Working with employee Datasets using Hive Complex Data types

[1]Dr.Sheikh Ghouse, [2]S.Sravya, [3]G.Charitha, [4]Harish sai
[1]Assoc. Prof., Department of IT, MLRIT, Hyderbad, India
[2, 3, 4]UG Scholar, Department of IT, MLRIT, Hyderbad, India

***Abstract* -** The one of a kind part of this exploration has been the utilization of five prescient information mining systems on a specimen information of 120 representatives in an association. The aftereffects of the investigation obviously demonstrate a relationship of representative turnover. The age and conjugal status rose as key statistic factors. The discoveries of this examination have suggestions for both research and practice. There is a need to grow the extent of this exploration to incorporate various associations and a substantial specimen, which will take into account more powerful forecasts. For specialists, it stresses the requirement for more noteworthy utilization of models and explanatory apparatuses in drawing in with human asset methodologies and plans, and specifically that HR expert should comprehend, acknowledge and apply such models in future to have the capacity to play out their parts as key business accomplices. Record Terms: Data Mining,Employee Turnover, Applications, Algorithm. We have learnt numerous lessons featuring the way that while the effect of representatives with positive personality can without a doubt put an association on the direction of progress, the nearness of workers with contrary mentality can diffuse the officially existing constructive air as well as can end the development motor out and out and push the association into grievously irretrievable state. Henceforth, understanding the psyches of the representatives is of foremost significance for starting proactive strides to manage the development force. The writers have developed a model, clarified in a well ordered instructional exercise way, to peruse and group representatives state of mind.

I. INTRODUCTION

BigData is a term that portrays the vast volume of information both organized and unstructured that immerses a business on an everyday premise. In any case, it's not the measure of information that is imperative. It's what associations do with the information that issues. BigData can be examined for bits of knowledge that prompt better choices and key business moves. 3 characteristics of defining big data is

Volume

Associations gather information from an assortment of sources, including business exchanges, online networking and data from sensor or machine-to-machine information. Previously, putting away it would've been an issue – yet new advances, (for example, Hadoop) have facilitated the weight.

Velocity

Information streams in at an extraordinary speed and should be managed in a convenient way. RFID labels, sensors and shrewd metering are driving the need to manage downpours of information in close ongoing.

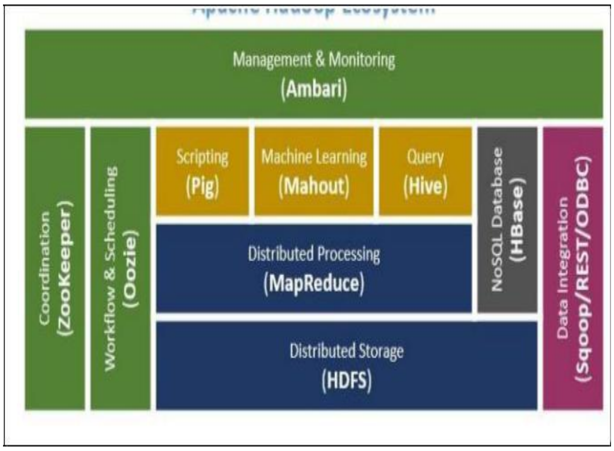
Variety

Information comes in a wide range of configurations – from organized, numeric information in customary databases to unstructured content reports, email, video, sound, stock ticker information and budgetary exchange.

HADOOP

Apache Hadoop has been the main thrust behind the development of the huge information industry. You'll hear it specified regularly, alongside related advances, for example, Hive and pig. Hadoop brings the capacity to inexpensively process a lot of information, paying little heed to its structure. By expansive, we mean from 10-100 gigabytes or more.

Existing endeavor information distribution centers and social databases exceed pectations at handling organized information and can store monstrous measures of information, however at a cost: This necessity or structure confines the sorts of information that can be prepared, and it forces an inactivity that makes information stockrooms unsuited for spry investigation of gigantic heterogeneous information. The measure of exertion required to distribution center information frequently implies that significant information sources in associations are never mined.



Map Reduce programming is not a good match for all problems. It's useful for straightforward data solicitations and issues that can be isolated into autonomous units, however it's not productive for iterative and intelligent diagnostic errands. Guide Reduce is record concentrated.

Since the hubs don't intercommunicate with the exception of through sorts and rearranges, iterative calculations require

II. RELATED WORK

Hive

It was initially introduced by face book in the year 2007 in order to full fill requirements with respect to ETL jobs. Later it became Hadoop sub project. Hive is a data warehousing frame work in Hadoop where we store data in the form of tables (structured format).Hive runs on the top of hdfs and MapReduce.
The back end storage for hive is hdfs and executing model is MapReduce. so it allows user to query data in Hadoop cluster without knowing java or MapReduce. Hive is designed to enable Easy data summarization

1) Ad-hoc querying

2) Analysis of large volume of data.

Hive provides SQL like language called Hive (HQL). HQL is very similar to SQL.

Hive is designed for scalability and easy of use. When you submit a hive query (hql), it is converted into MapReduce program, converted code will be submitted to jvm for execution, because Hadoop runs on jvm. Hive can process structured data, xml,json,urls(major content of web logs). numerous guide rearrange/sort-decrease stages to finish. This makes different records between Map Reduce stages and is wasteful for cutting edge diagnostic registering. But hive is weak for unstructured text data process. But hql is not for operating rows randomly. Such as reading randomly, insert /update/delete records randomly. In older versions (such as 0.7.*,0.8.*) of hive ,hive does not support update and delete operations. But latest versions (0.14.*..), hive supports update and delete.

Different types of modes to access hive

1) Command line interface

2) Web interface

3) Thrift server

UI – The user interface for users to submit queries and other operations to the system.

Driver –Receive the query from the UI, creates a session for the query and sends the query to the compiler to generate an execution plan..

Compiler – it does semantic analysis on the different query blocks and query expressions and generates an execution plan by getting necessary metadata from the metastore

Execution Engine – The component which executes the execution plan created by the compiler. The plan is a DAG of stages. The execution engine manages the dependencies between these different stages of the plan and executes these stages on the appropriate system components.

Meta Store

It is a service runs on same jvm in which hive started. it is used to manage hive metadata stored in Derby. Hive comes with embedded database called Derby. By default all hive tables metadata will be stored in Derby.

Meta-store provides schema information of the tables, tables location, data partitiones ....

Note: meta-store is key role in hive architecture.

If Meta data service in not running, hive can't process data.

We can configure other databases(MySQL or Oracle....) as a metastore of hive. this

configuration is done in the hive configuration file called hive-site.xml.

Hive Data Model:
When you create hive table, in hdfs, with table name one directory will be created.

When we load a file into table, the file will be copied into tables back end hdfs directory.This means, hive gives the table shape (structured shape) for the hdfs file.

Hive complex data types
Collection data types:-

1) Array -->collection of elements

2) Map --> collection of key& value pairs

3) STRUCT --> collection of attributes with different data types.

Array

The first complex type is an array. It is nothing but a collection of items of similar data type. i.e, an array can contain one or more values of the same data type.

Map

Map is a collection of key-value pairs where fiels are accessed using array notation of keys.

Struct

Struct is a record type which encapsulates a set of named fields that can be any primitive data type. An element in STRUCT type can be accessed using the DOT (.) notation.

III. PROPOSED SYSTEM

The proposed system requires integrating systems for employee datasets, Client management and Project management at one place. It makes data manipulation of projects & employees easy and fast. Its Less time consuming and provide efficient searching.

Complex Data Types for Employee

Step1 Employee DataSet:

Sravya nikki, venni, prasanna, sweetu, geethu personal: 123456789, Offical: 111222333 IBM,5608,no,90000.0

Charitha manasa, priya, niveda, nikila, ravali personal: 123456789, Offical: 111222333 Tech, 5608, no, 19856.0

Harish vinnu,sumanth,aditya,nikil,ram personal:123456789,Offical:111222333 MLR,5608,yes,60000.0

Step2 Create a table Employee:

Create table Sravya.emp1(name string,friends array<string>, mobile map<string,bigint>,

others

struct<Company:string,Pincode:int,Married:

string,Salary:float>) row format delimited fields terminated by '\t' collection items terminated by ',' map keys terminated by ':' lines terminated by '\n' stored as textfile;

Step3 Run

>hive –f employee.hql

Step4 Enter into Hive Terminal

>hive

Step5 Choose Database

hive> use Sravya;

OK

Time taken: 1.377 seconds

Step6 Check the table

hive> show tables**;**

OK

employee

Time taken: 0.126 seconds, Fetched: 9 row(s) **Step7** Load the data into table

hive> load data local inpath '/home/Sravya/work/hive/empdata.txt' into table employee;

Loading data to table Sravya.emp1

Table Sravya.emp1 status: [numFiles=1, totalSize=269]

OK

Time taken: 0.756 seconds

Step8 Query

hive> select name from employee;

OK

Sravya

Charitha

Harish

Time taken: 0.349 seconds, Fetched: 3 row(s)

hive> select name, mobile['personal'] from employee;

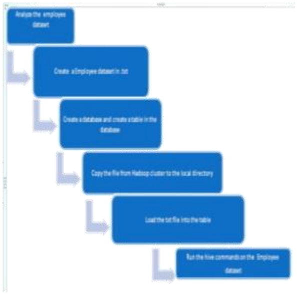
OK

Sravya 123456789

Charitha 123456789

Harish 123456789

Time taken: 12.985 seconds, Fetched: 3 row(s).


Analyze the employee dataset
Create a Employee dataset in .txt
Create a database and create a table in the database
Copy the file from Hadoop cluster to the local directory
Load the txt file into the table
Run the hive commands on the Employee dataset

IV. CONCLUSION

Hive is a Data Warehousing package built on top of Hadoop used for structure and semi structured data analysis and processing. It provides flexible query language such as HQL for better querying and processing of data. Therefore, we conclude that this project can be helpful for predicting the Employee

V. REFERENCES

[1]. Anand Bahety Department of Computer Science University of Maryland, College Park “Extension and Evaluation of ID3 – Decision Tree Algorithm”

[2]. Brijesh Kumar, Baradwaj, Saurabh Pal, “Mining Educational Data to Analyze Students Performance “ ,” IJACSA, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[3]. Ying Liu, et all , "Region-based image retrieval with high-level semantics using decision tree learning," Journal of Pattern Recognition, Vol. 41, No. 8, pp. 2554 – 2570, Aug 2008.

[4]. Breiman, Friedman, Olshen, and Stone. ― Classification and Regression Trees, Wadsworth, Mezzovico, Switzerland. 1984

[5]. Daniel Rodríguez “Making predictions on new data using Weka” University of Alcala.

[6]. Matthew N.Anyanwu, Sajjan G.Shiva, ―Comparative Analysis of Serial Decision Tree Classification Algorithms, International Journal of Computer Science and Security, volume 3.

[7]. Mitra S, Acharya T. Data Mining.Multimedia, Soft Computing, and Bioinformatics. John Wiley & Sons, Inc., Hoboken, New Jersey; 2003.

[8]. Parr Rud, O. Data Mining Cookbook.Modeling Data for Marketing, Risk, and Customer Relationship Management. John Wiley & Sons, Inc.; 2001.

[9]. Quinlan, J.R. Induction of decision trees. Machine Learning, volume 1. Morgan Kaufmann; 1876. p. 71-96.

[10]. Quinlan, J.R., (1883), C4.5:Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann

[11]. S.Anupama Kumar and Dr. M.N.Vijayalakshmi “Efficiency of Decision