**INDEX**

**Index of Figures and Algorithms**…………………………………………….…………………………….

**Abstract** ……………………………………………………………………………………………………………1

**Introduction** ……………………………………………………………..…………………………………………2

1.1 What is Big Data? ...............................................................................................2

1.1.1 Characteristics of Big Data………………………………………………………….……………….2

1.2 Big Data Significance in Industry and Challenges ……………………………………………2

1.3 Challenges in Handling Big Data……………………………………………………………………..3

1.4DataStream …………………………………………………………………………………………………..4

1.5 What is MapReduce? …………...…………………………………………………..……………………………….4

**ANONYMIZATION**………………………………………………………………………………………………..6

2.1 Few privacy prevention techniques used for static data……………………..………….6

2.1.1 .Computer Security …..……………………………………………………………………….……….6

2. 1.2 K-Anonymity ……………………………………….………………………………………………..…..6

2.1.3 l-diversity ……………………………………………………………………………………………………9

2.2 Dynamic Data Privacy Prevention Methods …………………………………………………10

2.2.1 Perturbation ………………………………………………………………………………………..…..10

2.2.2 Tree structure ………………………………………………………………………………..…………11

2.2.3. Adding l-diversity sensitive data…………………………………………………..…………..11

**RELATED WORK** ……………………………………………………………………………….……………….12

3.1 Faanst Algorithm …………………………………………………………………………………………12

3.2 Fads Algorithm …………………………………………………………………………………………….12

**PARAMETERS OF ALGORITHM** ………………………………………………………….………………13

4.1 Data Stream …………………………………………………………………………………………………13

4.2 Cluster ……………………………………………………………………………………….………………..13

4.3 K-anonymized cluster………………………………………………………………..………………….13

4.4 Generalization ……………………..………………………………………………………………………13

4.5 Distance ………….……………………………………………………………………………………………14

4.6 Distance between two tuples ……….……………………………………………………………..15

4.7 Information Loss ………………………………….………………………………………………………15

**PROPOSED ALGORITHM** ……………………………..…………………………………………………….16

5.1 Details of the Algorithm ………………………….……………………………………………………16

5.2 Algorithm …………………………………………………………………………………………………….17

**RESULT AND DISCUSSION** …………………………………………………………………………………19

6.1 Experiment Environment ………………………………………..…….…………………………….19

6.2 Dataset Description……………………………………………………..………………………………19

6.3 Parameters and Improvement ………………………………………………………………….…19

6.4 Future Work ………………………………………………………………..………………………………20

**REFERENCES** ………………………………………………………………………………………………………21

**Index of Figures and Algorithms**

Figure 1. Big data vs data warehouse growth over past 36 months ……………..……..3

Figure 2. Data growth and data handling problems ……………………………………..……..3

Figure 3. Map reduce working ………………………………………………..……………….….………5

Figure 4. University taxonomy tree ………………………………………………………….………..13

Figure 5. Categorical tree …………………………………………………………………………………..14

Figure 6. Taxanomy tree …………………………………………………………………………..……….17

Figure 7. Algorithm camparisions ……………………………………………………………….……..20

Figure 8. Result graph ……………………………………………………………………………..…….…….20

Table 1. Synthetic hospital data ……………………………………………………………………...…..7

Table 2. Synthetic voter data ……………………………………………………………………….………7

Table 3. Complimentary release data …………………………………………………………………..8

Table 4. Hospital data …………………………………………………………………..…………..………..8

Table 5. Random data ……………………………………………………………….……………….………..9

Table 6. Random data ………………………………………………………..……………………….……….9

Table 7. Generalized data with no diversity ………………………………………………..……..10

Table 8. University person …………………………………………………………………………………17

Table 9. Two Anonymized University-person ………………………………………..…………..17