International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763
Issue 03, Volume 3 (March 2016) www.ijirae.com

Managing Big data using Hadoop Map Reduce in Telecom Domain

V.Vaidhehi, Arun M.B
Department of Computer Science, Christ University
Bengaluru

Abstract- Map reduce is a programming model for analysing and processing large massive data sets. Apache Hadoop is an efficient frame work and the most popular implementation of the map reduce model. Hadoop's success has motivated research interest and has led to different modifications as well as extensions to framework. In this paper, the challenges faced in different domains like data storage, analytics, online processing and privacy/ security issues while handling big data are explored. Also, the various possible solutions with respect to Telecom domain with Hadoop Map reduce implementation is discussed in this paper.

Keywords - MapReduce, Data Storage, Analytics, Online processing, privacy and security issues.

1. INTRODUCTION:

Big data is typically characterized by 3 V's namely; volume, velocity and variety. The data expands in all these 3 forms. Volume refers to size or amount of data. Velocity refers to the speed at which data is created, stored, analysed and visualized.Variety refers to whether data created is structured data or unstructured data. In any domain, Big data can be seen. This paper is based on Big data in Telecom domain. Telecom domain is richly connected with technology and computers. Therefore, a large amount of data is generated by different stakeholders in the telecom industry. Thus it becomes very challenging to operate on the huge data. In the Big data domain, map reduce is one of the key approach used for handling large data sets. At the same time, map reduce faces a huge number of challenges when dealing with Big data such as lack of high language like SQL, support for stream processing, iterative ad-hoc data exploration and certain challenges in implementing iterative algorithms.

Though there are plenty of challenges in the Telecom domain with big data environment, the following are the list of major challenges to be addressed with ultimate care.

1. Challenges related to data storage
2. Challenges related to analytics
3. Challenges related to online processing
4. Challenges related to privacy and security

This paper has different sections. Section 2 explains about the overview of the Map-reduce. Section 3 explains about the literature review in this domain. Section 4 explains the challenges in detail. Section5 briefs about the various approaches to overcome the challenges.

1.1 Deriving business value from big data by Telecom operators:

The Telecom industry in India has crossed nearly two decades post privatization of its sector. Since then there has been a lot of consolidation, innovation and maturation has happened in the industry. Today there are 12 major mobile telecom operators operating in the country. These operators are facing disruptive technologies, rapidly changing business rules and intensified regulatory environment leading to eroding service margins. So, in this scenario the major stabilizing factor for telecom operators is the revenue generated from data provisioning and driving value from this data.

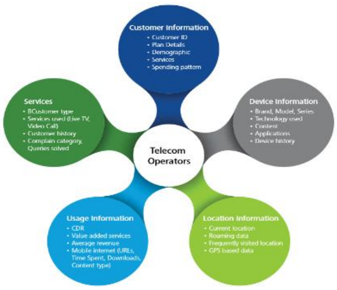


Fig 1. Sources of big data in telecom industry

It has been observed that Big data has the potential to place the communication service providers in a good position to win the battle for customers and generate new revenue streams. It provides them with a good amount of information about the customer's behaviours, preferences and movements. At some macro level, big data throws certain challenges to customers because of its variety, velocity and complexity as described above. The industry is looking for professional data scientists who can understand the trends in data analytics and merge it with the business objectives of telecom operators. Also the infrastructure requires high computational capabilities and data storage. It also requires flexibility to analyse different formats of data. Transforming the historical data from older system to new system is a big challenge. Data quality is also major challenge as different equipment provides data in the various formats. Data can be inaccurate and maintaining appropriate quality of data is a mammoth task for every company. The privacy issues and government policies are other challenges as customers prefer not to share their personal data.

II. MAP REDUCE OVERVIEW:

Map reduce is a programming framework for processing and analysing large data sets in a highly distributed environment. The map function is responsible for filtering and sorting whereas reduce function is responsible for grouping and aggregation operations. A map task can run on any compute node in cluster. Multiple map tasks can run in parallel across the cluster. The output of all the maps will be partitioned and each map will be partitioned and then each partition will be sorted. There will be only one partition for each reduce task. There can be multiple reduce tasks running in parallel on the cluster.

The reduce function will aggregate the information received from map functions. Map reduce framework supports scalability to a greater extent. Also, Map Reduce achieves a high level of parallel operations and distributed execution over a large number of nodes. In the map reduce model, task is distributed into a multiple jobs which are assigned to nodes in the network. Reliability is achieved by reassigning any failed node's job to another node. The widely used open source MapReduce implementation is Hadoop. Hadoop uses MapReduce on top of the Hadoop Distributed File System (HDFS).

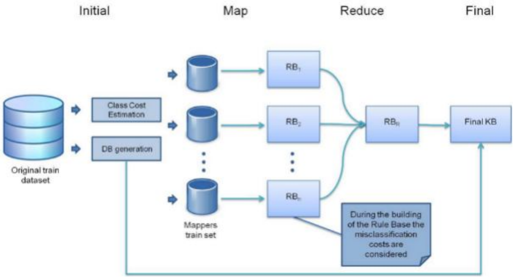


Fig 2. Map reduce flow

III. LITERATURE SURVEY

The process of research into complex data basically is concerned with revealing of hidden patterns. The critical and major issue about Big data is privacy and security. Big data samples provide the review about atmosphere, biological science, research and life sciences etc. Mapreduce is an important paradigm of Big data that helps to solve the problem of processing large unstructured data. The overall evaluation describes that the data is increasing randomly and becoming complex.

Analytics process in big data is discussed in [1]. The different approaches and data source for big data analysis is elaborated in [1]. Dean[2] the working of Map reduce programming paradigm along with the map reduce and map functions. The major challenges related to querying the big data is addressed by C.Doulkeridis[3]. The Hadoop HDFS with map reduce is discussed in[4]. The different approaches related to big data storage is briefed in[5]. Kumar[6] describes the programing model called Apache Hive which implements Big data environment in any domain. Y.Bu[7] explains about different programming model to deal with iterative computation of data.Q.He[8] explicates different approaches for parallel processing of data.J.Heer[9] talks about the methods for implementing interactive analysis of big data. Y.chen[10] speaks about interactive analytical processing of big data in online processing. The literature survey reveals that specific challenges and solutions are being addressed by different authors. In this paper, challenges and solutions to different domains like data storage, analytics, online processing and privacy/ security issues with respect to telecom domain are addressed.

IV. CHALLENGES

Processing and analyzing in big data environment using Map Reduce framework has the following challenges.

1. Challenges related to data storage
2. Challenges related to analytics
3. Challenges related to online processing
4. Challenges related to privacy and security

The details about the challenges are listed in figure 3.

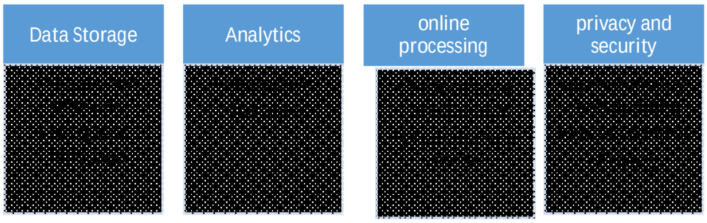


Fig 3. Challenges in big data

4.1 Challenges related to data storage:

4.1.1 Map reduce is schema independent and index free:

Map reduce is basically schema independent and index free. It provides a great flexibility and enables map reduce to work with several semi-structured and unstructured data. Map reduce will run as soon as data is loaded. But, due to lack of indexes on Map reduce, it may result in poor performance compared to relational databases. NoSQL and NewSQL have emerged as the new alternatives for Big data storage. NoSQL refers to not only SQL. Its main characteristics are schema flexibility and efficient scaling over a large number of commodity machines. NoSQL includes data storage with respect to scaling of read or write operations.

4.1.2Lack of standard SQL like language:

The next challenge to Map reduce and big data is the lack of standardized SQL like language. Therefore, there is a need to provide SQL like query language in map reduce. Suitable example is Apache Hive, a very popular frame work that provides SQL kind of language on top of Hadoop. Another apache framework, Mahout helps to build scalable machine learning libraries on top of Map reduce. Even though they provide powerful data processing capabilities, they lack features like advanced indexing and a sophisticated optimizer.

4.2 Challenges related to analytics:

4.2.1 Interactive analysis:

Interactive analysis is defined as an approach that helps data scientists to explore data in an interactive way. Methodologies to query and visualize data in big data environment at an interactive level is still a complex task to be solved. Interactive analysis for Map reduce involves processing many small and interactive jobs. As there is a shift from RDBMSs to Big data storage systems, some prior assumptions regarding map reduce are violated such as uniform data access and prevalence of large batch jobs. Interactive analytics on Big data is still a path to be explored by many researchers.

4.2.2 Iterative algorithms:

Map reduce involves multiple iterations in order to reach convergence and jobs are expensive in terms of startup time. Apparently, skews in the data forms stragglers in the reduce phase which causes backup execution to be launched, which increases the computational load.

4.3 Challenges related to online processing:

4.3.1 Performance and latency issues:

The velocity dimension as one of the V’s used to define Big data offers several new challenges to data processing techniques and especially to map reduce. So, in order to handle the velocity of big data, it often requires applications with online processing capabilities.

From the business perspective, the goal is to obtain inputs from these data streams and to enable prompt reaction to them. Areas such as algorithmic trading and financial fraud protection have been highly interested in this type of solutions. The mains reasons concerned for latency issues are:

- Map reduce computations are batch processes that start and finish. On the other hand computations over streams are continuous tasks that will only finish upon user request.
- The inputs of map reduce computations are snapshots of data stored on files and the content does not change during processing.
- In order to achieve fault tolerance, most Map reduce implementations write results of Map phase to the local files before sending them to reducers.
- It is not possible to express every computation result using the map reduce programming model and the model does not support composition of jobs.

4.3.2 Programming model:

In Map reduce, there is no standard programming model followed in its implementation. Due to the advert difficulty of expressing complex computations, there is great motivation for the development of programming model. MapUpdate is a new programming model that was conceived by Muppet project. The significant difference is that the update phase has access to slates. These slates are data structures that have persistent state related to each update key. Theoretically, these slates can help to achieve easier implementation of iterative algorithms.

4.4 Challenges related to privacy and security:

4.4.1 Accountability and auditing:

Accountability is the ability to know when someone performs the action and to hold them responsible for that action and is often tracked through auditing. Accountability in Mao reduce is witnessed only when mappers and reducers are held responsible for the tasks they have completed.

4.4.2 Access control:

An additional security challenge for Big data map reduce is that of providing access control which can be defined through 3 of Big data's defining V properties: volume, variety and velocity. When dealing with large volume of information, it often requires multiple access to storage locations and devices. So, multiple access requirements are required for any one task.

4.4.3 Privacy:

Privacy is also one of the major topic of concern whenever large chunks of information are used. Certain processes like data mining and predictive analytics can deduce information linkages. For certain organizations, information linkages are useful allowing them to understand better, target and provide for their clients or users. But, on an individual basis this can cause the identities of data providers to be exposed.

V. SOLUTIONS

The solutions for various challenges posted are discussed in detail in this section. They are

i. Solutions related to data storage
ii. Solutions related to analytics
iii. Solutions related to online processing
iv. Solutions related to privacy and security

5.1 Solutions related to data storage

The challenges related to data storage and its solutions are depicted in the figure 4.

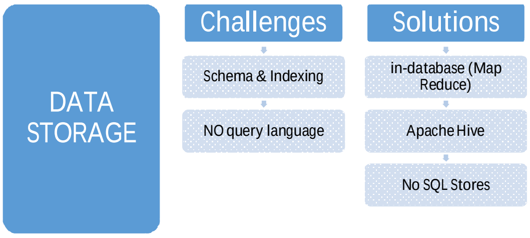


Fig 4. Solutions to Data Storage

5.1.1 Use of In-database Map reduce:

For most of the companies, big data feeds includes point of sale records, call detail records, financial, medical and intelligence data stored in oracle databases but map reduce algorithms such as clustering, classification and recommenders are very hard to implement with SQL alone. The In-database map reduce accesses and processes data in place, directly in database tables. Therefore, by using In-database map reduce we can overcome the schema independent and index free challenges of Map reduce.

The solution works really well when the organizations have invested heavily in RDBMSs and they still want to capture and analyse the unstructured data from different sources like social media in Hadoop systems. In the telecom domain, various service providers offer multiple services to its customers. It is essential for every service provider to understand the customer behaviour and usage patterns to achieve customer satisfaction. Most telecom companies would normally store the details of customer and CDRs using RDBMS tables. Due to wide variety of unstructured data, it is essential to integrate the RDBMS with map reduce function. With this, it is easy for service providers to get detailed analysis of customer behaviour and preferences. It is shown in figure 5.

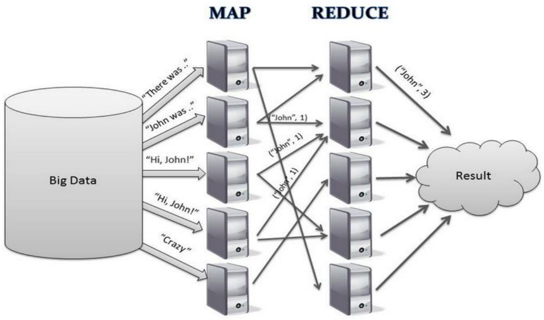


Fig 5. Unstructured data from different telephone conversations

5.1.2 Apache Hive:

It is a data warehouse infrastructure built on top of Hadoop. It supports data summarization, query and analysis. Apache Hive Apache Hadoop supports the analysis of large data sets stored in HDFS and compatible file system like Amazon S3 filesystem. It provides a language called HiveQL with schema on read feature and apparently converts queries to map or reduce. In order to accelerate queries it provides indexes including bit map indexes.

The features of Hive are that it is very familiar tool that allows query data with a SQL based language, next feature is robustness, which has got very interactive response times even over huge datasets and when the data grows, more commodity machines can be added without a reduction in performance. Most of the telecom companies store the customer details and CDR details in RDBMS tables. Every time, the data is fetched by applying SQL queries and procedural scripts. Since Apache Hive is built on top pf Hadoop, it is easy to provide details call summary, analysis of calls, fraud detection methods and other useful customer behaviours. Thus Apache Hive has emerged as a query language in Big data domain. It is shown in figure 6.

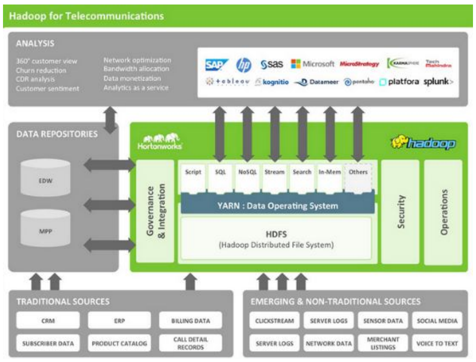
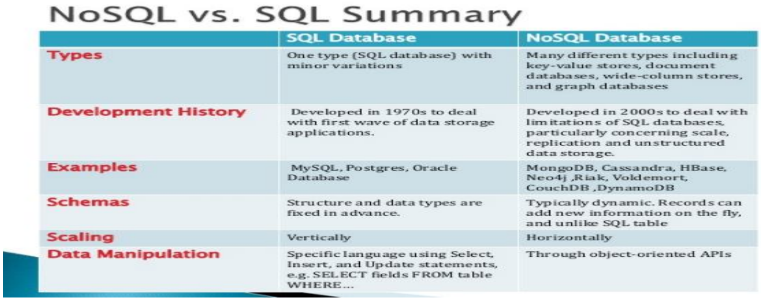


Fig 6. Apache Hive and Hadoop for telecommunication.

5.1.3 No SQL stores:

A No SQL (originally referring to “non SQL” or “non-relational”) provides an approach for storage and retrieval of data which is modelled in means other than tabular relations used in relational databases. Motivations are simplicity in the design, simple horizontal scaling to clusters of machines and good control over availability. Also, the data structures used by NoSQL databases are different from those used by default in relational databases making some operations faster in NoSQL. It is shown in figure 7.



NoSQL vs. SQL Summary

	SQL Database	NoSQL Database
Types	One type (SQL database) with minor variations	Many different types including key-value stores, document databases, wide-column stores, and graph databases
Development History	Developed in 1970s to deal with first wave of data storage applications.	Developed in 2000s to deal with limitations of SQL databases, particularly concerning scale, replication and unstructured data storage.
Examples	MySQL, Postgres, Oracle Database	MongoDB, Cassandra, HBase, Neo4j ,Riak, Voldemort, CouchDB ,DynamoDB
Schemas	Structure and data types are fixed in advance.	Typically dynamic. Records can add new information on the fly, and unlike SQL table
Scaling	Vertically	Horizontally
Data Manipulation	Specific language using Select, Insert, and Update statements, e.g. SELECT fields FROM table WHERE…	Through object-oriented APIs

Fig 7. NoSQL vs SQL summary

5.2 Solutions related to analytics

The challenges related to analytics and its solutions are depicted in the figure 8.

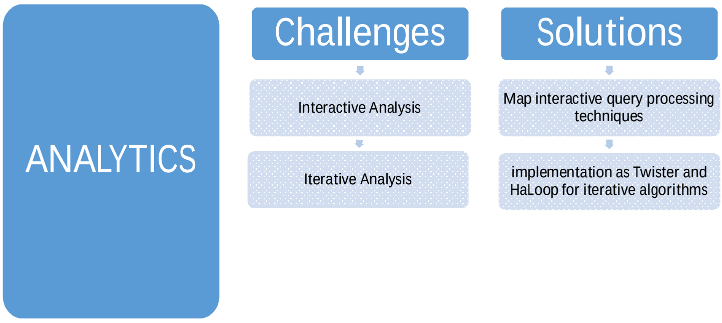

ANALYTICS
Challenges
Interactive Analysis
Iterative Analysis
Solutions
Map interactive query processing techniques
implementation as Twister and HaLoop for iterative algorithms

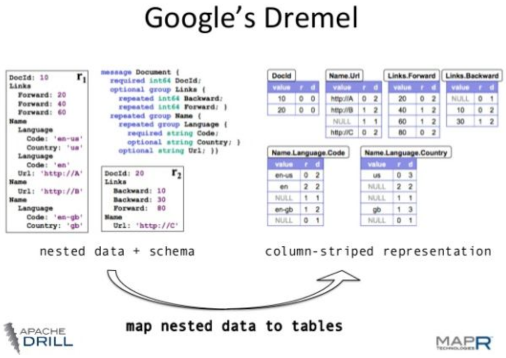
Fig 8.Solutions to Data Analytics

5.2.1 Map interactive query processing techniques for handling small data to map reduce:

Interactive analysis for big data is an extension of existing well researched area of interactive query processing. Based on this assumption there are many potential solutions available to optimize the interactive analytics with map reduce. Google’s Dremel system is one approach tuned for interactivity which acts in complement to Map reduce. Dremel is build by algorithms that constructs the columns and reconfigure the original data. Some of the highlights of Dremelssystem are:

- A real time interactivity support for scan based queries.
- Near linear scalability in the number of clusters.

In the telecom industry, customers are able to choose among multiple service providers and frequently switch from one service provider to other. In this competitive market customers expect tailored products and better services at less prices. So customer retention has become more important than customer acquisition. So, in order to meet this, many telecom companies deploy retention strategies in synchronizing programs and processes to retain customers longer by providing them with tailored products and services. This can be achieved through an interactive analysis coupled by innovative marketing approach.



Google's Dremel

```
DocId: 10                r1
Links
  Forward: 20
  Forward: 40
  Forward: 60
Name
  Language
    Code: 'en-us'
    Country: 'us'
  Language
    Code: 'en'
  Url: 'http://A'
Name
  Url: 'http://B'
Name
  Language
    Code: 'en-gb'
    Country: 'gb'
```

```
message Document {
  required int64 DocId;
  optional group Links {
    repeated int64 Backward;
    repeated int64 Forward; }
  repeated group Name {
    repeated group Language {
      required string Code;
      optional string Country; }
    optional string Url; }}
```

```
DocId: 20                r2
Links
  Backward: 10
  Backward: 30
  Forward:  80
Name
  Url: 'http://C'
```

DocId

value	r	d
10	0	0
20	0	0

Name.Url

value	r	d
http://A	0	2
http://B	1	2
NULL	1	1
http://C	0	2

Links.Forward

value	r	d
20	0	2
40	1	2
60	1	2
80	0	2

Links.Backward

value	r	d
NULL	0	1
10	0	2
30	1	2

Name.Language.Code

value	r	d
en-us	0	2
en	2	2
NULL	1	1
en-gb	1	2
NULL	0	1

Name.Language.Country

value	r	d
us	0	3
NULL	2	2
NULL	1	1
gb	1	3
NULL	0	1

nested data + schema

column-striped representation

map nested data to tables

Fig 9. Google's Dremel system

5.2.2 Implementation such as Twister and HaLoop:

HaLoop is an extension of Hadoop which along with processing of data provides an interesting way to perform iterative computation on data. It was originally developed by YingYi Bu, Bill Howe and Magda Balazinska at University of Washington.it is shown in figure 10.

Below are some of the enhancements that were incorporated in Hadoop to support iterative data analysis:

1. Supporting a new API to simplify and minimize the iterative expressions.
2. Automated generation of Map reduce program through the use of master node using a loop control module until the loop condition is met.
3. To efficiently compute iterative operations, a new task scheduler which supports data locality is incorporated.
4. The task scheduler and task tracker are often modified to manage execution and manage cache indices on slave module.

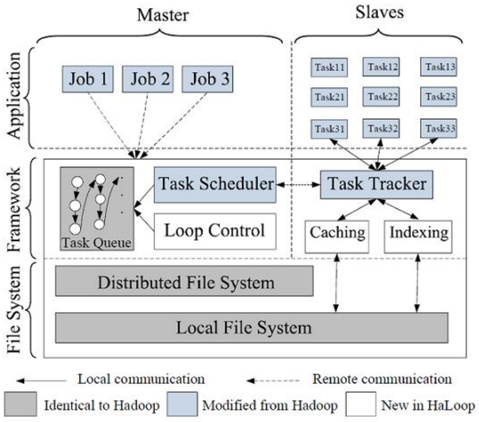


Fig 10. Haloop for Map reduce

Furthermore, Twister is a light weight Map reduce runtime paradigm. Twister provides the below features to support map reduce computations:

- Clear difference between the static and variable data.
- Configurable long running map/reduce tasks.
- Combine phase to collect all reduce outputs.
- Data access via local disks.
- Lightweight (Approximately 5600 lines of Java code).
- Support for map reduce computations.
- Sophisticated tools to manage data.

5.3 Solutions related to Online processing

The challenges related to Online processing and its solutions are depicted in the figure 11.

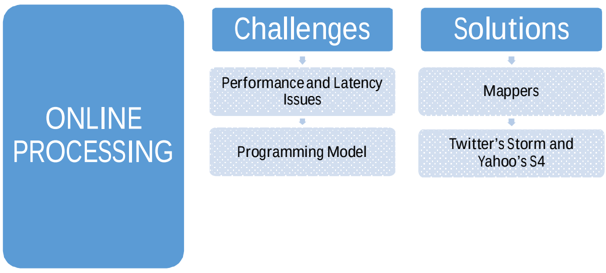

ONLINE PROCESSING
Challenges
Performance and Latency Issues
Programming Model
Solutions
Mappers
Twitter's Storm and Yahoo's S4

Fig 11. Solutions to Online Processing

5.3.1 Mappers

Several efforts have been laid out in order to overcome the performance and latency issues of Map reduce.In this case, the execution of functions map and reduce is normally managed by a data stream processing platform. To improve the processing latency, the mappers are frequently pushed with batches of tuples, instead of the input files and the results are pushed to reducers as soon as they are available.

Several telecom companies offer different broadband services to its customers. The main features of the broadband service include performance, speed and latency. It is essential for operators to offer optimum speed and bandwidth to its customers. Due to lot of data traffic and presence of various unstructured data, the speed may be compromised at some intervals. By implementing the above strategies, we can overcome these limitations. Major telecom companies adopted the data stream processing platform to address the continuous map reduce functions. As a result of that, the operators were able to achieve high performance and throughput. Below figure shows the statistics with respect to latency mapped against different organizations. Telecom domain, latency level computations are shown in figure 12.

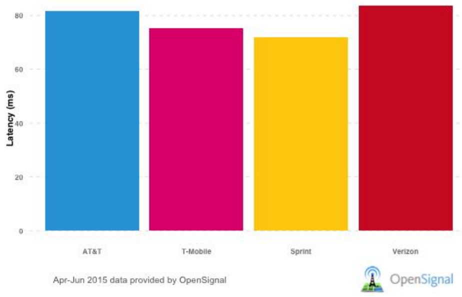


Fig 12. Latency levels computed against different operators

5.3.2 Twitter's Storm and Yahoo's S4

Certain programming frameworks like Twitter's Storm and Yahoo's S4 maintain runtime platforms inspired by the Map reduce implementations. In Twitter's Storm computation is defined by a topology which specifies the sequence of processing elements (bolts) that will contain user defined logic, number of threads (tasks) for every bolt and how to partition the input streams out of the many bolt tasks. In Yahoo's S4, a computation is expressed by a graph of processing elements which are equivalent to Storm's bolts. In both projects the runtime platform tries to manage many low level aspects of distributed computing like parallelization, messaged delivery and fault tolerance.

5.4 Solutions related to Privacy and Security

The challenges related to privacy and security and its solutions are depicted in the figure 13.

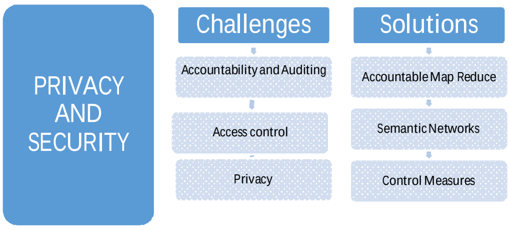


Fig 13. Solutions to Privacy and Security

5.4.1 Accountable Map Reduce

One solution that was provided to deal with this issue is the creation of Accountable Map reduce. The solution uses a set of auditors to perform accountability tests on mappers and reducers in real time. By monitoring the results of these tests, malicious mappers or reducers can be identified and hence accountability can be achieved.

5.4.2 Semantic Network

When dealing with large amount of data that has wide variety, semantic understanding of the data should play a role in the access control decision process. Finally, the velocity parameter of Map reduce and big data requires that whatever access control approach is used must be optimized to determine access control rights in a reasonable amount of time.

5.4.3 Control Measures

Privacy protection requires an individual to maintain control over their personal information. This can be achieved through transparency and allowing input from the data provider. Transparency is provided to an individual by the knowledge of how the private information is being used, how private information is collected, what private information is collected and who has access to it.

VI. CONCLUSION

In the real world, data processing and storage approaches are facing many challenges in meeting the continuously increasing demands of big data. This work focussed on Map reduce, one of the key approaches for meeting the big data demands through highly parallel processing on a large amount of commodity nodes. Challenges and solutions on four dimensions like data storage, analytics, online processing and privacy and security are elaborated in detail in this paper.

References

1. P. Zadrozny and R. Kodali, Big Data Analytics using Splunk, Berkeley, CA, USA: Apress, 2013.
2. J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), pp. 107-113, 2008.
3. C. Doulkeridis and K. Nørvåg, "A survey of large-scale analytical query processing in MapReduce," The VLDB Journal, pp. 1-26, 2013.
4. X. Su and G. Swart, "Oracle in-database hadoop: When MapReduce meets RDBMS," Proc. of the 2012 ACM SIGMOD International Conference on Management of Data, 2012.
5. J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, "MAD skills: New analysis practices for Big Data," VLDB Endowment, 2(2), pp. 1481-1492, 2009.
6. K. A. Kumar, J. Gluck, A. Deshpande and J. Lin, "Hone: Scaling down hadoop on shared-memory systems," Proc. of the VLDB Endowment, 6(12), pp. 1354-1357, 2013.
7. Y. Bu, B. Howe, M. Balazinska and M. D. Ernst, "HaLoop: Efficient iterative data processing on large clusters," Proc.VLDB Endow., 3(1-2), pp. 285-296, 2010.
8. Q. He, Q. Tan, X. Ma and Z. Shi, "The high-activity parallel implementation of datapreprocessing based on MapReduce," Proc. Of the 5th International Conference on Rough Set and Knowledge Technology, 2010.
9. J. Heer and S. Kandel, "Interactive analysis of Big Data," XRDS: Crossroads, the ACM Magazine for Students, 19(1), pp. 50-54, 2012.
10. Y. Chen, S. Alspaugh and R. Katz, "Interactive analytical processing in Big Data systems: A cross-industry study of MapReduce workloads," Proc. of the VLDB Endowment, 5(12), pp. 1802-1813, 2012.