# INTERNATIONAL JOURNAL
## FOR RESEARCH
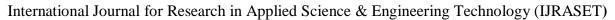
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

# Review on Big Data and Mining Algorithm

Shoban Babu Sriramoju[1]

[1]*Department of Computer Engineering, SR Engineering College, Warangal, India*

*Abstract*: *An information stream is a requested arrangement of examples that can be perused just once or a little number of times utilizing constrained processing and stockpiling abilities. Numerous applications take a shot at stream information like web pursuit, system activity, phone call and so on. In this application information is persistently changing in view of time. In this paper we will examine future patterns of information digging that are utilized for examination and expectation of big data. We will talk about difficulties while performing mining on big information. Stream information is likewise alluding as constant information. Constant information produced through web, every second a huge number of information created, so how to oversee and dissect this information, we examine in this paper.*
*Keywords*: *Network Traffic, Computation, Stream Data.*

## I.  INTRODUCTION

In these days the quantity of data created per second is very large. Data stream real time analysis is required to manage this large data, through proper analysis we can get crucial data, through this we can predict network traffic, intrusion related activity, weather. log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others [1].data generated from stream is just snapshot of stream data. Snapshot is based on time interval. Stream data mining algorithm processing based time and space constraint. The main of algorithms is usage of resources (resources can be Memory and time).In stream database, to perform stream mining we have to consider ,accuracy, amount of space, time required to learn from training examples for getting prediction. Data is large, and growing, there are important patterns and trends in the data. We don't fully know where to look or how to find them. Big data analysis is most important because the data is continuously changing based on interval of time to store big data most of companies are using cloud setup.

## II.  METHODS AND MATERIAL

### A.  Characteristics of Big Data

Big data can be described by following characteristics,

1) *Volume:* The quantity of data generated is very important in this context.it is the size of data which determines value and potential of data under consideration and whether it is can actually considered as big data or not.

2) *Variety:* The next aspect og big data is variety. means that the category to which big data belongs to is also a very essential fact that need to be known by data analyst. this information is very important for people, those will analyze this data. Through this they will understand importance of big data.

3) *Velocity:* Velocity refers to speed of generation of data and how fast data is generated and process to meet the demands of business.

4) *Variability:* Because of this characteristics, analysis of data is very difficult. This refers to inconsistency can be present in data at time .due to this the process required for analysis should manage and handle big data effectively

5) *Complexity:* The data collected from different sources, this data needs to be linked, collected and correlated with each other so throgh this data we will able to derive information that data wants to represent.[1]The major sources of big data are from the following

### B.  Archives

Archives are mainly maintained by organizations, to show the function of a particular person or organization functions. Accumulation of archives sometimes does not fit into the traditional storage systems and need systems with high processing capabilities. This voluminous archive contributes to big data.

### III.MEDIA

Users generate images, videos, audios, live streams; podcasts and so on contributes for big data.

## IV. BUSINESS APPLICATIONS

Huge volumes of data are generated from business applications as part of project management, marketing automation, productivity, customer relation management (CRM), enterprise resource planning (ERP) content management systems, procurement, human resource (HR), storage, talent management, Google Docs, intranets, portals and so on. These data contributes to big data.

## V. PUBLIC WEB

Many organizations under government sector, weather, competitive, traffic regulatory compliance, health care services, economic, census, public finance, stock, open source intelligence (OSINT), the world bank, electronic data gathering analysis and retrieval (Edgar), Wikipedia and so on uses web services for communication. These data contributes to big data

## VI. SOCIAL MEDIA

Nowadays users rely on social media sites such as twitter, linkedln, facebook, tumblr, blog, slideshare, youtube, google+, instagram, flickr, pinterest, vimeo, word press and so on for the creation and exchange of user generated contents. These social networking sites contribute to big data.

## VII. DATA STORAGE

Data storage in SQL, NoSQL, Hadoop, doc repository, file systems and so on also contributes to big data.

## VIII. SENSOR DATA

Accumulation of large quantitative data sets from distributed sensors are now becoming widely available online from medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery and so on. This contributes to big data [2].

## IX. EXAMPLES OF BIG DATA

The characteristics of Big Data are the huge volume of data represented by diverse dimensionality. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. Example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, blood group, other medical check-up reports.. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic- related test, micro- array expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data.

Imagine that different organizations of doctors may have their own schema to represent each patient, the data heterogeneity and diverse dimensionality issues become major. Other examples of big data are Calls, text, tweet, net surf, browse through various web sites each day and exchange messages via several means. Social media usage, several people use a social media for exchanging the information via several forms, also forms a part of big data. Transactions made through cards for various payment issues in large numbers every second across the world also constitute the big data. These application are facing the problems at the time of analysis.[2]

When data collected from web it comes from many sources. most of the web sites are maintain data on single sever. sometimes it is not possible to maintain all data at single server. If data is collected on single server, it does not rely on other server, but here the main issue is security, enormous amount of data is venerable to attack and malfunction.so the copy of server should be maintain on other server, there should be replicas present to serve services, if main server goes down. for example facebook, Google, twitter gives the nonstop services to their customers.[7]

Facebook, twitter are social media through this people can communicate, but when user is registering for these web sites they has to enter details such as gender, address, education, mobile number, hobbies, e-mail address and so on. These data fields are used to categorize each individual. These data fields represents individual as a separate entity. In daily life, generally friend circles are formed based on similar hobbies and through biological relation. The same concept for forming the friend is applied in cyberworld.as Facebook and twitter mainly characterized by social function such as friends connections and followers in

twitter.[2]

## X. BIG DATA (STREAM DATA) OPEN SOURCE SOFTWARE

Big Data is general term that is used to refer data; only for large size of data we called it as big data. The main issue is our normal data mining algorithm will not able to manage this big data.

There are two main strategies for dealing with big data: sampling and using distributed systems. In sampling, we obtain a smaller data set with the same structure. Sampling is based in the fact that if the dataset is too large and we cannot use all the examples, we can obtain an approximate solution using a subset of the examples. A good sampling method will try to select the best instances, to have a good performance using a small quantity of memory and time.[3]

Different sampling techniques are available to sample stream data. An alternative to sampling is the use of probabilistic techniques. Facebook uses friendship links to compute distance between two nodes through distance. Millions of user accessing a face book but they managed to compute the average distance between two users on Facebook only using one machine. when we are operating on big data that time memory required to store data is main constraint but, to store the big data does not required big machine, it needs big intelligence‖. A MapReduce program is composed of a Map() procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() procedure that deals with a summary operation such as counting the number of students in each queue, yielding name frequencies. MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed, algorithm on a cluster. The map-reduce methodology started in Google. Hadoop is a open-source implementation of map-reduce started in Yahoo! Then twitter, eBay. Big data phenomenon is intrinsically related to the open source software. Large companies as Facebook, Yahoo!, Twitter, and LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop. Hadoop is a software framework for processing and storing big data which accomplishes two tasks: massive data storage and faster processing.

Traditionally, open source software is free to download, use and contribute to, though more and more commercial versions of Hadoop are becoming available.

VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning and also uses different optimization technique.

### A. Moa

MOA( Massive On-Line Analysis) is a framework for data stream mining. It includes tools for evelution and collection of machine learning algorithm. It has implementation of classification, regression, Clustering, frequent pattern mining and frequent graph mining. Related to the WEKA project it also implemented in JAVA. It includes a collection of offline and online as well as tools for evaluation: classification and clustering. Easy to design, extend and run experiments. The goal of MOA framework for running experiments in data stream min ing context by proving, Storable setting for data streams for repeatable experiments, set of existing algorithm and measure from literature for comparison, an easily extendable framework for new streams, algorithms and evaluation methods.

### B. Appache Mahout

Scalable machine learning and data mining open source software based mainly in Hadoop. More specific to Big Graph mining we found the following open source tools:Graph Lab: High-level, Graph-parallel system built without using Map Reduce. Graph Lab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in Graph Lab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices. [3]ramework. In this case, it means everything you need to develop and run your software applications is provided programs, tool sets, connections, etc.The Hadoop framework can store huge amounts of data by splitting the data into blocks and storing it on clusters of lower cost commodity hardware.Hadoop will process huge amount of data in parallel across clusters of tightly connected low cost systems for quick results.

### C. Big Data Analysis Platforms and Tools

There are many open source software for big data analysis,Apache pig: A platform for processing and analyzing large data sets. It consists of a high-level language (Pig Latin) for expressing data analysis programs paired with the Map Reduce framework for processing these programs.

1) *Apache HBase:* A column-oriented NoSQL data storage system that provides random real-time read/write access to big data

for user applications. Non- relational columnar distributed database designed to run on top of Hadoop Distributed File system (HDFS). It is written in Java and modeled after Google's BigTable.

2) *Apache S4:* platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.

### D. Big Data Mining Open Source Software

R is a language and environment for statistical computing and graphics. Vowpal Wabbit: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning and also uses different optimization technique. Massive Online Analysis is the framework for data stream mining. It contains different tools for evaluation and collection of machine learning algorithm. It has implementation of classification, regression, Clustering, frequent pattern mining and frequent graph mining. Related to the WEKA project it also implemented in JAVA. It includes a collection of offline and online as well as tools for evaluation: classification and clustering. Easy to design, extend and run experiments. The goal of MOA framework for running experiments in data stream min ing context by proving, Storable setting for data streams for repeatable experiments, set of existing algorithm and measure from literature for comparison, an easily extendable framework for new streams, algorithms and evaluation methods. Appache Mahout : Scalable machine learning and data mining open source software based mainly in Hadoop. More specific to Big Graph mining we found the following open source tools:

## XI. BIG DATA MINING ALGORITHM

### A. Decision tree induction classification algorithms

In the initial stage different Decision Tree Learning was used to analyze the big data. In decision tree induction algorithms, tree structure has been widely used to represent classification models. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classifiers break a complex decision into collection of simpler decision. Hall. et al. [8] proposed learning rules for a large set of training data.

The work proposed by Hall et al generated a single decision system from a large and independent subset of data. An efficient decision tree algorithm based on rainforest frame work was developed for classifying large data set [6].

### B. Evolutionary based classification algorithms

Evolutionary algorithms use domain independent technique to explore large spaces finding consistently good optimization solutions. There are different types of evolutionary algorithms such as genetic algorithms, genetic,programming evolution strategies, evolutionary programming and so on. Among these, genetic algorithms were mostly used for mining classification rules in large data sets [3].

### C. Partitioning based clustering algorithms

In partitioning based algorithms, the large data sets are divided into a number of partitions, where each partition represents a cluster. K-means is one such partitioning based method to divide large data sets into number of clusters. Fuzzy- CMeans is a partition based clustering algorithm based on Kmeans to divide big data into several clusters[1]

### D. Hierarchical based clustering algorithms

In hierarchical based algorithms large data are organized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters. Birch algorithm is one such algorithm based on hierarchical clustering.

To handle streaming data in real time, a novel algorithm for extracting semantic content were defined in Hierarchical clustering for concept mining. This algorithm was designed to be implemented in hardware, to handle data at very high rates. After that the techniques of self-organizing feature map (SOM) networks and learning vector quantization (LVQ) networks were discussed in Hierarchical Artificial Neural Networks for Recognizing High Similar Large Data Sets . SOM consumes input in an unsupervised manner whereas LVQ in supervised manner. It subdivides large data sets into smaller ones thus improving the overall computation time needed to process the large data set.

### E. Density based clustering algorithms

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A

cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clustering.

### F. Grid based clustering algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning.

### G. Model based clustering algorithms

In model based clustering algorithms clustering is mainly performed by probability distribution. Expectation- Maximization is one such model based algorithm to estimate the maximum likelihood parameters of statistical models. In 2013, a new algorithm called scalable Visual Assessment of Tendency‖ (sVAT) algorithm was developed to provide high scalable clustering in big data sets. Afterwards a distributed ensemble classifier algorithm was developed in the field based on the popular Random Forests for big data. This proposed algorithm makes use of Map Reduce for improving the efficiency and stochastic aware random forests for reducing randomness. Later in the field, a mixed visual or numerical clustering algorithm for big data called Clusi VAT was developed to provide fast clustering.[3,5]

## XII. RESULTS AND DISCUSSION

### A. Challenges with Big Data

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Now Big Data analysis drives nearly every aspect of our modern society including financial services, mobile services, retail, manufacturing, life sciences, and physical sciences. As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. Moreover, the global mining can be featured with a two-step (global correlation and local mining) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. Through exchange of patterns between multiple sources, a new global patterns can be synthesized by aggregating patterns across all sites. The model correlation analysis investigates the difference between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.[4]The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature [6]. Documents posted on communication networks, Internet backbones, WWW servers, social networks, transportation networks and so on are all featured with complex data. For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly real time speed and very high accuracy. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. This connection is quadratic associated to number of nodes in a network, so a million node network may be subject to one trillion connections. For a large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an huge network is a big challenge for Big Data mining. If daily user actions/interactions are considered, the scale of difficulty will be even more astonishing.

### B. Complex Heterogeneous Data Types

In Big Data, data types include structured data, unstructured data, and semi structured data, and so on. Specifically, there are tabular data (relational databases), text, hyper-text, image, audio and video data, and so on.

## XIII. CONCLUSION

Analysis of large volumes of data is required to create business intelligence. Examination is obliged to improve development in business and experimental exploration. In this paper we are confronting numerous specialized difficulties for investigation of huge volume of information. The difficulties incorporates clear issues of scale, additionally heterogeneity, absence of structure, blunder taking care of, security, convenience, provenance, and representation, at all phases of the investigation pipeline from information procurement to result understanding. These specialized difficulties are basic over an expansive mixed bag of utilization areas, and in this way not practical to address in the setting of one space alone. Clear Different arrangements are obliged to investigate the huge information.

## REFERENCES

[1] Xindong Wu,Xingquan Zhu,Wei Ding,‖Data Mining with Big Data‖,IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014

[2] Albert Bifet‖Mining Big Data in Real Time‖Informatica 37 (2013) 15–20

[3] Nawsher Khan , Ibrar Yaqoob,1 Ibrahim Abaker Targio Hashem,1 Zakira Inayat,‖Big Data: Survey, Technologies, Opportunities, and Challenges‖The Scientific World Journal Volume 2014  (2014),
Article ID 712826, 18 pages

[4] Challenges and Opportunities with Big Data‖A community white paper developed by leading researchers across the United States

[5] Sherin A, Dr S Uma, Sarnya K, Sarnya Vani M, ‖Survey On Big Data Mining Platforms, Algorithms

And Challenges‖, International Journal of Computer Science & Engineering Technology (IJCSET)

[6] Thakur, "Data Mining for Big Data A Review" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014 ISSN: 2277 128X

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1{135, 2008.

[8] M. Helft, Google Uses Searches to Track Flu's Spread, The New York Times http://www.nytimes.com/2008/11/12/technology/ internet/12flu.html. 2008.

[9]  Babu, Sriramoju Ajay and Babu, S Shoban. "International Journal of Research and Applications Jan-Mar© 2016 Transactions 3 (9): 422-426 eISSN: 2349–0020." International Journal of Research In Science & Engineering e-ISSN: 2394-8299 Volume: 3 Issue: 4 July-August 2017 p-ISSN: 2394-8280 IJRISE JOURNAL| www.ijrise.org|editor@ijrise.org [56-65]

[10]  Sriramoju, Ajay Babu. "Analysis on Image Compression Using Bit-Plane Separation Method" (2014).
 Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays For Image Processing" (2013).

[11] Dr. Shoban Babu Sriramoju, Prof. Mangesh Ingle, Prof. Ashish Mahalle "Trust and Iterative Filtering Approaches for Secure Data Collection in Wireless Sensor Networks" in "International Journal of Research in Science and Engineering" Vol-3, Issue-4,July-August 2017 [ ISSN : 2394-8299 ].

[12] Dr. Shoban Babu, Prof. Mangesh Ingle, Prof. Ashish Mahalle "HLA Based solution for Packet Loss Detection in Mobile Ad Hoc Networks" in "International Journal of Research in Science and Engineering" Vol-3, Issue-4,July-August 2017 [ ISSN : 2394-8299 ].

[13] Shoban Babu Sriramoju.  "A Framework for Keyword Based Query and Response System for Web Based Expert Search" in "International Journal of Science and Research"  Index Copernicus Value(2015):78.96 [ ISSN : 2319-7064 ].

[14] Sriramoju Ajay Babu, Dr. S. Shoban Babu. "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications" Vol-1, Issue-1,Jan-Mar 2014 [ ISSN : 2349-0020 ].

[15] Dr. Shoban Babu Sriramoju, Ramesh Gadde. "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol-1, Issue-1,Jan-Mar 2014 [ ISSN : 2349-0020 ].

[16] Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju. "Risk Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management"
Vol-VI, Issue-I, Feb'2014 [ ISSN : 2249-4510 ]

[17] Mounica Doosetty, Keerthi Kodakandla, Ashok R, Shoban Babu Sriramoju. "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management" Vol-VI, Issue-I, Feb'2012 [ ISSN : 2249-4510 ]

[18] Shoban Babu Sriramoju. "An Application for Annotating Web Search Results" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2,Issue-3,March'2014  [ ISSN(online) : 2320-9801, ISSN(print) : 2320-9798 ]

[19] Shoban Babu Sriramoju. "Multi View Point Measure for Achieving Highest Intra-Cluster Similarity" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2,Issue-3,March'2014  [ ISSN(online) : 2320-9801, ISSN(print) : 2320-9798 ]

[20] Shoban Babu Sriramoju, Madan Kumar Chandran. "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering" Vol-4,Issue-2,February'2014  [ ISSN : 2277 128X ]

[21] Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao. "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications"  Vol-V,Issue-II,Febrauary'2014  [ ISSN : 2321-3469 ]

[22] Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Analysis in Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management" Vol-VII, Issue-X, Nov' 2014 [ ISSN : 2249-4510 ]

[23] Shoban Babu Sriramoju. "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering"  Vol-2, Issue-1,January'2014 [ ISSN(online) : 2320-9801, ISSN(print) : 2320-9798 ]

[24] Mr. Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Objective Quality Metric Design for Wireless Image and Video Communication" in "International Journal of Information Technology and Management" Vol-VII, Issue-I0, Nov' 2014 [ ISSN : 2249-4510 ]

[25] Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Procesing" in "International Journal of Information Technology and Management" Vol-V, Issue-I, Aug' 2013 [ ISSN : 2249-4510 ]

[26] Dr. Shoban Babu Sriramoju. "A Review on Processing Big Data" in "International Journal of Innovative Research in Computer and Communication Engineering"  Vol-2, Issue-1, January'2014 [ ISSN(online) : 2320-9801, ISSN(print) : 2320-9798 ]

[27] Shoban Babu Sriramoju, Dr. Atul Kumar. "An Analysis around the study of Distributed Data Mining Method in the Grid Environment : Technique, Algorithms and Services" in "Journal of Advances in Science and Technology"  Vol-IV, Issue No-VII, November'2012  [ ISSN : 2230-9659 ]

[28] Shoban Babu Sriramoju, Dr. Atul Kumar. "An Analysis on Effective, Precise and Privacy Preserving Data Mining Association Rules with Partitioning on Distributed Databases" in "International Journal of Information Technology and management" Vol-III, Issue-I, August'2012 [ ISSN : 2249-4510 ]

[29] Shban Babu Sriramoju, Dr. Atul Kumar. "A Competent Strategy Regarding Relationship of Rule Mining on Distributed Database Algorithm" in "Journal of Advances in Science and Technology"  Vol-II, Issue No-II,November'2011 [ ISSN : 2230-9659 ]

[30] Shoban Babu Sriramoju, Dr. Atul Kumar. "Allocated Greater Order Organization of Rule Mining utilizing Information Produced Through Textual facts" in "International Journal of Information Technology and management"

# INTERNATIONAL JOURNAL
# FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089   ◎ (24*7 Support on Whatsapp)