Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.26, NO.6, Dec. 2016 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2016.26.6.1593

스마트그리드 개인정보보호를 위한 미터링 데이터 비식별화 방안 연구*

이 동 혁,[1†] 박 남 제[2‡]
[1]제주대학교 일반대학원, [2]제주대학교 초등컴퓨터교육전공

A Study on Metering Data De-identification Method for Smart Grid Privacy Protection*

Donghyeok Lee,[1†] Namje Park[2‡]
[1]Graduate School, Jeju National University,
[2]Teachers College, Jeju National University

요 약

스마트그리드 환경에서는 기존 전력망과 정보통신기술이 접목됨에 따라 다양한 보안 위협 요소가 존재한다. 특히, 스마트 미터링 데이터는 사용자의 생활 패턴, 사용 기기 등 다양한 정보를 노출하며, 심각한 개인정보 침해로 이어질 수 있으므로 미터링 데이터에 적합한 비식별화 알고리즘이 필요한 상황이다. 따라서 본 논문에서는 미터링 데이터에 대한 새로운 비식별화 방안을 제안하였다. 제안한 방법은 시간정보와 수치정보를 각각 비식별화 데이터로 처리하여 해당 데이터만으로는 패턴 정보를 분석할 수 없도록 하였다. 또한, 통계처리 및 가용성을 위하여 비식별화된 상태에서도 데이터베이스에서 직접 범위검색, 집계처리 등의 질의가 가능하다는 장점을 가진다.

ABSTRACT

In the smart grid environment, there are various security threats. In particular, exposure of smart meter data can lead to serious privacy violation. In this paper, we propose a method for de-identification method of metering data. The proposed method is to de-identify the time data and the numeric data, respectively. Therefore, it can't analyze the pattern information from the metering data. In addition, there is an advantage that the query is available, such as the range of search in the database for statistical analysis.

Keywords: Smart Grid, Privacy Protection, De-Identification, Utilizing Personal Information

I. 서 론

스마트그리드는 기존의 전력망에 정보통신 기술을 접목한 것으로, 전력 소비자와 공급자간 실시간 정보 교환을 통해 에너지 효율을 높이는 기술이다. 이러한 스마트그리드 기술은 향후 생활에 밀접하게 다가올 전망이며, 효율적인 에너지 소비에 긍정적인 영향을 미칠 것이다. 향후에는 스마트그리드와 정보통신기술이 융합하여 스마트그리드에 필요한 고용량, 고가용성, 고효율성을 보장하기 위한 여러 기술이 도입될 것이다. 한 전력회사의 예비추정 결과 스마트그리드는 200만 인구 기준으로 하루에 22기가바이트 수준의 데이터를 생성할 것으로 예측되었다[1]. 아울러, 이러한 정보량은 시간이 지날수록 증가될 것으로 예

Received(06. 08. 2016), Modified(1st: 08. 30. 2016, 2nd: 10. 26. 2016), Accepted(10. 26. 2016)
* 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(과제번호:NRF-2016R1D1A3A03918513)
† 주저자, bonfard@jejunu.ac.kr
‡ 교신저자, namjepark@jejunu.ac.kr(Corresponding author)

측된다. 이러한 전망에 따라, 향후 스마트그리드 관련 클라우드 시장이 각광을 받을 것으로 예측된다[2,3,4]. 그러나, 향후에 현실화될 스마트그리드는 현재의 전력망 시스템에 비해 여러가지 불확실한 요소를 지닌다. 즉, 기존 전력망에서 가진 보안 위협 외에도 여러 가지 요인에 의해 더 많은 보안 위협이 발생 가능하다. 또한, 정보통신기술의 접목에 따라 새로운 공격 패턴이 발생할 수도 있다[5,6]. 특히, 스마트 미터기에서 얻어진 전기 사용량을 기반으로 언제 외출하고 있는지, 어떤 전자기기를 사용하는지 등 여러가지 라이프 스타일에 대한 유추가 가능하며, 이러한 점은 심각한 개인정보 침해로 이어질 수 있다.

따라서 스마트그리드 환경에서의 미터링 데이터는 중요하게 보호해야 할 데이터로서 인식될 필요가 있으며, 여기에 적합한 보안 기술을 확보하여 안전하게 관리해야 한다.

Fig.1.은 스마트그리드의 데이터 클라우드 모델을 나타낸다[7]. 이러한 모델은 스마트그리드 환경을 클라우드와 접목하여 데이터 관리상의 여러 가지 장점을 제공한다. 하지만 클라우드 기반의 스마트그리드 환경이 정착되기 위해서는 보안에 대한 고려, 특히 개인정보보호에 대한 철저한 대비가 반드시 필요하다. 스마트그리드에 클라우드 환경을 도입하면 클라우드 환경의 특성에 따른 여러가지 보안 위협을 그대로 답습할 수 있기 때문이다.

여기에서, 데이터를 암호화하는 방안을 생각해 볼 수 있으나, 암호화된 데이터는 기본적으로 질의를 통

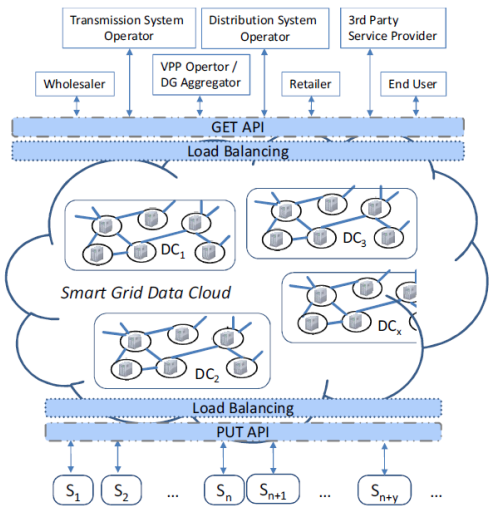

Fig. 1. Smart Grid Data Cloud Model[7]

한 통계 처리가 불가능하다. 암호화된 데이터의 정렬 순서는 평문과 상이하기 때문에, 범위검색, 전방일치 검색, 통계 질의에 대해서는 무의미한 결과를 리턴하기 때문이다. 이러한 문제로 인하여, 현실적으로는 통계 처리 및 가용성 등 개인정보 활용 측면에서 보안성 뿐만 아니라 효율성도 동시에 고려되어야 한다.

따라서, 본 논문에서는 스마트 미터링 데이터에 대한 새로운 비식별화 방법을 제안한다. 제안한 방법은 비식별화된 상태에서 범위검색 및 여러 통계 질의 처리가 가능하다는 장점이 있다. 또한, 미터링 데이터에 대해서 기존 평문에서는 사용자의 전력 소비 패턴 분석 공격이 가능했던 문제점이 존재하였으나, 제안한 방법을 통하여 이러한 문제를 해결할 수 있다.

II. 관련 연구

본 장에서는 먼저 스마트그리드 환경에서 개인정보보호의 필요성에 대해 살펴보고, 비식별화 기술 및 순서보존 암호화 기술에 대하여 살펴보고자 한다.

2.1 스마트그리드 개인정보보호의 필요성

스마트그리드 환경에서는 에너지 효율의 최적화를 위하여 전력 공급자와 소비자가 실시간 정보를 교환한다. 이러한 기능을 수행하려면 여러 관련 정보에 대하여 효율적으로 저장, 처리, 분석이 필요하다[8]. 스마트그리드의 원활한 서비스 제공을 위해, 서비스에 필요한 최소한의 개인정보는 불가피하게 수집하여야 한다. 소비자가 원활하게 서비스를 제공받으려면 공급자는 소비자의 개인정보를 사전에 알고 있어야 하기 때문이다.

'개인정보'는 개인 신상에 관한 모든 정보이며, 그 정보 자체만으로 특정 개인을 식별할 수 있거나, 다른 정보와 결합해서 개인을 식별할 가능성이 있는 정보를 의미한다. 스마트그리드 환경에서는 여러가지 다양한 개인정보가 노출될 수 있다. 대표적으로 발생할 수 있는 개인정보의 예는 Table 1.과 같다. 여기에서, 개인정보의 항목 중 하나로 스마트 미터링 데이터가 존재한다. 스마트 미터링 데이터는 그 자체만으로는 개인정보가 아니라고 생각할 수도 있으나, 사용자의 전력 소비 패턴 분석 등을 기반하여 특정 사용자의 식별 가능성이 존재할 수 있다. 한편, 개인의 식별 뿐만 아니라 개인의 전력 소비 성향, 라이프 스타일, 선호하는 가전기기 등 여러가지 정보도 함께

Table 1. Examples of Personal Information[9]

Item	Examples
Name	The name of user
Address	Where the services provided
Smart Meter Reading	Daily, monthly energy consumption
Finance	Arrears, unpaid charge
Life Cycle	Hour of rising, bedtime, used appliances
Identifier	IP address, network Identifier

파악될 소지가 있어 개인의 사생활 침해와 직결될 수 있으며, 심지어 범죄로의 악용으로도 연결될 소지가 있다. 따라서, 스마트 미터링 데이터도 넓은 의미에서 개인정보로써 함께 취급될 필요가 있다.

Fig.2 는 1일 동안의 스마트 미터링 데이터를 분석한 것만으로 파악된 여러가지 정보를 나타낸다. 여기에는 아침식사는 언제 하는지, 낮에는 어떤 전자제품을 사용하는지(외출여부 포함)등 특정 가정의 라이프 스타일이 그대로 나타난다. 이러한 정보는 심각한 사생활 침해로 이어질 수 있으며, 미터링 정보를 데이터베이스에 저장하고자 하는 경우, 최대한 안전한 방법을 동원하여 저장할 필요가 있음을 시사한다.

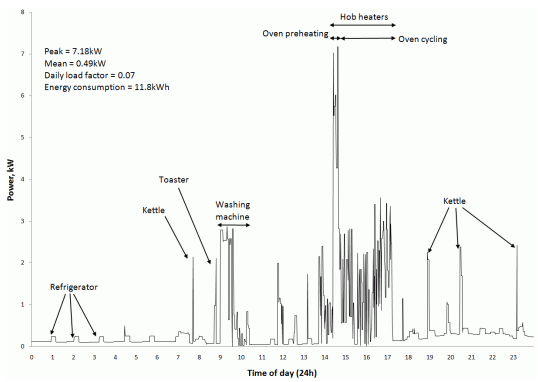

Fig. 2. Security Threats of Metering Data[10]

2.2 비식별화 및 순서보존 암호화 기술

2.2.1 개인정보 비식별화 기술

비식별화(De-Identification)기술이란 개인정보의 침해 위험을 최소화할 목적으로 데이터를 변경하거나 일부를 삭제하여 특정 개인을 식별할 수 없도록 하는 기술을 의미한다. 최근 빅데이터 환경이 도입되면서 방대한 개인정보를 실시간으로 분석하여 비즈니스 활용의 근거를 제공하는 등 개인정보의 침해 위험이 높아지고 있다. 이에 따라, 2013년 9월 안전행정부는 효과적으로 개인정보를 보호하기 위한 목적으로 개인정보 비식별화 기준을 발표한 바 있다.

또한, 최근 NIST에서는 개인정보 비식별화에 관한 보고서인 'De-Identification of Personal Information'을 발표하였다. 또한, 2012년 3월에는 개인정보의 비식별화 가이드라인인 'Protecting Consumer Privacy in an Era of Rapid Change'을 수립한 바 있으며, 개인을 식별가능한 장치와 연관될 수 있는 것은 어떤 경우라도 보호의 대상이 되어야 한다고 규정하였다. 특히, 개인 및 컴퓨터 및 장치에 대한 정보를 식별할 수 있는 데이터의 삭제, 수정, 'noise' 추가, 샘플링 등 적절한 방법으로 반드시 비식별 조치를 취해야 함을 명시하고 있다[11].

일반적으로 개인정보를 비식별화하는 방법은 가명처리(Pseudonymization), 총계처리(Aggregation), 데이터 값 삭제(Data Reduction), 범주화(Data Suppression), 데이터 마스킹(Data Masking) 등이 있다. 그러나 이러한 방법들만으로는 스마트그리드에서의 미터링 데이터로부터 개인정보를 안전하게 보호하기는 쉽지 않다. 알려진 비식별화 방법은 스마트 미터링 데이터의 특성과 적합하지 않다. 미터링 데이터는 사용자의 전력 사용 정보로서, 총계처리나 범주화 등으로 명확하지 않은 값으로 보관하면 서비스의 원활한 제공에 지장이 생길 수 있고, 통계 처리에 있어 명확하지 않은 부분도 발생하기 때문이다. 예컨데, 구체적으로 어떤 시간에 전력이 많이 소비된다는 것을 알고자 한다면, 미터링 데이터를 가급적 상세하고 구체적으로 가지고 있을 필요가 있다. 그러나, 미터링 데이터는 일종의 개인정보로서 평문 그대로 가지고 있는 것은 매우 위험하다. 미터링 데이터를 분석하면 사용자의 습관 등 라이프 스타일을 추정 가능하기 때문에, 결국 해당 데이터의 주체가 누구인지 판별이 가능할 수 있다. 따라서, 미터링 데이터에 적합한 보안 기술이 필요한 상황이다.

2.2.2 순서보존 암호화 기술

전통적인 암호화 알고리즘을 데이터베이스에 적용할 경우, 암호화된 데이터의 순서는 평문과 달라지므로 데이터베이스 인덱스를 구성할 수 없다는 문제가

있다. 이러한 배경에서 순서보존 암호화 기법(Order-Preserving Encryption)이 제안되었다.

이 방법은 암호화된 상태 그대로 인덱싱이 가능하고, 별도의 절차 없이 범위검색, 전방일치검색, 통계질의 등이 가능하게 되었다. 순서보존 암호화 기술 중 대표적인 연구로 Agrawal이 제안한 OPES가 있으며[12], 이는 수치 데이터에 대한 순서를 보존하면서, 분포를 변경하여 원본 데이터에 대한 정보 노출을 최소화한다는 장점이 있다. 그러나, 순서보존 암호화 알고리즘은 평문과 정렬 순서가 같다는 치명적인 단점이 존재한다. 극단적으로, 평문 데이터의 집합과 그 순서를 알고 있다면, 암호화된 데이터베이스에서도 평문의 순서대로 나열하면 결국 동일한 데이터를 얻을 수 있을 것이다. 한편, 순서보존의 특성상 여러 가지 다양한 공격이 가능할 수 있으며, 특정 두 값의 대소 비교만으로도 많은 정보가 노출될 수 있기 때문에 보안성 측면에서 많은 취약성이 존재한다.

2.2.3 미터링 데이터 비식별화의 필요성

향후 스마트그리드 기술이 보급되면서, 데이터 저장공간, 가용성 등 여러 이유로 인하여 스마트그리드가 클라우드 환경과 접목될 것으로 예상된다[13].

Fig.3.에서는 클라우드 환경에서의 스마트그리드 서비스 제공 모델을 나타낸다. 여기에서 서비스 제공자는 클라우드 서비스 제공자 및 신뢰된 제3자가 될 수 있으며 이는 신뢰된 구간 내에 속해 있다. 그러나, 실제 데이터는 신뢰된 구간 밖의 클라우드 데이터베이스에 저장되어 있다. 서비스 제공자는 클라우드 서버간 API 통신을 수행하여 클라우드 서버 내의 데이터에 대한 다양한 처리를 수행한다[14,15].

이러한 클라우드 환경에서는 미터링 데이터 등의 스마트그리드 개인정보가 신뢰된 영역이 아닌 외부에 존재하게 된다. 여기에서 클라우드 서비스는 신뢰되지 않은 영역으로 간주된다. 클라우드에 저장된 데이터의 다양한 보안 위협, 내부자에 의한 데이터 누출 등 여러 위험이 존재할 수 있기 때문이다.[16].

이러한 점은 개인정보에 대한 직접적인 위협요소로 작용한다. 따라서, 데이터를 클라우드 서버상의 데이터베이스에 보관시에는 적절한 방법을 통해 데이터를 보호할 필요가 있다. 즉, 서비스 제공자가 클라우드상에 데이터를 보관 시에는 비식별화한 상태로 보관하며, 평문으로의 복원은 클라우드 서버로부터 데이터를 가져온 이후에 신뢰된 구간 내에서 처리하는 방법이 안전하다.

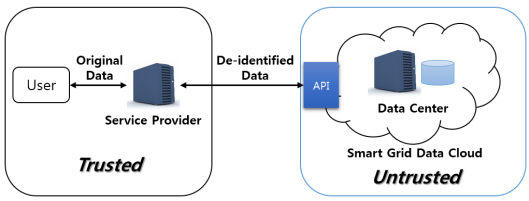

Fig. 3. Smart Grid Cloud Model

물론 가장 안전한 것은 데이터베이스에 있는 값 전체를 암호화하는 것이다. 그러나, 이러한 경우는 데이터베이스의 효율을 심각하게 떨어뜨리고, 실제로 가용성을 기대할 수 없는 상태가 된다. 한편, 가용성을 위해서 평문을 그대로 저장하는 것도 큰 문제가 된다. 개인을 식별할 수 있는 정보가 그대로 노출되는 것이나 마찬가지이기 때문이다[17,18,19].

따라서 일반적으로 강한 수준의 안전이 요구되는 주요 필드에 대해서는 암호화를 수행하고 있다. 그러나, 암호화된 해당 필드는 인덱스를 구성할 수 없고, 범위검색, 전방일치 검색, 집계질의가 되지 않는다는 문제가 존재한다. 이러한 특성은 암호화된 필드의 경우에는 통계처리가 불가능하게 만든다.

스마트그리드의 경우, 전력량에 대한 데이터를 집계처리함에 있어 순서보존 암호화 등의 보완책을 고려해 볼 수 있으나, 이는 순서보존 암호화의 특성상 전력데이터의 분석이 암호화된 상태 그대로 가능하다는 점에서 해결책이 될 수는 없다. 따라서, 스마트그리드 미터링 데이터 특성에 적합한 방법이 필요하다. 비식별화의 큰 장점은 비식별화된 상태 그대로 통계처리가 가능하다는 것이며, 데이터를 기반으로 개인정보의 추적이 어렵다는 특성이 있다. 따라서, 스마트그리드 미터링 데이터는 비식별화 방법으로 처리하는 것이 적합하다[20].

본 논문에서는 이러한 점을 고려하여 미터링 데이터를 비식별화하여 개인의 전력소비 패턴을 분석하기 어렵도록 변경하는 방법에 대하여 제안하고자 한다.

III. 제안 방식

본 장에서는 개인정보를 보호하기 위해 시간 데이터, 전력량 수치 데이터의 두가지 관점에서 미터링 데이터를 비식별화하는 방식을 제안한다.

3.1 제안 방식 개요

3.1.1 개요

Fig.4.에서는 본 논문에서 제안하는 방식의 대략적인 개념을 나타낸다. 클라우드 서버에는 스마트그리드의 미터링 데이터가 저장된다. 여기에는 고객이 사용한 구체적인 전력 사용 내역이 시간대별로 존재한다. 만약, 비식별화가 이루어지지 않은 경우에는 고객의 전력 사용 내역이 그대로 클라우드 서버상에 저장될 것이다. 그러나, 본 논문에서 제안한 방식을 기반으로 적절한 방식의 비식별화를 적용하면 클라우드 서버상에 저장된 정보만으로는 구체적으로 소비자의 사용 내역이 어떠했는지에 대해서는 파악하기 어렵다. 이러한 원본 미터링 데이터는 신뢰된 서버를 통해서만 가져올 수 있다. 신뢰된 서버는 클라우드 서버의 데이터베이스에 대한 질의를 통하여 비식별화된 데이터를 가져오고, 이에 대한 원본 데이터를 복원한다. 이러한 신뢰 서버를 기반으로 사용자는 원본 데이터에 대한 획득이 가능하며, 신뢰 서버는 쿼리 변환을 통해 집계검색 질의문을 구성할 수 있어 통계처리도 가능하다. 또한, 이 과정에서 신뢰된 서버와 클라우드 서버 간 전달되는 질의문 및 질의에 대한 결과값이 중간자공격 등에 의하여 노출이 발생하더라도 원본데이터로 복원할 수 없어 안전하다.

한편, 제안 방식에서는 신뢰된 서버는 안전한 영역으로 가정하고 있다. 따라서, 클라우드 서버와 신뢰된 서버 간에는 대칭키와 의사난수를 공유하고 있으며, 신뢰된 서버에 한해서는 특정 가입자에 대한 질의 및 복원이 가능하다. 따라서, 신뢰된 서버는 비밀정보를 안전하게 관리해야 할 의무를 갖는다.

전체적인 동작 절차는 아래와 같다.

① 미터링 데이터는 실시간으로 비식별화되어 클라우드 서버에 보관된다.
② 사용자의 클라이언트 측에서는 신뢰된 서비스 제공자(신뢰 서버)에게 데이터를 질의한다.
③ 신뢰된 서비스는 질의문을 변환하여 재구성한다.
④ 신뢰 서버 측에서는 변경된 수치 데이터를 가져올 수 있도록 변환된 질의문으로 클라우드 서비스 측의 데이터베이스에 질의한다.
⑤ 클라우드 서비스는 비식별화된 데이터 상태 그대로 신뢰 서버에 리턴한다.
⑥ 신뢰서버는 데이터에 대한 재식별화를 수행한다.
⑦ 재식별화된 데이터를 사용자 측에 전달한다.

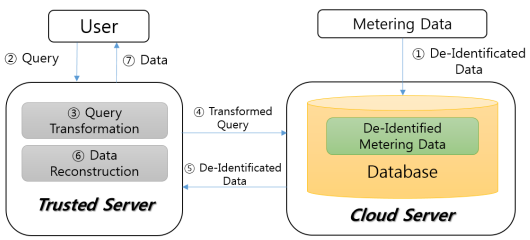

Fig. 4. Overview of Proposed Scheme

3.2 세부 절차

3.2.1 약어

원활한 설명을 위해, 약어를 Table.2에 제시한다.

Table 2. Notation

Abbreviation	Content
s	의사난수 초기 seed
P_n^s	n번째 의사난수 값
GID_n	n번째 그룹 아이디
G_n	n번째 그룹
$H(\cdot)$	해쉬한 결과값
K	신뢰된 영역간 사전 공유된 암호화 키
$E(\cdot)^K$	K를 키로 암호화한 결과값
DT_n	n번째 시간값

3.2.2 비식별화 처리 단계

(1) 시간 정보 암호화 및 그룹화 단계

미터링 정보에서 개인정보 노출 관점에서 유의미한 정보는 시간정보와 시간당 소비 전력량 값이다. 또한, 소비 전력량을 분석하려면 시간 정보에 기반한 분석이 필요하다. 다시 말해, 시간 정보를 알지 못하면 전력 데이터 자체만으로는 의미있는 정보 조합이 어렵게 된다. 따라서, 본 논문에서는 이러한 시간 정보를 암호화하여 저장하도록 한다.

그러나 시간 정보를 암호화하게 되면 시간 정보를 기준으로 범위검색을 수행할 수 없다. 암호화를 수행하면 평문과는 정렬순서가 완전히 달라지므로, 범위검색시 리턴되는 값은 의미가 없게 된다. 즉, 특정 시간 사이에 발생한 전력값에 대한 질의문을 구성하

기가 매우 어려워진다. 만약 데이터베이스의 모든 레코드에 대한 질의를 수행한다면 오버헤드가 매우 커지게 될 것이다.

이러한 문제를 해결하는데 시간단위 그룹화 방법을 사용한다. 즉, 인접한 여러 시간의 데이터를 하나의 단위로 묶어서 그룹으로 한다. 단, 여기에서 그룹 내 데이터의 개수가 균일한 경우는 데이터 분석 공격으로 해당 기간 동안의 전력 사용량이 얼마인지를 분석하여 해당 시간대가 얼마인지에 대한 정보에 대한 대략적인 추정이 가능하다. 이러한 위험을 방지하기 위해, 그룹 내 데이터 개수를 랜덤하게 설정한다. 이러한 데이터 개수는 의사난수를 기반으로 결정할 수 있으며, 여기서, 특정 그룹에 대한 사이즈(해당 그룹이 가진 데이터 개수) 값을 다음과 같이 정한다. 초기 seed와 n의 합, 즉, s+n을 seed로 정하고 이를 s'로 정한다. 이후, $P_n^{s'}$의 결과값으로 그룹의 사이즈를 결정할 수 있다. n번째의 그룹 아이디인 GID_n은 아래와 같이 구한다.

$$GID_n = H(\sum_{i=1}^{n} P_i^s)$$

GID_n을 구하기 위해서는 seed를 구체적으로 알고 있어야 한다. seed와 K는 신뢰된 영역간에 사전 공유하고 있는 값이며, 이러한 사전 정보를 알고 있지 않은 공격자는 정확한 GID_n을 생성할 수 없다.

Fig.5.에서는 그룹 아이디가 추가되고, 시간값이 암호화된 상태를 나타낸다.

Fig. 6에서는 랜덤한 사이즈의 그룹이 적용된 예를 나타낸다. 한 그룹 단위에서 가질 수 있는 데이터의 갯수, 즉, 그룹 사이즈는 랜덤하다.

클라우드 서버 측에서는 비식별화된 데이터의 특정 그룹 아이디에 해당되는 데이터를 카운팅함으로써 각 그룹에 속한 데이터의 개수가 몇 개인지를 알 수

Time period	Usage (KWH)
2/1/2016 1:00	0.385
2/1/2016 2:00	0.365
2/1/2016 3:00	0.425
2/1/2016 4:00	0.5

Group ID	Time period	Usage (KWH)
$H(\sum_{i=1}^{n} P_i^s)$	$E(DT_n)^K$	0.385
$H(\sum_{i=1}^{n+1} P_i^s)$	$E(DT_{n+1})^K$	0.365
$H(\sum_{i=1}^{n+2} P_i^s)$	$E(DT_{n+2})^K$	0.425
$H(\sum_{i=1}^{n+3} P_i^s)$	$E(DT_{n+3})^K$	0.5

Fig. 5. De-Identification of Time Field

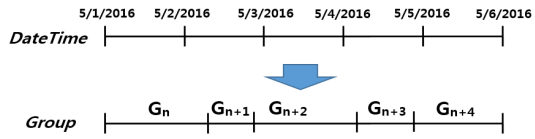

Fig. 6. Group Size Randomization

는 있으나, 그것 자체가 큰 의미를 갖지는 않는다. 그룹 아이디 자체는 연속된 일련의 숫자나 체계가 아니므로 특정 그룹과 다른 그룹의 순서를 연결 지을 수 없기 때문이다. 즉, 의사난수의 seed s를 알지 못하면, 특정 G_n 이후의 그룹 아이디가 G_{n+1}이라는 것을 알 수 없다. 따라서, 데이터는 각 그룹 단위로 독립적이라는 특성을 가지고 있다.

(2) 수치 데이터 변경 단계

이 단계는 다항식을 기반으로 실제 전력량에 대한 수치 데이터를 변경하는 단계이다. 참고로, 전력량은 0보다 큰 임의의 실수로 측정되는 특성을 가진다.

여기에서, $f: X \rightarrow Y$인 함수 f에 대해 X에 속하는 0보다 큰 임의의 실수 데이터 x_1, x_2가 있을 때, $x_1 < x_2$인 모든 x_1, x_2에 대해 항상 $f(x_1) < f(x_2)$ 가 성립하는 강한 단조증가함수가 필요하다. 여기에서, 변경된 수치 데이터는 $P_{n-2}^s, P_{n-1}^s, P_n^s$ 값을 기반으로 아래와 같은 식을 통하여 구한다.

$$f(x) = P_{n-1}^s x^2 + P_{n-2}^s + P_n^s$$

수치 데이터는 이와 같이 다항식을 기반으로 변경한다. 이는 특정 그룹 내에서는 원본 데이터의 분포를 그대로 가지게 되지만 실제 값은 변경되어 데이터를 기반으로 다양한 에너지 분석공격을 어렵게 만든다. 만약 변경 식을 높은 항을 갖는 다차함수로 구성하였을 경우 보안성은 높아지나, 변경된 데이터의 사이즈가 크게 증가할 수 있으며, 이는 데이터베이스 특성에 따른 필드 사이즈 확보 문제 및 연산의 비효율성 문제가 발생할 수 있으므로 적절한 절충이 필요하다. 따라서, 본 논문에서의 변경식은 이러한 목적에 적합한 이차함수로 구성하였다.

P_n^s의 값은 각 그룹단위로 동일하다. 따라서, 단일 그룹 내에 대해서는 순서의 분포는 그대로 유지하고 있다는 특징이 있다. 따라서, 특정 그룹 내부에서

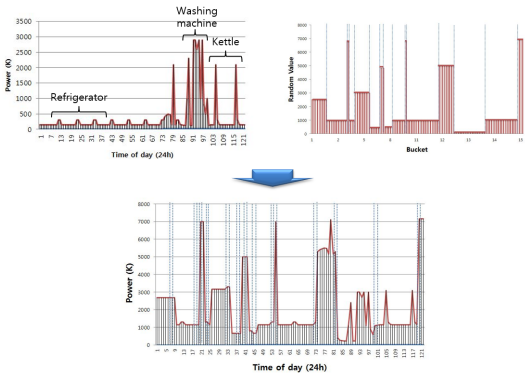

Fig. 7. Transformation of Numeric Data

는 범위검색 등 다양한 통계 질의가 가능하며, 실제로 유의미한 값을 전달받을 수 있다.

Fig.7.은 수치 데이터를 변경한 예를 나타낸다. 원본 데이터에서 냉장고, 세탁기, 전기주전자의 사용 현황을 미터링 데이터만으로 용이하게 분석이 가능하였으나, 변경된 데이터를 통해서는 이러한 분석이 매우 어렵다. 특히, 아래 데이터는 이해를 돕기 위하여 변경된 데이터도 시간 단위로 데이터를 나열한 것이나, 실제 공격자는 사전 공유된 키 K를 알지 못하므로 미터링 데이터를 그림과 같이 시간 순으로 나열할 수 없다. 따라서, 공격자는 전력 분석 공격을 할 수 없게 된다.

3.2.3 비식별화 데이터 질의 및 복원 단계

비식별화된 데이터베이스에 신뢰된 서버가 접근하여 질의를 하고자 할 경우, 우선 해당 데이터에 대한 그룹 아이디를 구해야 한다. 여기서 신뢰된 서버는 seed를 알고 있으므로, 이를 기반으로 그룹 아이디를 추출할 수 있다. 만약 질의하고자 하는 데이터가 특정 그룹 내에 속해 있을 경우, 단 1회의 질의만 필요하여 질의 횟수는 평문을 대상으로 질의할 때와 동일하다. 즉, 그룹 아이디를 검색 조건으로 하여 해당 그룹내의 전체 데이터를 가져올 수 있으며, 사전에 암호화되었던 시간정보는 복호화를 통하여 원본으로 되돌릴 수 있다. 이후, 변경된 수치데이터를 기반으로, 원본 수치 데이터로의 복원이 필요하다.

여기에서, $f(x) = y$ 로 하였을 때, 원본 수치데이터인 x 를 구할 수 있는 역함수 $f^{-1}(y)$는 아래와 같이 구할 수 있다.

앞서, 원본 수치데이터가 다음과 같은 식으로 변경되었다.

$$f(x) = P_{n-1}^{s} x^2 + P_{n-2}^{s} + P_{n}^{s}$$

여기에서, $y - P_{n-2}^{s} - P_{n}^{s} = P_{n-1}^{s} x^2$ 와 같이 이항하면, 아래와 같은 식으로 변환 가능하다.

$$\frac{y - P_{n-2}^{s} - P_{n}^{s}}{P_{n-1}^{s}} = x^2$$

따라서, 역함수 $f^{-1}(y)$ 은 아래와 같다.

$$f^{-1}(y) = \pm\sqrt{\frac{y - P_{n-2}^{s} - P_{n}^{s}}{P_{n-1}^{s}}}$$

수치 데이터 변경 및 복원의 구체적인 예를 들면 다음과 같다. 만약 P_n^s, P_{n-1}^s, P_{n-2}^s의 값이 각각 254, 691, 759라 가정하였을 때, Fig.5.에 나타난 첫번째 데이터인 0.385를 변경하고자 할 경우, 수치 변경 공식에 따라, 변경된 수치 데이터를 아래와 같이 계산할 수 있다.

$$(691)(0.385)^2 + 759 + 254 = 1115.42$$

변경된 수치 데이터는 P_n^s, P_{n-1}^s, P_{n-2}^s의 값을 모두 알지 못한다면 복원이 불가능하다. 즉, 해당 데이터는 비식별화된 값이며, 의사난수의 seed를 알지 못하면 원본 데이터로 되돌릴 수 없다. 이러한 변경된 수치 데이터는 seed를 기반으로 의사난수 생성이 가능한 신뢰된 서버 측에서만 제시된 역함수를 통하여 아래와 같이 복원이 가능하다.

$$\sqrt{\frac{1115.42 - 759 - 254}{691}} = 0.385$$

또한, 값에 대한 복원을 하지 않더라도, 비식별화된 데이터베이스 자체에 대한 질의만으로 통계처리에 일부 유의미한 결과를 얻을 수 있다. 원본 데이터가 반드시 필요한 경우는 복원이 필요하나, 만약 최대값 또는 최소값을 기록한 것이 언제인지, 혹은 상위 이용 구간 범위를 알고자 하는 경우, 비식별화된 데이터베이스 자체의 질의만으로 별도의 복호화가 필요없이 통계 처리가 가능하다는 장점이 있다.

IV. 분 석

4.1 질의 효율성 분석

본 절에서는 평문, 암호화 방식, 제안한 비식별화 방식의 세가지 관점에서 각각의 항목을 비교 분석한다.

4.1.1 범위검색

평문에 대한 범위검색은 1회의 SQL Select문을 구성함으로써 모든 정보의 검색이 가능하다. 한편, 제안한 비식별화 방식은 범위에 해당하는 그룹의 개수만큼 Select 문을 구성하는 것으로 처리한다. 만약, 데이터를 암호화 했을 경우는 범위내 모든 Row에 대한 개수만큼의 Select문을 구성해야 한다.

관리자의 입장과 가용성을 고려할때 간단한 SQL 질의만으로 정보를 가져오는 것이 효율적이고 편리하다. 그러나, 데이터의 안전성을 위해서는 암호화를 하는 것이 가장 안전하다. 하지만, 데이터의 암호화 수행 시 정렬순서의 범위검색이 불가능하게 된다는 부분에서 가장 큰 단점이 있다. 제안한 방법은 이러한 부분을 상호 보완해 주는 측면이 있다. 아래의 SQL은 질의로 가져와야 할 범위에 속하는 그룹이 두개일 때를 가정하고 있다.

◉ 평문의 경우 : Select {Field} from {Table} Where {Datetime} Between {A} and {B}
◉ 암호화 방식의 경우 : Select {Field} from {Table} Where $\{E(DT_1)\} = \{E(DT_1)\}$, … , Select {Field} from {Table} Where $\{E(DT_n)\} = \{E(DT_n)\}$
◉ 제안한 방식 : Select {Field} from {Table} Where GroupID = $\{GID_1\}$, Select {Field} from {Table} Where GroupID = $\{GID_n\}$

제안한 방식은 두개의 Select로써 구성되며, 암호화된 방식은 Row 개수만큼의 SQL 질의가 필요하다. 만약 두개 이상의 그룹을 갖는 범위를 질의할 경우에는 해당하는 그룹 개수만큼 $\{GID_1\}$, … , $\{GID_n\}$을 GroupID의 일치검색 조건으로 각각 질의하여 결과를 얻을 수 있다.

4.1.2 집계검색

집계검색은 대표적으로 MIN,MAX,AVG, COUNT,SUM 과 같이 특정 데이터의 집합에 대한 단일한 결과를 리턴하는 질의 방법을 의미한다. 집계검색은 통계처리시 유용하게 사용되는 질의이며, 평문의 경우는 모든 집계검색이 유의미하다. 그러나, 암호화 방식을 적용한 경우는 순서의 불일치로 인하여 집계검색 질의를 적용할 수 없으며, 질의를 한다 하더라도 무의미한 값이 리턴된다. 그러나, 제안한 방식은 비식별화된 데이터에 대해 그대로 질의하더라도 MIN,MAX,COUNT에 대해서는 정상적인 결과를 보장한다는 장점이 있다. 그러나, 수치데이터 자체가 변경되므로 AVG와 SUM에 대해서는 정확한 값이 발생하지 않는다.

4.1.3 질의회수 분석

범위검색과 집계검색시 각각 평문은 한번의 Select문의 질의만으로 데이터를 가져온다. 그러나, 암호화된 데이터는 범위내의 그룹 개수가 c라고 가정할 경우, c개의 그룹 내의 각각의 원소들의 개수를 모두 합산한 만큼의 질의가 필요하므로 매우 비효율적이다. 한편, 제안한 방식은 범위 내 그룹 개수인 c회 만큼의 질의만 수행하게 되므로 평문에 비해서는 다소 횟수가 증가하나 암호화 방식에 질의하는 것에 비하면 매우 효율적이라고 볼 수 있다.

Table 3. Analysis of Query Count

Item	Range Query	Aggregation Query
Plaintext	1	1
Encryption	$\sum_{i=1}^{c} P_i^{s\|\|i}$	$\sum_{i=1}^{c} P_i^{s\|\|i}$
The Proposed Method	c	c^*

* : MIN,MAX,COUNT Query

4.2 안전성 고찰

4.2.1 데이터베이스의 노출

평문 데이터의 경우는 데이터베이스가 노출되면 모든 정보가 노출된다. 따라서, 개인정보는 고스란히

침해될 수 밖에 없다. 만약, 순서보존 암호화가 된 경우라면 평문 자체의 노출은 되지 않으나, 데이터의 순서가 노출되며, 이는 미터링 데이터의 분석을 용이하게 한다. 예를 들어, 특정시간에 집에 있는지에 대한 정보는 순서보존 암호화를 하는 경우라도 그대로 나타나게 될 것이다. 그러나 제안하는 방식은 데이터베이스에서의 노출이 있더라도, 시간 정보를 연결할 수 없어서 비식별화된 데이터만으로는 쉽사리 개인정보를 판단할 수 없다. 모든 시간정보는 전체적으로 암호화가 되어 있으므로, 전력량의 정보와 시간정보를 매핑할 수 없어서 사용자의 패턴을 파악할 수 없기 때문이다.

4.2.2 데이터 카운팅

데이터 카운팅 공격은 일반적인 암호 알고리즘을 적용하더라도 문제가 그대로 발생한다. 즉, 특정 데이터의 분포가 알려져 있을 경우, 해당 데이터가 몇 개인지를 카운팅하는것 만으로 대략적인 원본 데이터를 유추할 수 있다. 이러한 공격은 동일한 평문에서도 각각 상이한 결과가 나타나는 알고리즘을 사용해야 해결이 가능하다. 제안하는 방식은 그룹단위로 랜덤값이 추가되어, 동일한 값이라 할지라도 버킷단위로 값이 다르게 발생한다. 만약, 그룹 사이즈가 일정하다면 데이터 카운팅에 대한 유효한 분석 방법이 존재할 수 있으나, 본 논문에서 제안하는 방식은 가변 사이즈의 그룹화를 적용하여 데이터 카운팅 공격을 어렵게 한다.

4.2.3 질의문 분석 공격

제안된 방식에서는 만약 신뢰된 서버와 클라우드 서버간 질의문을 모두 엿볼 수 있다고 해도, 원본 데이터에 대한 노출은 발생하지 않는다. 신뢰 영역과 클라우드 서버는 그룹 아이디 기반의 질의만 수행되므로 질의문 분석 자체로는 평문을 유추할 수 있는 정보나 어떤 위험한 정보도 노출하지는 않는다. 다만, 질의문상에서 어느 그룹 아이디가 얼마나 빈번히 질의되었는가 하는 부분에 대한 정보는 파악될 수 있다. 이러한 것이 비록 치명적인 부분은 아니나, 안전성의 향상을 위해 신뢰 서버와 클라우드 서버간 암호화된 채널을 통하여 질의문이 전달될 것이 권장된다.

4.2.4 안전성 비교

평문의 경우는 데이터베이스의 노출, 데이터 카운팅, 쿼리 분석 모두 취약함을 보인다. 이는 데이터베이스의 성능을 얻는 대신 안전성은 포기하는 것이라고 볼 수 있다. 한편, 대표적인 순서보존 암호화 알고리즘인 OPES의 경우는 데이터베이스 전체가 노출될 상황이 발생할 경우, 데이터의 순서가 그대로 노출되므로 결코 안전하다고 할 수는 없다. 그러나, OPES의 장점인 원본 데이터의 분포를 변경하는 특성에 따라, 일정 수준의 안전성은 가지고 있다. 한편, 데이터 카운팅에는 취약한 편이며, 이러한 특성은 일반적인 암호화 알고리즘을 적용하더라도 마찬가지이다. 일치검색시 정확한 데이터를 가져오려면 모두 같은 키로 암호화를 해야 하기 때문이다.

제안한 방법은 그룹 아이디 및 시간정보의 암호화 적용을 통하여, 데이터가 노출되어도 시간순으로 수치 데이터를 배열할 수 없으므로 OPES에 비해 안전하다. 또한, 수치 데이터 변경 시 그룹 단위로 랜덤 값이 변경되므로 타 그룹간은 데이터 카운팅 공격을 실시할 수 없으며, 질의문 분석 시에도 그룹아이디 이외에 특별한 정보는 획득할 수 없다는 특성이 있다.

Table 4. Security Analysis

	Database Exposure	Data Counting	Query Analysis
Plaintext	×	×	×
OPES[7]	△	×	△
Encryption	○	×	○
The Proposed Method	○	△	○

○ Support, △ : Semi-Support, × : Not Support

V. 결 론

스마트그리드 환경은 향후 취급해야 할 데이터 증가가 예상됨으로 인하여 앞으로 클라우드 환경을 도입할 가능성이 크다. 그러나, 클라우드 환경을 도입하기에 앞서 보안에 대한 고려가 반드시 필요하다. 보안이 없이는 서비스의 활성화도 기대할 수 없을 뿐더러, 개인정보의 침해로 인해 여러가지 큰 이슈가

발생할 수 있고, 재앙으로 이어질 수도 있다.

특히, 스마트 미터링 데이터는 개인의 사생활 패턴, 사용 기기 등 각종 정보를 노출하고 있으므로 신뢰되지 않은 클라우드 환경을 도입한다면 반드시 비식별화가 필요하다. 따라서 본 논문에서는 스마트그리드의 개인정보 보안취약점인 미터링 데이터에 대한 비식별화 기법을 제안하였다.

이를 위해 먼저 2장에서 스마트그리드 관점에서의 개인정보보호에 대한 필요성을 살펴보고, 비식별화 기법과 기존에 제안되었던 순서보존 암호화 기법에 대하여 살펴보았다. 그리고 3장에서는 새로운 미터링 데이터 비식별화 방법을 제안하였으며, 4장에서는 제안한 기법을 효율성과 안전성 측면에서 분석하였다. 빅데이터와 클라우드 환경이 본격화되면 여러 가지 다양한 보안 이슈들이 발생할 것이다. 본 논문에서는 향후 도입될 클라우드 환경을 고려하여, 스마트그리드에서의 미터링 데이터를 안전하게 보호하고자 하였다.

안전한 스마트그리드 환경을 위해서는 스마트그리드 환경에서 사용되는 여러가지의 다양한 개인정보에 대한 보호가 고려되어야 한다. 본 논문에서는 이 가운데 미터링 데이터의 보호 방법에 중점을 두고 해결방안을 제안하였다. 스마트그리드에서는 미터링 데이터 이외에도 다양한 개인정보들이 존재하며, 이에 적합한 보호 방안들이 연구된다면, 안전한 스마트그리드 환경의 실현의 가능할 것이다.

향후 제안한 방식에 대하여 속도측정을 통한 정량적 평가를 실시할 예정이며, 아울러 다양한 스마트그리드 개인정보의 보호 방법에 대해서도 지속적으로 연구할 예정이다.

References

[1] M. Shargal and D. Houseman, "The Big Picture of Your Coming Smart Grid", Smart Grid News, 5. 2009.

[2] Namje Park, Marie Kim, "Implementation of load management application system using smart grid privacy policy in energy management service environment", Cluster Computing, 17(3), pp. 653-664, Sep. 2014.

[3] Namje Park and Namhi Kang, "Mutual Authentication Scheme in Secure Internet of Things Technology for Comfortable Lifestyle", Sensors, 16(1), pp. 1-16, Dec. 2015.

[4] Dan Keun Sung, Nah-Oak Song, Kab Seok Ko, Jiyoung Cha, Kuk Yeol Bae, Hanseung Jang, "Convergence of Power System Technology and Information Communication Technology in Smart Grid", Communications of The Korea Information Science Society, 31(3), pp.10-21, Mar. 2013.

[5] Namje Park, "UHF/HF Dual-Band Integrated Mobile RFID/NFC Linkage Method for Mobile Device-based Business Application", Journal of KICS, 38(10), pp. 841-851, Oct. 2013.

[6] Namje Park, "Implementation of Terminal Middleware Platform for Mobile RFID Computing", International Journal of Ad Hoc and Ubiquitous Computing, 8(4), pp. 205-219, Nov. 2011.

[7] Rusitschka, Sebnem, Kolja Eger, and Christoph Gerdes. "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain." Smart Grid Communications (SmartGrid Comm), 2010 First IEEE International Conference on. IEEE, 2010.

[8] Namje Park, Jin Kwak, Seungjoo Kim, Dongho Won, and Howon Kim, "WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment", LNCS, Advanced Web and Network Technologies and Applications, 3842, pp. 741-48, Jan. 2006.

[9] NIST, "Guidelines for Smart Grid Cybersecurity, Volume 2 - Privacy and the Smart Grid," U.S. Department of Commerce, pp. 291-473, Sep. 2014.

[10] E. L. Quinn, "Privacy and the New Energy Infrastructure", Social Science Research Network (SSRN), Feb. 2009.

[11] Namje Park, Hongxin Hu, Qun Jin, "Security and Privacy Mechanisms for

Sensor Middleware and Application in Internet of Things (IoT)", International Journal of Distributed Sensor Networks, Volume 2016, 2016.
〔12〕 Agrawal, Rakesh, et al. "Order preserving encryption for numeric data." Proceedings of the 2004 ACM SIGMOD international conference on Management of data. ACM, 2004.
〔13〕 Donghyeok Lee, Namje Park, "Geocasting-based synchronization of Almanac on the maritime cloud for distributed smart surveillance", The Journal of Supercomputing, pp.1-16, 2016.
〔14〕 Namje Park, "Design and Implementation of Mobile VTS Middleware for Efficient IVEF Service", Journal of KICS, Vol. 39C(6), pp. 466-475, June 2014.
〔15〕 Daeseon Choi, Younho Lee, "Privacy Protection Technology for Public Information Open & Sharing", Journal of KIISE : Computer Systems and Theory 41(3), pp.109-115, Jun. 2014.
〔16〕 Hyun-A Park, Dong Hoon Lee, Taik Yeong Chung,"Comprehensive Study on Security and Privacy Requirements for Retrieval System over Encrypted Database", Journal of the Korea Institute of Information Security and Cryptology, 22(3), pp.621-635, Jun. 2012.
〔17〕 Namje Park, "The implementation of open embedded S/W platform for secure mobile RFID reader", The Journal of Korean Institute of Communications and Information Sciences, 35(5B), pp. 785-793, 2010.
〔18〕 Donghyeok Lee, Namje Park, "A Proposal of SH-Tree Based Data Synchronization Method for Secure Maritime Cloud", Journal of the Korea Institute of Information Security & Cryptology, 26(4), pp. 929-940, Aug. 2016.
〔19〕 Namje Park, Hyo Chan Bang, "Mobile middleware platform for secure vessel traffic system in IoT service environment", Security and Communication Networks, 9(6), pp. 500-512, April 2016.
〔20〕 Dongkook Kim, Hyeok Lee, "Personal Information De-Idenification Trends based on Big Data", Review of Korean Society for Internet Information, 16(2), pp.15-22, Dec. 2015.

〈저자소개〉

이 동 혁 (Donghyeok Lee) 정회원
2007년 2월: 동국대학교 전자상거래기술전공 공학석사
2007년 6월~2008년 5월: 한국전자통신연구원 정보보호연구단 연구원
2008년 11월~2015년 6월: KT 플랫폼개발단 과장
2015년 9월~현재: 제주대학교 컴퓨터교육전공 박사과정
〈관심분야〉 클라우드 보안, 스마트그리드 보안, 데이터베이스 보안, 해사클라우드 등

박 남 제 (Namje Park) 종신회원
2008년 2월: 성균관대학교 컴퓨터공학과 박사
2003년 4월~2008년 12월: 한국전자통신연구원 정보보호연구단 선임연구원
2009년 1월~2009년 12월: 미국 UCLA대학교 공과대학 Post-Doc, WINMEC 연구센터 Staff Researcher
2010년 1월~2010년 8월: 미국 아리조나(ASU) 주립대학교 컴퓨터공학과 연구원
2010년 9월~현재: 제주대학교 교육대학 초등컴퓨터교육전공 교수
〈관심분야〉 융합기술보안, 컴퓨터교육, 스마트그리드, IoT, 해사클라우드 등