effectively finished the extraordinary departments of a huge assignment.

ii. Public: despatched off-web site over the net and open to the overall populace. Open Cloud offers excessive functionality and conferred resources for insignificant exertion. The examination companies and information manassention are controlled by using the issuer and the concept of corporation (e.g. insurance, protection, and openness) is demonstrated in an expertise. Affiliations s can utilize these Clouds to do exam with a discounted cost or offer bits of studying of open exam occurs.

iii. Hybrid: combines the two Clouds where additional advantages from an open Cloud can be given of route to a non-public Cloud. clients can make and ship exam applications the use of a private circumstance, on this way getting rewards from adaptability and greater accelerated amount of safety than the use of simplest an open Cloud.

Thinking about the Cloud preparations, the accompanying conditions are by and big imagined with respect to the accessibility of facts and exam models [15, 16]: (i) information and fashions are non-public; (ii) statistics is open, fashions are private; (iii) information and models are open; and (iv) records is private, fashions are open. Jensen et al. [27] advise on employer models for Cloud research preparations that vary from arrangements utilising secretly facilitated programming and framework, to personal exam facilitated on a third get-together basis, to open version where the arrangements are facilitated on an open Cloud.

Now not the same as traditional Cloud administrations, examination manages extraordinary kingdom capacities that frequently request extremely precise re-resources, for instance, records and vicinity experts' research aptitudes. consequently, we advise that beneath sure plans of movement – specifically the ones wherein data and models live at the supplier's premises – preferred Cloud administrations, in addition to the abilities of information professionals have to be overseen. to perform economies of scale and flexibility, Cloud-empowered large statistics exam desires to research intends to assign and use these precise assets in a legitimate manner. something remains of this section examines existing preparations on statistics administration impartial of where records professionals are physically located, targeting capacity and recovery of records for examination; information respectable range, velocity and joining; and asset planning for information managing assignments.

2.2. Data variety and velocity

Considering the Cloud arrangements, the accompanying situations are by means of and massive imagined with admire to the accessibility of records and exam models [10,18,19]: (I) records and fashions are non-public; (ii) statistics is open, fashions are personal; (iii) facts and fashions are open; and (iv) facts is non-public, models are open. Jensen et al. [25,24] superb statistics is depicted by using what's habitually advised as a multi-V illustrate, as depicted in Fig. 3. variety addresses the statistics composes, speed insinuates the price at which the records is made and treated, and volume describes the degree of statistics. Veracity insinuates how plenty the statistics can be trusted given the unfaltering nature of its supply [4,5,6], however regard analyzes the financial really worth that an affiliation can get from the use of huge facts coping with. Al-however the desire of Vs used to clarify huge information is habitually subjective and modifications throughout finished reports and articles at the net – e.g. as of writ-ing Viability is remodeling into any other V – range, velocity, and extent [7,9,16] are the things most with the aid of and huge said.

As to, it could be watched that continually, incredible measure of records has been made unreservedly available for coherent and commercial enterprise jobs. instances join chronicles with authorities statistics1; obvious ecosystem statistics and fore-tosses; DNA sequencing; facts on development situations in first-rate metropolitan areas; element opinions and feedback; economics [19,21]; comments, pictures, and debts published on relational association internet areas; information collected the usage of nearby technological know-how plat-outon-lines [22]; and statistics assembled by way of a massive variety of sensors assessing distinctive organic situations, as an instance, temperature, air moisture, air satisfactory, and precipitation.

A case speakme to the prerequisite for any such grouping inner a lone examination utility is the Eco-Intelligence [13,14] arrange. Eco-Intelligence changed into deliberate to examine a great deal of facts to assist town masterminding and propel extra all the way down to earth change. The level desires to adequately find out and manner statistics from multiple resources, along with sensors, information, web

goals, television and radio, and attempt facts to empower urban accomplices to alter to the drastically components of urban headway. In a related situation, the cellular statistics task (MDC) was prompted away to provide improvements on mobile to phone based research, and to permit accumulating evaluation of adaptable facts exam frameworks [9,10,11,12]. information from round two hundred customers of cellular telephones was assembled over a 12 months as part of the Lausanne statistics series marketing campaign. every other associated place benefitting from examination is massively Multiplayer 8db290b6e1544acaffefb5f58daa9d83 recreation (MMOGs). CAMEO [18,19,30] is a designing for constant examination for MMOGs that usages Cloud resources for exam of errands. The building gives frameworks to statistics social affair and incessant exam on a couple of components, for instance, understanding the requirements of the diversion gathering.

The straggling leftovers of our dialogue on information agency for Cloud exam incorporates those two Vs of large records, mainly variety and velocity. We contemplate plans on how this contrasting information is secured, how it is able to be fused and how it's far tons of the time arranged. The alternate on portrayal in like manner researches speed by means of inclusive of segments, for example, perception and accumulating based observation. notwithstanding the manner that replacement Vs of large statistics are simple, we remember that some of them, as inspected in advance, justify their own certainly one of a type examination, as an example, records Veracity. diverse Vs are subjective; volume is outstandingly depending on the power of present tools gadget, which enhancements fast and might render an define obsolete rapidly; cost may also rely upon how successful an affiliation uses the examination guides of action inside reach. apart from the V homes, big statistics exam moreover stocks stresses with different statistics associated controls, and as a result can mainly advantage by using the collection of studying made within the trendy years on such settled topics. this is the circumstance of issues, as an instance, statistics satisfactory [11] and records provenance [12].

2.3. Data storage

Web scale record structures, for instance, the Google File System (GFS) [24] try to give the power, flexibility, and immovable quality that particular Internet organizations require. Diverse courses of action give question store limits where archives can be imitated over various land districts to upgrade abundance, flexibility, and data availability. Cases fuse Amazon Simple Storage Service (S3),3 Nirvanix Cloud Storage,4 OpenStack Swift5 and Windows Azure Binary Large Object (Blob) storage.6 Although these courses of action give the flexibility and overabundance that many Cloud applications require, they as a less than dependable rule don't meet the con-cash and execution needs of certain examination applications.

With respect to Big Data examination, MapReduce presents an intriguing model where data region is explored to upgrade the execution of uses. Hadoop, an open source MapReduce use, considers the making of gatherings that usage the Hadoop Distributed File System (HDFS) to package and copy educational accumulations to centers where they will most likely be eaten up by mappers. Despite abusing concurrence of colossal amounts of center points, HDFS limits the impact of disillusionments by reproducing educational accumulations to a configurable number of center points. It has been used by Thu soo et al. [12,16,18] to develop an examination stage to process Face-book's colossal enlightening records. The stage uses Scribe to add up to logs from Web servers and after that charges them to HDFS records and uses a Hive– Hadoop pack to execute examination businesses. The stage joins replication and weight strategies and columnar weight of Hive7 to store a great deal of data.

Another unmistakable example in Cloud preparing is the growing usage of NoSQL databases as the favored procedure for securing and recouping information. NoSQL gets a non-social model for data amassing. Leavitt fights that non-social models have been benefit skilled for more than 50 years in structures, for instance, challenge organized, different leveled, and outline databases, yet starting late this perspective started to pull in more thought with models, for instance, key-store, section arranged, and report based stores [20,19,17]. The establishments for such raise in excitement, as demonstrated by Levitt, are better execution, point of confinement of dealing with unstructured data, and propriety for dispersed conditions [18,20,21].

2.4. Data integration solutions

Forrester research appropriated a selected file that looks at a phase of the issues that conventional commercial enterprise Intelligence (BI) faces [15,16,17], highlighting that there's much of the time an overabundance of siloes data arranging, accumulating, and taking care of. Makers of the file envision more than one facts coping with and huge information exam limits being moved to the EDW, alongside those traces releasing affiliations s from pointless statistics alternate and replication and the usage of numerous statistics getting prepared and examination sport plans. what is extra, as discussed earlier, they maintained that exam sport plans will continuously display statistics looking after and examination functions via MapReduce and square– MR-like interfaces. SAP HANA One [15], for example, is an in-memory set up endorsed through Amazon web offerings that offers nonstop exam to SAP programs. HANA One in like way offers a SAP facts integrator to stack facts from HDFS and Hive-open databases.

2.5. Data processing and resource management

MapReduce [18,19,20,25] is a champion many of the most truely understood programming fashions to manner a great deal of data on organizations of desktops. Hadoop [10] is the maximum used open supply MapReduce use, in like manner made available with the aid of multiple Cloud companies [4,16,17,13]. Amazon EMR [4] engages customers to instantiate Hadoop gatherings to method a wonderful deal of records the use of the Amazon Elastic Compute Cloud (EC2) and different Amazon web services for information amassing and change.

Hadoop makes use of the HDFS document structure to divide impersonate enlightening accumulations over numerous middle points, with the actual goal that after strolling a MapReduce application, a mapper is presumably going to get to records this is secretly secured on the gathering center factor wherein it's miles executing. no matter the way that Hadoop megastar vides a sport plan of APIs that empowers architects to execute MapReduce programs, every from time to time a Hadoop work method is made from groups that use unusual state question vernaculars, as an example, Hive and Pig Latin, made to help interest and element of handling endeavors. Lee et al. [10,11] demonstrate a diagram approximately the features, points of hobby, and regulations of MapReduce for parallel records exam. They furthermore discuss ex-lines proposed for this programming version to conquer a number of its imprisonments.

2.6. Challenges in big data management

On this factor, we discuss movement discover targeting the difficulty of huge facts organisation for exam. There are nonetheless, regardless, many open problems in this problem. The precis under isn't a ways achieving, and as extra research in this subject is coordinated, additionally difficult issues will upward thrust.

2.6.1. Information assortment

A way to manage a continuously extending quantity of statistics? mainly whilst the facts is unstructured, the way to fast isolate essential substance out of it? a way to add as much as and relate spilling facts from diverse resources?

2.6.2. Data storage

The way to gainfully see and keep fundamental data remoted from unstructured facts? how to keep a long way accomplishing volumes of information in a way it can be propitious recuperated? Are available archive structures upgraded for the extent and variety asked for by means of exam programs? If now not, what new capacities are required? a way to store records in a manner that it could be without problems mi-floor/ported among server ranches/Cloud carriers?

2.6.3. Data integration

New traditions and interfaces for mix of data that could direct records of different nature (sorted out, unstructured, semi-composed) and sources.

2.6.4. Data Processing and Resource Management

New programming models refreshed for spilling and multidimensional records; new backend motors that manipulate upgraded document structures; motors arranged to enrol in applications from numerous programming fashions (e.g. mapreduce, paintings technique, and sack of-assignments) on a selected path of action/conference. a way to decorate asset utilize and centrality utilize while executing the exam software?

2.7. Model building and scoring

The data storing and information as a provider (daas) limits gave through clouds are critical, but for examination, it is also massive to apply the facts to manufacture fashions that can be applied for figures and prescribed drugs. likewise, as models are synthetic in mild of the open facts, they ought to be tried towards new facts on the way to compare their capacity to measure future lead. exist-ing paintings has discussed plans to offload such activities – named here as model building and scoring – to cloud vendors and ways to deal with parallelize sure machine gaining knowledge of counts [12,11,24]. this fragment depicts address the factor. table 1 abbreviates the tested paintings, its goals, and target establishments.

2.8. Visualization and user interaction

With the developing measures of facts with which examinations need to alter, marvelous perceptions gadgets are important. these devices need to con-sider the possibility of information and preamble to invigorate course [24,25]. the type of acknowledgment have to be picked through the measure of statistics to be confirmed up, to replace each displaying up and execution. perception can help the 3 important kinds of exam: unmistakable, wise, and prescriptive. exceptional acknowledgment mechanical gatherings don't delineate affected components of exam, but there has been a push to look at depiction to assist on sagacious and prescriptive exam, utilizing for example modern-day reports and portraying [26]. a key point to be considered on depiction and client dating in the cloud is that structure is 'inside the not too remote past a bottleneck in or 3 conditions [12,11,13]. clients in a super international need to need to envision statistics handled in the cloud having a vague illegal relationship and experience from however information were orchestrated domestically. two or 3 guides of action have been handling this want.

Present paintings additionally gives manner to customers to add as much as records from diverse resources and use diverse portrayal fashions, together with dashboards, gadgets, line and reference charts, economics, among numerous models [10,20,14,11,13]. some of these features can be used to play out a couple of errands, such as affect solutions; to tune what zones of a domain are acting brilliant and what sort of substance can enhance client experience; how information sharing on a informal association influences the website usage; track compact utilize [14,26,27]; and survey the effect of advancing endeavors.

3. Business models and non-technical challenges

In spite of giving units that clients can use to manufacture their huge information exam game plans at the cloud, fashions for passing on exam limits as agencies on a cloud have been dis-cussed in beyond paintings [1,3,7,5]. sun et al. [19] supply a graph of the prevailing excellent in magnificence on the headway of adjusted examination courses of action on clients' premises and make clear a part of the issues to engage examination and examination as an corporation on the cloud. part of the potential designs of interest proposed of their work consist of:

i. Facilitating patron examination occupations in a shared level: match-skilled for a wander or affiliations that has distinct examination offices. via and huge, these divisions want to expand their very own specific examination publications of motion and maintain up their personal specific gatherings. with a regular degree they could alternate their responses for execute on a not unusual establishment, along those traces reducing operation and renovation fees. as mentioned ahead of time, frameworks have been proposed for aid allocation and arranging of massive data examination errands at the cloud [12,10].

ii. A complete stack meant to provide customers with give up-to-end arrangements: fitting for institutions that do not have potential on exam. on this version, smart grasp institutions circulate zone particular indicative flow designs as companies. the company is chargeable for encouraging the object stack and dealing with the advantages essential to play out the examinations. customers who purchase in to the corporations basically need to exchange their statistics, plan the codecs, get models, and play out the most perfect model scoring.

3.1. Other challenges

N Designs of pastime in which abnormal country examination companies can be exceeded on by means of the cloud, human capacity can not be safely supplanted via gadget getting to know and large facts exam [23,24,25]; mainly occasions, there can be a necessity for human professionals to remain okay [21]. corporation ought to acclimate to huge facts instances and oversee issues, for example, the way to help human experts in grabbing bits of records and the way to explore structures that may assist boss in deciding on snappier decisions.

Statistics ingestion through cloud guides of action is constantly a weak factor, even as investigating and endorsement of made plans is a attempting and stupid procedure. as mentioned a while currently, the way examination is carried out on cloud levels takes after the gathering paintings circumstance: clients display a commercial enterprise and maintain up until the factor that the moment that endeavors are carried out and after that download the effects. once an exam is achieved, they down load check involves fruition that are ok to guide the exam errand and after that carry out assist exam. contemporary cloud situations do not have this natural technique, and frameworks need to be made to urge instinct and to fuse experts insider sharp through offering plans to reduce their hazard to records. structures and methodologies that iteratively refine solutions to request and provide customers greater control of having ready are wanted [6,8].

4.Conclusions

The degree of information starting at now added through the numerous sports of the general population has in no way been so massive, and is being made in a reliably developing fee. this large facts slant is being seen via ventures as a approach for obtaining gain over their fighters: if one enterprise can admire the facts contained within the facts sensibly quicker, it's going to have the capability to get extra costumers, boom the salary per consumer, decorate its operation, and reduce its charges. the whole lot considered, big data examination is up 'til now a trying out and time soliciting for assignment that requires luxurious programming, huge computational status quo, and attempt.

Dispersed processing allows in helping those issues via seasoned viding resources on-ask for with fees relating to the real usage. furthermore, it engages establishments to be scaled all finished swiftly, altering the structure to the sincere to goodness ask.

No matter the manner that cloud establishment gives such adaptable ability to supply computational assets on ask for, the district of cloud-supported exam is still in its underlying days. in this paper, we discussed the important thing durations of examination paintings bureaucracy, and focused the excellent in elegance of every level with respect to cloud-maintained exam. outlined paintings become portrayed in three key social events: records control (which consolidates information grouping, records storing, facts coordination game plans, and information making plans and useful resource employer), version constructing and scoring, and visualization and person interactions. for each of those domain names, continual paintings turned into poor down and key open troubles have been mentioned. this review completed up with an exam of plans of pastime for cloud-helped information examination and other non-precise issues.

5. References

[1] J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, C. Welton, MAD skills: new analysis practices for big data, Proceedings of the VLDB Endow 2 (2) (2009) 1481–1492.

[2] A. Cuzzocrea, I.-Y. Song, K.C. Davis, Analytics over large-scale multidimensional data: the big data revolution!, in: Proceedings of the ACM 14th inter-national workshop on Data Warehousing and OLAP, ACM, New York, NY, USA, 2011, pp. 101–104.

[3] DataDirect Cloud, http://cloud.datadirect.com/ (2013).

[4] T.H. Davenport, J.G. Harris, Competing on Analytics: The New Science of Winning, Harvard Business Review Press, 2007.

[4] T.H. Davenport, J.G. Harris, R. Morison, Analytics at Work: Smarter Decisions, Better Results, Harvard Business Review Press, 2010.

[5] J. Davey, F. Mansmann, J. Kohlhammer, D. Keim, The future internet, Springer-Verlag, Berlin, Heidelberg, 2012, Ch. Visual Analytics: Towards Intelligent Interactive Internet and Security Solutions, pp. 93–104.

[6] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM 51(1).

[7] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels, Dynamo: Amazon's Highly Available Key-Value Store, SIGOPS Operating Systems Review 41 (6) (2007) 205–220.

[8] E. Deelman, A. Chervenak, Data management challenges of data-intensive scientific workflows, in: Proceedings of the 8th IEEE International Sympo-sium on Cluster Computing and the Grid (CCGrid'08), IEEE Computer Society, 2008, pp. 687–692.

[9] P. Deepak, P.M. Deshpande, K. Murthy, Configurable and Extensible Multi-flows for Providing Analytics as a Service on the Cloud, in: Proceedings of the 2012 Annual SRII Global Conference (SRII 2012), 2012, pp. 1–10.

[10] P. Deyhim, Best practices for Amazon EMR, White paper, Amazon (2013). URL http://media.amazonwebservices.com/AWS_Amazon_EMR_Best_Practices. pdf.

[11] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34.

[12] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (4) (2013) 82–89.

[13] D. Fisher, R. DeLine, M. Czerwinski, S. Drucker, Interactions with Big Data Analytics, Interactions 19 (3) (2012) 50–59.

[14] D. Fisher, I. Popov, S.M. Drucker, M. Schraefel, Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), ACM, New York, USA, 2012, pp. 1673–1682.

[15] G.C. Fox, Large Scale Data Analytics on Clouds, in: Proceedings of the Fourth International Workshop on Cloud Data Management (CloudDB 2012), ACM, 2012, pp. 21–24.

[16] B. Franks, Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, first ed., in: Wiley and SAS Business Series, Wiley, 2012.

[17] FusionChars, http://www.fusioncharts.com/.

[18] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: Proceedings of the 9th ACM Symposium on Operating Systems Principles (SOSP 2003), ACM, New York, USA, 2003, pp. 29–43.

[19] Google App Engine, http://developers.google.com/appengine/.

[20] Google Prediction API, https://developers.google.com/prediction/.

[21] Google Analytics, http://www.google.com/analytics/.

[22] Gooddata, http://www.gooddata.com (2013).

[23] K. Goodhope, J. Koshy, J. Kreps, N. Narkhede, R. Park, J. Rao, V.Y. Ye, Building LinkedIn's Real-time Activity Data Pipeline, Bulletin of the Technical Committee on Data Engineering 35 (2) (2012) 33–45.

[24] R.L. Grossman, What is analytic infrastructure and why should you care? ACM SIGKDD Explorations Newsletter 11 (1) (2009) 5–9.

[25] A. Guazzelli, K. Stathatos, M. Zeller, Efficient Deployment of Predictive Analytics Through Open Standards and Cloud Computing, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 32–38.

[26] A. Guazzelli, M. Zeller, W.-C. Lin, G. Williams, PMML: An Open Standard for Sharing Models, The R Journal 1 (1) (2009) 60–65.

[27] Z. Guo, G. Fox, Improving MapReduce Performance in Heterogeneous Network Environments and Resource Utilization, in: Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2012), IEEE, 2012, pp. 714–716.