# SKRIPSI

# PENERAPAN ALGORITMA *ORDER PRESERVING ENCRYPTION* PADA LINGKUNGAN *BIG DATA*

**Kevin Arnold**

**NPM: 2016730038**

**PROGRAM STUDI TEKNIK INFORMATIKA**
**FAKULTAS TEKNOLOGI INFORMASI DAN SAINS**
**UNIVERSITAS KATOLIK PARAHYANGAN**
**2020**

UNDERGRADUATE THESIS

# APPLICATION OF ORDER PRESERVING ENCRYPTION ALGORITHM IN BIG DATA ENVIRONMENTS


**Kevin Arnold**

**NPM: 2016730038**

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2020

# PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

**PENERAPAN ALGORITMA *ORDER PRESERVING ENCRYPTION* PADA LINGKUNGAN *BIG DATA***

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 11 Juni 2020


Kevin Arnold
NPM: 2016730038

# LEMBAR PENGESAHAN

# PENERAPAN ALGORITMA *ORDER PRESERVING ENCRYPTION* PADA LINGKUNGAN *BIG DATA*

**Kevin Arnold**

**NPM: 2016730038**

**Bandung, 11 Juni 2020**

**Menyetujui,**

| Pembimbing Utama | Pembimbing Pendamping |
|---|---|
| **Mariskha Tri Adithia, P.D.Eng** | **Dr. Veronica Sri Moertini** |

| Ketua Tim Penguji | Anggota Tim Penguji |
|---|---|
| **Pascal Alfadian, Nugroho, M.Comp.** | **Husnul Hakim, M.T.** |

**Mengetahui,**

**Ketua Program Studi**

**Mariskha Tri Adithia, P.D.Eng**

# PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

**PENERAPAN ALGORITMA *ORDER PRESERVING ENCRYPTION* PADA LINGKUNGAN *BIG DATA***

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 11 Juni 2020

Kevin Arnold
NPM: 2016730038

# ABSTRAK

Pada zaman *big data* dihasilkan data yang berukuran sangat besar dengan sangat cepat. Hal ini menyebabkan beberapa masalah yaitu privasi data dan kebutuhan untuk menganalisis data yang dihasilkan. Privasi data merupakan hal yang harus diperhatikan agar pihak yang tidak berwenang tidak dapat mengakses data. Selain itu terdapat kepentingan untuk dapat mengolah data yang berukuran sangat besar agar dapat dicari pengetahuan dari data tersebut.

Oleh sebab itu diperlukan sebuah metode untuk melindungi privasi data tersebut dari pihak yang tidak berwenang. Salah satu metode untuk melindungi data adalah dengan menggunakan enkripsi. Namun penggunaan algoritma enkripsi pada umumnya tidak terlalu tepat untuk dilakukan untuk data yang berukuran sangat besar. Hal ini disebabkan oleh kebutuhan untuk mendapatkan pengetahuan dari hasil enkripsi tanpa perlu didekripsi. Salah satu algoritma yang dapat digunakan adalah *Order Preserving Encryption*. Algoritma ini menjaga keterurutan *ciphertext* sesuai dengan *plaintext*.

Pada zaman *big data*, sudah tidak tepat untuk dilakukan penambangan data dengan menggunakan cara tradisional. Volume data yang besar dan pertumbuhan data yang cepat tidak dapat diproses menggunakan *framework* pada umumnya. Untuk menghadapi masalah tersebut maka diperlukan *framework* Spark yang memungkinan proses komputasi paralel dalam melakukan penambangan data.

Oleh sebab itu penelitian ini dilakukan untuk mengimplementasikan algoritma *Order Preserving Encryption* pada Spark dan juga menguji hasil penambangan data dengan mencari nilai mean, median, modus, dan *clustering* dengan menggunakan Spark.

Berdasarkan pengujian yang dilakukan didapatkan bahwa algoritma *Order Preserving Encryption* dapat melakukan enkripsi dengan menggunakan kunci yang sudah pernah dibuat sebelumnya. Artinya proses enkripsi dapat dilakukan dengan benar walaupun terjadi penambahan ukuran data. Namun proses enkripsi bisa menjadi salah bila data baru terdapat nilai yang berada di luar dari jangkauan kunci.

Kemudian dari penambangan data sederhana yang menggunakan mean, median, dan modus didapatkan bahwa hasil *ciphertext* yang memanfaatkan keterurutan dan kemunculan elemen dapat langsung dicari hasilnya tanpa perlu dilakukan dekripsi. Namun pada saat menghitung nilai mean, *ciphertext* tidak dapat menghasilkan nilai yang benar. Oleh sebab itu perhitungan nilai mean tidak dapat langsung dilakukan penambangan data, karena nilai perhitungannya dilakukan antar *bucket* sehingga nilai perhitungannya tidak konsisten.

Pada pengujian hasil *clustering* dengan K-means didapatkan bahwa proses pemberian label tidak dapat dilakukan untuk menghasilkan label yang sama antara *plaintext* dengan *ciphertext* karena terjadi pembentukan model yang berbeda. Untuk percobaan terakhir dilakukan untuk mencari konfigurasi penambangan data yang optimal untuk K-means yang dilakukan. Dari percobaan didapatkan bahwa nilai $K$ yang semakin tinggi akan berpengaruh untuk mendapatkan nilai *error* yang lebih kecil. Untuk peningkatan jumlah iterasi yang dilakukan juga akan menyebabkan nilai *error* yang dihasilkan semakin kecil.

**Kata-kata kunci:** OPE,*Order Preserving Encryption*, Enkripsi, Penambangan Data, K-means, Spark

## ABSTRACT

In the era of big data, data generated very large and data growth very quickly. This causes several problems, namely data privacy and the need to analyze the data produced. Data privacy is a matter that must be considered so that unauthorized parties cannot access the data. In addition there is an interest in being able to process very large data so that knowledge can be sought from the data.

Therefore we need a method to protect the privacy of the data from unauthorized parties. One method to protect data is to use encryption. However, the use of encryption algorithms in general is not very appropriate to do for very large data. This is caused by the need to obtain knowledge from the results of encryption without the need for decryption. One algorithm that can be used is Order Preserving Encryption. This algorithm maintains ciphertext sequences according to the plaintext.

In the era of big data, it was not appropriate to do data mining using traditional methods. Large data volumes and fast data growth cannot be processed using a framework in general. To deal with these problems, Spark framework is needed that allows parallel computing processes in mining data.

Therefore this research was conducted to implement the Order Preserving Encryption algorithm in Spark and also test the results of data mining by finding the mean, median, mode, and clustering values using Spark.

Based on tests, it is found that the Order Preserving Encryption algorithm can encrypt using a key that has been made before. This means that the encryption process can be done correctly even if there is an increase in data size. But the encryption process can be wrong if the new data has a value that is outside the key range.

Then from simple data mining operation that uses mean, median, and modus, it is found that the results of ciphertext that utilize the order and appearance of elements can be directly searched for results without the need for decryption. However, when calculating the mean value, ciphertext cannot produce the correct value. Therefore the calculation of the mean value cannot be directly carried out mining the data.

In testing the results of clustering with K-means it was found that the labeling process could not be carried out to produce the same label between the text and the encryption text because a different model was formed. For the last experiment carried out to look for an optimal data mining configuration for K-means conducted. From the experiments it was found that the higher K value would have an effect on getting a smaller error value. Increasing the number of iterations performed will also cause the error value to be smaller.



**Keywords:** OPE, Order Preserving Encryption, Encryption, Data Mining, K-means, Spark

*Dipersembahkan untuk Tuhan YME, keluarga tercinta,*
*dosen-dosen, teman-teman seperjuangan, segala pihak yang terlibat*
*dalam penulisan skripsi ini, serta diri sendiri*

# KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan YME, karena karunia-Nya sehingga skripsi berjudul "Penerapan Algoritma *Order Preserving Encryption* pada Lingkungan *Big Data*" ini dapat selesai dengan baik. Selama penulisan skripsi ini, penulis banyak menghadapi kendala dan berbagai masalah, namun dengan berbagai dorongan dari lingkungan sekitar penulis, akhirnya skripsi ini dapat diselesaikan penulis. Oleh karena itu, penulis ingin mengungkapkan rasa terima kasih kepada orang-orang disekitar penulis, yaitu:

- Keluarga yang terus memberikan dukungan sehingga skripsi dapat diselesaikan dengan baik.
- Ibu Mariskha Tri Adithia dan Ibu Veronica Moertini sebagai pembimbing skripsi yang telah memberikan dukungan dan bimbingan kepada penulis dalam menyelesaikan skripsi ini.
- Bapak Pascal Alfadian dan Bapak Husnul Hakim selaku penguji yang telah memberikan kritik dan saran yang membuat skripsi ini lebih baik lagi.
- Segenap dosen, staf Tata Usaha, dan Pekarya yang terlibat dalam kegiatan pembelajaran selama penulisan skripsi ini.
- Teman seperjuangan skripsi yaitu Cantika Liem dan Chris Eldon yang selalu membantu dalam menyusunan skripsi.
- Seluruh teman mahasiswa yang ikut memberikan hiburan di saat penulis sendang mengalami kejenuhan dalam penulisan.
- Seluruh pihak lain yang secara langsung maupun tidak langsung mendukung penulis dalam menyusun skripsi ini.

Bandung, Juni 2020

Penulis

# DAFTAR ISI

# DAFTAR GAMBAR

# BAB 1

# PENDAHULUAN

## 1.1 Latar Belakang

Seiring dengan perkembangan teknologi, jumlah data yang dihasilkan semakin meningkat. Dari banyaknya data yang dihasilkan dapat dilakukan penambangan data untuk berbagai keperluan. Setiap data yang dihasilkan dapat memiliki privasi, maka data harus dijaga privasinya agar pihak-pihak yang tidak bertanggung jawab tidak dapat menggunakan informasi dari data tersebut. Oleh karena itu diperlukan sebuah metode untuk mengamankan privasi dari data yang ada. Salah satu metode untuk menjaga privasi data adalah dengan melakukan proses enkripsi. Enkripsi adalah sebuah metode yang digunakan untuk mengubah sebuah pesan dari data menjadi bentuk lain yang tidak bisa dimengerti [1]. Selain terdapat proses enkripsi terdapat juga proses dekripsi. Dekripsi adalah proses mengubah pesan yang sudah dienkripsi menjadi pesan yang dapat dimengerti oleh pihak yang berwenang. Dengan enkripsi dan dekripsi, privasi dari data dapat dijaga. Namun hasil dari proses enkripsi pada umumnya kurang tepat digunakan untuk melakukan penambangan data pada lingkungan *big data*. Hal ini dikarenakan setiap data yang sudah dienkripsi perlu dilakukan proses dekripsi terlebih dahulu ketika ingin dilakukan penambangan data.

Penambangan data adalah suatu bidang ilmu yang mempelajari proses pengolahan data menjadi informasi yang penting. Proses penambangan data yang dilakukan dengan mentransformasi data sehingga menghasilkan informasi disebut *knowledge discovery in Database* (KDD) [2]. Terdapat banyak hal yang dilakukan untuk melakukan KDD, misalnya dengan menggunakan metode statistika (mean, median, modus, standar deviasi, dan lainnya), klasifikasi, dan *clustering* atau pengelompokan data misalnya dengan K-means. Dari data yang diolah untuk mendapatkan informasi terdapat dua sifat data yaitu numerik dan kategorikal.

Dalam penambangan data terdapat dua jenis data. Data yang bersifat numerik akan bernilai angka yang dihasilkan dari perhitungan maupun pengukuran, sedangkan data yang bersifat kategorikal adalah data yang bernilai kata-kata untuk mengelompokan misalnya jenis kelamin, agama, ras, dan lainnya.

Salah satu algoritma enkripsi yang dapat digunakan untuk melakukan penambangan data adalah *Order Preserving Encryption*. Algoritma ini berfungsi untuk menjaga keterurutan *ciphertext* sesuai dengan *plaintext* [3]. Dengan algoritma *Order Preserving Encryption* proses penambangan data dapat dilakukan tanpa perlu dilakukan dekrpsi terlebih dahulu terhadap datanya. Hal ini bisa dilakukan sebab *ciphertext* dapat memanfaatkan keterurutan data tanpa perlu melakukan proses dekripsi terlebih dahulu.

Pada Algoritma *Order Preserving Encryption* akan dihasilkan sebuah kunci simetris terlebih dahulu. Sebuah kunci simteris akan digunakan dalam melakukan enkripsi maupun dekripsi. Oleh sebab itu diperlukan cara untuk memastikan sebuah kunci dapat didistribusikan kepada pihak-pihak yang berwenang.

Salah satu masalah dalam menjaga privasi adalah volume data yang besar dan pertumbuhan datanya yang cepat. Lingkup masalah data pada zaman sekarang sudah mengacu kepada era *big data*. *Big data* adalah sebuah terminologi yang digunakan untuk lingkungan yang menghasilkan data yang memiliki volume yang sangat besar dan juga memiliki pertumbuhan data yang sangat cepat [5]. Dengan besarnya volume dan pertumbuhan data yang dihasilkan proses pengolahan data tidak dapat dilakukan dengan cara tradisional.

Pada lingkungan *big data* proses pengolahan data perlu dikerjakan dengan menggunakan *framework* yang sesuai. Terdapat beberapa *framework* yang dapat digunakan untuk memproses *big data*. Framework tersebut adalah *Hadoop* dan *Spark*. Hadoop dan Spark dapat membantu proses pengolahan data pada lingkungan *big data* karena *framework* tersebut dapat melakukan proses komputasi secara paralel. Dengan proses komputasi yang dilakukan secara paralel, data tidak perlu diproses menggunakan satu komputer namun data dapat dipecah-pecah sehingga dikerjakan oleh banyak komputer secara paralel. Namun Spark memiliki keunggulan dibanding Hadoop dalam melakukan proses komputasinya. Hadoop akan melakukan proses komputasi dengan melakukan penulisan dan pembacaan pada penyimpanannya yang disebut *Hadoop Distributed File System* (HDFS), sedangkan pada Spark terdapat konsep *Resilient Distributed Datasets* (RDD) yang membuat *framework* Spark tidak perlu melakukan penulisan dan pembacaan dari HDFS. Hal ini akan membuat Spark dapat melakukan proses komputasi lebih cepat dan efisien dibanding Hadoop.

Oleh sebab itu pada skripsi ini akan dibuat sebuah perangkat lunak yang dapat mengimplementasikan algoritma OPE pada lingkungan *big data* dengan menggunakan *framework* Spark. Penelitian ini akan menguji penambangan data terhadap *ciphertext* yang dihasilkan oleh algoritma *Order Preserving Encryption* dan membandingkan hasilnya terhadap *plaintext*. Penambangan data yang dilakukan adalah dengan mencari nilai mean, median, modus, dan melakukan clustering terhadap *ciphertext*.

## 1.2 Rumusan Masalah

Rumusan masalah penelitian adalah sebagai berikut:

1. Bagaimana cara mengimplementasikan algoritma OPE pada lingkungan *big data* dengan menggunakan *framework* Spark?
2. Bagaimana hasil dari penambangan data terhadap *ciphertext Order Preserving Encryption* dengan mencari nilai mean, median, modus, dan clustering menggunakan K-means dengan menggunakan Spark?

## 1.3 Tujuan

Berdasarkan rumusan masalah, berikut merupakan tujuan yang harus dicapai:

1. Mengimplementasikan algoritma OPE pada lingkungan *big data* dengan menggunakan *framework* Spark.
2. Mengetahui hasil dari penambangan data terhadap *ciphertext Order Preserving Encryption* dengan mencari nilai mean, median, modus, dan clustering menggunakan K-means dengan menggunakan Spark.

## 1.4 Batasan Masalah

Batasan-batasan masalah pada penelitian ini adalah sebagai berikut:

- Seluruh data yang diproses dalam skripsi ini adalah data yang bersifat numerik.

- Penambangan data yang yang dilakukan akan mencari nilai mean, median, modus, dan melakukan *clustering* dengan K-means.
- Proses pendistribusian sebuah kunci tidak dilakukan secara aman.

## 1.5 Metodologi

Metodologi yang digunakan dalam penyusunan penelitian ini adalah:

1. Mempelajari algoritma Order Preserving Encryption.
2. Mempelajari sifat-sifat atau karakteristik dari *big data*.
3. Mempelajari framework Hadoop dan Spark.
4. Melakukan analisis masalah dan pengujian dengan kasus sederhana pada algoritma *Order Preserving Encryption*.
5. Melakukan perancangan perangkat lunak pada lingkungan *big data*.
6. Melakukan implementasi perangkat lunak pada lingkungan *big data*.
7. Melakukan pengujian eksperimental dengan mencari *mean*, *median*, *modus*, dan melakukan *clustering* menggunakan K-means.
8. Menarik kesimpulan dari pengujian yang dilakukan.

## 1.6 Sistematika Pembahasan

Penelitian ini ditulis dalam beberapa topik pembahasan yang disusun dengan sistematika sebagai berikut:

- **Bab 1 Pendahuluan**
  Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.
- **Bab 2 Landasan Teori**
  Berisi studi literatur tentang kriptografi, penambangan data, algoritma *Order Preserving Encryption*, *big data*, Hadoop, dan Spark.
- **Bab 3 Analisis**
  Berisi analisa masalah, analisa skema *Order Preserving Encryption*, contoh kasus *Order Preserving Encryption*, dan gambaran umum perangkat lunak.
- **Bab 4 Perancangan Perangkat Lunak**
  Berisi perancangan program utama dan perancangan program pengujian.
- **Bab 5 Implementasi Dan Pengujian**
  Berisi pengujian perangkat lunak, dataset yang digunakan dalam pengujian, rancangan eksperimen, dan pengujian eksperimen
- **Bab 6 Kesimpulan dan Saran**
  Berisi kesimpulan dari percobaan ayng dilakukan dan saran untuk penelitian selanjutnya.