УДК 004.67
DOI: https://doi.org/10.17721/AIT.2021.1.09

Liudmyla Zubyk, orcid.org/0000-0002-2087-5379,
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
Yaroslav Zubyk, orcid.org/0000-0002-0802-3552,
National University of Water and Environmental Engineering, Rivne, Ukraine

Architecture of modern platforms for big data analytics

Big data is one of modern tools that have impacted the world industry a lot of. It also plays an important role in determining the ways in which businesses and organizations formulate their strategies and policies. However, very limited academic researches has been conducted into forecasting based on big data due to the difficulties in capturing, collecting, handling, and modeling of unstructured data, which is normally characterized by it's confidential. We define big data in the context of ecosystem for future forecasting in business decision-making. It can be difficult for a single organization to possess all of the necessary capabilities to derive strategic business value from their findings. That's why different organizations will build, and operate their own analytics ecosystems or tap into existing ones. An analytics ecosystem comprising a symbiosis of data, applications, platforms, talent, partnerships, and third-party service providers lets organizations be more agile and adapt to changing demands. Organizations participating in analytics ecosystems can examine, learn from, and influence not only their own business processes, but those of their partners. Architectures of popular platforms for forecasting based on big data are presented in this issue.

Keywords: *big data, unstructured data, platforms for data analytic, data ecosystems, big data environment.*

Для цитування (for citation): L. Zubyk, Y. Zubyk. "Architecture of modern platforms for big data analytics," *Сучасні інформаційні технології*, vol. 1, p. 67–74, 2021.

INTRODUCTION

The world is changing, and the variations in living conditions that we have all seen lately play a key role in just such a resource as information. Modern information technologies are already forming new structures and creating enormous opportunities for development. Finally, only those who can adapt to new conditions faster will be able to win. The amount of data is growing exponentially, a clear example is the amount of information generated over the past 2 years by users of social networks. Currently, the daily data flow is more than 8 TB, and this number is growing daily. Millions of photos, videos, texts, hundreds of terabytes of information are uploaded to the public every minute. Information content in the form of documents, books, movies is constantly digitized. IoT devices and sensors regularly update the system with information from their owners, which allows them to build appropriate behavioral models and use them for future development. IoT and machine learning are causing the global Big Data market grow rapidly, and is forecasted to 103 billion U.S. dollars by 2027 [1].

FORMULATION OF THE PROBLEM

The big data resource is universal, dynamic and inexhaustible. Due to the availability of significant amounts of collected data, high quality digital information and its availability in developed countries among the world leaders in using of Big Data were primarily the United States, Great Britain, China, Switzerland, South Korea etc. [2].

Among the world-famous companies working with Big Data: iTechArt, ScienceSoft, Xplenty, IBM, HP Enterprise, Teradata, Oracle, SAP, EMC, Amazon, Microsoft Google, VMware, Splunk, Alteryx, Cogito, etc. [3].

This direction is quite new for Ukraine. Among the projects that have gone far beyond the country – Grammarly (a system for constructing spell-checking algorithms based on Big Data). Both government agencies (the Ministry of Digital Transformation) and business representatives are interested in big data analytics.

Big Data in the world is used in almost all areas: medicine, telecommunications, logistics, urban planning, retail, energy, agriculture and finance, space exploration. Customers of services for the

analysis of large arrays of information are both large holdings and representatives of medium and small businesses. Powerful streaming platforms (Netflix, Spotify, etc.) analyze video and audio content; Amazon studies shopping history; Tinder creates individual portraits of users. Most Big Data projects focus on either expanding the customer base or developing existing customers and improving existing services [10; 19].

Large amounts of customer data can be available for analysis only with the direct participation of their holders, ie banks, retailers, mobile operators and so on. Data sets of Internet service users always provide a good basis for building analytical models.

The processing of personal data of customers places strict requirements on companies to comply with current legislation in terms of protection of personal data of users, so all analytical models using Big Data are based solely on impersonal data [4; 5].

The first Big Data cases for customers appeared a long time ago and since then many companies have entered the market, specializing in the development of specific platforms and solutions for customers. Big Data analysts and specialists work with an internal platform for data processing and analysis based on Open Source technologies, constantly creating new products and improving existing ones.

2020 and 2021 tested many companies for strength in the face of constant uncertainty, abrupt changes in the business environment and unusually severe constraints. Successful examples of those companies that have managed not only to stay afloat, but also to improve their business performance, include the network of fuel filling stations UPG, Brain, online store ITbox.ua and more.

Data analysis for business customers is often aimed at developing and involving in business development targeting programs to organize the influx of customers through bonus offer systems; construction of portraits of clients to determine the target groups and their needs; expanding markets by forming Look-Alike audiences; active use of heatmap and geoanalytics in order to optimize the integration of new objects into the infrastructure; organization of communication with existing and potential customers depending on the events that occur with them (trigger mailing); replacing traditional SMS with Viber messages, given that they may contain useful multimedia interactive content, etc. [20].

Large amounts of information are only part of success. The data flow detached from the analytics does not provide any advantages. For further use, the information needs to be cleaned and reworked. The growing popularity of Big Data is largely due to changes in the technologies and infrastructure used for processing [17].

A data ecosystem is a collection of infrastructure, analytics, and applications used to capture and analyze data. Data ecosystems provide companies with data that they rely on to understand their customers and to make better decisions. Like real ecosystems, data ecosystems are intended to evolve over time. There is no one ‘data ecosystem’ solution. Every business creates its own ecosystem, sometimes referred to as a technology stack, and fills it with a patchwork of hardware and software to collect, store, analyse, and act upon the data.

METHODS OF THE RESEARCH

Methods of the research - empirical, or rather observation and comparison.

COMPARING OF DIFFERENT PLATFORMS ARCHITECTURE FOR DATA ANALYTICS

The best data ecosystems are built around a product analytics platform that ties the ecosystem together. Analytics platforms help teams integrate multiple data sources, provide machine learning tools to automate the process of conducting analysis, and track user cohorts so teams can calculate performance metrics.

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. This software is suitable for applications that have large data sets. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. A core architectural goal of HDFS is detection of faults and quick, automatic recovery from them. Software is designed more for batch processing. The emphasis is on high throughput of data access HDFS is tuned to support large files. HDFS applications need a write-once-read-many access model for files. A file once created, written, and closed need not be changed. This assumption simplifies data coherency issues and enables high throughput data access. A computation requested by an application is much more efficient if it is executed near the data it operates on. HDFS has been designed to be easily portable from one platform to another. It has a master/slave architecture (Fig. 1).

The NameNode and DataNode are pieces of software designed to run on commodity machines. The system is designed in such a way that user data never flows through the NameNode. HDFS supports a traditional hierarchical file organization.

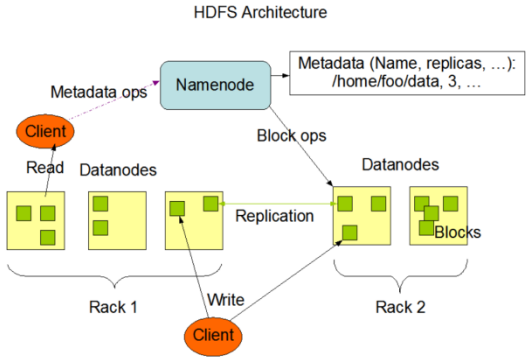


Fig. 1. HDFS Architecture

It doesn't yet implement user quotas. HDFS doesn't support hard links or soft links. However, the HDFS architecture doesn't preclude implementing these features.

An application can specify the number of replicas of a file that should be maintained by HDFS.

The number of copies of a file forms the replication factor of that file. This information is stored by the NameNode. It stores each file as a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are replicated for fault tolerance. The block size and replication factor are configurable per file. Optimizing replica placement distinguishes HDFS from most other distributed file systems. The NameNode uses a transaction log called the EditLog to persistently record every change that occurs to file system metadata.

All HDFS communication protocols are layered on top of the TCP/IP protocol.

The primary objective of HDFS is to store data reliably even in the presence of failures. The three common types of failures are NameNode failures, DataNode failures and network partitions.

DataNode sends a Heartbeat message to the NameNode periodically. The HDFS architecture is compatible with data rebalancing schemes. When a client creates an HDFS file, it computes a checksum of each block of the file and stores these checksums in a separate hidden file in the same HDFS namespace. The client can opt to retrieve that block from another DataNode.

NameNode can be configured to support maintaining multiple copies of the FsImage and EditLog. This may degrade the rate of namespace transactions per second that a NameNode can support. The NameNode machine is a single point of failure for an HDFS cluster. Currently, automatic restart and failover of the NameNode software to another machine is not supported. A typical block size used by HDFS is 64 MB. A client request to create a file does not reach the NameNode immediately. In fact, initially the HDFS client caches the file data into a temporary local file. When the local file accumulates data worth over one HDFS block size, the client contacts the NameNode. The NameNode inserts the file name into the file system hierarchy and allocates a data block for it.

HDFS can be accessed from applications in many different ways. Natively, HDFS provides a Java API for applications to use. A C language wrapper for this Java API is also available. In addition, an HTTP browser can also be used to browse the files of an HDFS instance. Work is in progress to expose HDFS through the WebDAV protocol.

It provides a command-line interface called FS shell that lets a user interact with the data in HDFS. The DFSAdmin command set is used for administering an HDFS cluster.

When a file is deleted by a user or an application, it isn't immediately removed from HDFS.

When the replication factor of a file is reduced, the NameNode selects excess replicas that can be deleted.

Azure Databricks is a platform optimized for the Microsoft Azure cloud services platform. This platform includes such environments for developing applications: SQL Analytics and Workspace (Fig. 2) [13].

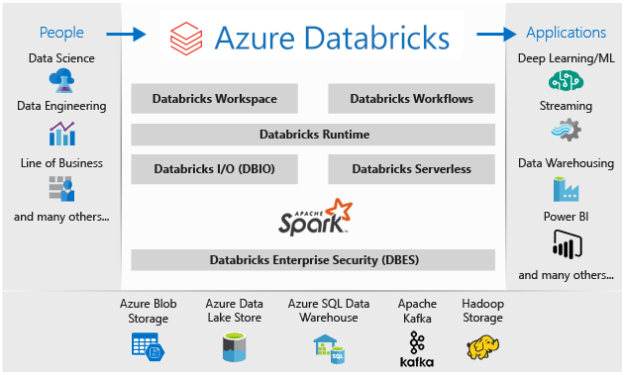


Fig. 2. Azure Databricks architecture

SQL Analytics in Azure Databricks provides an simple platform for data analysis in case of SQL queries against a data lake, supports a few types of visualizations needed results for special contexts, and using of dashboards.

The Azure Databricks Workspace provides an interactive workspace for collaboration between different specialists. In the big data pipeline, this data is received in Azure through Azure Data Factory as batches, or streamed in near real-time. This data ends up in a data lake for long-term storage in Azure Blob storage or Azure Data Lake Storage. As part of

your analytics workflow, you can use Azure Databricks to read data from a variety of data sources and get actionable insights using Spark.

The Azure Databricks workspace is a high-performance platform based on Apache Spark. The Azure Databricks workspace integrates with Azure, providing an interactive workspace, easy setup, and simplified workflows.

In the big data pipeline, this data (raw or structured) is received in Azure via the Azure Data Factory in packets or transmitted in near real-time streaming via Apache Kafka, the Event Hub, or the Internet of Things Center. This data enters a data lake for long-term storage in the Azure or Azure Data Lake Storage BLOB.

As part of the analytics workflow, you can use Azure Databricks to read data from multiple data sources, such as Azure BLOB, Azure Data Lake Storage, Azure Cosmos DB, or SQL Azure Data Storage, and retrieve useful statistics from them using Spark (Fig. 3).

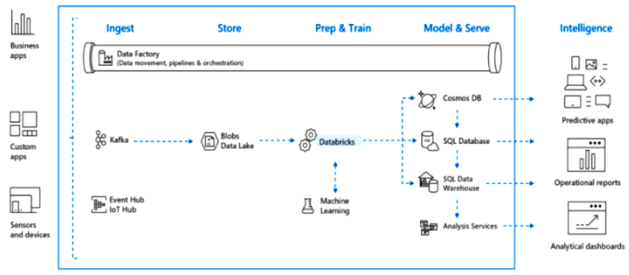


Fig. 3. Multiple data sources in Azure Databricks

The central component of **Amazon EMR** is the cluster. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is a node that has a role within the cluster, connected with the node type. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application alike Apache Hadoop (Fig. 4).

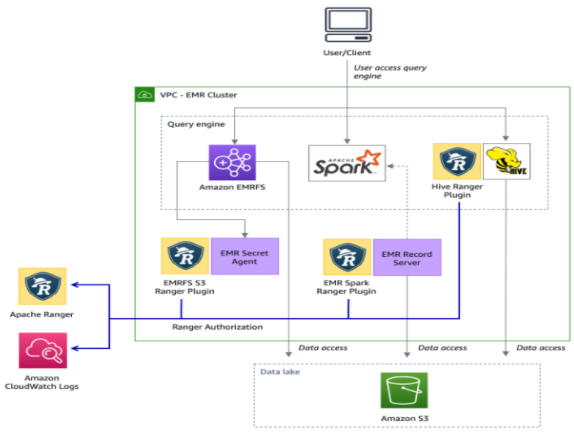


Fig. 4. Architecture of Amazon EMR

When you run a cluster on Amazon EMR, you have several options as to how you specify the work that needs to be done. When you launch your cluster, you choose the applications to install for your data processing needs. To process data in your Amazon EMR cluster, you can submit jobs or queries directly to installed software, or you can run steps in the cluster. A failure during the cluster lifecycle causes Amazon EMR to loss all data of it's instances in this cluster.

There are some benefits to using Amazon EMR: cost savings; AWS integration; deployment; scalability and flexibility; reliability; security; monitoring; management interfaces. But in this issue we are investigating to architecture's specific for different solutions first of all. Amazon EMR integrates with other AWS services to provide capabilities and functionality related to your cluster. Several examples of this integration: Amazon EC2 for the instances that comprise the nodes in the cluster; Amazon Virtual Private Cloud (Amazon VPC) to configure the virtual network in which you launch your instances; Amazon S3 to store input and output data; Amazon CloudWatch to monitor cluster performance and configure alarms; AWS Identity and Access Management (IAM) to configure permissions; AWS CloudTrail to audit requests made to the service; AWS Data Pipeline to schedule and start your clusters; AWS Lake Formation to discover, catalog, and secure data in an Amazon S3 data lake [14].

Azure HDInsight is a cloud distribution of Hadoop components. Azure HDInsight makes it easy, fast, and cost-effective to process massive amounts of data. HDInsight includes the most popular open-source Apache frameworks, such as: Hadoop, Spark, Hive with LLAP, Kafka, Storm, HBase and R (Fig. 5).

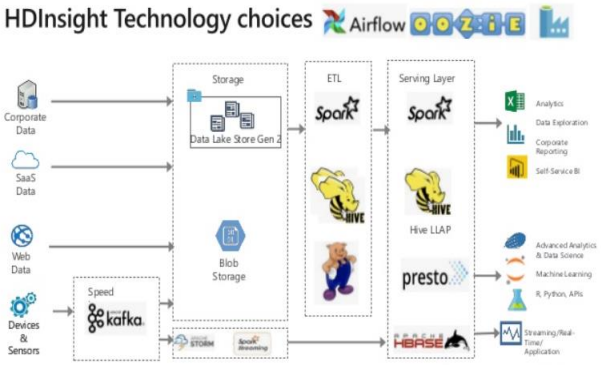


Fig. 5. Azure HDInsight Technology

Azure HDInsight advantages over on-premises Hadoop:

- Automated cluster creation requires minimal setup and configuration. Automation can be used for on-demand clusters.

- There's no need to worry about the physical hardware or infrastructure with an HDInsight cluster. Just specify the configuration of the cluster, and Azure sets it up.
- Azure takes care of data redistribution and workload rebalancing without interrupting data processing jobs.
- HDInsight is available in more regions than any other big data analytics offering.
- HDInsight enables you to protect your enterprise data assets with Azure Virtual Network, encryption, and integration with Azure Active Directory.
- A typical on-premises Hadoop setup uses a single cluster that serves many purposes. With Azure HDInsight, workload-specific clusters can be created. Creating clusters for specific workloads removes the complexity of maintaining a single cluster with growing complexity.
- You can use various tools for Hadoop and Spark in your preferred development environment.
- HDInsight clusters can be extended with installed components and can also be integrated with the other big data solutions by using one-click deployments from the Azure Market place.
- Azure HDInsight integrates with Azure Monitor logs to provide a single interface with which you can monitor all your clusters.
- HDInsight can easily be integrated with other popular Azure services such as the following: Data Factory (ADF), Blob Storage, Data Lake Storage Gen2, Cosmos DB, SQL Database, Analysis Services
- HDInsight constantly checks the infrastructure and open-source components using its own monitoring infrastructure. It also automatically recovers critical failures such as unavailability of open-source components and nodes. Alerts are triggered in Ambari if any OSS component is failed.

Snowflake Elastic Data Warehouse is an elastic system integrated with cloud storage on the Amazon Web Services platform, which can be expanded as needed as data, load, etc. increase

Snowflake provides the customer with a Data Warehouse as a Service. It is a high-performance column management system that supports standard SQL and meets ACID requirements. Data is accessed via Snowflake Web UI, Snowflake Client command-line interface, as well as ODBC and JDBC. The system consists of three components: Database Storage, Processing and Cloud Services.

The storage layer is responsible for the secure, secure and resilient storage provided by S3. Data is stored in S3, customers do not have direct access to it. To upload data to Snowflake, special S3 buckets (staging area) are created, where you need to put files, from which you can then upload data using Snowflake SQL. When loaded, the data is compressed (gzip) and converted to column format. Indexes in Snowflake are not provided. Data distribution is carried out automatically on the basis of statistics of their use. There is no partitioning. (Fig. 6).

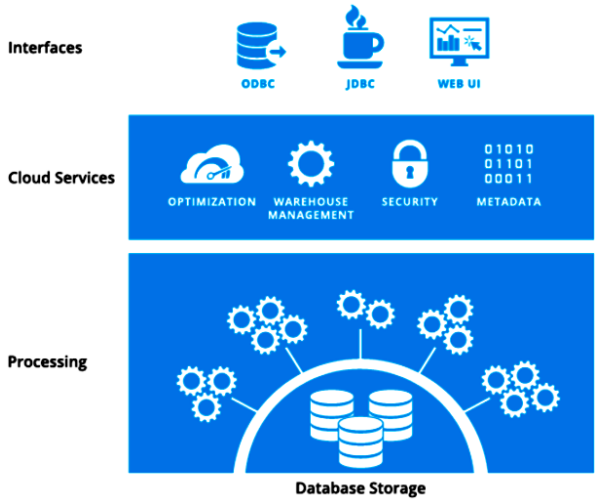


Fig. 6. Snowflake layers

Data is accessed through a data processing layer, a set of virtual servers that can access S3 files. A virtual cluster can consist of 1-16 virtual servers (EC2). All servers in the virtual cluster are the same and equivalent, the client does not have direct access to them. From the client's point of view, a virtual cluster is a whole. EC2 servers can be of two types: Standard and Enterprise. With Snowflake, it takes 1 to 5 minutes to create a new virtual cluster or expand an existing one. Data loss is not critical, in case of failure EC2 is simply re-created. SSD virtual servers are used as a cluster cache. Placing data in the cache speeds up queries up to 10 times. You can create multiple virtual clusters that will access the same data at the same time. This allows you to distribute the load.

Snowflake cloud services are used for data management. Using the Snowflake UI, the client creates databases, users and roles. Metadata is used at the data processing level (determining access rights, compiling queries, etc.). You can manage clusters: set the default cluster, specify the database to run, start the cluster on a schedule or when prompted, stop the cluster, change the number of servers. Snowflake provides installation, setup and maintenance. All you have to do is go to the site, create the tables, download the data, and run the query.

Modern systems and software stacks, (for example Apache Hadoop, for big data analytics are still complex to operate and far from perfect. As a result many organizations struggle to keep up with operating, optimizing and making their data infrastructure work to serve their data processing needs.

The complete data infrastructure solution includes many components among them:

Data Collection Service for both real time and bulk upload of data from different data sources such as applications, databases, web crawls etc.

Batch Computation Service such as Hadoop/Hive to process this data and transform it from data to information.

Real Time Computation Service to generate real time results on data streams and data captures for time sensitive and actionable reporting and monitoring.

AdHoc Query Service to answer one of queries sometimes exactly and other times approximately in a short amount of time.

Tools and Frameworks for job dependencies, data and query discovery, SLA and monitoring etc.

Qubole aims to provide all of the above components in the cloud. Qubole take care of optimizing, operating and evolving the data infrastructure for clients.

All data stored in S3. Adhoc Query and Batch Computation Service in the Cloud provides Apache Hive and Apache Hadoop as a service with close integration with Apache Oozie.

Hortonworks Data Platform (HDP) is open source Hadoop distribution that is based on a centralized architecture (Fig. 7).

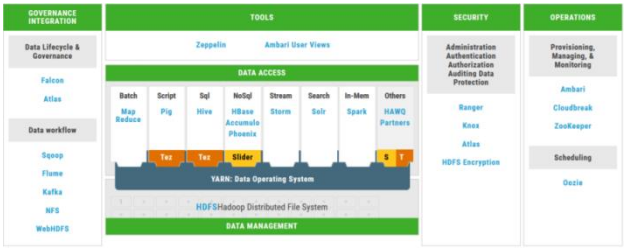


Fig. 7. Architecture Hortonworks Data Platform

HDP addresses a range of data-at-rest use cases, powers real-time customer applications, and delivers robust analytics that accelerate decision making and innovation.

The following sections describe the HDP components:

- Data management
- Data access
- HDP Operations
- Security and governance

Data management. The foundational components of HDP are Yet Another Resource Negotiator (YARN) and Hadoop Distributed File System (HDFS). While HDFS provides the scalable, fault-tolerant, cost-efficient storage for a Hadoop-powered Big Data lake, YARN provides the centralized architecture that enables organizations to process multiple workloads simultaneously.

YARN also provides the resource management and pluggable architecture to enable a wide variety of data access methods.

Data access. With YARN at its architectural center, HDP provides a range of processing engines that allow users to interact simultaneously with data in multiple ways. YARN enables a range of access methods to coexist in the same cluster against shared datasets to avoid unnecessary and costly data silos. HDP enables multiple data processing engines that range from interactive SQL, real-time streaming, data science, and batch processing. These processing engines leverage data that is stored in a single platform, unlocking an entirely new approach to analytics.

HDP Operations. HDP Operations enables IT organizations to bring Hadoop online quickly by taking the guesswork out of manual processes and replacing them with automated preconfigured best practices, guided configurations, and full operation control. Because of the rapid emergence of Hadoop, many users lack an optimal way to provision and operate the environment, leading them to waste time on inefficient troubleshooting, monitoring, and configuration. HDP Operations makes it easy to operate distributed multi-user, multi-tenant, and multiple data access engines and helps manage HDP clusters at scale through an integrated web UI or single pane of glass. Apache Ambari is an open source management platform for provisioning, managing, monitoring, and securing Hadoop clusters. Ambari removes the manual and often error-prone tasks that are associated with operating Hadoop. It also provides the necessary integration points to fit seamlessly into the enterprise and enables the IT operator to focus on delivering world-class service and support for their HDP consumers.

Security and governance. Hortonworks created Data Governance Initiative (DGI), a consortium of cross-industry leaders, to address the need for an open source governance solution to manage data classification, lineage, security, and data life cycle management. Apache Atlas, created as part of the DGI, empowers organizations to apply consistent data classification across the data ecosystem. Apache Ranger provides centralized security administration for Hadoop.

CONCLUSIONS

HDFS is designed to be deployed on low-cost hardware. It provides high throughput access to application data. HDFS has been designed to be easily portable from one platform to another. It has a master/slave architecture.

The primary objective of HDFS is to store data reliably even in the presence of failures.

Simplified architecture Databricks allows traditional analytics and data science to co-exist in the same system. The complete vision of lakehouse architecture deliver 9x better performance than traditional cloud data warehouses.

The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon EC2 instances. There are some benefits to using Amazon EMR? among them cost savings; AWS integration; deployment; scalability and flexibility; reliability; security; monitoring; management interfaces etc.

Azure HDInsight is a cloud distribution of Hadoop components. It makes it easy, fast, and cost-effective to process massive amounts of data.

The technology Snowflake is interesting, the ability to change the amount of resources on the fly looks very attractive. Cloud storage can be a decent option for new projects with limited data. But it's problem to transfer of all resources to this cloud.

Nowadays one of simplest and secure data lake platform for machine learning, streaming, and ad-hoc analytics was proposed by an Idera Inc. company and named Qubole.

HDP is the industry's only true secure, enterprise-ready open source Apache Hadoop distribution based on a centralized architecture (YARN). HDP includes a versatile range of processing engines that empower users to interact with the same data in multiple ways, at the same time. HDP extends data access and management with powerful tools for data governance and integration. Critical features for authentification, authorization, accountability and data protection are in place to help secure HDP across these key requirements. The users can integrate and extend their current security solutions.

REFERENCES

[1] (2021) Mlitz K. Forecast revenue big data market worldwide 2011–2027 [Online]. Available: https://www.statista.com/statistics/254266/global-big-data-market-forecast/

[2] Hajirahimova M. Sh., Aliyeva A. S. "Big Data initiatives of developed countries", *Problems of information society*, №1, pp. 10–15, 2017.

[3] (2020) Patrisio A. Top Big Data Companies. [Online]. Available: https://www.datamation.com/big-data/big-data-companies/

[4] (2021) Закон України про захист персональних даних [Online]. Available: https://zakon.rada.gov.ua/laws/show/2297-17#Text

[5] Bradlow E. T., Gangwar M., Kopalle P. & Voleti S. "The Role of Big Data and Predictive Analytics in Retailing", *Journal of Retailing*, 93(1), pp. 79–95, 2017.

[6] Chen S.-H., & Yu T. "Big Data in Computational Social Sciences and Humanities: An Introduction". *Big Data in Computational Social Science and Humanities*, pp. 1–25, 2018.

[7] Fernando Y., Chidambaram R. R. M. & Wahyuni-TD I. S. "The impact of Big Data analytics and data security practices on service supply chain performance". *Benchmarking: An International Journal*, 25(9), pp. 4009–4034, 2018.

[8] Gnizy I. "Big data and its strategic path to value in international firms". *International Marketing Review*, 36(3), pp. 318–341, 2019.

[9] Harrison-Walker L. J. & Neeley S. E. "Customer Relationship Building on the Internet in B2B Marketing: A Proposed Typology". *Journal of Marketing Theory and Practice*, 12(1), pp. 19–35, 2004.

[10] March Hofacker C. F., Malthouse E. C., & Sultan F. "Big Data and consumer behavior: Imminent opportunities". *Journal of Consumer Marketing*, 33(3), pp. 311–330, 2016.

[11] (2017) Big data text analytics: An enabler of knowledge management. [Online]. Available: https://doi.org/10.1108/JKM-06-2015–0238

[12] Kitchens B., Dobolyi D., Li J. & Abbasi A. "Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data". *Journal of Management Information Systems*, 35(2), pp. 540–574, 2018.

[13] Liu C., Yang C., Zhang X. & Chen J. "External integrity verification for outsourced big data in cloud and IoT: A big picture". *Future Generation Computer Systems*, 49, pp. 58–67, 2015.

[14] Liu X., Singh P. V. & Srinivasan K. "A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing". *Marketing Science*, 35(3), pp. 363–388, 2016.

[15] Mawed M. & Aal-Hajj A. "Using big data to improve the performance management: A case study from the UAE. *FM industry. Facilities*, 35(13–14, SI),pp. 746–765, 2017.

[16] Moorthy J., Lahiri R., Biswas N., Sanyal D., Ranjan J., Nanath K., & Ghosh P. "Big Data: Prospects and Challenges". *Vikalpa*, 40(1), pp. 74–96, 2015.

[17] Salehan M. & Kim D. J. "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics". *Decision Support Systems*, 81, pp. 30–40, 2016.

[18] Sanders N. R. "How to Use Big Data to Drive Your Supply Chain". *California Management Review*, 58(3), pp. 26–48, 2016.

[19] Szlezák N., Evers M., Wang J. & Pérez L. "The Role of Big Data and Advanced Analytics in Drug Discovery, Development, and Commercialization". *Clinical Pharmacology & Therapeutics*, 95(5), 492–495, 2014.

[20] Talón-Ballestero P., González-Serrano L., Soguero-Ruiz C., Muñoz-Romero S. & Rojo-Álvarez J. L. "Using big data from Customer Relationship Management information systems to determine the client profile in the hotel sector". *Tourism Management*, 68, pp. 187–197, 2018.

[21] Tan K. H. & Zhan Y. "Improving new product development using big data: A case study of an electronics company". *R&D Management*, 47(4), pp. 570–582, 2017.

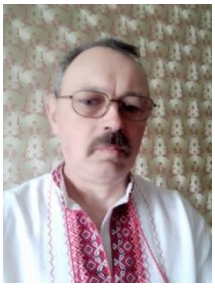
Стаття надійшла до редколегії 04.09.2021

Архітектура сучасних платформ для аналітики великих даних

Великі дані – це один із сучасних інструментів, який має великий вплив на світову індустрію, а також відіграє важливу роль у визначенні способів, якими підприємства й організації формулюють свою стратегію та політику. Проте кількість проведених наукових досліджень щодо прогнозування на основі великих даних є обмеженою через труднощі зі збиранням, обробленням та моделюванням неструктурованих даних, які зазвичай є конфіденційними. У статті великі дані розглядаються в контексті екосистеми для майбутнього прогнозування під час прийняття бізнес-рішень. Для однієї організації може бути важко володіти всіма необхідними даними, щоб розробити правильну стратегію. Ось чому різні організації створюватимуть та експлуатуватимуть власні аналітичні екосистеми або підключатимуться до існуючих. Екосистема аналітики, що складається із симбіозу даних, додатків, платформ, партнерств і сторонніх постачальників послуг, дозволяє організаціям бути більш гнучкими й адаптуватися до мінливих вимог сьогодення. Організації, які беруть участь в аналітичних екосистемах, можуть досліджувати, вчитися та впливати не лише на свої власні бізнес-процеси, а й на бізнес-процеси своїх партнерів. У цій публікації представлено архітектури популярних платформ для прогнозування на основі великих даних.

***Ключові слова**: великі дані, неструктуровані дані, платформи для аналізу даних, екосистеми даних, середовище великих даних.*

Людмила Зубик. Закінчила факультет кібернетики Київського національного університету імені Тараса Шевченка, кандидат технічних наук, доцент. Працює на посаді доцента кафедри програмних систем і технологій факультету інформаційних технологій Київського національного університету імені Тараса Шевченка, Україна.
Сфера наукових інтересів: педагогіка вищої школи, вебтехнології, штучний інтелект.
Liudmyla Zubyk. Graduated from the Cybernetics Faculty of Taras Shevchenko National University of Kyiv, PhD, Associate Professor. She works as an associate professor at the Software Systems and Technologies Department, Faculty of Information Technologies, Taras Shevchenko National University of Kyiv, Ukraine.
Research interests: high school pedagogy, web technologies, artificial intelligence.



Ярослав Зубик. Закінчив факультет кібернетики Київського національного університету імені Тараса Шевченка, старший викладач. Працює на кафедрі комп'ютерних наук і прикладної математики факультету кібернетики та комп'ютерних наук Національного університету водного господарства та природокористування, Рівне, Україна.
Сфера наукових інтересів: проблеми оптимізації, аналіз даних.
Yaroslav Zubyk. Graduated from the Cybernetics Faculty of Taras Shevchenko National University of Kyiv, Senior Lecturer. He works at the Department of Computer Sciences and Apply Mathematics, Institute of Automatics, Cybernetics and Computer Engineering, National University of Water and Environmental Engineering, Rivne, Ukraine.
Research interests: optimization problems, data analytics.