Business Intelligence를 지원하기 위한 Big Data 기반 Data Lake 플랫폼의 선행 연구

이 상 범[O]
[O]조선대학교 미술대학 디자인학부
e-mail:kinglee@chosun.ac.kr[O]

A Leading Study of Data Lake Platform based on Big Data to support Business Intelligence

Lee,Sang-Beom[O]
[O]Design ManagementCollage of Art,Chosun University

● 요 약 ●

We live in the digital era, and the characteristics of our customers in the digital era are constantly changing. That's why understanding business requirements and converting them to technical requirements is essential, and you have to understand the data model behind the business layout. Moreover, BI(Business Intelligence) is at the crux of revolutionizing enterprise to minimize losses and maximize profits. In this paper, we have described a leading study about the situation of desk-top BI(software product & programming language) in aspect of front-end side and the Data Lake platform based on Big Data by data modeling in aspect of back-end side to support the business intelligence.

키워드: 비즈니스 인텔리전스(Business Intelligence); 빅데이터(Big Data); 데이터레이크(Data lake)

Ⅰ. 서론

우리는 디지털 시대에 살고 있으며, 디지털 시대에는 고객의 특성이 끊임없이 변한다. 고객들은 항상 인터넷과 가까이 생활하며 데이터를 생성하기도 하고 데이터를 소비하기도 한다. 또한 고객의 요구사항은 점점 늘어나지만 브랜드에 대한 충성도는 점점 낮아지고 있다. 그러기 때문에 비즈니스적인 요구 사항을 이해하고 이를 기술적 요구 사항으로 변환이 필수적이며, 비즈니스 레이아웃 뒤에 숨겨진 데이터 모델을 이해해야만 한다.

데이터 모델 이해를 바탕으로 고객의 컨설턴트를 지원하기 위한 단순한 스프레드시트 이상의 대쉬 보드와 BI 애플리케이션이 필요한 상황이며, 데이터로부터 유용한 시각화를 지원하여야 한다. 따라서 급변하는 비즈니스 환경에 대응하고 경쟁력을 확보할 수 있는 고객 중심 비즈니스로의 전환이 요구되고 있으며, 이것이 바로 데이터 기반 비즈니스 혁신 전략이다.

누구나 손실을 최소화하고 이익을 극대화하고자 하며, 비즈니스 인텔리전스(BI, Business Intelligence)는 혁신적인 기업의 핵심이다. Big Data와 데이터 분석 방법이 개선되어 데이터 분석가와 데이터 과학자는 정보에 입각한 의사 결정을 내리기 위해 점점 더 많은 데이터를 사용하고 있으며, 데이터 분석 방법만 알면 안 되기 때문에 데이터를 비즈니스 자산으로 사용하는 방법을 생각한 다음 적절한 분석을 수행하여 통찰력 있는 BI 솔루션을 구축하고자 한다.

본 논문에서는 BI를 지원하기 위한 전단부의 Desktop의 BI 상황과 후단부의 데이터 모델링을 위한 Big Data 기반 Data Lake 플랫폼에 관한 선행 연구를 기술한다.

Ⅱ. 관련 연구

1. Big Data

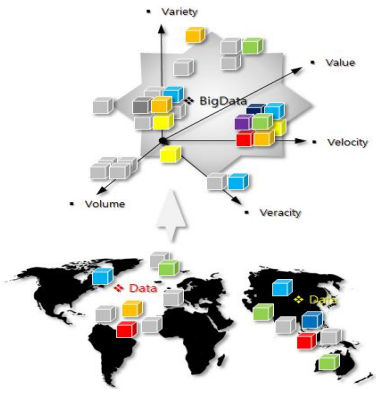

Fig. 1. Big Data의 5가지 속성 정의

Big Data는 시간이 흐름에 따른 다양한 정의가 부가되고 있으며, 추가적으로 사물 인터넷(IoT, Internet of Things) 분야와 Cloud Computing 분야 등에 의해서 가속화되고 있는 상황이다. Big Data의 일반적인 정의는 5 Vs로 정의되며, Volume(대용량 데이터 처리), Velocity(거의 실시간의 고속 데이터 스트림 처리), Variety(다양한 형식의 데이터 처리), Value(데이터의 가치) 및 Veracity(데이터 품질)이다. Big Data의 핵심 사항으로 고도로 구조화된 구조(highly structured)에서 완전히 구조화되지 않은 구조(completely unstructured)에 이르기까지 매우 다른 유형의 데이터를 관리하는 데 매우 적합한 조합을 제공한다[2].

2. BI(Business Intelligence)

비즈니스 인텔리전스는 비즈니스 환경 범위 내에서 분석적 조작 및 데이터 표현을 통해 실용적인비지니스 결정을 내리는 프로세스이며, 또한 BI는 여러 가지의 도구들을 사용해서 획득할 수 있다[1]. BI에 주력한 소프트웨어 제품 출시는 점차적으로 증가하고 있으며, 초기에는 BI를 위해 구축되지 않았지만 이후 업계에 필수 요소로 자리매김한 많은 소프트웨어 제품과 프로그래밍 언어가 있다.

BI를 위한 소프트웨어 제품으로는 Tableau[3], Qlik[4], Power BI[5] 등이 있다. Tableau는 Desktop 도구를 사용해 BI 솔루션을 제공하는데 특화된 소프트웨어이며, 간편한 설치 및 설정, 사용 가능한 데이터와의 연결성을 통한 전달 메커니즘으로 구성된다. Qlik 또한 Desktop 도구를 사용한 BI 솔루션을 제공하며, 자체적인 Desktop 애플리케이션을 통해 데이터와 쿼리(query)를 기반으로 빠른 시각화를 제공한다. 그리고 Power BI는 마이크로소프트의 비교적 새로운 BI 도구이며, Self-Service 솔루션으로 마이크로소프트 엑셀, Microsoft SQL Server와 같이 다른 데이터 소스와 원활하게 통합할 수 있다.

BI를 위한 프로그래밍 언어로는 R[6], Python[7], D3.js[8] 등이 지원한다. R은 무료 오픈소스 통계 프로그래밍 언어이며, 최근에는 데이터 과학과 머신 러닝(machine learning) 분야에서 널리 사용되고 있다. 이러한 이유로 R 언어는 효과적이고 실질적인 BI를 보여주고 제공하는 플랫폼으로 발전이 가속화되고 있다. 또한 R은 알고리즘 및 예측을 이용한 예측 분석을 시각화할 수 있다. Python은 전통적인 프로그래밍 언어이며, 강력한 데이터 분석 및 시각화 모듈을 제공하는 범용 프로그래밍 언어이다. D3.js는 JavaScript 라이브러리이며, 데이터 기반 문서(Data-Driven Documents)를 이용해 웹 기반의 시각화를 지원한다.

3. Data Warehouse, Data Silo, & Data Lake

BI를 지원하기 위한 후반부의 포괄적인 데이터센터 모델로 Data Warehouse, Data Silo, 그리고 Data Lake에 관한 간략하게 기술한다.

Data Lake는 대용량의 원시 데이터(raw data)를 필요로 할 때까지 원시 데이터 형식(raw data format)으로 저장할 수 있는 확장가능한 대용량 스토리지 저장소를 의미한다[9-11]. 계층적 Data Warehouse는 파일이나 폴더에 데이터를 저장하지만, Data Lake는 플랫 아키텍처를 사용하여 데이터를 저장하게 되며, Data Warehouse와 Data Lake의 차이점은 다음의 [표 1]과 같이 나타낼 수 있다.

Table 1. Key differences between Data Warehouse and Data Lake[12]

Items	Data Warehouse	Data Lake
Data	Structured, Processed	Structured / Semi-structured / Unstructured, Raw
Processing	Schema-on-write	Schema-on-read
Storage	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, configure and reconfigure as needed
Security	Mature	Maturing
Users	Business Professionals	Data Scientists et al.

Data Silo는 Data Lake와는 상대적으로 고정된 데이터의 폐쇄적인 저장소로, 한 부서 또는 조직의 통제 하에 있으며 Data Farm의 Silo가 외부 요소와 차단되어있는 것처럼 부서 또는 조직의 나머지 부분과 격리되어 있다[13].

III. BI를 위한 Big Data 기반 Data Lake Framework의 선행 연구 및 초안 설계

BI 컨설팅에서 모든 데이터와 협업 측면에서 요구되어지는 공통된 주제로는 후단부의 데이터 웨어하우스의 모델이 제대로 설계되지 않을 경우, 생산적인 BI 애플리케이션 구축을 위해 Desktop과 같은 전단부에서 얼마나 많은 기술과 노력이 사용했는지는 전혀 중요하지 않다는 것이다. 이에 따라 우선적으로 후반부의 BI를 위한 데이터 설계 및 이를 지원하기 위한 데이터 저장소 시스템의 구축이 매우 중요하다. 더불어 이러한 데이터를 기반으로 전반부의 BI를 위한 유용한 시각화를 개발할 수 있다는 것이다. 본 장에서는 BI를 위한 Data Lake의 정의와 문제점, Data Lake 아키텍처, AWS의 S3(Simple Storage Service) 기반 Data Lake 플랫폼, 그리고 시각화 기능의 사전 연구 내용을 기술한다.

1. Data Lake의 정의와 문제점

Data Lake 개념의 출현은 기업을 위하여 데이터베이스 또는 저장소가 아니라 초점을 거꾸로 뒤집어서 사물들(Things)로 부터 제안되었다. Data Lake는 일반적인 데이터베이스 구조를 먼저 정의한 다음, 이 구조에 맞는 데이터로 데이터를 채우는 대신에 모든 종류의 데이터를 저장 한 다음 필요할 때 이 데이터를 필요한 형식으로 사용할 수 있게 한다는 개념이며, 또한 장기간 패턴 분석을 수행하기 위해 이 데이터를 장기간 보관할 수 있어야 한다. Data Lake 저장소는 데이터 정제소와 함께 임시 기준으로 질의 될 수 있다[2].

Big Data 저장소를 Data Lake를 분류되기 위해서는 다음의 3가지 핵심 특징을 갖추어야 한다[14]:

- 일반적으로 DFS(Distributed File System)에 저장된 데이터의 단일 공유 저장소
- 오케스트레이션(Orchestration) 및 작업 스케줄링 기능

• 데이터의 사용, 처리 또는 작동하는 일련의 응용 프로그램 또는 워크플로우(Workflow)를 포함

Data Lake는 데이터를 원래 형태로 보존하고 데이터 수명주기(Data Lifecycle) 전반에 걸쳐 데이터 및 문맥적 의미(contextual semantics)에 대한 변경 사항 등을 기록하여야 한다. Data Lake는 오케스트레이션 기능과 YARN(Yet Another Resource Negotiator)[15] 등을 통한 작업 스케줄링 기능을 포함하여야 한다. 작업 부하 실행은 기업의 전제 조건이며 YARN은 자원 관리 및 Hadoop 클러스터 전체에 일관된 운영, 보안 및 데이터 거버넌스(Data Governance) 도구를 중앙 플랫폼에서 제공하며, 분석 워크 플로가 필요한 데이터 및 컴퓨팅 자원에 액세스 할 수 있도록 보장하여야 한다. 또한 쉬운 사용자 접근은 사실상 Data Lake의 특징 중 하나이며, 조직은 구조화, 비구조적 또는 반 구조화 여부에 관계없이 원래 데이터 형식으로 데이터가 적재되고, 보존 및 저장되어야 한다. 더불어, Data Lake를 위한 가장 일반적인 유스케이스(Use Cases)를 5 가지(EDW Augmentation, Agile Analytics, Enterprise Reporting, Data Monetization, 그리고 Data Science)를 기술하였다.

Data Lake에서 가장 큰 문제점으로 거론되고 있는 주제는 Oneway Data Lake라는 Garbage Dump 문제이며[11], 이를 해결하기 위한 방법으로 BI 프로세싱 과정에 다양한 수학적 이론을 기반으로 하는 알고리즘과 인공 신경망(Artificial Neural Networks) 기술의 적용을 제안하고자 한다.

2. Data Lake의 아키텍처

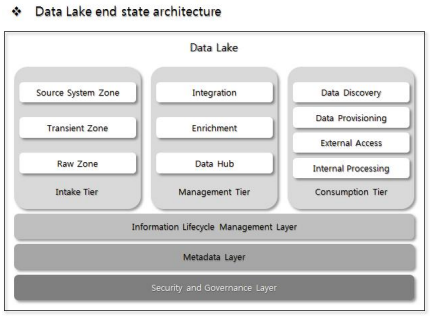

Fig. 2. Pradeep & Beulah의 Data Lake 아키텍처

Pradeep & Beulah은 Data Lake 아키텍처를 [그림 2]과 같이 나타냈으며, Security and Governance Layer, Metadata Layer, 그리고 Information Life-cycle Management Layer의 3개 계층과 Intake Tier, Management Tier, 그리고 Consumption Tier의 3개 티어로 구성하였다[10]. Intake Tier는 Source System Zone, Transient Zone, 그리고 Raw Zone([그림 3] 참조)으로 구성되며, Management Tier는 Integration Zone([그림 4] 참조), Enrichment Zone([그림 5] 참조), 그리고 Data Hube Zone([그림 6] 참조)으로 구성된다. 마지막으로, Consumption Tier는 Data Discovery, Data Provisioning, External Access, 그리고 Internal Processing으로 구성되었다. 그리고 [그림 7]은 Consumption Zone의 특성과 절차를 표현한 것이다.

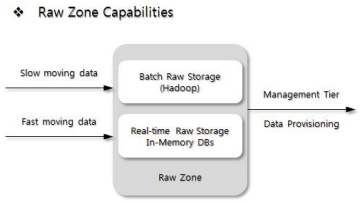

Fig. 3. Raw Zone Capabilities

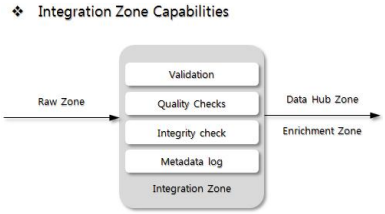

Fig. 4. Integration Zone Capabilities

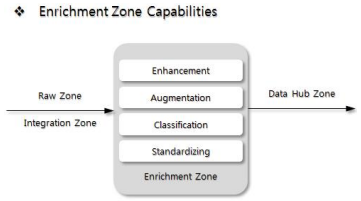

Fig. 5. Enrichment Zone Capabilities

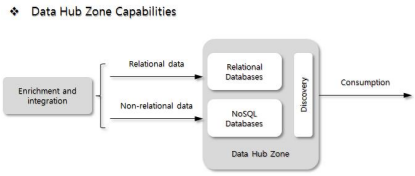

Fig. 6. Data Hub Zone Capabilities

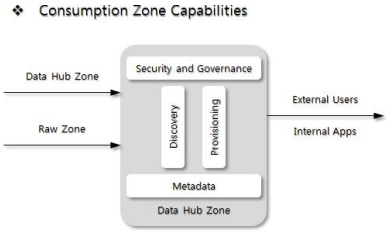

Fig. 7. Consumption Zone Capabilities

3. AWS 기반의 Data Lake 플랫폼

기업 또는 조직에서 점점 더 많은 양의 데이터를 수집 및 분석하면서 기존의 데이터 저장, 데이터 관리 및 분석을 위한 기존 솔루션으로는 더 이상의 속도를 낼 수 없는 상황이며, 특히 Data Silo의 특수성에 의하여 보다 포괄적이고 효율적인 분석을 위한 스토리지 통합을 어렵게 만든다. 이러한 상황은 기업 또는 조직의 민첩성, 데이터에서 더 많은 통찰력과 가치 시슬을 이끌어내는 능력, 기술 진화와 비즈니스의 요구 사항들이 변화됨에 따라보다 정교한 분석 도구와 다양한

프로세스를 원활하게 채택 할 수 있는 능력을 제한하게 된다.

이러한 상황의 타개책 중의 하나로 아마존 AWS(Amazon Web Service)는 Big Data 처리 가능한 Data Lake를 제안 및 BI를 위한 다양한 분석 모듈 및 라이브러리를 서비스로 제공하고 있다[16].

아마존의 S3(Simple Storage Service)[17] 기반 Data Lake 아키텍처는 다음의 항목들을 지원함:

- 다양한 소스의 데이터를 중앙 집중의 플랫폼으로 수집 및 저장
- 포괄적인 데이터 카탈로그를 작성하여 Data Lake 플랫폼에 저장된 데이터 자산을 찾아서 사용
- Data Lake에 저장된 모든 데이터를 보안, 보호 그리고 관리함
- 도구 및 정책을 사용하여 인프라 및 데이터를 모니터링, 분석 및 최적화함
- 원시 데이터 자산을 최적화된 사용 가능한 형식으로 변환 및 데이터 자산의 질의가 가능함
- Amazon Web Services - Amazon Web Services로 Data Lake 구축
- 광범위하고 심도 있는 데이터 분석, 데이터 과학, 기계 학습 및 시각화 도구 포트폴리오를 사용
- 현재 및 미래의 타사 데이터 처리 도구를 신속하게 통합가능
- 처리 된 데이터 세트 및 결과를 쉽고 안전하게 공유 가능

또한 AWS Data Lake 플랫폼은 Storage 계층(Amazon S3, AWS Glue Data Catalog), Serverless Compute 계층(Amazon Kinesis Firehose, AWS Glue, Amazon Redshift Spectrum, AWS Lambda), 그리고 Data Processing 계층(Amazon EMR, Amazon Reshift, Amazon Athena) 의 3 계층에 다양한 서비스들을 제공하고 있다.

4. BI의 시각화 기능

시각화를 위한 데이터 집합을 생성하는 대부분의 로직을 레포트 레벨이 아닌 데이터베이스 레벨로 가지고 갈 때 다양한 잇점을 연속적으로 갖게 된다. 궁극적으로 데이터베이스는 복잡한 로직을 처리하는데 능숙하고 정보 소스에 더 가깝다. 그래서 데이터 품질 및 데이터 유효성 검사 등을 더 수월하게 수행할 수 있게 된다.

이상적인 BI 도구들은 신속하게 데이터 소스에 연결한 후 유용하면서 실용적인 정보를 비즈니스에 신속하게 알려주는 방식으로 차원 및 특정 값을 분할하고 잘라낼 수 있다. 궁극적으로 개인 또는 조직의 BI 도구들의 선택은 도구 사용의 용이성뿐만 아니라 그래프, 챠트, 위젯(Widget), 인포그래픽스(Infographics)[18]와 같은 다양한 컴포넌트들을 통해 데이터를 보여줄 수 있는 유연성에 달려있다.

IV. 결론

디지털 시대에는 고객의 특성이 끊임없이 변한다. 비즈니스적인 요구 사항을 이해하고 이를 기술적 요구 사항으로 변환이 필수적이며, 비즈니스 레이아웃 뒤에 숨겨진 데이터 모델을 이해해야만 한다. 데이터 모델 이해를 바탕으로 고객의 컨설턴트를 지원하기 위한 대쉬 보드와 BI 애플리케이션이 필요한 상황이며, 데이터로부터 유용한 시각화를 지원하여야 한다.

본 논문에서는 BI를 지원하기 위한 전단부의 Desktop의 BI 상황(소프트웨어 제품과 프로그래밍 언어 측면)과 후단부의 데이터 모델링을 위한 빅데이터 기반 데이터레이크 플랫폼에 관한 선행 연구로 Data Lake의 정의와 문제점, Zaloni의 Data Lake 참조 모델, Pradeep & Beulah의 Data Lake 아키텍처, AWS 기반의 Data Lake 플랫폼, 그리고 시각화 기능에 대한 연구 내용을 기술하였다.

REFERENCES

[1] Ahmed Sherif, “Practical Business Intelligence - Learn to get the most out of your business data to optimize your business,” PACKT Publishing, Dec. 2016.

[2] Dirk Slama, Frank Puhlmann, Jim Morrish & Rishi M. Bhatnagar, "Enterprise IoT - Strategies & Best Practices for Connected Products & Services," O'Reilly, 2016.

[3] Tabeau, https://www.tableau.com/

[4] Qlik, http://www.qlik.com/

[5] Power BI, https://powerbi.microsoft.com

[6] R, https://www.r-project.org/

[7] Python, https://www.python.org

[8] D3.js, https://d3js.org/

[9] Natalia Miloslavskaya and Alexander Tolstoy, “Big Data, Fast Data and Data Lake Concepts,” Procedia Computer Science, Volume 88, 2016, Pages 300–305, 2016.

[10] Pradeep Pasupuleti, Beulah Salome Purra, "Data Lake Development with Big Data," PACKT Publishing, 2015.

[11] Bill Inmon, "Data Lake Architecture - Designing the Data Lake and Avoiding the Garbage Dump," Technics Publications, 2016.

[12] Tamara Dull, “Data Lakes vs Data Warehouse: Key Differences,” http://www.kdnuggets.com/ 2015/09/data -lake-vs-data-warehouse-key-differences.html

[13] Data Silo, http://searchcloudapplications.tech tar get.com /definition/data-silo

[14] Zaloni, "Defining the Data Lake - An Introduction to Big Data Lakes and Common Use Cases," http://www. zaloni.com

[15] YARN, https://hortonworks.com/apache/yarn/

[16] "Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility," Amazon Web Service, July 2017,

[17] S3(Simple Storage Service), https://aws.amazon. com/ko /s3/

[18] Infographic, https://en.wikipedia.org/wiki/Infogra phic