Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

Big Data Analytics on AWS Cloud

Using AWS Athena and QuickSight

Anand Mishra
M.TECH CSE Scholar
Shri Balwant Institute of Technology
Dcrust, Murthal,
Sonepat-131039, Haryana

Gajendra Kumar
Assistant Professor, CSE
Shri Balwant Institute of Technology
Dcrust, Murthal,
Sonepat-131039, Haryana

***Abstract*:- This paper concentrates upon the recent trends in Big Data analytics in the AWS clouds. How analytics on AWS Cloud changing the landscape of Big Data Analytics and how easy it is for small businesses to jump into Big Data Analytics space. Some light is also thrown into the future scope of this concept.**

Keywords— Big Data on Cloud, Big Data Analytics, Big Data Analytics on AWS Cloud, AWS Big Data tools, AWS.

I. INTRODUCTION

Big Data is large sets of structured and unstructured complex data which traditional processing techniques or algorithms cannot process. Big data helped in revealing hidden patterns and has led to an evolution from a model-driven science paradigm into a data-driven science paradigm. According to a study by Boyd & Crawford [5] it is based on the interplay of:

Technology: With the help of technology, we can maximize computation power and algorithmic accuracy. This will in return help us to gather, analyze, link, and compare large data sets.

Analytics: Large data set analytics helps to identify patterns to make economic, social, technical, and legal decisions.

Mythology: Large data sets provides knowledge that helps us to get insights and intelligence. These insights were previously not very accurate and objective.

Big Data has four areas: Volume, Velocity, Variety, and Veracity [7]. Big Data consists of high-volume, velocity and variety of information that need cost-effective and intelligent processing for insights which help in decision making [8].

II. BIG DATA DIMENSIONS

Big data has four dimensions as described below [19]:

- Volume – Current data existing is in petabytes, which is already problematic; it is predicted that in the next few years it is to increase to zettabytes (ZB) [39]. This explosion of data is mainly due to social media and mobile devices.
- Velocity – Refers to both the rate at which data is captured and the rate of data flow. As Live data is too large and continuously in motion, it causes challenges for traditional analytics.
- Variety – As data collected is not of a specific category or from a single source, Data exists in numerous raw data formats obtained from the web, texts, sensors, e-mails, etc. which are structured or unstructured. It is not from a specified source or from a single category. Traditional analytical methods cannot manage this kind of data known as big data.
- Veracity – Ambiguity within data typically from noise and abnormalities within the data is the primary focus in all four V's. Big Data helps enterprise to develop big data driven e-commerce architecture which aids in gaining extensive "insight into customer behavior, industry trends, more accurate decisions to improve just about every aspect of the business, from marketing and advertising, to merchandising, operations, and even customer retention [9].

III. ANALYTICS

Computational analysis of data which is done in systematic way is known as Analytics. [1] Analytics helps to discover, interpret, and communicate meaningful patterns in data. These data patterns help towards effective decision making. Analytics is valuable when there is abundant recorded information. Computer programming, operation research and simultaneous application of statistics is required quantify performance of Analytics.

To improve business performance, predict and describe; Organizations applies analytics to business data. Specifically, areas within analytics include Big Data Analytics, retail analytics, supply chain analytics, predictive analytics, web analytics, call analytics, speech analytics, prescriptive analytics, enterprise decision management, descriptive analytics, cognitive analytics, predictive science, graph analytics, credit risk analysis, and fraud analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modelling, sales force sizing and optimization, price and promotion modelling. Analytics requires extensive computation. All latest technologies in computer science, mathematics and statistics are used in algorithms and software for analytics.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

IV. CLOUD COMPUTING

Cloud computing basically provides databases, storage, servers, networking, software, intelligence, and analytics through the Internet known as cloud. Cloud computing gives the power to have faster innovation, scale economically and use resources as per your need [19]. Cloud computing helps you to lower your operating costs by giving flexibility to pay only for cloud services you use. Thus, running your infrastructure more efficiently and scale as your business according to the need.

A. Benefits of cloud computing

Cloud computing changed thinking of businesses about IT resources. Below are some benefits of cloud computing [19]:

- Cost – Cloud computing takes the hassle and expense of buying software and hardware from you. It eliminates the expense of setting up and running on-site datacenters which include the racks of servers, IT experts for the infrastructure, the 24/7 electricity for power and cooling. It adds up to huge amount.
- Speed – Cloud computing provides vast number of resources provisioned within minutes on demand. It gives business flexibility and ease of capacity planning.
- Global scale – Cloud computing can scale as and when required. It can provide resources from the right geographical location. It can provide scalable resources like storage, computing power and bandwidth from the location near to you.
- Productivity – Datacenters placed on-site requires multiple tasks like racking and stacking, hardware setup, software patching and other maintenance, which requires IT team to constantly involve in maintaining the Datacenter. Cloud computing does not require these tasks so that IT team can utilize their time on other tasks beneficial to organization.
- Performance - Cloud computing datacenters are always updated with latest, fastest, and efficient computing hardware. They have a worldwide network of secure datacenters. In comparison to single corporate, it gives more benefits in term of reduced network latency due to geographical availability of locations and cloud computing becomes more economical to scale [19].
- Reliability – Due to replication of data in cloud computing, businesses can relax in terms of data backup and disaster recovery. Business continuity is easier and less expensive because of data replication on cloud provider's network.
- Security - Cloud providers protect your data, apps, and infrastructure by providing various controls, polices and technologies.

B. Types of cloud computing

Different types of cloud computing suites different business demands. There are several different types, models, and services to help business find the right solution [19].

There are three types of cloud deployment or computing architecture namely as public, private and hybrid clouds. Let us see each of them [19].

- Public cloud
  In public clouds, computing resources like storage and servers are provided over the internet. They are operated and owned by third part cloud service provider. Google Cloud, AWS, Microsoft Azure are some examples of public cloud. All the supporting infrastructure, software and hardware managed and owned by the cloud provider. Using web browser, you can manage and access these services.
- Private cloud
  When a cloud computing resource is used and managed by only single organization and company has on-site datacenter then it is known as Private cloud. Private cloud can be purchased from third party service providers also. Private cloud as their own private network to interact with the resources.
- Hybrid cloud
  When public and private cloud are combined by technology so that it allows data and applications to be shared between them then it is known as Hybrid cloud. Hybrid cloud offers more flexibility, deployment options, helps to improve security, compliance and optimize infrastructure.

C. Types of cloud services: IaaS, PaaS, serverless and SaaS

Cloud computing services can be classified into four types: infrastructure as a service (IaaS), platform as a service (PaaS), serverless and software as a service (SaaS). They build on top of one another that is why they are also known as cloud computing stacks [19].

- Infrastructure as a service (IaaS)
  In IaaS, servers, storage, networks, and virtual machines are rented out from cloud provider. They are usually pay-as-you-go basis.
- Platform as a service (PaaS)
  PaaS helps developers to create web/mobile apps without worrying about managing underlying infrastructure of servers, storage, network, or database. It is cloud computing service which provides on-demand environment for managing, testing, developing, and delivering software applications.
- Serverless computing
  Serverless computing takes all the worries related to scalability and availability as it is managed by cloud provider. It uses resources automatically depending upon the application load. It overlaps with PaaS as developer only needs to focus on development of web/mobile application, capacity and server management is done by cloud provider.
- Software as a service (SaaS)
  In SaaS, everything is managed by the cloud provider, including managing software application, security patching, software upgrades, infrastructure handling and maintenance. It is given typically as Subscription basis.

D. Uses of cloud computing

Now a day's use of cloud computing is very common. Behind the scenes cloud computing is being used when you

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

play online games, send email, edit documents, watch movies, listen to music, or store pictures and other files. Cloud computing is only decade old but it still being used by various organization big or small, from government agencies to non-profit organizations [19].

Below are some use cases for the same:

- Cloud native applications
  Cloud native applications can easily be built, deploy and scale whether it is for web, mobile or API. Applications today take advantage of cloud-native technologies and approaches, such as containers, Kubernetes, microservices architecture, API-driven communication and serverless architecture.
- Test and build applications.
  Cloud infrastructure can easily be scaled depending upon the application requirement. Application development time and cost can easily be reduced when using cloud infrastructure.
- Store, back up and recover data.
  Data can be stored, backup and recovered from offsite cloud storage system. This data can be accessed from any device and location. By transferring your data to offsite, your data can be protected in cost efficient manner and at a massive scale.
- Analyze data.
  Using AI and machine learning, useful analytics can be uncovered which can help organizations to achieve better informed decision. Data from various teams, divisions and locations can be unified in the cloud.
- Stream audio and video
  HD video and audio can stream to audience anywhere anytime with cloud's global distribution.
- Embed intelligence.
  Data captured from various sources can provide valuable insights and engage customers using embed intelligent models.

V. THE AWS ADVANTAGE IN BIG DATA ANALYTICS

The type of analysis and the number of inputs in large data set demands significantly high computing capacity. Pay-as-you-go cloud computing model supports this characteristic of big data workloads. In this type of model applications can easily scale up and down based on demand. Environment can be resized horizontally or vertically on AWS. You do not have to wait for additional hardware or must over invest in provisioning enough capacity [20].

In traditional infrastructure, system designers must over provision to support mission critical applications in case of surge in data due to increase in business need. On AWS your system runs as efficient as possible because on AWS, according to your big data applications shrink and grows as per demand you can increase compute and capacity in matter of minutes.

You can build sophisticated big data application using AWS's scalable services which are available across different geographical regions. Customers have built successful big data analytics workloads on AWS because AWS provides capabilities which make it an ideal fit for solving big data problems.

The following services for collecting, processing, storing, and analyzing big data are described in order [20]:

- Amazon Kinesis
- AWS Lambda
- Amazon Elastic MapReduce
- Amazon Glue
- Amazon Machine Learning
- Amazon DynamoDB
- Amazon Redshift
- Amazon Athena
- Amazon Elasticsearch Service
- Amazon QuickSight

In addition to these services, Amazon EC2 instances are available for self-managed big data applications.

VI. AMAZON ATHENA

Amazon Athena helps to analyze data stored in S3 by providing interactive query service using standard SQL. Athena does not need any setup or manage as it is serverless, so just start analyzing data immediately without worrying about infrastructure. Athena does not need data to be loaded, it can directly work on the data stored in S3. Using Athena console, you can just start defining your table schema and query the data. All the major data formats like Apache Avro, CSV, ORC, JSON and Apache Parquet are supported by Amazon Athena. It uses Presto with full ANSI SQL support.

Ideal Usage Patterns

- Interactive ad hoc querying – When you must perform one-time interactive SQL queries on Amazon S3 data then Athena is a good tool. In Athena you can start querying data using standard SQL by just defining a table for your data. Amazon QuickSight provides easy visualization for AWS Athena [20].
- To query staging data before loading into Redshift – You can query data stored in S3 using Athena before processing and loading it into Redshift.
- Send AWS Service logs to S3 for Analysis with Athena – AWS provides various service logs through CloudTrail, CloudFront, ELB/ALB and VPC, these flow logs can be analyzed with Athena.
- Building Interactive Analytical Solutions with notebook-based solutions, e.g., RStudio, Jupyter, or Zeppelin - While using notebook-based solutions such as RStudio, Jupyter, and Zeppelin, Analysts and Data scientists are concerned about managing the infrastructure. By using Amazon Athena, you do not have to worry about managing infrastructure, you just must analyze your data using standard SQL. Amazon Athena integrates these notebook bases solutions to gives data scientists a powerful platform for building interactive analytical solutions [20].

A. Cost Model

Amazon Athena has no upfront costs or minimum fees; you simply pay for the resources you consume under pay-as-you-go pricing. Amazon Athena is priced per query. Charges are based on the amount of data scanned by the query. Using columnar data format saves 30% to 90% on per-query costs. It

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

also gives better performance if you compress and partition the columnar data. Athena reads only the columns it needs to process the query in columnar data format [20].

Charges are based on the number of bytes scanned by Amazon Athena, rounded up to the nearest megabyte, with a 10 MB minimum per query. CREATE/ALTER/DROP TABLE like Data Definition Language (DDL) statements, for managing partitions, or failed queries does not attract any charges. Cancelled queries are charged based on the amount of data scanned.

B. *Performance*

Partitioning, compressing, and converting your data into Columnar formats improves the query performance. Apache Parquet and ORC like open-source columnar formats are supported by Amazon Athena. Partitioning, compressing, and converting your data into Columnar formats also lowers cost of Athena as it scans less data from S3 when query is executed.

C. Durability and Availability

Amazon Athena routes queries automatically to available compute resources across regions when a particular facility is unreachable, this makes Athena highly available at all the times. Amazon S3, which is highly available, and is designed for durability of 99.999999999% of objects is used as the data store for Amazon Athena. S3 stores data redundantly in multiple devices in each facility across multiple facilities.

D. Scalability and Elasticity

Athena does not require any infrastructure manage or setup and it scales automatically depending upon the need because it is serverless. So, you can start analyzing data immediately.

E. Security, Authorization and Encryption

AWS Identity and Access Management (IAM) policies, Access Control Lists (ACLs), and Amazon S3 bucket policies provides access control to your data. Fine grained controls to your S3 bucket can be granted using IAM polices. Restricting user to access data in S3 restricts them to query data using Athena.

You can query data that has been protected by:

• Server-side encryption with an Amazon S3-managed key
• Server-side encryption with an AWS KMS-managed key
• Client-side encryption with an AWS KMS-managed key

Result sets of Amazon Athena can be encrypted directly with AWS Key Management System (KMS).

F. *Interfaces*

Athena supports CLI, JDBC and API via SDK to query data. Amazon QuickSight can also be used to visualize data based on Athena queries.

VII. AMAZON QUICKSIGHT

Amazon QuickSight is a cloud powered business analytics tool that can get insight from data and can perform ad-hoc query analysis and build visualization data on any device, it can access on premise databases like SQL Server, MySQL, and POSTGRESQL, it can connect to data sources [20].

Including flat files, CSV, and Excel, it can also connect to AWS resources like Amazon RDS databases, Amazon Redshift, Amazon Athena, and Amazon S3, Amazon QuickSight helps organizations for scaling their business analytics to hundreds and thousands of users and it also delivers fast query performance by using a robust in-memory engine (SPICE)

Amazon QuickSight is built for the cloud with SPICE. It is a super-fast, parallel, in-memory calculation engine. SPICE uses latest hardware innovations and machine code generations to run queries on large datasets and combination of columnar storage to get rapid response. SPICE derives valuable insights from your analysis using rich calculations without managing infrastructures. Data replication Is done by SPICE automatically for high availability and persisted until explicitly deleted by user and helps in performing fast interactive analysis across vide variety of AWS data sources.

A. Ideal usage patterns:

Amazon QuickSight is a business intelligence tool which helps users to provide data insight and to create visualizations to helps make business decisions [20].

It helps to do the following:

- optimized visualization of data
- Create dashboards and KPI's for data insight.
- Allows creating insights and specific analysis and stories to share with others.
- Data analyzing and visualization using logs stored in S3.
- Data analyzing and visualization in AWS resources like Amazon Redshift, Amazon S3, Amazon RDS databases, Amazon Athena.
- Analyze and visualize data in software as a service (SaaS) application like Salesforce.
- Analyze and visualize data in data sources that can be connected to using JDBC/ODBC connection.

B. Costing model:

Amazon QuickSight has two different pricing editions: enterprise edition and standard edition.

Standard edition:

Annual subscription - $9/user/month with 10GB of SPICE capacity

Additional SPICE capacity for $.25/GB/month

Month to month option -$12/GB/month

Enterprise edition:

1)Annual subscription - $18/user/month with 10GB of SPICE capacity

Additional SPICE capacity for $.38/GB/month

2)Month to month option -$24/GB/month

Both the editions offer full features for creating and sharing data visualizations. Enterprise editions offers encryption at rest and Microsoft Active Directory (AD) integration

C. Performance

Amazon QuickSight is built for the cloud with SPICE. It is a super-fast, parallel, in-memory calculation engine. SPICE uses latest hardware innovations and machine code

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

generations to run queries on large datasets and combination of columnar storage to get rapid response [20].

D. Durability and Availability

SPICE automatically replicates data for high availability and enables Amazon QuickSight to scale to hundreds of thousands of users who can all simultaneously perform fast interactive analysis across a wide variety of AWS data sources.

E. Scalability and Elasticity

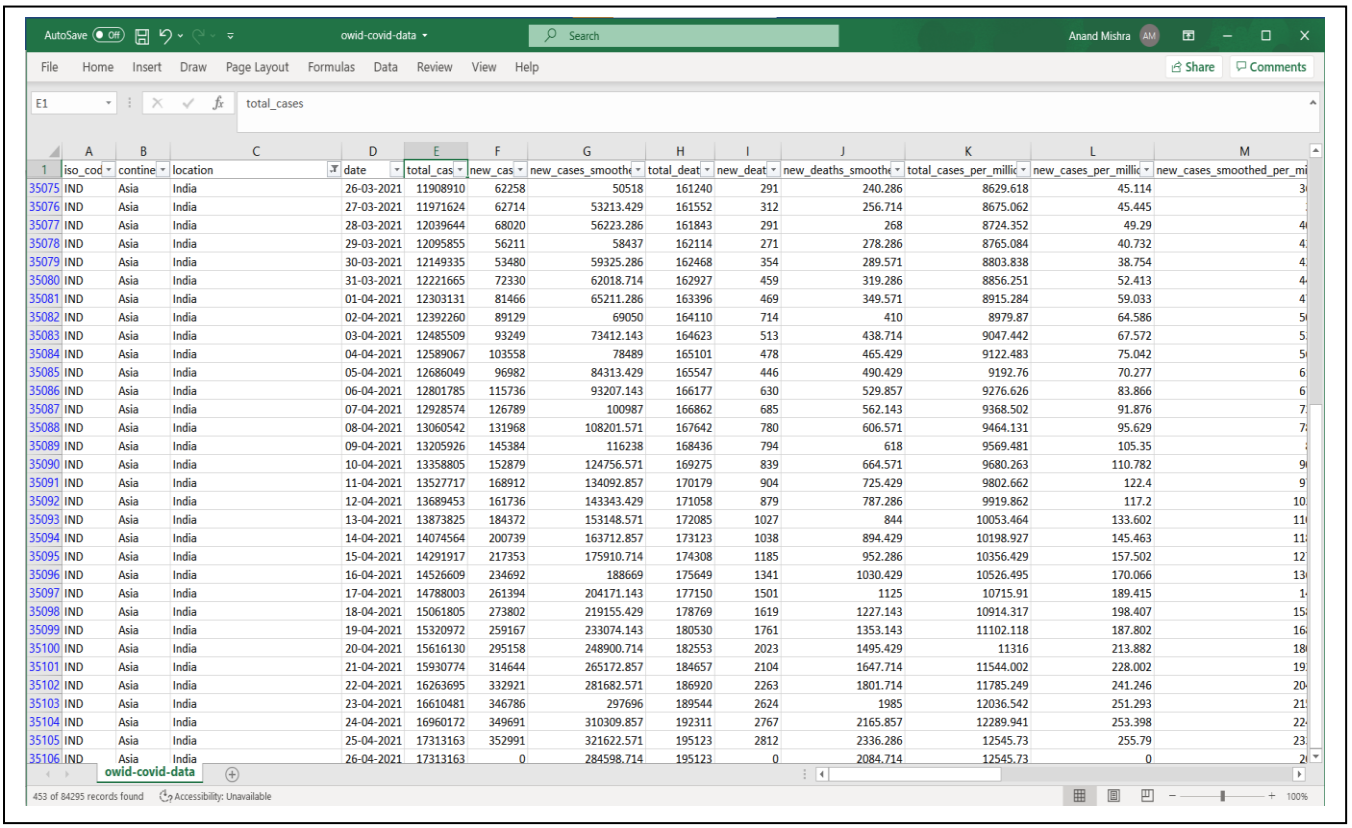
Amazon QuickSight is a fully managed service that internally takes care of scaling to meet the demands of end users, users can manage data from few hundred MB's to TB's without managing infrastructures.

F. *Interfaces*

Amazon QuickSight can connect to on-premises databases like SQL Server, MySQL, and PostgreSQL, and AWS data sources including Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon Athena, and Amazon S3, and SaaS, applications like Salesforce, it can also use variety of data sources that includes flat files (CSV, TSV, CLF, ELF), using a sharing icon you can share analysis and dashboards or story from Amazon QuickSight service interface. You could select (email address, username, or group name), permission levels, and other options before sharing the content with others [20].

VIII. CASE STUDY: AWS'S S3, ATHENA AND QUICKSIGHT FOR DATA PROCESSING AND ANALYTICS

For our case study we have taken Covid data till 26 April 2021, when India got hit by second wave of Covid's double mutant virus B.1.617. We will be analyzing world data which will be huge and complex and can be termed as Big Data related to Covid. We will perform analytics on this Big Data.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	iso_cod	contine	location	date	total_cas	new_cas	new_cases_smoothe	total_deat	new_deat	new_deaths_smoothe	total_cases_per_millic	new_cases_per_millic	new_cases_smoothed_per_mi
35075	IND	Asia	India	26-03-2021	11908910	62258	50518	161240	291	240.286	8629.618	45.114	3
35076	IND	Asia	India	27-03-2021	11971624	62714	53213.429	161552	312	256.714	8675.062	45.445	
35077	IND	Asia	India	28-03-2021	12039644	68020	56223.286	161843	291	268	8724.352	49.29	4
35078	IND	Asia	India	29-03-2021	12095855	56211	58437	162114	271	278.286	8765.084	40.732	4
35079	IND	Asia	India	30-03-2021	12149335	53480	59325.286	162468	354	289.571	8803.838	38.754	4
35080	IND	Asia	India	31-03-2021	12221665	72330	62018.714	162927	459	319.286	8856.251	52.413	4
35081	IND	Asia	India	01-04-2021	12303131	81466	65211.286	163396	469	349.571	8915.284	59.033	4
35082	IND	Asia	India	02-04-2021	12392260	89129	69050	164110	714	410	8979.87	64.586	5
35083	IND	Asia	India	03-04-2021	12485509	93249	73412.143	164623	513	438.714	9047.442	67.572	5
35084	IND	Asia	India	04-04-2021	12589067	103558	78489	165101	478	465.429	9122.483	75.042	5
35085	IND	Asia	India	05-04-2021	12686049	96982	84313.429	165547	446	490.429	9192.76	70.277	6
35086	IND	Asia	India	06-04-2021	12801785	115736	93207.143	166177	630	529.857	9276.626	83.866	6
35087	IND	Asia	India	07-04-2021	12928574	126789	100987	166862	685	562.143	9368.502	91.876	7
35088	IND	Asia	India	08-04-2021	13060542	131968	108201.571	167642	780	606.571	9464.131	95.629	7
35089	IND	Asia	India	09-04-2021	13205926	145384	116238	168436	794	618	9569.481	105.35	
35090	IND	Asia	India	10-04-2021	13358805	152879	124756.571	169275	839	664.571	9680.263	110.782	9
35091	IND	Asia	India	11-04-2021	13527717	168912	134092.857	170179	904	725.429	9802.662	122.4	9
35092	IND	Asia	India	12-04-2021	13689453	161736	143343.429	171058	879	787.286	9919.862	117.2	10
35093	IND	Asia	India	13-04-2021	13873825	184372	153148.571	172085	1027	844	10053.464	133.602	11
35094	IND	Asia	India	14-04-2021	14074564	200739	163712.857	173123	1038	894.429	10198.927	145.463	11
35095	IND	Asia	India	15-04-2021	14291917	217353	175910.714	174308	1185	952.286	10356.429	157.502	12
35096	IND	Asia	India	16-04-2021	14526609	234692	188669	175649	1341	1030.429	10526.495	170.066	13
35097	IND	Asia	India	17-04-2021	14788003	261394	204171.143	177150	1501	1125	10715.91	189.415	1
35098	IND	Asia	India	18-04-2021	15061805	273802	219155.429	178769	1619	1227.143	10914.317	198.407	15
35099	IND	Asia	India	19-04-2021	15320972	259167	233074.143	180530	1761	1353.143	11102.118	187.802	16
35100	IND	Asia	India	20-04-2021	15616130	295158	248900.714	182553	2023	1495.429	11316	213.882	18
35101	IND	Asia	India	21-04-2021	15930774	314644	265172.857	184657	2104	1647.714	11544.002	228.002	19
35102	IND	Asia	India	22-04-2021	16263695	332921	281682.571	186920	2263	1801.714	11785.249	241.246	20
35103	IND	Asia	India	23-04-2021	16610481	346786	297696	189544	2624	1985	12036.542	251.293	21
35104	IND	Asia	India	24-04-2021	16960172	349691	310309.857	192311	2767	2165.857	12289.941	253.398	22
35105	IND	Asia	India	25-04-2021	17313163	352991	321622.571	195123	2812	2336.286	12545.73	255.79	23
35106	IND	Asia	India	26-04-2021	17313163	0	284598.714	195123	0	2084.714	12545.73	0	2

Fig. 1. Latest Covid data

This CSV file contains World stats related to Covid. Around 17 lac records spread across the CSV. We will put this CSV file in Amazon S3 for storage.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

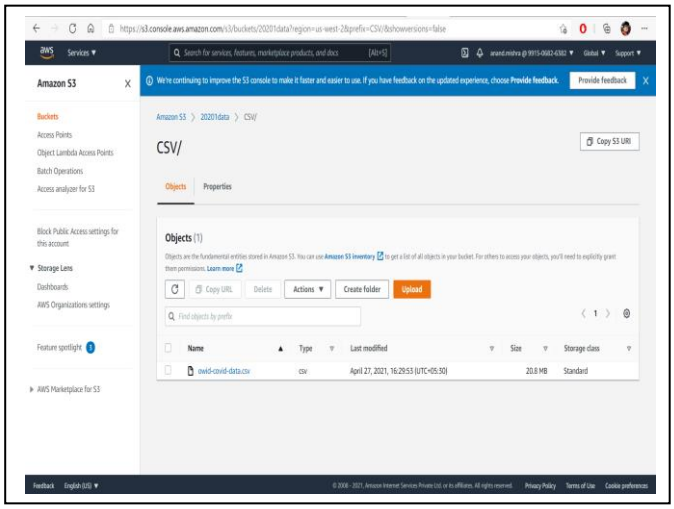


Fig. 2. Data put in Amazon S3.

Once CSV file is uploaded to Amazon S3, we will head over to Amazon Athena to process the data by making a table.

TABLE. I. RESULT GOT FROM SQL QUERY.

S. No.	Continent	Location	Date	Total_Cases	New_Cases
1	Asia	India	44311	17313163	352991
2	Asia	India	44310	16960172	349691
3	Asia	India	44309	16610481	346786
4	Asia	India	44308	16263695	332921
5	Asia	India	44307	15930774	314644
6	Asia	India	44306	15616130	295158
7	Asia	India	44304	15061805	273802
8	Asia	India	44303	14788003	261394
9	Asia	India	44305	15320972	259167

This data can also be visualized using Amazon QuickSight for better and quick understanding. Amazon QuickSight help to visualize your analytics data which you got by querying your Big Data.

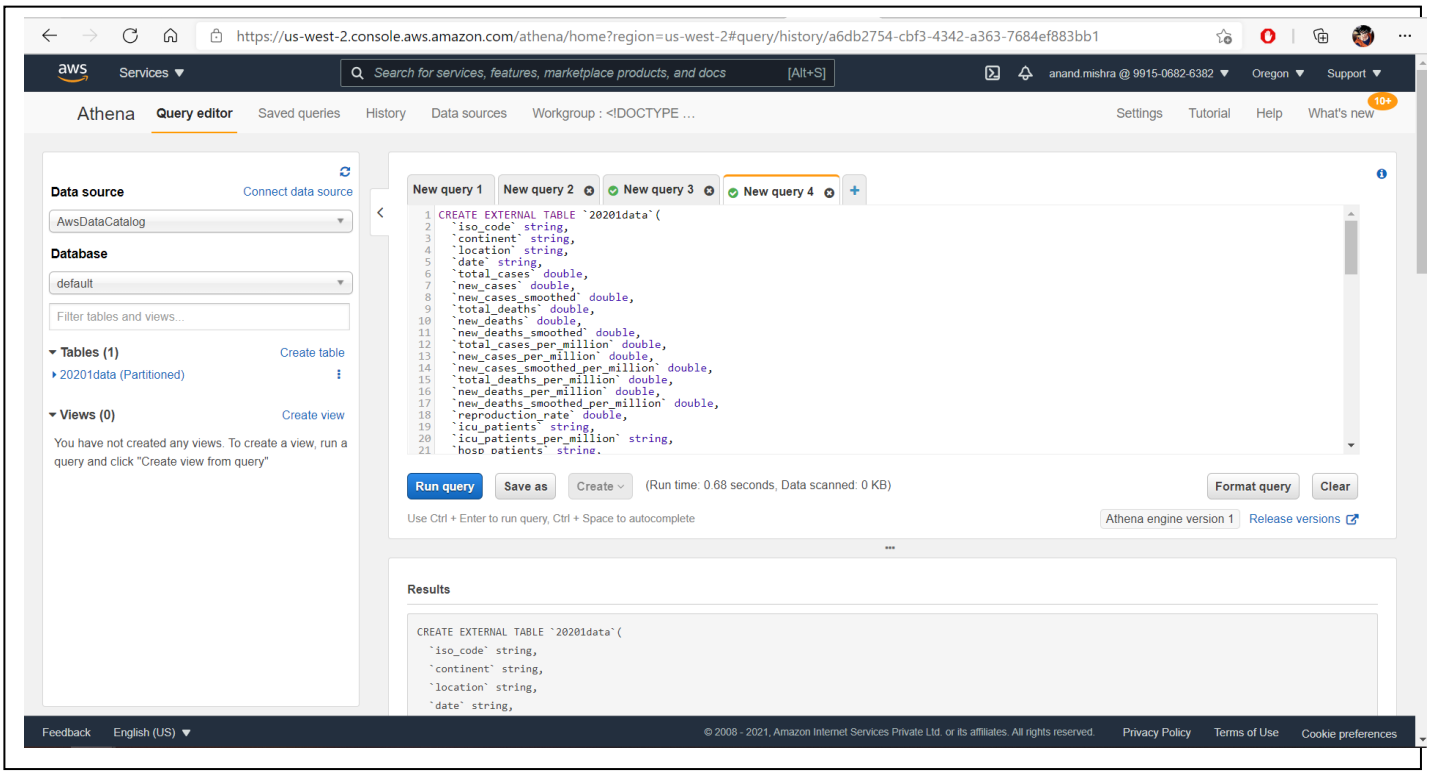


Fig. 3. Query to create table in Amazon Athena.

As shown in Fig. 3. Table has been successfully created. Now we can query the data using standard SQL. We have run the SQL query to analyze number cases in India where total deaths were more than 1000.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

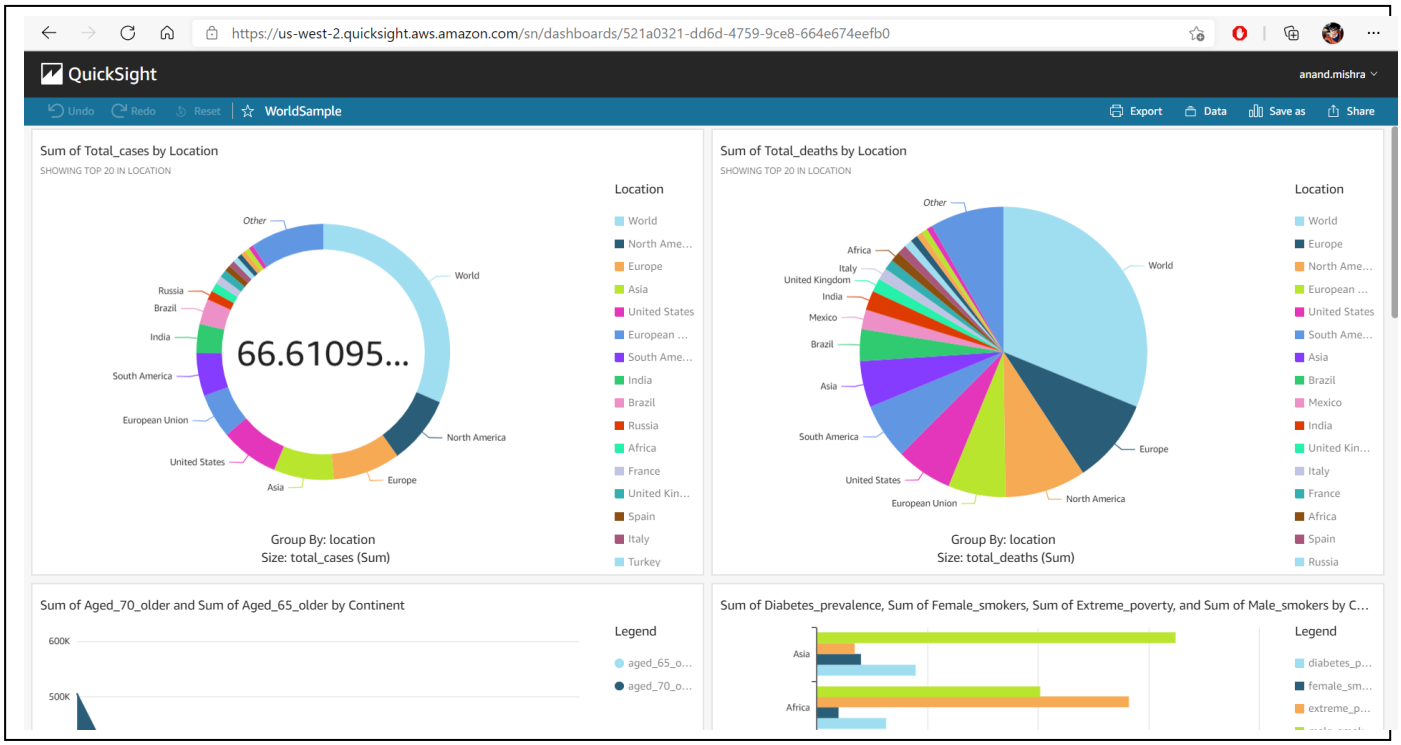


Fig. 4. Big Data Analytics data represented in Graphs and Charts format.

In Fig. 4. you can see the data which we got from analyzing the Big Data. It has been represented in charts and graphs using Amazon QuickSight. It gives better idea of overall statistics of the data.

Fig. 5. QuickSight supports different type of charts.

QuickSight supports various types of charts are shown in Fig. 5. It makes analysts life easy by providing various representations of Big Data Analytics.

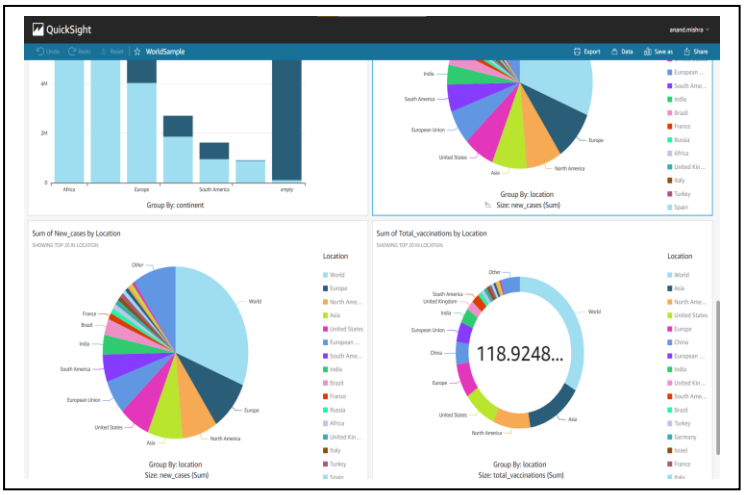


Fig. 6. Representing data by location.

As shown in the table below Table. II., performance of query results does not much depend on size of data. Of course, size of data matters but what matters the most is how much your query demands scanning of whole data. If your query is not optimized, is not partitioned properly, is not compressed or in columnar format then your query performance will take a hit.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

TABLE. II. PERFORMANCE OF QUERIES PERFORMED ON BIG DATA.

Query	State	Run time(sec)	Data scanned
SELECT continent, location, date, total_cases, new_cases, ...	Succeeded	1.32	6.51 GB
SELECT continent, location, date, total_cases, new_cases, ...	Succeeded	1.38	6 GB
SELECT continent, location, date, max(total_cases), new_cases, ...	Failed	0.15	600 MB
SELECT continent, location, date, total_cases, new_cases, ...	Succeeded	1.92	5.5 GB
SELECT continent, location, date, total_cases, new_cases, total_deaths, population FROM "default"."2...	Succeeded	1.55	5.7 GB
SHOW CREATE TABLE 20201data	Succeeded	0.68	0 KB

IX. CONCLUSION AND FUTURE SCOPE

Big Data is not a new term but has gained its spotlight due to the huge amounts of data that are produced daily from different sources. From our analysis we saw that big data is increasing in a fast pace, leading to benefits but also challenges. Cloud Computing is the best solution for storing, processing, and analyzing Big Data. Companies like Amazon, Google and Microsoft offer their public services to facilitate the process of dealing with Big Data.

From the analysis we saw that there are multiple benefits that Big Data analytics provides for many different fields and sectors such as healthcare, education, and business. We also saw that because of the interaction of Big Data with Cloud Computing there is a shift in the way data is processed and analyzed.

Data analytics requires flexible, scalable, and high-performance tools so that it can provide insights more quickly. As more and more data are generated and collected so new tools emerge every now and then, but it is difficult to choose the right tool and to keep pace as most of them “die” very soon.

AWS platform makes it easier to scale, deploy and build big data applications. It provides various managed services to collect, process, and analyze big data. AWS provides various solutions to help in your big data analytic requirements so that you can focus on business problems instead of updating and managing these tools. To achieve a flexible and scalable big data architecture most business use Multiple AWS tools to build a complete solution. This approach helps meet stringent business requirements in the most cost-optimized, performant, and resilient way possible.

In future scope, work can be done on AWS’s new Big Data tool called AWS Glue DataBrew. AWS Glue DataBrew is a new visual data preparation tool that features an easy-to-use visual interface that helps data analysts and data scientists of all technical levels understand, combine, clean, and transform data. It has 250 pre-built transformations so that you can automate filtering anomalies, correcting invalid values, converting data to standard formats and other tasks.

REFERENCES

[1] M. Hillbert and P. Lopez, ”The World’s Technological Capacity to Store, Communicate and Compute Information,” Compute Information.Science, vol. III, pp. 62-65, 2011.
[2] J. Hellerstein, ”Gigaom Blog,” 8 November 2019. [Online]. Available: https://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallelprogramming/. [Accessed 20 January 2021].
[3] Statista, ”Statista,” 2020. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accessed 21 January 2021].
[4] D. Reinsel, J. Gantz and J. Rydning, ”Data Age 2025: The Evolution of Data To-Life Critical,” International Data Corporation, Framingham, 2017.
[5] S. Kaisler, F. Armour and J. Espinosa, ”Big Data: Issues and Challenges Moving Forward,” Wailea, Maui, HI, s.n, pp. 995 - 1004., 2013.
[6] Wikipedia, ”Wikipedia,” 2018. [Online]. Available: https://www.en.wikipedia.org/wiki/Big data/. [Accessed 4 January 2021].
[7] J. Weathington, ”Big Data Defined.,” Tech Republic, 2012.
[8] PCMagazine, ”PC Magazine,” 2018. [Online]. Available: http://www.pcmag.com/encyclopedia/term/62849/big-data.. [Accessed 9 January 2021]
[9] D. Gewirtz, ”ZDNet,” 2018. [Online]. Available: https://www.zdnet.com/article/volume-velocity-and-varietyunderstanding-the-three-vs-of-big-data/. [Accessed 1 January 2021].
[10] S. M. F. Akhtar, Big Data Architect’s Handbook, Packt, 2018.
[11] WhishWorks, ”WhishWorks, ”, 2019. [Online]. Available: https://www.whishworks.com/blog/data-analytics/understanding-the3-vs-of-big-data-volume-velocity-and-variety/. [Accessed 23 January 2021].
[12] S. Yadav and A. Sohal, ”Review Paper on Big Data Analytics in Cloud Computing,” International Journal of Computer Trends and Technology (IJCTT), vol. IX, 2017.
[13] R. Kimball and M. Ross, The data warehouse toolkit: The definitive guide to dimensional modeling, 3rd ed. John Wiley & Sons, 2013.
[14] LaprinthX, ”LaprinthX, ”, 2018. [Online]. Available: https://laptrinhx.com/better-faster-smarter-elt-vs-etl-2084402419/. [Accessed 22 January 2021].
[15] Xplenty, ”XPlenty, ”, 2019. [Online]. Available: https://www.xplenty.com/blog/etl-vs-elt/#. [Accessed 20 January 2021]
[16] Forbes, ”Forbes, ”, 2018. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2019/11/06/fivebenefits-of-big-data-analytics-and-how-companies-can-getstarted/?sh=7e1b901417e4. [Accessed 13 January 2021]
[17] EDHEC, ”EDHEC, ”, 2019. [Online]. Available: https://master.edhec.edu/news/three-ways-educators-are-using-bigdata-analytics-improve-learning-process#. [Accessed 6 January 2021]
[18] Google Cloud, ” BigQuery, ”, 2020. [Online]. Available: https://cloud.google.com/bigquery. [Accessed 5 January 2021] [19] Forbes, ”Forbes ”, 2020. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-muchdata-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=5936b00460ba
[19] Building Big Data and Analytics Solutions in Cloud (ibm.com) By Wei-Dong Zhu, Manav Gupta, Ven Kumar, Sujatha Perepa, Arvind Sathi and Craig Statchuk Available: http://www.redbooks.ibm.com/redpapers/pdfs/redp5085.pdf [Online]
[20] Big Data Analytics Options on AWS (awsstatic.com) [Online] Available: https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf