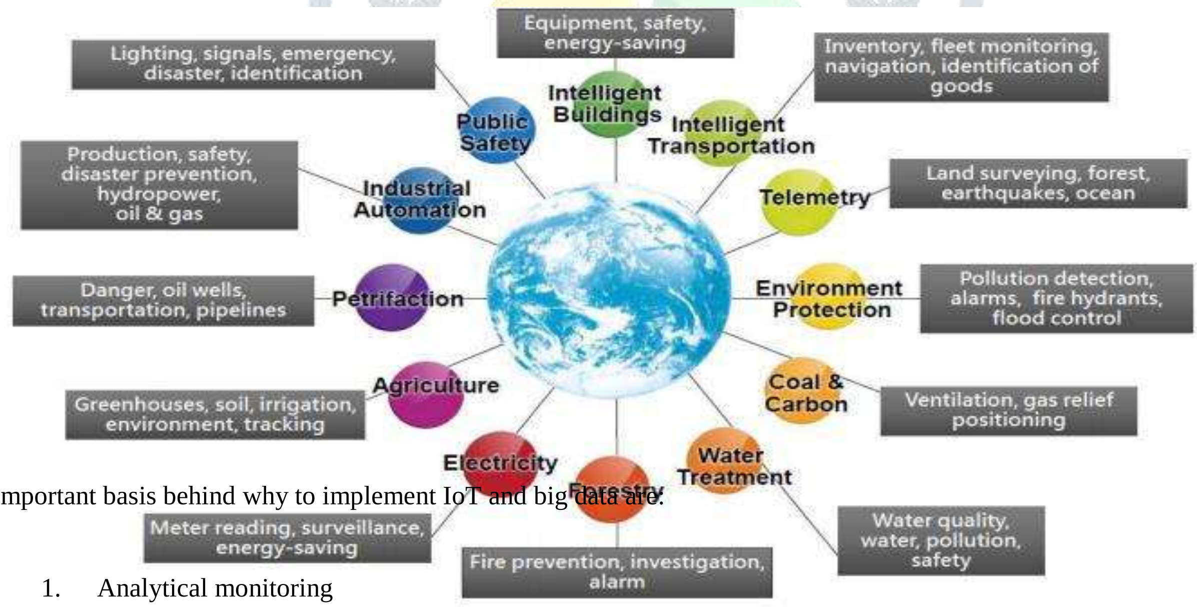
ENABLING IOT AND BIG DATA ANALYSIS

[1] CH.SRILAKSHMI [2] A.SHOBHARANI
[1] ASSISTANT PROFESSOR [2] ASSISTANT PROFESSOR
[1, 2] NARSIMHA REDDY ENGINEERING COLLEGE
Maisammaguda, Dhulapally, Kompally,Secunderabad -500 100.

Abstract: The Internet of Things (IoT), firstly coined by Kevin Ashton as the title of a presentation in 1999 , is a technological revolution that is bringing us into a new ubiquitous connectivity, computing, and communication era. The development of IoT depends on dynamic technical innovations in a number of fields, from wireless sensors to nanotechnology. For these ground-breaking innovations to grow from ideas to specifice products or applications, in the past decade, we have witnessed worldwide efforts from academic community, service providers, network operators, and standard development organizations, etc (see, e.g., the recent comprehensive surveys .Generally, current research on IoT mainly focuses on how to enable general objects to see, hear, and smell the physical world for themselves, and make them connected to share the observations. In this paper, we argue that only connected is not enough, beyond that, general objects in future IoT should have the capability to learn, think, and understand the physical world by themselves.

1. NECESSITY OF IOT AND BIG DATA IMPLEMENTATION

IoT will enable big data, big data needs analytics, and analytics will improve processes for more IoT devices. IoT and big data can be used to improve various functions and operations in diverse sectors. Both have extended their capabilities to wide range of areas. The figure below shows the areas of big data produced. Some or the other way, data is produced through connected devices.


Equipment, safety, energy-saving
Lighting, signals, emergency, disaster, identification
Inventory, fleet monitoring, navigation, identification of goods
Intelligent Buildings
Public Safety
Intelligent Transportation
Production, safety, disaster prevention, hydropower, oil & gas
Industrial Automation
Telemetry
Land surveying, forest, earthquakes, ocean
Danger, oil wells, transportation, pipelines
Petrifaction
Environment Protection
Pollution detection, alarms, fire hydrants, flood control
Agriculture
Greenhouses, soil, irrigation, environment, tracking
Coal & Carbon
Ventilation, gas relief positioning
Electricity
Forestry
Water Treatment
Meter reading, surveillance, energy-saving
Fire prevention, investigation, alarm
Water quality, water, pollution, safety

The important basis behind why to implement IoT and big data are:

1. Analytical monitoring
2. More Uptime
3. Lower reject rates
4. Higher throughput
5. Enhanced safety
6. Efficient use of labor

7. Enable mass customization
8. Analyze the activities for real-time marketing
9. Improved situational alertness
10. Improved quality
11. Sensor-driven decision analytics

2. IMPACTS OF IOT ON BIG DATA

The massive amount of revenue and data that the IoT will generate, its impact will be felt across the entire big data universe, forcing companies to upgrade current tools and processes, and technology to evolve to accommodate this additional data volume and take advantage of the insights all this new data undoubtedly will deliver.

Data Storage:

- When we talk about IoT, one of the first things that comes to mind is a huge, continuous stream of data hitting the data storage. In response to this direct impact on big data storage infrastructure, many organizations are moving toward the Platform as a Service model instead of keeping their own storage infrastructure, which would require continuous expansion to handle the load of big data. PaaS is a cloud-based, managed solution that provides scalability, flexibility, compliance, and a sophisticated architecture to store valuable IoT data. Cloud storage options include private, public, and hybrid models. If companies have sensitive data or data that is subject to regulatory compliance requirements that require heightened security, a private cloud model might be the best fit. Otherwise, a public or hybrid model can be chosen as storage for IoT data.

Open source:

- The IoT isn't built solely using open source software, but open source plays a key role. Linux serves as the operating system for many connected devices. Open source networking standards make it possible for devices from different vendors to communicate. Some IoT devices are even designed to be hack able by users in a way that extends the open source software concept to include open hardware. In all of these ways, the IoT plays on open source's strengths and brings open source to new frontiers.

Big Data:

- The IoT promises to take big data to a new level. IoT devices not only generate huge amounts of information, which can then be fed to data analytics tools. They also rely on data-based logic in order to perform many of their "smart" functions. Take your Nest thermostat, for example. It collects data from your home, then runs analytics based on the data it collected along with external information (like weather reports) to predict when to turn on your furnace.

Cybersecurity:

- Security and privacy aren't new concerns. But they are on the minds of consumers now more than ever, thanks to the seemingly never-ending reports of breaches at major organizations. The IoT serves both to feed and to alleviate those concerns. IoT devices raise huge new security challenges, especially when it comes to things like critical infrastructure. But they also offer ways to help keep users more secure by adding extra barriers of defense to data and persons.

3. CHALLENGES IMPLEMENTING IOT ON BIG DATA

IoT is not free from challenges. Issues of Governance, security, Interoperability, privacy, regulations, providing power to billions of sensors and standardization issues can slow down the progress of Internet of Things. Due to the absence of a generic governance, there are many confusions and inconsistencies. The absence of a universal numbering system is a bane for providing a true IoT environment. In the current context, systems like EPC Global and ubiquitous ID systems are used to address the issue of global ID systems. There is a challenge of implementing common security protocols. So, interoperability is an issue while interacting with IoT objects developed by different manufacturers.

Major challenges that can fetch momentous rewards when they are solved.

1. Huge data volumes
2. Difficulty in data collection
3. Incompatible standards
4. New security threats
5. No reliability in the data
6. Fundamental shifts in business models
7. Huge amount of data to analyze
8. A rapidly evolving privacy landscape

4. HETEROGENEOUS DATA PROCESSING

In practical IoT applications, the massive data are generally collected from heterogeneous sensors (e.g., cameras, vehicles, drivers, and passengers), which in turn may provide heterogeneous sensing data (e.g., text, video, and voice). Heterogeneous data processing (e.g., fusion, classification) brings unique challenges and also offers several advantages and new possibilities for system improvement.

Mathematically, random variables that characterize the data from heterogeneous sensors may follow disparate probability distributions. Denote z_n as the data from the n-th sensor and

$Z := fz_ng^N_{n=1}$ as the heterogeneous data set, the marginal's $fz_ng^N_{n=1}$ are generally non-identically or heterogeneously distributed. In many IoT applications, problems are often modeled as multi-sensor data fusion, distribution estimation or distributed detection. In these cases, joint probability density function (pdf) f(Z) of the heterogeneous data set Z is needed to obtain from the

marginal pdfs $ff(z_n)g^N_{n=1}$.

For mathematical tractability, one often chooses to assume simple models such as the product model or multivariate Gaussian model, which lead to suboptimal solutions . Here we recommend another approach, based on copula theory, to tackle heterogeneous data processing in IoT. In copula theory, it is the copulas function that couples multivariate joint distributions to their marginal distribution functions, mainly thanks to the following theorem: Sklar' Theorem : Let F be an N-dimensional cumulative distribution function (cdf) with continuous marginal cdfs

$F_1; F_2; :::; F_N$.

Then there exists a unique copulas function C such that for all $z_1; z_2; :::; z_N$ in

$$F\ (z_1; z_2; :::; z_N\ ) = C\ F_1(z_1); F_2(z_2); :::; F_N\ (z_N\ ) :$$

The joint pdf can now be obtained by taking the N-order derivative of

$$f(z_1; z_2; :::; z_N\ ) = \frac{@^N}{@z_1\ @z_2\ :::@z_N} C\ F_1(z_1); F_2(z_2); :::; F_N\ (z_N\ )$$

$$= f_p(z_1; z_2; :::; z_N\ )c\ F_1(z_1); F_2(z_2); :::; F_N\ (z_N\ ) ;$$

where $f_p(z_1; z_2; :::; z_N\ )$ denotes the product of the marginal pdfs $ff(z_n)g^N_{n=1}$ and c( ) is the copula density weights the product distribution appropriately to incorporate dependence between the random variables. The topic on the design or selection of proper copula functions is well summarized.

5. NONLINEAR DATA PROCESSING

In IoT applications such as multi-sensor data fusion, the optimal fusion rule can be derived from the multivariate joint distributions obtained in. However, it is generally mathematically intractable since the optimal rule generally involves nonlinear operations. Therefore, linear data processing methods dominate the research and development, mainly for their simplicity. However, linear methods are often oversimplified to deviate the optimality.

In many practical applications, nonlinear data processing significantly outperforms their linear counterparts. Kernel-based learning (KBL) provides an elegant mathematical means to construct powerful nonlinear variants of most well-known statistical linear techniques, which has recently become prevalent in many engineering application. Briefly, in KBL theory, data x in the input space X is projected onto a higher dimensional feature space F via a nonlinear mapping as follows:

$$: X\ !\ F; \quad x\ 7!\ (x):$$

For a given problem, one now works with the mapped data (x) 2 F instead of x 2 X . The data in the input space can be projected onto different feature spaces with different mappings. The diversity of feature spaces provides us more choices to gain better performance. Actually, without knowing the mapping explicitly, one only needs to replace the inner product operator of a linear technique with an appropriate kernel k (i.e., a positive semi-definite symmetric function),

$$k(x_i; x_j) := h\ (x_i); \ (x_j)i_F\ ; \ 8x_i; x_j\ 2\ X :$$

The most widely used kernels can be divided into two categories: projective kernels (functions of inner product, e.g., polynomial kernels) and radial kernels (functions of distance, e.g., Gaussian
kernal

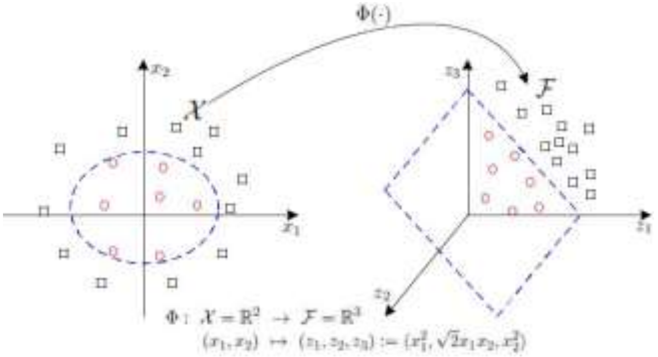

$\Phi(\cdot)$
x_2
$\mathcal{X}$
x_1
z_3
$\mathcal{F}$
z_1
z_2
$\Phi : \mathcal{X} = \mathbb{R}^2 \to \mathcal{F} = \mathbb{R}^3$
$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

Fig. 3. An introductory binary classification example . By mapping data x = (x_1; x_2) in 2-D input space X = R^2 (left) via nonlinear mapping ( ) onto a 3-D feature space F = R^3 (right), the data become linearly separable.

6. HIGH-DIMENSIONAL DATA PROCESSING

In IoT, massive or big data always accompanies high-dimensionality. For example, images and videos observed by cameras in many IoT applications are generally very high-dimensional data, where the dimensionality of each observation is comparable to or even larger than the number of observations. Moreover, in kernel-based learning methods discussed above, the kernel function nonlinearly maps the data in the original space into a higher dimensional feature space, which transforms virtually every dataset to a high-dimensional one. Mathematically, we can represent the massive data in a compact matrix form. Many practical applications have experimentally demonstrated the intrinsic low-rank property of the high-dimensional data matrix, such as the traffic matrix in large scale networks and image frame matrix in video surveillance , which is mainly due to common temporal patterns across columns or rows, and periodic behavior across time, etc.

Low-rank matrix plays a central role in large-scale data analysis and dimensionality reduction. In the following, we provide a brief tutorial on using low-rank matrix recovery and/or completion[1] algorithms for high-dimensional data processing, from simple to complex.

1)Low-rank matrix recovery with dense noise and sparse anomalies: Suppose we are given a large sensing data matrix Y, and know that it may be decomposed as

$$Y = X + V;$$

where X has low-rank, and V is a perturbation/noise matrix with entry-wise non-zeros. We do not know the low-dimensional column and row space of X, not even their dimensions. To stably recover the matrix X from the sensing data matrix Y, the problem of interest can be formulated as classical principal component analysis (PCA):

X jjXjj subject to jjY Xjj$_F$ ";
min
f g

where " is a noise related parameter, jj jj and jj jj$_F$ stands for the nuclear norm (i.e., the sum of the singular values) and the Frobenious norm of a matrix. Furthermore, if there are also some abnormal data A injected into the sensing data matrix Y, we have

$$Y = X + V + A;$$

where A has sparse non-zero entries, which can be of arbitrary magnitude. In this case, we do not know the low-dimensional column and row space of X, not know the locations of the nonzero entries of A, and not even know how many there are. To accurately and efficiently recover the low-rank data matrix X and sparse component A, the problem of interest can be formulated as the following tractable convex optimization: where is a positive rank sparsely controlling parameter, and jj jj$_1$ stands for the l$_1$-norm (i.e., the number of nonzero entries) of a matrix.

2) Joint matrix completion and matrix recovery: In practical IoT applications, it is typically difficult to acquire all entries of the sensing data matrix Y, mainly due to i) transmission loss of the sensing data from the sensors to the data center, and ii) lack of incentives for the crowd sources to contribute all their sensing data.

In this case, the sensing data matrix Ye is made up of noisy, corrupted, and incomplete observations,

$$Y := P (Y) = P (X + A + V);$$

where [M] [N] is the set of indices of the acquired entries, and P is the orthogonal projection onto the

e

linear subspace of matrices supported on , i.e., if (m; n) 2 , P (Y) = y$_{m;n}$; otherwise, P (Y) = 0. To stably recover the low-rank and sparse components X and A, the problem can be further formulated as

X;A jjX jj jj Ajj$_1$ subject to jjP (Y) P (X + A + V) jj$_F$ ":

min +

f g

All the problems formulated in _and _fall into the scope of convex optimization, efficient algorithms can be developed based on the results

7. PARALLEL AND DISTRIBUTED DATA PROCESSING

So far, all the data processing methods introduced above are in essence centralized and suitable to be implemented at a data center. However, in many practical IoT applications, where the objects in the networks are organized in an ad hoc or decentralized manner, centralized data processing will be inefficient or even impossible because of single-node failure, limited scalability, and huge exchange overhead, etc. Now, one natural question comes into being: Is there any way to disassemble massive data into groups of small data, and transfer centralized data processing into decentralized processing among locally interconnected agents, at the price of affordable performance loss?

In this subsection, we argue that alternating direction method of multipliers (ADMM) serves as a promising theoretical framework to accomplish parallel and distributed data processing. Suppose a very simple case with a IoT consisting of N interconnected smart objects. They have a common objective as follows

NXmin f(x) = f$_i$(x);

xi=1

where x is an unknown global variable and f$_i$ refers to the term with respect to the i-th smart object. By introducing local variables fx$_i$ 2 R^{n}g$^N{}_{i=1}$ and a common global variable z, the problem in _can be rewritten as

This is called the global consensus problem, since the constraint is that all the local variables should agree, i.e., be equal. The augmented Lagrangian of problem _can be further written as

The resulting ADMM algorithm directly from _is the following:

$x^{k}{}_{i}{}^{+1} := \mathrm{argmin}_{xi}\ f_i(x_i) + y_i^{kT}\ (x_i z^k) + {}_{2}\ jjx_i \qquad \_$

$z^k jj^2{}_{F--}N_z k+1\ {}_{:=\ N}\underline{{}^{1\ X}}$

${}_{x}k_i+1\ {}_{+\ 1=\ yi}ki=1$

$y_i^{k+1} := y_i^{k} + \ (x_i^{k+1}\ \ z^{k+1}):$

The first and last steps are carried out independently at each smart object, while the second step is performed at a fusion center.

8. CONCLUSION

In this paper, we have provided a high-level tutorial on massive data analytics in terms of heterogeneous, nonlinear, high-dimensional, and distributed and parallel data processing, respectively. Actually, in practical IoT applications, the obtained massive sensing data can be of mixed characteristics, which is much more challenging. Moreover, the development of practical and effective algorithms for specific IoT applications are also urgently needed. Since there is a major impact of IoT on big data we need to quickly improvise the complete structure to manage the daily changing circumstances. There are a few areas of concern and security and privacy and data collection efficiency are probably the most difficult problems we are facing. Security compromise and inefficiencies in data collection mechanisms result in a loss of status, money, time and effort. But there is hope because both the IoT and the big data are at an emerging stage and there will be upgrade.

9. REFERENCES

1)http://www.sas.com/en_us/insights/big-data/internet-of-things.html

2) http://www.kdnuggets.com/2015/07/impact-iot-big-data-landscape.html

3) http://data-informed.com/the-impact-of-internet-of-things-on-big-data/

4) http://rdi2.rutgers.edu/sites/rdi2/files/img/Greer_Rutgers_BigData_Apr_2014.pdf

5) K. Ashton, “That ‘internet of things’ thing in the real world, things matter more than ideas,” RFID Journal, June 2009, http://www.rfidjournal. com/article/print/4986 [Accessed on: 2013-10-25].

6)International Telecommunication Union (ITU), “ITU internet reports 2005: The internet of things,” ITU, Executive Summary, Nov. 2005, http://www.itu.int/pub/S-POL-IR.IT-2005 [Accessed on: 2013-10-25].

7) L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” Computer Networks, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

8) C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Context aware computing for the internet of things: A survey,” IEEE

Communications Surveys & Tutorials, Accepted for Publication.

9) M. R. Palattella, N. Accettura, X. Vilajosana, et al, “Standardized protocol stack for the internet of (important) things,” IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1389-1406, Third Quarter 2013.