Mobile Cloud Computing & Big Data Social Network

Adnan Majeed

(Master of Computer Science) Virtual University of Pakistan Lahore. Teaching Assistant Beacon house National University Tarogil Lahore, Lecturer Computer Learning Center Ali Park Lahore, adnanmajeed82@gmail.com +923004870104.

Abstract- **Mobile cloud computing is hot research topic in these days, social network data increase day by day, mobile socializing increasing with effect of 4G network. Big data exist in cloud data server Amazon web service, Google BigTable, MapReduce and Hadoop big data has a significance role to process large data sets in the cloud. The purpose of this research paper to highlight the importance of social network big data with mobile cloud. The primary data is collected from questionnaire, interviews, and previous research paper.**

Index Terms— Facebook Data Management, Big Data Management, YouTube Big Data

I. Introduction

Mobile cloud computing is a hot research topic now a days. Cloud computing solves mobile computing problems without using mobile features. iPhone is the best of all Smartphone's because the stability of its software is very impressive, it is very user friendly and camera is very good. People now a day's use iPhone to make movies. Amazon EC2 and Google App Engine are examples of cloud computing. Before cloud computing traditional business application have always been complicated and expensive. The selection of hardware and software required to run them frightening. Organization needs an expert team to configure, install, test, run, secure and update them. When organization multiply these efforts across dozens or hundreds of apps, it's easy to see why the biggest companies with the best IT departments aren't getting the apps they need. Small and mid-sized businesses don't place a chance. The arriving of cloud computing eliminate those headaches because organization are not managing hardware and software, it is the responsibility of an experienced vendor like Google cloud platform and Google Compute Engine, salesforce.com is the enterprise cloud computing company. Amazon offer Amazon Cloud Drive is web storage application. The storage space of Amazon Cloud Drive can be access from up to eight specific devices e.g. Mobile devices, computer and different browser on the same computer. Amazon also offers Amazon Elastic compute cloud platform e.g. Amazon web service by allowing users to rent virtual computers on which to run their own computer application. Apple's icloud stores customers, music, photos, apps, calendars, documents etc. Apple's icloud stores are hosted in Amazon EC2 and Microsoft Azure. Amazon has released its new cloud accelerated web browser split browser whose software resides both on stimulate fire and EC2.

Big data is a conceptual idea apart from lots of data it also has some other features, which determine the difference between itself and lots of data i.e. very big data. The latest progress of information technology (IT) makes it more easily to generate data e.g. on average 72 hours of videos are uploaded on YouTube in every minute. Consequently we are tackling with the main challenge of collecting and integrating lots of data from widely distributed data sources. [1] Facebook servers 570 billion page view per month, store 3 billion new photos every month and manage 25 billion pieces of content. Google search and ad business, facebook, flicker, YouTube, LinkedIn used a bundle of artificial intelligence tricks require parsing vast quantities of data and making decision immediately. Cloud computing is connected with new pattern for the provision of computing infrastructure and big data processing method for all kinds of resources. "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [2] . The objective of this research paper to provide the status of mobile cloud computing and big data social network related work. This paper is organized as follows section 2 provides review of the literature section 3 provides methodology of the research and section 4 provides data analysis and results and finally section 5 provides conclusion with future research direction and challenges.

II. Literature Review

1. Architecture of mobile cloud computing

PaaS: Platform as a service offers an advanced integrated environment for building testing and deploying custom applications. The example includes Google App Engine, Microsoft Azure, and Amazon Map Reduce Simple Storage Service.

SAAS: Software as a service sustains software sharing with precise requirements. The user can access application & information remotely via the internet and pay only for that they use. Microsoft Live Mesh and Salesforce is one the pioneering providing this model.

IAAS: Infrastructure as a service, it enables hardware, server, storage and networking components. The client usually pays on a per-use basis. The client can save cost as the payment is only based on how much resource they really use. The examples of IAAS are Amazon Elastic cloud computing and simple storage service. [3]

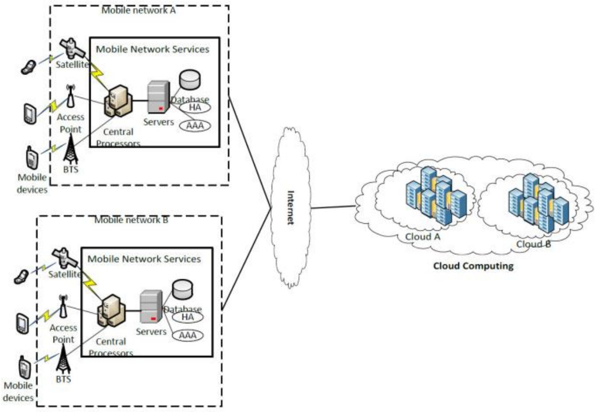


Figure 1: Architecture of Mobile Cloud

2. Big Data Management on Cloud

The increasing number of large scale date warehouse and the non-similarity of data structure complicate data management. Integrated huge distributed data and providing virtually unified storage for mobile users. Amazon simple storage service is online public storage web service offered by Amazon web service. The file system is targeted at cluster hosted on the Amazon Elastic compute cloud server on demand infrastructure.

3. Distributed Data Management on Cloud

Bigtable is a distributed storage system of Google for managing structured data that is designed to scale to a very large size (petabytes) of data across thousands of commodity servers. Bigtable does not support a full relational data model. Conversely it provides clients with a simple data model that supports dynamic control over data layout and format. PNUTS is a massive scale hosted database system planned to support Yahoo’s web applications. The dynamo is a extremely available and scalable distributed key/value based data store built for supporting internal Amazon’s applications. Facebook projected the design of a new cluster-based data warehouse system, LIama a hybrid data management system which combines the features of row-wise and column-wise database system. They also describe a new column wise file format for Hadoop called CFile, which provides better performance than other file formats in a data analysis. [4]

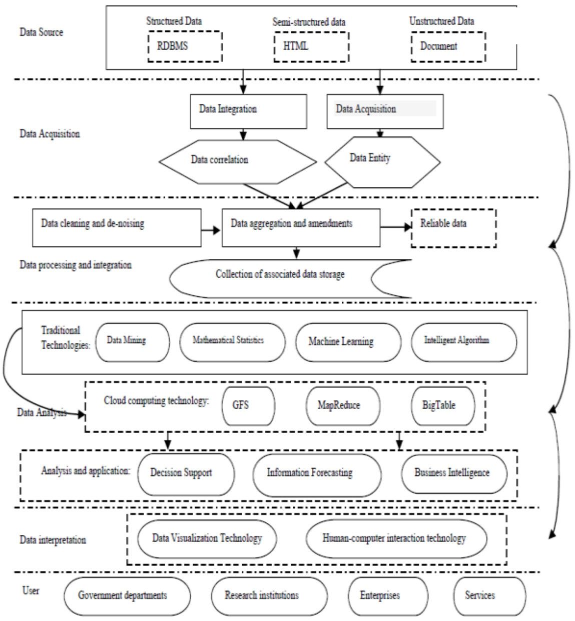


Figure 2: Basic Framework of Big data processing

A. Big Data Processing

MapReduce projected by Google is a very popular big data processing model that has rapidly been studied and applied by both industry and academia. MapReduce has become a popular framework for processing and generating large datasets in parallel over a cluster. Hadoop as an open source implementation of MapReduce is successfully applied in many applications such as web indexing, report generating, data mining, log file analysis. The MapReduce system runs on the top of the Hadoop Distributed File System(HDFS) in which data is loaded and partitioned into chunks, with each chunk replicated across multiple machines. [5]

Figure 3: Big data storage server

B. Microblogs Data Management

Amr Magdy (2015) proposed social media site to which a user makes short, frequent post. Microblogs e.g. Tweets, Facebook comments, news websites comments, are observer extraordinary successful period with the extensive of mobile users. Every day, 288+ million active Twitter users generate 500+ Million Tweets, while 1.39+ billion Facebook users post 3.2+ billion comments. The vast majority of microblogging activity comes from mobile users, specifically, 80% of Twitter users and 85% of Facebook users are mobile. In the development of big data management system Apache Spark and AsterixDB, DBMSs are primarily designed and optimized for efficient processing of big volume data, which can be supported through either in-memory lightweight distributed processing e.g. Spark, or disk-resident index-based distributed processing e.g. AsterixDB. Though supporting big volume data is necessary to handle the large number of microblogs, it is not sufficient as mircroblogs need inherent support for fast streaming data as well. [6]

C. Myria Big Data Management Service

University of Washington has developed a new online service for managing big data. Myria now runs on 100-node Amazon EC2 deployments and processes terabytes of data from applications in astronomy, oceanography, social media, and cyber security. Importantly Myria is setup as a cloud service that user's access directly from their browsers, dramatically reducing the "activation energy" required to be productive with big data. Myria programmability based on SQL for analytics, and window functions, pivoting, UDAs, UDFs, in-database analytics packages such as MADlib etc. Hadoop, GraphLab, Spark, and related systems require users to develop algorithms in low level imperative languages, reducing opportunities for algebraic optimization. [7]

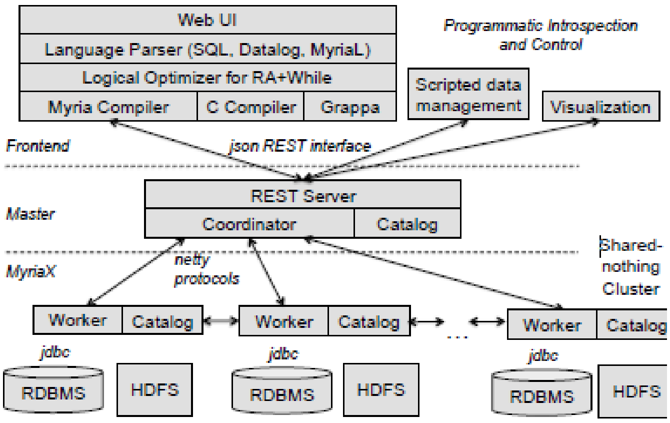


Figure 4: Myria System Architecture

D. Facebook Big Data Management & Storage

Facebook design its own server and networking. It design and builds its own data center. Its staff writes most of its own applications and creates virtually all of its own middleware. Everything about its operational IT unites it in one extremely large system that is used by internal and external individuals alike. According to vice president of infrastructure engineering at Facebook, it said every layout of our stack, server, storage, networking and data center as well as software operations the visibility the tools it all comes together in this one application that we have to provide to all our users. Facebook stores more than 240 billion photos, with users uploading and additional 350 million new photos every single day. To house those photos, Facebook's data center team deploys 7 petabytes of storage gear every month. Facebook builds Exabyte data centers for cold storage. Most importantly, each rack uses just 2 kilowatts of power instead of the 8 kilowatts in a standard Facebook storage rack. But Parikh said it will be able to store 8 times the volume of data of standard racks. Parikh said the system is architected so that different "chunks" of image data don't share same power supply or top-of-rack switch to avoid a single point of failure that would lose data. And if a user deletes a photo, it is deleted from cold storage as well. Not many companies face storage challenges at the kind of scale seen at Facebook. But Parikh believes more companies will be confronting these massive storage issues. "Our big data challenges that we face today will be your big data challenges tomorrow," he said. "We need to keep coming up with advanced solutions to our storage problems.[8] The most important innovations are the problems people solve before the scale of the problem emerges. I believe big data is one of those problems. And we won't keep up unless we work together."[9].

Question: How does Facebook manage the insane amount of data that one billion users pour into the service nearly every day?

Answer: Facebook has the world's largest Hadoop cluster — a group of servers connected using Hadoop's open-source software — with more than 4,000 machines containing over 100 petabytes of data. Even more impressive, it isn't Facebook's only cluster. The problem of managing the constantly swelling system requires some of the greatest engineering and computing minds to solve, but as database administration and storage systems manager Santosh Janardhan told *Wired*, "if you're a technical guy, this is like Candy Land." [10]

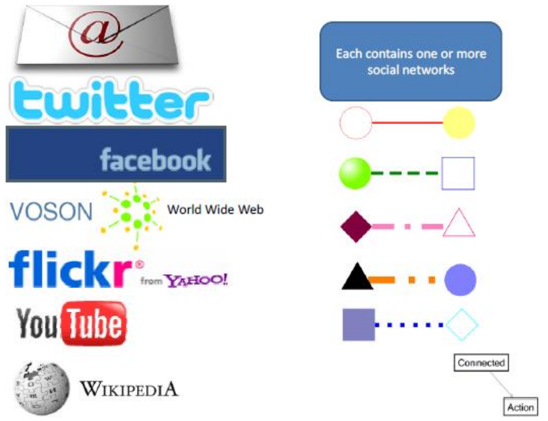


Figure 5: big data and social network

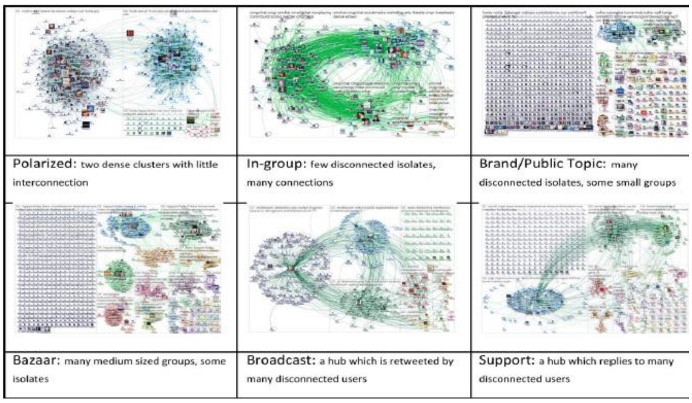
E. YouTube Big Data

Question: How does YouTube manage to store such a huge amount of data?

Answer: It is fairly easy to store large amounts of data on a distributed file system like Hadoop HDFS. This basically scatters blocks of data (typically 64-256MB) around a large amount of standard servers (bunch of xTB disks).[11]

F. Twitter Big Data

Twitters have 140M active users, 340M tweets/day. Twitter is social networking site, there are bundle of tweets has been done. IBM and Twitter work together and twitter data also managed by IBM. Twitter has widely used for education purpose as compare to Facebook twitter tweets are very small in size.



Polarized: two dense clusters with little interconnection

In-group: few disconnected isolates, many connections

Brand/Public Topic: many disconnected isolates, some small groups

Bazaar: many medium sized groups, some isolates

Broadcast: a hub which is retweeted by many disconnected users

Support: a hub which replies to many disconnected users

Figure 6: 6 kinds of social media twitter network

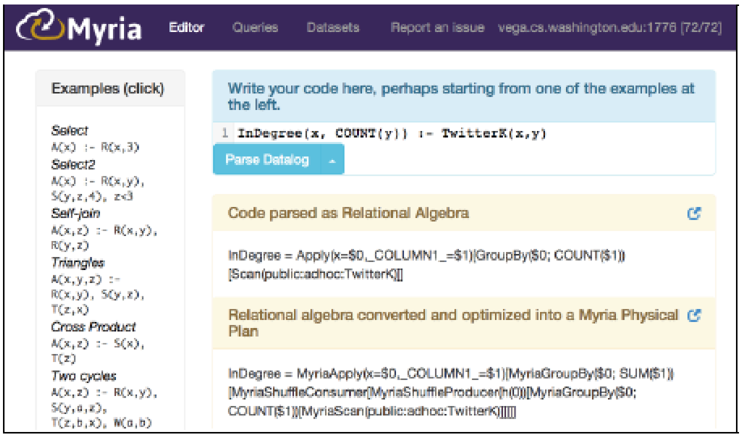
G. Google Gmail Big Data

Google Cloud Platform and Google Cloud storage can manage and store Gmail data on Google Cloud. Google Cloud Storage stores and replicates user data allowing a high level of persistence. Google Cloud Storage is built with a replicated storage strategy. All data is encrypted both in-flight and at rest. The Google security model is an end-to-end process, built on over 15 years of experience. Google Cloud Storage offers developers and IT organizations **durable and highly available object storage**. Google created three simple product options to help you address the needs of your applications while keeping your costs low. These three product options use the same API, providing you with a **simple and consistent method of access**. [12]

METHODOLOGY

The primary data has been collected based on qualitative approach e.g. research paper, case studies and human observation. For the investigation, the data was carried out from different researches, case study, analysis report and interviews was conducted from different author and professor. Questionnaire has been design to carry out big data analysis. Source of data is collected from latest research. Variable of research is the research problem e.g. social network big data. The purpose of this research paper to highlight the big data social network how it is managed and organized in a meaningful way.

1. DATA ANALYSIS & RESULTS



Write your code here, perhaps starting from one of the examples at the left.

```
InDegree(x, COUNT(y)) :- TwitterK(x,y)
```

Select
```
A(x) :- R(x,3)
```
Select2
```
A(x) :- R(x,y),
S(y,z,4), z<3
```
Self-join
```
A(x,z) :- R(x,y),
R(y,z)
```
Triangles
```
A(x,y,z) :-
R(x,y), S(y,z),
T(z,x)
```
Cross Product
```
A(x,z) :- S(x),
T(z)
```
Two cycles
```
A(x,z) :- R(x,y),
S(y,a,z),
T(z,b,x), W(a,b)
```

Code parsed as Relational Algebra

InDegree = Apply(x=$0,_COLUMN1_=$1)[GroupBy($0; COUNT($1))[Scan(public:adhoc:TwitterK)]]

Relational algebra converted and optimized into a Myria Physical Plan

InDegree = MyriaApply(x=$0,_COLUMN1_=$1)[MyriaGroupBy($0; SUM($1))[MyriaShuffleConsumer[MyriaShuffleProducer(h(0))[MyriaGroupBy($0; COUNT($1))[MyriaScan(public:adhoc:TwitterK)]]]]]

1. Myria Query processing Analysis

Graph reachability can be expressed in MyriaL

```
Edge = SCAN(user@uw.edu:edges_table);
Reachable = [1]; Delta = Reachable;
DO
NewNodes = [FROM Delta, Edge
WHERE Delta.addr == Edge.src
EMIT addr=Edge.dst];
Delta = DIFF(DISTINCT(NewNodes), Reachable);
Reachable = UNIONALL(Delta, Reachable);
WHILE [*COUNTALL(Delta) > 0];
```

Graph Reachibility in Myria

```
Edge = SCAN(user@uw.edu:edges_table);
Reachable = [1 AS addr]; Delta = Reachable;
DO
NewNodes = [FROM Delta, Edge WHERE Delta.addr =
Edge.src
EMIT Edge.dst AS addr];
Delta = DIFF(DISTINCT(NewNodes), Reachable);
Reachable = UNIONALL(Delta, Reachable);
WHILE [*COUNTALL(Delta) > 0];
```

Twitter Query

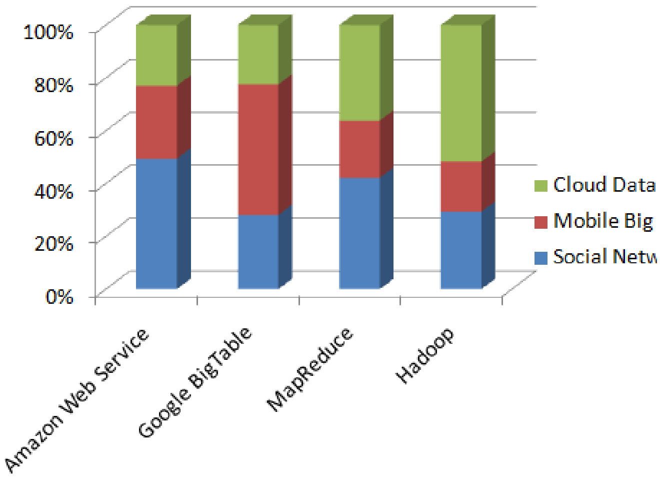
```
SELECT *
FROM twitter_stream
WHERE keyword CONTAINS ALL {Obama, Care}
ORDER BY Max(timestamp)
LIMIT 20 ON LAST ∞ DAYS
```

Example 2. The following continuous aggregate query retrieves the most frequent 10 keywords from tweets in Ukraine since February 18, 2014:

```
SELECT CONTINUOUS keyword, COUNT(*)
FROM twitter_stream
WHERE location WITHIN (52,44.7,39.91,21.8)
GROUP BY keyword
LIMIT 10 ON ("18 Feb 2014",∞)
_ SELECT [CONTINUOUS] attr_list
FROM stream_name1 [,stream_name2,...]
[WHERE condition]
ORDER BY F(arg_list)
LIMIT k
ON {LAST T {MINUTES|DAYS} | (T_start,T_end) }
```

```
_ SELECT [CONTINUOUS] grouping_attr_list,
COUNT(attr_list)
FROM stream_name1 [,stream_name2,...]
[WHERE condition]
GROUP BY grouping_attr_list
LIMIT k
ON {LAST T {MINUTES|DAYS} | (T_start,T_end) }
```

Big Data Analysis



Open Source Software for Big Data Management & Analysis

- Druid is an open-source analytics data store designed for business intelligence (OLAP) queries on event data. Druid provides low latency (real-time) data ingestion, flexible data exploration, and fast data aggregation. Existing Druid deployments have scaled to trillions of events and petabytes of data. Druid is best used to power analytic dashboards and applications.
- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
- Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project. Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk. In addition to Map and Reduce operations, it supports SQL queries, streaming data, machine learning and graph data processing. Developers can use these

Conclusion

In this research paper we investigate the big data prospectus at social media network. Data is big and very large day by day socializing increasing due to vast majority of mobile user increasing. Data is very fast and semi/unstructured Google MapReduce Hadoop scaling the big data at cloud. Amazon web service is very fast and accurate service at cloud computing to deliver accurate data at accurate time. Open source of version Google File system (HDFS) uses various high level language e.g. Pig, Jaql, Hive. Myria Big data management proposed by Washington University but researcher needed to work in this era to improve open source technology. Facebook data center increasing day by day due to vast majority of advertisement as well as education academia increasing their data, and it's a challenging task for upcoming researchers.

Adnan Majeed received his Master of Computer Science degree from Federal Govt University Virtual University of Pakistan Lahore. And won Best Teacher Award from High School, received certificate of Appreciation from Dean School of Computer & IT Beaconhouse National University Lahore.

References

1. Min Chen · Shiwen Mao · Yunhao Liu Big Data Survey Mobile Netw Appl (2014) 19:171–209
2. Douglas and Laney, "The importance of 'big data': A definition,"2008.
3. A survey of mobile cloud computing: architecture, applications, and approaches Hoang T. Dinh, Chonho Lee WIRELESS COMMUNICATIONS AND MOBILE COMPUTING 2011.
4. Changing Ji et al. Big Data Processing in Cloud Computing Environments International Symposium on Pervasive Systems, Algorithms and Networks, 2012.
5. Xia Tang, et al. A Reduce Task Scheduler for MapReduce with Minimum Transmission Cost Based on Sampling Evaluation International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.1-10.
6. Amr Magdy et al. Towards a Microblogs Data Management System (Invited Industrial Paper) University of Minnesota, Minneapolis, MN, USA.
7. Daniel Halperin et al. Demonstration of the Myria Big Data Management Service Computer Science & Engineering and eScience Institute, University of Washington.
8. http://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage/
9. http://www.eweek.com/c/a/Data-Storage/How-Facebook-Is-Handling-All-That-Really-Big-Data-423736.

10. http://www.theverge.com/2013/2/5/3954666/facebook-data-management-system-candy-land.
11. https://www.quora.com/How-does-Youtube-manage-to-store-such-a-huge-amount-of-data.
12. https://cloud.google.com/storage/