Integration of a Distributed Hadoop System into the Infrastructure of a Technology Startup Company

Dr. Stefka Petrova
University of Economics - Varna, Varna, Bulgaria
s.petrova@ue-varna.bg

PhD candidate Svetoslav Ivanov
University of Economics - Varna, Varna, Bulgaria
svetoslav_ivanov@ue-varna.bg

Abstract

The article reviews the problems and methods of combining big data processing work in a fully distributed mode of operation into the infrastructure of a technology startup company. In this direction, the methods of integrating an external distributed Hadoop system (cluster) are described in more detail. These methods are connected to the operating scenarios for infrastructure integration. The problems with traffic and latency are reviewed. A plan of the process of implementation and operation of a distributed Hadoop system is presented with the steps and indicative term. The goal it reaches is a suitable system for operational work on projects in a technology startup company. Then, it is connected to a list of measures in the different stages in a startup company's work. The problems with operational management of the teamwork with flexible methodology and system administration are listed, discussed and solutions proposed.

Keywords: distributed system, Hadoop, startup company, software infrastructure

JEL Code: C88, O32   DOI: https://doi.org/10.36997/IJUSV-ESS/2020.9.2.76

Introduction

When integrating an external distributed Hadoop system (cluster) into another existing Hadoop infrastructure, it is possible to use different approaches, depending on the objectives pursued. According to the Hadoop documentation, the operating modes of the system that are officially supported are three:

1. Local (stand-alone) mode. This is the default mode when installing on a computer, which does not use the HDFS file system, but the local file system of the computer when performing input/output operations. In this variant of using Hadoop, for the integration with existing infrastructure, it is necessary to make appropriate changes in the configuration files core-site.xml, hdfs-site.xml, and mapred-site.xml to move to a full distributed mode of operation (Apache, 2020a).

2. Pseudo-distributed mode. It is a single-node cluster that uses HDFS. In this mode, both the NameNode, which contains the list of all directories and files and the DataNode, where the data is stored, are located on the same computer. For the integration with existing infrastructure, it is necessary to make appropriate changes to the configuration files core-site.xml, hdfs-site.xml, and mapred-site.xml to switch to fully distributed mode (Apache, 2020b).

3. Fully distributed mode. It is a cluster with many nodes in which the data is distributed and processed in each of them. In this situation, the following configuration parameters are set in the main configuration files for connection with the other nodes in the cluster: File /etc/hadoop/conf/core-site.xml - HDFS settings; File /etc/hadoop/conf/hdfs-site.xml - settings for NameNode (and possibly Secondary NameNode) and multiple DataNode; File /etc/hadoop/conf/yarn-site.xml - settings for ResourceManager, NodeManager and History Server; /Etc/hadoop/conf/mapred-site.xml file - MapReduce settings.

Combined big data processing work (Petrivskyi, 2020; Bychkov, 2020; Baglaeva, 2019) in a fully distributed mode of operation of one cluster with another cluster can be performed in three main operating scenarios.

In the first scenario, the two clusters operate relatively independently of each other, with the data from one cluster being transferred (copied) for processing to the other, after which the result of the processing is eventually transferred back. The DistCp tool (Cloudera, 2019) can be used for this purpose when transferring large volumes of data.

In the second scenario, the two clusters merge statically to work as one large cluster and the resources of one distributed Hadoop system become part of the resources of the other Hadoop system (Ryu, 2018). Special care is required when setting network settings in terms of increasing security.

In the third scenario, the resources of the two clusters are dynamically pooled by specially designed systems that build on Hadoop (Wang et al., 2013; Jeon et al., 2014).

Each of the listed scenarios has its advantages and disadvantages, which makes it suitable for one situation or another. Each of the main scenarios may have different variations related to the details of a particular implementation.

1. Scenarios for Hadoop infrastructure integration

The scenario in which two clusters merge statically to work as a large "single" cluster and the resources of one distributed Hadoop system become part of the resources of the other Hadoop system sets certain requirements. In practice, the nodes of the new "single" Hadoop cluster are located in several data centers, geographically located in different places. It is necessary for all nodes in the cluster to be accessible through the network, i.e. their IP addresses to be available to other nodes in the network. There are several options for merging the networks of different geographically remote data centers:

1. To use a separate telecommunication line between the two networks for connection and thus to build a common LAN. This is a relatively expensive option in case you must pay rent for a leased line to a telecommunications company.

2. To use the public Internet for connection. It is possible to apply two main tactics:

a) All nodes must have public IP addresses - providing many IP addresses version 4 can be a problem, as their number is limited and scarce nowadays. The use of IP addresses version 6, where there are no problems with a shortage of IP addresses, is not possible at this stage, as it is not officially supported (HadoopIPv6, 2019). A significant problem is the protection of each node from various attacks specific to the Internet space. The guidelines provided by Cloudera (Ahluwalia, 2017) can be used for this purpose.

b) Use of virtual private networks - VPN (Virtual Private Networks). Virtual private network technology makes it possible to provide an encrypted connection between private networks by using a public network, such as the Internet, as the transmission medium. Possible options are the use of GRE (Generic Routing Encapsulation), PPTP (Point-to-Point Tunneling Protocol), L2F (Layer 2 Forwarding), L2TP (Layer 2 Tunneling Protocol), IPSec (Internet Protocol Security), and MPLS (MultiProtocol Label). Switching). Each of the options has its advantages and disadvantages in the implementation of the so-called. "Site-to-site VPN" and is recommended for use by various companies, such as Microsoft and Cisco.

One of the main problems in connecting several geographically remote clusters is that Hadoop generates a very large amount of traffic between nodes in the common cluster. Also, time latency reduces the speed of data processing, as there is more waiting between the individual nodes that work together. Time latency will very likely be the "bottleneck" in the cluster and limit overall performance (Petrov et al., 2020).

2. Planning the process of implementation and operation of a distributed Hadoop system

One of the main points in the implementation and operation of a distributed Hadoop system is a plan for the implementation of the information system that uses Hadoop. This includes planning the conversions, versions, and features that will be implemented in them, and allocating them over time (Panayotova et al., 2016; Dimitrov et al., 2016; Dimitrov et al., 2018; Kostadinova et al., 2018). Based on our experience and observation, we have developed an abbreviated version of a sample plan based on interaction scenarios. If necessary, it can be expanded and supplemented according to the needs and strategic views of the management.

The plan with the conversions is presented in Table 1. The set deadlines are indicative and are suitable for small to medium-sized organizations.

Table 1. Abbreviated project implementation plan.

Stage / Realization	Stage idea	Scenarios / characteristics included	Indicative term
Detailed design	Detail of project, 360° view, conventions.	Detailed design stage with project detailing based on the current project, concept, and selected approach	One week
Skeleton of the project	Skeleton of all components and minimum working site access, menu, and test editing of products.	Database with sample data and without optimization; Software project and physical components with class skeleton only; Server settings, added core components. Interface structure; User login; product review/definition;	Two weeks
Logic and data components	Development of data and business logic components	Product overview/definition; Data access; Advanced with new modules and playback modules;	Two weeks
Project and plan definition	All functions for defining project, content, process, and plan; testing components;	A site where a user can define a product, project, features, etc., ranking and evaluating features, selecting a process and scales for, defining conversions, allocating features, preparing a plan;	Three weeks
Work on tasks and testing	Planning and execution of iterations and tasks, testing, problems, defects	Defining iterations, tasks, list of risks and tasks on them, change of status, minimum plan and time report on them; defines problems, defects, conversion, defect handling and risks;	Three weeks
Detailed work report	Detail of work plan and report and iterations	Records time spent and remaining work on tasks. Personal status board. Reporting the cost of time and money. Notes and documentation for tasks, iterations, and all objects. Records conclusions from reflective improvement in iteration;	Two weeks
Completion of conversion	To be able to complete conversion and reporting. To form a reflective improvement.	Status reports and expected completion. Correction of estimates for iterations/conversions, the definition of realized products. Records conclusions from reflective improvement in conversion. Assigns tasks from them. Add customer requests, transfer in features.	Two weeks
Full Cycle and Administrati on	Clear for full-cycle development. Customer and employee administration.	Improved testing and notes, exporting a list of features for easy documentation. Tags to all objects. Improved time and resource reporting screens. Customers, representatives, user administration, settings, processes, scales, etc.	Three weeks

		Detailed data for employees - history, rates, skills, time calendar, etc.	
Negotiation	To be able to prepare an offer for a client and define a contract;	Adding a new contract or offer, determining the efforts for them, compliance with capacity. Preparation of an Offer, printing, defining a contract/plan with parameters; detailed realization/conversion time planning, and capacity compliance.	Two weeks
Reporting		Reporting component, standard performance reports, project completion. Prepares information boards and other reports.	Two weeks
Maintenance	Defining maintenance, monitoring, and maintenance work	Defining maintenance - coordinator/chief designer defines requirements, conditions, monitoring during maintenance; Tracking, and filling out support requests. Convert requests to defects, tasks, or update features. Description of implemented versions.	Two weeks
Final	Other extras; completion	The user describes an opinion on satisfaction with the implemented version and provided support. Defect reports, title corrected, maintenance	Two weeks

The plan is for about 26 weeks (or 6 months) and is subject to detail. The goal is to have a suitable system for operational work on projects in the first three months. When it is ready in three months, the team will be able to test the system with real data.

Regarding the implementation process, we have identified a set of basic steps and measures that must be performed when implementing the information system in a technology start-up company:

- Research and ensure that the technology start-up company and its team are ready to implement the approach;
- Development of an individual approach, process, techniques, work products in the company;
- Quick test play of the above to make sure they work;
- Hardware installation - servers, client stations, network, security, etc.
- Software installation - system and basic software, information system applications;
- Setting users, rights, process templates, scales, new options, etc.
- Input/transfer of data from current and test projects;
- Integration with existing software (if required)
- Play and test with the team in real conditions; final adjustments and adjustments;
- Approval of the implemented information system by the management;
- Enter all data for current projects and desired historical data;
- Protection of the environment and the information system, creation of a plan for archiving, and dealing with risks.

After implementation, each system requires appropriate measures before and during its operation. We have identified a set of measures for the use and maintenance of the information system. As our proposed approach to software development pays special attention to teamwork and the information system will work with team members, it should include measures to prepare the team and use the information system by him. These measures should also be included as part of the conventions (rules) for project work. For the former, we assume that they may include the following usual measures in the different stages of the individual roles:

- Administration and coordination: administrator with the help of chief designer and coordinator set up the information system, process templates, scales, and other options; administrator and coordinator provide each user with adequate access and data about employees, including qualifications, history, working hours, rates, etc .; coordinator and assignor take care of adequately defined products/projects;

- At the beginning of each project the assignment is reviewed according to the case and the project is planned: team/coordinator describes conditions, characteristics, and parameters; the team uses techniques to evaluate it; coordinator, assignor, and chief architect plan the realizations.

- Preparation of an offer, the definition of a contract, and first realization is planned in detail;

- Every day the team in the technology start-up company has the following obligations: after 15 minutes of the meeting everything is reflected from the meeting about the status of tasks/project, problems, and notes, personally or by a coordinator; review all outstanding news, problems, risks with, and others, process and appoint a person responsible for them; at the beginning of the work the beginning and the topic of the work are noted; at the end of the task and day the time worked and the expected time until the completion of the task(s) being worked on are recorded;

- At least once or twice a week: review the status of each project, all risks, problems, news; pending completion reports and forecasts are reviewed; in case of a problem (signaling in yellow or red, or other) a discussion is made about what to do and personal tasks are set to deal with it; in case of a significant problem with overtime work, delayed move, or necessary change of contract, part of the work is suspended until renegotiated; all pending news, problems, risks, defects, and other notes are reviewed, processed and a person responsible for them is appointed; The database is archived and stored on a separate server or media; there should be no unreported and unappreciated tasks.

- At the end of each iteration and realization: the product for realization is defined; the notes from the discussion with the client are recorded; reflected in the plan; iteration reporting, tasks, notes, characteristics, etc .; reports for presentation in the completed iteration/realization are displayed; the results of the reflective improvement are reflected; tasks are given; The database is archived on a technical medium and stored outside the office.

- At the end of each project: the incomplete reporting on the project is supplemented; describes realized products, version; financial statements; the results of the reflective improvement are reflected; tasks are given; if process and rule adjustments are needed, they are redefined; the project is completed and archived.

- When the information system cannot work: the team continues to work on the substance of the project; work products are used manually; the administrator corrects the information system immediately; the data from the working products are transferred to the information system shortly.

3. Problems of the operational management of the team working with the information system

Concerning the operational management of the team working with the information system, the importance of human resources and their effective management in the technology start-up company should not be underestimated, as well as the necessary qualities for working with flexible approaches. The ideology and spirit of flexible approaches are such that great weight is given to teamwork. As noted by some authors (Cockburn, 2005; Poppendieck, 2003), flexible approaches are not for everyone and require certain prerequisites and measures. It is necessary to clarify some considerations and requirements for the team of the technology start-up company and external participants, which must be considered.

First, participants must be prepared/appropriate to apply a flexible approach.

Second, people need to know the process and methodology. One technique for this is the so-called "Mini Process Training". It reproduces the process and methodology with the simplest sample product, possibly twice for 1-2 hours, so that the new members are aware of the methodology. The positive is if they can lose, along with that, and work with the information system.

Third, the information system must be known and usable. Although the entire team may be involved in the system, most of the work on the process and more complex functions concerns the roles of coordinator and chief architect. That is why their training is key. It is also crucial because they will take care of the team and guide it. Developers are expected to work side by side, have close communication, and use advice from others so that they can quickly learn on the spot or learn from each other. The system is designed to allow you to quickly receive brief help, as well as for instructions for specific options.

Fourth, users must have a minimum of technical skills to work with a client web application. It can be considered that in today's conditions this is not a problem for the internal team, and unlikely for external contractors/users. You only need a brief introduction to how to work with the application. If there is still a problem or users are not allowed into the information system, then the coordinator with appropriate communication with them could perform these tasks. The administrator and probably the coordinator need to be able to use the product administration and, possibly, to know the minimum administration of the MS SQL Server and IIS web server, which includes regular database backup, access settings, and optimization.

Fifth, the skills to define and write accurately, clearly, and concisely, and to avoid excessive formality are important. The information system and approach are designed as a simplified flexible tool and it is an unnecessary waste to overdo administrative work. The information system must be part of the process and used regularly and wisely so that the team can do its job successfully. Finally, dealing with stress and factors for it. For example, the presence of excessive workload, visibility of a too much-unfinished business, and more. Some simple measures in this regard are useful - the interface shows progress, the personal dashboard shows it and hides long lists over a certain size, focuses on work, avoids multitasking, and adding new features at least in the iteration, as well as rituals supporting progress.

4. Some problems with system administration

Regarding the administration of the system, we assume that it can be performed only by the roles of administrator and chief designer. The information system has an administrative part in the client application and the menus "Administration" and "Profile". Through the first we can do:

- user administration - adding, activating, changing data and password;
- administration of employees - complete data according to the model (Ivanov&Petrov, 2019; Petrov, 2008) for an employee and his history - contact details, qualification, remuneration, job parameters;
- assigning a user to a project with a specific role;
- defining a process template;
- defining a ranking scale;
- defining lists or adding options to expandable lists;
- correction of common errors - moving tasks between iterations, characteristics between implementations, problems between projects, log correction, and similar defects;
- logical archiving - exclusion from work of elements marked as archived and facilitated complete archiving of a project/product.

Through the "Account" menu the user can change his data and password.

In the course of operation of the information system, problems of various nature may arise - hardware, software, etc., and terminate its use. The possible risks and adequate countermeasures are the following:

- database crash - regular backup of MSSQL database and recovery procedure;
- server crash - replacement or relocation of applications with a new one; change settings; virtualization (Iliev et al., 2010);
- security breach - use of secure protocols - SSL, VPN, disabled caching, the security of

workstations and web server, use of strong passwords, etc. (Petrov, 2013);

- another system failure - work on projects can continue with templates of working products in files, and later be entered into the information system;

To successfully counteract the risks, it is necessary to regularly review the following measures: elaboration of a plan for dealing with risks regarding IP / approach; defining a special project for the technology start-up company and the process; regular documentation in the special project of problems, defects, risks and responsible setting of tasks for their solution; regularly reflecting improvements in risk plans and conventions.

The inclusion of risk management as a real project and its responsible perception in the form of tasks with direct responsibility is a prerequisite for more reliable work.

Users of the information system can be technology start-ups with a team size of 2 to 9-12 people. With a larger staff, the team for one project should be within the limit and there should be no dependencies between projects. One possible solution for a larger team working on one product is to divide it into smaller teams on separate projects with synchronized requirements (Poppendieck, 2003). For example, such projects may be about:

- common product components valid for all versions;
- development of separate versions of the product;
- a separate team to accompany the product;
- means for sales, automation, and customer service;
- individual implementations based on the main product.

Synchronization between projects can be done in the following several ways. The first is through a synchronized interface design. A similar idea was put forward by Poppendieck (2003) as a tool. A second idea is from the Scrum approach to hold regular daily meetings between the main designers of individual teams. There is a similar idea in the Lean approach, where to apply a tool for making such decisions based on multiple options. Also, it is possible to rotate team members in different areas to transfer experience, develop and gain diverse experience, diversify work, and more.

IP can be applied to different projects - external or own, with fixed or flexible parameters, with different criticality according to the Cockburn classification - categories C, D and possibly E with additional formal verification and validation procedures (Cockburn, 2005), security programming techniques (Petrov, 2017), etc. Application in L-category projects (life-threatening) is excluded. According to the author, it is possible to use in the development of intellectual products other than software, but only if the chosen approach is suitable for them. The more significant benefits of this are:

- visibility of the product and the project in one system;
- flexibility in development and other benefits of incremental adaptive approach;
- improvement of business processes;
- improving the quality of management through an advanced process;
- measuring efficiency with simplified reports;
- increasing the productivity and better efficiency of specialists;
- the simplified system of work;
- automation for most work products;
- formally built-in mechanisms for improvement, good quality, and customer satisfaction;
- good structuring of the staff and unloading from administrative work;
- means for distribution of work by burden over time;
- means for forecasting the observance of the deadlines;

Conclusion

The article reviewed the methods of combining big data processing work in a fully distributed mode of operation. In this direction, the methods of integrating an external distributed

Hadoop system (cluster) are presented and described in more detail. These methods are connected to the operating scenarios for infrastructure integration – separate telecommunication lines, public Internet connection, and VPN's. The problems with traffic and latency are reviewed. Accordingly, a plan of the process of implementation and operation of a distributed Hadoop system is presented with the steps and indicative term for about 6 months. The goal it reaches is a suitable system for operational work on projects. After the implementation, a list of measures in the different stages is defined per role in a start-up company. Finally, problems with operational management of the teamwork with flexible methodology and system administration are listed, discussed and solutions proposed.

References

1. Ahluwalia, M., (2017). How to secure "Internet exposed" Apache Hadoop, [Online] Available from: https://blog.cloudera.com/how-to-secure-internet-exposed-apache-hadoop/ [Accessed 10/10/2020]
2. Apache, (2020a). Hadoop 3.3.0: Setting up a Single Node Cluster. [Online] Available from: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html [Accessed 10/10/2020].
3. Apache, (2020b). Hadoop 3.3.0: Hadoop Cluster Setup. [Online] Available from https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html [Accessed 10/10/2020].
4. Baglaeva, E., Buevich, A., Sergeev, A., et al. (2019). Statistical Characteristics Calculation of the Natural Dust Size Distribution in the Air Surface Layer of Belyy Island. In Proceedings of the International Conference of Computational Methods in Sciences and Engineering 2019 (ICCMSE-2019), AIP, 2186. https://doi.org/10.1063/1.5137945
5. Bychkov, O., Merkulova, K., Dimitrov, G., et al. (2020). Using Neural Networks Application for the Font Recognition Task Solution. In 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST 2020), Nis, Serbia, IEEE, pp.167-170. https://doi.org/10.1109/ICEST49890.2020.9232788
6. Cloudera, (2019). Administering HDFS, Using DistCp to Copy Files, [Online] Available from https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/administration/content/using-distcp-to-copy-files.html [Accessed 10/10/2020].
7. Cockburn, A. (2005). Crystal Clear: A Human-Powered Methodology for Small Teams, Addison-Wesley, Second Edition.
8. Dimitrov, G., Panayotova, G., Garvanov, I., et al. (2016). Performance analysis of the method for social search of information in university information systems. 3rd International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland, IEEE, pp.149-153. doi:10.1109/ICAIPR.2016.7585228
9. Dimitrov, G., Panayotova, G., Kovatcheva, E., et al. (2018). One Approach for Identification of Brain Signals for Smart Devices Control. Journal of Software, 13(7), pp.407-413.
10. HadoopIPv6, (2019). Hadoop and IPv6 https://cwiki.apache.org/confluence/display/HADOOP2/HadoopIPv6
11. Iliev, P., Salov, V., & Petrov, P. (2010). Virtualni sistemi. Varna: Nauka i ikonomika,
12. Ivanov, S., & Petrov, P. (2019). Business Models for Starting Software Companies. In 9th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE-2019), pp.170-182.
13. Jeon, K., Chandrashekhara, S., Shen, F., Mehra, S., Kennedy, O., & Ko, S. Y. (2014). Pigout: Making multiple hadoop clusters work together. In 2014 IEEE International Conference on Big Data (Big Data) pp.100-109.
14. Kostadinova, I., Toshev, R., et al. (2018). Temporal Analysis of the Pedagogical Adoptions use and Application of the Augmented and Virtual Reality Technologies in Technical Subject

Areas. In 11th Annual International Conference of Education, Research and Innovation, ICERI2018 Proceedings, Seville, Spain: IATED, pp.4387-4393.

15. Panayotova, G., Dimitrov, G., et al. (2016). Modeling and data processing of information systems. 3rd International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland, IEEE, pp.154-158. doi:10.1109/ICAIPR.2016.7585229
16. Petrivskyi, V., Dimitrov, G., Shevchenko, V. et al. (2020). Information Technology for Big Data Sensor Networks Stability Estimation. Information and Security, 47(1), pp.141-154. https://doi.org/10.11610/isij.4710.
17. Petrov, P. (2008). Sarvarno programirane. Varna: Nauka i ikonomika.
18. Petrov, P. (2013). Trends In The Use Of Web Server Software In Bulgarian Banks. In International Conference On Application Of Information And Communication Technology And Statistics In Economy And Education (ICAICTSEE-2012), Conference Proceedings, UNWE, pp.359-364.
19. Petrov, P. (2017). *Obektnoorientirano programirane*. Varna: Nauka i ikonomika.
20. Petrov, P., Dimitrov, G., Bychkov, O. (2020). Real Time Big Data Analysis by Using Apache Kudu and NoSQL Redis in Web Applications. *Izvestia Journal of the Union of Scientists - Varna. Economic Sciences Series, 9*(1), pp.26-34. DOI: https://doi.org/10.36997/IJUSV-ESS/2020.9.1.26
21. Poppendieck, M., & Poppendieck, T. (2003). *Lean Software Development: An Agile Toolkit*. Addison-Wesley.
22. Ryu, W. (2018). Implementation of dynamic node management in Hadoop cluster. *International Conference on Electronics, Information, and Communication (ICEIC)*, Honolulu, HI, 2018, pp. 1-2, doi: 10.23919/ELINFOCOM.2018.8330612.
23. Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., & Chen, D. (2013). G-Hadoop: MapReduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems, 29*(3), pp.739-750.