Analyzing Natural Calamities Using Apache Hive

[1]Sandeep Iddalgave, [2]M.Tejashree, [3] G.Nikhil, [4]Dattu goud
[1]*Ass.professor, Department of IT, MLRIT,Hyderabad,India.*
[2,3,4] *Final year IT,Department of IT, MLRIT,Hyderabad,India.*

Abstrac-: Analysis of the severity of geographical events is the need of the hour. Using Apache hive in hadoop distributed file system (HDFS) environment, we will be analyzing various data recorded by the climate monitoring organization. Around the world various geographical events will be occurring, like: earthquakes, Tsunamis, landslides, volcanic eruptions and so on. To identify the amount of destruction or damage caused by these events, we will be developing hive queries.

Keyword- Bigdata,Apachehive

I. INTRODUCTION

Consistently, land, organic, hydrological, and climatic elements deliver characteristic risks, which now and again result in cataclysmic events that can devastatingly affect biological systems and human social orders. Dangers can be geophysical (e.g. seismic tremors, cyclonic tempests), organic (e.g. pervasion), or created by a blend of various elements (e.g. surges, rapidly spreading fires, and so forth).Huge Data advancements can assume a part in:

- monitoring risks
- mitigating vulnerabilities; and
- Strengthening flexibility of groups.

Especially intriguing is the part of Big Data for identifying quakes, surges, sea tempests, and in addition guaging future event of such dangers.

Cataclysmic events are extraordinary and unforeseen wonders coming about because of regular procedures of the Earth that, ordinarily, cause human and financial misfortunes. Among these damaging occasions, tremors, waves, volcanic ejections, typhoons, tornadoes or surges emerge.

As of late, enormous measure of information are put away in all orders. Geosciences are not a special case. Large time arrangement or high determination satellite and airborne pictures are wellsprings of profitable data. Be that as it may, the learning extraction from such gigantic information can't generally be performed by utilizing standard factual procedures.

Powerful methodologies have been created inside the setting of huge information investigation. These methodologies can manage expansive datasets, thinking about all examples and estimations. With its quick advancement, mechanized machine learning techniques for separating applicable examples, superior registering or information representation are in effectbroadly, and effectively, connected to catastrophic events related information.

For all the previously mentioned, we compassionately welcome the Scientific Community to add to this unique issue, by submitting novel and unique research tending to at least one of the accompanying themes, dependably with regards to enormous information:

1. New methodologies for cataclysmic events prior examples disclosure.
2. New methodologies for cataclysmic events expectation.
3. New methodologies for cataclysmic events information combination and incorporation.
4. New methodologies for cataclysmic events information representation from perceptions and models.
5. Case investigation portraying pertinent discoveries with clear enthusiasm to the Scientific Community.
   At long last, creators are urged to share codes and information so their examinations can be effortlessly reproducible and fill in as seed for future change.

II. METHODOLOGY

a) Technologies used HDFS: Hadoop File System was created utilizing conveyed document framework plan. It is keep running on item equipment. Not at all like other disseminated frameworks, HDFSis exceptionally faulttolerant and outlined utilizing minimal effort equipment.

b) HDFS holds huge measure of information and gives less demanding access. To store such enormous information, the documents are put away over different machines. These documents are put away in repetitive design to protect the framework from conceivable information misfortunes if there should be an occurrence of disappointment. HDFS additionally makes applications accessible to parallel preparing.

Highlights of HDFS

- It is appropriate for the disseminated stockpiling and preparing.
- Hadoop furnishes a charge interface to cooperate with HDFS.
- The worked in servers of namenode and datanode help clients to effortlessly check the status of bunch.
- Streaming access to record framework information.
- HDFS gives record consents and confirmation.

HDFS Architecture: Given beneath is the design of a Hadoop File System.
HDFS takes after the ace slave engineering and it has the accompanying components.

Namenode
The namenode is the item equipment that contains the GNU/Linux working framework and the namenode programming. It is a product that can be keep running on ware equipment. The framework having the namenode goes about as the ace server and it does the accompanying undertakings:Manages the record framework namespace. Regulates customer's entrance to records.It likewise executes document framework operations, for example, renaming, shutting, and opening records and catalogs.

Datanode
The datanode is an item equipment having the GNU/Linux working framework and datanode programming. For each hub (Commodity equipment/System) in a group, there will be a datanode. These hubs deal with the information stockpiling of their framework.
Datanodes perform read-compose operations on the record frameworks, according to customer ask.

- They likewise perform operations, for example, piece creation, cancellation, and replication as indicated by the guidelines of the namenode.

Square: By and large the client information is put away in the records of HDFS. The document in a record framework will be partitioned into at least one fragments and additionally put away in singular information hubs. These document sections are called as squares. At the end of the day, the base measure of information that HDFS can read or compose is known as a Block. The default square size is 64MB, however it can be expanded according to the need to change in HDFS setup.

Objectives of HDFS
Fault discovery and recuperation: Since HDFS incorporates an expansive number of item equipment, disappointment of parts is visit. Subsequently HDFS ought to have systems for speedy and programmed blame recognition and recuperation.Huge datasets: HDFS ought to have many hubs per group to deal with the applications having gigantic datasets.Hardware at information: An asked for errand should be possible productively, when the calculation happens close to the information. Particularly where immense datasets are included, it lessens the system movement and builds the throughput.

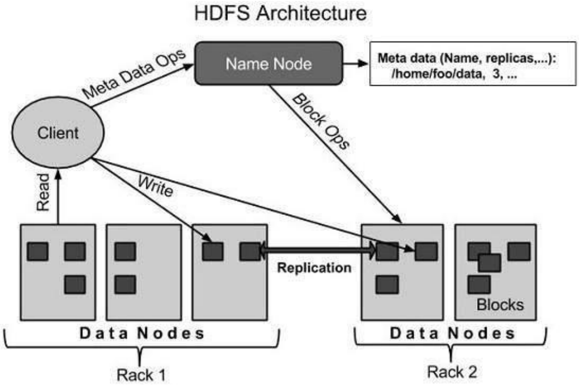
III. HIVE

Hive is an information stockroom framework device to process organized information in Hadoop. It lives over Hadoop to compress Big Data, and makes questioning and investigating simple.

This is a concise instructional exercise that gives an acquaintance on how with utilize Apache Hive HiveQL with Hadoop Distributed File System. This instructional exercise can be your initial move towards turning into an effective Hadoop Developer with Hive
Hive is an information distribution center framework apparatus to process organized information in Hadoop. It dwells over Hadoop to abridge Big Data, and makes questioning and investigating simple.
At first Hive was created by Facebook, later the Apache Software Foundation took it up and created it further as an open source under the name Apache Hive. It is utilized by various organizations. For instance, Amazon utilizes it in Amazon Elastic MapReduce.

Hive isn't

- A outline for OnLine Transaction Processing (OLTP)
- A dialect for continuous inquiries and column level



updates
- Highlights of Hive
- It stores diagram in a database and handled information into HDFS.
- It is intended for OLAP.
- It gives SQL compose dialect to questioning called HiveQL or HQL.
- It is natural, quick, versatile, and extensible.
- Architecture of HiveThe following segment outline delineates the design of Hive

b) Implementation:
- First we downloaded and installed winscp and putty in our working system.
- Then collected some datasets regarding landslides and stored the dataset in the form of csv file.
- We stored that dataset in our folder in winscp by dragging it to the local file.
- And implemented some queries to upload the dataset into the cluster like:
- We used Hadoop fs –copyFromLocal dataset/catalog.csv
- created a database by the command create database

naturalcalamitiesdb

- And created table by using the command

  Sample Queries:

  - Hive>create table naturalcalamities(id int, date date, time int, country_name string, country_code string, state string, population longint, city string, distance float, location_description string, latitude float, longitude float, geolocation double, hazard_type string, landslide_type string, landslide_size string,trigger string, strom_name string, injuries int, fatalities int, source_name string, source_link string)row format delimited fields terminated by ’,’;
  - to upload the dataset we used:
  - hive>load data inpath ‘/catalog.csv’ overwrite into table naturalcalamities.
  - And at last developed some hive queries to retrieve the data form the datasets

```
hive>
  > select state,count(state) from landslide where country_name="United States" group by state;
```

```
hive> select state,trigger,fatalities from landslide where country_name="United States";
```

```
Aguada	1
Ahuachapán	2
Alabama 9
Alajuela	20
Alta Verapaz	3
Amapá	1
Ancash	5
Antioquia	19
Aragua	1
Arizona 16
Arkansas	6
Artemisa Province	1
Artibonite	3
Atlántico Norte	1
Azuay	2
Baja California 4
Baja California Sur	1
Baja Verapaz	1
Bocas del Toro	7
Bolívar 1
Borough of Arima	1
Boyacá	3
Cabañas 1
Caldas	10
California	57
Caquetá 2
Carchi	2
Cartago 6
Cauca	5
Cayo	1
```

```
Alabama 9
Arizona 16
Arkansas	6
California	57
Colorado	108
Connecticut	6
Florida 2
Georgia 11
Idaho	37
Illinois	9
Indiana 5
Iowa	11
Kansas	2
Kentucky	124
Maine	2
Maryland	8
Massachusetts	10
Michigan	1
Minnesota	23
Mississippi	2
Missouri	9
Montana 1
Nevada	7
New Hampshire	7
New Jersey	11
New Mexico	11
New York	31
North Carolina	52
Ohio	61
Oklahoma	4
Oregon	1
Pennsylvania	97
South Carolina	2
South Dakota	2
Tennessee	39
Texas	4
Utah	65
Vermont 7
```

id	date	time	continent	country_n	country_c	state/prov	populatio	city/town	distance
34	3/2/2007	Night	NA	United Sta	US	Virginia	16000	Cherry Hil	3.40765
42	########		NA	United Sta	US	Ohio	17288	New Phila	3.33522
56	4/6/2007		NA	United Sta	US	Pennsylva	15930	Wilkinsbu	2.91977
59	########		NA	Canada	CA	Quebec	42786	ChÃ¢teau	2.98682
61	########		NA	United Sta	US	Kentucky	6903	Pikeville	5.66542
64	########		NA	United Sta	US	Kentucky	6903	Pikeville	0.23715
67	########		NA	United Sta	US	South Dak	2540	Dakota Du	2.48033
77	########		SA	Colombia	CO	Risaralda	440118	Pereira	0.62022
105	########		SA	Ecuador	EC	Zamora-Cl	15276	Zamora	0.47714
106	########		SA	Ecuador	EC	Loja	117796	Loja	0.35649
107	########		SA	Ecuador	EC	Pichincha	5114	SangolquÃ	33.94603
109	7/1/2007		NA	United Sta	US	Texas	42409	Haltom Ci	0.03668
115	7/4/2007		NA	Mexico	MX	Veracruz-I	1947	Laguna Ch	9.51003
119	7/8/2007		NA	Canada	CA	Ontario	812129	Ottawa	1.74759
124	########	Night	NA	Dominicar	DO	Distrito N	13456	San Carlos	1.70298
138	########		NA	United Sta	US	Texas	175396	Grand Pra	5.66936
165	8/9/2007		NA	Guatemal	GT	Guatemal	47247	San JosÃ©	4.74385
174	########		NA	Jamaica	JM	Portland	14400	Port Anto	7.79027
185	########		NA	United Sta	US	Colorado	2475	Meeker	10.87949

location_c	latitude	longitude	geolocatic	hazard_ty	landslide_	landslide_	trigger	storm_nar	injuries	fatalities	s
Unknown	38.6009	-77.2682	(38.60090(	Landslide	Landslide	Small	Rain				I
	40.5175	-81.4305	(40.51749!	Landslide	Landslide	Small	Rain				(
Urban are	40.4377	-79.916	(40.4377, -	Landslide	Landslide	Small	Rain				T
Above riv	45.3226	-73.7771	(45.32260(	Landslide	Riverbank	Small	Rain				L
Below roa	37.4325	-82.4931	(37.43249!	Landslide	Landslide	Small	Downpour			0	I
	37.4814	-82.5186	(37.48140(	Landslide	Landslide	Small	Rain				/
	42.4941	-96.4576	(42.49410(	Landslide	Landslide	Small	Rain				S
	4.8081	-75.6941	(4.808099!	Landslide	Mudslide	Large	Rain			13	I
	-4.065	-78.951	(-4.06500(	Landslide	Landslide	Medium	Downpour				I
	-3.99	-79.205	(-3.99, -79	Landslide	Landslide	Medium	Downpour				I
	-0.356	-78.148	(-0.355999	Landslide	Landslide	Medium	Downpour				I
	32.7995	-97.2688	(32.79950(	Landslide	Landslide	Medium	Rain				I
	18.5369	-96.8229	(18.53689!	Landslide	Landslide	Medium	Rain			7	r
	45.4257	-75.6896	(45.42569!	Landslide	Landslide	Small	Unknown				
	18.4757	-69.914	(18.4757, -	Landslide	Landslide	Small	Unknown				I
	32.7883	-97.0317	(32.7883, -	Landslide	Landslide	Small	Rain				T
	14.5667	-90.45	(14.56670(	Landslide	Mudslide	Medium	Rain			5	I

```
9994)"  Landslide       Landslide       Small   NULL    NULL
5350    NULL    Afternoon       NA      United States   US      Nevada  24085   N
ULL     23.60286                36.2423 -115.5421       "(36.2423        -115.54
21)"    Landslide       Debris flow     Medium  NULL    NULL            0
5351    NULL    Afternoon       NA      United States   US      Nevada  36441   N
ULL     31.0066         36.2566 -115.6442       "(36.2565999999999       -115.64
42)"    Landslide       Debris flow     Medium  NULL    NULL            0
5385    NULL    23:30:00        NA      United States   US      Utah    9555    N
ULL     1.96154         40.4679 -111.765        "(40.4679        -111.765)"     L
andslide        Mudslide        Medium  NULL    NULL            0
5387    NULL    10:00:00        NA      United States   US      Pennsylvania    1
048     NULL    3.0665          40.1504 -79.4592        "(40.1503999999998
-79.4591999999996)"     Landslide       Mudslide        Medium  NULL    NULL    0
5389    NULL    Afternoon       NA      United States   US      Idaho   54255   N
ULL     0.57038         42.8729 -112.4389       "(42.8729000000001       -112.43
89)"    Landslide       Mudslide        Medium  NULL    NULL            0
5390    NULL    Morning NA      United States   US      Utah    2129    NULL    2
2.93764         39.4714 -111.1546       "(39.4714000000003       -111.1546)"    L
andslide        Debris flow     Medium  NULL    NULL            0
5397    NULL            NA      United States   US      Tennessee       12714   N
ULL     5.71688         35.2865 -85.1789        "(35.2864999999997       -85.178
899999999999)"  Landslide       Mudslide        Small   NULL    NULL            0
5400    NULL    18:00:00        NA      United States   US      California      3
552     NULL    1.82682         34.0588 -116.5671       "(34.0587999999998
-116.5671)"     Landslide       Mudslide        Medium  NULL    NULL            0
5403    NULL    20:20:00        NA      Mexico  MX      Veracruz-Llave  30607   N
ULL     1.52983         20.5004 -97.4647        "(20.5003999999999       -97.464
699999999993)"  Landslide       Landslide       Medium  NULL    NULL            3
5405    NULL            NA      Mexico  MX      Veracruz-Llave  15800   NULL    2
.85382          19.7906 -97.2428        "(19.7906000000001       -97.24280000000
0003)"  Landslide       Landslide       Medium  NULL    NULL            1
5406    NULL            NA      Mexico  MX      Veracruz-Llave  3198    NULL    3
.7316           19.8413 -96.8005        "(19.8413        -96.8005)"     Landslid
e       Landslide       Medium  NULL    NULL            9
5408    NULL            NA      Costa Rica      CR      Alajuela        1015    N
ULL     4.87432         10.1181 -84.2146        "(10.1181        -84.2146000000
```

```
        Mudslide        1
Aguada  Landslide       1
Ahuachapán      Landslide       1
Ahuachapán      Mudslide        1
Alabama Landslide       7
Alabama Mudslide        2
Alajuela        Landslide       14
Alajuela        Mudslide        4
Alajuela        Rockfall        2
Alta Verapaz    Landslide       2
Alta Verapaz    Mudslide        1
Amapá   Landslide       1
Ancash  Complex 1
Ancash  Landslide       2
Ancash  Mudslide        2
Antioquia       Complex 1
Antioquia       Landslide       15
Antioquia       Mudslide        3
Aragua  Landslide       1
Arizona Complex 2
Arizona Debris flow     5
Arizona Landslide       4
Arizona Mudslide        5
Arkansas        Landslide       5
Arkansas        Mudslide        1
Artemisa Province       Landslide       1
Artibonite      Landslide       1
Artibonite      Mudslide        2
Atlántico Norte Complex 1
Azuay   Landslide       1
Azuay   Rockfall        1
Baja California Landslide       2
Baja California Mudslide        2
```

IV. CONCLUSION

Various geographical events like landslides, tsunamis, earthquakes and so on will be taking place around the world and the climate monitoring organisation will be recording the destructions/damages caused. The Analysis made by running hive queries on the datasets provided helps to reduce damages caused to the lives in those regions. We can also predict the future occurrence of such incidents based on the date and time recorded of the previous occurrences.

V. REFERENCES

[1]. Challenges and opportunities with Big Data, http://cra.org/ccc/wpcontent/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

[2]. Oracle: Big Data for Enterprise, June 201http://www.oracle.com/us/products/datab ase/bigdata-for-enterprise-519135.pdf

[3]. Marta C. González, César A. Hidalgo, and Albert-László Barabási. 5 June 2008 Understanding individual human mobility patterns. Nature 453, 779-782.

[4]. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. May 2011 Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[5]. Yuki Noguchi. Nov. 30, 2011 The Search for Analysts to Make Sense of Big Data.. National Public Radio.

[6]. Data set is taken from edureka http://www.edureka.co/my-course/big-data- and-hadoop