ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

BIG DATA ANALYTICS: 4A's

Prof. Sachchidanand Nimankar [1], Prof. Sushant Dagare[2]
[1]Assistant Professor, Mechanical Department, SSPM's College of Engineering, Kankavli, India
[2]Assistant Professor, Computer Science Department, SSPM's College of Engineering, Kankavli, India

DOI: 10.5281/zenodo.1173488

ABSTRACT

Basically, data process is seen to be gathering, processing and management of data for giving output of "new" information for end users [2]. Over time, key challenges are related to mining, storage, transportation and processing of high throughput data. It is different from Big Data challenges to which we have to add Volume, Velocity, Value, Veracity, variety, Visualization and Variability [4]. Consequently, these requirements imply an additional step where data are cleaned, tagged, classified and formatted. Big Data analysis currently splits into four steps: Acquisition or Access, Assembly or Organization, Analyze and Action or Decision. Thus, these steps are mentioned as the "4 A's".

KEYWORDS: Acquisition, Assembly, Analyze, Action

I. INTRODUCTION

We are awash in a flood of data today. And the horrible thing is that the data is becoming big and big. It is generated in multiples of terabytes and petabytes per day. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or estimation on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences, construction, defense. The paper's primary focus is on the 4 steps of big data analysis ie data Acquisition, data Assembly, data Analyze and data Action.

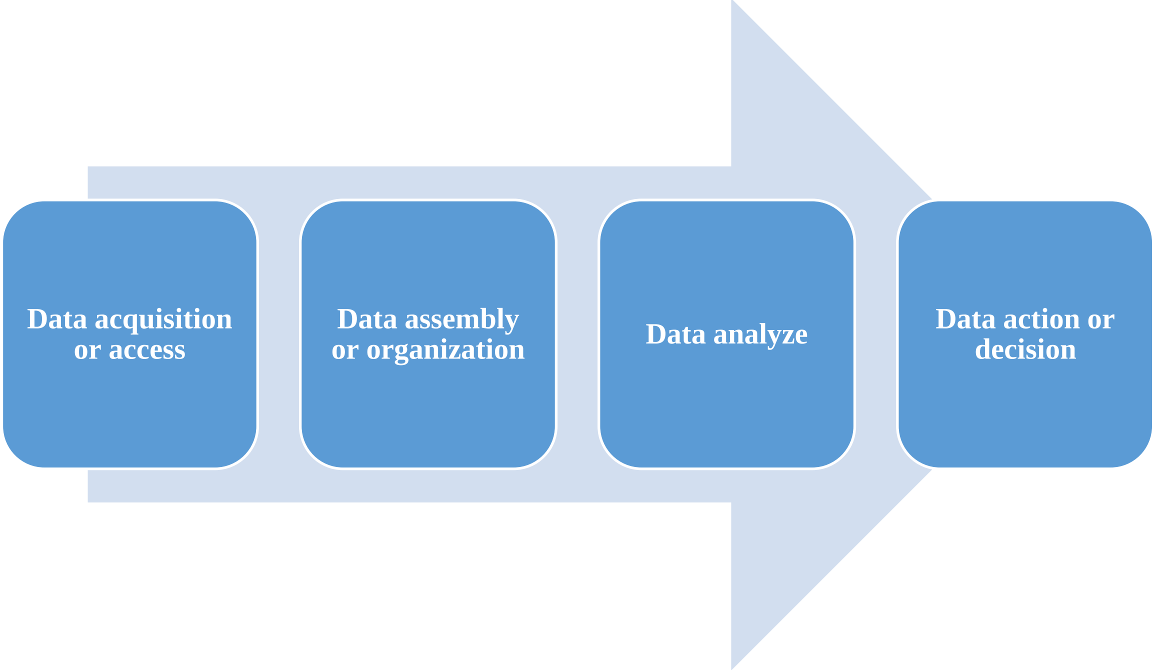

Figure-1: Major Steps in Big Data Analytics pipeline

II. DATA ACQUISITION OR ACCESS

Big Data architecture has to acquire high speed data from a variety of sources like web, DBMS(OLTP), NoSQL, HDFS and the data is also diverse in nature . It is required to store only data which could be helpful or "raw" data with a lower degree of uncertainty[1]. For that a filter could be established. In some applications, the conditions of generation of data are important, thus it could be interesting for further analysis to capture these metadata and store them with the corresponding data [1].

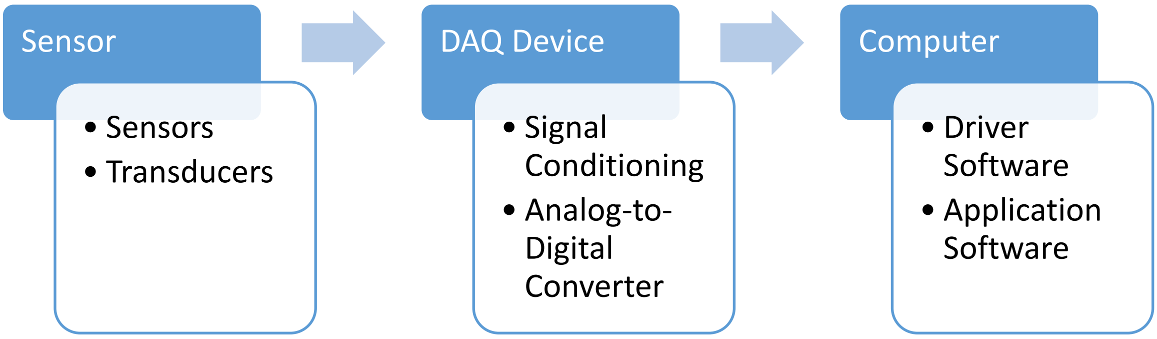

Figure-2: Major Steps in Data Acquisition System

Big Data is recorded from some data generating source like, ability to sense and observe the world, from the heart rate, presence of toxins in the air we breathe, which will produce up to 1million terabytes or more than that of raw data per day. Similarly, scientific experiments and simulation modeling can easily produce petabytes of data per day. The data can be filtered and compressed by orders of magnitude because much of this data is of no interest. But these filters do not discard useful information and it is the one challenge. We all need research in the science of Big Data analytics for data reduction that can smartly handle this raw data to a size that its users can handle while not missing the needle in the haystack. In addition to this, we require "real time" analysis techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward. The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data origin. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. Thus we need research both into generating suitable metadata and into data systems that carry the origin of data and its metadata through data analysis pipelines.

Frequently, the information collected will not be in a format ready for analysis. We cannot leave the data in this useless form even we unable to analyze it effectively. But we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is absolutely a continuing technical challenge. Note that this data also includes images and videos; such extraction is often highly application dependent In addition, due to the ubiquity of surveillance cameras and popularity of GPS-enabled smart phones, cameras, and other portable devices, rich and high fidelity location and trajectory, data can also be extracted. We are used to thinking of Big Data as always telling us the truth, but this is actually far from reality.[1] Existing work on data cleaning assumes well-recognized constraints on valid data or well-understood error models; for many emerging Big Data domains these do not exist .

III. DATA ASSEMBLY OR ORGANIZATION

At this point the architecture has to deal with various data formats like texts formats, compressed files, variously delimited, twits, mails, videos with structured, semi structured, unstructured nature and must be able to parse them and extract the actual information like named entities, relation between them, etc. [4]. Also this is the point where data have to be clean, put in a computable mode, structured or semi-structured, integrated and stored in the right location like existing data warehouse, data marts, Operational Data Store, Complex Event Processing engine, NoSQL database [1]. Thus, a kind of ETL (extract, transform, load) had to be done. Successful cleaning in Big Data architecture is not entirely guaranteed; in fact "the volume, velocity, variety, and variability of Big Data may preclude us from taking the time to cleanse it all thoroughly"

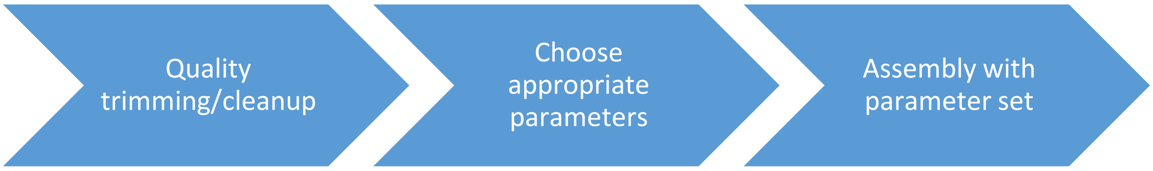

Figure-3: Major Steps in Data Assembly Process

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Data analysis is considerably challenging than simply locating, identifying, understanding, and citing data. For effective large scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error free difference resolution. Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information regarding substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

IV. DATA ANALYZE

Here we have running queries, modeling, and building algorithms to find new insights. Mining requires integrated, cleaned, trustworthy data; at the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions [1].

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big---data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. As noted previously, real life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems. The value of Big Data analysis in health care, to take just one example application domain, can only be realized if it can be applied robustly under these difficult conditions. On the flip side, knowledge developed from data can help in correcting errors and removing ambiguity. For example, a physician may write "DVT" as the diagnosis for a patient. This abbreviation is commonly used for both "deep vein thrombosis" and "diverticulitis," two very different medical conditions. A knowledge base constructed from related data can use associated symptoms or medications to determine which of two the physician meant. Big Data is also enabling the next generation of interactive data analysis with real time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot lists or recommendations, and to provide an ad-Hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non SQL processing such as data mining and statistical analyses. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process and bringing the data back. This is an obstacle to carrying over the interactive elegance of the first generation of SQL driven OLAP systems into the data mining type of analysis that is in increasing demand. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

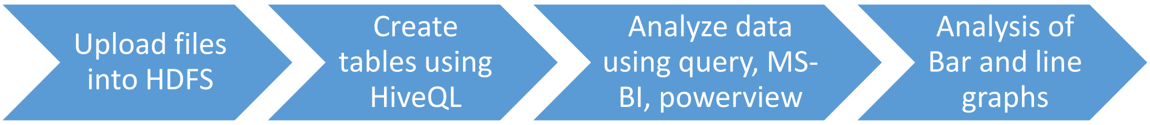
Upload files into HDFS
Create tables using HiveQL
Analyze data using query, MS-BI, powerview
Analysis of Bar and line graphs

Figure-4: Major Steps in Data Analyze Process

V. DATA ACTION OR DECISION

Being able to take valuable decisions means to be able to efficiently interpret results from analysis. Consequently it is very important for the user to “understand and verify” outputs [1]. Furthermore, origin of the data (supplementary information that explains how each result was derived) should be provided to help the user to understand what he obtains. If we can easily see how volume, velocity, veracity and variety influence the pipeline of Big Data architecture, there is another important aspect in data to handle in Big Data Architecture that is privacy. Privacy consideration is very important that it appears in a good place in his definition of Big Data. Privacy can cause problems at the creation of data(someone who wants to hide some piece of information), at the analysis on data [1] because if we want to aggregate data or to correlate it, we could have to access private data; and privacy can also cause inconsistencies at the purging of database. Indeed if we delete all individuals data, we can get in coherences with aggregate data. To sum up handle Big Data implies having an infrastructure linear scalable, able to handle high throughput multi-formatted data, fault tolerant, auto recoverable, with a high degree of parallelism and a distributed data processing [3]. It is important to note that, in this management, integrating data (i.e “access, parse, normalize, standardize, integrate, cleanse, extract, match, classify, mask, and deliver data.” represents 80% of a Big Data projects. There are various tools which can be used in Big Data management from data acquisition to data analysis. Most of these tools are parts of Apache projects and are constructed around the famous Hadoop. Written in Java and created by Doug Cutting, Hadoop brings the ability to cheaply process large amounts of data, regardless of its structure [2]. Hadoop is made up of two core projects: Hadoop Distributed File System(HDFS) and MapReduce.

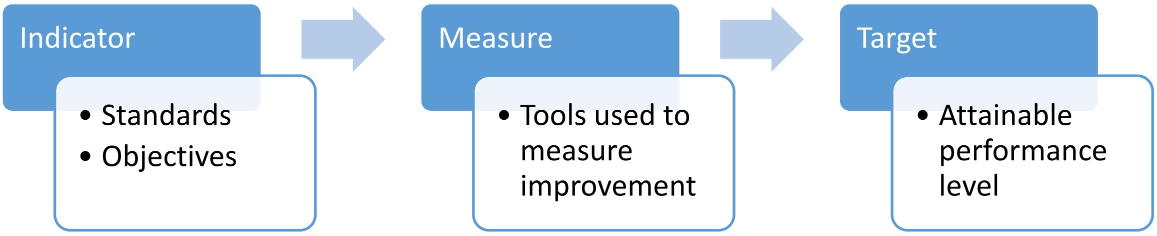
Indicator
• Standards
• Objectives
Measure
• Tools used to measure improvement
Target
• Attainable performance level

Figure-5: Major Steps in Data Action Process

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision maker, provided with the result of analysis, has to interpret these results. These interpretations cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error namely computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will surrender authority to the computer system. Rather user will try to understand, and verify, the results produced by the computer. The computer system must make it easy for user to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage related shock to the financial system dramatically under scored the need for such decision maker diligence rather than accept the stated solvency of a financial institution at face value a decision maker has to examine critically the many assumptions at multiple stages of analysis. In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the origin of the (result) data.

By studying how best to capture, store, and query provenance, in conjunction with techniques to capture adequate metadata, we can create an infrastructure to provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or data sets. Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain. Whereas early business intelligence systems' users were content with tabular presentations, today's analysts need to pack and present results in powerful visualizations that assist interpretation, and support user collaboration. Furthermore, with a few clicks the user should be able to drill down into each piece of data that user sees and understand its provenance, which is a key feature to understanding the data. That is, users need to be able to see not just the results, but also understand why they are seeing those results. However, raw provenance, particularly regarding the phases in the analytics pipeline, is likely to be too technical for many users to grasp completely. One alternative is to enable the users to "play" with the steps in the analysis make small changes to the pipeline, for example, or modify values for some parameters. The users can then view the results of these incremental changes. By these means, users can develop an intuitive feeling for the analysis and also verify that it performs as expected in corner cases. Accomplishing this requires the system to provide convenient facilities for the user to specify analyses.

VI. CONCLUSION

Data Acquisition, data Assembly, data Analyze and data Action are the 4 steps of big data analytics. This paper primary focuses on these 4 steps and gives the detail explanation. These 4 steps are very crucial from the data management point of view. [2]. It is important to note that, in this management, integrating data (i.e "access, parse, normalize, standardize, integrate, cleanse, extract, match, classify, mask, and deliver data." represents 80% of a Big Data project.

VII. REFERENCES

[1] Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin., ... Widom J. (2012). Challenges and Opportunities with Big Data : A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc - led - whitepapers/

[2] CheikhKacfahEmani, Nadine Cullot, Christophe Nicolle, "Understandable Big Data: A survey," ⓝ 2015 Elsevier Inc.,computer science review17 (2015) 70−81.

[3] GemaBello-Orgaza,JasonJ.Jungb,∗,DavidCamachoa,"Socialbigdata:Recent achievements and new challenges," in science direct Information Fusion 28 (2016) 45–59.

[4] Avita Katal, Mohammad Wazid, R H Goudar, Big Data: Issues, Challenges, Tools and Good Practices 978-1-4799-0192-0/13/$31.00 ©2013 IEEE

CITE AN ARTICLE

Nimankar , S., Prof., & Dagare, S., Prof. (n.d.). BIG DATA ANALYTICS: 4A's. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 7*(2), 328-332.