# Big Graph Data Management and Analytics

**R.Venkateswara Gandhi[a], Ayesha Naureen[b], M.Seshiah[c], Manjunath[d]**
[a,b,c,d]Research Scholar, Vtu, Belagavi,

## Abstract

Big data can be well visualized using graphs. Graphical analytics are the key for analyzing many big data applications. the management and analysis of huge amounts of graph data. Previous approaches for graph analytics is necessary in many business applications. The management of such applications can be done by graphical analysis such as graph databases and parallel graph processing systems .But due to lack of either sufficient scalability or flexibility and expressiveness big data is becoming complex. A new end-to-end approach for graph data management and analysis at the Big Data center of excellence ScaDS Dresden/Leipzig, called Gradoop (Graph analytics on Hadoop) is developed. Gradoop is designed around the so-called Extended Property Graph Data Model (EPGM) which supports semantically rich, schema free graph data within many distinct graphs. A set of high-level operators is provided for analyzing both single graphs and sets of graphs. The operators are usable within a domain specific language to define and run data integration workflows as well as analysis workflows. These are used for integrating heterogeneous source data into the Gradoop graph store. The Gradoop data store is currently utilizing HBase for distributed storage of graph data in Hadoop clusters. An initial version of Gradoop is operational and has been used for analyzing graph data for business intelligence and social network analysis.

**KEYWORDS:** Bigdata,EPGM,HDFS,Clustering

## I. INTRODUCTION

Big data analytics refers to large amounts of data to discover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze data and get answers from it as soon as possible. Pattern finding is a prominent way for data mining. Data predictions, decision making are done by data architects in big data focusing data decisions. Many methods are followed including pattern matching, deep learning help for predictive analysis. Though data can be collected easily, the concept managing the data for futuristic use is a major challenege.Many machine algorithms are being used for big data applications. The concept of analyzing complex data is very difficult. When big data becomes complex data the above mentioned methods are used for data analysis. Such complex data can be easily analyzed when the data take the form of graphs. The parameters present in the necessary data are to be analyzed .This leads to understanding of relationship between the parameters used. Thus we discuss the concept of

## II. BIG GRAPH DATA.

Graph Analytics is not limited to any data. Any data can be represented in a graph format. Graphical format is beneficial for representing numerical data. The social media is such sort of data where the data is represented in dynamic graphs. Relationship analysis is studied by assuming the computing paths among vertices[1].A sub graph pattern search problem are analyzed. Graph analytics involves google maps, identification of chemical

structures, social media and many.

## III. GRAPH ANALYTICS

The relationship between the data are represented by nodes and edges in the graph.Every vertice and edge are being labeled so as to the relation between every data item can be identified. Whenever we say simply "graph", we mean either directed or undirected graph; when it is important to distinguish whether the graph is directed or undirected[4]. Graph mining is also done using this labels.Graph mining is a concept of discovering the relation between data in each and every gap.This leads to proper predictive analysis.the concept of data analytics is being done by graph analysis.Multiple directed edges can be drawn between any two vertices,so long as they are differentially labeled.A labeled multidigraph can be defined as a four tuple represented as

G(V,E,L,l)

Where

V is the set of vertices

EcV*V*L (the set of labeled vertices) (1)

L is the set of labels

l: V->L(vertex labeling function)

A set of edges describe the connection between vertices.Such multigraph becomes a digraph when edge labels are not considered as Ec V*V.For a digraph uv € E represents a directed edge from vertex u to v which isalso applicable in the case of digraphSuch relationships can be easily understood in the case of parent and child relationship defined as

Child(u)= {v: uv € G.E}

Parent(u)={w: wu € G.E}

Thus every relation in the graph can analysed easily.The graph is traversed in a way called path. Finding paths, patterns or partitions in very large data graphs are major problems in graph analytics. Such situations like graphs with a billion edges. These problems are strongly interrelated. A path is considered as a simple linear pattern and partitioning is needed for both path and pattern problems, when graphs become too large to store or process on a single machine or single thread.

Some hurdles regarding paths involve the following

### 3.1 Reachability:

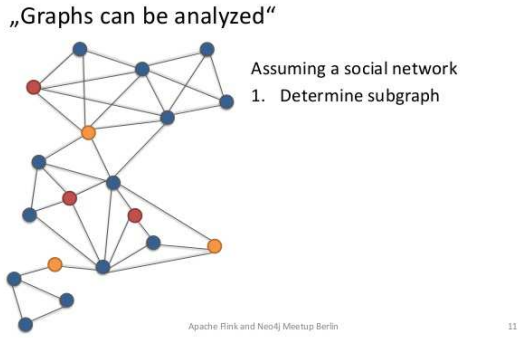
Rechability is the problem of disability to reach a vertex from another within a graph.A vertex can reach another vertex if there exist a sequence of adjacent The connected components of the graph are to identified in a undirected graph to overcome such problem[2]. It is also possible to define backward versions of the reachability graph, in the same way that backward reachable sets can be obtained. Any pair of vertices in such a graph can reach each other if and only if they belong to the same connected component. The connected components of an undirected graph can be identified in linear time. A path can be defined as

Path(u,w) = $uv_1l_i, v_1v_2l_{i2....}v_nwl_{in+1}$…. € G,E

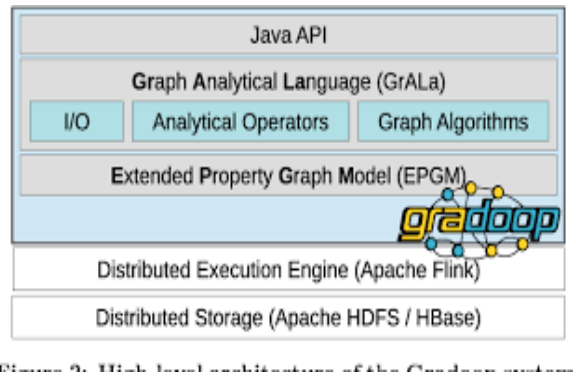
This can be generalized to return all paths between u and v.
a-paths(u,w)={p:p=path(u,w)}
reachability can be made simple as
reach(u,w)= ∃path(u,w).
Reachability analysis has applications in many domains including XML indexing and querying,home land security,navigation in road and root causes analysis in large scale based distributed systems.

3.2 Shortest Path:
A shortest path from vertex s to vertex t is a directed path from s to t with the property that no other such path has a lower weight[3].A cumulative edge weight is a solution to find the minimum distance path including all K vertices in the path.For a digraph i(e),the edge labels represent an edge weight.These weights help to find the minimal path and best way for traversing a graph.
For k = 3, three applications of Dijkstra's Algorithm (or equivalent) will suffice to find the short path connecting all three vertices. The all-pairs short path problem [5] is also of interest in Big Data Analytics.

3.3 Reachability graph algorithm
1) Label the initial marking m0 as the root and tag it "new".
2) While "new" markings exists, do the following:
a) Select a new marking m.
b) If no transitions are enabled at m, tag m "dead-end".
c) While there exist enabled transitions at m, do the following for each enabled transition t at m: that results from firing t at m:
that results from firing t at m.′i. Obtain the marking m .
ii. If m' doesnot appear in the graph add m'and tag it "new".
iii. Draw an arc with label t from m to m' .
3) Output the graph
For analyzing such graph data i.e.big graph data many tools are available.The hdfs and mapreduce techniques make the big data simple to analyse and when data is being converted to graph data its is necessary that some additions tools are used in the hadoop platform.One of them is such called the gradoop.when graphical data is analysed using hadoop,it is named as gradoop.the graphical analysis using hadoop is gradoop.The objectives of gradoop are discussed below.
IV. GRADOOP:
When data is connected as graphs it becomes more and more important in many different domains. Processing highly connected graphs are a challenge. social networks stand examples for such domains.e.g. facebook and Twitter, networks like the World Wide Web or biological networks.




One important similarity of these domain specific data is their inherent graph structure which makes them eligible for analytics using graph algorithms. Such data are huge in size, which result in making it hard or even impossible to process them on a single machine.As they grow over time, classifies them as dynamic graphs become complicated reaching the complex data category. With the objective of analyzing these large-scale, dynamic datasets, “Gradoop” (Graph Analytics on Hadoop) was started[6] . Gradoop is designed around the so-called Extended Property Graph Data Model (EPGM) is one of the main property of gradoop which is semantically rich, schema-free graph data within many distinct graphs[7].Gradoop has mainly these objectivesdefinition of analytical pipelines, developing a graph data model including operators.Heterogeneous source systems are integrated into graphs by data integration. Optimize the execution of distributed graph operators, Distribute and replication.




The gradoop is build on top of the distributed dataflow framework . The data model has been designed and the operators have been implemented. A first use case is the BIIIG project for graph analytics in business information networks[8]. Business Intelligence with Integrated Instance Graphs (BIIIG) [9].

## V. CONCLUSION:

Analysis of big data is an outstanding challenge.Data management and architecture design for big data analysis is to be designed well so that data can be utilized well.Data can be well understood and analysed in the form of graphs.The big graph data is to be managed and manipulated well for predictive analysis.Gradoop is one of the best tool for maintining graphical analysis.

VI. REFERENCES

[1] W. Fan, J. Li, S. Ma, N. Tang, Y. Wu, and Y. Wu, “Graph pattern matching: from intractable to polynomial time,” Proc.VLDB Endow., vol. 3, no. 1-2, pp. 264–275, Sep 2010.
[2] http://planning.cs.uiuc.edu/node735.html
[3] https://algs4.cs.princeton.edu/44sp
[4] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.309&rep=rep1&type=pdf
[5] D. Z. Ghent, “On the all-pairs Euclidean short path problem,”in Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms, vol. 76. SIAM, 1995, p. 292.
[6] www.gradoop.com
[7] https://www.researchgate.net/publication/277723414_GRADOOP_Scalable_Graph_Data_Management_and_Analytics_with_Hadoop.
[8] https://dbs.unileipzig.de/en/research/projects/gradoop.
[9] https://dbs.uni-leipzig.de/file/15_07_28_Gradoop_BIIIG_Dresden.pdf.