Big Data Reduction Framework for Value Creation in Sustainable Enterprises

Muhammad Habib ur Rehman[1], Victor Chang [2], Aisha Batool[3], Teh Ying Wah[1]

[1]Faculty of Computer Science and Information Technology, University of Malaya, KL, 50603, Malaysia
[2]Suzhou Business School, Xi'an Jiaotong Liverpool University, Suzhou, China
[3]Department of Computer Science, Iqra University, Islamabad, Pakistan

Abstract

Value creation is a major sustainability factor for enterprises, in addition to profit maximization and revenue generation. Modern enterprises collect big data from various inbound and outbound data sources. The inbound data sources handle data generated from the results of business operations, such as manufacturing, supply chain management, marketing, and human resource management, among others. Outbound data sources handle customer-generated data which are acquired directly or indirectly from customers, market analysis, surveys, product reviews, and transactional histories. However, cloud service utilization costs increase because of big data analytics and value creation activities for enterprises and customers. This article presents a novel concept of big data reduction at the customer end in which early data reduction operations are performed to achieve multiple objectives, such as a) lowering the service utilization cost, b) enhancing the trust between customers and enterprises, c) preserving privacy of customers, d) enabling secure data sharing, and e) delegating data sharing control to customers. We also propose a framework for early data reduction at customer end and present a business model for end-to-end data reduction in enterprise applications. The article further presents a business model canvas and maps the future application areas with its nine components. Finally, the article discusses the technology adoption challenges for value creation through big data reduction in enterprise applications.

Keywords: sustainable enterprises, value creation, big data analytics, data reduction, business model

1. Introduction

Research shows that 90% of enterprises fail, and that one of the key failure factors is invaluable products and services that do not meet customer expectations (Patel, 2015). A market research on 135 failed startups reveals that 42% of failures occurred because the products and services did not meet the market needs, 17% failed because of the lack of business models, and 14% of the enterprises failed because they ignored their customers (Insights, 2016). Therefore, enterprises must consider maintaining the right product development for the right customers at the right time, have a well-defined business model for revenue generation and profit maximization, and re-evaluate and customize their products and services according to customer requirements (Patel, 2015). Considering these facts, this article addresses the issue of value creation to create sustainable enterprises.

The adoption of Internet of Things (IoTs), big data, and cloud computing technologies by enterprises has led to better value creation at the customer and enterprise ends (Haile & Altmann, 2016; Mital et al., 2016). Value creation for the customer, called value to the customer (V2C), is the process of understanding customer needs and offering them products while considering the competitive advantage over rival enterprises (Cossío-Silva, Revilla-Camacho, Vega-Vázquez, & Palacios-Florencio, 2015; Verhoef, Kooge, & Walk, 2016). Value creation for enterprises, otherwise called value to firm (V2F), is

the process of searching for pitfalls inside enterprise operations and optimizing business process models accordingly (Qi, Qu, & Zhou, 2014). Big data analytics is becoming a key driver of value creation in modern enterprises, wherein enterprise applications are designed to collect direct customer feedback and information from internal business operations (Verhoef et al., 2016). The collected data streams are analyzed using a six-step big data analytic process that continuously evolves to meet the business dynamics and customer requirements. However, the acquisition of big data analytic services from cloud service providers increases financial burden on enterprises, which may lead to the failure of small and medium-sized enterprises (Verhoef et al., 2016).

The main contribution of this article is the concept of early data reduction at the customer and enterprise ends to reduce big data and achieve V2C and V2F objectives. The article presents the background of big data, cloud computing, and IoTs for enterprises to assist readers who may not be familiar with these concepts. A review of the big data analytic process and popular relevant tools for value creation is also provided. The article also presents a novel framework for early data reduction at the customer end wherein the analytic-driven data reduction approaches convert raw data streams into actionable knowledge patterns. The article presents a hypothetical business model to achieve the V2C and V2F objectives of enterprises. Finally, the article presents the business model canvas and maps 10 potential application areas on the business model canvas.

1.1 Big Data for Enterprises

Big data is defined as the set of structured, unstructured, and semi-structured data accumulated from heterogeneous data sources (Yaqoob et al., 2016). Conventionally, big data are presented in terms of 3Vs namely, i) volume, ii) velocity, and iii) value. Volume represents the size of the data whereas velocity represents the speed of data that is entering into big data systems. The value property of big data determines its usefulness to take actionable decisions after data analysis. However, big data is currently redefined with the addition of three new Vs: i) variety, ii) variability, and iii) veracity (Rehman & Batool, 2015; Gani, Siddiqa, Shamshirband, & Hanum, 2016). The variety property defines the multi-facet big data integrating with the different data types generated by various data sources. The variability property determines the internal variability in big data with multiple ‘information shifts’ as time passes. The information shift is defined as the difference between states of knowledge in big data systems. The veracity property shows that big data are collected from authentic and reliable data sources.

Despite considering the three basic Vs of big data, enterprises are adopting big data systems for innovative business models. Modern enterprises collect massive amounts of data from various direct and indirect sources to uncover hidden knowledge patterns and optimize the business process models (Gandomi & Haider, 2015). The direct data sources in enterprises generate operational information relevant to supply chain management, production, fleet management, marketing strategies, behavior analysis of employees, etc. Indirect information includes data collection from click streams, ambulation activities, geo-location information, health records, and many other types of customer-relevant data. Currently, most enterprises collect indirect data from third-party data providers, such as database marketers or market analysis firms. This strategy increases the operational cost of big data systems and creates serious privacy threats, resulting in customer churn and lowering the enterprises’ profits. Therefore, variability, veracity, and variety properties of big data require serious attention, particularly in terms of direct data collection to build trust between enterprises and customers.

Big data help enterprises in profit maximization by optimizing business process models for V2C objective (Vera-Baquero, Colomo-Palacios, & Molloy, 2013). To this end, enterprises use big data mainly for market analysis, customers' segmentations, and personalization. For example, enterprises collect social media data streams such as that provided by Twitter, Facebook, and YouTube. Similarly, enterprises acquire data from e-commerce websites to analyze customers' feedbacks and online product reviews. Big data are also used to perform segmentation of market data to optimize business process models. For example, customer segmentation can assist enterprises in offering products and services to a specific group of customers with similar characteristics. Moreover, big data can also aid in uncovering customer behaviors that enable the design of recommender systems that meet the personal needs of each customer. For example, enterprises analyze click streams of web browsers to uncover customer behaviors and recommend products and services accordingly.

Enterprises use big data to improve the internal business processes to achieve V2F objectives (Vidgen, 2014). On the production side, analysis of machine log files helps in improving the lifetime of machinery and other equipment. Similarly, big data acquired from supply chain management systems help in improving delivery time of products and services. The analysis of big data acquired from employee management systems assists in formulating better and competitive salary plan to retain productive employees. Enterprises integrate big data from multiple internal data sources to improve the overall business models. Big data help in increasing V2F in numerous perspectives; however, uncovering actionable knowledge from big data is a significant challenge that requires laborious efforts to meet value creation objectives.

1.2 Cloud Computing for Enterprises

Cloud computing is the provision of computational, networking, and storage resources to lessen the operational and financial burden of maintaining large-scale computing systems. Cloud computing service providers offer a plethora of services that enable enterprises to deploy business applications and benefit from large scale powerful data centers (Chang, 2014; Sharma et al., 2016). The typical infrastructure of a cloud computing system has three layers: i) infrastructure, ii) platform, and iii) application layer (Chang, Walters, & Wills, 2013). Cloud service providers offer services through all three layers. For example, they provide compute-only services through the infrastructure layer, virtualized platform for application deployment at the platform layer, and generalized application services at the application layer.

Enterprises adopt cloud computing systems to run their business applications optimally and efficiently. The adoption of cloud computing platforms for big data processing is increasing and many new cloud service providers offer big data processing services for enterprises. Big data processing models require huge amounts of computational, networking, and storage resources. Therefore, the adoption of cloud computing technologies for small- and medium-sized enterprises continues to be a challenge because of the high cost of service utilization. Cloud computing technologies can help enterprises in achieving V2C and V2F objectives for profit maximization (Chou, 2015). Cloud computing systems offer a high level of service availability as compared to in-house computing infrastructure, which could increase customers' trust. Alternatively, the enterprises do not need to worry about technology management and instead can focus on product development, customer retention, and operational activities (Chang & Wills, 2016).

1.3 IoTs for Enterprises

IoTs are key drivers for profit maximization through value creation in sustainable enterprises. IoTs systems interact with physical environments to collect useful behavioral and operational information and optimize business process models (Li, Darema, & Chang, 2016). IoTs also enable enterprises to achieve V2C and V2F objectives (Pang, Chen, Han, & Zheng, 2015). For V2C, IoTs aid in optimizing business processes and offering efficient services. For example, IoTs in retail stores help to minimize queuing time for customers. Similarly, IoTs enable shoppers to interact with products to maximize customer retention and build trust. For V2F, IoTs help in optimizing enterprise operations, such as manufacturing processes, supply chain management, and retail operations, to name a few. However, the adoption of IoTs by enterprises has led to the emergence of many use-cases for human-to-machine and machine-to-machine interactions. The convergence of IoTs with big data and cloud computing technologies has taken enterprises to the next level for value creation (Hashem et al., 2015). IoT systems collect massive amounts of data from the in-house and market levels of business operations and transfer them in big data systems, which utilize cloud services to determine actionable insights and improve business process models. Although enterprises collect big data in cloud computing environments, big data analytics remains a key challenge to achieving maximum value creation for customers and enterprises.

2. Big Data Analytics for Value Creation in Sustainable Enterprises

Big data analytics, also known as big data mining, is the process of uncovering actionable knowledge patterns from big data (Wu, Buyya, & Ramamohanarao, 2016). Modern enterprises benefit from big data processes as it provides insights from customer and business data. Big data analytics help in achieving business goals, leading towards customer retention and profit maximization. For example, Twitter uses big data analytics as customer retention tools, in which trending hashtags are mined to engage customers. Similarly, e-commerce enterprises, such as Amazon and Netflix, use big data analytics tools to recommend new and relevant products to users and to maximize their revenue generation (Konstan & Riedl, 2012). Big data analytics help in achieving V2C and V2F objectives because it allows enterprises to perform knowledge discovery operations and improve the internal and external business process models relevant to operations, marketing plans, and workforce and resource management, to name a few.

The big data analytics process, as shown in Figure 1, is based on six major steps: 1) big data are collected from multiple direct and indirect data sources, 2) data preprocessing and integration operations are performed to improve the quality of big data, 3) learning models are generated using statistical methods and machine learning-based data mining techniques, 4) the models are evaluated using test data, 5) the models are deployed in real applications, and 6) the performance of models is monitored in terms of prediction accuracies (Rehman, Khan, & Batool, 2016). The big data analytics process is a continuously evolving process wherein the learning models are regenerated with continuously incoming data to predict shifts in the information. The big data analytics process enables enterprises to uncover the continuously changing knowledge patterns and optimize their business process models accordingly.

Big data Collection: To achieve maximum value creation, enterprises collect maximum data on their customers, product reviews, services feedbacks, industrial monitoring applications, supply chain management systems, and other enterprise operations. The collected data are transferred to cloud data centers to search for actionable knowledge patterns. However, data collection processes should be sufficiently optimal to enable enterprises to avoid collecting irrelevant data. The optimal data collection

strategies help to lower the financial burden of enterprises and ease the computational and storage burden of cloud data centers.

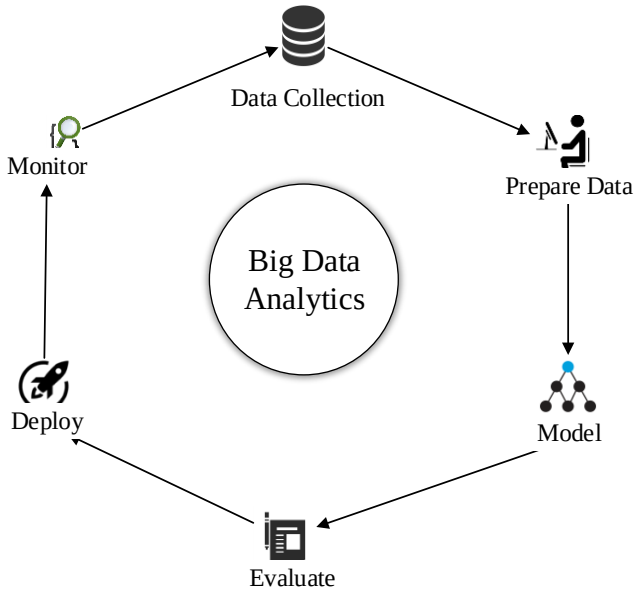

Figure 1: Big Data Analytics Process

Data Preparation: The most important stage of big data analytics is the data preparation stage, wherein data preprocessing and integration operations are performed to improve the quality of big data. Data preprocessing operations include a wide array of methods that are used mainly for the following purposes:

- **Noise Reduction:** Big data collection from IoTs-based sensory data sources and Internet-based social media data streams introduce massive amounts of unstructured and noisy information. Noise reduction methods are applied to remove noise and irrelevant data (Salmon, Harmany, Deledalle, & Willett, 2014).
- **Detecting Outliers:** The presence of outliers (i.e. undesired attributes/values) in big datasets degrades the quality of knowledge patterns and directly affects enterprise business models. Numerous methods are used to detect and remove outliers from big data to produce high-quality datasets (Aggarwal, 2015).
- **Removing Anomalies:** The presence of irregular, unusual, and unwanted data values in big dataset has a significant effect on knowledge quality. Anomaly detection and removal methods are used to improve the quality of big datasets (Moshtaghi et al., 2015).
- **Extracting Features:** Unstructured and continuous data streams in big data systems require considerable effort, and therefore, feature extraction methods are used to separate useful and structured data from raw big data. Depending on the nature and type of data, various statistical methods are used to identify time-domain and frequency domain features from big data (Grzegorowski & Stawicki, 2015).

- **Fusing Data Streams from Multiple Data Sources:** The velocity and data types of big data vary according to data sources, and therefore, intelligent data fusion operations are necessary for data integration and to improve data quality (Yaqoob et al., 2016).
- **Creating Uniform Datasets:** Big data systems collect data streams from multiple data sources in multiple formats. Therefore, data preprocessing operations are performed to convert raw, unstructured, and semi-structured data streams into structured formats.
- **Reducing Dimensions:** Big datasets usually contain thousands and millions of dimensions (i.e. attributes/columns in data tables). Therefore, analyzing such huge datasets can be a challenge. Dimension reduction methods are used to limit the datasets to produce highly relevant datasets for big data analysis (Zhai, Ong, & Tsang, 2014).
- **Handling Missing Values:** Despite the creation of uniformly structured big data sets, huge amount of missing values, which can lower the quality of uncovered knowledge patterns, continue to persist. Data elimination, sketching, and imputation-based methods are used to handle missing values in big datasets (Singh, Javeed, Chhabra, & Kumar, 2015).

Learning Model Generation: Learning models are based on statistical and machine learning theories that are used to study the nature of existing data and to recognize and predict the behaviors of unknown data in the future. Learning models are generated through training datasets that contain similar characteristics as that of future data. The model generation stage ensures the quality of knowledge patterns produced by big data systems.

Evaluation: The trained models are evaluated through different model evaluation methods to ensure that produced models can handle the maximum amount of unknown data.

Deployment: Once generated and evaluated, learning models are deployed in enterprise applications to determine the knowledge patterns from future big data.

Monitoring: The performance of learning models and produced knowledge patterns is monitored continuously through business intelligence (BI) dashboards and reporting tools (Larson & Chang, 2016). Based on the feedback during the monitoring phase, big data analytics process continues to evolve to ensure that information shifts can be handled and newly emerging knowledge patterns are uncovered.

Big data analytics processes vary in terms of descriptive, prescriptive, and predictive analytic models (LaValle et al., 2013). Some examples of big data analytics methods are presented in Table 1. Descriptive analytics are the simplest form of big data analytics, and involve the summarization and description of knowledge patterns using simple statistical methods, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events in big data streams. Descriptive analytics are used mainly at the data preprocessing stage of big data analytic processes to extract features from unstructured data. Predictive analytics methods are based on supervised, unsupervised, and semi-supervised learning models. Alternatively, for prescriptive analytics, enterprises optimize their business process models based on the feedback provided by predictive analytic models. Prescriptive analytics are performed to determine the cause-effect relationship among analytic results and business process optimization policies. Although difficult to deploy, prescriptive analytics contribute to handling the information shift and the continuous evolution of business process models.

Table 1: Data Analysis Methods for Big Data

Type	Methods	Description	Example Methods
Machine Learning	Supervised Learning	The supervised learning methods predict the future events from learning models that are trained using labeled data points. The supervised learning models are trained using labeled data points and tested with leave-one-out, cross-validation, and 5-fold validation methods. The supervised learning models are widely used for data classification and clustering. However, the supervised learning algorithms have the limitations to handle information shifts in big data.	• Neural Networks (Rojas, 2013) • Decision Trees (Barros, Basgalupp, De Carvalho, & Freitas, 2012) • Bayesian Networks (S. H. Chen & Pollino, 2012)
	Unsupervised Learning	The unsupervised learning models are trained using unlabeled data points to predict the future events. The unsupervised learning models are mainly used for data clustering.	• k-means (Jain, 2010) • DB-SCAN (Amini, Wah, & Saboohi, 2014)
	Semi-Supervised Learning	The semi-supervised learning models are initially developed from labeled data points and continuously updated on the feedback from positively predicted events. The adaptive behavior of semi-supervised learning models enables to handle information shift.	• Generative models (Xu, Zhang, Yu, & Long, 2012) • Graph-based • Heuristic-based
	Deep Learning	The deep learning models are a hierarchical representation of supervised and unsupervised learning models. The deep learning models are best suitable for large-scale high-dimensional data. The deep learning models are a good choice when analyzing big data.	• Deep belief Networks (DBNs) (X.-W. Chen & Lin, 2014) • Convolutional Neural Networks (CNNs)
Data Mining	Classification	The classifiers are built with or without learning models and are used to predict the object class of nominal data points.	• Linear Discriminant Analysis (LDA), • Boosting Methods
	Association Rules Mining	The association rule mining methods work in two steps. First, the frequent itemsets are outlined by setting a minimum support threshold value and then the association between itemsets is established by giving a minimum confidence threshold.	• Apriori (M. H. Rehman, Liew, & Wah, 2014) • FP-Growth • AClose
	Regression Analysis	The regression analysis methods are based on statistical theories and are used to establish a relationship between given data points.	• Linear RA (Draper & Smith, 2014) • Non-linear RA
Statistical Methods	Descriptive Statistics	The descriptive statistical methods are used to produce summary statistics using basic statistical operations over whole input data.	• Mean • Median • Standard Deviation
	Inferential Statistics	The inferential statistical methods help to infer the behavior of the whole population by analyzing representative sample data points.	• T-test • Analysis of Variance

Various big data analytics software tools have emerged, and the following is a list of the most commonly used tools:

- **Accenture:** Accenture is an advanced analytics platform that can configure other advanced analytics applications for its users. Accenture provides consultancy and systems integration services to enterprises. Accenture can also rapidly build a huge number of learning models.
- **Alpine Data:** Alpine Data is a native application development platform that offers analytic services for big data by running analytic workflows natively within existing Hadoop systems. Alpine data have high reference scores for innovation, collaboration capabilities, excellent speed in model development, and the ability to model efficiently against a wide range of datasets. Alpine Data provides solutions to clients from banking, services, government and manufacturing sectors.
- **Alteryx:** Alteryx provides data blending and an advanced analytics platform where analysts can integrate internal business processes, third-party tools, and cloud data centers, as well as enable data analytics using some in a single workflow.
- **Angoss:** Angoss provides a suite of advanced analytics tools, including spontaneous, easy-to-use software well-suited for citizen data scientists. Angoss has also achieved significant progress by enhancing its functionality, such as preparation of data in SAS compatible formats. Angoss is a flexible platform capable of providing end-to-end analytics pipeline.
- **BigML:** BigML provides cloud services for numerous machine learning algorithms for correlation analysis, statistical tests, regression analysis, classification, prediction, and clustering. It offers free usage if data tasks are under 16MB. Different subscriptions plans for pay-as-you-go services and virtual private cloud are also available.
- **BIME:** BIME is a visual analytic tool used for big data analysis and visualization. BIME supports 65 data sources, including social media websites, such as Twitter and Facebook, big data stores, such as IBM DB2 and MongoDB, cloud-based analytic tools, such as Amazon's Aurora Web Services, and online storage servers such as DropBox and Google Drive. BIME supports customized visual analytics tools for finance, marketing, product development, sales, and customer support operations of enterprises.
- **Clario:** Clario offers software-as-a-service for the integration of marketing information, web data streams, retail data and e-commerce transactions. Clario provides an online workbench for big data analytics and facilitates the design of online workflow for data analytics and visualization relevant to customer churn operations and market analysis.
- **CoolaData:** CoolaData is a powerful cloud-based big data analytics tool that offers a multitude of services for data integration, analysis, visualization, and prediction. CoolaData offers cloud services for real-time and historical data analysis, as well as support for various open-source big data management tools. CoolaData also offers its own version of SQL, which is tailored especially to meet the behavioral analytics needs of enterprises.
- **CoreMetrics:** IBM provides CoreMetrics, a Software-as-a-Service platform for big data analytics. CoreMetrics offers various services for customer data analysis for data acquired through customer marketing reports, social media, and online customer profile analysis.
- **Data Applied:** Data Applied is a web-based visual data analytics tool that offers Analytics-as-a-Service. Data Applied enables visualization of large datasets, perform data analytics to find correlations, detect anomalies, assess similarities, and uncover association rules. Data Applied

works with Web APIs and CSV files, and enables the deployment of analytic services in public, private, and personal clouds and Intranets.

- **Dell:** Dell addresses the widest set of use-cases for advanced analytics and includes a focus on IoTs and allowing edge deployment of analytic models on gateways or anywhere. Dell provides Hadoop-based execution models for data preparation, as well as for building an analytic model and finding knowledge patterns, and to reduce performance bottlenecks.
- **FICO:** FICO provides advanced analytics platform for big data as well as solutions for decision management, optimization problems, and various analytical applications. FICO's environment for management of models enables audit trial for organizations to track the creation and usage of models.
- **IBM:** IBM is well known for its SPSS statistics and SPSS modeling products. SPSS is a very strong and useful product with a huge user database that is constantly improved through innovations. IBM has high visibility in the advanced analytics platform, as well as through its messaging around intellectual computing and its Watson platform.
- **KNIME:** KNIME (Konstanz Information Miner) provides an open source, desktop-based advanced analytics environment. KNIME also ensures the availability of an additional platform for enterprise application services that could be deployed in a private cloud.
- **Kognitio:** Kognitio is a powerful analytic tool that provides functionalities for in-memory big data analytics. The platform provides three layers of operations: 1) a persistence layer for data storage in cloud data centers, Hadoop clusters, and legacy data warehouses, 2) an analytical platform layer to run SQL and NoSQL queries, and 3) applications and client layers for developing analytic applications.
- **Lexalytics:** Lexalytics is a Software-as-a-Service tool that performs sentiment analysis and named entity extractions from unstructured data. Laxalytics provides analytic services based on machine learning algorithms, text analytics methods, categorization, intention extraction, and summarization. The tool also provides support for natural language processing of 22 languages.
- **Microsoft:** Microsoft offers its predictive analytics capability, which is called SSAS and embedded in the SQL server. This platform provides efficiency in Azure's cloud data source's integration and deployments as a web service, as well as ease of use for data scientists.
- **MicroStrategy:** MicroStrategy provides a unified big data analytics platform whereby the data sets are stored in large-scale Hadoop clusters where users are given access to desktop computers and mobile devices. This tool supports real-time visualization and interactions with BI applications to perform quick decisions.
- **Predixion Software:** Predixion recently launched a product that provides the ability to deploy predictive models in mobile devices or network gateway devices. This capability makes Predixion considerably beneficial with its enhanced ability in extracting information from streamed data. Prediction platform focuses on data-intensive and asset-demanding industries, such as health care, marketing, transportation, and manufacturing.
- **Prognoz:** Prognoz provides software and services for advanced analytics, natively integrated BI, visual discovery functionality, and strong capabilities in forecasting, time series, economic modeling, and financial systems analysis.
- **RapidMiner:** RapidMiner proposes community and basic editions that are open source and free; however, it also offers a commercial edition with additional functionality that is capable of working with large data sets and can be connected to numerous data sources. RapidMiner also

provides a server platform for collaboration because of high performance during processing and integration with business applications.

- **SAP:** SAP's lead product is SAP Predictive Analytics, which has two major components, namely, expert analytics and automated analytics. Expert analytics is a visual workflow tool for customers working with data, whereas automated analytics is a wizard-driven user interface for native data scientist and analysts.
- **SAS:** SAS provides an advanced analytics platform, and has numerous customers and a large system of partners and user. SAS provides quality products with high flexibility and is able to efficiently model huge data sets.
- **SqlStream:** SqlStream offers real-time live analytics of big data, as well as supports Hadoop-based big data systems and legacy data warehousing architectures. SqlStream collects live data streams from machines, devices, and operational information to support real-time online data analytics services for enterprises.

3. Big Data Reduction: Key to Value Creation

Research shows that 57.5% of data scientists spend most of their work time on data preparation, thereby increasing resource conservation in enterprises (CrowdFlower, 2015). The adoption of effective data reduction strategies facilitates workload optimization. Similarly, enterprises can minimize the financial cost of data storage services (Chang & Wills, 2016). Cloud service providers can reduce operational costs by optimizing storage services and minimum in-network data movement (Fu, Jiang, & Xiao, 2012). Another perspective for data reduction is that historical knowledge about customers' behaviors and enterprise operations should be preserved instead of iteratively processing the same raw data. Security breaches and privacy compromises at the enterprises' end were observed recently, thereby posing serious threats to customers and decreasing the trust between enterprises and customers. These considerations indicate that big data reduction can achieve the V2C and V2F objectives.

Enterprises benefit from big data reduction methods in multiple manners. They perform preprocessing information to reduced big data streams before entering in cloud computing systems (Di Martino et al., 2014). They perform dimension reduction methods to address the curse of dimensionality and determine the substantially relevant big data sets (Zhai et al., 2014). They also perform compression and decompression methods to reduce in-network bandwidth utilization in cloud data centers (Yang et al., 2014). Network theory-based methods are used to uncover the semantic relationship between data points in big data to optimize the storage and processing operations in cloud computing systems (Trovati, 2015). Redundancy elimination methods are used to remove duplicated data to improve the value of big data (Fu et al., 2012). Data mining and machine learning methods are used to uncover the knowledge patterns for lateral utilization instead of iterative raw data processing (Jiang et al., 2014).

→ **Preprocessing Operations**: Data preprocessing methods facilitate big data reduction (Brown, 2012; Lin, Chiu, Lee, & Pao, 2013; Cheng, Jiang, & Peng, 2014; Di Martino et al., 2014). These methods are applied immediately after data acquisition. Enterprises can adopt numerous big data preprocessing methods depending upon the application needs. The ontology-based semantic analysis or linked data structures can facilitate intelligent data preprocessing. However, other methods (e.g., low-memory pre-filtration of data streams, filtration of URLs from web browsers data, or 2D peak detection methods) could also be adopted. Data preprocessing, particularly data filtration, could be applied at

the customers' end to minimize big data volume and velocity. Although preprocessing techniques are applicable, these methods are considerably dependent on the nature of big data and their intended use. Therefore, generalizing these methods for all types of enterprise solution is difficult.

→ **Dimension Reduction**: Enterprises collect big data from numerous internal and external data sources; therefore, the emergence of thousands to millions of variable data sets is the norm rather than the exception (Feldman, Schmidt, & Sohler, 2013; Hsieh et al., 2013; Vervliet, Debals, Sorber, & De Lathauwer, 2014; Weinstein et al., 2013). By considering high-dimensionality, big data reduction is mainly considered a dimension reduction problem. The curse of dimensionality necessitates enterprises to acquire additional cloud resources to store and process high-dimensional data. Numerous methods are applied for dimension reduction in big data systems, including online feature selection methods for big data streams, front-end data processing, clustering-oriented machine learning solutions, statistical methods, and implementation of fuzzy logic-based classification methods.

→ **Compression/decompression**: Compression and decompression of big data sets can facilitate the increase in V2F levels by reducing both in-network data movement and storage requirements for long storage objectives (Ackermann & Angus, 2014; Jalali & Asghari, 2014; Yang et al., 2014). However, enterprises still face processing challenges during big data analytics because decompressed big data maintains its original characteristics. The decompressed big data requires huge computational resources and adapts other data reduction methods to increase V2F for enterprises.

→ **Network Theory**: Network theory-based methods are grounded on graph theory. Network theory-based methods convert unstructured big data into structured form and maps on the graph data structures (Patty & Penn, 2015; Trovati, 2015; Trovati & Bessis, 2015). The data are reduced by determining and optimizing the semantic relationships among the graph nodes. Network theory-based methods are technical in nature; therefore, enterprises require highly skilled human resources to adopt these methods.

→ **Redundancy Elimination**: Cloud service providers create copies of big data and store in multiple places to ensure the considerable availability of data in case of network failures or natural disasters (Dong et al., 2011; Xia, Jiang, Feng, & Hua, 2011; Fu et al., 2012; Zhou, Liu, & Li, 2013). However, enterprises need to share the financial burden of duplicate data storage. Redundancy elimination and data deduplication methods are used by cloud service providers to minimize storage costs and improve storage efficiency in big data systems. Data deduplication schemes are applied at different levels in data centers, such as nodes, clusters, racks of clusters, and overall data centers.

→ **Data Mining and Machine Learning**: Enterprises use data mining- and machine learning-based big data reduction methods to achieve different value creation objectives (Jiang et al., 2014; Leung, MacKinnon, & Jiang, 2014; Rágyanszki et al., 2015; Stateczny & Wlodarczyk-Sielicka, 2014). These methods are applied for big data reduction during intelligent data collection, dimension reduction, feature extraction and selections, and artificial intelligence-based optimization techniques. Enterprises are also adopting deep learning models for big data reduction. Deep learning models are initially generated from certain and known big data sets and continuously evolve with uncertain and unknown data sets. However, deep learning models are computationally complex; therefore, enterprises need to acquire additional cloud resources to benefit from deep learning models as data reduction tool.

→ **Data Filtration**: Enterprises adopt data filtration methods for big data reduction whereby the data streams are filtered at the data sources end before even entering big data systems (Antonic, et al., 2014; Papageorgiou, Schmidt, Song, & Kami, 2013). This approach facilitates the increase of V2F for

enterprises by lowering their storage and processing requirements. However, data filtration strategies differ according to the data requirements of enterprises. For example, IoT application filter sensors data streams and web applications perform URL-level filtration. Similarly, manufacturing applications perform filtration operation before data transmission to big data systems.

→ **Representative Data Sensing**: For enterprises that perform data collection using crowd sensing applications, the representative data sensing strategies facilitate the reduction of big data streams (Liu, Iwai, Tobe, & Sezaki, 2013). The enterprise applications need to select considerably useful and quality data producers from the crowd of customers and devices to collect relevant and quality data. The representative data collection strategies facilitate big data reduction in terms of volume and velocity. However, data collection strategies must be adaptive and dynamic enough that the value of big data should not be compromised.

Despite the availability of extensive data reduction methods, enterprises still need to go with high cost service-level agreements. In addition, big data analytics tools are suffering from handling low-quality and substantially unstructured big data sets. This limitation led us to propose a novel big data reduction framework for value creation in sustainable enterprises. This framework is designed to reduce big data at multiple stages without losing its quality.

4. Big Data Reduction Framework for Value Creation in Sustainable Enterprises

The objective of the proposed framework is to enable knowledge-driven data sharing in big data systems to replace raw data sharing (see Figure 2) (Rehman & Batool, 2015). Accordingly, we consider IoT-based big data systems as a case study in establishing the context of the proposed framework.

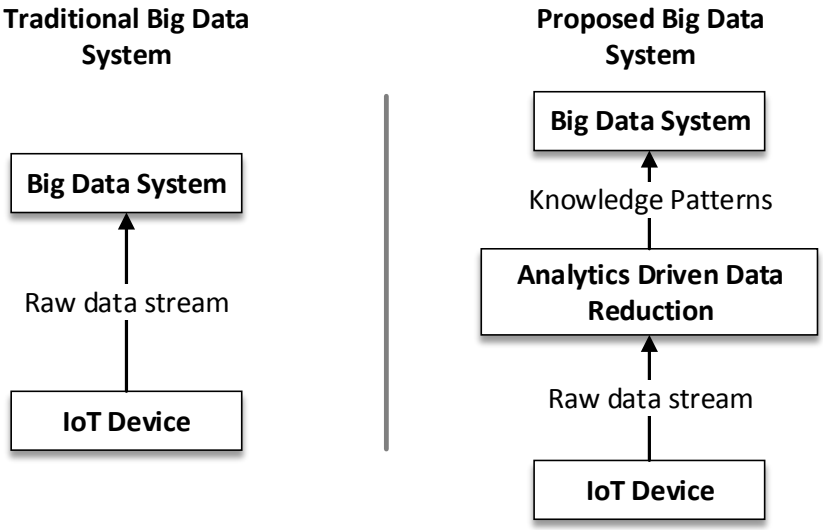

Figure 2: Analytics-driven Big Data Reduction

Recent advancements in cloud computing technologies created opportunities for enterprises to reduce big data streams before data storage in cloud data centers. The mobile edge cloud computing systems that extend the centralized cloud resources near the data sources are being adopted rapidly. Mobile edge cloud computing architectures involve three layers of data processing whereby mobile devices near the data sources provide computational facilities for local data reduction (Akhbar, Chang, Yao, & Muñoz, 2016).

The edge servers reside at the second layer, thereby extending the cloud services from centralized servers and provide cloud services for data reduction on the edge. At the third layer, the centralized cloud computing systems provide computational, networking, and storage services for big data reduction.

We consider the five-layer IoT reference architecture of Fog computing systems introduced by Cisco (see Figure 3) (Luan, Gao, Li, Xiang, & Sun, 2015). The physical layer at the lowest level facilitates in data collection from IoT devices using onboard and off-board sensory and non-sensory data sources. The communication layer at the second level enables connectivity and data transfer from IoT devices to Fog edge servers. The big data aggregation layer provides functionality to aggregate data streams from connecting devices, as well as performs data filtration operations to transfer useful raw data streams in cloud computing systems. The big data analytics layer ensures the availability of data analysis services through cloud service providers. Finally, the application layer provides functionalities to interact with IoT and big data applications.

In the Fog computing architecture, the data sources produce raw data streams that are directly transferred to mobile edge servers. The edge servers perform aggregation and filtration operations to provide distributed intelligence to local mobile devices and reduced data transfer between edge servers and centralized cloud data sources. The centralized cloud computing systems collect data streams from geographically distributed edge servers to perform big data analytics. However, the Fog computing architecture has multiple issues relevant to data reduction. The raw data transfer between mobile devices and edge servers increases the cost of data communication, network traffic, and energy consumption for data transfer in mobile devices. Despite data filtration in edge servers, big data systems still collect raw data and increase the cost of big data analytics. We envision a new knowledge-driven framework for big data reduction. This framework enables the reduction of big data through the provision of analytic support in mobile devices, edge servers, and cloud computing systems.

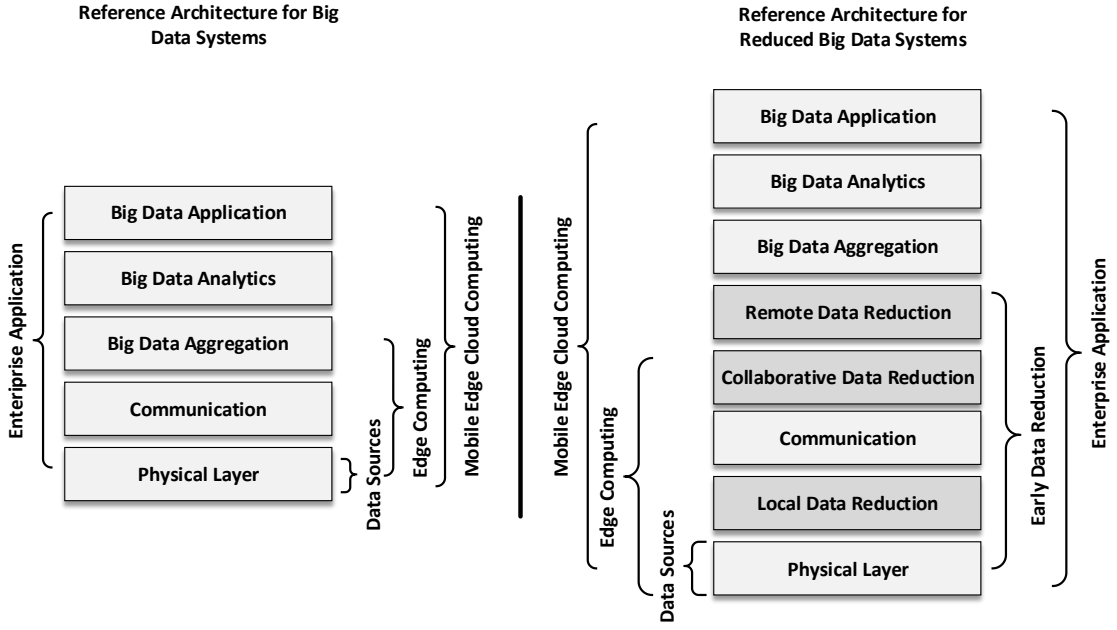

Figure 3: Reference Architectures for Big Data Systems

The proposed framework enables three layers for (1) local data reduction, (2) collaborative data reduction, and (3) remote data reduction. Local data reduction is achieved by deploying analytic components in mobile devices whereby the mobile applications collect, preprocess, analyze, and store knowledge patterns locally. The collaborative data reduction is achieved by deploying analytic components in edge servers whereby the edge servers execute analytics process on locally aggregated knowledge patterns and produce collaborative knowledge patterns. For remote data reduction, the knowledge patterns from edge servers are aggregated and analytics services are executed to determine new knowledge patterns. The resultant knowledge patterns are aggregated in big knowledge stores inside cloud data centers whereby big data applications can access and perform further analytics for value creation.

The proposed framework can facilitate in multiple methods to achieve the V2C and V2F objectives. Traditional business models facilitate value creation at the enterprise side whereby the customers are unaware about the usage and level of the produced knowledge. The proposed framework facilitates local knowledge availability to customers by enabling local analytic components in customers' devices and system. The customers can witness the level of knowledge being produced and collected by enterprises. In addition, local knowledge availability increases the customers' trust on enterprises and creates a direct link between the two entities. Privacy preservation is the main concern of customers; thus, the proposed framework effectively handles the privacy issue. The customer should be provided complete control over knowledge patterns by enabling sharing and subscription to different big data applications. In addition, enterprises should design and develop end-to-end secure data sharing applications to improve customers' trust and acquire maximum knowledge patterns.

The proposed data reduction strategies enable knowledge availability at multiple levels. For example, the customer can benefit from local knowledge availability and enterprises can benefit from the collective knowledge of a group of users connected to the same edge server. Similarly, the global knowledge about customers and their collective behaviors is available at the cloud level. The early data reduction in mobile edge cloud computing systems reduces the computational cost and the cost of data communication and data movement in cloud computing environments. Therefore, early data reduction effectively eases the financial burden of enterprises. Conventionally, small- and medium-level enterprises suffer in adopting cloud-based big data analytics because of the high financial cost of service-level agreements. However, the proposed framework can assist cloud service providers in lowering the cost of cloud services to capture the market for small- and medium-level enterprises. The proposed framework facilitates analytics-driven big data applications; therefore, the BI dashboards and reporting tools experience low latency and improved real-time visualization of operational data and customers' insights.

4.1 Design of a Win–Win Business Model for Sustainable Enterprises

Enterprises strive to achieve a high level of value creation by considering the cost structures of different business operations, such as product development, marketing, and human resource management. However, early data reduction at the customers' end facilitates the cost minimization of IT operations in enterprises (see Figure 4) (Osterwalder, Pigneur, & Tucci, 2005). The enterprise applications collect customer data from multiple data sources and perform data reduction thereafter using customers' devices, edge servers, and cloud computing services. To enable data reduction at the customers' end, the main hurdles include attracting, pursuing, and providing incentives to customers to reduce raw data and share the knowledge patterns. Incentivizing strategies should be adopted to deploy beneficial and feature-rich

applications at the customers' end that fulfill their needs and offer them additional services and benefits in exchange for knowledge patterns. In addition, enterprises need to build trust by ensuring privacy preservation and secure data sharing channels for customer–enterprise relationships. Similarly, the early data reduction of internally produced data can assist enterprises in decreasing financial costs for cloud services acquisitions.

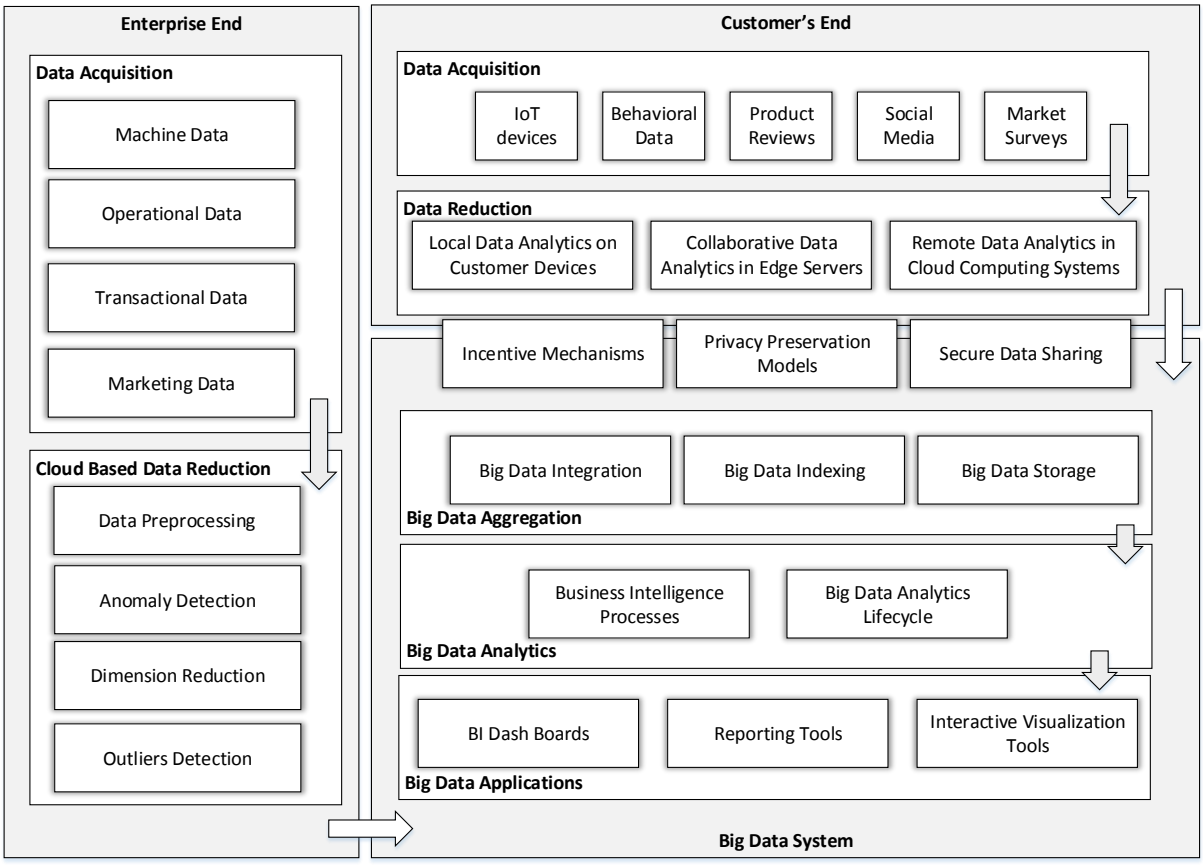

Figure 4: Win–Win Business Model for Value Creation

The early data reduction at the customers and enterprises' ends reduces the cost of cloud service utilization in big data systems. Accordingly, big data management operations (e.g., integration of data from multiple sources and indexing and storage of big data in cloud data centers) become easy. Similarly, big data reduction facilitates the improvement of big data analytics whereby reduced, high-quality, and information-rich data increase the value creation for enterprises and their customers. In addition, BI applications work efficiently with minimum latency, thereby enabling entrepreneurs to immediately optimize the business process models. In addition, the reduced data streams improve the performance of real-time big data applications by cutting down the execution time of handling raw big data.

Figure 5 presents the traditional business model canvas (Zolnowski, Weiß, & Bohmann, 2014). Early data reduction assists in value creation for all components, as well as benefits the key partners (e.g., customers, enterprises, cloud service providers, and other third-party partners) in achieving the V2C and V2F objectives. Such data reduction enables the maximization of the efficiency of performing key activities (e.g., collecting and analyzing customers and organizational data) and optimization of business processes.

The early data reduction facilitates the optimized resource utilization of IT-bound (e.g., big data applications and cloud computing services) and non-IT bound (e.g., human resources and enterprise assets) resources.

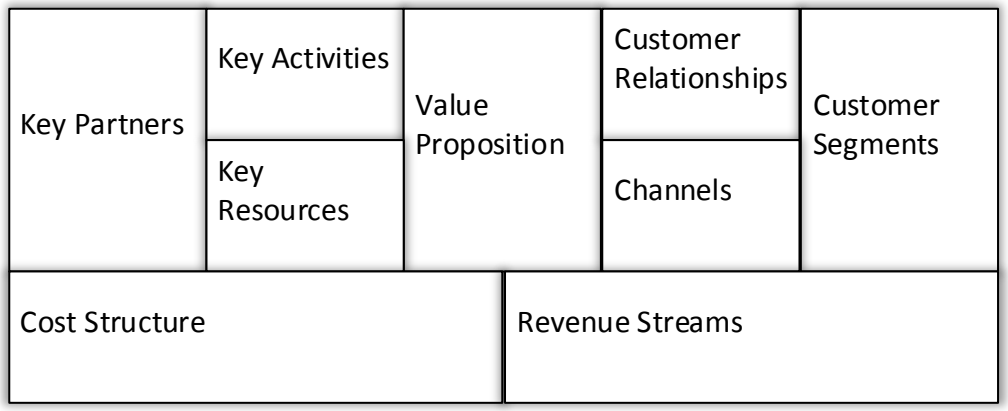

Figure 5: Business Model Canvas

Early data reduction assists in achieving value creation and balances it with the cost of cloud service utilization. Such data reduction facilitates the development of trust between customers and enterprises and opens direct channels of communication and product and service delivery. Similarly, knowledge-based data sharing facilitates the improved behavioral analysis of customers' data at a fine-grained level, thereby facilitating the delivery of personalized products and services to maximize the revenues and build long-term relationships (Chuang & Chen, 2015). Moreover, early data reduction improves the overall cost structure of enterprises by balancing the operational and optimization costs of business process models. Similarly, such reduction increases the revenue streams and profit maximization because enterprises can devise the appropriate strategies for potential customers at the opportune time. Table 2 maps the business canvas model with 10 potential application areas for the proposed data reduction framework. However, the business models could be applied to any other application domain involving direct product and service provisions to customers, such as assisted living, waste management, smart energy management, education, banking, mobile health, e-government, and smart buildings, among others.

Enterprises need to address the following issues for maximum value creation and technology adoption.

Customer and Market Knowledge: Enterprises must perform a detailed customer and market survey to determine the key requirements of customers and the viability of products and services. The survey could be performed via third party market analysis firms or releasing the premature testing versions of the products. The early product releases assist enterprises in immediately improving value creation and introduce customers with upcoming releases.

End-to-End Analytic Services for Data Reduction: Enterprises must adopt a complete end-to-end framework for data reduction, thereby enabling the immediate reduction of the data stream. IT managers and the data scientists should design a complete execution model whereby all components must be defined and selected before the actual product development. The proposed framework enables analytic-driven data reduction; therefore, sufficiently accurate and relevant data must be acquired to train the learning models at every stage of data reduction. The learning models for local, collaborative, and remote

Table 2: Mapping Potential Application Areas on Business Model Canvas

Application Area	Key Partners	Key Activities	Key Resources	Value Proposition	Customer Relationship	Channels	Customer Segments	Cost Structure	Revenue Streams
Smart Parking	Drivers, Parking Contractors, Government Agencies	Monitoring Parking Spaces, Recommending Parking Lots, Managing Parking Tickets	Mobile Applications	Minimize Time to Park, Find Best Place to Park	Personal Assistance, Automated Car Parking	Mobile Devices, Internet, Call Centers	Drivers	Channels Cost, Operations Cost	Parking Fees
Smart Healthcare	Patients, Physicians, Healthcare Centers, Insurers, Government Agencies	Provision of Cost-effective Healthcare Facilities	Application Services, Personalized Healthcare Services, Insurance Services	Improved and Efficient Healthcare Services, Better Insurance Plans	Personal Assistance, Automated Healthcare Services, Customized Insurance Plans	Mobile Devices, Wearable Systems, Patient Support Systems	Patients	Channels Cost, Operations Cost	Multiple Healthcare Plans
Telecom	Customers, Telecom Service Providers	Provision of Quality Telecom Services	Call Services, Internet Services	Improved Call Quality, Better Connectivity, Fast and high bandwidth communication channels	Personal Plans, Package plans, Pay-as-you-use plans	Mobile Devices, Call Centers, Customer Support Services	Mobile Users	Channels Cost, Operations Cost	Multiple Call, SMS, and Internet Usage Packages
Smart Retailers	Retail Shoppers, Retailers	Facilitation in Better Shopping Experience	Retail Services	Improved Shopping Experience, Fast Operations	Sales, Discount Offers	Mobile Devices, Internet, Shops	Retail Customers	Channels Cost, Operations Cost	General Retailers

Supply Chain Management	Drivers, Supply Chain Management Staff, Company Executives, Stores	Facilitation in Optimal Supply Chain Management	Business Operations, Supply Chain Management	Fast and Cost-effective operations	Personal Assistance, Automated System Operations	Mobile Devices, Internet, Enterprise Application Servers	Stores, Drivers, Supply Chain Management Staff	Channels Cost, Operations Cost	Stores and Internal Cost Minimization
Smart Agriculture	Farmers	Facilitation in Monitoring Climate and Agricultural Crops	Field Monitoring	Continuous Field Monitoring	Personal Assistance	Mobile Devices, Call Centers	Farmers	Channels Cost, Operations Cost	Farmers
Farm-to-Market Smart Solutions	Farmers, Market Vendors	Facilitation in Farm-to-Market Delivery of Goods	Business Operations	Fast and Effective Operations	Personal Assistance, Automated System Operations	Mobile Devices, Internet Servers	Farmers, Traders, Purchasers	Channels Cost, Operations Cost	Transportation, Trade, Commissions
E-commerce	Online Purchasers, Online Retailers,	Facilitation in Personalized Shopping	Business Operations	Fast, Effective, and Highly Personalized Operations	Personal Assistance, Personalized Recommendations	Mobile Devices	Online Shoppers	Channels Cost, Operations Cost	Sales, Discounts, Product Offers
Smart Transportation	Commuters, Transporters, Transportation Authorities, Drivers	Facilitation in Better Commute Services	Business Operations	Fast, Efficient, and Useful Transportation Services	Personal Assistance, Automated Operations	Mobile Devices, Internet Servers	Commuters	Channels Cost, Operations Cost	Ticketing, Passes, Discounts and Offers
Smart Water Management	Domestic Users, Water Supply, and Management Authorities	Better Water Quality and Timely Water Services	Business Operations	Timely and Better Quality Water Provision	Personal Assistance, Automated Services	Mobile Devices, Internet Servers	Domestic Users	Channels Cost, Operations Cost	Pricing Plans, Discounts Products, Offers

data reduction must be able to reduce data at the customer end. Meanwhile, the data collected at the enterprise end must be reduced properly by adopting the appropriate data reduction methods without compromising on V2C and V2F.

Detecting and Handling Information Shifts through Learning Models: Customers' behaviors and the information produced by enterprises constantly change with the passage of time. Therefore, the learning models should be designed to be adaptive to detect information shifts in the recently collected data. Accordingly, the enterprise application should be able to update all of its learning models.

Incentive Mechanisms: Enterprise customers are reluctant to share knowledge patterns because of privacy and security concerns. Therefore, enterprises should provide valuable incentive mechanisms to attract customers for participatory data sharing. However, the incentive mechanism should be designed in addition to the V2C objectives. In addition, enterprises should design transparent privacy preservation models whereby the control of knowledge sharing is in the hands of the customers instead of enterprises. Moreover, the security models should be robust enough that customers' personal information could be inaccessible.

Balancing cost structures with V2C and V2F: The deployment of the proposed data reduction framework brings the technical and financial overhead for enterprises. However, for the sustainable growth of enterprises, effort should be exerted to maintain this overhead lower compared to the perceived benefits gained from the V2C and V2F strategies.

5. Conclusion

This study presents the concept of big data reduction for value creation to achieve the V2C and V2F objectives. The current study discusses the adoption of big data analytics as a value creation tool for enterprises in determining hidden knowledge patterns in operational data and data collected from customers. This research proposes an early big data reduction framework at the customer end. The proposed framework enables enterprises to reduce the cost of cloud service utilization to perform big data analytics. In addition, this framework enables local knowledge availability, privacy preservation, and secure data sharing functions to build trust between customers and enterprises. In addition, the business model blueprint for early data reduction is presented and the key components of a few application areas are mapped on the business canvas model. Finally, a few challenges relevant to technology adoption are discussed in this study. In the future, we will develop a software component-based architecture for the proposed framework and will test it for real-world applications to assess the performance of the proposed framework and quantify the achieved levels of V2C and V2F.

Acknowledgment: The work presented in this paper is supported by University of Malaya Research Grant No. VOTE RP028C-14AET. In addition the authors would like to acknowledge Bright Spark Unit of University of Malaya for providing incentive support under grant no. BSP/APP/1634/2013.

References

Ackermann, K., & Angus, S. D. (2014). A resource efficient big data analysis method for the social sciences: the case of global IP activity. *Procedia Computer Science, 29*, 2360-2369.

Aggarwal, C. C. (2015). *Outlier analysis.* Paper presented at the Data Mining.

Akhbar, F., Chang, V., Yao, Y., & Muñoz, V. M. (2016). Outlook on moving of computing services towards the data sources. *International Journal of Information Management, 36*(4), 645-652.

Amini, A., Wah, T. Y., & Saboohi, H. (2014). On density-based data streams clustering algorithms: A survey. *Journal of Computer Science and Technology, 29*(1), 116-141.

Antonic, A., Roankovic, K., Marjanovic, M., Pripuic, K., & Zarko, I. P. (2014). *A mobile crowdsensing ecosystem enabled by a cloud-based publish/subscribe middleware.* Paper presented at the Future Internet of Things and Cloud (FiCloud), 2014 International Conference on.

Barros, R. C., Basgalupp, M. P., De Carvalho, A. C., & Freitas, A. (2012). A survey of evolutionary algorithms for decision-tree induction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 42*(3), 291-312.

Brown, C. T. (2012). BIGDATA: Small: DA: DCM: Low-memory Streaming Prefilters for Biological Sequencing Data. *DCM, July de*.

Chang, V. (2014). The business intelligence as a service in the cloud. *Future Generation Computer Systems, 37*, 512-534.

Chang, V., Walters, R. J., & Wills, G. (2013). The development that leads to the Cloud Computing Business Framework. *International Journal of Information Management, 33*(3), 524-538.

Chang, V., & Wills, G. (2016). A model to compare cloud and non-cloud storage of Big Data. *Future Generation Computer Systems, 57*, 56-76.

Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. *Environmental Modelling & Software, 37*, 134-145.

Chen, X.-W., & Lin, X. (2014). Big data deep learning: Challenges and perspectives. *Access, IEEE, 2*, 514-525.

Cheng, Y., Jiang, P., & Peng, Y. (2014). *Increasing big data front end processing efficiency via locality sensitive Bloom filter for elderly healthcare.* Paper presented at the Computational Intelligence in Big Data (CIBD), 2014 IEEE Symposium on.

Chou, D. C. (2015). Cloud computing: A value creation model. *Computer Standards & Interfaces, 38*, 72-77.

Chuang, H.-M., & Chen, Y.-S. (2015). Identifying the value co-creation behavior of virtual customer environments using a hybrid expert-based DANP model in the bicycle industry. *Human-centric Computing and Information Sciences, 5*(1), 1-31.

Cossío-Silva, F.-J., Revilla-Camacho, M.-Á., Vega-Vázquez, M., & Palacios-Florencio, B. (2015). Value co-creation and customer loyalty. *Journal of Business Research*.

CrowdFlower. (2015). CrowdFlower 2015 Data Scientist Report. Online.

Di Martino, B., Aversa, R., Cretella, G., Esposito, A., & Kołodziej, J. (2014). Big data (lost) in the cloud. *International Journal of Big Data Intelligence, 1*(1-2), 3-17.

Dong, W., Douglis, F., Li, K., Patterson, R. H., Reddy, S., & Shilane, P. (2011). *Tradeoffs in Scalable Data Routing for Deduplication Clusters.* Paper presented at the FAST.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis*: John Wiley & Sons.

Feldman, D., Schmidt, M., & Sohler, C. (2013). *Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering.* Paper presented at the Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms.

Fu, Y., Jiang, H., & Xiao, N. (2012). A scalable inline cluster deduplication framework for big data protection *Middleware 2012* (pp. 354-373): Springer.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. *Knowledge and Information Systems, 46*(2), 241-284.

Grzegorowski, M., & Stawicki, S. (2015). *Window-based feature extraction framework for multi-sensor data: a posture recognition case study.* Paper presented at the Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on.

Haile, N., & Altmann, J. (2016). Value creation in software service platforms. *Future Generation Computer Systems, 55*, 495-509.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98-115.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., & Poldrack, R. (2013). *BIG & QUIC: Sparse inverse covariance estimation for a million variables.* Paper presented at the Advances in Neural Information Processing Systems.

Insights, C. (2016). The Top 20 Reasons Startups Fail. Online.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Jalali, B., & Asghari, M. H. (2014). The anamorphic stretch transform: Putting the squeeze on "big data". *Optics and Photonics News, 25*(2), 24-31.

Jiang, P., Winkley, J., Zhao, C., Munnoch, R., Min, G., & Yang, L. T. (2014). An intelligent information forwarder for healthcare big data systems with distributed wearable sensors.

Konstan, J., & Riedl, J. (2012). Deconstructing Recommender Systems: How Amazon and Netflix predict your preferences and prod you to purchase. *IEEE Spectrum, 49*.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700-710.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT sloan management review, 21*.

Leung, C. K.-S., MacKinnon, R. K., & Jiang, F. (2014). *Reducing the search space for big data mining for interesting patterns from uncertain data.* Paper presented at the Big Data (BigData Congress), 2014 IEEE International Congress on.

Li, C.-S., Darema, F., & Chang, V. (2016). Distributed behavior model orchestration in cognitive internet of things solution. *International Journal of Information Management*.

Lin, M.-S., Chiu, C.-Y., Lee, Y.-J., & Pao, H.-K. (2013). *Malicious URL filtering—A big data application.* Paper presented at the Big Data, 2013 IEEE International Conference on.

Liu, G., Iwai, M., Tobe, Y., & Sezaki, K. (2013). *REPSense: On-line sensor data reduction while preserving data diversity for mobile sensing.* Paper presented at the 2013 IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob).

Luan, T. H., Gao, L., Li, Z., Xiang, Y., & Sun, L. (2015). Fog computing: Focusing on mobile users at the edge. *arXiv preprint arXiv:1502.01815*.

Mital, M., Chang, V., Choudhary, P., Pani, A., & Sun, Z. (2016). Adoption of cloud based Internet of Things in India: A multiple theory perspective. *International Journal of Information Management*.

Moshtaghi, M., Bezdek, J. C., Leckie, C., Karunasekera, S., & Palaniswami, M. (2015). Evolving fuzzy rules for anomaly detection in data streams. *Fuzzy Systems, IEEE Transactions on, 23*(3), 688-700.

Osterwalder, A., Pigneur, Y., & Tucci, C. L. (2005). Clarifying business models: Origins, present, and future of the concept. *Communications of the association for Information Systems, 16*(1), 1.

Pang, Z., Chen, Q., Han, W., & Zheng, L. (2015). Value-centric design of the internet-of-things solution for food supply chain: value creation, sensor portfolio and information fusion. *Information Systems Frontiers, 17*(2), 289-319.

Papageorgiou, A., Schmidt, M., Song, J., & Kami, N. (2013). *Smart m2m data filtering using domain-specific thresholds in domain-agnostic platforms.* Paper presented at the Big Data (BigData Congress), 2013 IEEE International Congress on.

Patel, N. (2015). 90% Of Startups Fail: Here's What You Need To Know About The 10%. Retrieved 11-05-2016, 2016, from http://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/#62c11ac955e1

Patty, J. W., & Penn, E. M. (2015). Analyzing big data: social choice and measurement. *PS: Political Science & Politics, 48*(01), 95-101.

Qi, J.-Y., Qu, Q.-X., & Zhou, Y.-P. (2014). How does customer self-construal moderate CRM value creation chain? *Electronic Commerce Research and Applications, 13*(5), 295-304.

Rágyanszki, A., Gerlei, K. Z., Surányi, A., Kelemen, A., Jensen, S. J. K., Csizmadia, I. G., & Viskolcz, B. (2015). Big data reduction by fitting mathematical functions: A search for appropriate functions to fit Ramachandran surfaces. *Chemical Physics Letters, 625*, 91-97.

Rehman, M., Khan, A., & Batool, A. (2016). Big Data Analytics in Mobile and Cloud Computing Environments. In Q. Hussain (Ed.), *Handbook of Research on Next-Generation High Performance Computing* (Vol. 1): IGI Global.

Rehman, M. H., & Batool, A. (2015). The Concept of Pattern based Data Sharing in Big Data Environments. *International Journal of Database Theory and Application, 8*(4), 11-18.

Rehman, M. H., Liew, C. S., & Wah, T. Y. (2014). *Frequent pattern mining in mobile devices: A feasibility study*. Paper presented at the Information Technology and Multimedia (ICIMU), 2014 International Conference on.

Rojas, R. (2013). *Neural networks: a systematic introduction*: Springer Science & Business Media.

Salmon, J., Harmany, Z., Deledalle, C.-A., & Willett, R. (2014). Poisson noise reduction with non-local PCA. *Journal of mathematical imaging and vision, 48*(2), 279-294.

Sharma, S., Chang, V., Tim, U. S., Wong, J., & Gadia, S. (2016). Cloud-based emerging services systems. *International Journal of Information Management*.

Singh, N., Javeed, A., Chhabra, S., & Kumar, P. (2015). Missing Value Imputation with Unsupervised Kohonen Self Organizing Map *Emerging Research in Computing, Information, Communication and Applications* (pp. 61-76): Springer.

Stateczny, A., & Wlodarczyk-Sielicka, M. (2014). Self-organizing artificial neural networks into hydrographic big data reduction process *Rough Sets and Intelligent Systems Paradigms* (pp. 335-342): Springer.

Trovati, M. (2015). Reduced topologically real-world networks: a big-data approach. *International Journal of Distributed Systems and Technologies (IJDST), 6*(2), 13-27.

Trovati, M., & Bessis, N. (2015). An influence assessment method based on co-occurrence for topologically reduced big data sets. *Soft Computing*, 1-10.

Vera-Baquero, A., Colomo-Palacios, R., & Molloy, O. (2013). Business process analytics using a big data approach. *IT Professional, 15*(6), 29-35.

Verhoef, P. C., Kooge, E., & Walk, N. (2016). *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*: Routledge.

Vervliet, N., Debals, O., Sorber, L., & De Lathauwer, L. (2014). Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in big data analysis. *Signal Processing Magazine, IEEE, 31*(5), 71-79.

Vidgen, R. (2014). Creating business value from Big Data and business analytics: organizational, managerial and human resource implications.

Weinstein, M., Meirer, F., Hume, A., Sciau, P., Shaked, G., Hofstetter, R., . . . Horn, D. (2013). Analyzing big data with dynamic quantum clustering. *arXiv preprint arXiv:1310.2700*.

Wu, C., Buyya, R., & Ramamohanarao, K. (2016). Big Data Analytics= Machine Learning+ Cloud Computing. *arXiv preprint arXiv:1601.03115*.

Xia, W., Jiang, H., Feng, D., & Hua, Y. (2011). *SiLo: A Similarity-Locality based Near-Exact Deduplication Scheme with Low RAM Overhead and High Throughput.* Paper presented at the USENIX Annual Technical Conference.

Xu, T., Zhang, Z., Yu, P. S., & Long, B. (2012). Generative models for evolutionary clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD), 6*(2), 7.

Yang, C., Zhang, X., Zhong, C., Liu, C., Pei, J., Ramamohanarao, K., & Chen, J. (2014). A spatiotemporal compression based approach for efficient big data processing on cloud. *Journal of Computer and System Sciences, 80*(8), 1563-1583.

Yaqoob, I., Chang, V., Gani, A., Mokhtar, S., Hashem, I. A. T., Ahmed, E., . . . Khan, S. U. (2016). Information fusion in social big data: Foundations, state-of-the-art, applications, challenges, and future research directions. *International Journal of Information Management*.

Zhai, Y., Ong, Y.-S., & Tsang, I. W. (2014). The Emerging" Big Dimensionality". *Computational Intelligence Magazine, IEEE, 9*(3), 14-26.

Zhou, R., Liu, M., & Li, T. (2013). *Characterizing the efficiency of data deduplication for big data storage management.* Paper presented at the Workload Characterization (IISWC), 2013 IEEE International Symposium on.

Zolnowski, A., Weiß, C., & Bohmann, T. (2014). *Representing Service Business Models with the Service Business Model Canvas--The Case of a Mobile Payment Service in the Retail Industry*. Paper presented at the System Sciences (HICSS), 2014 47th Hawaii International Conference on.