Federated Analytics: Opportunities and Challenges

Dan Wang, *Senior Member, IEEE,* Siping Shi, Yifei Zhu, *Member, IEEE,* and Zhu Han, *Fellow, IEEE*

***Abstract*—In this paper, we present federated analytics, a new distributed computing paradigm for data analytics applications with privacy concerns. With the advances of sensing, communication, and edge computing technologies, data are massively generated, transmitted and analyzed in an edge-cloud computing environment. In many applications, the edge devices and the data generated in the edge belong to heterogeneous owners. The data privacy and confidentiality have become increasing concerns to these owners. The current edge-cloud computing paradigm for data analytics, where data are sent to a central server for analytics, can no longer match the application requirements. Federated analytics is a newly proposed computing paradigm where raw data are kept local with local analytics and only the insights generated from local analytics are sent to a server for result aggregation. Federated analytics differs from the recent federated learning paradigm in the sense that federated learning emphasizes on collaborative model training, whereas federated analytics emphasizes on drawing conclusions from data. In this paper, we first clarify what federated analytics is and its position in the research literature. We then present why we need federated analytics, i.e., the motivation and application case studies. Finally, we discuss the opportunities and challenges of federated analytics.**

I. Introduction

Data have become important assets. In 2017, Economist claimed that "The world's most valuable resource is no longer oil, but data." [1] Data analytics is the computing task that draws conclusions from data. To increase the dimension, quantity and coverage of data, data analytics are commonly performed on the data contributed from multiple parties nowadays. In another dimension, there is also increasing awareness on privacy issues. Privacy regulations have been established and strengthened. In 2018, the General Data Protection Regulation (GDPR) were introduced. As such, the conventional collaborative data analytics paradigm where data are sent to a centralized server for analytics is no longer viable.

In this paper, we present federated analytics, a new distributed computing paradigm for data analytics applications with privacy concerns. In federated analytics, data are kept local. Local data analytics computing is performed and only the derived insights will be sent out to a coordination server for aggregation. Consequently, privacy can be protected.

The federated analytics paradigm differs from the recent federated learning paradigm [1] in the sense that federated learning is a distributed computing paradigm where the computing task is model training for a machine learning model, and in particular, a deep learning model. Federated analytics can be considered from a generalized perspective and a specialized perspective. From a generalized perspective, federated analytics include all computing tasks that draw conclusions from data, and thus include federated learning. From a specialized perspective, federated analytics specifically refers to the model inference phase of machine learning, and thus is the sequel of federated learning. For example, a federated video analytics architecture [2] assumes the machine learning model has been established and the participating peer devices collaboratively perform model inference.

The term federated analytics was first proposed by Google in May 2020 [3] for its Gboard application. Gboard is a keyboard app for smartphones providing AI features such as next word prediction, keyboard search suggestion, etc. With privacy concerns, Gboard trains deep learning models through federated learning. To improve the trained models, they should be evaluated periodically, which again needs to use the user data. Gboard leverages federated analytics to measure the overall quality of the trained deep learning models.

With the increase in collaborative data analytics applications and the increase in privacy concerns, it is foreseeable that there will be great demands on federated analytics. Unfortunately, there is a lack of clarification on federated analytics, its position in the research literature, possible case studies, and opportunities and challenges.

This paper fills in this gap. We organize this paper as follows. In Section II, we first discuss what is federated analytics. We present a definition of federated analytics and a taxonomy of federated analytics. There are related work summarizing the topics of distributed intelligence, edge computing, collaborative intelligence, federated learning, etc. We will clarify their differences from us in the taxonomy. In Section III, we discuss why we need federated analytics. We discuss in details the demands from applications and the technology readiness. These are the two forces to pull and push the development of federated analytics. We further present two case studies, Google Gboard and video analytics. In Section IV, we discuss the opportunities and challenges in federated analytics. We study the architecture designs, privacy management, resource management, analytics design and optimization, and business models in federated analytics. To the best of our knowledge, we are the first to provide a comprehensive overview on federated analytics, its opportunities and challenges.

Dan Wang and Siping Shi are with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong (email: {csdwang,cssshi}@comp.polyu.edu.hk).

Yifei Zhu is with the University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, China (email: yifei.zhu@sjtu.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, 446-701, South Korea. (email: zhan2@uh.edu).

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

II. What is Federated Analytics

In [3], Google called federated anaytlics as "Collaborative data science without data collection". More academically, *federated analytics* is a collaborative computing paradigm that performs data analytics computing tasks (i.e., to draw conclusions from data) across multiple decentalized devices where the raw data should be kept local.

Federated analytics can be seen as a branch of the distributed computing paradigm. There are many distributed computing paradigms, such as collaborating computing, distributed intelligence, coordinated intelligence, federated learning, collaborative analytics, to name but a few. We classify these from the perspectives of the collaboration model and the computing model. The collaboration model refers to the relationship of the distributed nodes. The computing model refers to the computing task undertaken by the system. We show a taxonomy in Fig. 1.

From the perspective of collaboration model, we further classify it into three sub-categories: the distributed, collaborative and federated models. A collaborative collaboration model is a special distributed collaboration model since a distributed collaboration model does not require active collaboration. For example, a divide and conquer computing task falls into a distributed collaboration model, not a collaborative collaboration model. The federated collaboration model is a special collaborative collaboration model since the federated systems emphasize multiple ownership among nodes. For example, the servers of a data center belonging to a single owner and collaboratively perform a computing task and thus belong to collaborative computing. In the collaboration model, there are sub-areas of collaborative computing, coordinative computing, collective computing, and cooperative computing. A summarize of their differences is in [4]. We do not further differentiate these models in this paper.

From the perspective of the computing model, we further classify it into three sub-categories: computing, intelligence, and learning/analytics tasks, and it focuses on learning some knowledge from the data or experimental observations. An intelligence computing task is a special form of the general computing task. A learning/analytics computing task is a special intelligence computing task as the latter emphasizes computing that involves data; which is not necessary for a general intelligence computing task. The analytics computing task can be referred from a general perspective or specialized perspective. A general analytics computing task refers to a data analytics task that draws conclusions from data, and it includes learning, inference, and non-machine learning analytics. As a result, learning is part of analytics. A specialized analytics computing task specifically refers to the model inference computing task.

An abstract federated analytics model is shown in Fig. 2. There are local devices and a coordination server. The execution process of federated analytics is usually divided into four main steps. First, a global model or analytics task is distributed to the local devices. Second, each local device performs local analytics using its own data. Third, the local analytics results are reported to the coordination server. Fourth, the server aggregate local results to generate a final analytical result. This abstract model shows the key elements of federated analytics: an untrusted environment with the necessity of local data analytics.

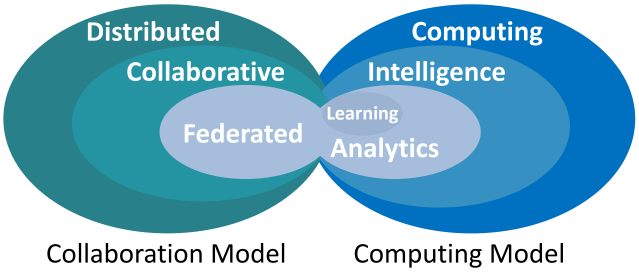


Fig. 1: A Taxonomy for federated analytics. Under this taxonomy, there can be distributed intelligence, collaborative computing, collaborative analytics, federated analytics, etc.

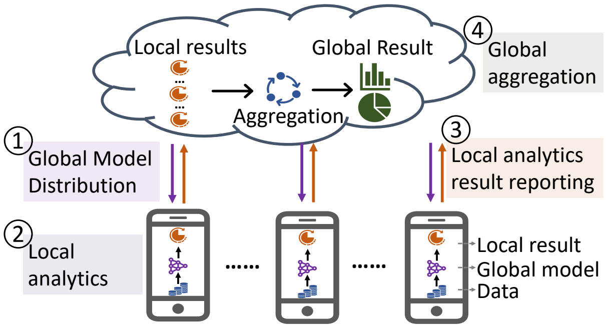


Fig. 2: An abstract federated analytics model.

III. Why Federated Analytics

Federated analytics is, on the one hand, primarily pulled by the demands of applications. On the other hand, the enabling technologies and the infrastructure are becoming mature and ready, pushing the emergence of federated analytics.

A. Application Demands

Increasing demands on collaborative data analytics: Data analytics have been proven effective in every aspect of our life today. Companies can turn raw data into relevant trends, predictions, projections with unprecedented accuracy and gain insights that drive intelligent decision-making. A Bain & Company survey of over 400 businesses found that organizations that make the most use of data tend to pull ahead of industry peers. Such companies are twice more likely to be in their industry's top quartile of financial performance and five times more likely to make decisions faster than market peers.

Collaborative data analytics is essential to improve data coverage, especially for companies that routinely rely on big data to inform business practices but have exhausted their internal data silos. American Airlines and Citi, for example, have been sharing their customer data for over thirty years to increase customer loyalty and boost credit card spending. During COVID-19, effective decisions about public health action are made due to data sharing of clinical trials, observational studies, operational research, routine surveillance, information on the virus and its genetic sequences, as well as the monitoring of disease control programmes.

Data-driven analytics methods have been proposed in industry. For example, a Bayesian inference method [5] has been proposed for the status of high-speed rails. Through periodic posterior updates with real-time data using onboard sensors, it is shown that fault diagnosis can be much improved as compared to currently fixed schedule fault inspections. Machine learning models [6] have also been developed for the status prediction of Heating, Ventilation, and Air Conditioning (HVAC) systems so that HVAC operation can be individualized to reduce the wastes and overprovisioning of a preset operation procedure.

Increasing Concerns on Privacy and Confidentiality: In another dimension, there is also increasing awareness on privacy issues along with the adoption of data analytics. In 2018, GDPR were introduced. The GDPR contains provisions and requirements related to the processing of personal data of individuals who are located in the European Economic Area (EEA), and applies to any enterprise—regardless of its location and the data subjects' citizenship or residence. Around the world, more than 60 jurisdictions have enacted or proposed postmodern privacy and data protection laws.

Within the first year of GDRP, its fines totalled €56 million, with more than 200,000 investigations. The tech giant Facebook was fined £500,000 by the UK's data protection watchdog for the Cambridge Analytica data scandal, in which the personal data of millions of Facebook users was predominantly used for political advertising without their consent. Facebook CEO Mark Zuckerberg testified in front of Congress on April 12th 2018 to respond to the scandal.

In addition to personal data privacy, business vendors also have confidentiality concerns. It has been reported that while data analytics can help business to improve the efficiency in their operation and maintenance, the industry operation procedure details are sometimes a core competitive advantage of the company, and thus sensitive to be released [7].

B. Technology Readiness

Increasing Data Generation in Edge: There is an increasing number of edge devices such as smartphones, tablets, wearable devices, industry sensor systems, AR/VR, etc. The average time usage is increasing. A recent report shows that an average user spent 4.3 hours on their mobile device in 2020. Therefore, the amount of data generated in the edge is exploding. The IDC Data Age 2025 report forecasts that the edge devices are expected to create over 90 zettabytes of data by the year of 2025, which is almost 52% of the total data.

Increasing Edge Resources: The edge resources keep increasing. One pillar is the hardware advances, in particular the processor for edge devices, such as FPGA, edge GPU, accelerator, etc. Many hardware companies participate in the chip development, e.g., Nvidia developed Jetson Nano to support embedded IoT applications [8] and apple developed a Bionic chip Xnor to power their flagship iPhones. The other pillar is the continuous upgrade in communication and networking, e.g., recent advances in 5G enable edge devices to communicate in high speeds, high reliability and low latency.

Platform Development: Recently, along with the development of edge computing and edge learning, a number of platforms for decentralized computing and learning emerge. These platforms help reduce the barrier in software development, allowing developers to emphasize on data analytics, rather than facing bare metal hardware and low level details. As an example, TensorFlow Federated[2] (TFF) was developed and released by Google. It is an open source platform for experimenting with machine learning and other computations on decentralized data. To enhance the privacy of edge devices, TensorFlow Privacy was developed and released. It is an open source library based on the differential privacy technique.

Privacy: Privacy has become an important research issue well before the explosion of big data today. The general technique is to add noise so that the original data cannot be easily differentiated. Common privacy concerns such as identity privacy, location privacy, etc have been heavily studied. Quantitative metrics have been developed, such as differential privacy and K-anonymity. There are development in specific fields such as computer vision, for example, visual privacy protection methods are summarized in [9].

C. Case Studies

Google Gboard: Gboard is a virtual keyboard app developed by Google for smartphones. Gboard provides many intelligent features to assist user typing, such as next word prediction, keyboard search suggestion, emoji prediction, etc. Deep learning models are trained for these features. However, user typing behaviors are privacy sensitive. This has led to the proposal of the federated learning paradigm, where the deep learning model training is carried out by local training and only exchanging the gradients among peers.

To improve the quality of service of Gboard, the applied deep learning models should be evaluated iteratively to reflect the change of the smartphone users. In a traditional approach, machine learning model evaluation is executed in a centralized server with collected data. As said, the user data are privacy sensitive, making such an approach again not viable.

This has led to the proposal of federated analytics by Gboard engineers [10] to measure the overall quality of the next word prediction models against raw typing data held on user phones: the participating phones locally computed a metric on how well the model predictions match the words that are actually typed, and then the engineers obtain a population-level summary (i.e., the analytics results, not the raw data) of the model performance by averaging the metrics uploaded by the participating phones. Experiments show the accuracy of mean relative prediction increases for 14.5%.

Federated analytics has also been used in other smartphone applications. For example, Google Pixel is a music recommendation system. In this application, user behaviors are also sensitive data. The music recommendation measurement can apply federated analytics to find the high rank musics without revealing which musics are preferred by an individual phone.

Video Analytics: Video analytics using networked smart cameras has become a core function for many applications

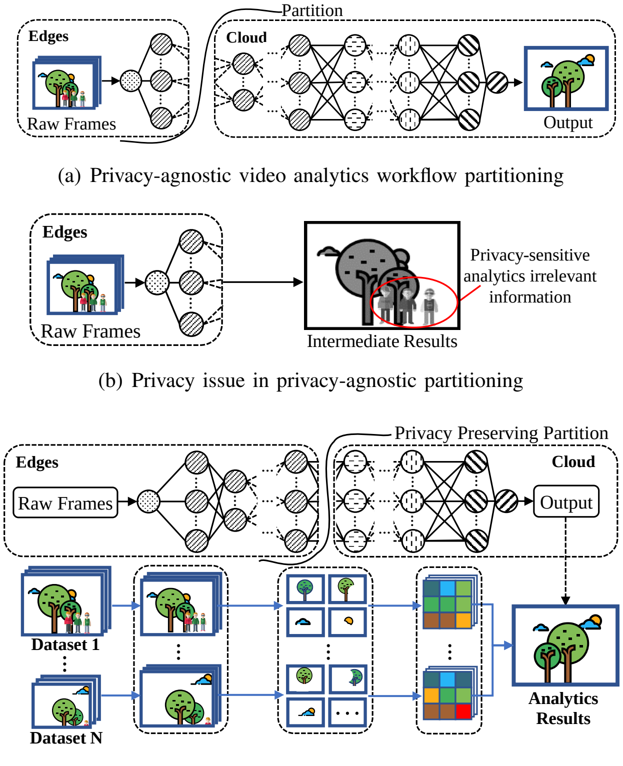
[2] https://www.tensorflow.org/federated

such as surveillance, object detection, AR/VR, etc. Many applications require multiple cameras to collaboratively complete a video analytics task to avoid the problems on limited view scopes, image missing and errors, low-resolution videos, etc. Typical examples include 3D Reconstruction, Multi-view Object Re-identification, etc.

Unfortunately, image sharing can lead to privacy concerns in many applications. One example is the High Definition Map (HD Map) used for autonomous driving. An HD Map has a highly dynamic layer of real-time objects. Vehicles can collectively contribute videos from their on-board cameras to construct such a layer, yet the video images can contain private information, e.g., the plate number of front cars.

Existing edge-cloud video analytics schemes emphasize on resource constraints, e.g., the computing and communication resources in an edge camera can be limited. Video analytics workload partition between edge devices and the cloud is a common theme in optimizing resource utilization. A recent study showed that existing privacy-agnostic workload partition can leak sensitive information and a new federated video analytics architecture, FEVA, was proposed for privacy-preserving collaborative video analytics [2].

We illustrate the core idea of FEVA in Fig. 3. Multiple edge cameras collaboratively contribute images to construct a new comprehensive image, see Fig. 3 (a). The FEVA study observed that current edge-cloud architecture partitions the video analytics workload in a privacy-agnostic way, which can lead to the leakage of analytics irrelevant yet privacy-sentitive information, see Fig. 3 (b). A FEVA architecture is proposed to effectively address such an issue, see Fig. 3 (c). Intrinsically, FEVA keeps the video image data local to the edge for analytics and transmits the analytics results to the cloud for aggregation. It partitions the video analytics computing tasks in a way that is privacy-preserving and maximizes the overall analytics accuracy under the computing and communication resource constraints of the edge devices. It is implemented by extending the open-source platform TensorFlow Federated from Google. The benefit of FEVA is two-fold: (1) an accuracy improvement of 1.90 times as compared to a privacy masking video analytics method which simply removes the privacy information and (2) a 16.80% reduction in communication since the size of features streamed to the server can be reduced significantly. Their measurement also shows that, as compared to the method that performs video analytics in the server, while the computation time increases for 15.9%, the overall latency remains the same due to the reduced communication time.



(a) Privacy-agnostic video analytics workflow partitioning

(b) Privacy issue in privacy-agnostic partitioning

(c) Privacy-preserving video analytics workflow partitioning

Fig. 3: Video anlaytics computing workflows.

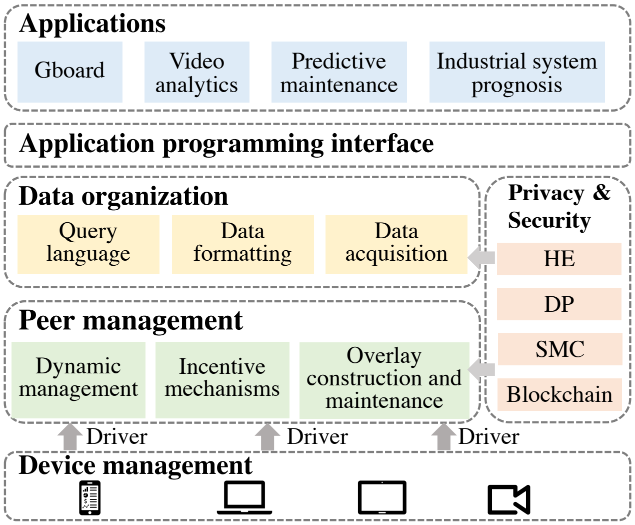


Fig. 4: A federated analytics architecture.

IV. Opportunities and Challenges

A. Architecture

To support the emerging federate analytics applications, it is important to develop architectures with abstractions at various layers. Such architectures can facilitate functional division and accelerate application development. Federated analytics are unique in the sense that the peer devices are organized in a federated manner, the computing tasks are data analytics, and there are various privacy mechanisms. These lead to the necessities on peer management, data organization and privacy management. We show an example federated analytics architecture in Fig. 4.

At the bottom, we can face the problem of heterogeneous edge devices. It is thus necessary to have a device management layer with an abstraction that can provide a uniform logic view of peer devices. Drivers or adapters can be developed so that the heterogeneous edge devices can be registered into the federated analytics system.

On top of the device management layer is a peer management layer, where an overlay of peers is created. There are a set of problems on peer selection, peer searching, peer dynamics, incentive mechanism design, etc. Overlay network and overlay management have been studied in past decades in the research of peer to peer networks. The difference here is that the computing task is not file or video broadcasting, but data analytics with privacy concerns. This brings about new challenges in the peer management designs.

On top of the peer management layer is a data organization layer. One critical part of federated analytics is to have a uniformed view of data. As such, developers only need to worry about what data is needed for the analytics design, rather than how to reach the data and how to handle the heterogeneity of data, e.g., in data format, etc. A new logic data organization is thus necessary. There are a set of problems on data naming, data formatting, data retrieval, etc. A new data querying language can be developed to facilitate data acquisition for the federated analytics application development.

There is a privacy and security management module, which interact with each of the aforementioned layers. For example, Secure multi-party computation (SMC) is a cryptographic method, where computation is carried out on encrypted data; thus data can be organized in encryption to accelerate data analytics computing. We will discuss privacy and security management in detail shortly.

On top of the data organization layer are APIs for application development. The applications can further abstract context related functions. For better service and efficient development, unified management of APIs is necessary. One aspect is to manage the life cycle of APIs, including accessibility, reliability, performance, etc. API management tools such as Apigee by Google can be used. The service If-This-Then-That (ITFFF) helps applications connect for better service. Another aspect is to provide flexibility of APIs, specifically, to make the interactions between applications easier. For example, feature extraction function is served for both 3D reconstruction and target tracking, thus flexible design is necessary for the utilization of both applications. Solutions similar to Home Assistant and openHAB can be considered as options.

The proposed architecture shares similarities with a federated learning architecture. For example, in Tensor Flow Federated, there is also a peer management module, called Client Manager. We illustrate some differences. One example is the data in federated learning are organized in the same format since the computing task is to train a model using all data. For example, to train a YoLov2 model in TensorFlow, the images are formatted into 224x224 image size and divided into a batch with batch size 10. In federated analytics, the computing task may need to select one out of multiple models for analytics. For example, in FEVA, a YoLov2 model and an Alexnet model can be selected during the federated analytics execution. Here the YoLov2 model needs the image size to be 224x224 and the Alexnet model needs the image size to be 128x128. To hid such details from the applications above, indexing structures can be established to organize the data to adapt to lower layer specifics and improve the performance of federated analytics. There are a number of structures such as Skip List, FITing-tree, developed in the past.

B. Privacy and Security Management

There are privacy and security issues in all stages of federated analytics, e.g., in peer management, raw data and intermediate data protection.

There are two broad types of threats in federated analytics: observe threats and tamper threats. In observe threats, the adversary tries to infer the sensitive raw data over the process of federated analytics. For example, inference attacks can learn the characteristic of the private data from the intermediate results. In tamper threats, the adversary tries to degrade the performance (i.e., accuracy or execution time) of the analytics tasks. For example, poisoning attacks attempt to tamper the raw data or intermediate results with some poisoning functions to significantly degrade the accuracy of the final results.

Cryptographic methods are classical privacy and security mechanisms that can be used for observe threats. The basic workflow is as follows. The messages are encrypted by the source participants before transmission. The intermediate operations are carried out on the encrypted messages. Finally, decryption of the encrypted output leads to the final result. SMC and homomorphic encryption (HE) are typical cryptographic methods used in distributed systems. One disadvantage of the cryptographic methods is that the encryption and decryption operations can have high computation overheads, which needs to be addressed in practical systems.

Differential privacy (DP) is also a popular privacy and security mechanism for observe threats. It guarantees that one single record will not influence much on the output of a function. By injecting random noises to the data or the intermediate results, differential privacy provides statistical privacy guarantees for individual records and protection against inference attacks. One disadvantage of DP, e.g., Apple's Sequence Fragment Puzzle (SFP) mechanism, is that the system utility may be largely decreased if heavy noise is injected for high privacy guarantee. Classical privacy mechanisms may need to be properly re-designed to balance the trade-off between privacy degree and system utility, see Google's trie-based heavy hitters (TrieHH).

Blockchain is a trust mechanism that can be used for tamper threats. It is a distributed digital ledger of cryptographically signed transactions with a consensus mechanism. With the proof of work consensus mechanism, the record of intermediate results in blocks cannot be tampered. Thus, tamper threats such as poisoning attacks can be prevented. One disadvantage of blockchain is the computation complexity in solving the puzzles when publishing new transactions. Estimates show that the Bitcoin blockchain network consumes the same amount of electricity as the country of Ireland [11].

There can be special privacy issues in an individual domain. For example, visual privacy has been heavily studied in computer vision. There can be three major categories of visual privacy protection methods. First, the intervention methods, i.e., to prevent sensitive image from being captured from the very first place. Second, the blind vision and secure processing methods, i.e., to conduct privacy-preserving computation, for example, using the SMC method discussed above. Third, the redaction and data hiding methods, i.e., image filtering, image modification, image de-identification, and removal, replace or hide the sensitive section of the image.

C. Resource Management

In federated analytics, edge devices can have resource constraints. Optimization of a number of resources is necessary. We discuss four types of resources: computing power, communication, energy, and monetary cost.

Analytics tasks can bring about non-trivial computing workloads, e.g., a deep learning model inference. There can be different types of computing workloads. For example, a video analytics task can consist of an image filtering operation and a neural network model inference operation. The former brings about instruction intensive computing workloads since it consists of a large number of pairwise addition and subtraction of the two sets of pixels in adjacent image frames in a video. The later brings about data intensive computing workloads since it needs to fit an image into a neural network model. There are different types of computing resources, e.g., CPU, GPU, state-of-the-art FPGAs, and dedicated AI chips from Nvidia, XILinx, Intel, each of which is suitable for certain types of computing workload acceleration. Computing resource management is critical for the system performance in latency, energy consumption, etc.

Communication can be a bottleneck factor for a federated analytics application. For example, in [2], a GAN model is used to generate intermediate insights, which are then delivered to the coordinator for aggregation. However, it is shown that the intermediate data generated by the GAN model can sometimes even be greater than the raw data, and bring about significant communication delay. On the other hand, there are multiple communication methods to choose at the edge devices, in the form of wireline and wireless, with significant difference in costs, delay, reliability, bandwidth variances, etc. Communication choices and optimization are important to the overall system performance.

Battery is a precious resource for edge devices. Both computation and communication consume energy. In the past decades, we have seen a large number of research on low power and energy efficient designs. In federated analytics, a trade-off of the energy and the objectives of the analytics computing task, e.g., accuracy and privacy, is necessary.

There are rental costs in the edge side. Wireless communications have costs, e.g., NB-IoT, 3G/4G, etc. Dedicated communication channels, such as Agora can provide private lines to the cloud with certain QoS guarantee; yet they also have rental costs. Edge side services, e.g., Amazon AWS Panorama with SDK, also have rental costs to provide the computing infrastructure services and software services. Dedicated analytics models have also been developed for sale.

D. Analytics Design and Optimization

The federated analytic mechanism design and optimization are gradually emerging. We present a few examples as follows.

Frequent items discovery: In many applications, discovering heavy hitters (most frequent items) in user-generated data streams is important for improving user experiences, but this operation can also incur substantial privacy risks. A federated analytics scheme TrieHH [12] was proposed. TrieHH is a distributed algorithm using sampling and thresholding techniques. It is proven to inherently support DP without adding additional noise. The trade-off between privacy and utility was also examined; here the utility is the discovery rate of the heavy hitters. TrieHH shows excellent utility while also achieving DP guarantees. A significant advantage of TrieHH is that it eliminates the need to centralize raw data while also avoiding the significant loss in utility incurred by DP. TrieHH was validated both theoretically, using worst-case analyses, and practically, using a Twitter dataset with 1.6M tweets and over 650k users. Compared to Apple's DP method for discovering heavy hitters, the recall and precision of TrieHH are improved up to 1.6 times and 1.0 times, respectively.

Defense of local model poisoning attack: Local model poisoning (LMP) attack is an attack that manipulates the shared local models during the process of distributed learning. Existing defense methods are passive in the sense that they try to mitigate the negative impact of the poisoned local models instead of eliminating them. A recent study [13] show that with federated analytics, proactive analytics can be done for a suspicious peer, while preserving privacy of all peers. More specifically, a Federated Anomaly Analytics enhanced Distributed Learning (FAA-DL) framework was proposed. FAA-DL firstly analyzes all the uploaded local models and observes the potential malicious ones. Then, it requires potential malicious peers to upload raw data with functional encryption, which protects privacy. It verifies each potential malicious local model with HE. Finally, it removes the verified anomalies and aggregates the remaining to produce the global model. Through theoretical analysis, FAA-DL is shown to be robust to strong attacks with high accuracy and low time complexity. The convergence rate is $O\left(\frac{1}{T}\right)$, where T is the training iterations. Experiments show that FAA-DL improves the accuracy of the learned global model under strong attacks for 6.90 times and outperforms the state-of-the-art defense methods with a robustness guarantee.

Data skewness mitigation in federated learning: Non independent and identically distributed (Non-IID) data distribution is a major problem in federated learning: the data distribution is heterogeneous among edge devices, termed as skewness; and since model training is performed locally on the edge devices and cannot be shuffled among peers, the performances of the trained machine learning model can be significantly degraded. A Federated skewness Analytics and Client Selection mechanism (FedACS) [14] is proposed. In principle, FedACS quantifies the skewness of the participating peers by privacy preserving sampling; and selects peers with small skewness. More specifically, FedACS has three steps. First, FedACS derives insights from the local data uploaded to the server. Second, the central server aggregates these insights to estimate the skewness of the clients using Hoeffding's inequality which has provable properties. Third, based on the estimations on drifting, FedACS select clients with milder skewness using a dueling bandit approach. Experiment results show that FedACS reduces the accuracy degradation by 65.6%, and speeds up the convergence 2.4 times.

Privacy preserving distribution estimation: Distribution estimation on a global dataset is a foundation task for a wide range of applications. Existing distribution estimation methods commonly require to have access to the overall population, yet this can lead to privacy concerns in many applications. A recent proposal [15] overcame this issue using federated analytics. More specifically, it models and formulates a joint distribution discovery based federated analytics problem. Then, a federated Markov Chain Monte Carlo with delayed rejection (FMCMC-DR) method is proposed to estimate the representative parameters of the global distribution, which combines the rejection-acceptance sampling and delayed rejection techniques, so that the proposed method is able to explore the entire state space. The method is applied to a digital twin case, where distribution estimation in the digital twin space of the physical space is necessary. Numerical experiments show the high performance of FMCM-DR. Compared to a standard Metropolis-Hastings (MH) algorithm and a random walk Markov Chain Monte Carlo method (RW-MCMC), the proposed algorithm has an improved contour accuracy of 50% and 95%, respectively.

E. Business Model

In addition to the architecture and performance challenges, a good business model is also important for the success of federated analytics applications. We discuss the pricing model and the incentive mechanisms.

From the perspective of federated analytics service providers, federated analytics provide users potential improvement in user experience and quality improvement, e.g., a high accuracy HD map, system robustness against attacks, an accurate grouping of users with similar interests in Floc, or prognosis of industrial systems. Such value-added services should be carefully priced to ensure an overall healthy eco-system in the federated analytics development. Existing pricing models, e.g., pay-as-you-go, time dependent pricing, etc, may not be appropriate for federated analytics services. New pricing models, e.g., model-based pricing, need to be developed to guarantee the profit of the service provider, acceptability of users, as well as other data-related properties, such as data arbitrage-free.

Within the federated analytics peer participants, there should be appropriate incentive mechanisms to ensure the peers to have enough motivation to contribute to federated analytics. Incentive approaches include entertainment, service, and money. Entertainment incentives turn federated analytics into playable games to attract peers. Service incentives are inspired by the principle of mutual benefits. Monetary incentives give participants payments for their contributions. Typical incentive mechanism designs should have properties such as truthfulness, individual rationality, and budget balance. In federated analytics, a special challenge issue is data heterogeneity, and careful design is needed to ensure incentives given the data heterogeneity in peers.

V. Conclusion

In this paper, we presented federated analytics, a new distributed computing paradigm for collaborative data analytics with privacy concerns. We present the definition of federated analytics and clarify its position in the research literature. We showed the two triggers of federated analytics. The first is the real demands on collaborative data analytics and the increasing concerns and restrictions on privacy and confidentiality issues. The second is the technology readiness on edge data and edge resources, and software platform support. These will lead to foreseeable booming of federated analytics. Nevertheless, the application scenarios, design and optimization techniques as well as the business models specifically on federated analytics, are still immature and need clarification. We presented two case scenarios and we presented the opportunities and challenges in the designs of architecture, privacy management, resource management and business models. We presented some examples on federated analytics design and optimization. To expand application development and enrich the federated analytics technologies, collected efforts are clearly needed. This paper serves one effort towards this direction.

Acknowledgment

Dan Wang's work is supported in part by GRF 15210119, 15209220, 15200321, ITF-ITSP ITS/070/19FP, CRF C5026-18G, C5018-20G, PolyU 1-ZVPZ, and a Huawei Collaborative Project. Yifei Zhu's work is partially supported by a SJTU Explore-X Research Grant. Zhu Han's work is partially supported by US NSF CNS-2107216, EARS-1839818 and Toyota.

References

[1] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, 2021.

[2] C. Hu, R. Lu, and D. Wang, "Feva: A federated video analytics architecturefor networked smart cameras," *To appear in IEEE Network*, 2021.

[3] D. Ramage. Federated Analytics: Collaborative Data Science without Data Collection. (2020, May 27). [Online]. Available: https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html

[4] L. E. Parker, "Distributed intelligence: Overview of the field and its application in multi-robot systems," in *Proc. AAAI'07*, Arlington, VA, Nov. 2007.

[5] Y. Wang, Y. Ni, and X. Wang, "Real-time defect detection of high-speed train wheels by using bayesian forecasting and dynamic model," *Mechanical Systems and Signal Processing*, vol. 139, p. 106654, May 2020.

[6] Z. Zheng, Q. Chen, C. Fan, N. Guan, A. Vishwanath, D. Wang, and F. Liu, "Data driven chiller sequencing for reducing HVAC electricity consumption in commercial buildings," in *Proc. ACM e-Energy'18*, Karlsruhe, Germany, Jun. 2018.

[7] Y. Guo, D. Wang, A. Vishwanath, C. Xu, and Q. Li, "Towards federated learning for hvac analytics: A measurement study," in *Proc. ACM e-Energy '20*, virtual, Melbourne, Australia, Jun. 2020.

[8] S. Cass, "Nvidia makes it easy to embed ai: The jetson nano packs a lot of machine-learning power into diy projects - [hands on]," *IEEE Spectrum*, vol. 57, no. 7, pp. 14–16, Jul. 2020.

[9] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "Visual privacy protection methods: A survey," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, Jun. 2015.

[10] K. Wang, R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage, "Federated evaluation of on-device personalization," *arXiv preprint arXiv:1910.10252*, 2019.

[11] A. de Vries, "Bitcoin's growing energy problem," *Joule*, vol. 2, no. 5, pp. 801–805, 2018.

[12] W. Zhu, P. Kairouz, B. McMahan, H. Sun, and W. Li, "Federated heavy hitters discovery with differential privacy," in *Proc. AISTATS'20*, Virtual event, August 2020.

[13] S. Shi, C. Hu, D. Wang, Y. Zhu, and Z. Han, "Federated anomaly analytics for local modelpoisoning attack," *To appear in IEEE Journal on Selected Areas in Communications*, 2021.

[14] Z. Wang, Y. Zhu, D. Wang, and Z. Han, “Fedacs: Federated skewness analytics in heterogeneous decentralized data environments,” in *Proc. IEEE/ACM IWQoS’21*, virtual, Tokyo, Japan, Jun. 2021.

[15] D. Chen, D. Wang, Y. Zhu, and Z. Han, “Digital twin for federated analytics using a bayesian approach,” *To appear in IEEE Internet of Things Journal*, 2021.

BIOGRAPHIES

Dan Wang [S’05, M’07, SM’13] (csdwang@comp.polyu.edu.hk) received his M.Sc degree from Case Western Reserve University in 2004 and his Ph.D. degree in computer science from Simon Fraser University, Canada in 2007. He is currently an associate professor in the Department of Computing at Hong Kong Polytechnic University. His research interests broadly include computer networking and smart energy systems.

Siping Shi (cssshi@comp.polyu.edu.hk) received her M.S. degree in computer applied technology from the University of Chinese Academy of Sciences in 2017. She is currently a Ph.D. candidate at Hong Kong Polytechnic University. Her research interests include edge computing, federated learning and analytics.

Yifei Zhu (yifei.zhu@sjtu.edu.cn) received his MPhil degree from Hong Kong University of Science and Technology in 2015, and his Ph.D. degree in computer science from Simon Fraser University, Canada in 2020. He is currently an assistant professor at Shanghai Jiao Tong University. His research areas include edge computing and multimedia networking.

Zhu Han [S’01, M’04, SM’09, F’14] (zhan2@uh.edu) received his M.S. and Ph.D. from the University of Maryland in 1999 and 2003, respectively. He is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless networking, game theory, big data analysis, security, and smart grid.