Joint Institute for Nuclear Research
Meshcheryakov Laboratory of Information Technologies
10
GRID2023
3-7 July 2023
10th International Conference
"Distributed Computing and Grid Technologies in Science and Education"

Contribution ID: **258** Type: **not specified**

# Horizontally Scalable Digital Footprints Storage And Processing Technologies As An Integral Part of IT-Professional Training For Accelerating Digital Transformation

*Friday 7 July 2023 10:00 (30 minutes)*

## Abstract

The article (proceeding) explores the importance of horizontally scalable technologies for storing and processing digital footprints, a crucial component of IT-professional training for accelerating digital transformation. It begins by defining digital footprints, subsequently addressing their increasing role in modern IT- education and digital transformation. The discussion progresses to the pivotal role of horizontal scalability in digital footprints management and introduces the CAP theorem as a fundamental principle affecting the design of distributed systems. An overview of cutting-edge scalable storage and processing technologies follows, including a discussion on the trend towards relaxing ACID properties for scalability, as implied by the CAP theorem. A comparative analysis of NoSQL databases is presented, highlighting their suitability for storing digital footprints considering CAP constraints. The unique capabilities of Intel DAOS for digital footprint management are also examined. The significance of distributed message brokers in the efficient stream processing of digital footprints is addressed, followed by a brief review of the most popular scalable brokers. The article underscores the role of the Virtual Computer Lab in the training process and its potential impact on digital transformation. It concludes by emphasizing the need for partnerships with leading data centers for integrating High Performance Computing (HPC) solutions into the educational process and outlines potential challenges and solutions in this domain.

## Introduction

Digital transformation refers to the integration of digital technology into all aspects of a business or organization, fundamentally changing how it operates and delivers value to its customers. It’s more than just a change in external business processes—it’s a cultural shift that requires organizations to continually challenge the status quo, experiment, and be comfortable with failure. The transformation may involve changes to business models, ecosystems, and customer engagement, among others, with the end goal of improving operational efficiency and meeting changing customer needs.

The digital transformation journey involves the use of innovative technologies such as cloud computing, big data, artificial intelligence (AI), and the Internet of Things (IoT) to enhance business operations. It also includes the digitization of information, increased use of software and applications, and the use of data analytics to drive decisions. The drive for digital transformation is fueled by changing customer expectations, increased competition, and the need for businesses to stay relevant in a rapidly evolving digital landscape.

Horizontally scalable digital footprints storage and processing technologies refer to systems that can handle increased data load by adding more machines or nodes to the network, rather than upgrading the existing infrastructure. These technologies are designed to accommodate the rapid and often unpredictable growth of digital footprints, which represent the data created and left behind because of individuals’ and organizations’ digital activities.

In the context of digital footprints, storage refers to the technologies used to store the vast amounts of data that these footprints generate. This can include anything from traditional database systems to modern cloud storage solutions. The key is that these technologies need to be scalable, allowing for the addition of more storage capacity as the volume of digital footprints grows.

On the other hand, processing technologies are those that are used to analyze and extract valuable insights from these digital footprints. These technologies include tools and frameworks for big data analytics, machine learning, and other advanced data processing methods. Just like storage technologies, these processing technologies also need to be horizontally scalable to keep up with the increasing volume and complexity of digital footprints.

These horizontally scalable digital footprints storage and processing technologies are crucial in today's digital age, where the volume of data is growing at an unprecedented rate. They enable organizations to effectively manage and gain insights from their digital footprints, thereby driving innovation, improving decision-making, and ultimately accelerating digital transformation.

IT professional training plays a pivotal role in driving successful digital transformation initiatives. As businesses continue to evolve in response to technological advancements, the need for skilled IT professionals who are conversant with emerging technologies and methodologies is paramount.

Digital transformation often involves the implementation of new technologies and processes that may be unfamiliar to an organization's existing IT staff. Professional training helps bridge this skills gap, enabling IT teams to effectively manage, maintain, and optimize these new systems. Training provides IT professionals with the knowledge and skills to not only manage new technologies but also to identify opportunities for their application. This can drive innovation, as employees use their training to find new ways to solve problems and create value.

With the surge in digital activities, cybersecurity risks have also increased. IT professional training in the latest security practices and technologies is critical to safeguarding an organization's digital assets during and after the transformation process.

Training in areas such as data analytics, AI, and machine learning can equip IT professionals to better understand and respond to customer needs, leading to improved customer experiences –a key objective of many digital transformation initiatives. Digital transformation often requires organizations to be more agile and responsive. IT professional training in areas such as DevOps, agile methodologies, and cloud computing can foster this agility, enabling quicker responses to changing market dynamics.

### Definition and overview of digital footprints

Digital footprints refer to the trail of data that individuals and organizations create and leave behind while using the internet and digital services. These footprints can be broadly categorized into two types: active and passive.

*Active Digital Footprints*: These are intentionally created and shared by individuals or organizations. For instance, social media posts, emails, online articles or blogs, and website content all form part of an active digital footprint. When an organization maintains a website or a social media presence, it's creating an active footprint. Similarly, when an individual posts a photo, updates their status, or writes a review online, they contribute to their active digital footprint.

*Passive Digital Footprints*: These are created without the direct intentional action of the user. They are usually generated when different digital services and platforms collect and store data about user activities. Examples include browsing history, location data, search logs, and other metadata that can be collected through cookies, tracking pixels, or other similar technologies.

Both types of digital footprints are valuable sources of data. For individuals, they represent their online identity and behavior, which can impact personal reputation, privacy, and even security. For businesses, digital footprints provide a wealth of information about customers, competitors, and market trends. This data can be analyzed to gain valuable insights, informing strategic decisions, improving products and services, and enhancing customer engagement.

As the volume of digital footprints grows with increased use of digital services, effective management, storage, and processing of this data become increasingly critical. This is where horizontally scalable digital footprints storage and processing technologies come into play, enabling organizations to effectively handle the growing data load and extract valuable insights.

### The role of scalable digital footprints storage and processing in digital transformation

Scalable digital footprints storage and processing technologies play a crucial role in data management. As businesses generate and collect more data, managing this data effectively becomes increasingly challenging. Scalable technologies enable businesses to store and process larger volumes of data, thereby improving data management. They also ensure that as data volumes grow, businesses can continue to store, access, and analyze this data efficiently and effectively. This improved data management capability can help businesses make more informed decisions and gain a competitive advantage.

With the ability to handle larger volumes of data, scalable storage and processing technologies can significantly enhance data analytics capabilities. They enable businesses to analyze larger, more complex datasets, thereby generating more accurate and comprehensive insights. These insights can inform strategic decision-making, improve operational efficiency, and drive business growth. Additionally, scalable technologies can support real-time or near-real-time analytics, enabling businesses to respond quickly to changing conditions and opportunities.

Scalable digital footprints storage and processing technologies can also support the delivery of more customer-centric services. By enabling businesses to collect, store, and analyze large volumes of customer data, these technologies can provide a more detailed understanding of customer behaviors, preferences, and needs. This can inform the development of more personalized, relevant, and responsive services, thereby enhancing the customer experience and promoting customer loyalty.

Finally, scalable digital footprints storage and processing technologies can drive innovation and business growth. By providing the capacity to handle large volumes of data, these technologies enable businesses to explore new ways of using this data, potentially leading to the development of new products, services, or business models. They also support business growth by enabling businesses to manage increasing data volumes as they expand their operations. Furthermore, they can facilitate the identification of trends and opportunities that can drive business growth [1–37].

### The importance of digital footprints in the modern IT-education

A digital footprint becomes an essential aspect of digital citizenship. As our world becomes more digitally interconnected, understanding, and managing digital footprints is becoming an increasingly important skill for students. Education plays a vital role in preparing IT-professionals for this digital reality.

The importance of digital footprints in education is multi-faceted. Here are several ways they can be significant:

- *Learning Opportunities*: Students can use digital footprints to learn about the importance of online safety, privacy, and ethical behavior. The concept of a digital footprint can serve as a real-world example of the consequences of online activities. Educators can use this topic to teach students about these concepts and discuss their implications.
- *Personal Branding*: A digital footprint can be viewed as a personal brand. It's the accumulation of your online activities, including your social media posts, blog entries, comments, and more. This brand can be a positive reflection of a student's personality, skills, and accomplishments. Students can learn how to create a positive online presence that can be beneficial for college applications, scholarships, or job prospects.
- *Critical Thinking and Media Literacy*: Understanding and managing digital footprints can help students develop critical thinking skills. They learn to consider the potential long-term impacts of their online activities and make more informed decisions. This is also tied to media literacy –understanding how information is created, shared, and perceived online.
- *Online Safety and Privacy*: By learning about digital footprints, students can become more aware of their online safety and privacy. They can better understand how their personal information can be accessed, used, and potentially misused, leading to safer online practices.
- *Cyberbullying Prevention*: Understanding digital footprints can help prevent cyberbullying. Students learn that their online activities are traceable, potentially leading to consequences if they engage in harmful behaviors. It can also help victims of cyberbullying understand that there are ways to trace and report harmful actions.
- *Future Opportunities*: Today, colleges and employers often look at the digital footprints of applicants. A well-managed, positive digital footprint can open up opportunities, while a poorly managed one can close them.

By recognizing the significance of digital footprints in education, students can develop crucial skills and knowledge that will serve them well in navigating the digital landscape responsibly and effectively.

As of now, IT professionals training typically covers a broad range of topics, including programming, systems analysis, cybersecurity, and database management. There's a strong emphasis on understanding the fundamentals of computing, problem-solving, and developing software applications. While these subjects are crucial, the rapid growth in data generation and digital transformation initiatives necessitates a shift in focus towards modern data management and processing techniques.

Given the proliferation of data and the increasing reliance on data-driven decision making, it's critical for IT professionals to understand how to manage, store, and process large volumes of data effectively. Businesses are looking for professionals who are familiar with modern, scalable technologies like distributed file systems, NoSQL databases, cloud storages, and distributed computing frameworks. Therefore, incorporating these subjects into IT professional training is crucial to prepare the workforce for the demands of the modern business environment.

Integrating scalable digital footprints storage and processing technologies into IT professional training offers several benefits. It provides IT professionals with the skills needed to manage and analyze large volumes of data, which are critical for driving digital transformation. This training can improve job prospects, as there's a high demand for professionals with these skills. It can also enable professionals to contribute more effectively to their organizations, supporting data-driven decision making and innovation.

System Analysis and Control Department of the Dubna State University have successfully integrated these technologies into their curriculums and offers master's programs that covers scalable data storage and processing technologies, which include courses on distributed computing and machine learning at scale. Graduates of this program have gone on to work in a variety of data-intensive roles, including data scientist, data engineer, and machine learning engineer [38–40].

Similarly, many businesses are investing in internal training programs to upskill their existing staff. For instance, global retail companies such as X5 Retail Group, implemented training programs covering scalable storage and processing technologies as part of its digital transformation initiative. These programs help to build a team capable of leveraging Big Data to improve customer insights, operational efficiency, and decision making.

### The role of horizontal scalability in footprints management

Horizontal scalability, also known as "scaling out," is a method of adding more machines or nodes to a system to improve its performance and capacity as demand increases. This contrasts with vertical scalability, or "scaling up," which involves increasing the capacity of a single machine, such as adding more memory or a faster processor.

In the context of digital footprints storage and processing, horizontal scalability allows a system to handle larger volumes of data by spreading the load across multiple machines. When the system reaches its limit, more machines can be added to continue scaling its capacity. This is typically done in a distributed computing environment, where multiple machines work together to perform a task.

The advantage of horizontal scalability is that it can, theoretically, allow for infinite scaling, as you can continue adding machines as long as you have the resources to do so. It also offers better fault tolerance: if one machine fails, the system can continue to operate by relying on the remaining machines.

Horizontal scalability is a critical feature of modern storage and processing technologies. As the volume and velocity of data generation continue to grow, being able to scale systems horizontally ensures they can handle the increasing load while maintaining performance. This capability is especially important in the realm of big data and real-time processing, where systems must be able to process large volumes of data quickly and efficiently.

Horizontally scalable technologies involve adding more machines or nodes to a system to increase capacity, offer several significant advantages. Here are some of the key benefits:

- *Improved Performance*: Horizontally scalable technologies can improve system performance by distributing workloads across multiple nodes or machines, reducing the load on any single node, and potentially speeding up processing times.
- *Increased Capacity*: By adding more machines or nodes, horizontally scalable systems can handle larger volumes of data or transactions. This is particularly valuable in the age of big data, where the volume and velocity of data generation can be massive and unpredictable.
- *High Availability and Fault Tolerance*: In a horizontally scalable system, if one node fails, the system can continue to operate by relying on the other nodes. This contributes to high availability and fault tolerance, ensuring that services remain up and running, and data loss is minimized.
- *Cost-Effective Scaling*: While the initial setup of a horizontally scalable system can be complex, it can be more cost-effective to scale over time. Rather than replacing existing hardware with more powerful (and often more expensive) machines, we can simply add relatively inexpensive machines or nodes as our needs grow.
- *Flexibility*: Horizontal scalability provides flexibility, allowing you to scale your systems based on demand. It becomes possible to add resources during peak times and reduce them when they're not needed, leading to more efficient use of resources.
- *Better Load Balancing*: Horizontal scalability improves load balancing, as requests can be distributed across multiple servers, reducing the chance of any single server becoming a bottleneck.
- *Easier to Manage*: While managing a distributed system can have its own complexities, in many ways, adding more similar machines can be easier than constantly upgrading a single machine to a more powerful version.

By leveraging these advantages, horizontally scalable technologies can help businesses effectively manage their digital footprints, improve system performance, and ensure high availability –all critical factors in today's fast-paced digital world.

**Overview of topical scalable storage technologies**

Scalable storage technologies are designed to handle a growing amount of data while maintaining performance and reliability. These technologies allow for both horizontal and vertical scalability, but for the sake of this discussion, we will focus on those that are horizontally scalable. Here's an overview of some key scalable storage technologies:

- *Distributed File Systems*: Distributed file systems like Hadoop's HDFS, Google's Cloud Storage, and Amazon's S3 are designed to store large volumes of data across multiple machines in a network. They allow for horizontal scaling by simply adding more machines to the network, thereby increasing storage capacity. They also provide redundancy, ensuring data is not lost even if a machine fails.
- *NoSQL Databases*: Unlike traditional SQL databases, NoSQL databases like Cassandra, Green Plum, MongoDB, Couchbase, etc. are designed to scale horizontally. They distribute data across multiple nodes, and as data volume grows, more nodes can be added to the network. NoSQL databases are particularly well-suited for handling large volumes of unstructured or semi-structured data.
- *Object Storage*: Object storage systems like Intel DAOS, Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage store data as objects rather than in a file hierarchy or block addresses. This makes them highly scalable and ideal for storing unstructured data like multimedia files, which can vary greatly in size.
- *Distributed Block Storage*: Distributed block storage systems like Ceph, GlusterFS, CVMFS break data into blocks and distribute them across multiple nodes. They can scale horizontally by adding more nodes, and they offer high performance and reliability.
- *Cloud Storage Services*: Cloud storage services like Google Cloud Storage, Amazon S3, and Microsoft Azure Storage provide scalable, on-demand storage capacity. They allow businesses to easily scale their storage capacity up or down as needed, without having to invest in additional hardware.
- *Software-Defined Storage (SDS)*: SDS solutions, as Ceph or VMware Virtual SAN separates storage hardware from the software that manages the storage infrastructure. This allows for greater flexibility and scalability,

as storage resources can be managed and allocated dynamically based on application needs.

- *Hyper-converged Infrastructure (HCI)*: HCI combines storage, computing, and networking into a single system to reduce data center complexity and increase scalability. HCI systems use software and x86 servers to replace expensive, purpose-built hardware.
- *Persistent Memory (PMEM)*: PMEM, such as Intel's Optane DC, blurs the line between memory (RAM) and storage. It can retain data even when powered off, like storage, but can be accessed at speeds comparable to memory. This can significantly improve the performance of data-intensive applications.
- *Automated Storage Tiering*: This technology automatically moves data between different types of storage media based on its usage, value, and performance requirements. It helps optimize storage resources and reduce costs.
- *Flash Storage (SSDs)*: Flash storage devices or solid-state drives (SSDs), store data on flash memory chips. They offer faster data access speeds and are more energy-efficient than traditional hard disk drives (HDDs). They are widely used in data centers and for high-performance applications. Of course, they are not new, but nowadays we see the high growth of their storage capacity, durability, high speed, and power –what is very important for green economy and sustainable development.

These scalable and modern storage technologies are integral to managing the large and rapidly growing volumes of data generated in today's digital world. By providing the ability to easily scale storage capacity, they enable businesses to effectively manage their digital footprints and leverage this data to drive insights and innovation.

### Overview of the most popular scalable processing technologies

Scalable processing technologies are designed to handle increasing amounts of data and computational tasks efficiently. As data volume grows, these technologies can distribute the load over more machines or resources, improving performance and ensuring tasks are completed in a timely manner. Here are some key scalable processing technologies:

- *Distributed Computing Frameworks*: Frameworks such as Apache Hadoop and Apache Spark allow for distributed processing of large data sets across clusters of computers. They're designed to scale up from a single server to thousands of machines, with a high degree of fault tolerance.
- *Stream Processing Engines*: Technologies like Apache Kafka and Apache Flink are designed for processing high-volume, real-time data streams. They allow for horizontal scaling and provide capabilities to handle large influxes of data in real-time.
- *NoSQL Databases*: NoSQL databases, such as MongoDB or Green Plum, are not only built to manage large volumes of data across many servers, providing high performance and availability but have integrated MapReduce and other processing functionality.
- *In-Memory Databases*: In-memory databases like Redis and SAP HANA (known as DataMarts) store data in memory rather than on disk for faster processing. They can scale horizontally to handle larger data volumes.
- *Container Orchestration Systems*: Kubernetes, an open-source system for automating deployment, scaling, and management of containerized applications, allows for horizontal scaling based on the demand or load on the system.
- *Serverless Computing*: Serverless computing platforms like AWS Lambda and Oracle or Google Cloud Functions allow for automatic scaling of application functionality. They can run code in response to events and automatically manage the resources required by the code.
- *GPU-Accelerated Computing*: GPU-accelerated computing leverages the parallel processing capabilities of GPU (Graphics Processing Units) for computational tasks. This can dramatically speed up workloads like machine learning, data analysis, and computational science.
- *Machine Learning Frameworks*: Machine learning frameworks like TensorFlow and PyTorch have capabilities to distribute computation across multiple GPUs, multiple machines, or large-scale cloud-based deployments, enabling scalable data processing and model training.

These scalable processing technologies enable organizations to handle the growing volume and complexity of data, supporting data-driven decision-making, real-time insights, and advanced analytics. They play a critical role in managing and gaining value from digital footprints in the era of big data and digital transformation.

### Horizontal scalability in favor of ACID relaxing

The decision not to use ACID (Atomicity, Consistency, Isolation, Durability) for digital footprints could be driven by the need for scalability, real-time processing, analytics efficiency, compatibility with distributed systems, and specific application requirements.

Digital footprints often generate a large volume of data. ACID transactions can introduce overhead and impact performance when processing and storing such high-volume data. By relaxing ACID properties, systems can achieve higher scalability and performance by prioritizing data ingestion and processing speed over transactional consistency.

Digital footprints often capture events and activities that occur in real-time. Achieving strong consistency in such scenarios, where data is continuously changing and distributed across multiple systems, can be challenging. By relaxing ACID properties, systems can adopt eventual consistency, where data consistency is guaranteed over time, but not necessarily at the exact moment of data ingestion.

Digital footprints are frequently used for analytics and reporting purposes, where complex queries and aggregations are performed on the data. ACID transactions may hinder the performance and efficiency of these analytical processes, as they often involve large-scale data processing. By loosening ACID guarantees, systems can optimize query performance and improve overall analytics capabilities.

In modern corporate or social environments, digital footprints are often generated and stored across distributed and decentralized systems, such as cloud-based platforms, consumer, banking, medical, transport or learning management systems, and mobile applications. Coordinating ACID transactions across these disparate systems can be complex and resource intensive. Embracing more relaxed consistency models, like eventual consistency, can simplify the integration and synchronization of data from multiple sources.

The requirements for data consistency and transactional guarantees vary across different applications and use cases. For some digital footprint scenarios, a certain level of inconsistency or data staleness may be tolerable without significantly impacting processes or decision-making. By tailoring the consistency requirements to specific use cases, systems can optimize performance and resource utilization.

**Comparative analysis to validate the choice of NoSQL databases for storing digital footprints**

Most of NoSQL solutions are designed to handle large amounts of data, but they have different focuses, strengths, and weaknesses, and are designed for different types of workloads.

In our courses at the Institute of System Analysis and Control, we often prefer Apache Cassandra due to its descriptive and demonstrative circle architecture and gossip protocol of metadata exchange, as well as it handles large volumes of data and thousands of concurrent users or operations per second. It allows easily add more servers to increase capacity, providing high availability with no single point of failure and capable of handling a high write load.

It could be a good choice for digital footprints collecting and storage because of scalability, high availability, and performance are critical as well as Cassandra is a column-oriented database, which is excellent for storing and querying large amounts of structured, semi-structured, or unstructured data. That's possible due to ACID relaxing, through Cassandra is not designed to support complex transactions with multiple operations or joins like a relational database. In terms of CAP (Consistency, Availability, Partition Tolerance) theorem, Apache Cassandra, it is designed to prioritize Availability and Partition Tolerance (AP), but not Consistency, which is common to most NoSQL databases. Cassandra offers eventual consistency, meaning that if no new updates are made to a given data item, eventually all accesses to that item will return the latest updated value. This is a relaxation of the consistency guarantee in favor of availability and partition tolerance. However, despite being primarily AP database, Cassandra allows the consistency level to be tuned per operation. For example, it's possible to specify that a write must be sent to two, three, or all nodes in a replica set. This provides some flexibility, but still falls short of the full consistency guarantee that CP systems provide, as well as in case of a network partitioning, Cassandra chooses to remain available, accepting writes even if they cannot be immediately replicated to all nodes (hinted handoff feature). Also, Cassandra has write availability option, as long as a single replica for the data being written is up and reachable, the write can succeed.

However, other NoSQL databases can also be used to store digital traces, considering their advantages and disadvantages.

*MongoDB*:

MongoDB is a document-oriented NoSQL database, making it highly flexible and adaptable. It supports a rich and dynamic data model, which can be an advantage when dealing with unstructured or semi-structured data.

*Pros*: MongoDB is a document-oriented database that supports a rich and flexible data model. It's easy to scale horizontally and offers automatic sharding. It also provides robust support for developer productivity with multiple SDK as well as high performance for read and write operations, especially for operations that involve large volumes of data. MongoDB supports multiple indexes, including secondary indexes, which can greatly improve the speed of data retrieval.

*Cons*: MongoDB might not perform as well with transaction-heavy applications. Also, tuning MongoDB for performance can sometimes be complex. It can consume a lot of system memory, especially under heavy load, which might be a concern in resource-constrained environments. MongoDB's query language and indexing options are powerful, but they can also be complex to understand and use correctly, especially for complex queries and aggregations.

*Redis*:

Redis, which stands for Remote Dictionary Server, is an in-memory data structure store that can be used as a database, cache, or message broker. It supports various types of data. Redis offers data persistence, so it's possible to snapshot the in-memory database onto disk either by time or by the number of writes since the last snapshot.

*Pros*: Redis provides very fast data access as it's an in-memory data store, making it ideal for caching and real-time analytics. It supports various data structures like strings, hashes, lists, and sets. Redis has a built-in publish/subscribe messaging system, which is useful for real-time messaging use cases.

*Cons*: Being an in-memory database, Redis can be limited by memory size. For persistence, it requires periodic saving of the dataset to disk which might impact performance.

*CouchDB*:

CouchDB is one more NoSQL database developed by Apache, which focuses on ease of use and embracing the web. It is a single-node database that works just as well on a shared server as it does on a large distributed system where multi-master replication, allowing to have multiple copies of your data, thus ensuring

high availability and disaster recovery. CouchDB is not designed to handle complex relationships between documents. It's best for use cases where documents can stand alone. In certain use cases, such as large-scale writes or complex queries, CouchDB may not perform as well as other NoSQL databases.
*Pros*: CouchDB supports a multi-master replication system, making it a good choice for distributed systems. It also provides a RESTful interface for interaction.
*Cons*: CouchDB might not be the best option for applications that require complex querying or aggregations.

*Couchbase*:
Couchbase is a NoSQL document database with a distributed architecture for performance, scalability, and availability. It enables developers to build applications easier and faster by leveraging the power of SQL with the flexibility of JSON. Couchbase has an in-memory-first architecture, offering high speed for read and write operations. It provides horizontal scalability with a distributed architecture where data is automatically partitioned across all available nodes. Couchbase offers a SQL-like query language, making it easier for developers coming from a SQL background to create and manage data and has built-in full-text search capabilities, making it easier to find relevant information in a large dataset. It stores data in flexible JSON documents, providing the flexibility to modify the schema on-the-fly.
*Pros*: Couchbase provides powerful indexing and querying capabilities. It's known for its high performance, scalability, and flexible JSON model.
*Cons*: Couchbase can be resource intensive compared to some other databases, meaning it might require more powerful hardware to run effectively. Also, the learning curve can be a bit steep due to its unique architecture and features where some features come with a cost, and it might be more expensive than other solutions.

*HBase*:
HBase is a distributed, scalable, big data store and a part of the Apache Hadoop ecosystem that provides random, real-time read/write capabilities on top of the Hadoop Distributed File System (HDFS). HBase is designed to scale linearly with the addition of more hardware. It can host large tables on top of clusters of commodity hardware.
*Pros*: HBase, built on Hadoop, is designed for large tables with billions of rows. It provides real-time read/write access and integrates well with Hadoop ecosystem tools. HBase is designed to scale linearly with the addition of more hardware. It can host large tables on top of clusters of commodity hardware. Unlike many other Hadoop tools which are oriented towards batch processing, HBase provides real-time read and write access to your big data. HBase guarantees strong consistency for reads and writes, which can be a critical requirement for certain types of applications. HBase integrates well with other Hadoop ecosystem tools. It uses Hadoop's distributed file system for storing its data and can be a source or destination for MapReduce jobs.
*Cons*: HBase is not suitable for low-latency applications due to its write-ahead log design. It also requires a fair amount of setup and maintenance.

*Kudu*:
*Pros*: Kudu is excellent for fast scans due to its design for columnar storage, which makes it ideal for analytical queries and real-time analytics. Unlike many other Hadoop-compatible storage options, Kudu supports real-time data insertion, updates, and deletes, making it suitable for scenarios requiring fast data modifications. Kudu is designed to integrate well with Hadoop ecosystem tools, like MapReduce, Spark, and Impala, providing a flexibility of choice for processing frameworks as well as it's designed with a distributed architecture that is meant to scale and handle failures.
*Cons*: Kudu is not the best choice for storing large objects or blobs and may not perform as well as some other data stores for heavy write workloads. Like other distributed systems, managing and configuring Kudu can be complex.

*Greenplum*:
Greenplum Database owned by VMware is an open-source, massively parallel processing (MPP) SQL database management system. It's designed to manage large-scale analytic data warehouses and business intelligence workloads.
*Pros*: The MPP architecture of Greenplum enables it to scale linearly, both in terms of data volume and query performance, by simply adding more nodes to the system. As a relational database management system, Greenplum fully supports SQL, including many advanced features. This makes it easy for users familiar with SQL to use Greenplum. Also, Greenplum offers data compression techniques which allow it to store large amounts of data efficiently and integrates well with various data formats and sources, including CSV, Avro, and Parquet files, as well as external databases via JDBC or ODBC.
*Cons*: Greenplum is optimized for analytical workloads and large queries across massive amounts of data. It's not designed for transactional workloads (OLTP). Compared to some more widely adopted databases, there may be less community support and fewer readily available resources for troubleshooting and optimization. As with most distributed systems, Greenplum can be complex to set up and manage.

*Neo4j*:
Neo4j is a highly scalable, native graph database purpose-built to leverage not only data but also the connections between data. It's designed to handle high-complexity queries with ease.
*Pros*: Neo4j, as a graph database, is excellent for handling data where relationships are key. It supports ACID

properties and provides a powerful query language, Cypher.
*Cons*: Neo4j might not scale horizontally as easily as some other NoSQL databases. Also, it can be more resource-intensive for storing and querying data compared to other types.

*Elasticsearch*:
Elasticsearch is a distributed, open-source search and analytics engine built on Apache Lucene. It's designed for horizontal scalability, reliability, and easy management, and is often used for log and event data analysis, as well as search functionality in applications. Elasticsearch can easily scale horizontally to handle large amounts of data while maintaining fast response times and Apache Lucene library allows providing of powerful full-text search capabilities with a very comprehensive set of querying and filtering options.
*Pros*: Elasticsearch is excellent for searching and analyzing large amounts of data in near real-time. It is scalable, distributed, and can index many types of content.
*Cons*: Elasticsearch might be overkill for simple search use-cases. Also, managing and maintaining an Elasticsearch cluster can be complex.

*InfluxDB*:
InfluxDB is an open-source database written in Go language and developed by InfluxData. It's specifically designed for time-series data, which are data points that are timestamped. This makes it highly suitable for logging, sensor data, real-time analytics, and monitoring systems. Influx DB can handle high write loads and still query effectively, making it a good choice for applications that need to write and read data rapidly. With the introduction of InfluxDB 2.0, InfluxData introduced a new scripting and query language called Flux, which is more powerful and flexible than the InfluxQL used in InfluxDB 1.x. Influx DB is a part of InfluxData Stack, a larger set of tools developed by InfluxData, including Telegraf for data collection, Chronograf for visualization, and Kapacitor for real-time streaming data processing and alerting.
*Pros*: InfluxDB is designed specifically for time-series data, making it a good fit for applications such as monitoring systems, IoT sensor data, real-time analytics, and metrics collection. It offers high write and query performance (for example, in some situations it could be 5x faster than Cassandra, or 1.5x faster than MongoDB). The database is optimized for fast, high-availability storage and retrieval of time series data. InfluxDB uses a lossless data compression, which reduces the amount of storage necessary for large volumes of data. Built-in HTTP API allows for direct interaction with the database without a need for a separate server or middleware, making integrations easier. Flux language is specifically designed for time series data and includes many built-in functions for time series analysis. It follows a functional programming model, which can be more intuitive for certain types of data manipulation, particularly time-based and streaming data. Flux not only retrieves data, but it also offers extensive capabilities for transforming and processing that data and allows for joining of data across different buckets (equivalent to databases in the relational model), which is beneficial for complex queries in Influx DB.
*Cons*: InfluxDB is a time-series database and is not designed to store complex, relational data. Thus, it may not be suited for applications requiring complex joins or transactions. However, the open-source version of InfluxDB does not support for authentication (SAML/SSO), data replication and scaling, automated backups, high availability, disaster recovery, data encryption, etc.

*Riak KV (key-value) and TS (time series)*:
Riak KV is a distributed NoSQL key-value database with advanced local and multi-cluster replication that guarantees reads and writes even in the event of hardware failures or network partitions. Riak TS is a key-value NoSQL database that has been optimized for time-series data.
*Pros*: Riak KV is known for its high availability, fault tolerance, and operational simplicity. It offers excellent scalability and easy data recovery. Riak TS supports linear and horizontal scalability, making it suitable for applications that need to grow over time or handle large spikes in traffic as well as special features as automated data co-location, which can improve the efficiency of range queries. Riak is designed to survive network partitions and server failures with no single point of failure, which makes it highly reliable and available.
*Cons*: Riak KV may not be as efficient for use cases that require complex queries or transactions. Also, its community support is considered less robust compared to other NoSQL databases. Riak doesn't support complex querying capabilities out of the box. Queries are limited to key-value pairs and range queries on keys. As Riak is an AP (Available and Partition-tolerant) system as per the CAP theorem, there can be temporary inconsistencies in data during network partitions. However, it does offer eventual consistency. Compared to other databases, the community and ecosystem around Riak and Riak TS might not be as large, which could lead to fewer resources for troubleshooting and learning.

Storing digital footprints, a task that encompasses the collection, storage, and analysis of varied and extensive sets of user behavior data, requires a database solution that's not only robust and scalable but also flexible enough to handle complex, semi-structured data.
NoSQL databases are particularly well-suited for this task, due to their ability to store non-relational data, horizontal scalability, and flexibility in terms of the schema. However, it's important to consider that while these databases are powerful, they each have their trade-offs. For example, while Redis can provide extremely quick access to data, its in-memory nature might not be best for persistent storage of large datasets. On the other hand, HBase could handle vast datasets but may not be suitable for low-latency applications.

In the end, the ideal database for storing digital footprints will depend on various factors, such as the volume, variety, and velocity of data being generated, the need for real-time processing and analytics, the complexity and type of queries you'll need to perform, and the resources available for database management and optimization. When choosing a database for your specific use case, consider conducting a comprehensive analysis that takes these factors into account to ensure the technology aligns well with your project's requirements.

### Intel DAOS capabilities for digital footprint management

Distributed Asynchronous Object Storage (DAOS) is an open-source software-defined object store that provides high bandwidth, low latency, and high I/O operations per second (IOPS) storage containers to HPC applications and workflows. It's developed by Intel and primarily designed to leverage next-generation NVM (Non-Volatile Memory) technologies like Storage Class Memory (SCM), NVMe (Non-Volatile Memory express), Optane Persistent Memory (3D XPoint).

DAOS has strong self-healing capabilities powered by placement maps that are stored on each storage target and I/O node. In case of storage target failure, it can rebuild the target in the background to maintain data redundancy. It achieves fault tolerance using erasure coding and replication. It is designed for extreme-scale storage and supports an almost unlimited number of Pools (storage clusters), Containers (user-defined storage units), and Objects (data units) and allows flexible and efficient resource utilization. It follows a Software-Defined Storage (SDS) approach, separating the data path from the control path. This allows it to bypass the kernel in the data path and make full use of the capabilities of NVM Express SSDs and Optane Persistent Memory.

Also, DAOS is meant to be a part of a larger ecosystem. It can be used in combination with other components like middleware libraries (HDF5, MPI-IO), distributed file systems (like Lustre, NFS), and data management services (like Apache Hadoop and Spark).

In the context of digital footprints, the data structure would likely be event-based, where each event represents a person interaction with a digital tool or platform. Each event could be stored as an object in DAOS, with attributes such as the Identifier (ID), the timestamp of the event, the type of event, and any additional data associated with the event.

### The role of distributed message brokers in footprints stream processing

Each action users take (like logging in, viewing a lesson, completing a quiz) can be considered a digital footprint and can be sent as a message to a broker. Multiple consumers (like analytics systems, monitoring tools, recommendation engines) can then independently process these messages.

That's why message brokers can be especially useful in handling digital footprints for several reasons:

• *Decoupling*: Message brokers allow different parts of a system to communicate without being directly connected. This can help decouple the system, making it easier to modify, scale, and maintain.

• *Reliability*: Message brokers often provide features like message persistence, delivery acknowledgments, and retry mechanisms, which help ensure that messages aren't lost even if some parts of the system fail.

• *Scalability*: Message brokers can help distribute work among multiple consumers. If the volume of digital footprints increases, additional consumers can be added to handle the load.

• *Asynchronous Processing*: The processing of digital footprints can be done asynchronously, which is especially useful if the processing is time-consuming. The system can continue to accept new digital footprints while processing others.

• *Ordering and Timing*: Some message brokers can ensure that messages are processed in the order they were sent, or schedule messages to be processed at a certain time.

• *Buffering*: In the case of spikes in data, message brokers can act as a buffer, holding onto messages until the consumers are ready to process them.

Popular message brokers include Apache Kafka, RabbitMQ, Amazon SQS, etc. Each has its own strengths and is suited to different types of tasks, so the choice of broker would depend on the specific requirements of the system handling digital footprints.

### Brief review of the most popular scalable message brokers

Message brokers play a crucial role in modern distributed systems as well as in digital footprints processing. They enable applications to communicate with each other, often in a publish-subscribe model, making them essential for event-driven architectures and real-time data processing tasks.

*Apache Kafka*:

*Pros*: Kafka is designed to handle real-time, high-volume data streams. It can be scaled horizontally to handle more data by adding more machines to the network. Kafka stores streams of records in categories called topics. Each topic is replicated across a configurable number of Kafka brokers to ensure data is not lost if a broker fails. Also, Kafka can be used with real-time processing systems like Apache Storm or Apache Samza.

*Cons*: Kafka's distributed system, while powerful, brings complexity and can be challenging to set up and manage. It's possible to encounter with lack of advanced message routing, due to Kafka primarily relies on topic-based routing. Also, Kafka relies on ZooKeeper for managing and coordinating brokers, which adds to its complexity.

*RabbitMQ*:

*Pros*: RabbitMQ supports several messaging protocols, including AMQP, STOMP, MQTT, and HTTP. It allows

advanced message routing and offers a variety of message routing options through exchanges, including direct, topic, headers and fanout. RabbitMQ is developer-friendly and has a large and active community and propose excellent developer support and client libraries in many languages.
*Cons*: RabbitMQ has lower throughput and may not perform as well as Kafka under high volumes of data. Also, It stores messages in memory, which can lead to high memory usage.

*Apache Pulsar*:
*Pros*: Pulsar is a unified messaging and streaming systems that provides both messaging (comparable to RabbitMQ) and event streaming (comparable to Kafka) capabilities, making it versatile for different use cases. Pulsar's architecture separates serving and storage layers, allowing for independent scaling and potentially improving performance and stability as well as it supports multi-tenancy for special needs to isolate different teams or applications within the same cluster. Pulsar supports configuring message replication across multiple datacenters out of the box, which is useful for creating distributed and resilient applications.
*Cons*: Pulsar is not as mature as Kafka, RabbitMQ, or Amazon SQS, which means it may not have as large of a community or as many resources available for troubleshooting and support. While Pulsar has more built-in features compared to some other systems, managing these features can add operational complexity.

Of course, it's possible to use fully managed Cloud Message Queuing for microservices, distributed systems, and serverless applications such as Amazon SQS, Azure Event Grid, Notifications Hub, or Google Cloud Pub/Sub etc., but this is beyond the scope of our review.

Message brokers act as a central hub to collect, integrate, and route data from various sources and facilitate seamless data integration, ensuring that digital footprints are captured efficiently. Also, they enable real-time processing of digital footprints. As data is ingested into the message broker, it can be immediately processed, transformed, and analyzed in real time as well as routed or filtered based on specific criteria or rules. Message brokers can store digital footprints data for a certain period or until consumed by the consuming applications or systems. This provides a temporary storage mechanism that ensures data availability and fault tolerance. Additionally, it allows replaying or reprocessing of data in case of failures or the need for historical analysis.

**Role of the Virtual Computer Lab in training process and its anticipated impact on Digital Transformation**

Open educational cloud datacenter «Virtual Computer Lab» created in the Institute of System Analysis and Control by Mikhail Belov (https://belov.global) in 2007. Nowadays it's being actively developed by all the institute's leading professionals and plays a crucial (and possibly a critical) role in IT-professional training, particularly in the context of learning scalable digital footprints storage and processing technologies. Virtual Computer Lab provides a virtual environment where students can learn, practice, and experiment with these technologies. Here are some ways in which Virtual Computer Lab contributes:

- *Practical Experience*: Virtual Computer Lab allows students to gain hands-on experience with the technologies they are learning about. They can run experiments, troubleshoot issues, and see the effects of their actions in real-time, which can enhance their understanding and skills.
- *Accessibility*: With the Virtual Computer Lab, students can access the lab environment from anywhere, at any time. This makes learning more flexible and convenient, as students can practice and learn at their own pace, without being constrained by the physical availability of lab resources.
- *Scalability and Flexibility*: Virtual Computer Lab can be easily scaled up or down to accommodate different numbers of students or different learning needs. They can also be easily updated or reconfigured to incorporate new technologies or tools, making them a flexible learning resource.
- *Safe Environment for Learning*: In the Virtual Computer Lab, students can experiment freely without the risk of causing damage to physical equipment. When students make a mistake, they can simply reset the virtual environment and start over. This encourages experimentation and learning from mistakes, which is critical for mastering new technologies.
- *Real-world Simulation*: Virtual Computer Lab is designed to mimic real-world scenarios, providing students with practical experience that is directly applicable to the workplace. For example, students can learn how to manage and analyze large volumes of data in a simulated business environment, preparing them for real-world data management tasks.

In the context of learning scalable digital footprints storage and processing technologies, the Virtual Computer Lab provides a powerful platform for developing practical skills and understanding. It helps prepare IT-professionals to drive digital transformation initiatives effectively and efficiently.

Virtual Computer Lab contributes a better understanding of scalable digital footprints storage and processing technologies among IT-professionals. As more professionals are equipped with the necessary skills to handle large volumes of data and leverage these for insights, businesses can more effectively and efficiently transition their operations to digital platforms, resulting in an overall acceleration in the pace of digital transformation.

When IT-professionals are trained to use scalable technologies effectively, it can lead to significant improvements in efficiency and productivity. These technologies enable businesses to manage and analyze large volumes of data more efficiently, leading to faster decision-making and improved operational efficiency. This can result in higher productivity and better business outcomes.

By leveraging scalable digital footprints storage and processing technologies, companies can gain a competitive edge in the market. With more professionals trained in these technologies, businesses can more effectively harness their data for insights, leading to innovations in products, services, and business models. This can help businesses differentiate themselves from competitors and gain a significant competitive advantage [41–64].

### Fundamentals of strategy for digital footprints integration into the training of IT-professionals

Developing a comprehensive curriculum is the first step towards integrating scalable digital footprints storage and processing technologies into IT professional training. The curriculum should cover key topics such as distributed computing, NoSQL databases, cloud storage, and data analytics at scale. It should also include modules on emerging technologies and trends to keep students abreast of the latest developments. Moreover, the curriculum should be designed in a way that builds on foundational IT knowledge and progressively introduces more complex concepts and skills [40].

Hands-on practical training and simulations in the Virtual Computer Lab are critical for effective learning. They allow students to apply the theoretical knowledge they gain in a practical context, enhancing their understanding and skills. Training programs should include lab sessions, projects, and simulations where students can work with real-world data and use scalable storage and processing technologies. These practical experiences can help students understand the challenges of managing and processing large volumes of data and learn how to overcome them.

Collaborating with industry partners can enrich IT professional training. Industry partners can provide valuable insights into the real-world applications of scalable digital footprints storage and processing technologies, helping to ensure that the training is relevant and practical. They can also offer internships, projects, and guest lectures, providing students with practical experience and exposure to industry practices. Such collaborations can help bridge the gap between academia and industry and ensure that students are job-ready when they graduate.

Given the rapid pace of technological advancement, continual learning and upskilling are crucial. IT-professionals need to regularly update their knowledge and skills to stay relevant. Training programs should therefore provide opportunities for continual learning, such as advanced courses, workshops, and seminars on emerging technologies and trends. They should also encourage students to pursue industry certifications, which can enhance their skills and employability. Moreover, a culture of lifelong learning should be fostered, encouraging students to take responsibility for their own professional development.

### The importance of partnership with leading data centers to introduce HPC solutions into the educational process

Partnerships with leading data centers such as JINR (Joint Institute for Nuclear Research) or CERN (European Organization for Nuclear Research) and are of great importance for the implementation of HPC (High-Performance Computing) solutions in preparing IT-professionals. Here are some reasons why such partnerships are significant:

• *Cutting-edge Infrastructure*: Collaborating with renowned data centers like JINR or CERN provides access to state-of-the-art infrastructure and supercomputing resources. These institutions invest heavily in high-performance computing systems, enabling advanced computational capabilities that are essential for training IT professionals in complex and data-intensive tasks.

• *Expertise and Knowledge Sharing*: Partnering with these leading data centers allows for valuable knowledge sharing and collaboration. JINR and CERN are home to some of the brightest minds in scientific research and computational science. Working closely with their experts provides an opportunity to exchange ideas, best practices, and innovative techniques in HPC, thus enriching the training of IT professionals.

• *Reputation and Credibility*: Partnering with internationally recognized institutions like CERN and JINR enhances the credibility and reputation of an organization involved in IT professional training. It signifies a commitment to excellence and cutting-edge technologies, attracting talented individuals and establishing credibility among potential employers.

We can see the greatest example, when *Vladimir V. Korenkov*, the legendary IT-expert in Russian Federation, the Scientific Director of Meshcheryakov Laboratory of Information Technologies at JINR is responsible for setting strategic direction for the integration of HPC in education, determining what resources are necessary and how they could best be deployed to benefit students and researchers. He leads project teams in the development and implementation of HPC solutions. *Vladimir Korenkov* provides significant technical insights and guidance, helping to solve problems and make decisions on which technologies to use. He also plays a tremendous role in developing educational materials and courses that teach students and researchers how to use and benefit from these HPC resources.

It's very important because the proliferation of digital technologies and increased internet penetration globally has led to an explosion in the quantity of digital footprints. Every click, like, comment, share, download, or upload that we perform online leaves a trace. These traces, known as digital footprints, are generated at an unprecedented volume, velocity, and variety. This not only includes social media interactions but also extends to e-commerce transactions, web searches, and even sensor data from IoT devices.

Every day, billions of people around the world use the internet, each leaving their unique digital footprints. The sheer scale of this data is enormous and still growing. According to estimates, the global data sphere will grow to 175 zettabytes by 2025, up from 33 zettabytes in 2018. A significant portion of this data will be digital footprints. Processing and making sense of this vast ocean of data using traditional methods or standard computing systems is not feasible due to the size and complexity of the data.

HPC solutions are designed to process and analyze massive amounts of data efficiently. They use parallel processing to perform high-speed computation tasks, making them well-suited for handling the volume, velocity, and variety of digital footprints. HPC can help in the real-time processing of these data, spotting trends, patterns, and anomalies.

Furthermore, the use of HPC is not just about handling the sheer size of the data. It's also about the need for sophisticated, high-speed analytics. This might involve complex machine learning algorithms to predict future behavior based on digital footprints or advanced graph analytics to understand the relationships between different entities. These tasks can be computationally intensive, further justifying the need for HPC solutions.

In essence, the enormous and ever-growing number of digital footprints necessitates the use of HPC solutions. Not only can HPC manage the scale of the data, but it can also facilitate the type of high-speed, advanced analytics needed to extract meaningful insights from these footprints. In this way, HPC becomes not just desirable, but essential in the era of Big Data.

### Challenges and Potential Solutions

Integrating scalable digital footprints storage and processing technologies into IT-professional training comes with its own set of challenges. Technologies evolve rapidly, and staying current can be a daunting task. Training providers must constantly update their curriculum and teaching methods to ensure relevance. Some educational institutions might struggle with limited resources, both in terms of finances and expertise. Investing in the necessary tools and technologies, as well as training the trainers, could be a challenge. Bridging the gap between theoretical knowledge and practical skills is a significant challenge. Without adequate practical exposure, learners might struggle to understand the real-world applications of the technologies.

Institutions must adopt a dynamic approach towards curriculum design, constantly updating their content to keep up with technological advancements. Collaborating with industry partners can help mitigate the resource constraint issue. Industry partners can provide the necessary financial support, tools, and expertise. blend of theoretical instruction and practical exposure can ensure that students acquire both knowledge and hands-on skills. Virtual Computer Lab, project-based learning, and internships can help facilitate practical learning.

As AI and machine learning continue to advance, these technologies will play a crucial role in processing and analyzing digital footprints. The rise of online learning and Virtual Computer Labs will provide more flexible and accessible learning opportunities for IT professionals worldwide.

### Conclusion

This article delved into the concept of digital transformation and how horizontally scalable digital footprints storage and processing technologies are integral to it. It discussed how IT-professional training is a key factor in accelerating global digital transformation efforts. The benefits of horizontal scalability were outlined, along with a deep dive into scalable storage and processing technologies.

The crucial role of these technologies in enhancing data management, fostering customer-centric services, influencing data analytics, and promoting business growth was highlighted. It was demonstrated how integrating these technologies into IT professional training could significantly enhance the readiness of IT professionals and thus contribute to digital transformation efforts.

The importance of strategies for effective integration, including curriculum development, hands-on practical training, industry partnerships, and continual learning opportunities, was also discussed. The anticipated impact of integrating these technologies into IT-professional training on global digital transformation was examined, pointing towards an accelerated pace of digital transformation, improved efficiency and productivity, and fostered innovation.

In conclusion, the importance of horizontally scalable digital footprints storage and processing technologies in IT-professional training cannot be overstated. The ability to effectively manage, store, and process ever-increasing volumes of data is a core skill set for the IT professionals of today and tomorrow.

As businesses increasingly turn to data for decision-making and innovation, having IT-professionals trained in these technologies is a critical element in accelerating global digital transformation. It provides businesses with the necessary technical expertise to leverage their data effectively and can drive significant improvements in efficiency, productivity, and innovation.

In an era characterized by rapid technological change, it is crucial for IT professional training programs to continually evolve and incorporate the latest technologies and practices. In doing so, they will equip the IT-professionals with the skills they need to navigate the digital landscape, drive digital transformation efforts, and ultimately contribute to the growth and success of their organizations.

Under the continued leadership of professor *Evgenia N. Cheremisina*, the Institute of System Analysis and Control has emerged as a leading institution in the realm of information technology, systems analysis, and control systems. With a focus on cutting-edge research and innovative educational practices, the institute is making notable contributions to the scientific community and the wider world.

*Evgenia Cheremisina*'s guidance has been pivotal in driving the institute towards excellence. Her vision and commitment to innovation have steered the institute's focus towards pivotal technologies like horizontally scalable digital footprints storage and processing technologies, positioning the institute at the forefront of the digital transformation era.

With a strong emphasis on high-quality education and industry-relevant training, *Evgenia Cheremisina* and her successor *Elena Yu. Kirpicheva*'s leadership have played an instrumental role in preparing the next generation of IT-professionals. Under their guidance, the institute has designed comprehensive IT-professional training programs that integrate the latest technologies and practices, preparing students to drive digital transformation efforts in their future roles.

Overall, the Institute of System Analysis and Control has positioned itself as a pioneering institution in the field of IT, continually advancing knowledge, driving innovation, and shaping the future IT-professionals. *Evgenia Cheremisina*, *Elena Kirpicheva*,*Nadezhda Tokareva*, *Snezhana Potemkina*'s unwavering commitment to excellence, innovation, and student success has been instrumental in this regard and promises an exciting future for the institute.

**References**

1. Muhammad S.S., Dey B.L., Syed Alwi S.F., Kamal M.M., Asaad Y. Consumers'willingness to share digital footprints on social media: the role of affective trust // Inf. Technol. People. 2023. Vol. 36, № 2. P. 595–625.

2. Jayasuriya D.D., Ayaz M., Williams M. The use of digital footprints in the US mortgage market // Account. Finance. 2023. Vol. 63, № 1. P. 353–401.

3. Tucakovic L., Bojic L. Computer-based personality judgments from digital footprints: theoretical considerations and practical implications in politics // Srp. Polit. Misao. 2022. Vol. 74, № 4/2021. P. 207–226.

4. Shiells K., Di Cara N., Skatova A., Davis O., Haworth C., Skinner A., Thomas R., et al. Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. // Int. J. Popul. Data Sci. 2022. Vol. 5, № 3.

5. Rowe F. Using digital footprint data to monitor human mobility and support rapid humanitarian responses // Reg. Stud. Reg. Sci. 2022. Vol. 9, № 1. P. 665–668.

6. Pereverzeva E., Komov A. The mechanism for digital footprints formation // Vestn. St Petersburg Univ. Minist. Intern. Aff. Russ. 2022. Vol. 2022, № 1. P. 128–133.

7. Loutfi A.A. A framework for evaluating the business deployability of digital footprint based models for consumer credit // J. Bus. Res. 2022. Vol. 152. P. 473–486.

8. Grove W., Goldin J.A., Breytenbach J., Suransky C. Taking togetherness apart: From digital footprints to geno-digital spores // Hum. Geogr. 2022. Vol. 15, № 2. P. 163–175.

9. Gorbatov S., Krasnova E.A., Samara State Transport University. A digital footprint as a mechanism of individualizing a student's educational trajectory (on the case of the "Digital technologies of self-education" course) // Perspect. Sci. Educ. 2022. Vol. 58, № 4. P. 193–208.

10. Feng S., Chong Y., Yu H., Ye X., Li G. Digital financial development and ecological footprint: Evidence from green-biased technology innovation and environmental inclusion // J. Clean. Prod. 2022. Vol. 380. P. 135069.

11. Dyachenko M., Leonov A. Digital footprint in education as a driver of professional growth in the digital age // E-Manag. 2022. Vol. 5. P. 23–30.

12. Vayndorf-Sysoeva M.E., Pchelyakova V.V. Prospects for Using the Digital Footprint in Educational and Scientific Processes // Vestn. Minin Univ. 2021. Vol. 9, № 3.

13. Shmatko A., Barykin S., Sergeev S., Thirakulwanich A. Modeling a Logistics Hub Using the Digital Footprint Method—The Implication for Open Innovation Engineering // J. Open Innov. Technol. Mark. Complex. 2021. Vol. 7, № 1. P. 59.

14. Pozdeeva E., Shipunova O., Popova N., Evseev V., Evseeva L., Romanenko I., Mureyko L. Assessment of Online Environment and Digital Footprint Functions in Higher Education Analytics // Educ. Sci. 2021. Vol. 11, № 6. P. 256.

15. Pavlenko D., Barykin L., Dadteev K. Collection and analysis of digital footprints in LMS // Procedia Comput. Sci. 2021. Vol. 190. P. 666–669.

16. Liu X., Huang Q., Gao S., Xia J. Activity knowledge discovery: Detecting collective and individual activities with digital footprints and open source geographic data // Comput. Environ. Urban Syst. 2021. Vol. 85. P. 101551.

17. Lapchik D.M., Fedorova G.A., Gaidamak E.S. Digital Footprint in the Educational Environment as a Regulator of Student Vocational Guidance to the Teaching Profession // J. Sib. Fed. Univ. Humanit. Soc. Sci. 2021. Vol. 14, № 9. P. 1388–1398.

18. Bushuyev S., Onyshchenko S., Bushuiev D., Bushuieva V., Bushuyeva N. Dynamics and impact of digital footprint on project success // Sci. J. Astana IT Univ. 2021. № 6. P. 15–22.

19. Songsom N., Nilsook P., Wannapiroon P., Fung L.C.C., Wong K.W. System Design of a Student Relationship Management System Using the Internet of Things to Collect the Digital Footprint // Int. J. Inf. Educ. Technol. 2020. Vol. 10, № 3. P. 222–226.

20. Olinder N., Tsvetkov A., Fedyakin K., Zaburdaeva K. Using Digital Footprints in Social Research: an Interdisciplinary Approach // WISDOM. 2020. Vol. 16, № 3. P. 124–135.

21. Hicks B., Culley S., Gopsill J., Snider C. Managing complex engineering projects: What can we learn from

the evolving digital footprint? // Int. J. Inf. Manag. 2020. Vol. 51. P. 102016.

22. Sürmelioğlu Y., Seferoğlu S.S. An examination of digital footprint awareness and digital experiences of higher education students // World J. Educ. Technol. Curr. Issues. 2019. Vol. 11, № 1. P. 48–64.

23. Songsom N., Nilsook P., Wannapiroon P., Chun Che Fung L., Wong K.W. System Architecture of a Student Relationship Management System using Internet of Things to collect Digital Footprint of Higher Education Institutions // Int. J. Emerg. Technol. Learn. IJET. 2019. Vol. 14, № 23. P. 125.

24. Martin F., Gezer T., Wang C. Educators'Perceptions of Student Digital Citizenship Practices // Comput. Sch. 2019. Vol. 36, № 4. P. 238–254.

25. Boudlaie H., Nargesian A., Keshavarz Nik B. Digital footprint in Web 3.0: Social Media Usage in Recruitment // AD-Minist. Universidad EAFIT, 2019. № 34. P. 131–148.

26. Muhammad S.S., Dey B.L., Weerakkody V. Analysis of Factors that Influence Customers'Willingness to Leave Big Data Digital Footprints on Social Media: A Systematic Review of Literature // Inf. Syst. Front. 2018. Vol. 20, № 3. P. 559–576.

27. Micheli M., Lutz C., Büchi M. Digital footprints: an emerging dimension of digital inequality // J. Inf. Commun. Ethics Soc. 2018. Vol. 16, № 3. P. 242–251.

28. Gutiérrez Puebla J. Big Data y nuevas geografías: la huella digital de las actividades humanas // Doc. Anàlisi Geogràfica. 2018. Vol. 64, № 2. P. 195.

29. Guha S., Kumar S. Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare: Past Contributions and Future Roadmap // Prod. Oper. Manag. 2018. Vol. 27, № 9. P. 1724–1735.

30. Azucar D., Marengo D., Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis // Personal. Individ. Differ. 2018. Vol. 124. P. 150–159.

31. Buchanan R., Southgate E., Smith S.P., Murray T., Noble B. Post no photos, leave no trace: Children's digital footprint management strategies // E-Learn. Digit. Media. 2017. Vol. 14, № 5. P. 275–290.

32. Önder I., Koerbitz W., Hubmann-Haidvogel A. Tracing Tourists by Their Digital Footprints: The Case of Austria // J. Travel Res. 2016. Vol. 55, № 5. P. 566–573.

33. Lewis K. Three fallacies of digital footprints // Big Data Soc. 2015. Vol. 2, № 2. P. 205395171560249.

34. Thatcher J. Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data // Int. J. Commun. 2014. Vol. 8. P. 19.

35. Lambiotte R., Kosinski M. Tracking the Digital Footprints of Personality // Proc. IEEE. 2014. Vol. 102, № 12. P. 1934–1939.

36. Gantz J., Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east // IDC IView IDC Anal. Future. 2012. Vol. 2007, № 2012. P. 1–16.

37. Weaver S.D., Gahegan M. Constructing, Visualizing, and Analyzing A Digital Footprint // Geogr. Rev. 2007. Vol. 97, № 3. P. 324–350.

38. CC2020 Task Force. Computing Curricula 2020: Paradigms for Global Computing Education. New York, NY, USA: Association for Computing Machinery, 2020. 205 p.

39. Куприяновский В.П., Сухомлин В.А., Добрынин А.П., Райков А.Н., Шкуров Ф.В., Дрожжинов В.И., Федорова Н.О., Намиот Д.Е. Навыки в цифровой экономике и вызовы системы образования // Int. J. Open Inf. Technol. 2017. Vol. 5, № 1.

40. Сухомлин В.А., Зубарева Е.В. Куррикулумная парадигма - методическая основа современного образования // Современные Информационные Технологии И Ит-Образование. 2015. Vol. 11, № 1.

41. Grishko S., Belov M., Cheremisina E., Sychev P. Model for creating an adaptive individual learning path for training digital transformation professionals and Big Data engineers using Virtual Computer Lab // Creativity in Intelligent Technologies and Data Science / ed. Kravets A.G., Shcherbakov M., Parygin D., Groumpos P.P. Cham: Springer International Publishing, 2021. P. 496–507.

42. Belov M., Grishko S., Cheremisina E., Tokareva N. Concept of peer-to-peer caching database for transaction history storage as an alternative to blockchain in digital economy // CEUR Workshop Proc. 2021. Vol. 3041. P. 494–497.

43. Belov M.A., Korenkov V.V., Potemkina S.V., Lishilin M.V., Cheremisina E.N., Tokareva N.A., Krukov Y.A. Methodical aspects of training data scientists using the data grid in a virtual computer lab environment // CEUR Workshop Proc. 2019. Vol. 2507. P. 236–240.

44. Belov M.A., Krukov Y.A., Mikheev M.A., Lupanov P.E., Tokareva N.A., Cheremisina E.N. Essential aspects of it training technology for processing, storage and data mining using the virtual computer lab // CEUR Workshop Proc. 2018. Vol. 2267. P. 207–212.

45. Cheremisina E.N., Belov M.A., Tokareva N.A., Nabiullin A.K., Grishko S.I., Sorokin A.V. Embedding of containerization technology in the core of the Virtual Computing Lab // CEUR Workshop Proc. 2017. Vol. 2023. P. 299–302.

46. Belov M.A., Tokareva N.A., Cheremisina E.N. F1: The cloud-based virtual computer laboratory - An innovative tool for training // 1st Int. Conf. IT Geosci. 2012. 2012. P. undefined-undefined.

47. Belov M., Korenkov V., Tokareva N., Cheremisina E. Architecture of a compact Data GRID cluster for teaching modern methods of data mining in the Virtual Computer Lab // EPJ Web Conf. / ed. Adam Gh., Buša J., Hnatič M. 2020. Vol. 226. P. 03004.

48. Белов М.А., Живетьев А.В., Подгорный С.А., Токарева Н.А., Черемисина Е.Н. Подход к управлению виртуальной компьютерной лабораторией на основе концептуальной модели операционных рисков // Моделирование Оптимизация И Информационные Технологии. 2023. Vol. 11, № 1 (40).

49. Белов М.А., Лишилин М.В., Черемисина Е.Н., Стифорова Е.Г. Роль проектно-ориентированного технологического предпринимательства в стратегии развития ит-образования в условиях цифровой трансформации // Современные Наукоемкие Технологии. 2022. № 11. P. 86–96.

50. Белов М.А., Гришко С.И., Живетьев А.В., Подгорный С.А., Токарева Н.А. Применение методов нечеткой логики для формирования адаптивной индивидуальной траектории обучения на основе динамического управления сложностью курса // Моделирование Оптимизация И Информационные Технологии. 2022. Vol. 10, № 4 (39). P. 7–8.

51. Белов М.А., Гришко С.И., Черемисина Е.Н., Токарева Н.А. Подготовка ИТ-специалистов в условиях глобальной цифровой трансформации. Концепция автоматизированного управления профилями компетенций в образовательных программах будущего // Современные Информационные Технологии И ИТ-Образование. 2021. Vol. 17, № 3. P. 658–669.

52. Белов М.А., Гришко С.И., Лишилин М.В., Осипов П.А., Черемисина Е.Н. Стратегия подготовки ИТ-специалистов с применением инновационного учебного дата-центра “Виртуальная компьютерная лаборатория”для эффективного решения задач цифровой трансформации и акселерации цифровой экономики // Современные Информационные Технологии И ИТ-Образование. 2021. Vol. 17, № 1. P. 134–144.

53. Белов М.А., Лупанов П.Е., Минзов А.С., Токарева Н.А. Система управления виртуальной инфраструктурой на основе визуальных моделей в среде виртуальной компьютерной лаборатории // Современная Наука Актуальные Проблемы Теории И Практики Серия Естественные И Технические Науки. 2019. № 6–2. P. 41–46.

54. Белов М.А., Крюков Ю.А., Михеев М.А., Лупанов П.Е., Токарева Н.А., Черемисина Е.Н. Повышение продуктивности освоения распределённых информационных систем в виртуальной компьютерной лаборатории на основе применения технологий контейнеризации и оркестровки контейнеров // Современные Информационные Технологии И ИТ-Образование. 2018. Vol. 14, № 4. P. 823–832.

55. Белов М.А., Крюков Ю.А., Лупанов П.Е., Михеев М.А., Черемисина Е.Н. Концепция когнитивного взаимодействия с виртуальной компьютерной лабораторией на основе визуальных моделей и экспертной системы // Современная Наука Актуальные Проблемы Теории И Практики Серия Естественные И Технические Науки. 2018. № 10. P. 27–35.

56. Белов М.А., Лупанов П.Е., Токарева Н.А., Черемисина Е.Н. Концепция усовершенствованной архитектуры виртуальной компьютерной лаборатории для эффективного обучения специалистов по распределённым информационным системам различного назначения и инструментальным средствам проектирования // Современные Информационные Технологии И ИТ-Образование. 2017. Vol. 13, № 1. P. 182–189.

57. Лишилин М.В., Белов М.А., Токарева Н.А., Сорокин А.В. Концептуальная модель системы управления знаниями для формирования профессиональных компетенций в области ит в среде виртуальной компьютерной лаборатории // Фундаментальные Исследования. 2015. № 11–5. P. 886–890.

58. Белов М.А., Лишилин М.В., Токарева Н.А., Антипов О.Е. От виртуальной компьютерной лаборатории к управлению знаниями. Итоги и перспективы // Качество Инновации Образование. 2014. № 9 (112). P. 3–14.

59. Черемисина Е.Н., Белов М.А., Лишилин М.В. Анализ ключевых активностей жизненного цикла управления знаниями в вузе и формирование концептуальной модели архитектуры системы управления знаниями // Открытое Образование. 2013. № 3 (98). P. 34–41.

60. Черемисина Е.Н., Митрошин П.А., Белов М.А. Комплексные системы электронного обучения как инструментарий оценки компетенций учащихся // Наука И Бизнес Пути Развития. 2013. № 5 (23). P. 113–122.

61. Черемисина Е.Н., Белов М.А., Антипов О.Е., Сорокин А.В. Инновационная практика компьютерного образования в университете “Дубна”с применением виртуальной компьютерной лаборатории на основе технологии облачных вычислений // Программная Инженерия. 2012. № 5. P. 34–41.

62. Белов М.А., Антипов О.Е. Контрольно-измерительная система оценки качества обучения в виртуальной компьютерной лаборатории // Качество Инновации Образование. 2012. № 3 (82). P. 28–32.

63. Антипов О.Е., Белов М.А. Технология применения виртуальной компьютерной лаборатории в учебных курсах ВУЗа // Естественные И Технические Науки. 2012. № 1 (57). P. 260–268.

64. Антипов О.Е., Белов М.А., Токарева Н.А. Архитектура виртуальной компьютерной лаборатории для подготовки специалистов в области информационных технологий // Компьютерные Инструменты

В Образовании. 2011. № 4. P. 37–44.

## Summary

This material describes the importance of horizontally scalable technologies for storing and processing digital footprints in IT professional training, aiming to expedite digital transformation and shows the amazing perspectives of partnership between Institute of System Analysis and Control (Dubna State University) and Meshcheryakov Laboratory of Information Technologies (JINR).

**Primary author:** BELOV, Mikhail (Dubna State Univeristy)

**Co-authors:** Mr ZHIVETIEV, Alex (Dubna State University); NECHAEVSKIY, Andrey (JINR); MILOVIDOVA, Anna; Mr SMIRNOV, Dmitry (Dubna State University); REZVAYA, Ekaterina; KIRPICHEVA, Elena (Dubna International University of Nature, Society, and Man. Institute of system analysis and management); Dr CHEREMISINA, Evgenia (Dubna State University); KIROV, Evgenii (State University Dubna); Mrs EKATERINA, Goryunova (Dubna State University); Ms BALASHOVA, Marina (Dubna State University); Mr LISHILIN, Mikhail (Dubna State University); TOKAREVA, Nadezhda (Dubna Univeristy); KREIDER, Oksana (Крейдер Оксана); STRELTSOVA, Oksana (JINR); TYATYUSHKINA, Olga (Dubna Univeristy); Mr MITROSHIN, Pavel (Dubna State University); Mr ZORIN, Roman (Dubna State University); POTEMKINA, Snezhana (Dubna State University); KORENKOV, Vladimir (JINR); ЖАТКИНА, Кристина (Dubna State University)

**Presenter:** BELOV, Mikhail (Dubna State Univeristy)

**Session Classification:** Plenary

**Track Classification:** Plenary