# General layout and chapter’s summary for the thesis

Title: Protecting Privacy in Big Data by Implementing Sensitivity-based Anonymity Framework

Chapter’s summary:

1. Chapter one:
   1. Research Question
   2. Thesis Objectives
   3. Thesis Contribution
   4. Thesis Layout
2. Chapter Two: Background and Challenges
   1. Big Data: Research Challenges
      1. Data Analytics
   2. Big Data analytics Challenges.
      1. MapReduce and Hadoop
      2. Batching and Pig
      3. Streaming and Spark
   3. Security Challenges in Big Data Analytics.
      1. Protecting Privacy by Differential Privacy
         1. Differential Privacy Frameworks
            1. Airavat
            2. GUPT
         2. Possible Attacks in Differential Privacy
      2. Protecting Privacy by K-anonymity
         1. Impairments in k-anonymity
   4. K-Anonymity Frameworks
      1. K-Anonymity Methods for Traditional Data
         1. Generalization
            1. Incognito
            2. L-diversity
         2. Specialization
            1. LKC Privacy
            2. (α, k)-anonymization
         3. Critic of Traditional Data Anonymization Methods
      2. K-Anonymity Method for Big Data
         1. Generalization
            1. Advanced BUG
         2. Specialization
            1. Two-Phase Top-Down Specialization
         3. Critic of Big Data Anonymization Methods
   5. Summary
3. Chapter Three: Multi-Dimensional Sensitivity-Based Anonymization Method
   1. Requirements for Big Data Anonymization Method
      1. Equivalency Increase
      2. Information Gain and Security
      3. Parallel Algorithm
      4. Gradual Access
   2. Multi-Dimensional Sensitivity-Based Anonymization Method Concept
      1. Quasi Identifiers and Classes
      2. Probability Concept and Anonymization Masking
         1. Interval and Taxonomy Tree Masking
         2. Suppression Masking
      3. Grouping Data Vertically and Horizontally
         1. Grouping Data Vertically
         2. Grouping Data Horizontally using Q-ID Groups
      4. Mathematical Equations to Calculate the Sensitivity Level
         1. Sensitivity Level and Time Factor
         2. Anonymization Operations
         3. Taxonomy-tree masking
         4. Discretization Masking
   3. Summary
4. Chapter Four: implementing Sensitivity-Based Anonymization method by Hadoop Ecosystems.
   1. Hadoop in Data Analytics
      1. Hadoop Core
      2. Hadoop Ecosystems
   2. Hadoop security
      1. Implementing Hadoop in LDAP Domain
      2. Applying Kerberos for Hadoop Secure Mode
   3. Deploying Sensitivity-Based Anonymization by using Hadoop Ecosystems
      1. Pig Latin Core Structure
      2. Anonymization by Pig Latin scripts
         1. Pig Latin Script Example
         2. User-Defined Function
         3. Taxonomy Tree Anonymization Algorithm
         4. Interval Anonymization Algorithm
      3. Comparison between Sensitivity-Based anonymization and Other Methods in Big Data
         1. UDF algorithm for MDTDS
         2. UDF algorithm for BUG
         3. Small Data Size experiments
         4. Large Data Size experiments
      4. Anonymization Classification
         1. Using Disruption to Compare between MDSBA and the Other Anonymization Methods
      5. Performance Comparison
   4. Summary
5. Chapter Five: Framework for Sensitivity-Based Anonymization
   1. Security Assertion Markup Language (SAML)
   2. MDSBA and Granular Access Control
      1. Live Data and Archived Data
      2. MDSBA *k̄* percentage and Business roles
      3. MDSBA Three Services
         1. Core Service
            1. Keep or Remove Decision
            2. Initial Service
            3. Subordinate Access
         2. Initializer Service
            1. Generating Pig scripts in Initializer Service
         3. Anonymizer Service
            1. Service Provider Log File and Purge Process
   3. Improvement to MDSBA Security
      1. Obvious Guess
      2. Across Groups Unique Identifiers (AGUI)
      3. Resolving Obvious Guess
      4. Resolving Across Groups Unique Idntifier (AGUI)
   4. Experimenting Data Disruption in MDSBA Framework
   5. Summary
6. Chapter six: Towards Optimal Sensitivity-Based Anonymization for Big Data
   1. Previous Solution to Find the Optimal *k* Value
   2. MDSBA Grouping and the Gradual Access
   3. Possible attacks against MDSBA
   4. Finding the optimal *k* value
   5. Cumulative Frequency
   6. Linear Regression
      1. Three security levels
   7. Finding the Optimal *k* percentage
   8. Dynamic G(QID) groups
   9. Summary
7. Chapter seven: Comparison between MapReduce and Spark in Processing Sensitivity-based Anonymization Framework
   1. Analytics in Big Data
   2. Spark Structure
   3. MapReduce and Spark
   4. Data Streaming Versus Batching
   5. Implementing MDSBA in Spark
      1. User Defined Function in MDSBA
      2. Differences between Pig and Spark Algorithms
   6. Comparison between Hadoop ecosystems and Spark
      1. Spark Tuning in MDSBA
   7. Summary
8. Chapter 8: Conclusion