ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

An Integrated approach for Big Data Analytics through Data Lake

Harsh Arora[1], Deepti Gupta [2], Govind Murari Upadhyay[3]

Assistant Professor, Department of Computer Applications, IINTM, Janakpuri, New Delhi, India[1]

Assistant Professor, Department of Computer Applications, IINTM, Janakpuri, New Delhi, India [2]

Assistant Professor, Department of Computer Applications, IINTM, Janakpuri, New Delhi, India [3]

ABSTRACT: Data Lake under Big Data Analytics is a new concept that is more than a simple repository of data. Also Data Lake is different from and completely refined than Data ware house or Enterprise Data Ware House. As there are some storage and capability limitations in Data Warehouse. So from organization's point of view, introducing the concept of Data Lake in EDW or the collaboration of both has given tremendous and efficient outputs in context of information storage capabilities, extraction and availability of timely requirement of data which leads to easy and efficient data processing. Data Lakes lead to the change from simple repository of data or in other words from organization's move from batch to data processing and also integration with Hadoop platform gives seamless and immense results to the system. Basically Data Lakes are being used for quick response of information from user's and organization's point of view. The continuous and fast accessing of data and as per the instant requirements of user or the system is the purpose of using Data Lake. Its role is to provide such an interface which is not bounded or restricted to the predefined schema as in the case of Data Warehouse. Also along with Hadoop and other big data handling technologies Data Lake utilization is great to handle collectively multiple and parallel data sets for quick reply and immediate response and action to the user without waiting so much. Hence Data Lake overcomes all the limitations of Data Warehouse hence so efficient and quick for big data handling though multiple softwares.
.

KEYWORDS: Data Lake, Big Data, EDW, Hadoop, Big data analytics.

I. INTRODUCTION

As we know "Big Data is high volume ,high velocity and high variety information assets that demand cost effective, innovative forms of information processing that enable enhanced insight, decision making and process automation[1].In order to handle Big Data, the concept of Data Lake has been introduced. A Data Lake is more than a simple dumping ground for data. Data Lake has a big role in big data analytics. Big data analytics is required to be executed on frequently changing data. As storage of data and the corresponding related activities in warehouse or in any particular schema is a cumbersome job. Hence in this regard, Data Lake gives a solution to this problem in Data Warehouse. We follow a flat architecture to store big amount of data in a repository that is present in structured manner in Data Lake. "Although data is stored in raw format in Data Lake but each data element is assigned a unique identifier and given a tag with related meta tags" [2].As far as structured analysis of big data is concerned, in context of Data Lake when data is needed to be fetched by users then Data Lakes provide small data sets which are analysed and categorized accordingly. Moreover, multiple types of data can be stored under single repository that is known by Data Lake. Data Lake has many advantages over Data Warehouse. ETL (Extract, Transform and Load) methods are used in Data Lake in order to make data structured and integrated instead of using normal traditional database approach. Semi structured, unstructured and structured all kind of data can be handled and managed by Data Lake. "Data Lake provides agility and can work properly even if some data is unavailable" [2].In Data Lake, data is stored and analysed at low cost with great efficiency. Also under the same umbrella and infrastructure, multiple types of data whether it is text, video, audio or any other document can be stored and analysed. Data Lakes are useful in those environments where analysis is being done on dynamic data. Data Lakes provide quick analysis of data as per the requirement of user's

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

project under consideration. Data Lakes provide accurate results and always update data as per the need. Hence it becomes easy for an organization or users .When we talk about big data analytics, to manage different kinds of data and utilization of data as and when required. Data Lakes provide the platform to the users to access the data as per their needs quickly with accurate results and in structured manner.

II. WHY DATA LAKE & NOT DATA WAREHOUSE: A BIG DATA AND FAST DATA APPROACH

The key reason to include the concept of Data Lake while we are having managed Data Warehouse for keeping Big data is to make data available without doing much tedious work and unnecessary work. We all know that there are certain rules to store data in data warehouse or in other words we can say Data Warehouse is built upon several principles:

- Based on global data model to store data
- Data in read only mode, data cannot be changed.
- Clean and consistent data.
- Particular loading of data, cleaning, reformatting and integration is done.

On the basis of these principles of Data Warehouse, data accessing cannot be done so much quickly and efficiently as we need today as this approach of Data Warehouse we have been using for many years but only IT professionals could get and access data as they want but moreover in this approach, data analysis becomes so expensive and it becomes difficult to access. Now there must be some solution to this problem which has come into the picture in form of "DATA LAKE". All constraints and problems of Data Warehouse data accessing have been changed. Many tools to analyze data now are available in the market. Moreover data can be present and found everywhere in form of cloud based applications, private and public sources and also open data sources that anyone can access[3].The most important aspect in case of Data Lake is that no special defined schema is required. Schema is imposed and transformations are done at the query time that is schema-on-read. Also in Data Lakes, all data has potential value plus apps and users interpret the data as they see fit. On the other side if we talk about Data Warehouse-it follows traditional business analytics process that starts with user requirements, then corresponding database schema and queries are defined. After that required data sources are identified. Moreover ETL pipeline is used to extract required data (curation) and transform it to target schema(schema-on-write) and finally reports are created and data is analyzed accordingly.

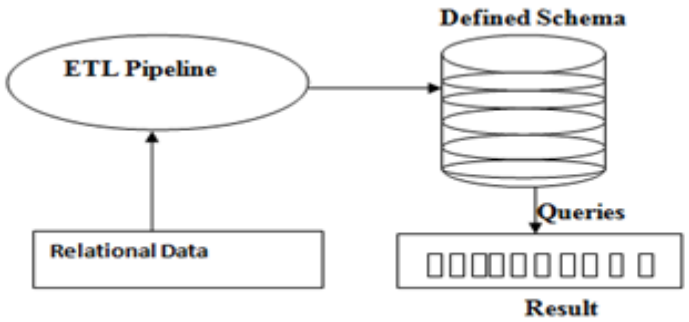

Fig1: A traditional approach for data analysis

As discussed earlier, data wherever present can never makes it into the Data Warehouse, nor should it[3].However it takes much more time whenever data accessing is required in case of data ware house. Data must go through the processes every time before it is available or it can be made available to the user. The processes are assessing, modelling, sourcing, cleaning and loading of data into a warehouse. It creates the task complicated and more time is wasted every time when specific data is required for a particular application by the user. Also it becomes expensive for the system to access the data in such a way.

Here the solution for this tiring and tedious job is Data Lake where no particular defined schema is required and its cost is very low as results are updated and analysed quickly.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The most common reason to build a Data Lake is need for analysis. Data Lake creates a balance between doing work in advance and deferring it until later. Sometimes we want to do the work in advance as the warehouse requires because the data will be used and reused often and by different people. Sometimes we want to defer the work because we are not sure how the data will be used or whether the data has value. Instead of always doing all of the work up front, a Data Lake allows us to choose when to do it and how much to do, speeding the availability of data [3].

A Data Lake is designed to support the environment and work where integration preparation and deriving new data is required as per the situation. While Data Warehouse is a platform which is based on the concept of read only repository .No alteration of data can be done in this regard. Hence Data Lake is more flexible approach and more supportive.

III. CORE SERVICES AND ANGLES OF ARCHITECTURE POINTING TO DATA LAKE:DIFFERENT PERSPECTIVES

There are various categories of services that make up the functionalities of Data Lake. The combination of all these core services gives a new direction and angle to a simple data repository that comes into the picture in the form of Data Lake. These services that define the Data Lake architecture are data persistence, data movement, data access, processing engines, data flow tools, scheduling and work flow management, Meta data, data curation and platform services.

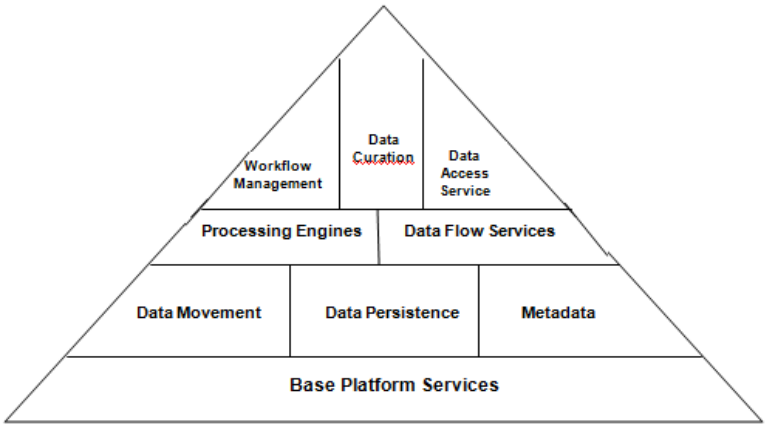

Fig2: Core Services in Data Lake

- Data persistence: Developers, tools and applications need access to higher level services that persist data for different durations from long term raw data storage to short lived output caching[3]
- Data movement: In Data Lake, data movement is bidirectional. Data can be added to the lake as required or it can be fetched back from elsewhere by the lake. There are various services to collect the data like from streams, services, files or database.
- Data access: data accessing is much easier in Data Lake although access services are different from the data movement services. Also there are many ways to access data from the large set of data to the individual objects.
- Processing engines: Depending on the applications we can have many different engines as one single engine cannot solve all the data problems. Few engines support more than one type. Engines are added in Hadoop for example MapReduce for batch, Spark or Tez for parallel pipelines.
- Dataflow tools: These are very useful in the way to abstract the problem of building pipelines so the developer can spend more time on the valuable work that is logic of dataflow and less on the redundant infrastructure [3].
- Scheduling and workflow management: This is a great aspect of Data Lake where excellent coordination is needed for different execution engines as multiple jobs are being done at the same time. So workflow is definitely a

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

necessary aspect of any processing beyond the simple persisting of data. Also this feature leads to the flexibility of defining and resolving scheduling priorities for jobs within and across workflows.

- Metadata: The problem of finding all kind of data and related information over the time is solved through metadata indexing and access services. Schema metadata is also important. All details are recorded with a dataset. In Data Lake, repository of metadata is being needed in order to store all information that leads to easy accessing of data over the period of time.
- Data curation and platform services: Interfaces and many other services are required to lead a mechanism for tracking all kind of data sets as there are multiple data sets present in a Data Lake where some data sets are created by developers and others are created by end users. So in this manner we can conclude that self service analysis works only in curated environment where a platform of multiple services is being coordinated. The idea is that Data Lake is built on the platform having all those services which are responsible for easy storage, execution, security and further more processing of data present in multiple data sets. The services are storage management, process coordination, resource management, encryption authentication, monitoring, logging and security.

IV. ANALYTICS FRAMEWORKS: IMPLEMENTING A DATA LAKE FOR FRAME WORK FLEXIBILITY

If we put light on data driven enterprise, we see Data Lake as an operational foundation that handles and coordinates all the data into a single location with perfect big data analytics tools in one or another way. To handle the volume and velocity of Big data, a Data Lake must be an order of magnitude more scalable than existing approaches for data warehousing and business intelligence[4].Data Lake has changed various analytical decisions by empowering the IT infrastructure as per the rapidly changing needs of the business ,multiple data sets and various other big data tools. Also it provides a platform for protection, backup and security of data as well. Importantly, the flexibility of a Data Lake fosters data analysis with any analytics application, from new frameworks like spark to established languages like Python and R.[DL3].When we talk about enterprise Data Lake ,it has the ability to provide the storage platform to security segregate the business units and their data sets .Although there are certain requirements for storage platform for secure Multitenancy(related to secured segregated business units and datasets).In order to implement a Data Lake for analytical framework flexibility ,these requirements should be satisfied as these requirements make a group of issues analytically that a storage solution for big data analytics must address. These important requirements can be depicted as follows:

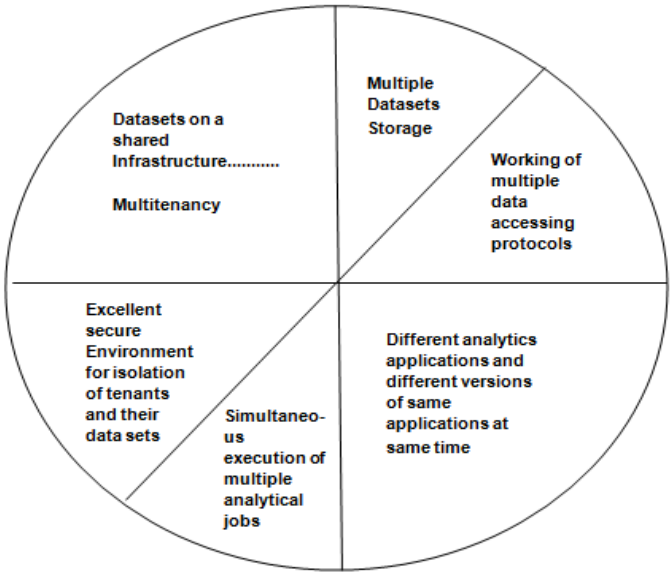

Fig3: Requirements to support Multitenancy

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

V. DATA LAKE CREATION: FROM ORGANIZATION'S PERSPECTIVE

From organization point of view, when we talk about big data repository essential stage required, Data Lake provides the flexibility to enhance the integrated services in an organization and how to handle all data queries from large data sets according to the demand of end users and naive users also. It is very important to have Data Lake that can be called as enterprise Data Lake .For Data Lake creation four essential stages are required. It is a gradual procedure to build or make a robust Data Lake. With right tools, a clearly planned platform, a strong and uniform vision and a quest for innovation an organization can architect an integrated, rationalized and rigorous Data Lake repository.[5].There are following 4 key steps or stages which are needed to create a Data Lake for an organization:

- Scalable data handling and ingestion:

First stage talks about the building block that puts architecture together. Simple transformations are handled over it. As we know how difficult is to handle and manage the bulk of data these days .However depending on the current transactions and user requirements, clean data and actual data must be abstracted from the large data sets. In this regard, Data Lake helps in enterprise data in context of data handling and ingestion. For smooth functioning and fetching of data, the first stage creates the base or building block of the platform. Also it is very important step for understanding and facilitating Hadoop services and work in an organization. Hadoop has become the biggest platform today to manage complex and bulk of data called big data. With the introduction of Data Lake, it has become so easy to handle data as there are various challenges and problems in the platform of Data Warehouse which have overcome by the efficient concept of Data Lake.

- Analytical ability enhancement:

Data analysis and interpretation is being focused through Data Lake and various tools are used to combine and merge EDW and Data Lake .Although enterprise Data Warehouse provides the mechanism of storage of bulk of data and also provides various ways to retrieve the required data as per the system requirement but there are various problems with EDW as discussed earlier. Hence there is a need of second step to amalgamate EDW with new concept of Data Lake.

- EDW and Data Lake collaboration:

The collaboration of enterprise Data Warehouse and Data Lake is very useful and important to facilitate the seamless flow of data where data pool is created for analyzing the data intelligently across the organization. Moreover Hadoop based Data Lake provides the way to efficient flow of data through large no of data sets.

- End to end adoption and maturity acquisition:

The fourth stage is the final and highest stage of maturity. This ties together enterprise capabilities and large scale unification covering information governance, compliance, security, auditing, metadata management and information life cycle management capabilities. [5]

VI. CONCLUSION

EDW leads to certain problems that are overcome by introducing the concept of Data Lake.Although the collaboration of EDW and Data Lake also gives a tremendous output to the organization. It provides a natural process of capabilities that gives fast returns in form of accessing of data. For getting quicker and driving business results, an organization must used Data Lake inspite of using Data Warehouse. In order to enrich our enterprise knowledge warehouse, implementing Data Lake gives immense, seamless and quick response from huge sets of complicated data present in the form of multiple parallel data sets. Data Lake concept has been emerged due to the need or necessity to manage and exploit new sets of data. Data Lake provides the method to preserve data relevance. Along with Hadoop platform with handling of big data enterprise gains more and more advantage by using Data Lake as an interface. Data Lake that is undoubtedly built upon several design principles is more than a simple dumping ground for data. Data Lakes do not necessarily to be Hadoop based always. Data Lakes are becoming more and more central to enterprise data strategies.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The biggest advantage of Data Lake is flexibility .Moreover, raw data storage is completely easy, the best part is that we can refine it as ours understanding and requirements. There are unlimited ways to query the data. Hadoop based Data Lake is important because it can extend the life and capabilities of a Data Warehouse .Hadoop based Data Lake is gaining in popularity because it can capture the volume of big data and other new sources that enterprises want to leverage via analytics and it does so at a low cost and with good interoperability with other platforms in DWE. In this sense, Hadoop and Data Lake have added the refinement and value to Data Warehouse and its environment without ripping and replacing mature investments [6].

REFERENCES

[1]Gartner, Big data definition

[2]Remya Sasidharan Panicker, "Data lake: A next generation data storage systemin Big data analytics", CSI communications, Volume no. 41, Issue no. 1

[3]Mark Madsen, "How to build an enterprise Data Lake: Important considerations before jumping in" Third nature Inc., Snaplogic,Inc.,www.snaplogic.com

[4]REDEFINE EMC White Paper: "Data lakes for data science: Integrating Analytics tools with shared infrastructure for Big data", May 2015.

[5]Article :Impetus: "Five steps to implement an enterprise Data Lake",www.impetus.com

[6] Philip Russom, "Benefits of the Hadoop Based Data Lake", November3, 2016, www.upside.tdwi.org