Algorithm for Greening Big Data Networks in Iraq Under Veracity Dimension

Howraa M. Mohammad Ali¹, Muayad Saleem Kod², Ali M. A. Al-Salim³, Caesar Al-amiri⁴

¹ University of Kerbala-Iraq, Howraa.alhaddad@gmail.com
² University of Kerbala-Iraq,
³ University of Kerbala-Iraq,
⁴ University of Glasgow, UK

Abstract—This work examines ways to improve energy efficiency in big data networks in Iraq under the veracity dimension by designing an IP/WDM Network in Iraq for processing big data and developing an algorithm that condenses the distinctive features of veracity. The analysis is based on preparing big data chunks for progressive processing after performing cleansing operations as proposed in this work. This is done in Processing Nodes (PNs) that consist of several servers and their associated network equipment, such as switches and routers. The PNs are distributed in the network for each city, and the cleansed data is handled at each PN, reducing the size of valuable data to be transmitted. Each cleansed chunk is also stored in a Backup Node (BN) for future use. Around a 50% power saving is achieved using this technique compared to the traditional method with no progressive processing.

Index Terms—Big data veracity, data cleansing, energy efficient IP over WDM networks.

INTRODUCTION

Veracity of big data refers to the biases, clutter, and irregularity frequently seen in such data. Veracity is a big challenge for the software used to analyse big data, as it is difficult to differentiate between valuable information and this waste data. Action must thus be considered to keep such grey data out of organisations' databases; the U.S. economy alone wastes around three trillion dollars every year because of poor data [1].

Data cleansing [2] is the process of sensing and eliminating errors, overlaps, contradictory material, and repetition in data to enhance data quality. Data cleansing should be considered prior to all big data analysis preparations to provide access to precise, reliable, and consolidated data [3], [4], [5].
In this article, a heuristic is developed to explore the influence of big data veracity on the energy consumed in a proposed IP over WDM core network in Iraq. An IP/WDM network for big data processing in Iraq is designed and cleansing operations performed for the unprocessed big data chunks. The cleansed data is then backed up and stored before a copy of the cleansed chunks is optimally forwarded to a Backup Node (BN) within the network. The Energy Efficient Big Data Networks (EEBDN) technique developed in [6] is then applied to allow energy aware big data progressive processing operations. Hence, the network resources and the computational resources (i.e. servers) are energy efficiently employed by combining the veracity dimension and the EEBDN processing technique.

A. Related Work

In [7], the authors proposed a method of Computing on Masked Data (CMD) to allow calculations to be achieved using concealed data, guaranteeing that only approved receivers can reveal the data. The authors in [8] proposed a method to mitigate two problems in selecting true data by combining the results of a set of actuality detection methods and initial tests to improve quality performance. The authors in [9] presented a tutorial to enable re-evaluation of the entity resolution process to determine the best ways to handle data veracity; this tutorial bonded entity resolution and focused on sources of ambiguity in the entity resolution process. The authors in [10] presented an inclusive big data research programme for information systems focusing on collaboration between the characteristics of big data. In this research, the work in [6],[11],[12], and [13] is extended.

B. The Impact of Veracity: An Example

The main processing platform for big data applications is Hadoop-MapReduce [5]. Raw data of huge size (Chunk) enters the Map process, and the output is data of a smaller size (Info) after the Reduce process. The size of output data after processing is much smaller than the input data, as in the following Equation [6]:

\[ \text{Size of Info} = \text{Reduction Rate} \times \text{Size of Chunk} \]  

(1)
where a chunk represents big data input before processing. On the other hand, Info is the extracted knowledge from the corresponding chunks after processing and it is much smaller in size compared to the corresponding chunk. The Reduction Rate (RR) is the factor of reducing the huge amount of unwanted dirty data such as noise data. An example of the dirty data is capturing hundreds of thousands of pictures in the ocean to find a specific crashed airplane part. The result could be only one useful image of size 1 KB from thousands of noise images of size 1 TB.

The example depicted in Fig. 2 explains the concept of veracity impact on greening big data networks. All SPNs receive unprocessed data generated by multiple big data applications from the corresponding users. Each SPN can then perform the cleansing process offline during off-peak periods. These cleansing operations help to clean the data and reduce the volume from the raw data (a reduction percentage of around 20% of the originated data is assumed in these calculations). In this approach, the cleansing process prepares the data before optimising the processing location of the cleansed chunks based on the optimisation technique used in [6]. Edge processing is implemented at the SPNs to process as many cleansed chunks as possible, with any extra unprocessed chunks transferred through energy efficient routes to the nearest IPN to be processed there. This chunk movement produces a big data traffic type known as CHT. Each cleansed chunk is also forwarded to a location-optimised backup node (BN) for future use, producing extra big data traffic known as BCH. The extracted knowledge from the processed cleansed chunks produces a smaller quantity of big data traffic (INF), which refers to the movement of cleansed information towards a location-optimised DC. Note that any unprocessed cleansed chunk that is not processed in the edge stage (in the SPNs) or the progressive stage (in the IPNs) is processed through the central stage inside the central DC.
Fig. 2. The impact of veracity: an illustrative example.

Although the main goal is to design and implement an energy efficient network for big data, an energy efficient procedure for big data processing is also offered. This procedure consists of two main operations: a Bin-Packing algorithm to handle as many cleansed chunks as possible in a single processor in conjunction with a slicing algorithm that can process a single chunk in multiple servers.

C. Problem Statement

The problem under the impact of the veracity dimension is stated below:

Define a physical topology that consists of a set of nodes and links, PNs, DCs, and a BN. A PN is attached to each core node, and two DC locations are optimised alongside one BN location; the latter is independent of the DC locations.

Each cleansed chunk is backed up and forwarded to an optimally selected BN for future use. The movement of the backup cleansed chunks generates specific big data traffic, Backup Cleansed Chunks Traffic (BCH). Each SPN is also given several uncleansed chunks (β) that are cleansed offline during off-peak periods. After the cleansing process is complete, the cleansed chunks are optimally processed inside the SPNs, IPNs, and DCs using the proposed energy efficient progressive central processing scheme. However, identifying the optimal cleansing process for real time applications is left for future work.
The objectives are as follows: Find an optimal location for processing the cleansed chunks. Find an optimal location for DCs and a BN. Find an optimal energy efficient route for all the types of big data traffic (CHT, INF, and BCH). Find the optimal minimum number of servers inside the PNs and DCs employed to process big data applications, along with an optimal minimum number of IP router ports, fibre links, and EDFAs for each physical link.

D. Heuristic

Fig. 3 shows the selected heuristic. Its objectives are to cleanse, back up, and process the received data chunks while exploiting the minimum amount of resources. The heuristic is implemented using MATLAB R2015b on a PC with 8 GB Ram and a Core i7 processor. The execution time was less than a second for each run. The heuristic was initiated by determining network topology. Each node in the network receives unprocessed data from the corresponding users. The heuristic then starts the cleansing process and backs up the cleansed data before forwarding a copy of the cleansed data to the identified optimal backup node. The heuristic then applies the algorithm from the work in [6] to complete the operation.

![fig:3](Veracity heuristic)

E. Results of Veracity Scenarios

The heuristic was assessed using the network depicted in Fig. 4 that contains 18 nodes and 18 bidirectional links. A mini data center is built at each node, with the central datacenter assumed to be located at node 1, in Baghdad, the capital city of Iraq. However, the BN is located at node 4, in Babylon, as this is closer to the
Data analysis is based on the trace file for the Hadoop-MapReduce of FaceBook, as seen in Table I [14].

| Input in GB | Output GB | PRR |
|------------|-----------|-----|
| 6.9        | 0.0060    | 0.0086 |
| 1500       | 2.20      | 0.0014 |
| 2100       | 2.70      | 0.0012 |
| 2700       | 260.0     | 0.096  |

To implement the proposed IP over a WDM network topology in Iraq, the first step is to install an IP over WDM node at each city; all cities are already connected via fibre optics. This connection would make it easy to implement the proposed topology. Once this is achieved, an optimal location for the central datacentre should be selected for each location using the optimisation algorithm in [6]. Each city has a headquarters (HQ) building for national telecommunications and Internet run by the Iraqi Telecommunications and Post Company (ITPC); these buildings should offer available space for the proposed PNs. These PNs would then operate specifically to support big data processing and networking. Once these points are addressed, this work can be applied to open a new era in energy efficient big data processing and networking in Iraq.
Fig. 4 The proposed Iraq IP/WDM network for big data processing.

Table II shows the input parameters for analysis.

| TABLE II | Input Parameters of the Algorithm [6][15][16]. |
|-----------|-----------------------------------------------|
| RR        | 0.001-1 (random uniform)                      |
| Uncleansed Chunk volume in Gb | 50-300 |
Each chunk is assumed to need 1 to 4 CPU units of workload to be processed, based on chunk size, in a server of 4 GHz capacity. The proposed work is evaluated as follows:

**F. 1 Veracity scenario #1 with DCs=1, and BNs=1.**

This scenario is evaluated using the network proposed in Fig. 5; the number of chunks generated per node is assumed to be 25.
Fig. 5 Proposed network topology for scenario 1.
There are 5 to 15 servers per PN, and each chunk requires 1 to 4 GHz of the CPU for processing. The size of each chunk before cleansing is 100 Gb; after cleansing, it is 80 Gb and the RR is assumed to be 0.001 for all chunks. Table III shows the input parameters for this scenario.

| Number of Chunks | Number of servers per PN | Chunk size in Gb | Cleansed chunk size in Gb | RR |
|------------------|--------------------------|-----------------|--------------------------|----|
| 25               | 5-15                     | 100             | 80                       | 0.001 |

The results in Fig. 6 are based on the proposed heuristic and compare the energy efficient power consumption approach with the Transmission Power Consumption (TPC) of the classical approach. The TPC is decreased when the Number of Servers (NS) increases as more chunks are processed locally. At NS = 0, the transmission power consumption of both the classical and the EEBDN-IQ are similar as no additional servers exist in the cities and thus there is no transmission power saving as all chunks are forwarded to the nearest DC according to the Open Shortest Path First (OSPF) algorithm, and are thus still processed in the central DC in Baghdad. At NS = 5, big data traffic becomes smaller due to some of the chunks generated per node being processed locally inside the servers of the corresponding cities, leading to a power saving of around 25%. The remaining unprocessed chunks are forwarded to the nearest DC to be processed there. At NS = 10, most chunks are processed locally and only the corresponding information is transmitted over the network along with a few remaining unprocessed chunks. This is because an increase in the number of servers per city increases the likelihood of locally processed chunks using the Bin-Packing algorithm. This leads to around an 80% transmission power saving. At NS=15, all the generated chunks are processed locally and only the corresponding information is transferred over the network to the nearest DC through the OSPF path. At this number of servers, there is enough processing capacity to process all chunks inside their corresponding cities, without the need for a DC. Thus, the best power saving is obtained, as the transmitted information is of very small size compared to the data quantities sent in the classical approach.
Fig. 6. Classical vs Green network power consumption for veracity scenario 1.

F.2 Veracity scenario #2 with DCs=3, and BNs=1

The number of central DCs is increased to three in this scenario. The topology depicted in Fig. 7, with DCs built at nodes, 1, 3, and 5 is thus considered. Each DC in this topology is connected to several cities chosen for closest distance to the corresponding DC. It is also assumed that the DCs are connected directly, allowing cooperation between them in terms of extra big data analysis and the sharing of information. This assumption is made to study the effects of increasing the number of DCs on the energy efficiency of the network without as used in scenario A except for the change in the number of DCs.

| TABLE IV: Veracity scenario 2 parameters. |
|--------------------------------------------|
| Number of Chunks per node | Number of servers per PN | Chunk volume in Gb | Cleansed chunk volume in Gb | RR |
|----------------------------|--------------------------|-------------------|-----------------------------|----|
|                            |                          |                   |                             |    |

Power saving % heuristics
Fig. 7. Proposed network topology for scenario 2.

Fig. 8 shows the power consumption of the classical network and green one. The general performance is relatively similar to scenario A in terms of power saving, as the number of servers per node and other

parameters are similar. Thus, the same number of chunks are processed locally and centrally using the Bin-Packing algorithm. However, in this scenario, the number of DCs is increased to 3 rather than 1 to reduce the average journey (routing path) of the transmitted chunks of data in the network, thus reducing the transmission power consumption for such traffic. At the point where the number of servers per node in both scenarios is 0, (i.e., the classical approach), the power consumption is still decreased by around 14% from 2,373,100 W in scenario A to 2,039,100 W in scenario B, as seen in Fig. 9. This is because, as mentioned earlier, the transmitted chunks travel from the source to the destination through a shorted path in scenario B compared to scenario A, reducing the number of IP over WDM devices used in the network.

![Graph showing power saving percentage and transmission power consumption](image-url)

- CBN_IQ
- EEBN_IQ
- Power saving % heuristic
Fig. 8: Classical vs Green power consumption for veracity scenario 2.

![Graph showing power consumption comparison between scenarios A and B]

Fig. 9. Scenario 1 vs scenario 2 power consumption.

### CONCLUSIONS

This work presented a heuristic to study an energy efficient big data network in Iraq under the veracity dimension. Several network topologies to employ the proposed method of handling big data traffic locally, intermediately, and centrally were designed. The first scenario assumed the existence of a big central Data Centre (DC) in Baghdad, with all other provinces connected to Baghdad DC via an IP over WDM backbone. Each province is thus not capable of processing big data individually and sends huge volumes of big data to the Baghdad DC for processing and analysing purposes. The transmission power consumption for sending this quantity of data to the central DC in Baghdad was thus calculated. A new topology was then created by adding mini-DC nodes in each city that consist of only a small number of servers compared to the central DC. These servers help to process big data in those cities and reduce the transmission and processing load in the central DC. For each scenario, all the unprocessed chunks are cleansed before processing and a copy of each cleansed chunk forwarded to and stored in a Backup Node (BN). The results of this scenario were compared with the centralised model where no mini-DCs exist in the network. Another possible scenario...
evaluated in the proposed EEBD-IQ is to build two more big DCs to support the central DC in Baghdad, such as in Erbil and Basrah. All scenarios were evaluated in terms of how much electricity is needed to transmit and process big data applications. An average 73% of transmission power saving was found in the two different veracity scenarios compared to the processing power consumption in the initial scenario.

References

[1] B. d. v. K. Normandeau, "variety, velocity and veracity", insideBIGDATA, 2016. [Online]. Available: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/. [Accessed: 09- Nov- 2016].

[2] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3-13, 2000.

[3] W. Raghuopathy and V. Raghuopathy, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, journal article vol. 2, no. 1, pp. 1-10, 2014.

[4] B. Larson, Delivering Business Intelligence. New York, 2009.

[5] P. Zikopoulos and C. Eaton, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[6] A. M. Al-Salim, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficient big data networks: impact of volume and variety," IEEE Transactions on Network and Service Management, 2017.

[7] J. Kepner et al., "Computing on masked data: a high performance method for improving big data veracity," in 2014 IEEE High Performance Extreme Computing Conference (HPEC), 2014, pp. 1-6: IEEE.

[8] L. Berti-Equille, "Data veracity estimation with ensembling truth discovery methods," in 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 2628-2636: IEEE.

[9] A. J. P. o. t. V. E. Gal, "Uncertain entity resolution: re-evaluating entity resolution in the big data era: tutorial," vol. 7, no. 13, pp. 1711-1712, 2014.

[10] A. Abbasi, S. Sarker, and R. H. J. o. t. A. f. I. S. Chiang, "Big data research in information systems: Toward an inclusive research agenda," vol. 17, no. 2, p. I, 2016.

[11] A. M. Al-Salim, A. Q. Lawey, T. El-Gorashi, and J. M. Elmirghani, "Energy Efficient Tapered Data Networks for Big Data Processing in IP/WDM Networks," in Transparent Optical Networks (ICTON), 2015 17th International Conference on, 2015, pp. 1-5: IEEE.

[12] A. M. Al-Salim, H. M. Mohammad Ali, A. Q. Lawey, T. El-Gorashi, and J. M. Elmirghani, "Greening Big Data Networks: Volume Impact," in Transparent Optical Networks (ICTON), 2016 17th International Conference on, 2016, pp. 1-6: IEEE.

[13] A. Al-Salim, T. El-Gorashi, A. Lawey, and J. M. Elmirghani, "Greening Big Data Networks: Velocity Impact," IET Optoelectronics, 2017.

[14] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz, "The case for evaluating mapreduce performance using workload suites," in Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on, 2011, pp. 390-399: IEEE.

[15] A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Distributed energy efficient clouds over core networks," Journal of Lightwave Technology, vol. 32, no. 7, pp. 1261-1281, 2014.

[16] GreenTouch, "GreenTouch Final Results from Green Meter Research Study Reducing the Net Energy Consumption in Communications Networks by up to 98% by 2020," A GreenTouch White Paper, vol. Version 1, 15 August 2015.